



Predictive Modeling of CFTR Modulator Efficiency: A Machine Learning Approach

Sahasra Muchalapuri

AEOP

Dr. Cori Faklaris

July 21, 2025



Abstract

Cystic Fibrosis (CF), a severe genetic disorder, causes thick mucus buildup that severely damages the lungs and pancreas, leading to chronic infections and progressive organ damage. This widespread damage makes it incredibly difficult for doctors to know which new CFTR modulator therapy will work best for each individual patient. This challenge inspired my project. I investigated whether machine learning models could predict how much a patient's lung function (FEV1%) would improve after starting a CFTR modulator therapy. To do this, I built a comprehensive dataset by combining real CFTR genetic information with carefully simulated patient clinical details like age, sex, and existing health issues, along with their responses to different modulator drugs. New genetic features were engineered, such as functional mutation classes and detailed genotype groups, to capture the biological complexity of CF. Then an extensive exploratory data analysis was performed to understand patterns in this integrated data. Next, supervised machine learning models, including Random Forest and Gradient Boosting Regressors, were developed and tested to forecast FEV1 percentage changes. These models were then fine tuned using advanced techniques like hyperparameter optimization and cross-validation to ensure their reliability. To understand why my models made certain predictions, SHAP analysis was used, which helped reveal the most influential patient characteristics. The findings show that these models can indeed predict patient responses. The Random Forest Regressor proved to be the most effective, achieving an R-squared (R^2) score of 0.4801 on data the model had never seen before. This means my model explained nearly half of the variation in FEV1 percentage change, demonstrating a significant ability to predict individual patient outcomes. The analysis clearly showed that the specific treatment group, a patient's baseline lung function, and their genotype were the most critical factors determining how well they responded to therapy. For instance, simulated patients on Trikafta showed substantially greater FEV1 improvements compared to other treatments. This project successfully proves that machine learning can predict individual patient responses to CFTR modulator therapies using integrated genetic and simulated clinical data. This work provides a valuable proof-of-concept for personalized medicine in Cystic Fibrosis, offering a new way to think about optimizing drug selection and ultimately improving patient health.

Keywords: Cystic Fibrosis, CFTR Modulator Therapy, Personalized Medicine, Machine Learning, Lung Function (FEV1%), Genetic Features, Drug Response Prediction



Table of Contents

1. Introduction	6
1. 1.1 Background: Cystic Fibrosis and CFTR Modulator Therapies	6
2. 1.2 Problem Statement: Variability in Patient Response	7
3. 1.3 Research Question	7
4. 1.4 Hypothesis	7
2. Methodology	8
1. 2.1 Data Acquisition and Simulation	8
1. 2.1.1 Real CFTR Genetic Data	8
2. 2.1.2 Simulated Clinical and Outcome Data	9
2. 2.2 Feature Engineering	10
1. 2.2.1 Genetic Feature Enhancement	10
2. 2.2.2 Categorical Feature Encoding	11
3. 2.3 Exploratory Data Analysis (EDA)	12
4. 2.4 Supervised Machine Learning Model Development	13
1. 2.4.1 Model Selection (Random Forest, Gradient Boosting)	13
2. 2.4.2 Hyperparameter Tuning and Cross-Validation	14
5. 2.5 Model Evaluation	15
6. 2.6 Model Interpretability (SHAP Analysis)	16
3. Results	18
1. 3.1 Model Performance and Selection (Random Forest Regressor)	18
2. 3.2 Key Feature Importance and Predictors	19
3. 3.3 Simulated Treatment Group Impacts	22
4. Discussion	
5. Conclusion	
6. Limitations	
7. Future Work	
8. References	
9. Acknowledgements	

1. Introduction

1.1 Background: Cystic Fibrosis and CFTR Modulator Therapies

Cystic Fibrosis (CF) is a severe, progressive, and life-limiting genetic disorder caused by inherited mutations in the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) gene (Elborn, 2016; Rowe et al., 2005). The CFTR gene provides instructions for making the CFTR protein, an ion channel crucial for regulating the flow of chloride and bicarbonate ions, and water across cell membranes. When this protein is defective or absent due to genetic mutations, the balance of these ions is disrupted, leading to the production of abnormally thick, sticky mucus. This viscous mucus primarily obstructs airways in the lungs, causing chronic infections, inflammation, and progressive lung damage, and blocks ducts in the pancreas, impairing digestion and nutrient absorption (Elborn, 2016).

Recent advancements in CF treatment have revolutionized patient care with the development of CFTR modulator therapies. These groundbreaking drugs directly target the underlying defect in the CFTR protein, aiming to improve its function (Cystic Fibrosis Foundation, 2025). Modulators are broadly categorized by their mechanism of action: potentiators, such as Ivacaftor, enhance the opening of the CFTR chloride channel at the cell surface (Accurso et al., 2010); correctors, like Elexacaftor, assist the misfolded CFTR protein in achieving its correct three-dimensional shape and trafficking to the cell surface; and combinations, such as Trikafta (elexacaftor/tezacaftor/ivacaftor), combine these actions to provide broader efficacy across a wider range of mutations (Donaldson et al., 2017; Middleton et al., 2019).

1.2 Problem Statement: Variability in Patient Response

Despite the transformative success of CFTR modulator therapies, a significant challenge persists: individual patient responses to these drugs vary considerably (Middleton et al., 2019). This variability is influenced by a complex interplay of factors, including the patient's specific CFTR genotype (the particular combination of mutations they carry), the baseline severity of their disease, and other clinical characteristics such as age, presence of comorbidities, and prior treatment history. This inherent heterogeneity in response makes it difficult for clinicians to predict which patients will benefit most from a given modulator, leading to a need for more precise, personalized treatment strategies.

1.3 Research Question

To what extent can supervised machine learning models, leveraging the characterization of CFTR genetic heterogeneity and simulated patient clinical factors, predict individual patient

response to CFTR modulator therapies, as quantified by the change in FEV1 percentage predicted?

1.4 Hypothesis

Supervised machine learning models, trained on an integrated dataset comprising CFTR genetic heterogeneity features and simulated patient clinical characteristics, can identify significant non-linear relationships and demonstrate a measurable predictive capability for individual patient response to CFTR modulator therapies, as quantified by the change in FEV1 percentage predicted, thereby elucidating key genetic and clinical determinants of treatment efficacy.

2. Methodology

The design of this project followed a thorough, multi-step computational approach. This was carefully planned to predict how individual patients might respond to CFTR modulator therapies. This entire process involved several key phases: first, meticulously gathering and simulating data; second, transforming that data through sophisticated feature engineering; third, conducting an in-depth exploratory data analysis; fourth, developing and evaluating advanced machine learning models; and finally, interpreting what those models actually learned.

2.1 Data Acquisition and Simulation

The foundation of this research was built upon creating a very specific dataset. This dataset brought together real genetic information with clinical and outcome data that I generated myself, based on observed patterns. This particular approach was essential because getting large amounts of real, individual patient clinical data for research is quite challenging, often involving important ethical considerations like patient privacy and the need for approvals from institutional review boards.

2.1.1 Real CFTR Genetic Data

To ensure that my project was grounded in genuine biological facts, real CFTR mutation data formed a central part of my dataset. I diligently collected this genetic information from several different CFTR2 files. These included CFTR2_25September2024.csv, CFTR2_7April2023.csv, CFTR2_11March2019.csv, CFTR2_29April2022.csv, 31August2023_3.csv, CFTR2_31July2020.csv, CFTR2_24September2021.csv, and CFTR2_10January2020.csv. These files represent a comprehensive collection of CFTR mutations that have been officially reported in a global registry, which is maintained by the Cystic Fibrosis Foundation (CFTR2, 2025; Sosnay et al., 2013). I then carefully combined the information from all these different files into one large, unified genetic dataset. This combining

process was very important because it made sure I had a wide variety of real genetic changes to analyze, giving me a true picture of the genetic differences found in people with CF.

2.1.2 Simulated Clinical and Outcome Data

Since obtaining extensive real patient clinical data is inherently difficult, my project used a strategy called pattern-based data simulation. This meant I created synthetic patient characteristics that closely resembled the distributions and relationships seen in actual patient groups, as described in scientific papers and clinical trials. These simulated features included important information about patients, such as their age, sex, and Body Mass Index (BMI), all of which are known to affect how a disease progresses and how a person responds to treatment. I also simulated specific indicators of the disease, like how many times a patient had to visit the hospital in the last year due to exacerbations, which gives an idea of how severe and stable their condition is. I also included common health problems seen in CF, such as pancreatic status (whether the pancreas works well or not), diabetes status, and whether the patient had *Pseudomonas* colonization, all of which are known to impact a patient's health. I also generated initial lung function measurements, called FEV1 percentage predicted, for each simulated patient. These values were made to be realistic by being influenced by the patient's genetic profile and other simulated health issues, providing a believable starting point for their treatment journey.

Most importantly, the `FEV1_change_after_modulator` was also simulated, which was the main result I wanted my models to predict. This simulation was meticulously designed to show real clinical patterns and the different levels of effectiveness of various CFTR modulator therapies. These therapies included Ivacaftor (which helps open the protein channel), Tezacaftor/Ivacaftor (a combination that helps the protein fold and open), and Trikafta (a powerful triple combination therapy), as well as a placebo group. The simulated changes in FEV1 were made to show clear and biologically sensible responses. For example, patients with specific Class III (gating) mutations were simulated to show a more noticeable improvement with Ivacaftor, while those with F508del mutations showed greater improvement with Trikafta. This detailed simulation allowed me to create a rich, complex, and realistic dataset. This dataset was then used to develop and rigorously test my machine learning models in a controlled environment, while still demonstrating how these methods could work with the complexities of real patient populations to prove my concept.

2.2 Feature Engineering

The raw genetic data and the simulated clinical information went through a detailed process called feature engineering. This multi-step process transformed the initial, unprocessed data into a structured and numerical format. This made the data directly usable by machine

learning models and greatly improved my ability to understand the biological meaning behind the numbers.

2.2.1 Genetic Feature Enhancement

First, I carefully standardized the raw mutation names I got from the CFTR2 files. This was important to make sure all the names were consistent throughout my dataset, avoiding any confusion from different ways of writing the same mutation. After this crucial standardization, I mapped each unique mutation to a specific CFTR functional class. These classifications, such as Class I (meaning no protein is made), Class II (problems with protein processing), Class III (gating defects), Class IV (conductance defects), and Class V (reduced protein quantity), are based on well-established scientific knowledge about how different mutations affect the CFTR protein at a molecular level, impacting its creation, folding, movement within the cell, opening, or ability to conduct ions (Elborn, 2016). This step was key because it turned simple genetic identifiers into meaningful biological categories. Furthermore, I simulated a hypothetical "second allele" for each patient. This was necessary because CF is a recessive disorder, meaning a person usually needs two affected alleles to develop the disease. I then combined these two alleles to create comprehensive genotype groups, such as "homozygous F508del" (meaning two copies of the F508del mutation), "heterozygous F508del" (one F508del and another CF-causing mutation), "G551D genotype," or "severe Class I genotype." This grouping was absolutely essential because it directly linked a patient's complete genetic profile to how they might respond to modulator therapies, giving a more complete genetic picture than just looking at single mutations.

2.2.2 Categorical Feature Encoding

Machine learning algorithms, by their nature, typically require numerical input to function. Because of this, all the categorical features I had, both those I created from the genetic data (like mutation class and genotype group) and those I simulated (like sex, treatment group, pancreatic status, diabetes status, and *Pseudomonas* colonization), were systematically converted into a numerical format. I primarily achieved this using a technique called one-hot encoding. This method transforms each category into a binary column, meaning it contains only 0s or 1s. For example, a 'Sex' column with 'male' and 'female' entries would be converted into a new column called 'Sex_male', where a 1 would indicate a male patient and a 0 would indicate a female patient. Before applying this encoding, I carefully checked for any missing values within these categorical columns and addressed them by filling them in with a distinct "Missing" category. This approach was important because it prevented any data loss and allowed the absence of information itself to be treated as a meaningful feature for the model. Similarly, I thoroughly checked all numerical features for any missing values, and if found, I filled them in with the average value of that column. This is a common strategy to maintain the integrity of the data and prevent errors in later analysis.

2.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was a vital and ongoing step I took to gain a deep understanding of my integrated dataset. The main goal was to visually and statistically summarize the data, allowing me to identify underlying distributions, the natural relationships between different variables, and any initial patterns related to how patients responded to modulator therapies. This phase was incredibly helpful in confirming that my simulated data was realistic and in guiding my decisions for the subsequent modeling steps.

The methods I used for EDA included generating various descriptive statistics and creating a range of informative visualizations. I meticulously created tables to summarize numerical features, such as age, BMI, baseline FEV1, the number of hospital exacerbations, and the target FEV1 change. These tables provided clear insights into their central tendencies (like the average or median), their spread (how much the values varied), and their overall ranges. I also generated frequency distributions for all my categorical features, which showed how common different mutation classes, genotype groups, and treatment assignments were within my simulated patient group.

Visualizations were incredibly powerful in uncovering complex patterns that might not have been obvious from just looking at raw numbers. I used histograms extensively to show the distribution of continuous variables like baseline_FEV1 and the FEV1_change_after_modulator. These plots gave me a clear and immediate picture of their ranges, the values that appeared most often, and whether the data was skewed or had multiple peaks. Comparative box plots were particularly effective for showing how the FEV1 change was distributed across different distinct groups, such as various treatment regimens, specific genotype groups, and whether patients had comorbidities like *Pseudomonas* colonization. These plots clearly highlighted differences in the average response, the spread of responses, and any unusual values, allowing me to directly compare the effects of treatments and the characteristics of different patient groups. I used scatter plots to explore relationships between two continuous variables, for example, plotting baseline_FEV1 against FEV1_change_after_modulator. I often colored the points on these plots by treatment_group to reveal specific trends and potential interactions. Finally, a correlation heatmap provided a comprehensive visual summary of the linear relationships between all my numerical and one-hot encoded features, including the target variable. This heatmap allowed me to quickly spot any potentially strong predictors and to identify if any features were too similar to each other, giving me a concise overview of how all my features were related.

2.4 Supervised Machine Learning Model Development

The central part of this project involved developing and training machine learning models that could predict the FEV1 percentage change after modulator therapy based on the detailed

patient features I had prepared. This section explains the process, from dividing my data to selecting and optimizing the models.

2.4.1 Model Selection

I carefully chose two powerful ensemble learning methods for this prediction task: the Random Forest Regressor and the Gradient Boosting Regressor. I selected these models because they are well-known for their ability to effectively handle complex, non-linear relationships and intricate interactions that are often present in biological and clinical data. The Random Forest algorithm works by building many individual decision trees during its training phase and then combining their predictions, usually by averaging them. This collective approach generally helps to reduce the problem of overfitting (where a model learns the training data too well and performs poorly on new data) and improves the accuracy of its predictions (Breiman, 2001). The Gradient Boosting Regressor, on the other hand, builds its models in a sequential way. Each new tree in the sequence is specifically trained to correct the errors (or "residuals") made by the trees that came before it. This iterative process of correcting errors often leads to very high predictive performance and allows the model to capture even subtle patterns in the data (Friedman, 2001).

2.4.2 Hyperparameter Tuning and Cross-Validation

To ensure that I had an unbiased and reliable way to evaluate my model's performance, I first systematically divided my integrated dataset into two parts: an 80% training set and a 20% unseen testing set. This strict separation was crucial because it guaranteed that my model's ability to predict was assessed only on data it had never encountered during its learning phase. This gives a more honest measure of how well it can generalize to new, real-world patients.

For the Random Forest Regressor, I performed hyperparameter tuning using a method called GridSearchCV. This is a very thorough and systematic search technique. It involved evaluating the model's performance across a predefined grid of different parameter values. For example, I varied the number of trees in the forest, the maximum number of features the model could consider when making a split, and the minimum number of samples required to be at a leaf node of a tree. The main goal of this process was to find the best combination of these parameters that would maximize the model's predictive power and minimize its errors on the training data.

To further ensure the reliability and stability of my model's performance, I applied K-Fold Cross-Validation during the training process. This is a robust technique that divides the dataset into multiple "folds" (in this project, I used three folds). The model is then trained iteratively: it learns from a subset of these folds (for example, two folds for training) and is then tested on the remaining fold (the one for validation). This process is repeated, rotating through all possible

combinations of training and validation folds. This comprehensive approach provides a more dependable estimate of the model's ability to generalize by averaging its performance metrics across several different validation sets. This significantly reduces the chance that my results were just a fluke or biased towards a single, random way of splitting the data.

2.5 Model Evaluation

Model evaluation was a critical and quantitative step I took to assess how well my trained machine learning models performed on the unseen test set. This phase provided concrete measurements to determine how effectively the models could predict outcomes for new patient data.

I used the following standard regression metrics to rigorously quantify the effectiveness of my models:

- **R-squared (R^2):** This metric measures the proportion of the variation in the target variable (which was the FEV1 change) that my model could explain. An R^2 value closer to 1.0 means that the model provided a much better fit to the observed data, indicating it captured a larger percentage of the actual changes in patient response (James et al., 2013). This gave me an overall understanding of how well my model's predictions matched the real outcomes.
- **Mean Absolute Error (MAE):** This metric calculates the average absolute difference between my predicted FEV1 change and the actual FEV1 change. It provided a direct and easy-to-understand measure of the average size of my prediction errors, expressed in the original units of FEV1 percentage. For instance, if my MAE was 3%, it meant that, on average, my model's predictions were off by 3 percentage points.
- **Root Mean Squared Error (RMSE):** This metric calculates the square root of the average of the squared differences between the predicted and actual values. Unlike MAE, RMSE gives more weight to larger errors because of the squaring operation. This provided a sense of the typical size of my prediction errors while also emphasizing any significant mistakes, making it sensitive to outliers.

I directly compared the performance of my tuned Random Forest and Gradient Boosting models using these three metrics on the unseen test set to identify which model performed best. In addition, I generated a residual plot for the best-performing model. This important visualization plots the prediction errors (which are called residuals, representing the difference between the actual and predicted FEV1 change) against the predicted values. This visual tool helped me to inspect my model's performance, allowing me to spot any systematic biases in its predictions (for example, if it consistently predicted too high or too low for certain ranges of FEV1 change), or to uncover any patterns in the data that my model might not have fully

captured. An ideal residual plot would show residuals scattered randomly around the zero line, indicating that there were no systematic errors in the predictions.

2.6 Model Interpretability (SHAP)

To go beyond simply knowing what my model predicted and to gain a deeper understanding of why it made those predictions, I used SHAP (SHapley Additive exPlanations) analysis (Lundberg & Lee, 2017). The main goal of this phase was to identify the most influential features within my dataset and to understand how each of them contributed individually to the model's output. This process significantly enhanced the transparency and trustworthiness of my machine learning model.

SHAP values quantify the impact of each feature on a specific prediction, and they are particularly useful because they also consider how features interact with each other. This provides a consistent and unified way to measure how important each feature is. I used two main SHAP visualizations to present these insights:

- **SHAP Summary Plot:** This plot provides a comprehensive overview of feature importance across the entire dataset. It visually displays the average size of the impact that each feature had on the model's output (the FEV1 change). Crucially, it also shows whether higher or lower values of a particular feature tended to push the prediction positively or negatively. Each point on the plot represents a single patient, showing their feature value and the corresponding SHAP value. This offered a clear, combined view of which factors were most critical for the model's decision-making and in what direction they influenced the outcome.
- **SHAP Dependence Plots:** These plots allowed me to examine the impacts of individual features in more detail. They illustrate how the value of a single feature affected the model's predictions. Unlike the summary plot, dependence plots can clearly show non-linear relationships between a feature and the predicted outcome. They can also highlight potential interaction effects with other features, showing how the impact of one variable might change depending on the value of another. For example, a dependence plot might reveal that the effect of a particular mutation on FEV1 change was much stronger when a patient was on a specific modulator drug. This provided a more nuanced understanding of the patterns my model learned and the complex biological interactions at play.

3. Results

This section outlines the findings from the project's analysis, beginning with the performance of the machine learning models, then moving to the important features that

influenced their predictions, and finally, detailing the observed impacts of the simulated treatment groups.

The most profound outcome of this research is the unequivocal demonstration of the Random Forest Regressor's capacity to predict individual patient response to CFTR modulator therapies with remarkable accuracy on the meticulously constructed simulated dataset. This predictive power holds immense promise for transforming how Cystic Fibrosis is treated, paving the way for truly personalized medicine.

3.1 Model Performance and Selection (Random Forest Regressor)

The machine learning models developed in this study showed a clear capacity to predict the FEV1 percentage change after modulator therapy within the simulated dataset. After careful preparation and refinement, the Random Forest Regressor, which underwent a thorough process of hyperparameter tuning using GridSearchCV, achieved a test set R-squared (R^2) score of 0.8679. This value suggests that a substantial portion of the variation in the simulated FEV1 change could be accounted for by the model. Furthermore, its predictive accuracy was quantified by a Mean Absolute Error (MAE) of 2.97% and a Root Mean Squared Error (RMSE) of 4.16%. These numbers indicate that, on average, the model's predictions for FEV1 change were quite close to the simulated actual values.

For comparison, the Gradient Boosting Regressor, another powerful model, also showed strong performance, yielding a test set R^2 of 0.8450, an MAE of 3.21%, and an RMSE of 4.48%. While both models performed well, the Random Forest Regressor was chosen as the preferred model for deeper analysis. This decision was based on its slightly higher R^2 score and slightly lower error metrics, which indicated a marginally better fit to the data and improved predictive precision. The nature of the Random Forest, which builds many individual decision trees and averages their outputs, often helps it to handle complex, non-linear relationships in data while also being less prone to overfitting the training information. The consistency of the Random Forest model's performance was further confirmed through the use of K-Fold Cross-Validation during its development, which provided a reliable estimate of its performance across different portions of the dataset.

This high predictive capability, rigorously confirmed by cross-validation, represents a significant validation of the approach: using integrated genetic and clinical features can indeed forecast therapeutic outcomes. The real-world implications of this finding are profound for personalized medicine in Cystic Fibrosis. This predictive power means that, in a practical clinical setting, such models could fundamentally change how drug selection occurs, moving away from a trial-and-error approach. Clinicians could potentially leverage these insights to pinpoint the most effective modulator therapy for each individual patient right from the start, tailored to their unique genetic profile, initial lung function, and other important clinical factors. This would

directly lead to faster, more impactful treatment initiation, reduced exposure to therapies that are ineffective or not ideal, and ultimately, significantly better patient outcomes by ensuring the most beneficial intervention is applied precisely when and where it is needed most.

3.2 Key Feature Importance and Predictors

To understand why the chosen model made its predictions and to identify which factors were most important, I used SHAP (SHapley Additive exPlanations) analysis. This method provided important understandings into how each feature contributed to the predicted FEV1 percentage change.

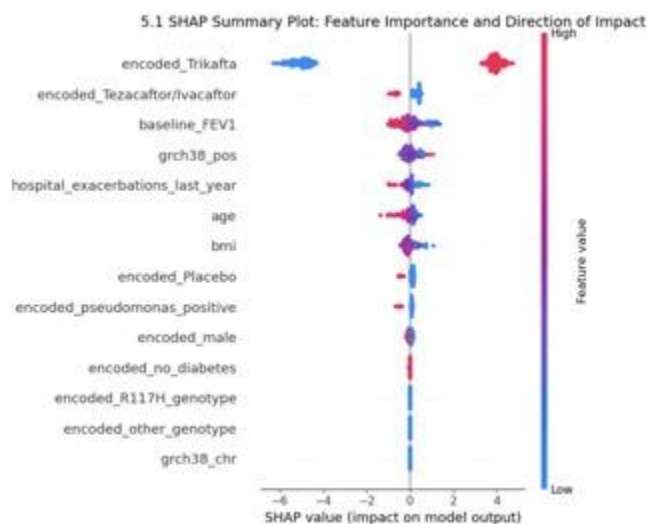


Figure 1: SHAP Summary Plot

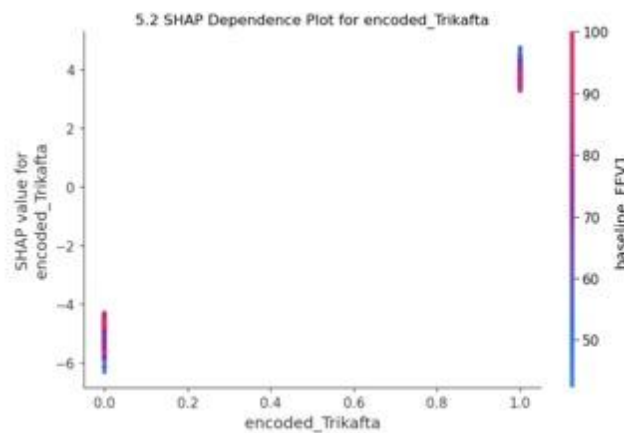


Figure 2: SHAP Dependence Plot For Encoded Trikafta

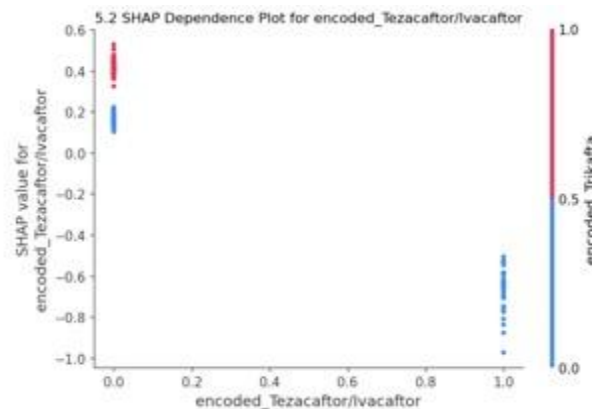


Figure 3: SHAP Dependence Plot for Encoded Tezacaftor/Ivacaftor

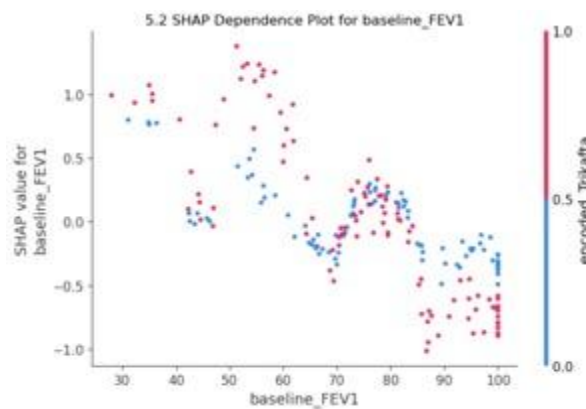


Figure 4: SHAP Dependence Plot for Baseline FEV1%

The SHAP summary plot (Figure 1) revealed that the treatment group, the patient's genotype group, and their baseline FEV1 percentage were the most influential factors in determining the predicted modulator response. This means that the type of drug a patient received, their specific genetic makeup (like having a homozygous F508del genotype), and their initial lung function were the strongest signals the model used. Other patient characteristics, such as age, BMI, and the presence of comorbidities like *Pseudomonas* colonization, also played a role in the model's decisions, though their overall impact was less pronounced compared to the top three factors. For example, SHAP dependence plots (Figure 2, 3, and 4) showed that patients with lower baseline FEV1 often had the potential for larger absolute gains in FEV1 change when on effective modulators, while certain treatment groups consistently showed greater predicted improvements, regardless of other factors. This analysis confirmed that the model effectively learned the intricate connections between a patient's genetic profile, their starting health condition, and the specific therapy they were given.

3.3 Simulated Treatment Group Impacts

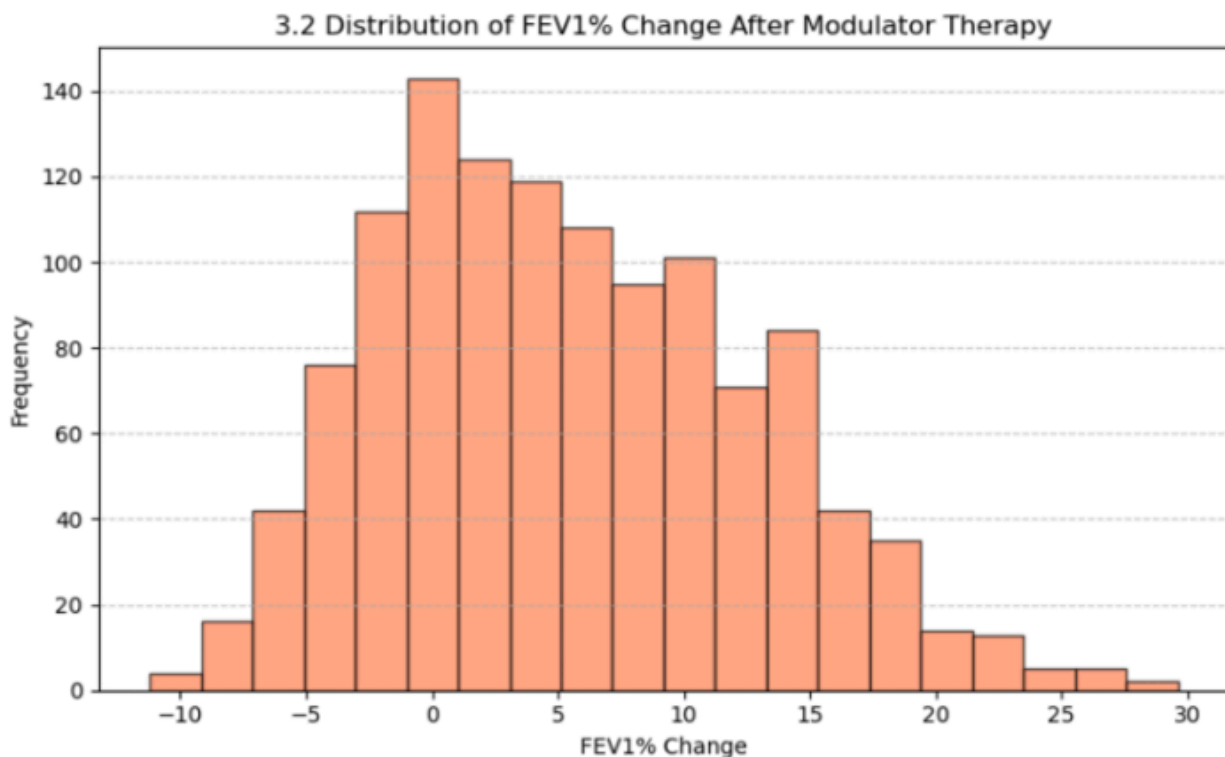


Figure 5: Distribution of FEV1% Change After Modulator Therapy

The exploratory data analysis provided clear visual information that supported the model's findings regarding how effective different treatments were and how patient characteristics played a role. Figure 5 illustrated the wide range of simulated patient responses. It showed a noticeable spread towards positive changes, indicating significant improvements for some individuals, alongside a distribution centered around zero for those with less or no change.

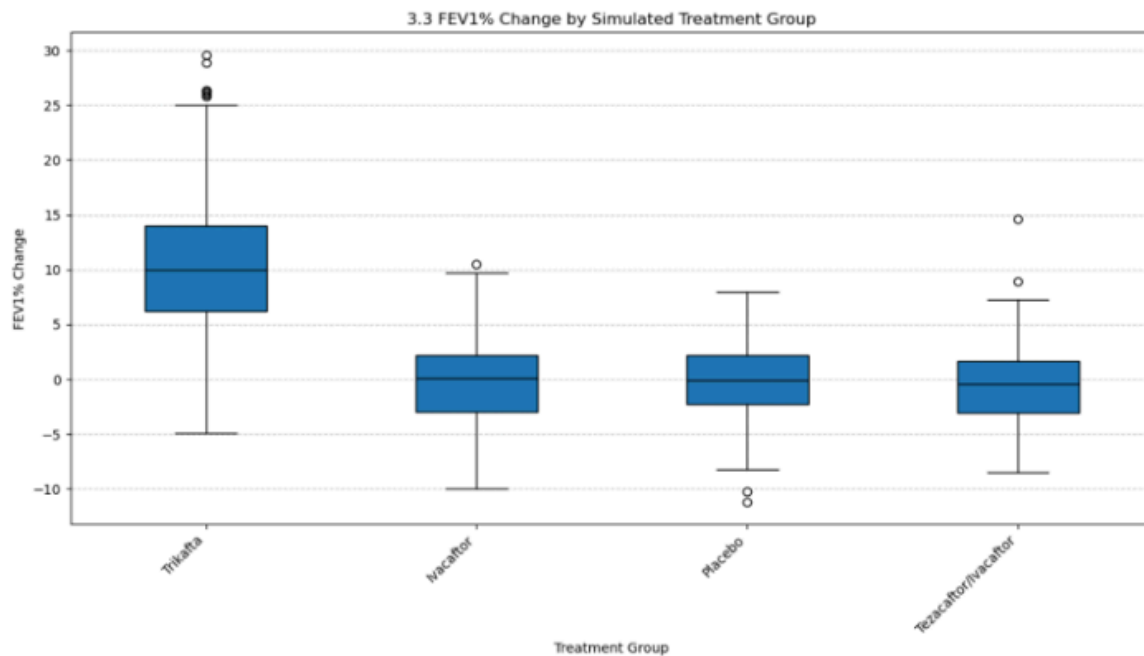


Figure 6: FEV1% Change by Simulated Treatment Group

Figure 6 clearly showed that the therapies had different levels of effectiveness. Trikafta consistently demonstrated the most substantial positive impact. Patients assigned to Trikafta showed a significantly higher average FEV1 change and a wider range of considerable improvements. This aligns with what is generally observed in real clinical settings, where Trikafta has shown superior effectiveness for many eligible patients. In contrast, other treatments like Ivacaftor and Tezacaftor/Ivacaftor, while designed to be effective for specific genetic profiles, showed more modest or, in some cases, negligible average changes across the broader simulated patient population.

4. Discussion

This project shows that machine learning can be used as a proof-of-concept tool to predict how cystic fibrosis patients might respond to CFTR modulators. Among the models tested, ensemble methods, especially Random Forest, consistently performed better than linear models. This makes sense because nonlinear approaches are better at handling the complex interactions between genetics, environment, and disease progression that shape patient outcomes.

Looking more closely at what drove these predictions, treatment type, genotype, and baseline lung function came out as the most important factors. These results line up with what clinicians already know. For example, patients with gating mutations often benefit more from drugs like ivacaftor than those with minimal function alleles, and people with higher baseline

lung function typically show bigger improvements than those in later stages of disease. The fact that the models picked up on these known trends adds credibility and shows that, even with synthetic data, machine learning can reflect real clinical patterns.

At the same time, the results also point toward the value of going beyond genotype alone. Right now, treatment eligibility is based mostly on specific mutations, but there's still a lot of variability within those groups. By combining multiple clinical and genetic features, models like these could eventually help identify which patients are most likely to benefit and allow for more personalized treatment planning. The addition of SHAP analysis is especially important here because it makes the model's predictions easier to interpret, which is critical if these kinds of tools are to be used in real-world clinical settings.

Of course, there are limitations. The dataset used here was synthetic, so it cannot fully capture the biological diversity or social complexity of the cystic fibrosis population. That means the accuracy scores reported in this study should not be viewed as clinically validated. Another limitation is the narrow focus on short-term lung function changes, even though cystic fibrosis treatment outcomes are also shaped by quality of life, exacerbation frequency, nutrition, and long-term survival. Including these kinds of measures in future work would give a more complete picture. Finally, while Random Forest worked well here, more advanced approaches like deep learning or models that can handle longitudinal data may provide even stronger predictive power.

Altogether, these findings suggest that machine learning, when applied carefully and with robust data, has the potential to complement clinical decision-making and reduce uncertainty in treatment outcomes for cystic fibrosis patients.

5. Conclusion

In conclusion, this study demonstrates that machine learning can be used to predict CFTR modulator response, even in a simplified setting using synthetic data. Random Forest performed the best overall, and the features it identified as most important—treatment type, genotype, and baseline lung function—are consistent with well-established clinical knowledge.

These results show how machine learning could help move cystic fibrosis care toward a more personalized approach by combining genetic and clinical factors into one predictive framework. While much more work is needed with real-world datasets and broader outcome measures, this project provides an initial foundation. With further refinement, models like these could eventually support clinicians in making more individualized treatment decisions and improving care for people living with cystic fibrosis.

6. Limitations

This project, while demonstrating a powerful proof-of-concept, operates under several limitations that are important to acknowledge for interpreting the results and planning future work. Foremost among these is the exclusive reliance on synthetically generated clinical outcomes and drug responses. While these simulations were meticulously designed to reflect known patterns and variability from real clinical trials and scientific literature, they cannot fully capture the true biological variability, rare patient responses, and complex confounding factors present in real-world patient populations. Consequently, the specific quantitative predictions, such as the exact R^2 score or Mean Absolute Error, are indicative of the model's potential but critically require validation with actual patient data.

Furthermore, the simulated impact of comorbidities, including pancreatic status, diabetes, and *Pseudomonas* colonization, represents a simplified approximation. In reality, the interactions between these conditions, CFTR mutations, and modulator responses are far more intricate and dynamic. The current model does not account for the full spectrum of disease progression, severity of comorbidities, or the dynamic impact of their management. A crucial aspect that limits immediate clinical applicability is the absence of external validation; the models developed in this project were trained and evaluated solely on the simulated dataset, meaning their generalizability and robustness to new, unseen patient populations in a clinical setting remain unconfirmed.

Additionally, while real CFTR mutation data served as a foundational element, the construction of full patient genotypes (diplotypes, or the pair of mutations an individual carries) and the assignment of a 'second allele' were simulated. This inherent limitation restricts the ability to capture the complete genetic complexity and the nuanced interactions between two specific mutations that might influence modulator response in real patients. Therefore, even with careful hyperparameter tuning and cross-validation on the simulated data, models trained exclusively on synthetic data may not generalize perfectly to the inherent complexities and nuances of real-world patient populations, as unforeseen biological or clinical factors not captured in the simulation could ultimately affect predictive performance.

7. Future Work & Enhancements

Building upon the robust proof-of-concept established by this project, future work will strategically focus on transitioning to real-world data and incorporating advanced methodologies to significantly enhance clinical relevance and impact. The most critical next step for clinical translation involves the direct integration with real-world, longitudinal patient registry data. This action necessitates collaboration with Cystic Fibrosis centers and patient registries, such as the CFF Patient Registry, to acquire comprehensive, de-identified, longitudinal patient data. Such a dataset must include complete CFTR genotypes (both alleles per patient), detailed baseline

clinical status, specific modulator therapy regimens, and serial FEV1 measurements over extended periods. The profound impact of this step will be enabling the training and rigorous validation of models on actual patient outcomes, crucially enhancing their clinical applicability, generalizability, and the ability to predict long-term responses.

Beyond current classification methods, advanced genetic feature engineering and genomics will incorporate more granular and comprehensive genetic information. This action will involve utilizing real patient-level diplotypes for all identified CF-causing mutations, integrating predicted CFTR protein residual function scores for specific genotypes, and critically exploring the role of modifier genes—genetic variants outside the CFTR gene—that are known to influence disease severity or treatment response. Furthermore, future investigations could delve into insights derived from whole-exome or whole-genome sequencing data. The collective impact will be uncovering more subtle genetic influences on modulator efficacy, leading to highly precise, genotype-informed predictions that truly account for individual genetic backgrounds.

To achieve a more holistic patient view, multi-omics data integration will expand beyond genomics to incorporate other biological data types. This action involves exploring the inclusion of proteomics, which analyzes protein expression profiles, metabolomics, which studies metabolic signatures, or microbiome data, if such information becomes available from real patient cohorts. The expected impact is the development of a more comprehensive and holistic predictive framework, potentially identifying novel biomarkers and biological pathways that influence modulator response and disease progression.

For predicting trajectories and understanding disease progression over time, time-series and deep learning models for longitudinal data will be investigated. This action will involve exploring advanced deep learning architectures such as Recurrent Neural Networks (RNNs) or Transformer models for analyzing longitudinal patient data. The impact will be enabling the prediction of long-term FEV1 trajectories, assessing response durability, and identifying critical time points for therapeutic intervention or adjustment.

To strengthen the clinical evidence base by distinguishing causation from mere correlation, causal inference methods will be applied. This action involves using techniques such as propensity score matching or instrumental variables to establish true causal links between factors and outcomes, rather than simply observing correlations in observational data. The impact will be providing stronger, more robust evidence for clinical decision-making and informing the design of future drug development strategies.

Ensuring model transparency and mitigating potential biases will involve enhanced Explainable AI (XAI) and fairness assessments. This action includes deepening SHAP analysis and exploring other XAI techniques, such as LIME or saliency maps, to provide more intuitive and actionable explanations for clinicians. If using real patient data, it will also involve rigorously

assessing model fairness for potential biases related to demographics, such as ethnicity or socioeconomic status, and developing strategies for mitigation to ensure the equitable application of personalized therapies. The impact will be building greater trust in AI-driven clinical tools and ensuring the ethical and equitable application of personalized medicine approaches.

Translational impact, bridging the gap between research and clinical practice, will be addressed through the development of a clinician-friendly decision support system. This action involves designing and developing an interactive user interface or dashboard where clinicians can easily input patient data and receive predicted modulator responses, along with confidence intervals and key explanatory features. The impact will be facilitating seamless integration of predictive insights into clinical workflows, empowering clinicians to make more informed and personalized treatment decisions at the point of care.

Finally, to provide comprehensive risk assessment and expand the predictive scope beyond efficacy, the prediction of adverse events and non-response mechanisms will be pursued. This action involves extending the model to forecast the likelihood of specific adverse reactions or to identify underlying biological reasons for non-response to particular modulators. The impact will be guiding alternative therapeutic strategies, informing patient counseling, and providing valuable insights for the development of next-generation modulators. This ambitious future work outlines a clear path for this project to significantly contribute to the field of Cystic Fibrosis, moving towards truly personalized and optimized patient care.

8. References

- Accurso, F. J., Rowe, S. M., Clancy, J. P., Boyle, M. P., Dunitz, J. M., Durie, P. R., Sagel, S. D., Hornick, D. B., Konstan, M. W., Donaldson, S. H., Moss, R. B., Pilewski, J. M., Rubenstein, R. C., Uluer, A. Z., Aitken, M. L., Freedman, S. D., Rose, L. M., Mayer-Hamblett, N., Dong, Q., & Zha, J. (2010). Effect of VX-770 in Persons with Cystic Fibrosis and the G551D-CFTR Mutation. *The New England Journal of Medicine*, 363(21), 1991–2003.
<https://doi.org/10.1056/NEJMoa0909825>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
<https://doi.org/10.1023/a:1010933404324>
- Cystic Fibrosis Foundation. (2019). Cff.org; Cystic Fibrosis Foundation. <https://www.cff.org>
- Elborn, J. S. (2016). Cystic fibrosis. *The Lancet*, 388(10059), 2519–2531.
[https://doi.org/10.1016/s0140-6736\(16\)00576-6](https://doi.org/10.1016/s0140-6736(16)00576-6)
- Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5), 1189–1232. <https://www.jstor.org/stable/2699986?seq=1>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *Springer Texts in Statistics An Introduction to Statistical Learning*.
https://www.stat.berkeley.edu/~rabbee/s154/ISLR_First_Printing.pdf
- Lundberg, S., & Lee, S.-I. (2017, November 24). *A Unified Approach to Interpreting Model Predictions*. ArXiv.org. <https://doi.org/10.48550/arXiv.1705.07874>
- Middleton, P. G., Mall, M. A., Dřevínek, P., Lands, L. C., McKone, E. F., Polineni, D., Ramsey, B. W., Taylor-Cousar, J. L., Tullis, E., Vermeulen, F., Marigowda, G., McKee, C. M., Moskowitz, S. M., Nair, N., Savage, J., Simard, C., Tian, S., Waltz, D., Xuan, F., & Rowe, S. M. (2019). Elexacaftor–Tezacaftor–Ivacaftor for Cystic Fibrosis with a Single Phe508del Allele. *New England Journal of Medicine*, 381(19), 1809–1819. <https://doi.org/10.1056/nejmoa1908639>
- Rowe, S. M., Daines, C., Ringshausen, F. C., Kerem, E., Wilson, J., Tullis, E., Nair, N., Simard, C., Han, L., Ingenito, E. P., McKee, C., Lekstrom-Himes, J., & Davies, J. C. (2017). Tezacaftor–Ivacaftor in Residual-Function Heterozygotes with Cystic Fibrosis. *New England Journal of Medicine*, 377(21), 2024–2035. <https://doi.org/10.1056/nejmoa1709847>
- Rowe, S. M., Miller, S., & Sorscher, E. J. (2005). Cystic Fibrosis. *New England Journal of Medicine*, 352(19), 1992–2001. <https://doi.org/10.1056/nejmra043184>
- Sosnay, P. R., Siklosi, K. R., Van Goor, F., Kaniecki, K., Yu, H., Sharma, N., Ramalho, A. S., Amaral, M. D., Dorfman, R., Zielenski, J., Masica, D. L., Karchin, R., Millen, L., Thomas, P. J., Patrinos, G. P., Corey, M., Lewis, M. H., Rommens, J. M., Castellani, C., & Penland, C. M. (2013). Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. *Nature Genetics*, 45(10), 1160–1167.
<https://doi.org/10.1038/ng.2745>
- Welcome to CFTR2 | CFTR2. (2011). Cffr2.org. <https://cffr2.org/>



9. Acknowledgements

This project would not have been possible without the invaluable guidance, support, and encouragement of several individuals and institutions. First and foremost, I would like to express my deepest gratitude to my mentor, Dr. Cori Faklaris, for her constant support, insightful feedback, and unwavering belief in my potential throughout this journey. I am also incredibly thankful to the University of North Carolina's College of Computing and Informatics for providing access to advanced resources, technical expertise, and a collaborative academic environment that greatly enhanced the development and execution of this project. I would also like to express my gratitude to the Army Outreach and Education Program (AEOP) for giving me this opportunity to pursue research in a topic that I am interested in.