A Logistic Regression Model for Predicting Treatment Response to Elexacaftor/tezacaftor/ivacaftor in Cystic Fibrosis Patients with the F508Del Mutation

Rishika Kurma[1], Sahiti Marella[2]
[1]Downingtown High School East Campus; Exton, Pennsylvania, United States
[2]Department of Dermatology, University of Michigan; Ann Arbor, Michigan, United States

Abstract

Cystic Fibrosis (CF) is a genetic disease that affects the lungs and other organs, causing mucus and other fluids to become excessively thick. The F508Del mutation is the most common variant of CF, and the Elexacaftor/tezacaftor/ivacaftor (ETI) therapy is frequently used to treat this specific mutation. Despite this recent advancement, patient variability leads to differences in individual response to ETI, regardless of sharing the F508Del mutation. This study addresses the gap in CF machine learning models to predict treatment response by developing a logistic regression model to detect the responsiveness of a CF patient following the ETI Treatment. Specifically, the research will answer the question: to what extent can a Logistic Regression model accurately classify responsive and unresponsive cases of CF patients with the F508Del Mutation following ETI treatment? The study implements a two-part quantitative method, including a Differential Gene Expression (DGE) Analysis experimental approach and a Logistic Regression Model evaluation approach. The DGE Analysis found that LDLR, TNF, and PSMD5 were the differentially expressed genes (DEGs) across the CF genetic data. The model evaluation leveraged a confusion matrix, McNemar's Test, and an ROC Curve. The model achieved an 85.71% accuracy, a 66.67% sensitivity, and a 100% specificity. The Area Under the ROC Curve (AUC) was 91.67%. The study concluded these evaluation factors as statistically significant (p-value = 0.00548). These findings suggest that machine learning can assist in personalized treatment prediction, and further validation with larger and more diverse cohorts is warranted to enhance generalizability.

Keywords

Cystic Fibrosis, Machine Learning, Personalized Medicine, Logistic Regression, Differential Gene Expression (DGE), Phenylalanine 508 Deletion (F508Del) Mutation, Elexacaftor/tezacaftor/ivacaftor (ETI), Cystic Fibrosis Transmembrane Regulator (CFTR)

1. Introduction

Approximately 3.5% to 5.9% of individuals worldwide have 1 of nearly 7, 000 rare or genetic conditions (Lichstein et al., 2022). Cystic Fibrosis (CF) is one of these rare genetic diseases, affecting fewer than 200,000 patients globally. Nevertheless, its infrequency does not diminish the urgency to address the substantial clinical issue. CF is a genetic disease caused by a mutation in the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Gene. Due to this genetic alteration, defective proteins are generated and affect the production of bodily fluids, such as sweat and mucus. In CF patients, mucus is significantly thicker, causing blockages, damage, and internal organ infections, and as a result, diminishes the efficiency of critical

functions of the airways and digestive tract. CF is a serious issue for carriers, as it is a life-threatening disease that decreases the average life expectancy from 71.3 years to 61 years as recorded by the National Heart, Lung, and Blood Institute ("What is Cystic Fibrosis?", 2024). Beyond life longevity, CF significantly reduces life quality. The American Lung Association stated that patients affected with CF require an above-average amount of calories for weight maintenance and growth. CF patients are not allowed to smoke, consume alcoholic beverages, and are required to maintain physical health to maintain respiratory functionalities ("Learn About Cystic Fibrosis", 2024). Though several genetic mutations lead to the development of CF, the most common mutation is the F508Del Mutation–Phenylalanine at Position 508 of the CFTR gene, accounting for roughly 82% of all CF cases (Lopes-Pacheco, 2020).

2. Literature Review

2.1 The Problem of Patient Variability

Treatments for CF, including antibiotics, airway clearance techniques, lung transplants, and breathing support, help alleviate CF's effects as therapies advance. In more recent endeavors, the Elexacaftor/Tezacaftor/Ivacaftor (ETI) Treatment is a triple combination therapy consumed orally, and is widely used for patients with at least one F508Del Mutation. Although emerging treatments and other triple combination therapies beyond ETI are being introduced, patient variability prohibits these medications from being effective for every individual. Batsheva Kerem and Eitan Kerem, researchers at the Department of Genetics at Hebrew University, described in their review of CF that "there is a substantial variability in disease expression among patients carrying the same mutation" (Kerem & Kerem 1996). Moreover, Mei-Zahav et al., experienced researchers in genetics, CF, and lung research, conducted a study to determine the variability in disease severity of CF patients with mutations that exhibit residual function. Within their study, they tested patients with a heterozygous version of the F508Del mutation (one mutated copy) against patients who had a homozygous version of the F508Del mutation (two mutated copies). The study found significant variability in disease severity, despite containing the same CF variant (Mai-Zahav et al., 2025). This suggests that CF patients with the same mutation may respond differently to ETI, as factors like modifier genes, environmental factors, allelic variation, and complex genetic and environmental interactions also influence variability (Lobo, 2008).

The matter of patient variability is not only noted in CF but also found in several uncured genetic and rare conditions. Thus, many researchers have sought to leverage the AI revolution to develop machine-learning models to predict a patient's response to a treatment. Implementing these predictive algorithms substantially improves the clinical diagnosis of patients. To elaborate, Mohamed Khalifa, a researcher at multiple medical institutions in Australia, and Mona Albadaway, a researcher at the School of Population Health at the University of New South Wales, conducted a study on AI in clinical prediction, and discussed that treatment response prediction with AI "helps healthcare professionals in selecting the most appropriate treatment plan for each patient, maximising efficacy and reducing the risk of complications" (Khalifa & Albadaway, 2024, p. 4).

Because the ETI treatment may not be effective for every patient with the F508Del mutation, the proposed study aims to answer the following research question: to what extent can a Logistic

Regression model accurately classify responsive and unresponsive cases of CF patients with the F508Del Mutation following ETI treatment? The conclusions yielded from this study will lead to a new understanding of the capability of this model to predict the response to ETI of F508Del patients, which will indicate their application in the broader context of clinical diagnosis and treatment prognosis.

2.2 The Research Gap

2.2.1 AI Treatment Response Algorithms for Other Genetic Diseases

AI has been increasingly applied in healthcare, particularly for genetic and rare conditions. In oncology, AI models are being developed to predict treatment response. Postdoctoral fellow at NYU Langone Health Theodore Sakellaropoulos and their colleagues designed a workflow of a series of DNNs (Deep Neural Networks)–a type of AI Algorithm–to forecast drug response and patient survival. The DNN model retrieved up to a 70% Area Under the ROC Curve (AUC), or the model's ability to distinguish between two categories (Sakellaropoulos et al., 2019).

Beyond cancer, other conditions and diseases have been addressed, particularly with Logistic Regression Models, to predict treatment response. As a binary classification tool, logistic regression has demonstrated strong utility in identifying patient outcomes. For instance, Bisaso et al., researchers with strong backgrounds in data science, HIV, and infectious diseases at Makerere University, developed a logistic regression model to predict early virological suppression, which describes how effectively and quickly the treatment reduces HIV in patients. The model was proven functional with an accuracy of 92.9% and an AUC of 0.878 (Bisaso et al., 2018). Similarly, Bilancia et al., researchers at the University of Foggia in fields including regenerative medicine and medical and surgical sciences, programmed a logistic regression model to characterize the response to therapy in severe eosinophilic asthma, facilitating personalized treatment strategies (Bilancia et al., 2024).

Moreover, logistic regression models have even been implemented to predict CF patients' response to ETI. Molin Yue, a researcher at the UPMC Children's Hospital of Pittsburgh, worked with colleagues with backgrounds in Cystic Fibrosis and Pulmonology to develop transcriptomic risk scores (TRSs) that significantly improved the ability to forecast changes in lung function and BMI following ETI treatment for CF patients, achieving an ~85% accuracy rate (Yue et al., 2024). The researchers develop a regression model, however, it does not use binary classification. In fact, it provides the extent to which the patient responded to the treatment rather than whether the patient was a responder or a non-responder.

While prior studies have applied AI to predict treatment response in diseases such as cancer, HIV, and asthma, a gap remains in the context of CF. Existing CF research has focused primarily on the degree of response to ETI therapy, rather than categorizing patients as responsive or unresponsive. This study addresses the gap by applying binary classification to predict ETI responsiveness, with the potential to enhance clinical conclusions and treatment efficiency.

2.2.2 The Demand for CF Treatment Response Prediction

Several previous research articles use different computational algorithms to develop an understanding of treatment response for CF variability. For instance, researcher Hermann Bihler, a scientist at the Cystic Fibrosis Foundation, and colleagues published a study that elaborates on the influence of the ETI treatment on ~655 CF variants. The results yielded the variants with a clinical benefit, had a negligible difference, and showed an extension of disability. In this study, researchers used computer diagnostic tools to determine this information and yield statistical treatment responses (Bihler et al., 2024). Accordingly, my research fits into the conversation of predictive computational algorithms and their assistance in determining treatment response to meet the experimental demands.

In addition to the mentioned research, multiple studies in this scope have been conducted to understand the effects of different CF medications on the CF-affected population. To elaborate, Karina Kleinfelder, a researcher under the Department of Medicine at the University of Verona, collaborated with colleagues to determine the ability of in silico tools, or computational methods, in matching the ultra-rare CFTR genotypes and variants with an appropriate therapy. Their study retrieved the effect of the treatment on these CF patients based on the in silico analysis yield. Within this study, computational method accuracy and efficiency were key components in ensuring the validity of this conclusion (Kleinfelder et al., 2023). While this study utilizes a type of computational algorithm (in silico analysis), it does not implement a form of binary classification or logistic regression to retrieve its results. Hence, the proposed model could potentially provide greater clinical benefit as well as efficiency. Moreover, the model can be applied to various medical conditions or CF variants upon customization of the model's trained and tested data. Therefore, it would be valuable to this particular study, as the researchers attempt to conclude the result of a specific group of CF variants aside from the F508Del mutation population.

The unique contribution of this work includes the specific focus on the common F508Del mutation across the CF population, intending to improve the treatment diagnosis process by leveraging logistic regression machine learning to develop a model to predict the patient's response to the frequent CF treatment, ETI. The application of this model to predict treatment response for diseases beyond CF. By modifying the genetic data for a different genetic disease, the logistic regression can adapt to the specific data input and provide treatment response predictions for other diseases which can aid clinical diagnosis.

3. Methodology

3.1 Addressing Ethical Concerns: The Use of Open-Source Genetic Data

To acquire the genetic data of CF patients, I will consult the GEO (Gene Expression Omnibus) Dataset, an open-source platform that shares genetic data for all users from previous research studies. In this case, no sensitive or personal information of the patients will be disclosed in the following method.

3.2 The Resources, Variables, and Groups of the Research

The major resources used for the study include RStudio, the GEO Database, the STRING (Search Tool for the Retrieval for Interacting Genes/Proteins) Database, and the *DESeq2* R Package. RStudio is a programming Integrated Development Environment (IDE), mainly utilized to perform statistical tests with the programming language R. This IDE was applied in this study to normalize the datasets, train and test the model, run evaluation tests, and create visual representations of the model's performance.

Furthermore, the GEO Database was used in this study to retrieve both datasets. These datasets were published by established studies, including research conducted by Cinek et al., established researchers at the Department of Medical Microbiology at Second Faculty of Medicine, Charles University, and Motol University Hospital, on the genetic expression of intestinal cells of ETI-treated CF patients (Cinek et al., 2025). The data provided by Cinek et al. were multiple samples of the gene expression levels of various genes in response to ETI treatment. The second dataset used was from De Jong et al.'s study, done by researchers at the Telethon Kids Institute Respiratory Research Centre, in which the difference in Ivacaftor and the Lumacaftor/ivacaftor treatment for F508Del CF patients was noted. (De Jong et al., 2021). The dataset included samples of the F508Del patients treated with Ivacaftor, which is a part of ETI and sustains similar characteristics to it. These datasets are similar, as they both include gene expression levels, which would later be used to determine the differentially-expressed genes (DEGs) and leverage their expression levels to predict response to ETI.

The STRING Database was used to analyze the differentially expressed genes. STRING is a biological database that provides visual representations of protein-protein interactions based on genetic data. STRING will be used in this study to determine the genes with the greatest expression in the treated group of CF patients.

Finally, *DESeq2* is an in-built package in R, which includes functions, data, and compiled code that act as an extension to the default capabilities of R, which is generally used to identify DEGs from RNA-Seq data. DEGs refer to certain genes that demonstrate the most significant differences in expression levels among two or more cohorts. In the case of this study, *DESeq2* will run a DGE Analysis on the most significant genes, which will be the results from STRING, to determine their expression levels. This is important to provide the model with a concept of comparison, allowing it to predict the cases with accuracy and evidence.

In this study, the treated CF cohort served as the experimental group and the untreated cohort as the control. Gene expression was the dependent variable, while treatment status functioned as the independent variable. A key limitation was the inability to control cell type, due to limited data availability across the datasets. Still, this minorly influences the outcome of the study, as the gene expression seemed to stay constant amongst the datasets, and three of the most significant genes in the treatment responsiveness were identified. The genders and ages of the CF patients also varied among the datasets, but it did not cause a significant discrepancy in the model's predictive accuracy.

3.3 The Research Design

The proposed study will feature a two-part quantitative method to gather the data. The two parts of this design are as follows: (1) Differential Gene Expression Analysis (DGE Analysis) and (2) a Logistic Regression Model Evaluation to analyze the performance of the model. The DGE Analysis enacts an experimental method approach, while the Logistic Regression Model Evaluation implements an evaluation method approach. Both will leverage the use of computational tools to gather the results.

The DGE Analysis involved two fundamental factors. First, STRING was used to identify genes with the strongest interactions with CFTR, highlighting those most likely to influence response to ETI and serving as the differentially expressed genes (DEGs). Subsequently, *DESeq2* was applied to quantify the expression levels of these DEGs before and following ETI treatment, providing a reference point for model prediction. To substantiate this decision of leveraging DGE Analysis, Yue et al. used a DGE Analysis along with the STRING Analysis of genes to investigate the underlying biological processes triggered by ETI treatment (Yue et al., 2024). Furthermore, Rosati et al., researchers in Biotechnology and Microbiology at the University of Siena and Temple University, conducted a literature review on DGE Analysis, and the researchers stated, “This tool [DGE Analysis] can help in identifying genes involved in a particular biological process, disease, or response to treatment, thereby providing information on gene regulation and underlying biological mechanisms” (Rosati et al., 2024, p. 1155). The DGE Analysis will assist in answering the research question and developing my new understanding by rationalizing the model's performance, as a poor analysis of the DEGs could result in a poor model performance, while a sufficient DGE Analysis will provide the model with more specified expression levels and improved precision.

The Logistic Regression Model Evaluation includes three aspects. First, a confusion matrix is generated to compare predictions to the correct outcomes, yielding key performance metrics, such as accuracy. Following, McNemar’s Test is applied to assess the statistical significance of the model’s predictions and eradicate bias or random chance, ensuring balanced classification of responders and non-responders. The use of McNemar’s Test is justified by Andrew C. Leon, a biostatistician and professor at Weill Cornell Medical College, “The McNemar test is used to examine paired dichotomous data. For example, one might compare the symptomatology pretreatment and post-treatment” (Leon, 1998, p. 243). The example provided by Leon is similar to my study, as I will be assessing the responsive and unresponsive cases and whether they had the same significance during model prediction, eradicating model bias. Finally, an ROC Curve is created to visualize the model’s discriminative abilities, with the Area Under the ROC Curve (AUC) used as a standard metric to assess performance. This metric was also employed by Yue et al. to evaluate the performance of their TRS, or scoring algorithm, as stated, “We evaluated TRS performance using leave-one-out cross-validation and calculated the area under the receiver operating characteristic curve (AUC) and the corresponding performance or confusion matrices” (Yue et al., 2024, p. 732). Model evaluation is central to addressing the research question, as it determines the extent to which the model can accurately classify treatment outcomes. An accuracy above 80% and an AUC near 90% would indicate strong predictive performance, comparable to previous successful models. These metrics also contribute to a new understanding of AI’s reliability in clinical decision-making contexts.

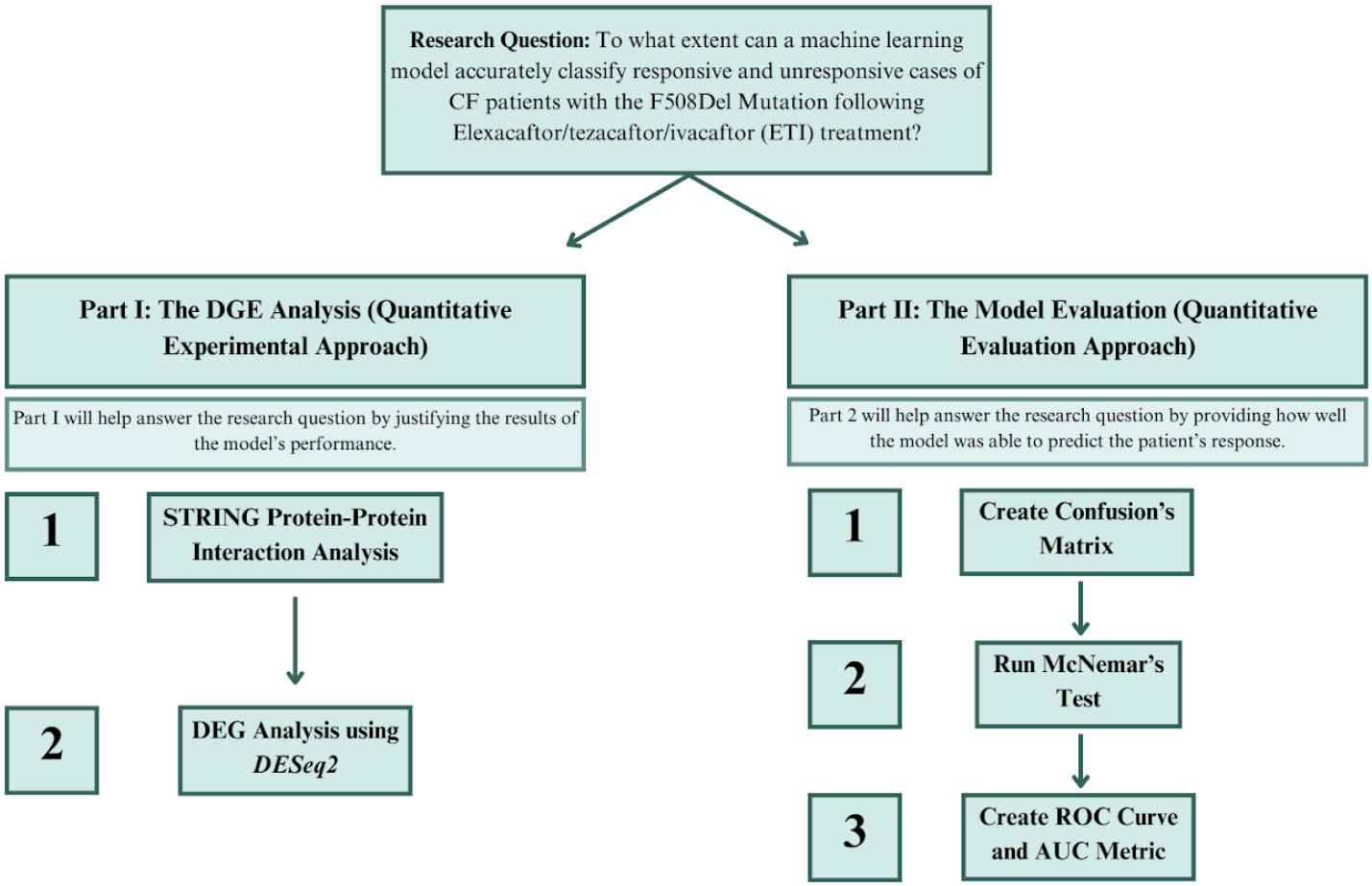

Figure 1: This figure is a flowchart representing the design of the research, including Part I: The DGE Analysis, and Part II: The Model Evaluation. The procedures for each part are outlined in the order they will be done, including the relation of each part to the initial Research Question.

3.4 Research Procedures

The procedure of the research included the following steps in their respective order: (1) retrieving and normalizing the data, (2) running a DGE analysis, (3) training and testing the model, and (4) completing the Logistic Regression Model Evaluation.

First, the two datasets will be retrieved from two established studies on the GEO Database. Once these datasets have been acquired, the data will be normalized and cleaned, to select the relevant samples, reduce the sample size, identify the treated and untreated CF cohorts of the data, and label the samples that were treated with ETI with a '1', and the untreated samples with a '0'. This process will be completed in an Excel file for both datasets, which will later be imported into RStudio as a .csv file. Both .csv files used for this specific study can be found in Appendix A.

Next, a DGE analysis will be performed on the normalized data using STRING and *DESeq2*. Essentially, the data of the treated group, or the samples containing a numerical value of '1', will

be loaded into STRING to identify which genes' proteins interact directly with the mutant CFTR gene protein. This procedure will be completed for the treated group of each dataset, and will later be compared to determine the common DEGs amongst both datasets to identify the most accurate different genes. Then, the common DEGs will be analyzed by *DESeq2* through the code in RStudio.

Upon this preliminary process of identifying the relevant data for the model to learn, training and testing datasets will be made by the code in RStudio. These datasets are generated from the '.csv' file, in which 75% of the data will be designated for training the model, while the remaining 25% is dedicated to testing the model. The training and testing datasets ensure a class balance, or contain the same proportion of responders to non-responders.

The evaluation factor is the next and final aspect of this procedure. Once the model has been trained and tested, the confusion matrix table will be made to test the model's predictions against the expected outcomes, overall demonstrating the model's predictive accuracy. It will also determine the performance rates: accuracy, specificity, and sensitivity. Sensitivity refers to the ability of the model to correctly identify positive cases, while specificity implies the ability of the model to identify negative cases. The accuracy of the model is recognized by its ability to make correct predictions, whether negative or positive. These metrics are automatically calculated using technical equations, which can be found in Appendix B. Then, McNemar's Test will be used to ensure the model does not contain bias during predictions. In addition to the performance rates, an ROC (Receiver-operating Characteristic) Curve will be created to demonstrate the significance of each gene's expression and the ability of the model to correctly distinguish between responsive and unresponsive, respectively. The AUC is generated upon creation of the ROC statistic.

The entire R Program used to analyze the DEGs, train and test the model, and create statistical representations can be found in Appendix A.

4. Results

4.1 The DGE Analysis

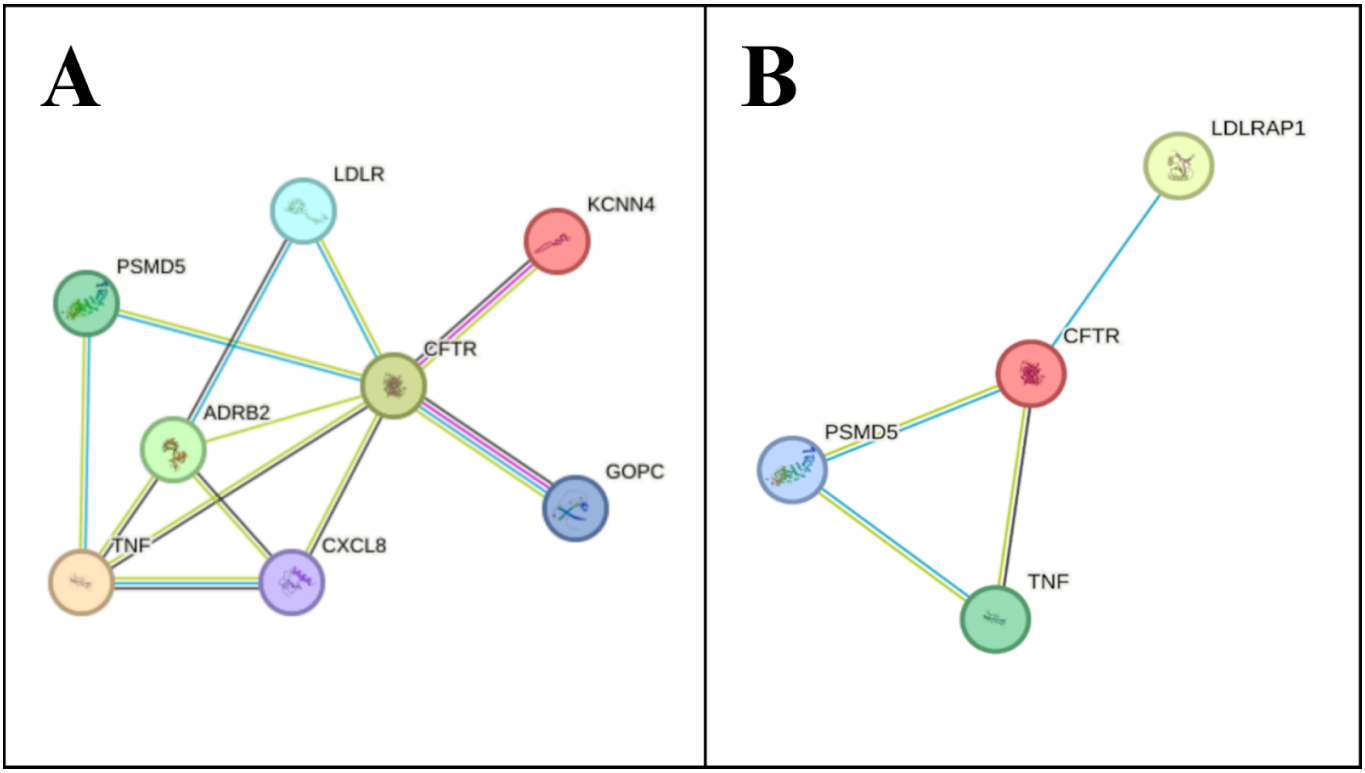

Figure 2: Figure 2A represents dataset 1, in which the genes that directly interacted with CFTR were LDLR, PSMD5, TNF, ADRB2, CXCLB, GOPC, and KCNN4. Figure 2B represented 2, in which the genes that directly interacted with CFTR were LDLRAP1, PSMD5, and TNF.

Based on these protein interactions with CFTR (Figure 2), genes LDLR, TNF, and PSMD5 are common amongst both datasets, implying that they are the most significantly expressed in a treated and responsive CF patient. Consequently, these three genes were chosen for the DGE Analysis, which will be done by DESeq2.

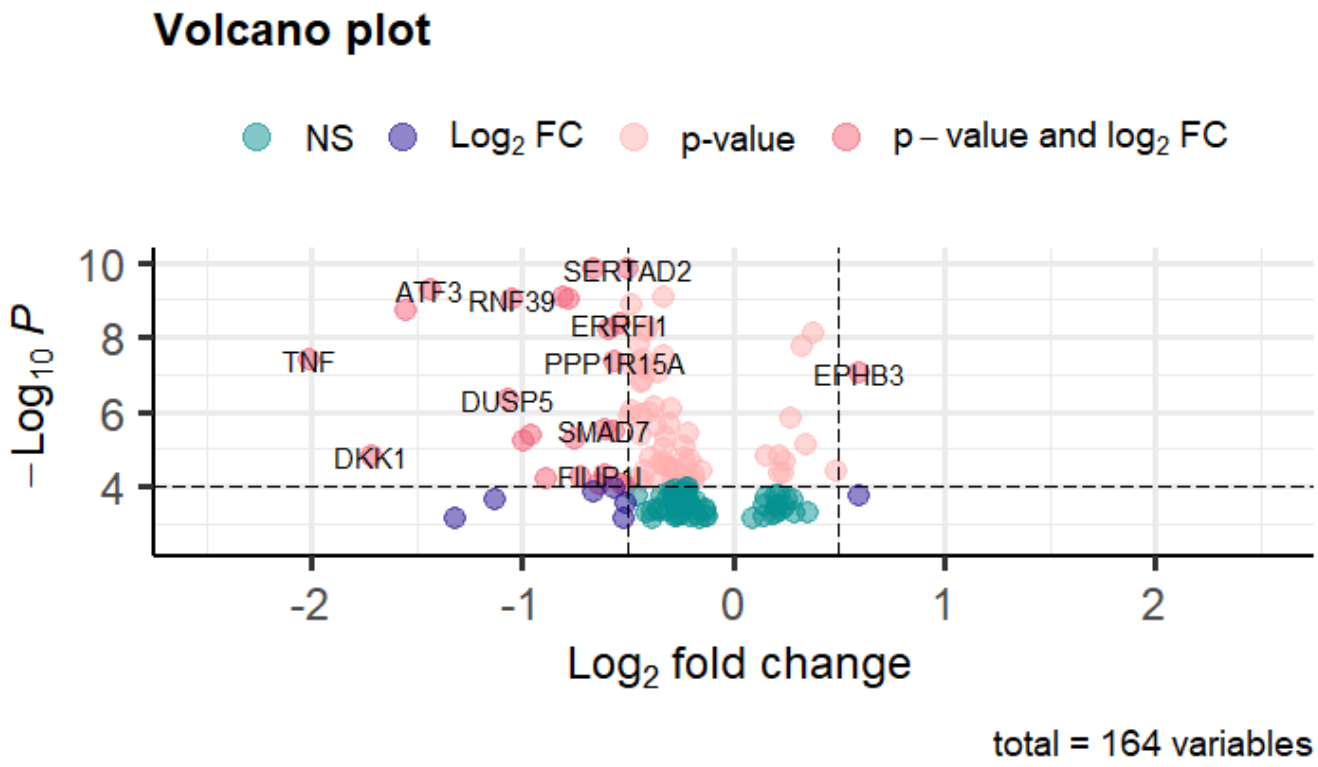

Figure 3: This figure is a Volcano Plot that depicts the upregulation and downregulation of the genes.

A Volcano Plot was created to further depict the highly expressed, or upregulated, genes, and the genes with lower expression, or downregulated genes. The -Log10P (y-axis) represents the p-value and statistical significance, while the Log2 Fold Change (x-axis) represents the amount of upregulation or downregulation of the plotted gene. The plot includes genes with significant p-values and insignificant log2 Fold Changes (light pink), significant log2 Fold Changes and insignificant p-values (dark purple), significant p-values and log2 Fold Changes (dark pink), and insignificant p-values and log2 Fold Changes (dark green).

TNF and ATF3 are significantly upregulated in the plot, and this is consistent with published studies. Researchers at the University of Leeds found that mutations in the CFTR Gene often lead to exaggerated production of the TNF Gene, which causes the chronic inflammation characteristic of CF (Lara-Reyna et al., 2019). Additionally, researchers with strong backgrounds in Hepatology and Histology found that ATF3 is upregulated in human and mouse fibrotic livers, and the overexpression of the gene is directly associated with CF traits (Shi et al., 2020). This validates that the data collected is consistent with the established norm of CF gene regulations and that it can be used to determine treatment response.

4.2 Model Evaluation and Performance Metrics

Once the model was trained and tested with relevant datasets, a Confusion Matrix was used to evaluate performance. The p-value of the model was 0.00548, which is significantly lower than the general p-value threshold of 0.05. This indicates that the model's accuracy is significantly higher than that of simply guessing the most common case amongst the data. The McNemar's Test p-value was 0.24821, which is significantly higher than the universal p-value threshold of 0.05. This implies that there is no strong evidence that the types of misclassifications, such as false positives and false negatives, are imbalanced. Therefore, the model has no favoritism in predicting a false negative or a false positive over the other, and the model is not biased. The entirety of the model evaluation results can be found in Appendix B. The three relevant performance rates, accuracy, sensitivity, and specificity, of the model evaluation were 85.71%, 66.67%, and 100%, respectively. This indicates that the model was 85.71% accurate when predicting the patient's response to the treatment, which demonstrates a well-performing model, as it achieved an accuracy above 80%. The model had a rather low sensitivity rate of 66.67%, implying that it was not able to correctly predict the case of the responders. Despite this, the model correctly predicted every sample of the unresponsive case, with a 100% specificity rate.

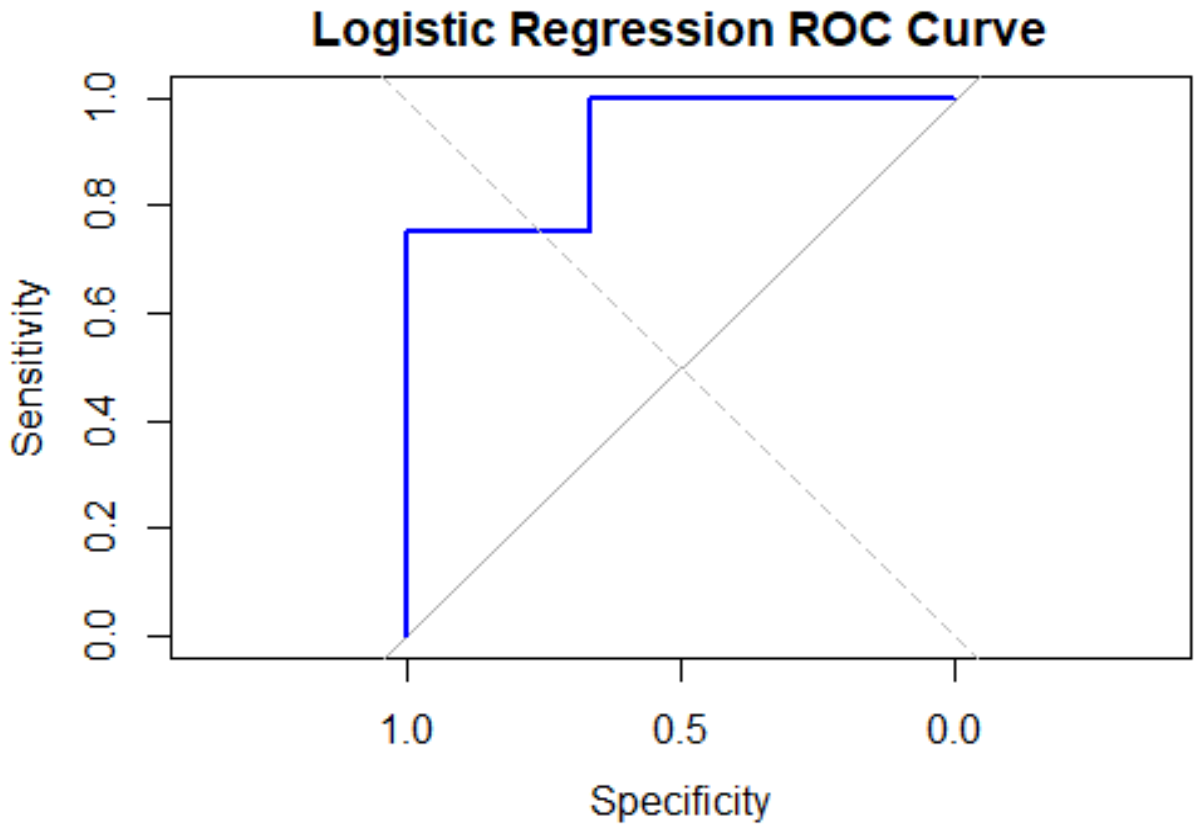

Figure 4: The figure demonstrates the ROC Curve with an AUC of 91.67%.

The ROC Curve (Figure 4) indicates that, although the model contained flaws, it was overall a confident and accurate algorithm. The gray lines running diagonally across the graph represent areas in which the false positive rate, or the specificity, is equal to the true positive rate, or the sensitivity. The ideal ROC Curve would never drop below these gray lines, creating a 90-degree

angle, which implies there is 100% area under the curve, or a 100% AUC. Nonetheless, because this model falls below the gray lines (near coordinates (1.0, 0.7) to (0.6, 1.0)), this implies that certain threshold ranges disabled the model from distinguishing between the responsive and unresponsive cases. The ROC Curve resulted in a 91.67% AUC, demonstrating that the model was highly functional, as it was above 90% and often made the correct predictions of the treatment response.

5. Discussion

5.1 Fulfillment of the Gap in the Research

By developing a model to predict the treatment response of CF patients, this model further elaborates on the identified research gap. Several predictive tools are recognized for CF treatment response, yet fail to develop ML models that can be applied to clinical environments and used to quickly assess a patient's response. For instance, Yue et al.'s study, although it develops a scoring mechanism to provide the extent to which the patient responded, it may pose a few gray areas. This score by Yue et al. does not depict a clear line between the responsive and unresponsive patients, and thus, may be difficult for a medical official to provide the patient's next steps. However, the model aims to bridge this gap by developing a logistic regression binary classification algorithm, which classifies the patient's responsiveness into one of two categories: responsive or unresponsive. The proposed model can initiate research in this gap and allow for future advancements.

5.2 Limitations

It is significant to recognize the limiting factors of this study and interpret the results in the broader context of the presented research. To begin, the limited accessibility of CF patients undergoing ETI treatment provided by the GEO Database was identified as a constraint, as few published datasets contained the necessary data. As a result of restricted accessibility, the sample size was smaller than ideal. With greater availability of relevant datasets, the predictive model may have achieved a superior performance. A larger training dataset would have reinforced the model with greater predictive power, thereby enhancing accuracy.

Furthermore, the disparity between the datasets utilized constitutes another limitation. Dataset 1 comprised data from the affected intestinal organoid of the CF patients, whereas Dataset 2 concentrated on affected nasal epithelial cells. Although this study does not inquire into the impact of varying cell types, homogenizing this factor may have led to marginal improvements in the model's performance. Consistency in cell type could have reduced slight variations in gene expression, as similar cell types often share comparable genetic profiles for analogous functions. Nonetheless, the examined cell types exhibit similar characteristics, and expression is not substantially divergent.

5.3 Implications

The proposed research and the results can motivate similar developments in clinical AI. Regarding communal impact, the study proposes a novel approach and diagnostic tool that,

upon refinement, can establish an accurate diagnosis and treatment for F508Del CF patients. This can assist experts in the medical industry and improve clinical treatment depending on the individuality of the CFTR mutation. The application of this tool also proposes greater access to treatment and improved patient survival rates. The application and further research upon this model and others can improve the ability for patients to seek assistance and gain valuable insight into personalized future medication. By leveraging binary classification, doctors and healthcare professionals can assess the patient's response to ETI to provide future directions for the patient to promote optimum care.

Moreover, DGE analysis demonstrated a profound impact on the performance metrics of the genetic scaling method. Through comprehension of the proposed research, DGE Analysis can be employed before predictive model test runs, allowing for greater accuracy of the model when examining the treatment response of genetic disease.

5.4 Future Directions

An important aspect of this model is that the variables can be modified to suit a larger range of patients. For instance, the model can be generated to produce the responsiveness to treatments aside from ETI. In addition, the algorithm can be modified to complement other CF variants, which can clinically benefit the broader CF community.

Finally, this specific algorithm could improve accuracy through the integration of multi-omics data. Multi-omics data refers to the combination of genomics, transcriptomics, proteomics, etc., to create a multi-layered dataset deriving from protein, genetic, RNA, and other biological agent expressions. Considering the accessibility to relevant data, multi-omics was not a possibility for the framework of this model. Nonetheless, multi-omics has been proven to significantly enhance predictive model performance. For example, in a study by Hua Chai and fellow researchers at Sun-Yat Sen University in China, the paper aimed to develop a multi-omics approach to accurately predict cancer prognosis. The study retrieved notable results on the implementation of multi-omics data for predictive model accuracy, as the model accuracy improved by 6.5% with the addition of multi-omics data (Chai, 2021).

6. Conclusion

Ultimately, the research adds to the cornucopia of treatment response prediction, specifically concentrating on CF F508Del mutants. The study addresses the gap for predictive models to aid CF treatment diagnosis and permits further research to resume ML algorithmic prediction for treatment personalization. My research also intensified my initial assumption that AI is beneficial in healthcare, due to its high accuracy despite a small sample size and limited accessibility to optimum CF Genetic data. Due to the high performance of the model, with an accuracy rate of ~86%, I was able to develop my new understanding that a logistic regression model using binary classification can generally predict F508Del-mutated CF patients' response to ETI. This can help assist further research in this niche field by also leveraging this tool and elaborating on this tool to improve clinical diagnoses, treatment prognosis, and understanding differences amongst responders and non-responders. It is significant to acknowledge the limitation of a minimal sample size, which emphasizes the requirement to optimize this study in future

research to ensure that a functional logistic regression model persists with a high accuracy rate when substantiated with a greater data population and multiple datasets. Despite this, the accomplishment demonstrates in a broader context that AI is capable of automating current systems to eradicate human error and time-consuming operations in the healthcare field.

7. Acknowledgements

I would like to express my sincere gratitude to Dr. Sahiti Marella for her invaluable mentorship and encouragement and providing me with the relevant tools and knowledge to perform this research. I would also like to extend my appreciation to my AP Research teacher, Mrs. Jaime Marcakis, for her unwavering support and guidance throughout this project. I am grateful to the National Center for Biotechnology Information's Gene Expression Omnibus (GEO) and the STRING database for providing open-access resources, which made this study possible, particularly for youth researchers like myself.

8. References

Bihler, H., Sivachenko, A., Millen, L., Bhatt, P., Patel, A. T., Chin, J., Bailey, V., Musisi, I., LaPan, A., Allaire, N. E., Conte, J., Simon, N. R., Magaret, A. S., Raraigh, K. S., Cutting, G. R., Skach, W. R., Bridges, R. J., Thomas, P. J., & Mense, M. (2024). In vitro modulator responsiveness of 655 CFTR variants found in people with cystic fibrosis. *Journal of Cystic Fibrosis*, *23*(4), 664-675. https://doi.org/10.1016/j.jcf.2024.02.006

Bilancia, M., Nigri, A., Cafarelli, B., & Di Bona, D. (2024). An interpretable cluster-based logistic regression model, with application to the characterization of response to therapy in severe eosinophilic asthma. *The International Journal of Biostatistics*, *20*(2), 361-388. https://doi.org/10.1515/ijb-2023-0061

Bisaso, K. R., Karungi, S. A., Kiragga, A., Mukonzo, J. K., & Castelnuovo, B. (2018). A comparative study of logistic regression based machine learning techniques for prediction of early virological suppression in antiretroviral initiating HIV patients. *BMC Medical Informatics and Decision Making*, *18*(1). https://doi.org/10.1186/s12911-018-0659-x

Chai, H., Zhou, X., Zhang, Z., Rao, J., Zhao, H., & Yang, Y. (2021). Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. *Computers in Biology and Medicine*, *134*, 104481. https://doi.org/10.1016/j.compbiomed.2021.104481

Cinek, O., Furstova, E., Novotna, S., Hubackova, K., Dousova, T., Borek-Dohalska, L., & Drevinek, P. (2025). Gene expression profile of intestinal organoids from people with cystic fibrosis upon exposure to elexacaftor/tezacaftor/ivacaftor. *Journal of Cystic Fibrosis*. https://doi.org/10.1016/j.jcf.2024.09.005

De Jong, E., Garratt, L. W., Looi, K., Lee, A. H., Ling, K.-M., Smith, M. L., Falsafi, R., Sutanto, E. N., Hillas, J., Iosifidis, T., Martinovich, K. M., Shaw, N. C., Montgomery, S. T., Kicic-Starcevich, E., Lannigan, F. J., Vijayasekaran, S., Hancock, R. E., Stick, S. M., Kicic, A., & Arest, C. (2021). Ivacaftor or lumacaftor/ivacaftor treatment does not alter the core CF airway epithelial gene response to rhinovirus. *Journal of Cystic Fibrosis*, *20*(1), 97-105. https://doi.org/10.1016/j.jcf.2020.07.004

Kerem, B., & Kerem, E. (1996). The molecular basis for disease variability in cystic fibrosis. *European Journal of Human Genetics*, *4*(2), 65-73. https://doi.org/10.1159/000472174

Khalifa, M., & Albadaway, M. (n.d.). Artificial Intelligence for Clinical Prediction: Exploring Key Domains and Essential Functions. *Computer Methods and Programs in Biomedicine Update*. https://doi.org/10.1016/j.cmpbup.2024.100148

Kleinfelder, K., Lotti, V., Eramo, A., Amato, F., Lo Cicero, S., Castelli, G., Spadaro, F., Farinazzo, A., Dell'Orco, D., Preato, S., Conti, J., Rodella, L., Tomba, F., Cerofolini, A., Baldisseri, E., Bertini, M., Volpi, S., Villella, V. R., Esposito, S., . . . Sorio, C. (2023). In silico analysis and theratyping of an ultra-rare CFTR genotype (W57G/A234D) in primary human rectal and nasal epithelial cells. *IScience*, *26*(11), 108180. https://doi.org/10.1016/j.isci.2023.108180

Lara-Reyna, S., Scambler, T., Holbrook, J., Wong, C., Jarosz-Griffiths, H. H., Martinon, F., Savic, S., Peckham, D., & McDermott, M. F. (2019). Metabolic reprograming of cystic fibrosis macrophages via the ire1α arm of the unfolded protein response results in exacerbated inflammation. *Frontiers in Immunology*, *10*. https://doi.org/10.3389/fimmu.2019.01789

Learn About Cystic Fibrosis. (2024, October 30). American Lung Association. Retrieved April 21, 2025, from https://www.lung.org/lung-health-diseases/lung-disease-lookup/cystic-fibrosis/learn-about-cystic-fibrosis

Leon, A. C. (1998). 3.12 - Descriptive and Inferential Statistics. In A. C. Leon (Author), *Comprehensive Clinical Psychology* (pp. 243-285). https://doi.org/10.1016/B0080-4270(73)00264-9

Lichstein, J., Riley, C., Keehn, A., Lyon, M., Maiese, D., Sarkar, D., & Scott, J. (2022). Children with genetic conditions in the united states: Prevalence estimates from the 2016-2017 national survey of children's health. *Genetics in Medicine*, *24*(1), 170-178. https://doi.org/10.1016/j.gim.2021.09.004

Lobo, I. (2008). *Same Genetic Mutation, Different Genetic Disease Phenotype*. https://www.nature.com/scitable/topicpage/same-genetic-mutation-different-genetic-disease-phenotype-938/

Lopes-Pacheco, M. (2020). CFTR modulators: The changing face of cystic fibrosis in the era of precision medicine. *Frontiers in Pharmacology*, *10*. https://doi.org/10.3389/fphar.2019.01662

Mei-Zahav, M., Orenti, A., Jung, A., & Kerem, E. (n.d.). Variability in disease severity among cystic fibrosis patients carrying residual-function variants: data from the European Cystic Fibrosis Society Patient Registry. *ERJ Open Research*. https://doi.org/10.1183/23120541.00587-2024

Rosati, D., Palmieri, M., Brunelli, G., Morrione, A., Iannelli, F., Frullanti, E., & Giordano, A. (2024). Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review. *Computational and Structural Biotechnology Journal*, *23*, 1154-1168. https://doi.org/10.1016/j.csbj.2024.02.018

Sakellaropoulos, T., Vougas, K., Narang, S., Koinis, F., Kotsinas, A., Polyzos, A., Moss, T. J., Piha-Paul, S., Zhou, H., Kardala, E., Damianidou, E., Alexopoulos, L. G., Aifantis, I., Townsend, P. A., Panayiotidis, M. I., Sfikakis, P., Bartek, J., Fitzgerald, R. C., Thanos, D., . . . Gorgoulis, V. G. (2019). A deep learning framework for predicting response to therapy in cancer. *Cell Reports*, *29*(11), 3367-3373.e4. https://doi.org/10.1016/j.celrep.2019.11.017

Shi, Z., Zhang, K., Chen, T., Zhang, Y., Du, X., Zhao, Y., Shao, S., Zheng, L., Han, T., & Hong, W. (2020). Transcriptional factor atf3 promotes liver fibrosis via activating hepatic stellate cells. *Cell Death & Disease*, *11*(12). https://doi.org/10.1038/s41419-020-03271-6

What is Cystic Fibrosis? (2023, November 21). National Heart, Lung, and Blood Institute. Retrieved October 11, 2024, from https://www.nhlbi.nih.gov/health/cystic-fibrosis

Yue, M., Weiner, D. J., Gaietto, K. M., Rosser, F. J., Qoyawayma, C. M., Manni, M. L., Myerburg, M. M., Pilewski, J. M., Celedón, J. C., Chen, W., & Forno, E. (2024). Nasal epithelium transcriptomics predict clinical response to elexacaftor/tezacaftor/ivacaftor. *American Journal of Respiratory Cell and Molecular Biology*, *71*(6), 730-739. https://doi.org/10.1165/rcmb.2024-0103oc

9. Appendix A

9.1 CSV Files with Relevant Samples

Dataset ID	Citation	Link to Original Dataset	Link to Modified Version Used in This Study
GSE263022	Cinek et al., 2025	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE263022	https://docs.google.com/spreadsheets/d/18n6Rfkw0D9Y9z0fOzZNVBMH7FvzUr3HIVr1xeflZ3Ks/edit?gid=0#gid=0
GSE139078	De Jong et al., 2021	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE139078	https://docs.google.com/spreadsheets/d/1L0BBPjMAcZU022e0d3zXcqMB_GnqG4cv2vEEG-Tu5To/edit?usp=sharing

9.2 GitHub Repository of the R Program for Model Development

Github repository link: https://github.com/rk-hue/Cystic-Fibrosis-Treatment-Response-Model

9.3 Figure of The R Program

```r
### cross-validation and improvement
#Load in the data and format for deseq2 analysis
# Load the file without specifying row.names
countdata <- read.table("/Users/rk092/Downloads/GSE139078_main(2).txt", header = TRUE)
# Make row names unique by appending a suffix
rownames(countdata) <- make.unique(as.character(countdata[, 1]))
countdata <- countdata[, -1]  # Remove the column used for row names
head(countdata)
count_matrix <- data.matrix(countdata)
#count_matrix <- count_matrix[,-1]
head(count_matrix)
## condition is the experimental setup so thats why reordering samples would be important to have all samples of one group together
condition <- factor(c(rep("Control", 5), rep("Virus_and_Drug_Treated", 5)))

coldata <- data.frame(row.names=colnames(count_matrix), condition)
dds <- DESeqDataSetFromMatrix(countData=count_matrix, colData=coldata, design=~condition)
dds$condition <- relevel(dds$condition, ref = "Control")

#Run deseq2
dds <- DESeq(dds)
resultsNames(dds)
res <- results(dds, alpha = 0.2)

#get results
res <- res[order(res$padj), ]
summary(res)
ressig <- subset(res, padj < 0.2)
summary(ressig)
resdata <- merge(as.data.frame(ressig), as.data.frame(counts(dds, normalized=TRUE)), by= 'row.names', sort=FALSE)
names(resdata)[1] <- "gene"
head(resdata)
summary(resdata)

write.csv(as.data.frame(resdata), file="/Users/rk092/Downloads/GSE139078_validation_sig_genes.csv")
EnhancedVolcano(ressig,
                lab = resdata$gene,
                x = 'log2FoldChange',
                y = 'pvalue',
                xlim = c(-3, 3),
                ylim = c(0, 10),
                pCutoff = 0.05,
                FCcutoff = 1.5,
                labSize = 4,
                pointSize = 4.0,
                labCol = 'black',
                colAlpha = 1,
                caption = NULL,
                gridlines.major = FALSE,
                gridlines.minor = FALSE,
                drawConnectors = TRUE,
                title = NULL,
                axisLabSize = 12,
                subtitle = NULL,
                col = c('#089392FF', '#260F99FF', '#FFB2B2FF', '#F55F7BFF'))

# Load your data (replace with your file path)
data <- read.csv("/Users/rk092/Downloads/input_main.csv", row.names = 1)  # Replace with your file
# Ensure data contains DE genes and a 'Response' column
#'ADRB2', 'CXCL8', 'GOPC', 'KCNN4',
# Define DE genes identified by DESeq2
de_genes <- c('LDLR', 'PSMD5', 'TNF')  # Replace with your DE genes
features <- data[, de_genes]
response <- as.factor(data$Response)  # Response: Binary (e.g., 0 = Non-responder, 1 = Responder)

# Split data into training and testing sets (75/25 split)
set.seed(42)
train_index <- createDataPartition(response, p = 0.75, list = FALSE)
train_data <- features[train_index, ]
train_response <- response[train_index]
test_data <- features[-train_index, ]
test_response <- response[-train_index]

# Normalize the data (center and scale)
preProc <- preProcess(train_data, method = c("center", "scale"))
train_data <- predict(preProc, train_data)
test_data <- predict(preProc, test_data)

# Train a logistic regression model
log_model <- train(x = train_data, y = train_response, method = "glm", family = "binomial")

# Summarize the model
summary(log_model$finalModel)

# Predict on the test set
log_pred <- predict(log_model, newdata = test_data)  # Class predictions
log_prob <- predict(log_model, newdata = test_data, type = "prob")[, 2]  # Probabilities

# Evaluate model performance
# Confusion matrix
conf_matrix <- confusionMatrix(log_pred, test_response)
print(conf_matrix)

# ROC Curve and AUC
roc_curve <- roc(test_response, log_prob)
auc_value <- auc(roc_curve)
cat("AUC:", auc_value, "\n")

# Plot ROC Curve
plot(roc_curve, col = "blue", main = "Logistic Regression ROC Curve")
abline(a = 0, b = 1, lty = 2, col = "gray")
```

Figure 5: This program is the same code as in the GitHub Repository. The figure was added to address the potential inability to access the GitHub link.

10. Appendix B

10.1 Technical Equations to Calculate the Performance Rates

Performance Evaluator	Equation
Accuracy Rate	(True Positives + True Negatives)/(True Positives + True Negatives + False Negatives + False Positives)
Sensitivity Rate	True Positives/True Positives + False Negatives
Specificity Rate	True Negatives/True Negatives + False Positives

Table 1: A True Positive is an actual responder, while a False Positive is the model's incorrect prediction of a responder. A True Negative is an actual non-responder, while a False Negative is the model's incorrect prediction of a non-responder.

10.2 Extended Results From the Model Evaluation

Metric	Value	Interpretation
Confidence Interval	95% CI: (0.6366, 0.9695)	The program is 95% confident that the accuracy of the model lies between 63.66% and 96.95%. This is a wide range due to the small sample size.
Kappa Value	0.6957	This implies that the agreement between the model's predictions and the true labels is on a good level. The closer the kappa value is to 1, the less likely that the agreement of the predictions with the true labels was by chance.
Positive Predictive Value	1.0000 (100.00%)	Whenever the model predicts a positive case or a responsive case, it is correct 100% of the time.
Negative Predictive Value	0.8000 (80.00%)	Whenever the model predicts a negative case, or an unresponsive case, it is correct 80% of the time.
Prevalence	0.4286 (42.86%)	~42.86% of the data belongs to the positive class, or is a responsive case. There is about a 50/50 split between both cases, which eradicates possibilities for the model to guess on random chance or bias.

Detection Rate	0.2857 (28.57%)	28.57% of the data were true positives.
Detection Prevalence	0.2857 (28.57%)	The model predicted 28.57% of the cases as positive, or responsive. Because this value is equivalent to the detection rate, it implies that the model was able to predict all of the positive cases correctly.

Table 2: This table presents a comprehensive summary of the mode evaluation, incorporating additional factors not discussed in the main text to maintain focus and relevance.