Universiteit Antwerpen

Multimodal Data Integration for Predictive Modelling of Measles Vaccine Response with Cross-Vaccine Marker Validation

Elias Dams

Promotor: Dr. Pieter Meysman
Supervisor: Fabio Affaticati

University of Antwerp
Faculty of Science

2024-2025

Submitted in fulfilment of the requirements for the degree of
Master in Computer Science: AI & Data Science

June 2025

Contents

List of Figures

List of Tables

List of Acronyms

- TCR: T-cell receptor

Summary

Acknowledgements

Abstract

CHAPTER 1

Introduction

Vaccination is widely recognized as one of the most cost-effective public health strategies, yet individuals exhibit significant variability in their immune responses. Foundational studies in systems immunology [3, 2] have shown that while vaccines like yellow fever and influenza generally trigger robust immune responses, the intensity of these responses (such as antibody production) can vary considerably. This variability is especially noticeable among very young or very old individuals, or those with other health issues, because their immune systems tend to be less robust and more unpredictable.

This thesis takes a data-driven, computer science approach to predict responses to the MMR (Measles, Mumps, and Rubella) trivalent vaccine using multimodal data integration, meaning that it combines different types of data (each representing a distinct aspect of the immune response) into a single, unified analysis. Measles is a highly contagious viral disease characterized by fever and a red rash, and it can lead to serious complications like pneumonia, particularly in infants and young children [7]. In this thesis, the immune response is primarily evaluated based on antibody titers, which are regarded as the gold standard for assessing vaccine effectiveness because they offer a precise, quantifiable measure of the immune system's capacity to generate protective antibodies [9] (see Section 2.1.2 for further details). The goal is to develop predictive models that not only forecast individual vaccine responses but also identify specific immune markers that consistently correlate with vaccine effectiveness across various vaccines. A process known as cross-vaccine marker validation.

Predictive modeling is essential because it allows anticipation of an individual's vaccine response well before the full immune reaction is measured. For example, Van Tilbeurgh et al. (2021) [11] demonstrated that combining high-throughput technologies (such as transcriptomics, proteomics, and *in vivo* imaging) with computational models can reveal immune signatures linked to vaccine efficacy. Such models support personalized vaccination strategies by allowing healthcare providers to tailor vaccine schedules, dosages, or even select alternative vaccines based on predicted responses. For instance, if a model indicates a low immune response based on specific genetic markers, healthcare providers might opt for an additional booster or an adjusted formulation. Ultimately, this targeted strategy not only optimizes vaccine efficacy but also ensures better resource allocation.

The initial measles dataset used is derived from the study by Bartholomeus et al. (2020) [1]. In this study, adult volunteers (23 females and 17 males, aged between 19 and 29 years, all of whom were previously primed with MMR vaccines in childhood) received a booster dose of Priorix® and were monitored at four time points (Day 0, Day 21, Day 150, and Day 365) to measure antibody titers and gene expression profiles. In this thesis, the focus will be concentrated specifically on the measles-related data from this dataset.

Nevertheless, several challenges must be addressed to realize the full potential of predictive modeling. One challenge in this research is that the dataset comprises only 40 samples. This small sample size limits statistical power and increases the risk of overfitting, making it difficult to generalize the model to a broader population. In addition, the project involves integrating heterogeneous data types, such as cytokine levels, cytometry data, T cell receptor sequences, and RNA profiles. These different modalities come with varying scales and units, which adds complexity to data normalization and feature selection. Moreover, any noise or missing values in a small dataset can have a large impact on the model's performance, potentially leading to biased results. The high number of features (in the RNA data) relative to the number of samples may require careful feature selection or dimensionality reduction to prevent the models from becoming too complex and overfitted. Finally, an additional challenge arises from the risk of data leakage during the feature selection process. Selecting features based on their relevance to the Hepatitis B response prior to model evaluation can introduce bias, artificially inflating performance metrics and ofcourse reducing generalizability.

CHAPTER 2

Background

TODO: Cover key concepts in immunology and data science, review existing work on vaccine response modeling.

To understand the work presented in this thesis, it is essential to have a basic grasp of concepts from both immunology and data science. As a computer scientist, my approach is mainly data-driven, focusing on extracting, integrating, and analyzing various types of biological data. However, a foundational understanding of the underlying biological processes is critical to meaningfully interpret the results and validate the predictive models developed in this research.

2.1 Background in Biology

2.1.1 Immune System Overview

A key element of these biological processes is the functioning of the immune system, which can be broadly divided into three main components, as depicted in Figure 2.1:

Physical barriers (top)
Physical barriers, including the skin and mucous membranes, serve as the body's first line of defense by blocking most pathogens (microorganisms such as viruses, bacteria, fungi, and parasites that can cause disease) from entering.

Innate Immunity (right)
In case a pathogen still crosses the barriers, innate immunity comes into action. This defense is rapid and non-specific. Think of macrophages and neutrophils as cells that engulf invaders through a process called phagocytosis. Eosinophils and other granulocytes also attack pathogens or initiate inflammatory responses. Natural killer cells (NK cells) are also part of innate immunity and can directly destroy infected or abnormal cells. Although this response is very rapid, it does not recognize pathogens in the same specific way as the next branch. [6]

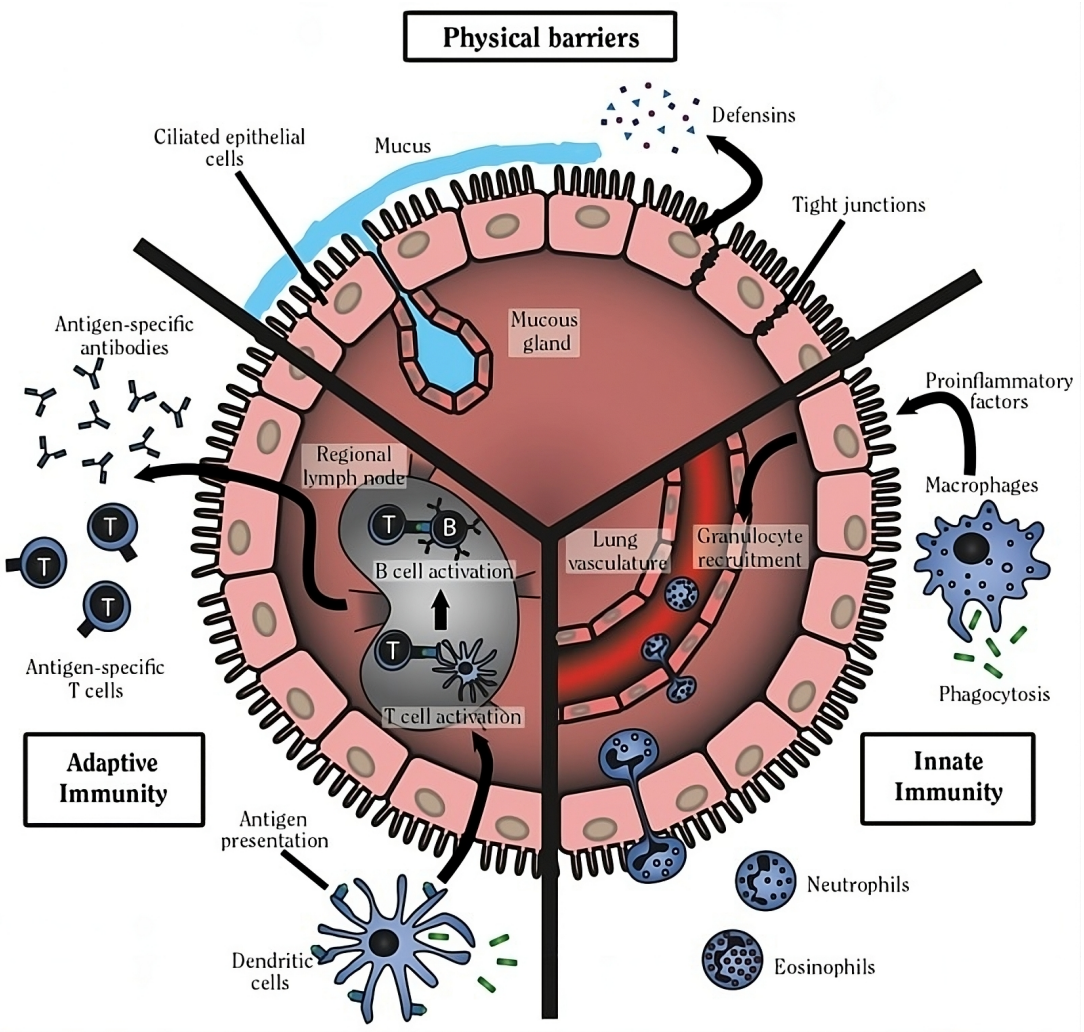

Figure 2.1: Diagram showing physical barriers, innate immune cells (e.g., macrophages, dendritic cells, natural killer cells) and adaptive immune components (B and T cells) working together. Reproduced from Figure 10.5 in *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General* (2014) [8].

Adaptive Immunity (left)
Acquired or adaptive immunity is the "slower but more targeted" defense. B cells play a crucial role by producing antibodies, which are proteins that bind to specific non-self antigens. These antibodies neutralize pathogens and tag them for destruction by immune cells such as phagocytes. The concentration of these antibodies in the blood is measured as "antibody titers," as mentioned earlier. Higher titers generally indicate a stronger immune response. T cells are also crucial and have several roles. They help coordinate the immune response (often referred to as "helper T cells") and can directly kill infected cells (cytotoxic T cells). T cell receptors (TCRs), located on the surface of T cells and responsible for recognizing peptides presented by MHC I/II molecules, can be sequenced (see Figure2.2) to determine which T cell clones are activated in response to a vaccine. A major advantage of adaptive immunity is that it "learns" from previous exposure, allowing for much faster and more powerful immune responses in the event of repeated infection. This ability to form memory also underlies how vaccines work. [6]

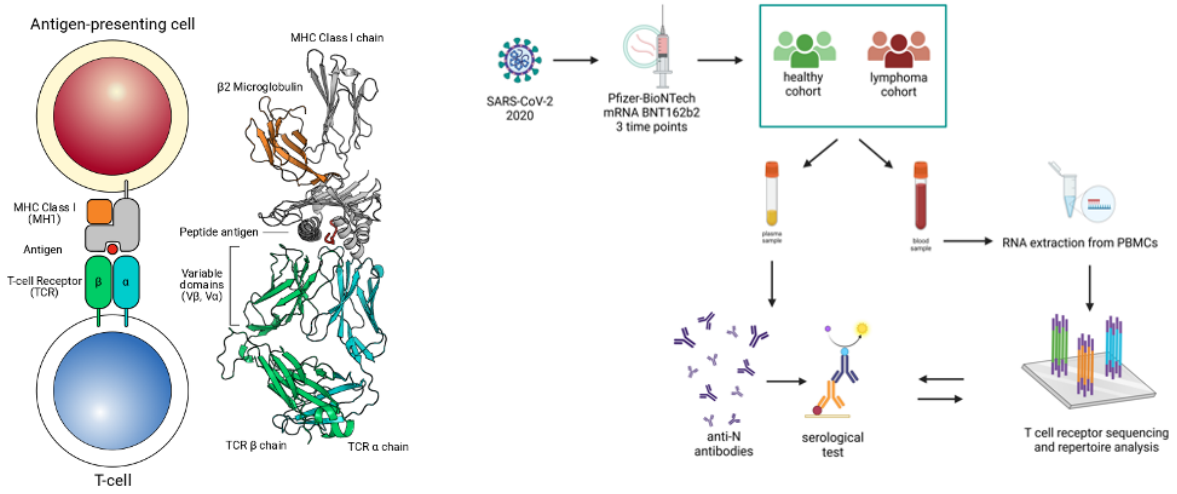

Figure 2.2: Schematic outline of how TCR data is obtained from peripheral blood mononuclear cells (PBMCs) after vaccination. By extracting and profiling TCR sequences, it becomes possible to identify which T cell clones expand in response to the vaccine. This information provides insights into the breadth (diversity) and Depth (magnitude or intensity) of the adaptive immune response.

Together, these three pillars provide a robust defence system that is able to successfully fight off most infections. It is also this dynamic between innate and adaptive immunity that determines the degree of vaccine response. The innate branch prepares the way, while the adaptive branch provides targeted antibodies and memory cells. This context is important because the predictive models developed here integrate features from both the innate and adaptive immune systems. For example, innate markers such as cytokine levels and certain cell counts can provide early signals about the body's general readiness to respond. Meanwhile, adaptive markers, such as TCR sequences of T cells, directly reflect the specific immune response that leads to antibody production after vaccination. Combining this data gives us a more complete picture of the immune landscape, allowing for better predictions of how effectively an individual will respond to the measles vaccine.

2.1.2 Antibody Titers

As mentioned above, antibody titers provide a quantitative measure of the concentration of specific antibodies in the blood, making them a reliable indicator of the immune system's functional response against a pathogen [9]. In essence, they reflect the "signal strength" of the response, with high titers indicating robust protection and low titers suggesting a weaker response. In this thesis, antibody titers serve as a crucial biomarker for classifying vaccine responses into strong and weak responders.

2.2 Background in Computer science

TODO: info about the used computer science techniques.

CHAPTER 3

Data Description and Preprocessing

TODO: Detail the measles dataset (Cytokines, Cytometry, Clonal Breadth/Depth, RNA data) along with preprocessing steps taken. Maybe i'll include a brief section here on hepatitis B preprocessing to set up the validation later.

In this chapter, I provide an overview of the measles dataset used in this study and outline the preprocessing steps required to prepare the data for analysis. The dataset integrates multiple biological modalities (including cytokine profiles, cytometry measurements, clonal breadth and depth of T cell receptors, and RNA sequencing data) to capture various aspects of the immune response to the measles vaccine. Additionally, I explain how response labels were assigned based on antibody titer trajectories, which serve as a key indicator of vaccine effectiveness.

3.1 Label Assignment Strategy

The original dataset included antibody titer measurements at four time points (Day0, Day21, Day150, and Day365), along with detailed qualitative classifications for each subject. These classifications, as noted in the original study [1], were derived using a hierarchical clustering method that groups individuals based on similar patterns in antibody titer evolution, effectively avoiding potentially biased cut-off values. Specifically, the study identified four response groups: High Ab, Low Ab, Long response, and Peak response.

It is important to acknowledge that these classifications, and the responder definition used in this study, are specific to measles. A significant factor in this dataset is the presence of pre-existing antibody titers due to prior exposure or vaccination. Consequently, some individuals with high pre-existing titers did not show an increase after vaccination, thus they don't technically respond to the stimulus.

For the initial stages of modeling, I simplified the labels to a binary classification (`responder` and `non-responder`). Responders were defined as those whose antibody titers increased

by at least 120 mIU/mL from Day0 to Day21, reflecting a widely accepted threshold for protective immunity against measles [5]. This simplification was done to create a more straightforward prediction task. However, the complexity of the original four classifications will be considered for future modeling refinements.

3.2 Antibody Titer Trajectories

3.2.1 By Response Label

Figure 3.1 shows the antibody titer trajectories over four time points, separated by each subject's final response label (`response` or `no response`). On the x-axis, we have the days at which titers were measured (0, 21, 150, 365). The y-axis represents the titer level. Each line corresponds to a single subject's progression across these time points. From this visualization, it is evident that "responders" generally exhibit a marked increase in titers between Day 0 and Day 21 (and sometimes up to Day 150), whereas "non-responders" show either a small rise or a plateau. This distinction validates the simplified two-class approach used for model building. In contrast, Figure 3.2 displays the antibody titer trajectories based on the original, more granular labels (i.e., `responder`, `no response - high ab`, and `no response - low ab`).

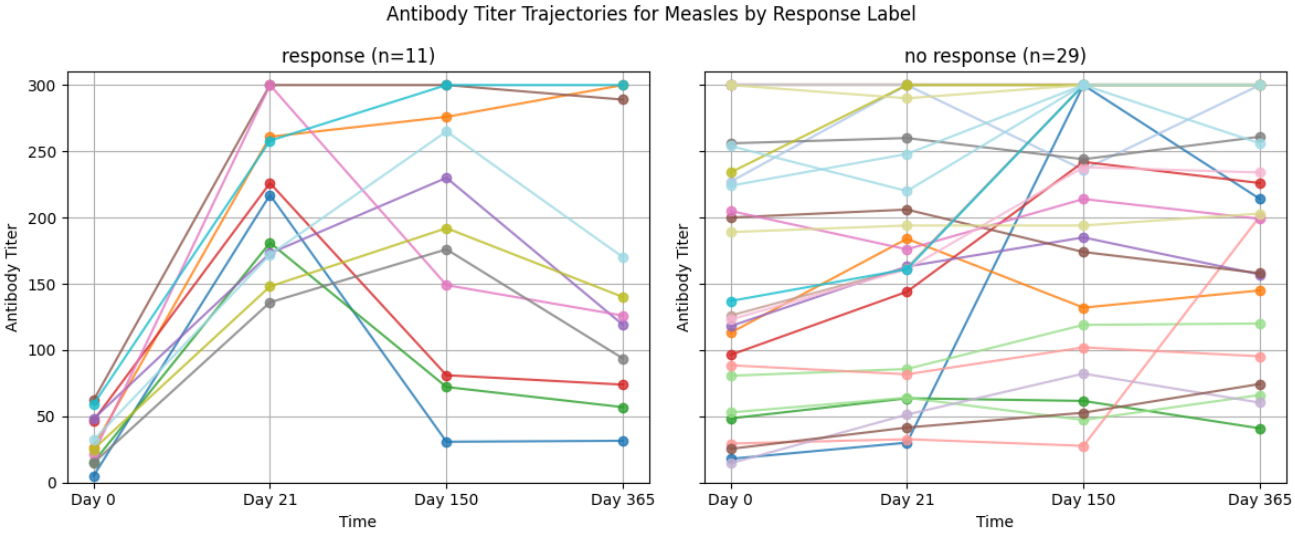

Figure 3.1: Antibody titer trajectories for each subject, colored by final response label. Subjects labeled as `response` are shown in the plot on the left, while those labeled as `no response` are shown on the right. The initial trajectory can be traced by following the color.

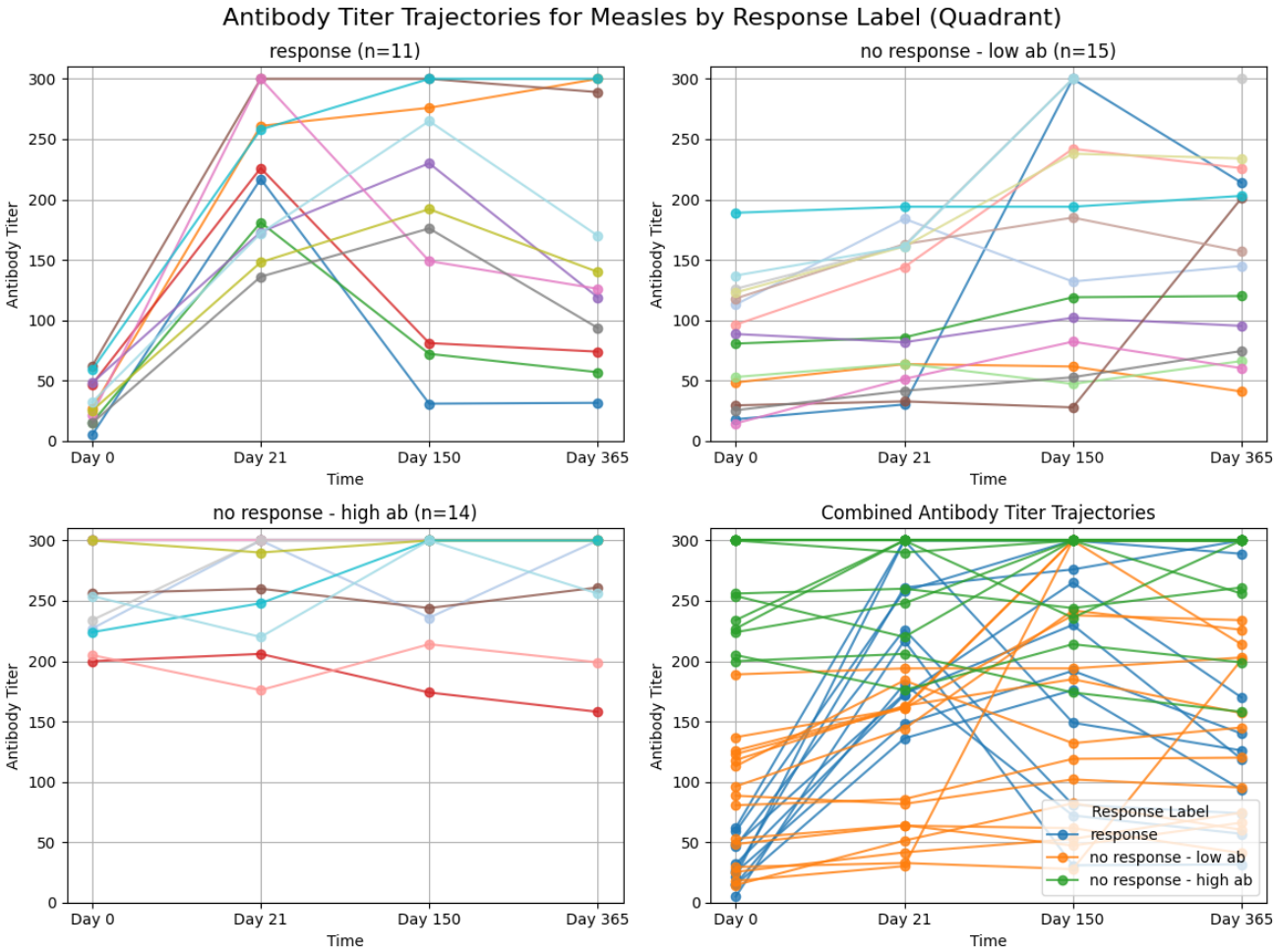

Figure 3.2: Antibody titer trajectories for each subject, colored by the original qualitative labels (quadrant). The plots show a clear differences in antibody responses, with responders displaying a sharp initial increase followed by gradual decline, while non-responders maintain consistently low or high titers.

The simplified `response_label` variable is used throughout the thesis as the target for classification tasks. This choice reduces the interpretability concerns. While the dataset initially contained more nuanced labels (e.g., distinguishing `no response - high ab` from `no response - low ab`), combining them into a single `no response` category improves the feasibility of training robust machine learning models given the limited sample size.

3.3 Correlation Analysis Within Individual Datasets

In an exploratory phase to gain familiarity with the data, I played around with Principal Component Analysis (PCA) to the full feature set, particularly focusing on the cytokines dataset. Although this analysis was not a definitive or robust evaluation, it provided valuable insights by showing that the first 10 principal components captured the majority of the variance. This finding suggests that much of the dataset's information can be summarized in fewer dimensions and indicates a high degree of redundancy among features. This observation led me to believe that similar correlations likely exist in the other datasets as well. Additionally, when I performed cross-validation using a Random Forest classifier on both the full and reduced feature sets, I observed that the balanced accuracy scores hovered near 50%. The model tended to predict only the majority class, even though the overall accuracy appeared acceptable due to class predominance. These preliminary findings underscore the challenges of high dimensionality, multicollinearity, and class imbalance that need to be addressed in subsequent predictive modeling efforts.

Following this, I delved deeper into understanding how the variables interrelate across the different data sources. As said earlier the study comprises five distinct datasets capturing various aspects of the immune response: cytokines, cytometry, clonal breadth (TCR metrics), clonal depth (TCR metrics), and RNA data. Since each dataset represents a unique facet of immunity, I first investigated correlations within each individual dataset. Next, I concatenated the datasets to perform an integrated correlation analysis. This strategy allowed me to uncover both within-modality and cross-modality relationships, facilitating effective clustering of the data and providing the models with informative, explanatory features.

3.3.1 Methodology

TODO: Check if correct...

In my analysis, I use a Weighted Gene Co-expression Network Analysis (WGCNA) framework to identify modules of highly correlated features. First, I compute a correlation matrix from the data, which quantifies the pairwise relationships between features. This correlation matrix is then transformed into a distance matrix by taking one minus the absolute correlation value, ensuring that strongly correlated features are considered close. Instead of relying on a fixed linkage method like Ward.D2, WGCNA integrates hierarchical clustering with network analysis to detect modules, or clusters, of co-expressed features. In the implementation, I use the `flashClust` package for efficient hierarchical clustering, and then apply a dynamic tree cut procedure to define the modules. The modules are further visualized by assigning each a unique color using WGCNA's labeling functions and displaying the resulting dendrogram and heatmap plots. This approach captures complex correlation patterns, effectively groups similar features.

3.3.2 Cytokine Data

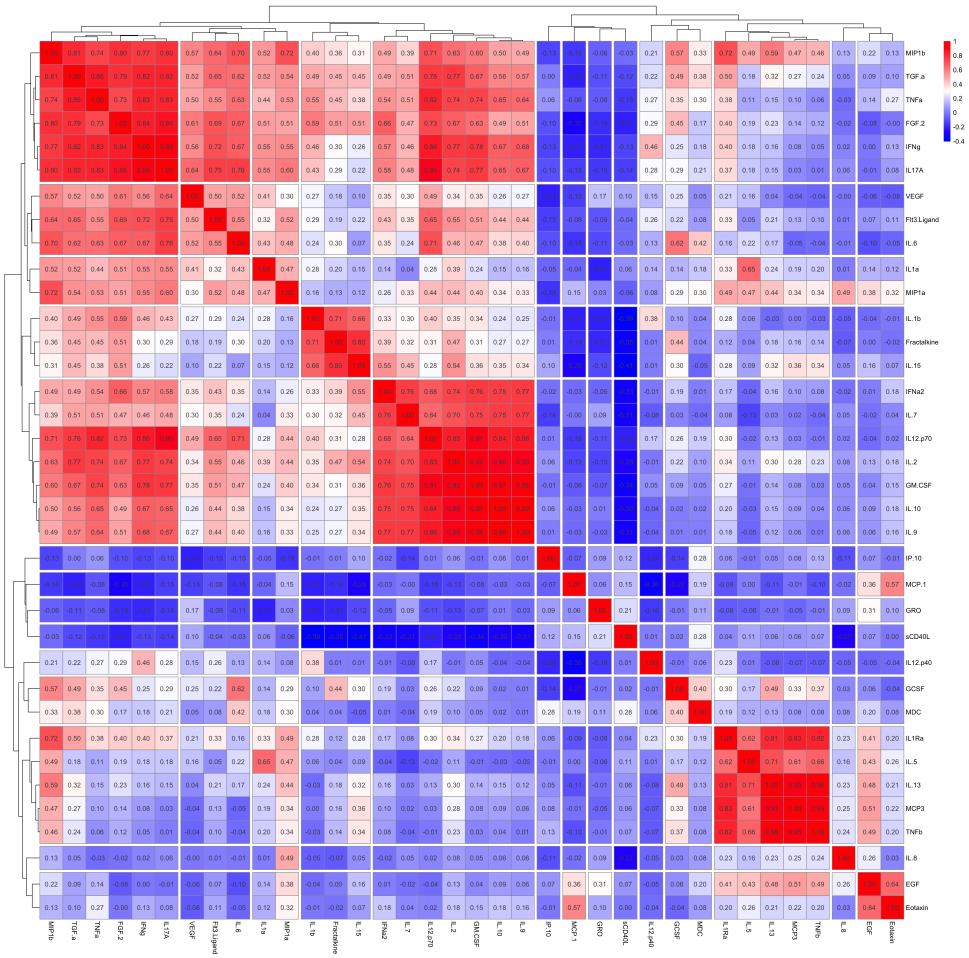

Figure 3.3: Correlation heatmap of the cytokine data using WGCNA hierarchical clustering.

The following cytokine clusters were obtained from the WGCNA hierarchical clustering analysis. The clustering reveals groups of cytokines with high inter-correlation, suggesting potential co-regulation or shared functional pathways.

Cluster 1:	MIP1β, TGF-α, TNF-α, FGF-2, IFN-γ, IL17A.
Cluster 2:	VEGF, Flt3 Ligand, IL-6.
Cluster 3:	IL1α, MIP1α.
Cluster 4:	IL-1β, Fractalkine, IL-15.
Cluster 5:	IFNα2, IL-7, IL12-p70, IL-2, GM-CSF, IL-10, IL-9.
Cluster 6:	IL1Ra, IL-5, IL-13, MCP3, TNFβ.
Cluster 7:	GCSF, MDC.
Cluster 8:	EGF, Eotaxin.

Table 3.1: Cytokine Clusters

3.3.3 Cytometry Data

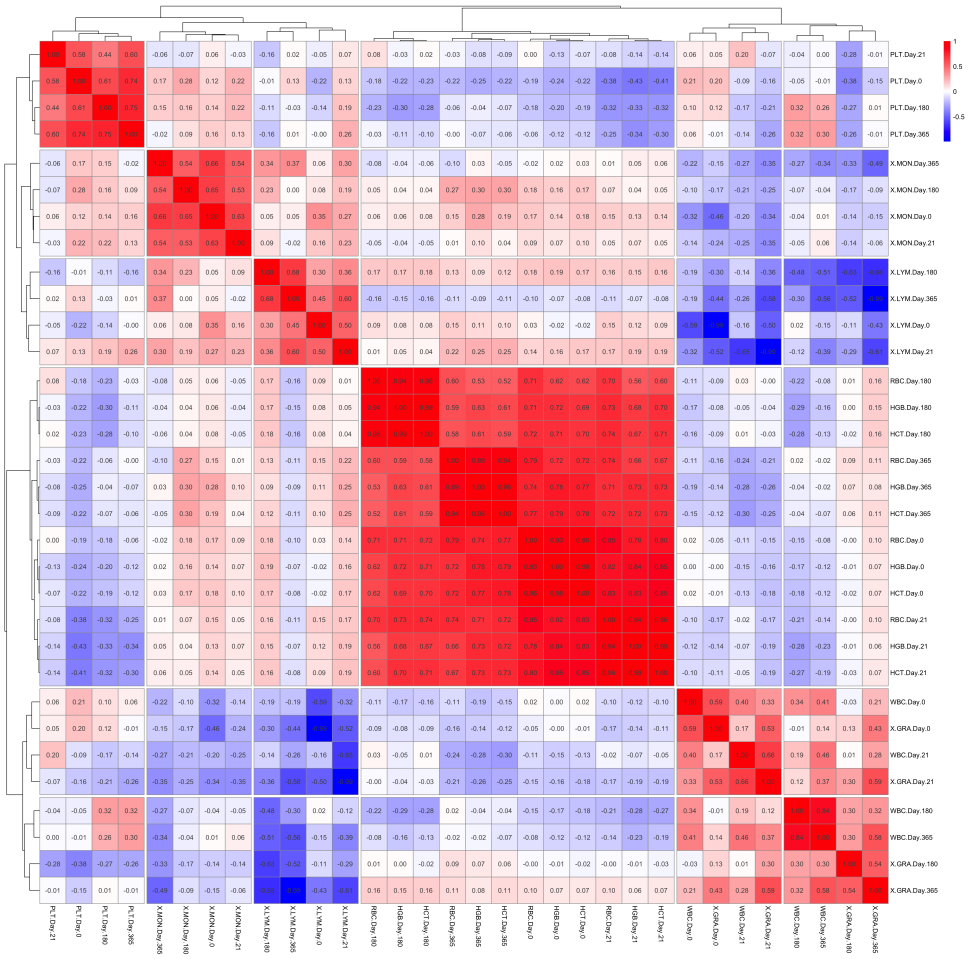

Figure 3.4: Correlation heatmap of the cytometry data using WGCNA hierarchical clustering.

3.3.4 TCR Metrics

At Day 1, the heatmap reveals high correlation coefficients (e.g., 0.89–0.99) among RBC, HGB, and HCT, and similar correlation patterns emerge at Day 0, Day 21, Day 180, and Day 365, reinforcing their consistent co-variation over time. Even when only Day 1 is considered.

Cluster 1: RBC, HGB, HCT

Table 3.2: Cytokine Clusters

3.3.5 RNA Data

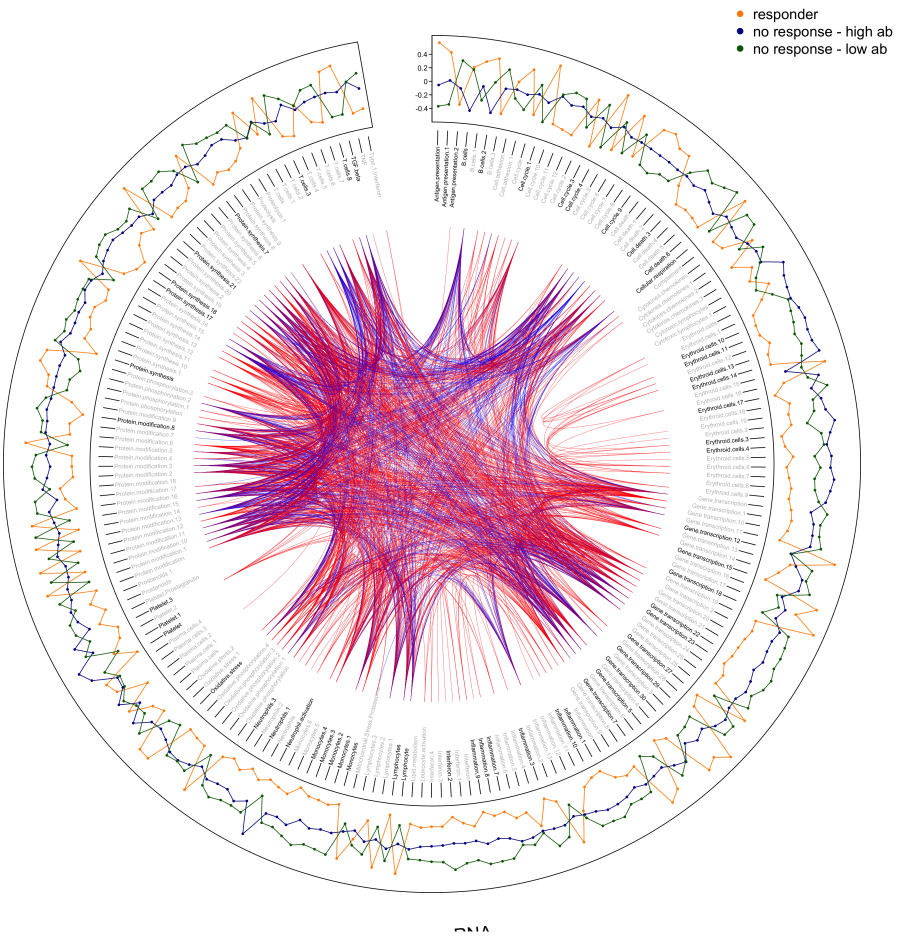

Figure 3.5: Circular correlation plot of the RNA data using. Each feature is represented along the outer edge, with colored lines indicating correlations between features (blue for positive correlations, red for negative correlations). The dense network of lines highlights the complexity of the dataset and the difficulty in visually identifying specific patterns or clusters.

The RNA dataset presented a greater challenge for correlation analysis due to the high dimensionality, with a total of 382 features. To explore potential relationships within this data, I employed various visualization techniques, including the circular plot shown above.

Although the plot provides a comprehensive overview of all possible correlations, the dense web of lines makes it difficult to discern specific patterns or clusters visually. However, the primary advantage of this visualization is that it allows for the identification of broad trends and the detection of features that may be particularly well-connected or influential.

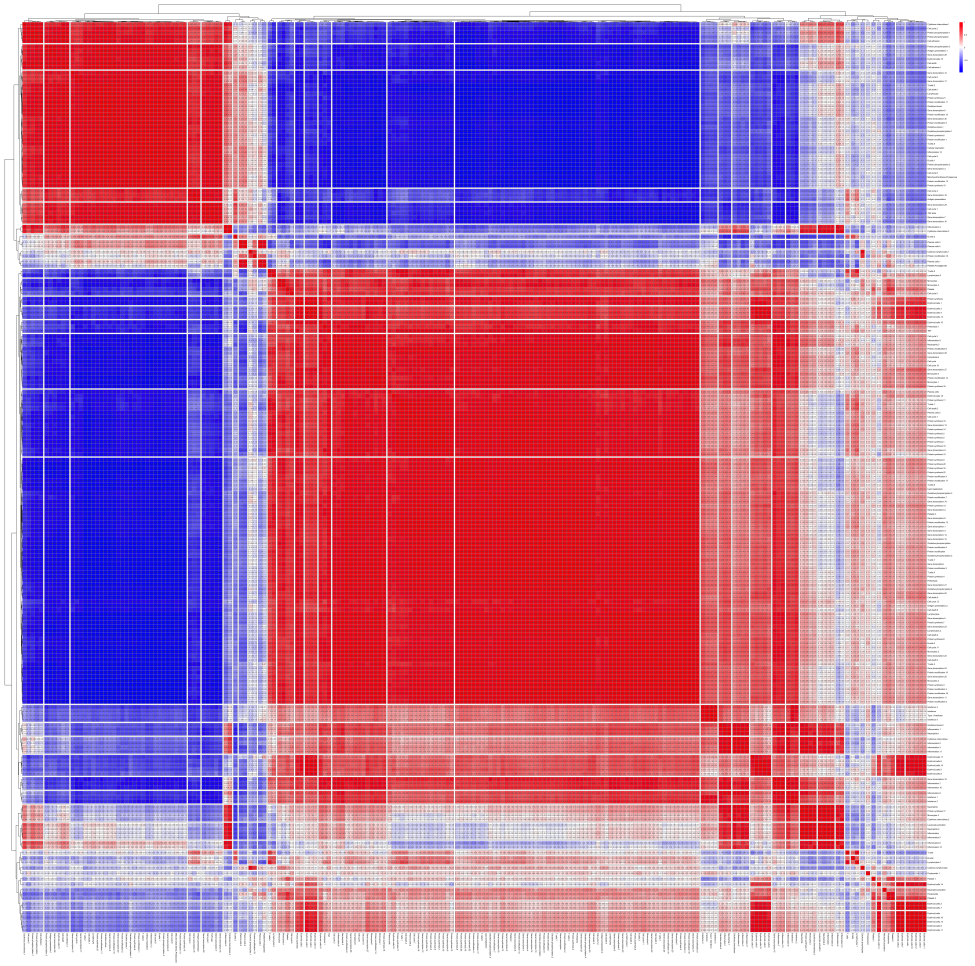

Figure 3.6: Correlation heatmap of the RNA data using WGCNA hierarchical clustering.

The same heatmap visualization technique as used before was applied to the RNA dataset, using a total of 36 cuts to correspond to the 36 different RNA modules identified through WGCNA hierarchical clustering. Selecting 36 cuts was challenging, but this number provided a reasonable balance between capturing distinct clusters and maintaining interpretability. The heatmap displays clear blocks of correlated features, with red indicating positive correlations and blue indicating negative correlations.

TODO: @Fabio - Do I include a huge table of the clusters, because that does not seem that usefull.

!REMARK: I think you can put it in the supplemental

CHAPTER 4

Methodology: Modeling and Feature Selection for Measles

4.1 Consensus Model Approach

In the initial phase of the modeling process, I opted for a consensus model approach to leverage the complementary strengths of multiple predictive models. This technique involves training several distinct models independently on the different datasets and then concatenating their responses to form a unified prediction. The reason behind this approach is that different models may capture unique aspects of the data or be sensitive to different feature sets, thus enhancing the overall robustness and generalizability of the predictions.

TODO: today-: differ from conventional consensus model

By combining model outputs, the consensus approach aims to mitigate individual model biases and reduce the risk of overfitting, particularly given the high-dimensional nature and small size of the dataset. This method also allows for the identification of stable predictive features across models, providing insights into which features consistently contribute to predictive performance. However, the complexity introduced by this approach requires careful consideration of how model outputs are integrated and evaluated.

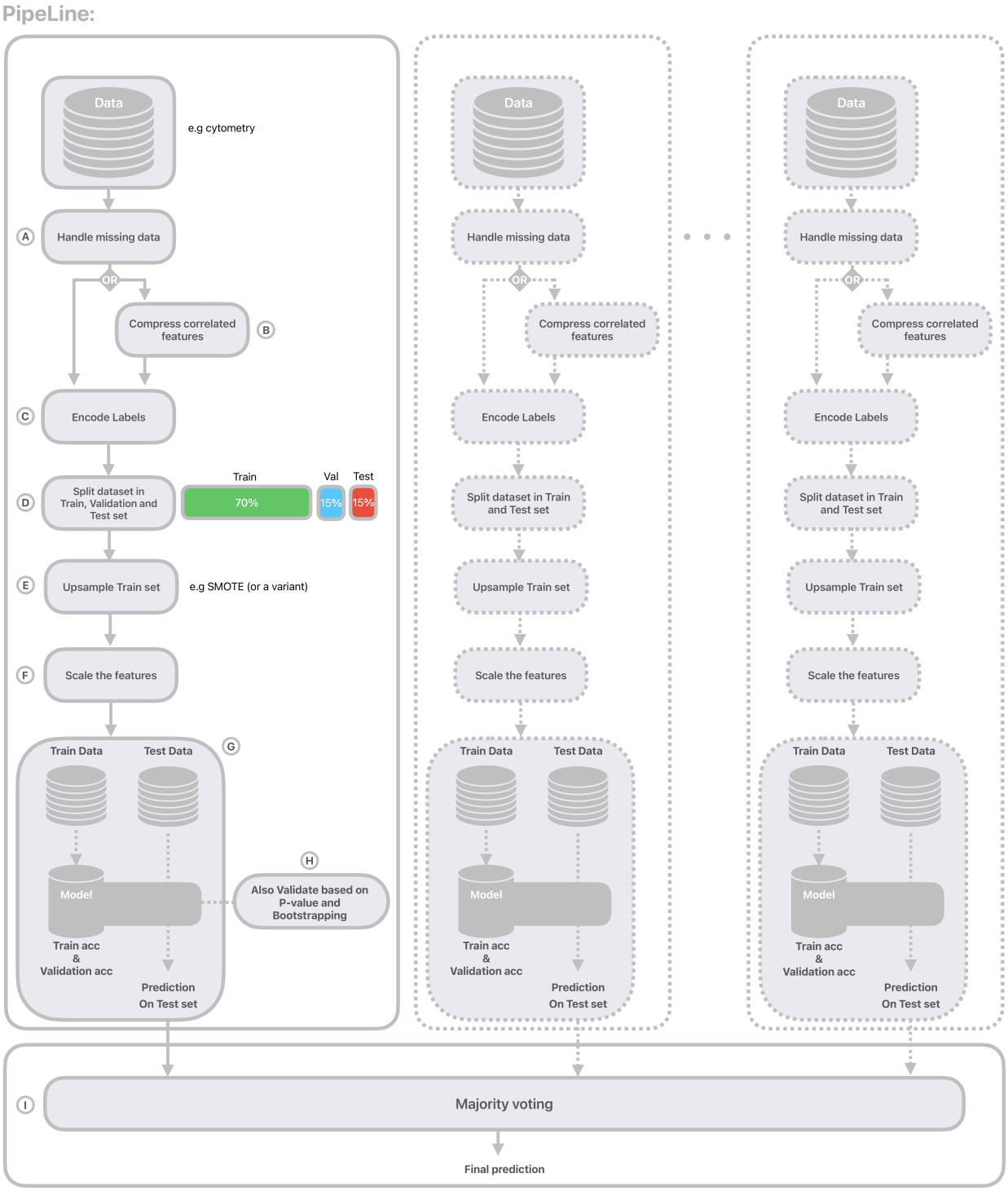

Figure 4.1: The diagram illustrates the multi-model pipeline used for prediction. Each individual model follows a standardized preprocessing workflow: handling missing data, compressing correlated features, encoding labels, splitting the dataset, upsampling the training set (e.g., using SMOTE or a variant), and scaling features. The trained models then produce predictions, which are evaluated based on training and testing accuracy. Additionally, validation is performed using p-value assessment and bootstrapping to ensure robustness. The final prediction is obtained through majority voting, aggregating predictions from all models to produce a consensus output.

4.1.1 Pipeline Structure for Heterogeneous Datasets

This section details the structure applied within the consensus modeling framework. Specifically, five parallel pipelines were executed, each dedicated to one of the following distinct data types: Cytokine measurements, Cytometry data, T-cell clonal breadth metrics, T-cell clonal depth metrics and RNA sequencing data.

Although each pipeline starts with different data to capture unique biological insights, maintaining consistency across them is essential. Therefore, the same standardized workflow is applied to all five pipelines after loading their specific data. Using this uniform process ensures each distinct data type is handled the same before the final predictions are combined. As illustrated in Figure 4.1, the standardized workflow applied to each dataset involves the following sequential steps:

Handle missing data (A)

There was incomplete data in the T-cell clonal breadth and depth datasets. Unlike the cytokine, cytometry, and RNA datasets, which included records for the full cohort of 40 patients, these two datasets were missing records for entire patients. To enable consistent downstream analysis across all five pipelines using the same set of 40 samples, a procedure was employed to impute the complete feature vectors for these missing patient records.

The core strategy leveraged the known response labels (e.g., responder/non-responder) for the complete patient cohort. First, patients absent from the clonal breadth or depth dataset but present in the full 40-sample cohort were identified. These missing patients were then conceptually grouped by their known response label. The rationale was that imputation for a missing patient should ideally reflect patterns observed in existing patients with the same outcome (e.g., imputing a missing 'responder' based on existing 'responders'). To perform the imputation, a SimpleImputer model using the mean imputation strategy was employed. This model was trained on the feature data from the available patients within the dataset and subsequently used to generate the feature values for the missing patients within their respective response groups.

While this imputation method ensures a consistent sample size, several potential limitations must be acknowledged. Primarily, generating entire patient records introduces synthetic data that may not fully capture true biological variability or complex feature interdependencies. Furthermore, guiding the imputation by response label relies on the assumption that feature profiles differ sufficiently between response groups; if this assumption is weak, the approach could introduce bias. The reliability of the imputed values also depends fundamentally on the quality and representativeness of the existing data used to train the imputer. Consequently, incorporating these imputed samples into downstream analyses carries the risk of influencing results, potentially by altering feature distributions, reducing variance (a known effect of mean imputation), or even overestimating model performance if the imputation unintentionally increases correlations with the response variable. Therefore, the findings from these 2 models are interpreted with these limitations in mind.

Compress correlated features (B)

This pipeline step focuses on addressing highly correlated features within each dataset, based on the findings presented in Section 3.3. To mitigate potential multicollinearity issues, these identified features were subsequently compressed into a single dimension using Principal Component Analysis (PCA), resulting in one principal component that represented them in the 'compressed' feature set.

To investigate whether feature compression impacted the final consensus model's performance, the modeling pipeline following this step was executed using two parallel approaches for each dataset. One approach utilized the original, full set of features, while the other used the reduced feature set. While addressing correlated features can potentially reduce redundancy and improve model stability, the main purpose of this parallel analysis was to empirically determine if this compression step offered a measurable advantage to the predictive accuracy of the overall consensus model for each specific data type.

Encode labels (C)

As machine learning algorithms typically require numerical inputs, the categorical target variable was converted into a numerical format in this step. This encoding simply assigned a unique integer (e.g., 0 and 1) to each response class.

Split dataset (D)

To properly evaluate model performance and ensure generalization to new data, the dataset was partitioned into three independent subsets: a training set, a validation set, and a test set.

The partitioning followed a 70% / 15% / 15% ratio, allocating the data as follows:

- **Training Set (70%):** This largest subset was used solely for training the models. The algorithms learn patterns, relationships, and parameters from this data.
- **Validation Set (15%):** This independent subset was crucial during the model development phase, specifically for selecting the optimal combination of model algorithm and data preprocessing choices. Since multiple distinct model types were evaluated in parallel across the pipelines, and different up-sampling techniques were considered (detailed in the next step), this validation set was used to compare their performance. The combination of model algorithm and up-sampling method yielding the best results on the validation set was selected for final evaluation. Performing the model and up-sampling selection using the validation set is critical to ensure that the test set remains completely untouched and unseen, thereby preserving its integrity for the final, unbiased assessment of generalization performance.
- **Test Set (15%):** This subset was strictly held out until all model development and selection were finalized based on the training and validation sets. It was used only once at the very end to provide a final, unbiased estimate of how the chosen model configuration is expected to perform on new, unseen data.

Up-sample Train set (E)

A common challenge in biological datasets is class imbalance, where one response class (in this case, responders) may be significantly less prevalent than the other within the training data. This may cause the model to favor the majority class during training. To counteract this and evaluate different mitigation approaches, three distinct strategies for handling class imbalance were applied exclusively to the 70% training set:

1. **No Resampling with Class Weighting:** In this strategy, the training data distribution was not modified. Instead, imbalance was addressed algorithmically during model training by assigning higher weights to the minority class samples. This typically adjusts the model's loss function, making errors on minority class examples more costly and forcing the model to pay more attention to them.

2. **Random Over-Sampling (ROS):** This method directly modifies the training set by increasing the representation of the minority class. It works by randomly selecting and duplicating existing samples from the minority class until a desired level of balance (an equal number of samples per class) is achieved.

3. **SMOTE (Synthetic Minority Over-sampling Technique):** Generating new, synthetic minority class samples through feature space interpolation, as detailed further below and illustrated in Figure 4.2.

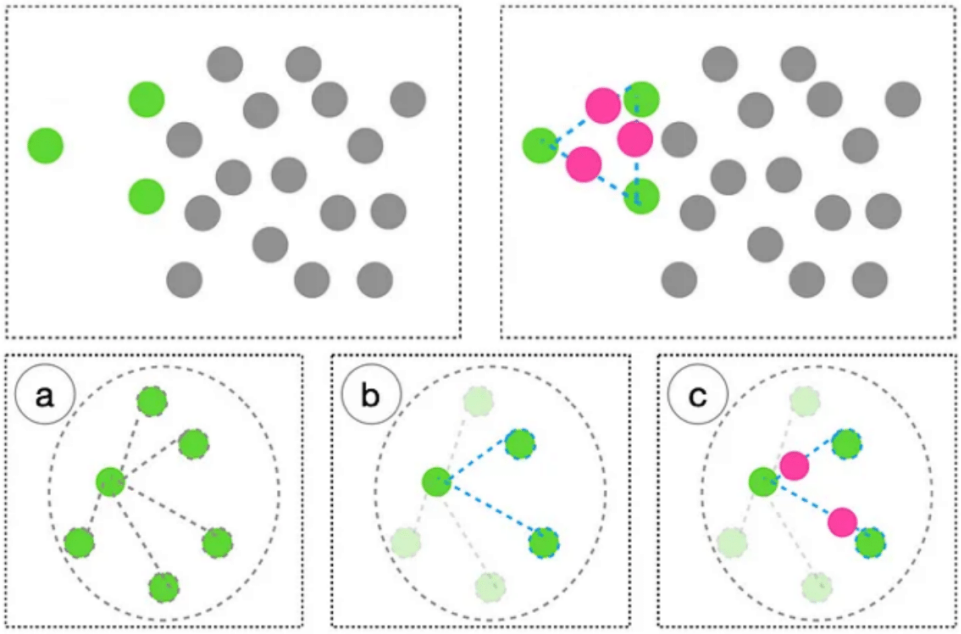

Figure 4.2: Illustration of the Synthetic Minority Over-sampling Technique (SMOTE) mechanism. Top left: Initial imbalanced data distribution (minority in green, majority in gray). Top right: Generation of synthetic samples (pink) along lines connecting a minority instance to its nearest minority neighbors. Bottom row: Detailed steps showing (a) identification of k-nearest minority neighbors (here k=5), (b) selection of neighbors for synthesis, and (c) creation of synthetic samples (pink) along the vectors to selected neighbors. Figure based on figures 1 and 2 presented in [10].

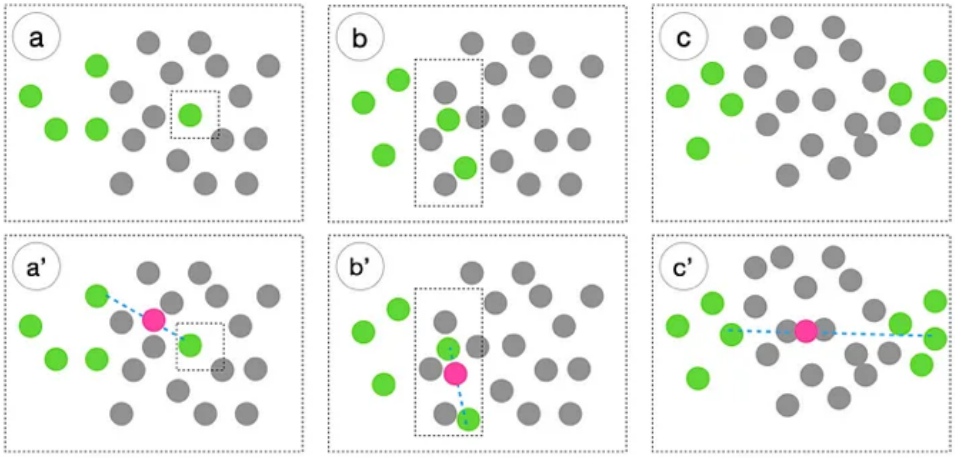

Figure 4.3: Conceptual illustration of potential weaknesses associated with SMOTE. The bottom row (a', b', c') depicts synthetic sample generation (pink) relative to regions shown in the top row (a, b, c). Potential issues illustrated include: (a/a') Generation influenced by potential noise or outliers (isolated minority sample). (b/b') Synthetic samples potentially overlapping with dense majority class regions due to disregard for majority sample proximity. (c/c') Generation possibly bridging distinct minority clusters, which could ignore underlying data structure or cause overgeneralization. Figure adapted from figure 4 presented in [10].

SMOTE [4], visually explained in Figure 4.2, generates synthetic minority samples rather than simply duplicating existing ones like ROS. It selects a minority instance, finds its k-nearest minority class neighbors, and creates new samples along the line segments joining the instance to some of these neighbors. This interpolation can potentially create a more diverse minority representation and smoother decision boundaries.

However, SMOTE also has known limitations (illustrated conceptually in Figure 4.3), as discussed in literature [10]. Potential drawbacks include overgeneralization (creating samples that blur class distinctions or ignore sub-clusters), amplification of noise if based on outlier samples, and possible generation of samples too close to or overlapping with the majority class, as the original algorithm doesn't explicitly consider majority proximity. Despite these points, SMOTE was included for comparison as a prominent synthetic data generation technique.

The relative effectiveness of these three imbalance-handling strategies (Class Weighting, ROS, SMOTE) was assessed empirically for each model pipeline. As established during data splitting (Section **??** - *ensure this label is correct*), the performance metrics achieved on the 15% validation set were used to select the optimal strategy for each model before final assessment on the test set. It is crucial to reiterate that these imbalance adjustments were applied strictly to the training data partition to maintain the integrity of the validation and test sets.

Scale the features (F)

TODO: today-...

Train the models (G)

TODO: today-...

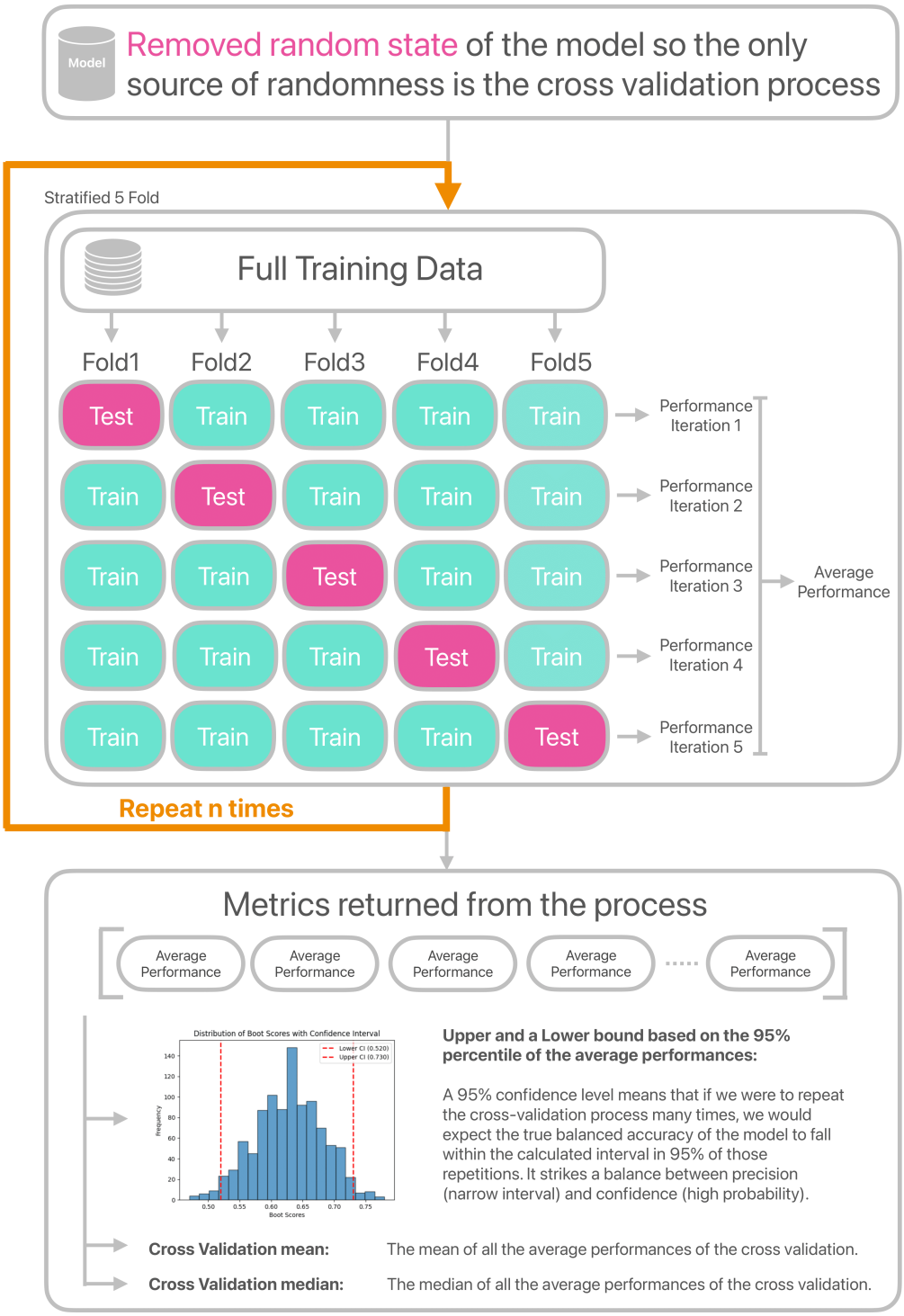

Figure 4.4: This diagram illustrates the process of estimating the confidence interval of a model's performance using repeated stratified k-fold cross-validation. Initially, the model's random state is removed to ensure the cross-validation process is the sole source of randomness. Stratified k-fold cross-validation (k=5) is used to evaluate model performance, with the average score across folds representing a single iteration's result. This entire cross-validation procedure is repeated 'n' times (e.g., 1000) to simulate the variability in performance across different random splits. The resulting 'n' average performance scores are then used to calculate the confidence interval. These bounds indicate the range within which we expect the model's true performance to lie with a high degree of confidence, providing a robust measure of the model's reliability.

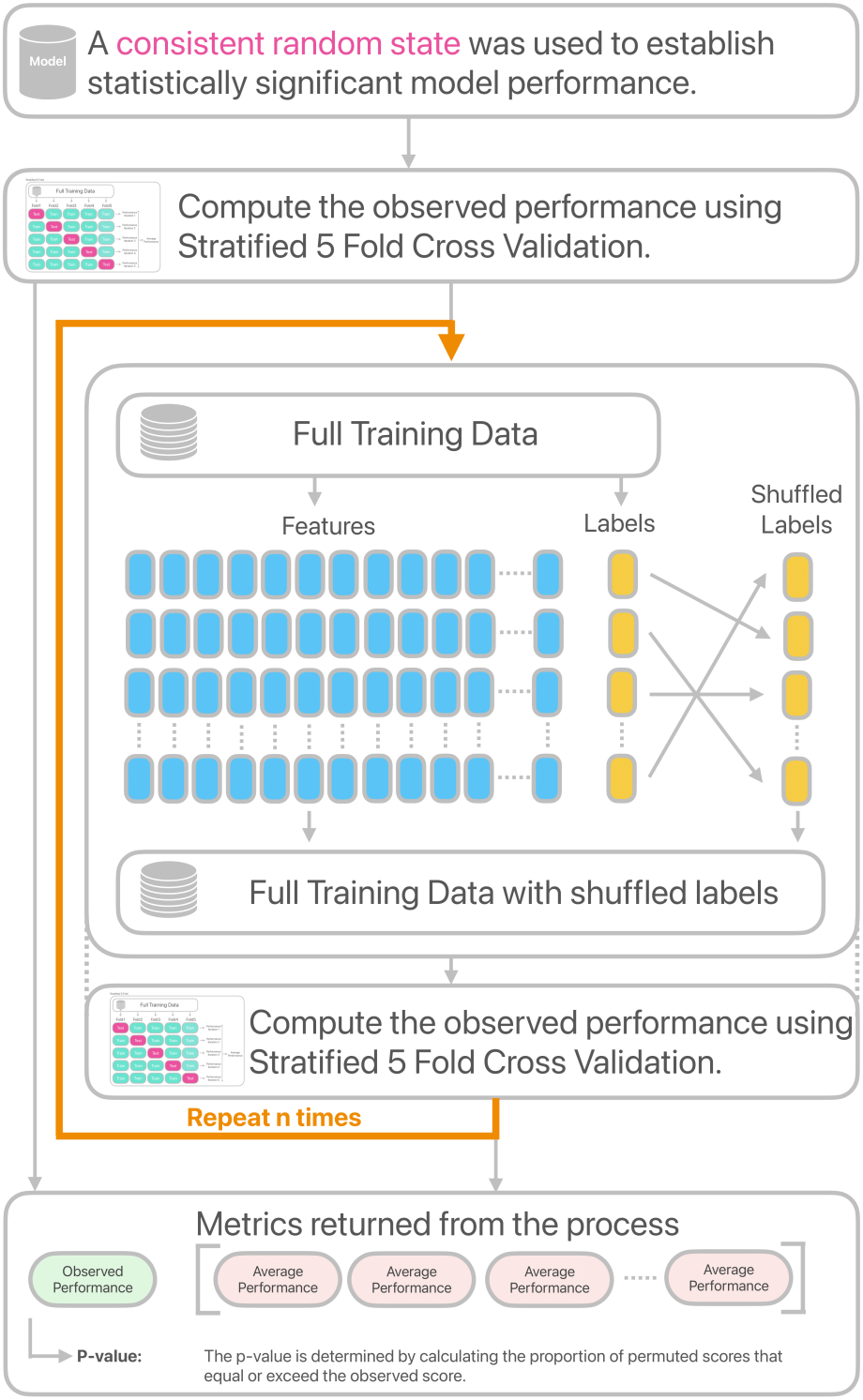

Figure 4.5: This diagram illustrates a permutation test designed to evaluate the statistical significance of a model's performance. Initially, a consistent random state is set to ensure reproducible results. The model's performance is first assessed on the original data using stratified 5-fold cross-validation, yielding the 'observed performance.' Subsequently, to determine if this performance is merely due to chance, the labels are repeatedly shuffled. For each shuffled dataset, the model's performance is re-evaluated using the same stratified 5-fold cross-validation. This process is repeated 'n' times, resulting in a distribution of 'permuted performances.' Finally, the p-value is calculated as the proportion of permuted performances that equal or exceed the observed performance, indicating the likelihood of obtaining such performance by chance alone.

CHAPTER 5

Results for the Measles Pipeline

TODO: Present model performance, feature importance, and interpretation of findings from the measles data.

CHAPTER 6

Cross-Vaccine Marker Validation with Hepatitis B

TODO: Describe the application of the same pipeline to the hepatitis B dataset, compare predictive features, and discuss the validation process.

CHAPTER 7

Discussion and Conclusion

TODO: Summarize insights, implications for vaccine response prediction, and potential future work based on the comparative analysis.

CHAPTER 8

Future work

Bibliography

[1] E. Bartholomeus, N. De Neuter, A. Suls, G. Elias, S. van der Heijden, N. Keersmaekers, H. Jansens, V. Van Tendeloo, P. Beutels, K. Laukens, B. Ogunjimi, G. Mortier, P. Meysman, and P. Van Damme. Transcriptomic profiling of different responder types in adults after a priorix® vaccination. *Vaccine*, 38(16):3218–3226, Apr 3 2020. Epub 2020 Mar 9; PMID: 32165045.

[2] Petter Brodin and Michael M. Davis. Human immune system variation. *Nature Reviews Immunology*, 17(1):21–29, 2017. Epub 2016 Dec 5; PMID: 27916977; PMCID: PMC5328245.

[3] M R Castrucci. Factors affecting immune responses to the influenza vaccine. *Human Vaccines & Immunotherapeutics*, 14(3):637–646, Mar 2018. Epub 2017 Jul 21; PMID: 28617077; PMCID: PMC5861809.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, jun 2002.

[5] R. T. Chen, L. E. Markowitz, P. Albrecht, J. A. Stewart, L. M. Mofenson, S. R. Preblud, and W. A. Orenstein. Measles antibody: reevaluation of protective titers. *Journal of Infectious Diseases*, 162(5):1036–1042, 1990.

[6] C. A. Janeway, P. Travers, M. Walport, and M. J. Shlomchik. *Immunobiology (9th Edition)*. Garland Science, 2001. innate vs adaptive immunity.

[7] Walter J. Moss. Measles. *Lancet*, 390(10111):2490–2502, Dec 2 2017. Epub 2017 Jun 30; PMID: 28673424.

[8] National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General*. Centers for Disease Control and Prevention (US), Atlanta, GA, 2014. Figure 10.5, Diagram of innate and adaptive immunity. Available from: https://www.ncbi.nlm.nih.gov/books/NBK294318/figure/ch10.f5/.

[9] Stanley A. Plotkin. Correlates of protection induced by vaccination. *Clinical Vaccine Immunology*, 17(7):1055–1065, Jul 2010.

[10] An Truong. Imbalanced Data ML: SMOTE and its variants. Blog post in TotalEnergies Digital Factory on Medium, jun 2022.

[11] Matthieu Van Tilbeurgh, Katia Lemdani, Anne-Sophie Beignon, Catherine Chapon, Nicolas Tchitchek, Lina Cheraitia, Ernesto Marcos Lopez, Quentin Pascal, Roger Le Grand, Pauline Maisonnasse, and Caroline Manet. Predictive markers of immunogenicity and efficacy for human vaccines. *Vaccines (Basel)*, 9(6):579, 2021.