Error Profiling of Machine Learning Models: An Exploratory Visualization

Jeffrey Feng* J64FENG@G.UCLA.EDU
Medical Imaging Informatics Group, Department of Radiological Sciences
University of California Los Angeles
Los Angeles, CA, USA

Al Rahrooh* ARAHROOH@G.UCLA.EDU
Medical Imaging Informatics Group, Department of Radiological Sciences
University of California Los Angeles
Los Angeles, CA, USA

Alex A.T. Bui, PhD BUIA@MII.UCLA.EDU
Medical Imaging Informatics Group, Department of Radiological Sciences
University of California Los Angeles
Los Angeles, CA, USA

Abstract

While data-driven predictive models are increasingly used in healthcare, their clinical translation remains limited, in part due to challenges in evaluating model performance across design choices. Comparing model performance based on aggregate performance metrics is useful for benchmarking, but it can obscure important details such as differences in decisions and which patient groups are affected by those differences. We present a visualization-based error profiling method that facilitates comparative evaluation by highlighting overlaps and differences in model predictions. This approach enables deeper insight into which (sub)populations are consistently (in)correctly classified across models, helping uncover patterns of model (dis)agreement to assess the impact of modeling decisions. We demonstrate our visualization method in four healthcare use cases: 1) missing data imputation in a longitudinal nutritional dataset; 2) feature set analysis using randomized controlled trial data; 3) end-model technical performance in cardiac morbidity prediction; and 4) data modality comparison using a lung cancer dataset with longitudinal and radiomic features. To evaluate the visualization, we obtained expert feedback and qualitative assessments of decision-making insights. Survey results across clinicians, computer scientists, and medical informaticians indicated that our method provides an interpretable and intuitive way to compare model error distributions by highlighting patterns within (in)correctly classified subpopulations across model types. Our error profiling approach represents an initial step toward a systematic framework for improving model assessment in clinical tasks. Through this framework, both model developers and end users can better understand when and where a given model is appropriate for real-world clinical deployment.

1. Introduction

The proliferation of health-related and clinical data promises to unlock ways to develop data-driven approaches to improve patient care Rajpurkar et al. (2022). But the existence of this data does not guarantee its utility, as the path from raw data to actionable insights is

fraught with challenges McCausland (2020). Moreover, given the complexity of the human condition and disease, neither expert physicians nor artificial intelligence (AI) models are correct *all the time*. When deploying clinical AI models, understanding *how* and *when* they fail is just as important as knowing how well they perform. Indeed, given the potential consequences of incorrect predictions, it is essential to characterize the nature of a model's errors to determine when it is appropriate for use – and when it is not. Error analysis is therefore critical, particularly for uncovering issues such as bias, subgroup disparities, or inconsistent decision boundaries. However, most current evaluations emphasize aggregate performance metrics, such as the area under the receiver operating characteristic curve (AUROC), which summarize performance at the population level. These metrics often obscure important nuances: two models may have similar AUROCs yet make different predictions on individual patients. When differences in performance are observed, it is crucial to understand why they arise, and which patient subgroups are affected. Even when metrics align, model behavior may diverge in clinically meaningful ways. Just as drug labels outline contraindications, AI tools should be transparent about the scenarios in which they may be unreliable. Without this level of insight, trust in AI-supported clinical decision-making will remain limited.

In this light, we describe an error profile visualization strategy that enables comparisons of different models to inform downstream assessment and deployment decisions. Our approach draws inspiration from existing standards and guidelines in AI evaluation that emphasize the need for rigorous and transparent assessment Ayers et al. (2020); WHO (2021); IMDRF (2025); Tabassi (2023). We focus on three common aspects of model development to highlight the flexibility of error analysis. First, different machine learning (ML) algorithms often yield varying performance metrics, making it challenging to select the most suitable model for clinical applications Flach (2019); Rainio et al. (2024). Second, various data imputation techniques can lead to significant variance in reconstructed data, impacting the reliability of a learned predictive model Jadhav et al. (2019); Joel et al. (2025); Li et al. (2024). Third, the use of different sets of features to train and evaluate models can introduce additional uncertainty in model performance Li et al. (2017); Noroozi et al. (2023); Mohtasham et al. (2024). Despite the need for systematic error analyses in these contexts, there are no standard practices Beck et al. (2022), obstructing the adoption of ML models in clinical settings Patel et al. (2009); Kelly et al. (2019); Petersen et al. (2023). To address this gap, we aim to provide a structured framework for error profiling that enhances the interpretability and explainability of predictive models in healthcare.

Generalizable Insights about Machine Learning in the Context of Healthcare

This work contributes a model-agnostic framework for error analysis that advances how we evaluate ML models in healthcare. Rather than relying on aggregate metrics, our method visualizes how models agree (or diverge) in their errors across patient subpopulations, exposing meaningful patterns that traditional metrics obscure. This ability enables:

- *Subpopulation-level model comparison.* We show that models with similar performance scores can behave very differently, and our method identifies those patient groups that are consistently misclassified or handled inconsistently across models.

- *Design-informed evaluation.* By applying the method across use cases involving different architectures, data imputation strategies, feature sets, and multimodal inputs, we reveal how modeling decisions impact real-world performance.
- *Deployment-relevant insights.* Our approach helps to surface clinically-meaningful contraindications for model use – analogous to drug warnings where it is clear when a drug should not be used – empowering practitioners to make better-informed decisions about when and where to rely on AI tools.
- *A step toward standardizing model auditability.* The method supports a broader shift toward transparent and structured model assessment pipelines that are both interpretable to clinicians and actionable for developers.

Together, these insights offer a general-purpose tool for evaluating not just *how well* a model performs, but *when*, *where*, and *why* it may succeed or fail in clinical practice – key factors for safe, equitable deployment.

2. Related Work

Error analysis has received growing attention in ML as practitioners increasingly recognize that high-level performance metrics can obscure important model behaviors. Existing tools such as Microsoft Azure's Error Analysis module within the Responsible AI dashboard Ignite (2025) focus on *intra-model* assessment, helping developers identify cohorts where a single model underperforms. These tools are valuable for improving reliability and fairness, particularly in detecting subpopulations with disparate error rates. However, they typically do not support *inter-model* comparisons, which are crucial when selecting among multiple candidate models or understanding the tradeoffs introduced by different modeling choices – especially in safety-critical domains like healthcare.

Several systems have been developed to improve model interpretability, including visual analytics platforms Li et al. (2020), DeepCompare Murugesan et al. (2019), VMS He et al., and LEGION Das and Endert (2020). While these tools support exploration of feature contributions, saliency, or decision boundaries, their focus remains primarily on explaining a model's internal logic rather than comparing *how different models behave on the same data*. For instance, DeepCompare visualizes differences in learned representations between deep models, but does not explicitly characterize differences in misclassification patterns across patient subgroups. LEGION emphasizes layered explanation for model reasoning but is not designed to identify when multiple models succeed or fail on the same patients.

In the healthcare domain, comparative model assessment has typically relied on aggregate performance metrics such as AUROC or precision-recall curves Rajkomar et al. (2019); Haenssle et al. (2018); Cabitza et al. (2017). While useful for benchmarking, these metrics do not capture whether different models make the same or different decisions, nor do they indicate which patient groups are affected by those differences. Some recent work in model auditing and fairness — such as subgroup performance disaggregation and disparity metrics — has highlighted the need for more granular analysis Obermeyer et al. (2019); Chen et al. (2020), but these approaches often stop short of providing actionable visual tools for comparison across modeling pipelines.

While the general concept of reviewing misclassified cases is well-established in the clinical domain, our contribution lies in enabling inter-model error comparisons through a model-agnostic, population-level error profiling visualization that highlights areas of agreement and disagreement across multiple models. Unlike intra-model cohorting tools or feature-level visualizations, our method captures the *interaction between models and patient subpopulations*, revealing which patient subgroups are consistently problematic across modeling approaches. This enables a structured comparison of model behaviors across design axes such as architecture, input modality, imputation strategy, and feature selection. Our framework augments existing error review processes, specifically targeting the gap in comparative model evaluation.

3. Methods

3.1. Rationale for the Error Profile

Unlike existing error analyses in ML that concentrate on aggregate metrics or techniques to uncover explanations or errors within a single model He et al.; Levman et al. (2023); Wang et al. (2019); Lundberg and Lee (2017); Xuan et al. (2022); Ming et al. (2019); Ribeiro et al.; Kwon et al. (2019); Liu et al. (2018), our approach of comparing multiple models presents a unique perspective on assessing model performance. A key motivation is that we expect that different models may yield different decisions when applied to the same data. Some instances may be incorrectly classified by all models, while others may be incorrectly classified by some models, one model, or none of the models.

We employ error profiles to observe patterns within the (in)correctly classified (sub)-populations by different models. By visualizing the overlaps and differences between the decisions of different ML models, we can not only discern which models are faltering and on what specific instances, but also which predictions are consistently (in)accurate across different models. For example, if some models tend to misclassify the same subset of individuals, it might suggest commonalities in the data that the models find challenging to interpret (and in comparison to human experts, may have also been a difficult case to judge), revealing areas where the data may lack the necessary signal or where the models may require adjustments to better capture the underlying patterns. If models misclassify different sets of individuals, it might suggest that the models preferentially identify different nuances in subpopulations of the data and are more suitable for specific use cases. Any standout subgroups of patients can be potentially isolated for stratified analysis, with an opportunity for case reviews with clinical experts to identify confounders or missing information.

Importantly, error profiling aims to provide a window into answering the question, *all else being equal, what are the differences between different models' decisions*? Error analysis, in general, provides a way to not only find opportunities for underlying model improvement, but provides explanatory information to aid decision-making during clinical implementation.

3.2. Creating and Visualizing Error Profiles

The error profiles illustrate sets of instances (e.g., subjects, patients, etc.) that different models classify incorrectly. Conventionally, Venn diagrams can succinctly visualize the relationships (e.g., the intersections and differences) between the decisions of two or three

sets using circles. It is also possible to use spheres to produce a 3D visualization comparing four sets, while ellipses can yield 2D visualizations comparing four or five sets Venn (1880); Grünbaum (1975, 1992a); Hamburger and Pippert (2000). When comparing any higher number of sets, Venn diagrams quickly become visually uninterpretable with several limiting characteristics (Figure A1). Notably, higher-dimensional symmetric Venn diagrams are only possible using complex shapes when comparing a prime number of sets Henderson (1963). Examples were only available for five, seven, and eleven sets Grünbaum (1975); Hamburger (2002); Mamakani et al. (2012); Edwards (1998); Grünbaum (1992b) before discovering the procedure that generalizes to any prime number Griggs et al. (2004). In contrast, the limiting number of sets for simple-symmetric Venn diagrams remains at eleven Mamakani and Ruskey (2012); Ruskey et al. (2006). Nonetheless, while it is theoretically conceivable to draw higher-dimensional Venn diagrams, they lack the scalability to visually interpret the potentially complex intersections of multiple models' decisions.

Instead of Venn diagrams, we use a scalable tabular structure representing the intersections of an arbitrary number of sets (Figure A1). To motivate this idea, we revisit the origin of Venn diagrams as a schematic for representing logical combinations of terms without enumerating all possibilities in a table Venn (1881). Extending upon the notion of succinctly visualizing information, we adapt a scalable tabular structure capable of simplifying the representation of complex Venn diagrams. The template shown in Figure 1 illustrates our visualization and highlights the shared components that can generalize over different use cases and compare different models' decisions.

To generate the profile, we require that the decision space of each ML model denoted ω results in either a correct or incorrect classification of each sample, with the incorrect classification of the ith sample in a dataset by one model expressed as $\omega_k \models \alpha_i$. A set of models incorrectly misclassifying the same sample is denoted $Mods(\alpha_i) = \{\omega : \omega_k \models \alpha_i\}$. We define a *group* G_j as a tuple consisting of: 1) a unique set β_j of n_j incorrectly classified patients; 2) a unique set g_j of $|g_j|$ ML models.

$$G_j \equiv (\beta_j = \{\alpha_1, \alpha_2 \ldots \alpha_{n_j}\}, g_j) \quad s.t. \; \forall \; i = 1, 2, \ldots n_j \; , \; Mods(\alpha_i) = g_j$$

For a succinct error profile, given $j = 1, 2 \ldots m$ groupings, each group should have a unique set of worlds such that $g_1 \neq g_2 \neq g_3 \ldots \neq g_m$, and each sample should belong to only one group such that $(\beta_1 \cap \beta_2 \cap \beta_3 \ldots \cap \beta_j) = \emptyset$. Given these definitions, we create the visualization as follows:

- As the goal of the error profile is to compare different models' decisions, the rows of the table represent the different models ω_k.
- The columns represent the groups G_j consisting of a distinct set of incorrectly classified patients β_j, where all patients are misclassified by the same set of models g_j.
- The number of incorrectly classified samples n_j in the set β_j associated with the group G_j is displayed in the bottom panel. The number of models $|g_j|$ incorrectly classifying the samples associated with the group G_j is displayed in the top panel.
- Each colored cell indicates that the model from the corresponding row incorrectly classified the set of samples associated with the cell's column. When viewing a column,

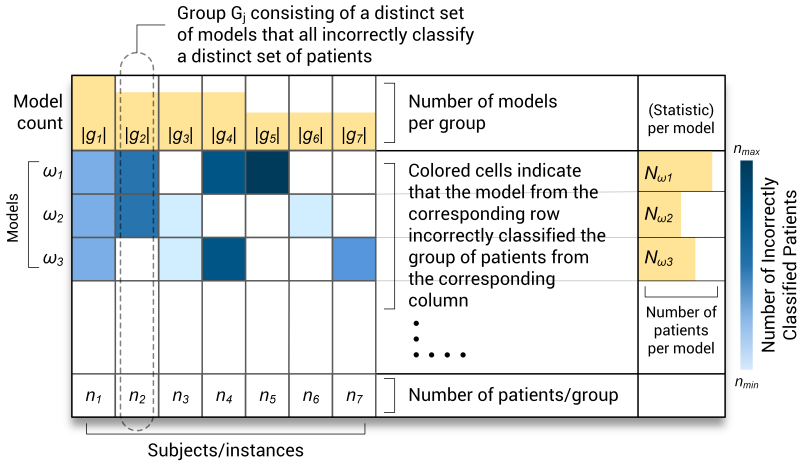

Figure 1: General template of the error profiling visualization.

the colored cells depict the models that belong to the set g_j. When viewing a row, the colored cells depict the sets of samples β_j incorrectly classified by one model. The color intensity reflects the number of incorrectly classified samples associated with each cell's column and maps to the color bar on the right of the profile.

- The right panel shows the number of samples N_{ω_k} that a model incorrectly classifies.

Through this visualization, it becomes possible to understand the extent to which: 1) models agree through common errors in classification decisions; 2) challenging (sub)-populations preclude the successful classification from one or more models; and 3) models differ when they disagree on decisions despite potentially similar performance metrics. A full formulation of the error profile with preliminaries is described in the Appendix. An example of error profile generation from data is shown in Figure A2.

3.3. Data and Use Cases

We explore four different use cases of classification-based models with the intent of capturing different stages of AI model development and demonstrating the flexibility of the error profile. Similar model development pipelines are used in each case. We present these use cases as follows, with further details of each dataset in the Appendix:

1. *Use Case 1: Error profiling the use of different ML algorithms.* Using an extra corporeal membrane oxygenation (ECMO) dataset, we tested six different ML architectures: random forest, RF; extreme gradient boosting, XGBoost; light gradient-boosting, LGBoost; logistic regression, LR; multilayer perceptron, MLP; and support vector machine, SVM. In theory, each model may learn a slightly different feature space and hence we would expect that there are situations that all ML models can readily learn, some cases that different models may learn, and a potential subset that no model may learn. For this application, each ML architecture is considered a distinct world ω and is represented in the rows of the profile. Our visualization helps

to elucidate common errors shared across different sets of architectures g_j, identify (sub)populations β_j that are particularly challenging for different sets of models, and appreciate differences between models' decision occluded by performance metrics.

2. *Use Case 2: Error profiling the effect of data imputation techniques on the performance of a ML model.* As the availability of information in a self-reported nutritional dataset (e.g., from MyFitnessPal, MFP) is highly inconsistent, we seek to understand how different imputation techniques would influence model performance at predicting if individuals would meet their weight goals. It is not always clear when to use simple techniques (e.g., means) instead of more sophisticated and exact imputation (e.g., multiple imputation by chained equations, MICE; k-nearest neighbors imputation, KNN), as simple methods may perform equally well depending on the domain, problem, and reason(s) for missingness Nijman et al. (2022). Comparing these methods within a single type of model and/or across model types would therefore provide insight into the choice of imputation technique. When applying our error profile, the missingness techniques are the worlds ω represented in the rows of the profile. The visualization helps to understand the degree to which the utilization different sets of imputation methods g_j lead to the successful classification of certain populations β_j.

3. *Use Case 3: Error profiling of feature sets utilized in the training and evaluation of a ML model.* With the range of observations now possible in healthcare, multimodal models that combine many inputs to generate a prediction are increasingly common. Still, it is often difficult to know which features are pertinent to a given predictive task and the sensitivity of the resultant model to a given feature. Given the breadth of data collected in the Habitual Diet and Avocado Trial (HAT) study, it is important to appreciate which data is necessary for making a robust prediction. By excluding one feature set at a time from the model training process, we seek to elucidate the relative importance of each feature set and the degree to which its exclusion would skew the model's predictions (e.g., would removal of a feature induce bias?). In this use case, we use error profiling to substantiate the data discovery process. The different combinations of features are the worlds ω represented in the rows of the error profile, and the visualization illustrates if different inclusion/exclusion of features g_j are appropriate for predicting the outcomes of target (sub)populations β_j.

4. *Use Case 4: Error profiling to evaluate sequential decision-making models in lung cancer screening.* This use case evaluates multiple modeling variants of a sequential decision-making approach to lung cancer screening using the National Lung Screening Trial (NLST) dataset. The key challenge ad-dressed is optimizing the trade-off between true and false positives in clinical recommendations based on low-dose computed tomography (LDCT). We compare nine models, including baseline clinical assessments (physician), modularized partially observable Markov decision procsses (POMDPs) (modPOMDP), and hybrid models that integrate classifiers and radiomic features. Each variant represents a distinct modeling world (ω), enabling comparison across different modeling strategies (temporal modularization, radiomic augmentation, and ensemble learning techniques) to evaluate their respective influence on prediction error. The visualization facilitates comparison of how often and in which

patient subgroups specific models error, and whether certain combinations of features and learning techniques systematically improve or de-grade performance. Here, the error profile is particularly valuable in revealing sets of patients (β_j) where models consistently misclassify, indicating shared limitations in the modeling assumptions or feature space. These insights help clinicians and model developers better understand where particular modeling decisions might fail, even when overall performance metrics appear favorable.

3.4. Evaluating the Error Profiles

Our visualization aims to answer several key questions from the different use cases: on which sample do ML models learn something different and/or overlap (Use Case 1); how do different imputation methods influence end model performance (Use Case 2); to what extent does the inclusion/exclusion of specific features change model performance (Use Case 3); and to what extent different modeling strategies affect model agreement and error across patient subgroups (Use Case 4)? To this end, we evaluate users' ability to employ the visualizations to answer these questions and their perception of usefulness. An anonymous survey was deployed via Qualtrics, resulting in the recruitment of (n=22) participants through professional networks with classification based on institutional affiliation and self-reported expertise. Respondents self-identified across three categories (computer scientists, medical informaticians, clinicians) without subspecialty stratification. Participants were given a brief description of the goal of the study and then presented with a series of error profiles for the different use cases (Figures 2-5). The participants were then asked 58 quantitative and qualitative questions (Table A8) to obtain: 1) objective measures of the ability of the users to successfully read and interpret the profiles; and 2) subjective measures of user feedback on the influence the visualization has on trust in the model, and overall clinical utility. Median completion time of the survey was 25 minutes. Free-text responses were analyzed through iterative thematic coding. Feedback from surveys was only obtained for the first three use cases, as the development of the fourth use case was done after the design and deployment of the survey.

4. Results on Real Data

Use Case 1: ECMO ML algorithm error profiling. By comparing the performance and error distributions of these models, we aimed to identify patterns in the (in)consistencies in decisions despite variations in model design. Table 1 reveals respectable performance across all architectures, with RF demonstrating the best performance across all aggregate metrics. Figure 2 illustrates the error profiles across different architectures. The error profiles reveal that despite the high performance across all the models, they make different decisions. There are noticeable subpopulations that the tree-based models (i.e., RF, XGBoost, LGBoost) and non-tree models (i.e., LR, MLP, SVM) disagree upon, with the tree-based models being more successful at correctly classifying a particular subpopulation of patients – these subpopulations could have unique characteristics presenting contraindications for model usage. Remarkably, even though RF boasts the best aggregate performance with an AUROC of 0.931 (0.008), there are identifiable subpopulations where another model may be more suitable. Similarly, while MLP does not yield the worst performance metrics, it

Table 1: Models trained on the ECMO dataset. Performance on the test set is measured by AUROC, PRAUC, F1-score, precision, and recall. Standard deviations are obtained by applying models trained using 5-fold cross-validation on the test set.

Model architecture	**AUROC (std.)**	**PRAUC (std.)**	**F1-Score (std.)**	**Recall (std.)**	**Precision (std.)**
LGB	0.893 (0.016)	0.864 (0.024)	0.809 (0.022)	0.792 (0.023)	0.827 (0.028)
LR	0.854 (0.016)	0.830 (0.020)	0.695 (0.154)	0.660 (0.213)	0.785 (0.030)
MLP	0.824 (0.041)	0.804 (0.042)	0.714 (0.051)	0.668 (0.088)	0.774 (0.020)
RF	0.931 (0.008)	0.918 (0.012)	0.848 (0.031)	0.823 (0.047)	0.875 (0.018)
SVM	0.818 (0.045)	0.798 (0.047)	0.718 (0.072)	0.709 (0.100)	0.730 (0.050)
XGBoost	0.877 (0.033)	0.856 (0.032)	0.766 (0.064)	0.743 (0.088)	0.794 (0.046)

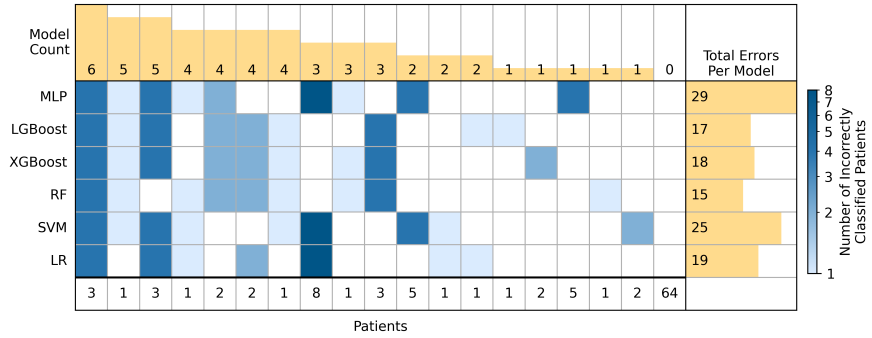

Figure 2: Error profile comparing the models described in Table 1.

noticeably struggles on a particular group of patients, highlighting a potential subpopulation where its use may be inappropriate. Different scopes of comparison are also possible, such as across more models (Figure A3), different feature selection methods (Figure A4), within one model architecture (Figure A5), and across subgroups (Figures A6-A8).

Use Case 2: MFP imputation error analysis. After analyzing the performance and errors after applying different missing data mechanisms, we identified patterns in error and populations in which models were effective and ineffective. Though metrics are relatively consistent across all missing data methods (Table 2), the error profiles reveal subpopulations that are challenging to classify when utilizing missing data mechanisms (Figure 3). For example, censoring or KNN imputation resulted in many incorrect classifications. However, other subpopulations were only correctly classified by models utilizing the same respective missingness techniques. This observation highlights that though performance metrics are similar, the models are still different and potentially optimized for different populations.

Use Case 3: HAT feature set error analysis. Assessing the model's performance with the omission of each feature set pinpoints which sets of feature data were most critical to achieving high predictive accuracy. For instance, the removal of data pertaining to labs markedly impacted the model's accuracy, highlighting the substantial predictive value

Table 2: Models trained on the MFP dataset. Performance on the test set is measured by AUROC, PRAUC, F1-score, precision, and recall. Standard deviations are obtained by applying models trained using 5-fold cross-validation on the test set.

Missing data method	**AUROC (std.)**	**PRAUC (std.)**	**F1-Score (std.)**	**Recall (std.)**	**Precision (std.)**
Censoring	0.713 (0.006)	0.771 (0.007)	0.720 (0.007)	0.805 (0.043)	0.653 (0.026)
EM Impute	0.729 (0.011)	0.786 (0.013)	0.723 (0.014)	0.797 (0.075)	0.667 (0.036)
Indicators	0.726 (0.009)	0.782 (0.011)	0.726 (0.009)	0.808 (0.075)	0.664 (0.040)
KNN Impute	0.725 (0.007)	0.782 (0.009)	0.728 (0.009)	0.815 (0.062)	0.660 (0.028)
MICE	0.733 (0.010)	0.791 (0.010)	0.716 (0.008)	0.757 (0.067)	0.685 (0.036)
Simple Impute	0.726 (0.009)	0.782 (0.011)	0.726 (0.008)	0.809 (0.074)	0.665 (0.040)

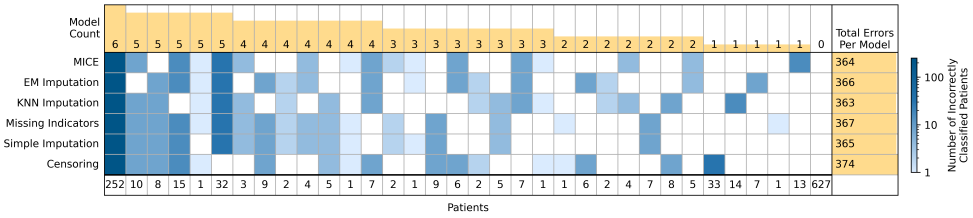

Figure 3: Error profile comparing the models described in Table 2.

contained within these feature sets (Table 3). Conversely, the exclusion of other data types resulted in less pronounced decreases in predictive performance, suggesting a diminished reliance on these variables. When further interpreting the instance of removal of labs using the error profile (Figure 4), we find that a large set of patients are exclusively misclassified by this one model (therefore suggesting the predictive value of labs for these particular patients); however, there is also a large patient set that was exclusively correctly classified by this one model (the other eight models were incorrect and were only correct when removing lab values).

Use Case 4: POMDP error analysis. The error profiling results (Figure 5) and aggregate metrics (Table 4) reveal critical distinctions across POMDP model variants. All models achieve similarly low false negative rates (FN $\leq$ 5) but vary substantially in false positives (FP) and true positives (TP), emphasizing the importance of balancing early detection with the risk of overdiagnosis and overtreatment. Notably, the modPOMDP radiomics XGBoost model achieves the highest TP rate (75.2), outperforming both the physician benchmark (57.2) and the base modPOMDP (57.4), while keeping FN comparably low (0.8). However, this model also maintains a moderate FP count (1,776.6), which must be contextualized clinically. From a visualization point of view (Figure 5), our error profiling matrix reveals clusters of patients misclassified by multiple models – evident from darker vertical bands – indicating challenging subpopulations across modeling strategies. The modPOMDP radiomics B2B model, with the lowest FP count (557.4), appears to consistently avoid overdiagnosis in certain patient clusters, although this comes at the cost of lower TP (54.4). In

Table 3: Multi-outcome models trained on the HAT dataset. Performance on the test set is measured by AUROC, PRAUC, F1-score, precision, and recall. Metrics are reported for HDL prediction. Standard deviations are obtained by applying models trained using 5-fold cross-validation on the test set.

Feature category	**AUROC (std.)**	**PRAUC (std.)**	**F1-Score (std.)**	**Recall (std.)**	**Precision (std.)**
All categories	0.657 (0.013)	0.618 (0.021)	0.580 (0.009)	0.603 (0.015)	0.560 (0.020)
No acid	0.674 (0.012)	0.600 (0.028)	0.578 (0.009)	0.595 (0.014)	0.563 (0.016)
No HEI	0.661 (0.013)	0.609 (0.020)	0.585 (0.030)	0.616 (0.041)	0.558 (0.029)
No labs	0.530 (0.035)	0.477 (0.056)	0.497 (0.027)	0.551 (0.081)	0.457 (0.028)
No metabolome	0.662 (0.019)	0.593 (0.034)	0.576 (0.021)	0.584 (0.018)	0.569 (0.026)
No MR	0.667 (0.038)	0.619 (0.048)	0.568 (0.024)	0.578 (0.029)	0.559 (0.024)
No survey	0.664 (0.021)	0.610 (0.021)	0.589 (0.020)	0.597 (0.032)	0.582 (0.017)
No UAC	0.662 (0.020)	0.617 (0.039)	0.586 (0.023)	0.605 (0.029)	0.569 (0.025)
No vitals	0.674 (0.026)	0.628 (0.023)	0.580 (0.018)	0.595 (0.019)	0.566 (0.019)

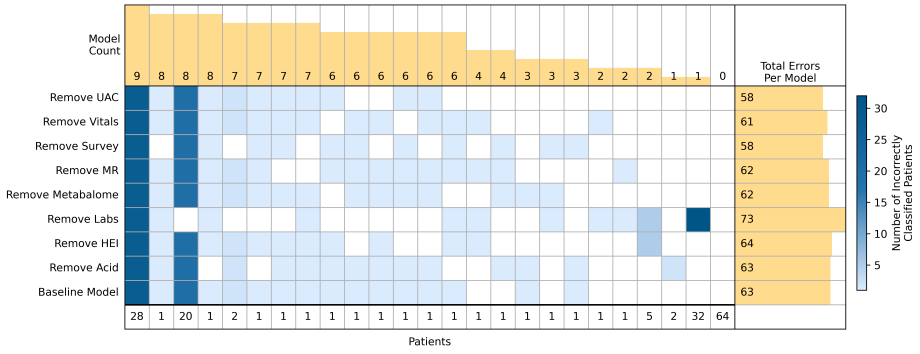

Figure 4: Error profile comparing the models described in Table 3.

contrast, the modPOMDP BRF model (TP: 54.8, FP: 1,463.6) shows less specificity, sharing error overlaps with both simpler and more complex models. Inclusion of radiomic features generally improves sensitivity, especially when paired with advanced learners like XGBoost, as seen in the more isolated error patterns compared to models without radiomics. This use case demonstrates that even with comparable AUROC or accuracy metrics, the visualization allows finer-grained insight into how models behave on the same patient subgroups. The profiles help identify where radiomics or ensemble methods add value and where they may introduce new risk, offering a lens for evaluating real-world model readiness.

General survey feedback. Overall, participants successfully interpreted the visualizations, achieving a median score of 75% across all use cases and user types, with insignificant (p=0.508) differences between the different types of users (Table A5). Participants responded positively to general questions about the visualization irrespective of a particular use case (median response of 3.65), with computer scientists (median response of 3.91,

Table 4: Models trained on the NLST dataset using the POMDP framework and its variants. Performance is evaluated using confusion matrix components: true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP). Standard deviations are obtained by applying models trained using 5-fold cross-validation on the test set.

Model	**TN (std.)**	**FP (std.)**	**FN (std.)**	**TP (std.)**
Physician	1009.6 (3.2)	1926.4 (1.7)	1.4 (0.9)	57.2 (1.3)
modPOMDP	960.0 (1.0)	1976.0 (1.4)	0.6 (0.5)	57.4 (1.7)
modPOMDP MLP	1127.6 (2.2)	1808.4 (1.8)	2.4 (1.3)	56.2 (0.8)
modPOMDP BRF	1472.4 (1.3)	1463.6 (1.8)	3.2 (0.8)	54.8 (1.3)
modPOMDP B2B	2238.6 (1.3)	697.4 (0.9)	4.8 (0.4)	53.2 (1.1)
modPOMDP Radiomics	1146.0 (2.0)	1782.8 (2.0)	1.0 (1.0)	64.4 (2.2)
modPOMDP Radiomics BRF	1490.0 (2.9)	1445.2 (2.5)	2.8 (1.8)	56.2 (0.8)
modPOMDP Radiomics B2B	2378.6 (2.7)	557.4 (3.8)	3.6 (0.9)	54.4 (0.9)
modPOMDP Radiomics XGBoost	1141.4 (2.1)	1776.6 (1.8)	0.8 (1.1)	75.2 (1.3)

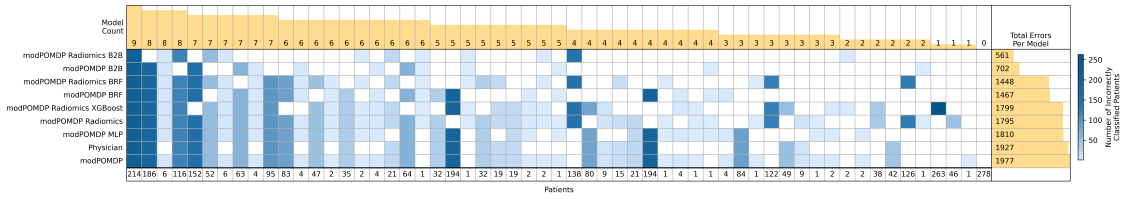

Figure 5: Error profile comparing the models described in Table 4.

p<0.001) experiencing the greatest ease and benefit (Table A6). There was mostly equivalent performance and subjective impressions across the use cases (Table A7); however, users found it more difficult (p=0.002) to interpret the profiles from the second use case demonstrating different imputation methods than the third use case depicting impact of feature omission. Comments highlighted a strong sentiment that the error profiles provide a deeper understanding when comparing multiple models, with representative quotes in Table A4. Notably, participants mentioned that these visualizations made it easy to understand which populations the models (dis)agree on, with color-coding effectively identifying which models struggle with specific patient groups. Details such as totals for each model and patient group were praised for their ability to summarize complex data.

5. Discussion

As stated by the statistician, George Box, "*All models are wrong, but some are useful*" Box (1976). Model errors are expected, and understanding the nature of the errors is critical to inform viable uses of a model. Our visualization method provides an interpretable way to compare model error distributions by highlighting the patterns within the correctly and incorrectly classified subpopulations from different models. The use of the error profile to

identify different subpopulations that are misclassified by a single model, unique sets of models, or all models is a key component of its utility and offers critical insights that would otherwise require assumptions. Without error analysis through a method such as error profiling, the identification of these subpopulations would not be possible. With knowledge of these subpopulations, one can interpret and characterize differences through additional analyses ranging from statistical comparisons to manual clinical review, ultimately providing more informed considerations for model usage.

Our visualization can lead to further conclusions that are specific to each use case. For example, in the ECMO use case, the visualization helped identify which model architectures made different decisions for certain patients, informing model selection and practical deployment considerations in clinical settings. In the MFP use case, it enabled us to identify the most effective imputation techniques for accurately predicting patient outcomes. It highlighted the importance of employing diverse data imputation techniques to ensure the robustness of models Wang et al. (2022). For the HAT use case, our visualization helped understand potential vulnerabilities due to our model's reliance on specific data types and identified critical feature sets necessary for accurate predictions. Such findings could enhance the robustness of our predictions through future mitigation of vulnerabilities, as well as refinement of a parsimonious model that prioritizes essential data while maintaining high levels of accuracy to balance model complexity and interpretability Frasca et al. (2024). For POMDP, the visualization helped disentangle the nuanced tradeoffs between sensitivity and specificity across sequential decision-making models for lung cancer screening. Despite similar aggregate performance metrics, the error profiles revealed that certain models — especially those augmented with radiomic features — consistently performed better on specific patient subgroups while introducing new error patterns in others. This enabled a more granular understanding of when particular modeling strategies (e.g., ensemble learning, temporal modularization, radiomic integration) improve outcomes versus when they may lead to overdiagnosis or reduced generalizability. In doing so, the visualization supported a more clinically grounded interpretation of model behavior, guiding both model selection and risk-aware deployment decisions. These findings accentuate the value of error profiling as a complementary evaluation tool that highlights when a model is useful — and when it may be wrong in ways that matter.

Visualization analytics tools are invaluable for making ML models more trustworthy and comprehensible Chatzimparmpas et al. (2020); Sacha et al. (2019); Alicioglu and Sun (2022); Yuan et al. (2021); Liu et al. (2017). Error profiling is merely one component that can embed into multiple stages of the larger framework of AI evaluation for clinical translation, to ensure that models are resilient to data inaccuracies and used when appropriate:

1. *Reassessment of data and objectives.* Error profiling helps to understand limitations in the current data and inform the acquisition of additional data that is better aligned with the modeling objectives.

2. *Exploration of advanced modeling techniques.* Error profiling can guide the selection and tuning of models by highlighting areas where either simpler or more complex techniques fall short, to help balance model complexity, predictive performance, and interpretability.

3. *Iterative model development and validation.* Error profiling should be a continuous component of validation and refinement, providing feedback on model performance to ensure that each iteration brings us closer to a clinically effective model.

4. *Engagement with clinical experts.* Error profiling can inform experts about appropriate use cases of a model, facilitating co-design and ensuring that models are technically sound and clinically useful.

5. *Transparency and documentation.* Error profiles and their impact on model adjustments or usage should be included in the documentation of the modeling processes as a record of challenges encountered and consequential decisions, to improve the future development of models in healthcare.

Limitations Alternative deep learning methods might better capture the underlying patterns and subtle relationships within complex datasets Bhatt et al. (2021). These "black box" approaches, while potentially more powerful, were set aside for this analysis in favor of simpler, more interpretable models von Eschenbach (2021). The motivation was to prioritize explainability within the clinical context – a critical consideration that ensures any deployed model can be understood, trusted, and effectively used by healthcare professionals Reboussin et al. (2021). However, our visualization method is model-agnostic and could still be used to compare the output behavior of deep learning models, even if their internal decision processes remain opaque. For example, it could be applied to compare per-patient predictions from a convolutional neural network and a Transformer model trained on the same dataset, highlighting which subpopulations they misclassify differently.

A key requirement of our approach is access to individual-level prediction outputs from all models under comparison. This may limit applicability in settings where only summary statistics (e.g., AUROC, accuracy) are reported, such as published benchmarks or regulatory documentation. For instance, many FDA submissions or multi-site clinical studies report only aggregate performance metrics, making fine-grained error profiling infeasible. However, in light of the FDA's predetermined change control plans and continuous learning paradigms for AI/ML-based SaMD, our error profiling framework could support: 1) premarket validation by identifying subpopulations requiring enhanced clinical validation; 2) post-market surveillance of model drift through longitudinal error pattern monitoring; and 3) algorithmic impact assessment revealing systematic biases across protected subgroups.

Although the utility of the visualization was qualitatively assessed through expert interpretation tasks, we recognize that its primary role is as a diagnostic rather than prescriptive tool. A broader range of empirical validation across different institutions and user roles would strengthen claims about its influence on trust or decision-making. It is also important to note that our method is fundamentally *descriptive* in nature. It reveals *where* models may be under-performing and highlights subpopulations where predictions diverge or consistently fail, but it does not explain *why* those failures occur or suggest how to fix them. In this way, it serves more like a radiograph than a treatment — it flags problems but requires additional analyses to diagnose and remediate underlying issues. However, we underscore the value of identifying which patient groups exhibit differential model performance as an important prerequisite to characterization and then actionable guidance. The method does not replace the need for causal inference or fairness audits, but complements

these techniques by revealing error patterns that may warrant deeper investigation. We provide a preliminary example of a follow-up analysis to provide additional characterization of subgroups in Figure A9, and will continue to pursue in future work a systematic approach to deriving actionable insights from these profiles. While our method is prevalence-agnostic and could be valuable for rare conditions where traditional aggregate metrics may obscure critical failure modes, there is still dependence on sufficient sample sizes to ensure interpretability of visual patterns and rigor of statistical analyses.

The small sample size of the user study limits statistical power and the findings present preliminary insights rather than definitive conclusions. In future work, we are committed to conducting a more rigorous evaluation of our framework. This future study would use validated survey instruments with neutral phrasing, recruit a larger and more diverse sample with clear participant characterization, include all four use cases, implement structured protocols for collecting and analyzing qualitative feedback (such as think-aloud sessions or semi-structured interviews), and potentially employ mixed methods to capture both quantitative metrics and rich qualitative insights about the framework's utility in practice.

Finally, the current implementation of our approach is *ad hoc* and disconnected from standardized ML development pipelines. Integrating the error profiling into established MLOps platforms could include presenting the profiles with model versioning systems to track error pattern evolution, and real-time error profiling alongside traditional performance monitoring. Real-time clinical deployment will require EHR integration with HL7 FHIR-compliant pipelines. Such integration could enhance usability, improve reproducibility, and enable continuous monitoring throughout the model lifecycle. This would support its adoption not only as a standalone interpretability tool, but as a core component of systematic model auditing in healthcare.

6. Conclusions

Although in its early stages, error profiling of data-driven models offers promising directions for systematic comparisons. Insights from our approach underscore the challenges of predictive modeling with complex health datasets and provide a roadmap for overcoming these obstacles. By examining different algorithms, imputation approaches, and feature set contributions, we can inform decisions about model architectures, feature engineering, and data prioritization, and when they are appropriate to use in clinical practice. This method advances our objective of leveraging clinical datasets for predictive insights and contributes to the broader field of predictive healthcare analytics, where model trustworthiness significantly impacts patient care and outcomes.

References

Gulsum Alicioglu and Bo Sun. A survey of visual analytics for explainable artificial intelligence methods. *Computers Graphics*, 102:502–520, 2022. doi: 10.1016/j.cag.2021.09.002.

Brian Ayers, Katherine Wood, Igor Gosev, and Sunil Prasad. Predicting survival after extracorporeal membrane oxygenation by using machine learning. *The Annals of Thoracic Surgery*, 110(4):1193–1200, 2020. doi: 10.1016/j.athoracsur.2020.03.128.

Christian Beck, Arnulf Jentzen, and Benno Kuckuck. Full error analysis for the training of deep neural networks. *Infinite Dimensional Analysis, Quantum Probability and Related Topics*, 25(02):2150020, 2022. doi: 10.1142/S021902572150020X.

Chandradeep Bhatt, Indrajeet Kumar, V. Vijayakumar, Kamred Udham Singh, and Abhishek Kumar. The state of the art of deep learning models in medical science and their challenges. *Multimedia Systems*, 27(4):599–613, 2021. doi: 10.1007/s00530-020-00694-1.

George E. P. Box. Science and Statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976. doi: 10.1080/01621459.1976.10480949.

Federico Cabitza, Raffaele Rasoini, and Gian. F. Gensini. Unintended consequences of machine learning in medicine. *JAMA*, 318(6):517–518, 2017. doi: 10.1001/jama.2017.7797.

Angelos Chatzimparmpas, Rafael M. Martins, Ilir Jusufi, Kostiantyn Kucher, Fabrice Rossi, and Andreas Kerren. The state of the art in enhancing trust in machine learning models with the use of visualizations. *Computer Graphics Forum*, 39(3):713–756, 2020. doi: 10.1111/cgf.14034.

Irene Y. Chen, Shalmali Joshi, and Marzyeh Ghassemi. Treating health disparities with artificial intelligence. *Nat Med*, 26(1):16–17, 2020. doi: 10.1038/s41591-019-0649-2.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. KDD '16, pages 785–794. Association for Computing Machinery, 2016. doi: 10.1145/2939672.2939785.

Subhajit Das and Alex Endert. Legion: Visually compare modeling techniques for regression. In *2020 Visualization in Data Science (VDS)*, pages 12–21, 2020. doi: 10.1109/VDS51726.2020.00006.

Anthony W. F. Edwards. Seven-set venn diagrams with rotational and polar symmetry. *Combinatorics, Probability and Computing*, 7(2):149–152, 1998. doi: 10.1017/S0963548397003143.

Peter Flach. Performance evaluation in machine learning: The good, the bad, the ugly, and the way forward. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9808–9814, 2019. doi: 10.1609/aaai.v33i01.33019808.

Maria Frasca, Davide La Torre, Gabriella Pravettoni, and Ilaria Cutica. Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review. *Discover Artificial Intelligence*, 4(1):15, 2024. doi: 10.1007/s44163-024-00114-7.

Jerrold Griggs, Charles E. Killian, and Carla D. Savage. Venn diagrams and symmetric chain decompositions in the boolean lattice. *The Electronic Journal of Combinatorics*, pages R2–R2, 2004. doi: 10.37236/1755.

Branko Grünbaum. Venn diagrams and independent families of sets. *Mathematics Magazine*, 48(1):12–23, 1975. doi: 10.2307/2689288.

Branko Grünbaum. Venn diagrams i. *Geombinatorics*, 1(4):5–12, 1992a.

Branko Grünbaum. Venn diagrams ii. *Geombinatorics*, 2(2):25–32, 1992b.

Holger A. Haenssle, Christine Fink, R. Schneiderbauer, Ferdinand Toberer, Timo Buhl, A. Blum, A. Kalloo, A. B. H. Hassen, Luc Thomas, A. Enk, and Lorenz Uhlmann. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol*, 29(8):1836–1842, 2018. doi: 10.1093/annonc/mdy166.

Peter Hamburger. Doodles and doilies, non-simple symmetric venn diagrams. *Discrete Mathematics*, 257(2):423–439, 2002. doi: 10.1016/S0012-365X(02)00441-7.

Peter Hamburger and Raymond E. Pippert. Venn said it couldn't be done. *Mathematics Magazine*, 73(2):105–110, 2000. doi: 10.2307/2691081.

Chen He, Vishnu Raj, Hans Moen, Tommi Gröhn, Chen Wang, Laura-Maria Peltonen, Saila Koivusalo, Pekka Marttinen, and Giulio Jacucci. Vms: Interactive visualization to support the sensemaking and selection of predictive models. In *IUI '24: 29th International Conference on Intelligent User Interfaces*, pages 229–244. ACM. doi: 10.1145/3640543.3645151.

David W. Henderson. Venn diagrams for more than four classes. *The American Mathematical Monthly*, 70(4):424–426, 1963. doi: 10.2307/2311865.

Elizabeth Hutchins, Al Rahrooh, Jeffrey Feng, Neha Chandra, Jeffrey J Hsu, and Alex Bui. Abstract 13293: Predicting Successful ECMO Decannulation - A Novel Machine Learning Approach. *Circulation*, 148(Suppl_1):A13293–A13293, 2023. doi: 10.1161/circ.148.suppl_1.13293.

Microsoft Ignite. Assess errors in machine learning models - azure machine learning. 2025. URL https://learn.microsoft.com/en-us/azure/machine-learning/concept-error-analysis?view=azureml-api-2.

IMDRF. *Characterization Considerations for Medical Device Software and SoftwareSpecific Risk*. IMDRF Software as a Medical Device (SaMD) Working Group, 2025.

Anil Jadhav, Dhanya Pramod, and Krishnan Ramanathan. Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 33(10):913–933, 2019. doi: 10.1080/08839514.2019.1637138.

Luke Oluwaseye Joel, Wesley Doorsamy, and Babu Sena Paul. A comparative study of imputation techniques for missing values in healthcare diagnostic datasets. *International Journal of Data Science and Analytics*, 2025. doi: 10.1007/s41060-025-00825-9.

Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17:195, 2019. doi: 10.1186/s12916-019-1426-2.

Bum Chul Kwon, Min-Je Choi, Joanne Taery Kim, Edward Choi, Young Bin Kim, Soonwook Kwon, Jimeng Sun, and Jaegul Choo. Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):299–309, 2019. doi: 10.1109/TVCG.2018.2865027.

Jacob Levman, Bryan Ewenson, Joe Apaloo, Derek Berger, and Pascal N. Tyrrell. Error consistency for machine learning evaluation and validation with application to biomedical diagnostics. *Diagnostics*, 13(7):1315, 2023. doi: 10.3390/diagnostics13071315.

JiaHang Li, ShuXia Guo, RuLin Ma, Jia He, XiangHui Zhang, DongSheng Rui, YuSong Ding, Yu Li, LeYao Jian, Jing Cheng, and Heng Guo. Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. *BMC Medical Research Methodology*, 24(1):41, 2024. doi: 10.1186/s12874-024-02173-x.

Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM Computing Surveys*, 50(6): 94:1–94:45, 2017. doi: 10.1145/3136625.

Yiran Li, Takanori Fujiwara, Yong K. Choi, Katherine K. Kim, and Kwan-Liu Ma. A visual analytics system for multi-model comparison on clinical data predictions. *Visual Informatics*, 4(2):122–131, 2020. doi: 10.1016/j.visinf.2020.04.005.

Shixia Liu, Xiting Wang, Mengchen Liu, and Jun Zhu. Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*, 1(1):48–56, 2017. doi: 10.1016/j.visinf.2017.01.006.

Shixia Liu, Jiannan Xiao, Junlin Liu, Xiting Wang, Jing Wu, and Jun Zhu. Visual diagnosis of tree boosting methods. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):163–173, 2018. doi: 10.1109/TVCG.2017.2744378.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. 2017. doi: 10.48550/arXiv:1705.07874.

Khalegh Mamakani and Frank Ruskey. A new rose : The first simple symmetric 11-venn diagram. 2012. doi: 10.48550/arXiv.1207.6452.

Khalegh Mamakani, Wendy Myrvold, and Frank Ruskey. Generating simple convex venn diagrams. *Journal of Discrete Algorithms*, 16:270–286, 2012. doi: 10.1016/j.jda.2012.04.013.

Tammy McCausland. The bad data problem. *Research-Technology Management*, 64(1): 68–71, 2020. doi: 10.1080/08956308.2021.1844540.

Yao Ming, Huamin Qu, and Enrico Bertini. Rulematrix: Visualizing and understanding classifiers with rules — ieee journals magazine — ieee xplore. *IEEE Transactions on Visualization and Computer Graphics*, 25, 2019. doi: 10.1109/TVCG.2018.2864812.

Farideh Mohtasham, MohamadAmin Pourhoseingholi, Seyed Saeed Hashemi Nazari, Kaveh Kavousi, and Mohammad Reza Zali. Comparative analysis of feature selection techniques for COVID-19 dataset. *Scientific Reports*, 14(1):18627, 2024. doi: 10.1038/s41598-024-69209-6.

Sugeerth Murugesan, Sana Malik, Fan Du, Eunyee Koh, and Tuan Manh Lai. Deepcompare: Visual and interactive comparison of deep learning model performance. *IEEE Computer Graphics and Applications*, 39(5):47–59, 2019. doi: 10.1109/MCG.2019.2919033.

Andrew Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. ICML '04, page 78. Association for Computing Machinery. doi: 10.1145/1015330.1015435.

Steven W. J. Nijman, Artuur M. Leeuwenberg, Inés Beekers, Inge Verkouter, John J. L. Jacobs, M. L. Bots, Folkert W. Asselbergs, K. G. M. Moons, and Thomas P. A. Debray. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *Journal of Clinical Epidemiology*, 142:218–229, 2022. doi: 10.1016/j.jclinepi.2021.11.023.

Zeinab Noroozi, Azam Orooji, and Leila Erfannia. Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. *Scientific Reports*, 13(1):22588, 2023. doi: 10.1038/s41598-023-49962-w.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453, 2019. doi: 10.1126/science.aax2342.

Vimla L. Patel, Edward H. Shortliffe, Mario Stefanelli, Peter Szolovits, Michael R. Berthold, Riccardo Bellazzi, and Ameen Abu-Hanna. The coming of age of artificial intelligence in medicine. *Artificial intelligence in medicine*, 46(1):5–17, 2009. doi: 10.1016/j.artmed.2008.07.017.

Eike Petersen, Sune Holm, Melanie Ganz, and Aasa Feragen. The path toward equal performance in medical machine learning. *Patterns*, 4(7), 2023. ISSN 2666-3899. doi: 10.1016/j.patter.2023.100790.

Oona Rainio, Jarmo Teuho, and Riku Klén. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1):6086, March 2024. doi: 10.1038/s41598-024-56706-x. Publisher: Nature Publishing Group.

Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *N Engl J Med*, 380(14):1347–1358, 2019. doi: 10.1056/NEJMra1814259.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J. Topol. Ai in health and medicine. *Nature Medicine*, 28(1):31–38, 2022. doi: 10.1038/s41591-021-01614-0.

David M. Reboussin, Penny M. Kris-Etherton, Alice H. Lichtenstein, Zhaoping Li, Joan Sabate, Nirupa R. Matthan, Kristina Petersen, Sujatha Rajaram, Mara Vitolins, and Nikki Ford. The design and rationale of a multi-center randomized clinical trial comparing one avocado per day to usual diet: The habitual diet and avocado trial (hat). *Contemporary Clinical Trials*, 110:106565, 2021. doi: 10.1016/j.cct.2021.106565.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM. doi: 10.1145/2939672.2939778.

Frank Ruskey, Carla D. Savage, and Stan Wagon. The search for simple symmetric venn diagrams. *Notices of the American Mathematical Society*, 53(11):1304–1311, 2006.

Dominik Sacha, Matthias Kraus, Daniel A. Keim, and Min Chen. Vis4ml: An ontology for visual analytics assisted machine learning. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):385–395, 2019. doi: 10.1109/TVCG.2018.2864838.

Elham Tabassi. Artificial intelligence risk management framework (ai rmf 1.0). *NIST Trustworthy and Responsible AI, National Institute of Standards and Technology*, 2023. doi: 10.6028/NIST.AI.100-1.

John Venn. On the diagrammatic and mechanical representation of propositions and reasonings. *Philosophical Magazine*, 10(59):1–18, 1880. doi: 10.1080/14786448008626877.

John Venn. *Symbolic logic*. London : Macmillan, 1881.

Warren J. von Eschenbach. Transparency and the black box problem: Why we do not trust ai. *Philosophy Technology*, 34(4):1607–1622, 2021. doi: 10.1007/s13347-021-00477-0.

Huimin Wang, Jianxiang Tang, Mengyao Wu, Xiaoyu Wang, and Tao Zhang. Application of machine learning missing data imputation techniques in clinical decision making: taking the discharge assessment of patients with spontaneous supratentorial intracerebral hemorrhage as an example. *BMC Medical Informatics and Decision Making*, 22(1):13, 2022. doi: 10.1186/s12911-022-01752-6.

Junpeng Wang, Liang Gou, Wei Zhang, Hao Yang, and Han-Wei Shen. Deepvid: Deep visual interpretation and diagnosis for image classifiers via knowledge distillation. *IEEE Transactions on Visualization and Computer Graphics*, 25(6):2168–2180, 2019. doi: 10.1109/TVCG.2019.2903943.

WHO. *Ethics and governance of artificial intelligence for health: WHO guidance*. World Health Organization, 2021.

Xiwei Xuan, Xiaoyu Zhang, Oh-Hyun Kwon, and Kwan-Liu Ma. Vac-cnn: A visual analytics system for comparative studies of deep convolutional neural networks. 2022. doi: 10.48550/arXiv.2110.13252.

Jun Yuan, Changjian Chen, Weikai Yang, Mengchen Liu, Jiazhi Xia, and Shixia Liu. A survey of visual analytics techniques for machine learning. *Computational Visual Media*, 7(1):3–36, 2021. doi: 10.1007/s41095-020-0191-7.

Appendix A. Datasets

Extra corporeal membrane oxygenation (ECMO) has emerged as a crucial life support intervention, yet the decision-making process for safe decannulation remains challenging due to a paucity of data Hutchins et al. (2023). An ECMO dataset representing 213 patients seen between 2015-2021 who were successfully or unsuccessfully decannulated was collected at our institution. Successful decannulation was defined as survival without relapse to mechanical circulatory support or heart transplant within 30 days. Demographic, hemodynamic, laboratory, and echocardiographic data obtained within 24 hours of decannulation were collected, totaling 74 unique features.

MyFitnessPal (MFP) is an online application used worldwide by many individuals to record daily food intake and exercise-related activities, typically for achieving weight-related health objectives. A dataset of daily logs of nutritional intake, activity, and app usage was accrued from 549,381 users between January 1, 2021 to May 1, 2021 (120 days), for a total of 65,925,840 unique rows of data. The dataset was reduced to 547,010 patients from 12,063,756 total non-empty rows of data. Daily tracking consisted of nutritional, activity, and app usage features. 12 nutritional features were tracked on the aggregate nutritional content consumed across all logged foods that day - calories, protein, fat, carbohydrates, sugar, trans fats, polyunsaturated fats, fiber, potassium, sodium, iron, and calcium. Activity features included: weight, number of foods logged, number of water logs, number of weight logs, number of meals scanned, number of food barcodes scanned, number of exercise entries, and number of steps logged. App usage features included the number of emails opened and the number of app logins. Information about the subject was also available - sex, age, country, ZIP code, height, BMI, weight change goal (weight loss, weight gain, or maintenance), starting weight at January 1, use of MFP's premium features, and start date if premium was used.

The Habitual Diet and Avocado Trial (HAT) is a randomized controlled trial (RCT) conducted on 1,008 patients between June 27, 2018, and March 4, 2020 at four clinics to investigate the impact of providing one avocado per day for 26 weeks, compared to a habitual diet with minimal avocado consumption, on visceral adiposity and associated cardiometabolic risks. The primary outcome measure involves the use of MRI to assess changes in visceral fat in individuals with an increased waist circumference (WC) Reboussin et al. (2021). The study collected myriad data including bio-samples (e.g., for labs and metabolic assays yielding additional parameters such as hepatic lipid content, plasma lipid profiles, blood pressure, and high sensitivity C-reactive protein), magnetic resonance imaging (MRI), and other self-reported health assessments (e.g., sleep). Follow-up assessments were completed in October 2020, with final MRI scans obtained from 936 participants. In total, 263 features from nine categories of features were available with the corresponding number of features: 1) demographics, six features; 2) vitals, three features; 3) laboratory test results, eight; 4) acids, 40; 5) survey responses, 151; 6) HEI, 15; 7) MRI, 6; 8) metabolome profile, 32; 9) uric acid (UAC), two.

The National Lung Screening Trial (NLST) is a multicenter randomized controlled trial of 53,454 participants that evaluated the impact of low-dose computed tomography (LDCT) versus chest X-ray in reducing lung cancer-specific mortality. The trial included high-risk individuals aged 55–74 with a ¿=30 pack-year smoking history and no recent history

Table A1: ECMO dataset population and subset of variables used in analysis, stratifying across outcomes. Data are reported as mean (standard deviation), unless otherwise indicated. Two-sample t-test or Chi-square test are utilized as appropriate to compare distributions across outcomes, with Bonferroni correction for multiple testing.

	Died/relapsed on ECMO (n=110)	Alive after ECMO (n=103)	p-value
Admission age (years)	53.3 (15.7)	56.1 (14.6)	1.000
Sex, n (%)			
Female	29 (26.4)	36 (35.0)	1.000
Male	81 (73.6)	67 (65.0)	
Non-cardiac PMH, n (%)			
Cirrhosis	1 (0.9)	1 (1.0)	1.000
CKD	12 (10.9)	13 (12.6)	
DM	27 (24.5)	25 (24.3)	
ESRD	4 (3.6)		
HLD	8 (7.3)	13 (12.6)	
HTN	8 (7.3)	5 (4.9)	
RV systolic function, n (%)			
Mild	41 (37.3)	51 (49.5)	0.033
Moderate	23 (20.9)	13 (12.6)	
Normal	18 (16.4)	30 (29.1)	
Severe	28 (25.5)	9 (8.7)	
LVEF, mean (SD)	25.6 (17.1)	38.5 (17.4)	<0.001
Days in hospital	26.0 (34.6)	36.8 (33.6)	0.793
ECMO flow rate (L/min)	3.5 (1.2)	2.2 (1.1)	<0.001
Heart rate (bpm)	91.1 (15.7)	91.6 (17.4)	1.000
SBP (mmHg)	93.0 (21.5)	116.9 (21.5)	<0.001
DBP (mmHg)	63.7 (13.2)	61.1 (11.4)	1.000
MAP (mmHg)	72.5 (14.2)	77.7 (10.5)	0.107
CO (L/min)	3.7 (1.5)	4.9 (1.7)	0.012
CI	2.6 (1.0)	2.7 (0.7)	1.000
mPA (mmHg)	25.0 (8.8)	21.0 (5.9)	0.103
Cr (mg/dL)	1.4 (1.0)	1.5 (0.9)	1.000
Lactate (mg/dL)	45.6 (56.2)	15.2 (18.5)	<0.001
MvO2 (mL/min)	55.9 (33.4)	85.0 (14.5)	<0.001

of lung cancer. Participants were screened annually for up to three rounds and followed longitudinally. For this study, we focused on a subset of 2,994 patients from the NLST dataset for whom radiomic features were available, enabling enhanced evaluation of decision-making models that incorporate longitudinal and imaging-derived data.

Table A2: MyFitnessPal dataset population and subset of variables used in analysis, stratifying across outcomes. Study population is restricted to those included in the predictive modeling analysis (the subset of patients who reported both initial weight and weight within 10 days from the end of the 120 day horizon and indicated goal of weight loss). Data are reported as mean (standard deviation), unless otherwise indicated. Two-sample t-test or Chi-square test are utilized as appropriate to compare distributions across outcomes, with Bonferroni correction for multiple testing.

	No 4% weight loss (n=9,375)	4% weight loss (n=5,551)	p-value
Sex, n (%)			
Male	1877 (37.7)	2314 (41.7)	0.006
Female	3100 (62.3)	3237 (58.3)	
US, n (%)			
No	2731 (54.9)	3095 (55.8)	1.000
Yes	2246 (45.1)	2456 (44.2)	
Premium user, n (%)			
No	4373 (87.9)	4712 (84.9)	0.002
Yes	604 (12.1)	839 (15.1)	
Percent weight change	-0.0 (3.6)	-9.3 (4.4)	<0.001
Age (years)	35.2 (13.2)	37.3 (13.6)	<0.001
Initial height (inches)	66.9 (3.8)	67.2 (3.8)	0.344
Initial BMI	28.6 (5.9)	27.8 (5.3)	<0.001
Initial weight (lb)	183.6 (42.8)	200.1 (44.0)	<0.001
Total number of entries	61.4 (27.4)	75.2 (22.3)	<0.001
Total range of entries (days)	81.1 (18.8)	86.4 (11.9)	<0.001
Number of days with food logs	31.1 (28.6)	55.4 (32.1)	<0.001
Average number of daily steps logged	6478.2 (4081.4)	6715.4 (4895.6)	1.000
Average daily net calories	1240.0 (430.0)	1281.0 (397.5)	<0.001
Standard deviation daily net calories	455.5 (182.3)	415.4 (139.2)	<0.001
Average daily protein	60.5 (28.2)	65.5 (28.2)	<0.001

Table A3: HAT dataset population and subset of variables used in analysis. Data are reported as mean (standard deviation), unless otherwise indicated.

Variable	Overall (n=850)
Intervention, n (%)	
No intervention	430 (50.6)
Intervention	420 (49.4)
Sex, n (%)	
Male	237 (27.9)
Female	613 (72.1)
Initial weight	93.5 (19.3)
SFA	46.7 (3.0)
MUFA	17.9 (2.1)
Total trans fat	0.6 (0.2)
Age	50.7 (13.9)
BMI	32.5 (5.3)
Diastolic blood pressure, mmHg	76.6 (10.5)
Systolic blood pressure, mmHg	122.9 (16.5)
Pulse	69.3 (10.2)
Waist circumference, inches	109.6 (13.2)
Insulin	17.7 (16.0)
hsCRP	6.1 (7.7)
Glucose	107.2 (29.2)
Cholesterol	188.2 (39.5)
Triglycerides	124.5 (83.7)
HDL	52.0 (12.8)
VLDL	24.9 (16.7)
LDL	112.1 (33.7)
Volume of visceral adipose tissue, litres	0.1 (0.1)
Hepatatic fat fraction	0.3 (0.1)
Proton density fat fraction	0.0 (0.0)
Calories, kcal	1855.7 (739.8)
HEI score	54.0 (14.8)
12-week change in waist circumference	-0.0 (3.6)
12-week change in weight	0.3 (2.4)
12-week change in insulin	1.2 (16.9)
12-week change in hsCRP	0.1 (6.1)
12-week change in glucose	0.2 (19.6)
12-week change in cholesterol	-3.4 (25.3)
12-week change in triglycerides	3.4 (64.8)
12-week change in HDL	0.1 (7.2)
12-week change in VLDL	0.7 (13.0)
12-week change in LDL	-3.7 (22.8)
12-week change in Calories	109.7 (868.1)
12-week change in HEI	4.6 (17.6)

Appendix B. Modeling Details

Modeling details for all datasets. We apply a similar general experimentation pipeline to all datasets. For each model, we utilize a nested cross-validation design with the following steps: 1) split the dataset into 80/20 stratified splits, and keep the 20% as a hold-out test set, 2) the remaining 80% is divided into five stratified folds, 3) perform a grid-search over the hyperparameters and train (including the fitting of standardization and imputation parameters) on 4/5 folds and validate on the remaining fold, 4) Test the optimal parameters from (3) on the hold-out test set, 5) Report classification performance measures such as AUROC, PRAUC, F1-score, precision, and recall, 6) Repeat (2)-(5) for the five folds.

Modeling details for ECMO. The objective with the ECMO datasets is to predict successful decannulation. Summary statistics of key variables from the 213 patients in the dataset are shown in Table A1. All features are standardized, and any missing features are imputed via mean imputation. We assess different model architectures for the same classification task, each with their own grid of hyperparameters: XGBoost, LGBoost, LR, RF, SVM, and MLP. For each of the six architectures, we run four variations of experiments pertaining to different input features: 1) baseline performance using all possible features; 2) performance using all features except for three clinical features (MVO2, PASP, PAD) that are known to be useful, but difficult to measure in practice; 3) parsimonious model when limiting the model to the top ten most important features identified using SHAP feature importance values; and 4) parsimonious model when limiting the model to the top ten features after removing the three features in (2). We visualize error profiles of test results of the 24 models, and demonstrate different possibilities for scopes of comparison. For example, we can identify all unique sets of patients that were incorrectly classified by a corresponding unique sub-set of all models, but we can also subset the profiles to features variations of one architecture.

Modeling details for MFP. The goal with the MFP data is to predict whether users achieve a 4% decrease in weight over four months. We use aggregated data (mean, standard deviation, minimum, maximum, difference) from the first 90 days to predict a 4% decrease in weight after four months. Due to sparsity in the data, we limited analysis to 10,528 subjects who had identifiable outcomes, meaning they recorded their weight at four months. Summary statistics of this population are illustrated in Table A2. With the MFP data, we considered the different model architectures (same as the ECMO models) as hyperparameters themselves. Instead of running different experiments for different model architectures, we run different experiments to assess different mechanisms of handling missing data. As such, we select a hold-out test set that consists of complete samples of data, while the train set consists of a mixture of incomplete and complete samples of data. Mechanisms for handling missing data include: 1) censoring which only trains on the subset of complete data from the train set; 2) simple (mean) imputation; 3) simple imputation with missing indica-tors; 4) k-nearest neighbors imputation; 5) Regression imputation; 6) Expectation-maximization imputation; 7) iterative imputation (single iteration); 8) iterative imputation (multiple imputation), and 9) Multiple imputation with chained equations. We generate error profiles from the optimal models after applying the nine different techniques for handling missing data.

Modeling details for HAT. We use the HAT dataset for secondary analyses to assess if a multimodal ML model can predict various outcomes. Our main outcome measures are change in weight, HEI, triglycerides, high-density lipoprotein (HDL), and low-density lipoprotein (LDL) over 26 weeks; all outcome measures are binarized to gain=1/lose=0 over 26 weeks. We limit analysis to 850 patients with at least one outcome recorded at the Week 12, and Week 26 timepoints, summarized in Table A3. For variables available at Week 12, we engineer a feature for the change in value between Weeks 0 and 12. All features are standardized, and any missing features are imputed via mean imputation. The classification method that we use is XGBoost with a multi-label vector output Chen and Guestrin (2016). Given high-dimensionality of the feature space, we include hyperparameters for L1 and L2 regularization of the XGBoost model, with coefficients ranging from one to 25 in increments of five, as well as a-priori LASSO feature selection with a coefficient ranging from zero to 0.001 in increments of 0.0002 Ng. First, we run the entire training pipeline using all possible input features. Then, for each of the eight non-demographic feature categories, we train a separate multi-outcome XGBoost model while excluding the features associated with the respective category. Demographic features are present across all dataset variations, including: intervention arm, age, sex, weight, waist circumference, body mass index (BMI). We then perform error analysis on the test results of the eight models that each do not train on one of the feature categories, as well as the model that has access to all features. In this analysis, we identify all unique sets of patients that were incorrectly classified by a corresponding unique subset of the nine models.

Modeling details for POMDP. We applied a series of sequential decision-making models based on the partially observable Markov decision process (POMDP) framework to this cohort. The POMDP structure allows modeling of uncertainty over time and is well-suited for the iterative nature of cancer screening. We developed and evaluated multiple variants of the modularized POMDP (modPOMDP) framework, each representing different combinations of features and prediction mechanisms. This included:

- **Base modPOMDP**, which optimized positive predictions independently across each screening time point;
- **modPOMDP2**, which integrates classic machine learning classifiers post hoc to filter out likely false positives from modPOMDP's output;
- **modPOMDP + Radiomics**, which incorporates imaging-derived radiomic features to enhance prediction;
- **modPOMDP + Radiomics + Ensemble Classifiers**, which includes variations such as MLP, balanced random forest (BRF), B2B (Brock2B), and XGBoost.

In each chained model (e.g., modPOMDP2), the modPOMDP identifies candidate positive cases, which are subsequently filtered using the secondary classifier trained on baseline or radiomic features. Final predictions are formed by accepting only the modPOMDP positives that also receive a positive label from the classifier, effectively reducing false positives while preserving true positives. The experiment was conducted using a nested stratified cross-validation framework consistent with the modeling pipelines used in other use cases.

This involved an 80/20 split for training and testing, followed by five-fold inner validation with grid search hyperparameter tuning. Classifier performance was evaluated based on metrics such as AUROC, precision, recall, and F1-score. The modeling process included ensemble-based classifiers that were previously selected through a classifier selection study on the full NLST dataset (N=5,089) without radiomic features. This study guided the selection of MLP and BRF due to their favorable tradeoffs between true positives and false positives across three screening rounds. The Brock2b (B2B) model, a well-established logistic regression model using radiologic features from the time of screening, was also included in the ensemble comparison. The final model evaluation focused exclusively on the 2,994-patient radiomic subset. Here, error profiling was performed to assess model disagreement and to determine which combinations of features and modeling choices led to better generalization or reduced error overlap across patient subgroups. These analyses are described in the main text (Use Case 4) and visualized in Figure 5. Performance results for each model variant are detailed in Table 4.

Appendix C. Additional Survey Results

Use Case 1: ECMO ML algorithm error profiling. Users were able to correctly interpret the error profiles from this use case, with a median score of 80% across all participants (Table A5). There was an insignificant (p=0.270) difference in performance between the different types of users and they were able to effectively use the visualization to analyze the differences and similarities between model architectures. Impressions were positive (median response of 3.78), with computer scientists (median response of 3.89, p<0.001) finding the visualization the easiest to understand, while physicians/clinicians (median response of 2.78, p=0.06) experienced more subjective difficulties interpreting the profiles (Table A6).

Use Case 2: MFP imputation error analysis. Users were successful at interpreting the error profiles, with a median score of 75% across all participants (Table A5). There was an insignificant (p=0.694) difference in performance between the different types of users, and they were able to effectively assess the impact of different missing data mechanisms on model's decisions. Impressions were positive (median response of 3.40), with computer scientists (median response of 3.60, p=0.005) again finding the visualization the easiest to understand, while both medical informaticians (median response of 3.05 p=0.65) and physicians/clinicians (median response of 2.80, p=0.06) experienced more subjective difficulties interpreting the profiles (Table A6).

Use Case 3: HAT feature set error analysis. Participants were able to use the error profiles to understand the impact of removing individual features groups on predictive performance, with a median score of 80% across all participants (Table A5), and an insignificant (p=0.77) difference in performance between the different types of users. Impressions were positive (median response of 3.68), with computer scientists (median response of 4.09, p<0.001) finding the visualization the easiest to understand, while physicians/clinicians (median response of 2.91, p=0.43) experienced more subjective difficulties interpreting the profiles (Table A6).

Table A4: Representative free-text survey responses.

What do you like most about this visualization?	What do you like least about this visualization?
Succinct way of showing every error made - highlights key discrepancies.	It's hard to group and cluster different models that seem to perform similarly.
The visualization is highly detailed and provides a comprehensive view of the error profiles across multiple models and scenarios.	The visualization can be somewhat overwhelming due to the dense amount of information presented.
The use of color-coding to represent incorrectly classified patients makes it easy to quickly identify which models are struggling with specific patient groups. Additionally, the inclusion of totals for each model and patient group helps to summarize the data effectively.	Hard to read multiple variables at once if they are not next to each other. Searching through the figures to find patterns at the same time means you are looking for a pattern of squares that can be hard to identify when there is so much going on in the visualization. It takes a bit of getting used to initially.
Easy to understand where models differ and exactly what populations they differ on as well as making it easy to see where models perform the same and what the actual difference between these models are. Generally a super useful and interesting visualization and definitely good for comparing lots of models at the same time. Once you know how to read it, it's genuinely useful.	It does not consider the natural grouping of the patients instead groups the patients by the model's prediction which might not be as meaningful.
The tabular format effectively simplifies the complexity of understanding error overlaps among various machine learning models. Additionally, the visualization's use of color-coding and clear labeling helps in quickly identifying trends and discrepancies.	The visualization might be a bit overwhelming at first glance, especially for users who are not familiar with the data or the models being analyzed.

Table A5: Aggregated scores of the multiple-choice survey questions that assess the ability to correctly interpret the error profile visualizations. Results are shown as 'mean (std.), median [IQR].' In addition to total scores, isolated scores are shown for each use case and participant type. The Kruskal-Wallis test was used to compare the multiple participant types. Abbreviations: Computer scientist, CS. Medical informatician, MI. Physician/clinician, PC.

	CS (n=9)	MI (n=8)	PC (n=5)	All (n=22)	p-value
Use case 1	0.60 (0.37)	0.80 (0.30)	0.52 (0.36)	0.66 (0.35)	
	0.80 [0.20,0.80]	1.00 [0.60,1.00]	0.40 [0.20,0.80]	0.80 [0.25,1.00]	0.270
Use case 2	0.64 (0.31)	0.75 (0.35)	0.60 (0.45)	0.67 (0.35)	
	0.50 [0.50,1.00]	0.88 [0.69,1.00]	0.75 [0.25,1.00]	0.75 [0.50,1.00]	0.694
Use case 3	0.60 (0.30)	0.70 (0.24)	0.60 (0.37)	0.64 (0.29)	
	0.80 [0.60,0.80]	0.80 [0.60,0.80]	0.60 [0.60,0.80]	0.80 [0.60,0.80]	0.768
Overall	0.61 (0.28)	0.75 (0.27)	0.57 (0.35)	0.65 (0.29)	
	0.71 [0.36,0.86]	0.82 [0.71,0.88]	0.50 [0.36,0.93]	0.75 [0.39,0.86]	0.508

Table A6: Aggregated results of the Likert-scale survey questions that assess the subjective opinions of each participant regarding the clarity and utility of the error profile visualizations. Results are shown as 'mean (std.), median [IQR].' In addition to overall responses, isolated results are shown for each use case and each participant type. The Kruskal-Wallis test was used to compare the multiple participant types (p-value column). The results from each participant type, as well as aggregated results from all participants, were compared against a neutral response (three in a five-point Likert scale) using a one-sample t-test, with * denoting significance at a level of p=0.05. Abbreviations: Computer scientist, CS. Medical informatician, MI. Physician/clinician, PC.

	CS (n=9)	MI (n=8)	PC (n=5)	All (n=22)	p-value
Use case 1	3.99 (0.28)	3.49 (0.76)	2.53 (0.40)	3.48 (0.76)	
	3.89 [3.89,4.00]*	3.67 [2.89,3.94]	2.77 [2.22,2.78]	3.78 [2.89,3.97]*	0.004
Use case 2	3.63 (0.49)	3.14 (0.81)	2.78 (0.19)	3.26 (0.66)	
	3.60 [3.50,3.80]*	3.05 [2.68,3.58]	2.80 [2.70,2.90]	3.40 [2.73,3.60]	0.039
Use case 3	4.12 (0.39)	3.52 (0.70)	2.82 (0.47)	3.61 (0.73)	
	4.09 [3.91,4.36]*	3.27 [2.98,4.05]	2.91 [2.73,3.00]	3.68 [3.00,4.09]*	0.006
General	3.91 (0.29)	3.51 (0.76)	2.66 (0.72)	3.48 (0.75)	
	3.90 [3.80,4.00]*	3.45 [3.03,4.08]	2.30 [2.20,3.10]	3.66 [3.10,4.00]*	0.019
Overall	3.92 (0.27)	3.42 (0.63)	2.71 (0.18)	3.46 (0.63)	
	3.83 [3.73,3.95]*	3.21 [2.93,3.95]	2.68 [2.60,2.88]*	3.70 [2.88,3.89]*	0.005

Table A7: Outcomes (p-values) of paired Wilcoxon sign-rank tests assessing for differences in responses for both objective multiple-choice questions and subjective Likert-scale questions across the different use cases. Isolated comparisons are performed for each participant type. Abbreviations: Computer scientist, CS. Medical informatician, MI. Physician/clinician, PC. Multiple-choice, MC. Likert-scale, LS.

Group	Comparison	CS (n=9)	MI (n=8)	PC (n=5)	All (n=22)
MC	Use 1 vs Use 2	0.020	0.033	0.188	0.001
	Use 1 vs Use 3	0.581	0.586	1.000	0.887
	Use 2 vs Use 3	0.020	0.022	0.438	0.001
LS	Use 1 vs Use 2	0.055	0.383	0.813	0.137
	Use 1 vs Use 3	0.293	0.641	0.438	0.135
	Use 2 vs Use 3	0.008	0.023	0.855	0.002

Appendix D. Formulation of the Error Profile

We can provide a formal description of the error profile using notation from propositional logic. In the following sections we: 1) outline basic definitions from propositional logic and describe their analogies to classification error; 2) formally describe the process of generating the error profile.

Definitions

The syntax of propositional logic builds around *propositional sentences*, which are used to express events. In the context of error profiling, the event of interest is the incorrect classification of a sample. As such, if Y_i is the ground truth for the ith sample and $\widehat{Y}_i$ is the prediction, then we can define incorrect classification as α_i using the following syntax. In the next sections, we focus on α_i as the base sentence that expresses the incorrect classification of a sample.

$$\neg((Y_i \wedge \widehat{Y}_i) \vee \neg(Y_i \vee \widehat{Y}_i)) \;\equiv\; \alpha_i$$

In further alignment with propositional logic, *worlds* assign values (e.g., true/false) to sentences. The following notation generally means that a sentence α is true at world ω. In the context of error profiling, a world can be construed as the decision space of a single ML model. Thus, we also consider ω to be a *model*. In other words, for $k = 1, 2... K$ worlds, the following can be interpreted as: "the kth ML model incorrectly classifies the ith sample."

$$\omega_k \;\models \alpha_i$$

Finally, the set of worlds that satisfy a sentence α_i are called the *models* of α_i. The appropriate term *models* of α_i can be interpreted as the set of ML models that incorrectly classified the ith sample.

$$Mods(\alpha_i) \;\equiv\; \{\omega :\; \omega_k \models \alpha_i\}$$

To summarize, a *sentence* α_i represents the incorrect classification of the *ith* sample. The incorrect classification of the *ith* sample holds true in a *world* ω_k, when a corresponding ML model indeed classifies the sample incorrectly. The set of worlds that incorrectly classify a given sample are $Mods(\alpha_i)$ and described as the *models* of α_i, which appropriately details the set of ML models that all classify a sample incorrectly.

Error profile generation

As part of the process of creating the error profile, we identify sets of patients that were incorrectly classified by the same ML models. Reiterating the definition in the Methods, we define a *group* G_j as a tuple consisting of: 1) a unique set β_j of n_j incorrectly classified patients; 2) a unique set g_j of $|g_j|$ worlds (i.e., a set of ML models).

$$G_j \equiv \ (\beta_j = \{\alpha_1, \alpha_2 \ ... \ \alpha_{n_j}\}, \ g_j) \ \ s.t. \ \forall \ i = 1, 2, \ ...n_j \ , \ Mods(\alpha_i) = g_j$$

For a succinct error profile, the following properties should hold, with $j = 1, 2... \ m$ groupings:

1. Each group has a unique set of worlds:

$$g_1 \neq g_2 \neq g_3 \ ... \ \neq g_m$$

2. Each sample belongs to only one group:

$$(\beta_1 \cap \beta_2 \cap \beta_3 ... \ \cap \beta_j) = \ \emptyset$$

Corollary 1 The total number of incorrect classifications N_{ω_k} from a model ω_k can be calculated from the following:

$$N_{\omega_k} = \textstyle\sum_{j=1}^{m} n_j e_j \quad \text{s.t. } e_j = \begin{cases} 1 \text{ if } \omega_k \in g_j \\ 0 \text{ else} \end{cases}$$

Corollary 2 If separate error profiles are created for two distinct groups (denoted A and B) of sizes U_A and U_B, then we can calculate odd ratios that represent the odds that a model incorrectly classifies group A vs incorrectly classifies group B. We can first represent the total odds ratio for a given model, where $N_{\omega_k,A}$ denotes the total number of incorrect classifications from model ω_k in group A.

$$OR_{\omega_k} \ = \frac{N_{\omega_k,A}/U_A}{N_{\omega_k,B}/U_B}$$

We can further define the relative contribution of each group G_j to the total odds ratio of a model ω.

$$OR_{\omega_k,j} \ = \frac{n_{j,A}e_{j,A} \ /U_A}{N_{\omega_k,B}/U_B}$$

Appendix E. Supplementary Figures

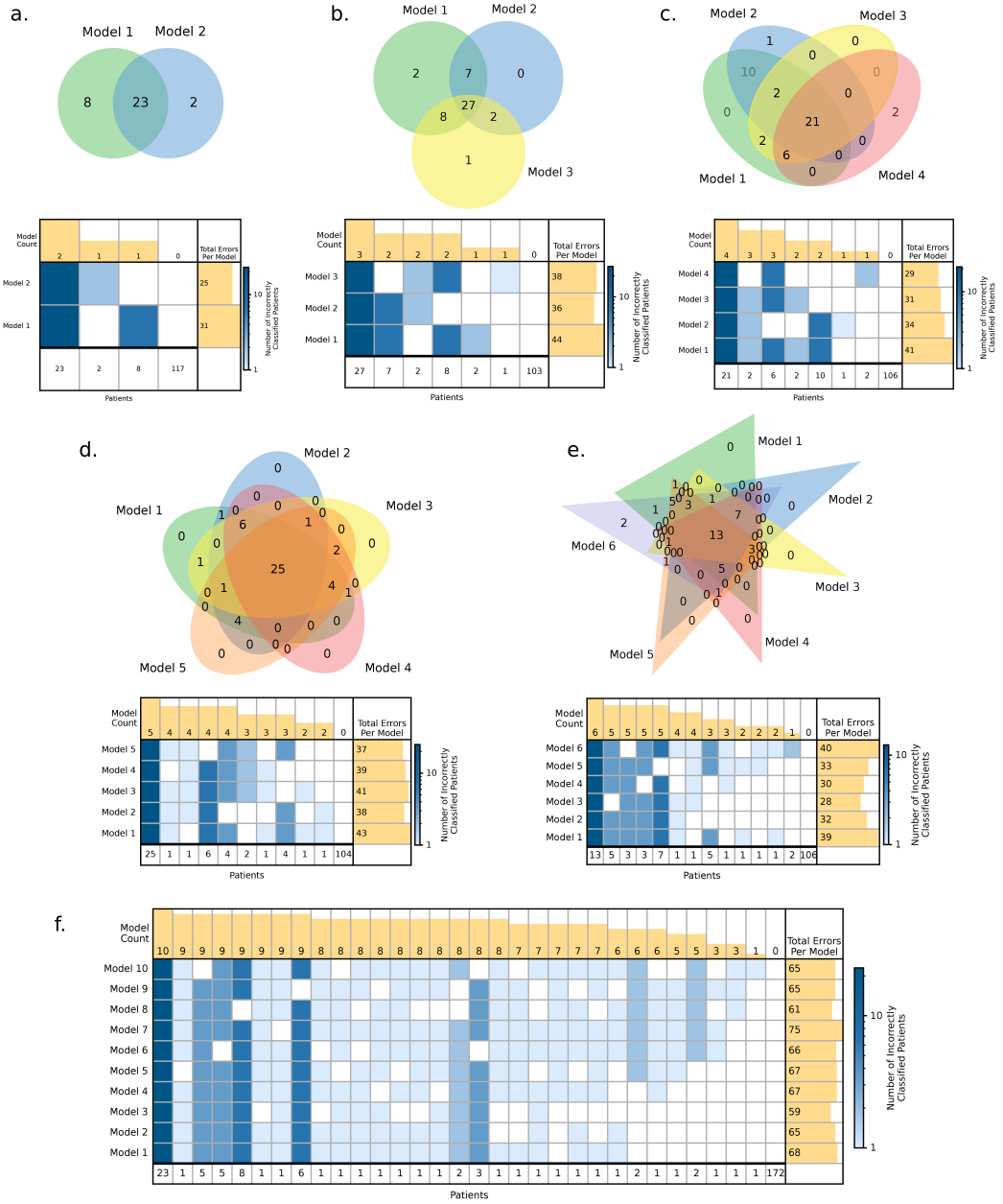

Figure A1: Qualitative comparisons between: a) two models; b) three models; c) four models; d) five models; and e) six models using Venn diagrams and our error profile with randomly generated data. Only the error profile can scale to a higher number of models such as 10 in f) while remaining interpretable.

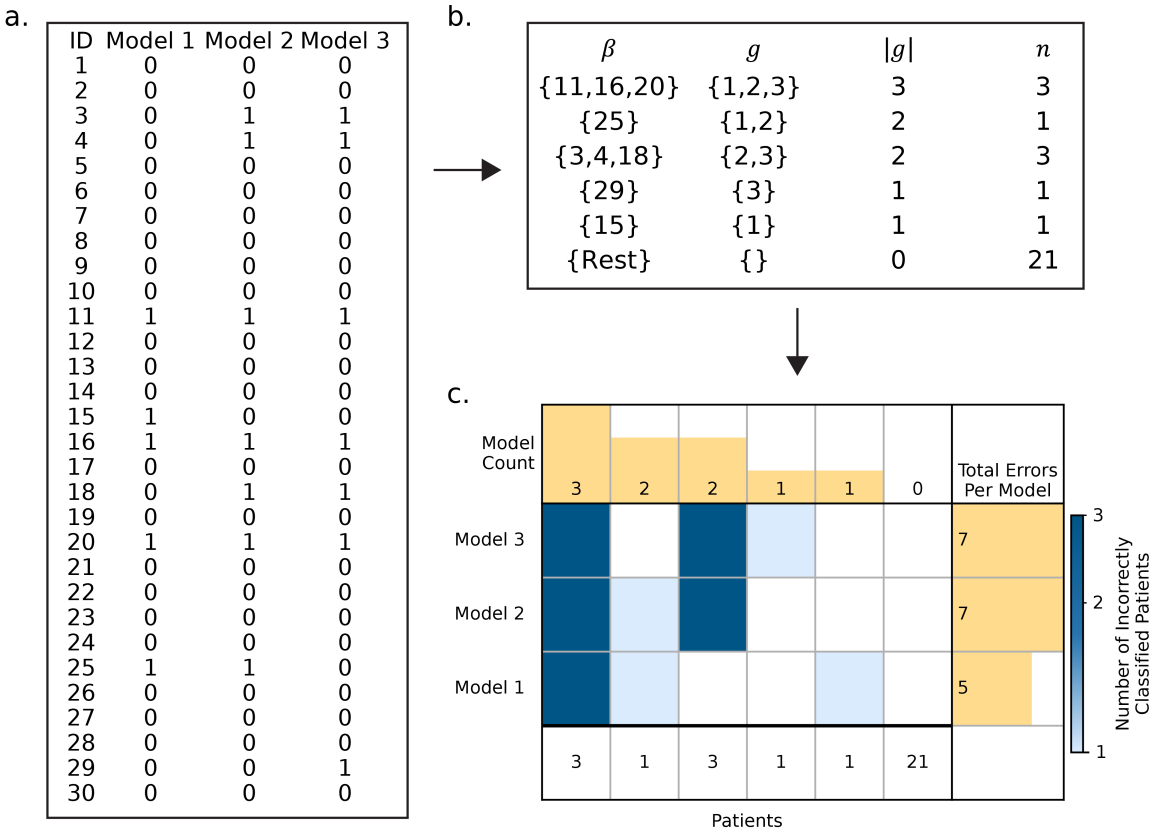

Figure A2: Example of error profile generation from random data with three models and 30 samples. Starting with results shown in a) (1/0 indicate incorrectly/correctly classified samples), we derive the set relationships in b) which ultimately form the final error profile in c).

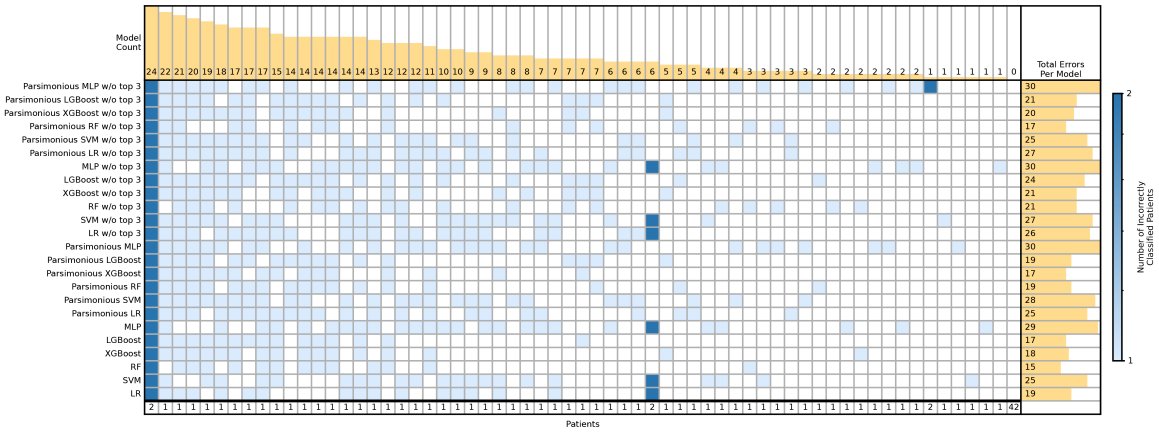

Figure A3: Error profile comparing the all models trained on the ECMO dataset, described in Table 1.

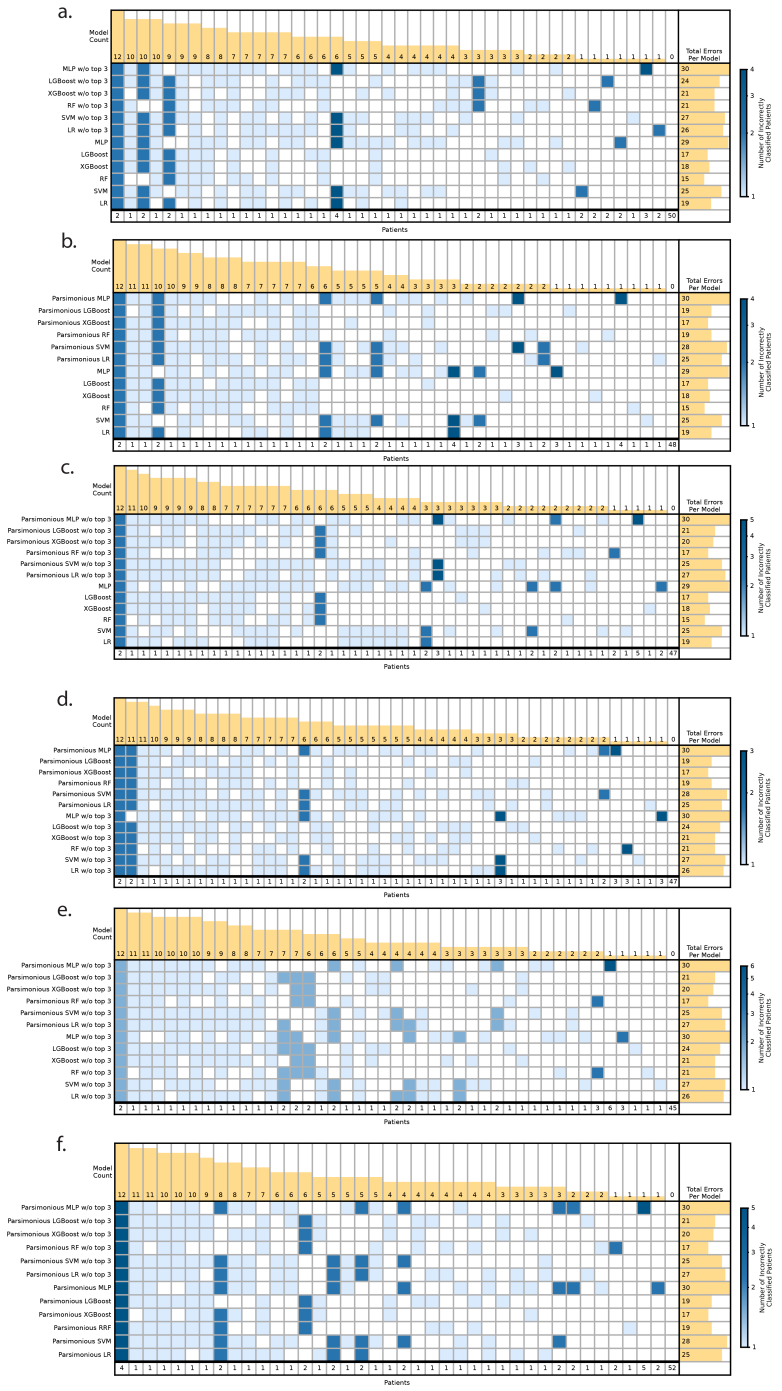

Figure A4: Error profile comparing the models trained on the ECMO dataset, described in Table 1. Each subfigure illustrates the isolated error analysis for pairs the model variations: a) baseline vs w/o top three; b) base-line vs parsimonious; c) baseline vs parsimonious w/o top three; d) parsimonious vs w/o top three; e) parsimonious w/o top three vs w/o top three; and f) parsimonious w/o top three vs parsimonious.

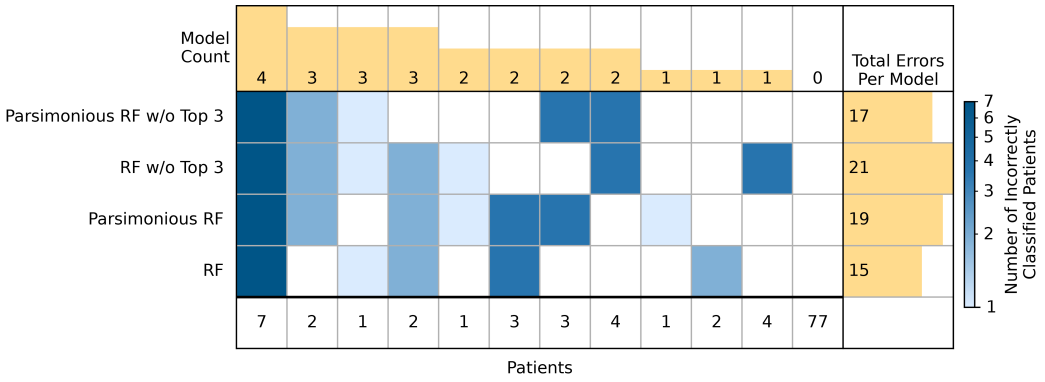

Figure A5: Error profile comparing the model variations for the optimal architecture (Random Forest) trained on the ECMO dataset, described in Table 1.

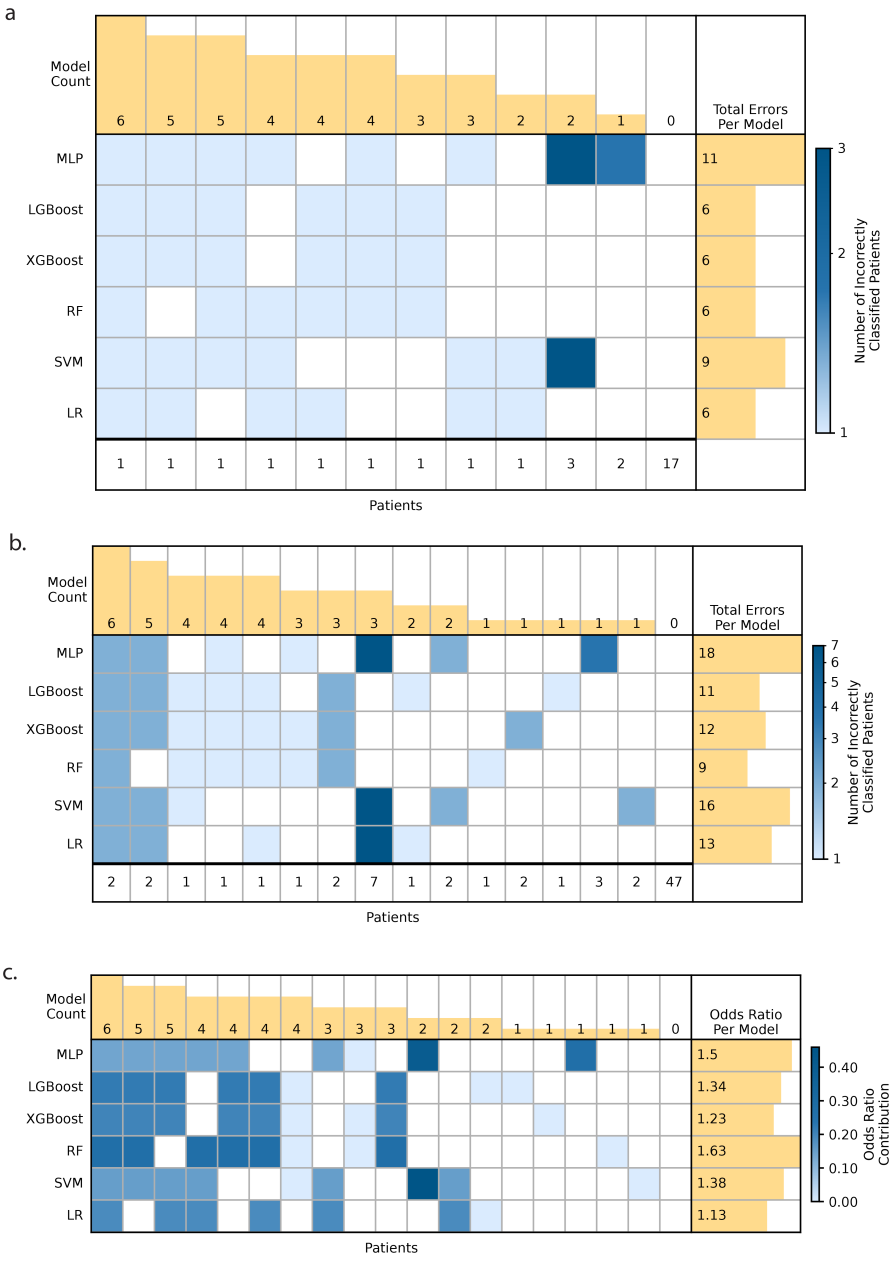

Figure A6: Exploratory extension of error profiling to subgroups using odds ratios. This analysis compares the base-line models trained on the ECMO dataset described in Table 1. a) Illustrates the error profile for the sub-set of females in the test set of the ECMO dataset. b) Illustrates the error profile for the subset of males in the test set of the ECMO dataset. c) Presents a comparison of the error profiles for females and males. Concretely, the right panel shows odd ratios, representing the odds that a model incorrectly classifies female patients vs incorrectly classifies male patients. Each cell, rather than representing the number of samples a given model classified incorrectly, now represents the relative contribution of the sample of patients towards the odds ratio (see A1 Corollary 2).

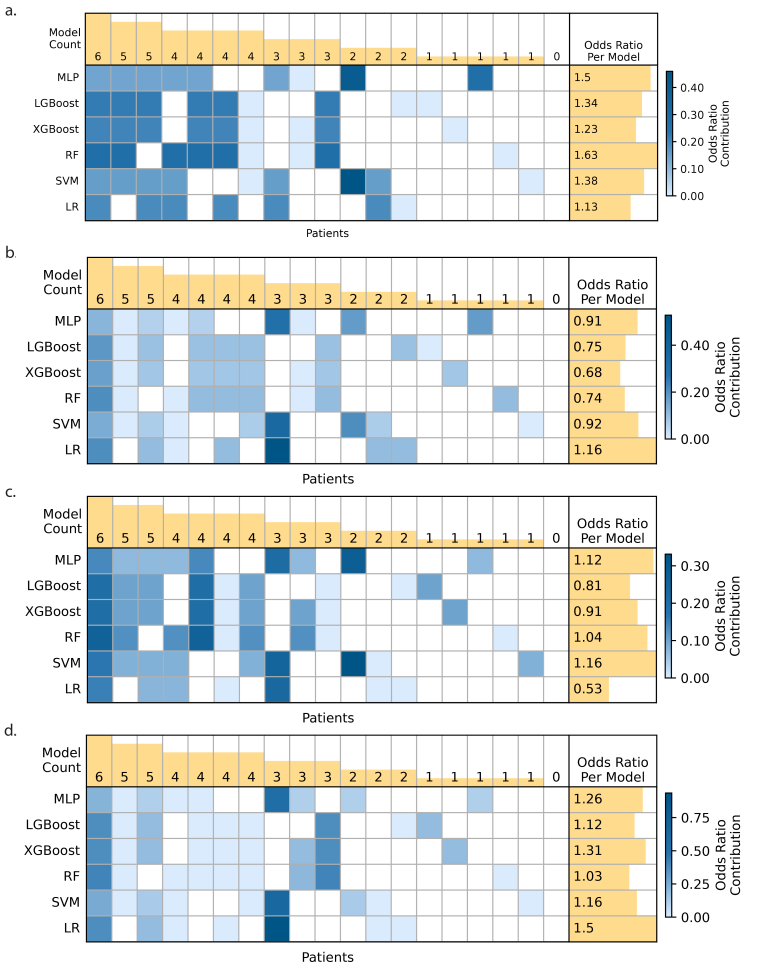

Figure A7: Exploratory extension of error profiling to subgroups using odds ratios. This analysis compares the baseline models trained on the ECMO dataset described in Table 1. Extending on Figure A6, this figure presents a comparison of the error profiles for different variables. While all subfigures are visually similar, they illustrate different odds ratios, as well as different contributions of samples to the total odds ratios. a) Odds that a model incorrectly classifies female patients vs incorrectly classifies male patients. b) Odds that a model incorrectly classifies patients greater or equal to 57 years of age vs incorrectly classifies patients less than 57 years of age (57 years is the median age in dataset). c) Odds that a model incorrectly classifies patients on ECMO for greater or equal to 105 hours vs incorrectly classifies patients on ECMO for less than 105 hours (105 hours is the median duration of ECMO prior to decannulation). d) Odds that a model incorrectly classifies patients with bleeding complications vs incorrectly classifies patients without bleeding complications.

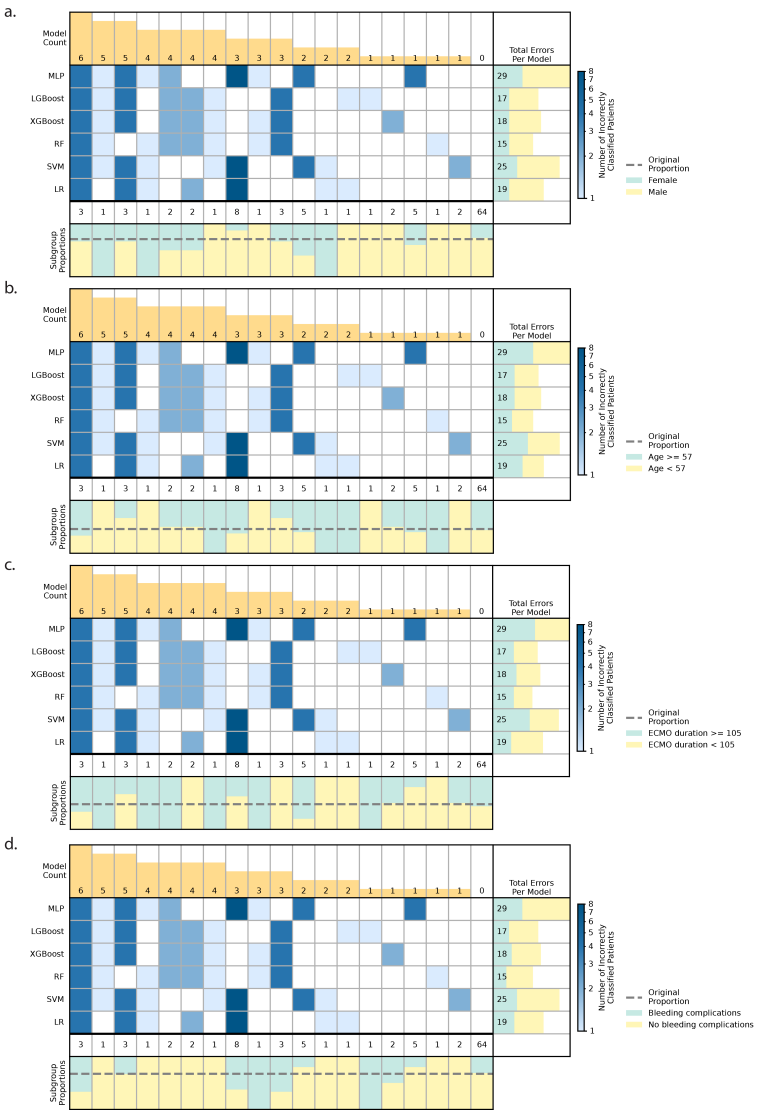

Figure A8: Exploratory extension of error profiling with visualization of subgroup distributions. This analysis compares the baseline models trained on the ECMO dataset described in Table 1. Each subfigure illustrates the distribution of patient subgroups within the incorrectly classified samples for each model (row, ω). For each group (column, G_j), the distribution of categories within the patients associated with the group is also visualized as proportions, with the original proportions in the test set indicated with the dotted horizontal line. While all subfigures are visually similar, they illustrate different subgroup distributions. a) Distribution of males and females. b) Distribution patients greater or equal to 57 years of age and patients less than 57 years of age (57 years is the median age in dataset). c) Distribution of patients on ECMO for greater or equal to 105 hours and patients on ECMO for less than 105 hours (105 hours is the median duration of ECMO prior to de-cannulation). d) Distribution of patients with bleeding complications and patients without bleeding complications.

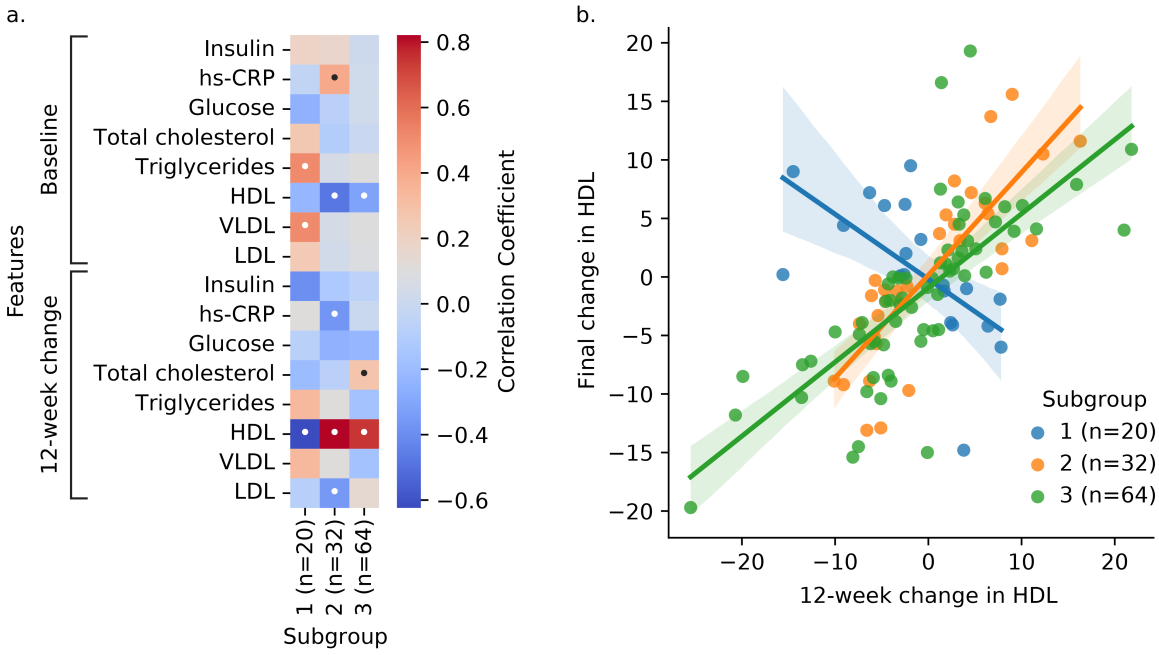

Figure A9: Exploratory extension of error profiling with statistical comparisons of subgroups. From the model with laboratory variables removed in Figure 4, we consider the subgroups that: 1) the model uniquely classified correctly; 2) the model classified incorrectly; and 3) all models classified correctly. a) Illustrates the correlation between the removed variables (y-axis) and the target variable (final change in HDL) for each subgroup, with the points indicating a statistically significant correlation. We draw attention to the 12-week change in HDL in b) as a likely predictor of final change in HDL. Notably, in the correctly classified subgroup 3, there is a strong correlation between 12-week change in HDL and final change in HDL. Subgroup 2 shares a similar relationship, but as the model does not have access to this information, incorrectly classifies the patients in this group. Remarkably, Subgroup 1 demonstrates an inverse relationship; however, as the model does not have access to this variable, it is not influenced by this strong signal, and identifies other predictors to successfully classify the patients in this group.

Appendix F. Survey Questions

Table A8: Survey questions. Abbreviations: Multiple choice, MC. Likert scale, LS

Use Case	Question Type	Question
1	MC	What number of patients were incorrectly classified only by the random forest model?
1	MC	What number of patients were incorrectly classified by both the SVM and MLP?
1	MC	How many models incorrectly predicted a unique set of 8 patients?
1	MC	Which model accurately predicted the outcome for the most patients?
1	MC	How many patients were classified incorrectly for the XGBoost model?
1	LS	How clear and understandable do you find the presented visualization?
1	LS	Does the visualization effectively convey the differences in model performance?
1	LS	How useful do you find the figure in providing insight into the errors made by different models?
1	LS	How effective is the visualization in highlighting subpopulations where different models disagree on predictions?
1	LS	To what extent does the visualization aid in selecting the most reliable model for clinical decision-making?
1	LS	How useful is the visualization in understanding the conditions under which different models fail?
1	LS	How well do you understand the concept of groups of distinct patients incorrectly classified by distinct sets of models as shown in the figure?
1	LS	How much does the figure help increase your confidence in understanding the reliability of different models?
1	LS	How much do you trust the information presented in the visualization?
2	MC	Which imputation technique classified the most patients incorrectly?
2	MC	How many patients were classified incorrectly by all imputation techniques?
2	MC	How many patients were classified incorrectly across KNN Imputation, EM Imputation, and MICE?
2	MC	Which imputation technique solely classified 7 patients incorrectly?
2	LS	Does the visualization effectively convey the differences in model performance?
2	LS	How useful do you find the figure in providing insight into the errors made by different models?
2	LS	To what extent does the figure help you compare the performance of different models?
2	LS	How effective is the visualization in highlighting subpopulations where different imputation methods influence model errors?
2	LS	How well do you understand the differences between imputation methods, such as MICE and k-nearest neighbors, as depicted in the visualization?

2	LS	Do the visualizations help in understanding the impact of different imputation techniques and feature sets on model performance?
2	LS	To what extent does the visualization aid in making better preprocessing choices for handling missing data?
2	LS	How much does the visualization help increase your confidence in the data imputation techniques used?
2	LS	How much does the figure help increase your confidence in understanding the reliability of different models?
2	LS	How much do you trust the information presented in the visualization?
3	MC	How many patients are classified incorrectly when removing Vitals?
3	MC	How many models distinctly classified the same 20 patients incorrectly?
3	MC	How many unique patients are incorrectly classified when you remove Labs?
3	MC	How many shared patients are incorrectly classified among the models that remove MR, HEI, and Acid data?
3	MC	How many patients were never incorrectly classified?
3	LS	How clear and understandable do you find the presented visualization?
3	LS	Does the visualization effectively convey the differences in model performance?
3	LS	How useful do you find the figure in providing insight into the errors made by different models?
3	LS	To what extent does the figure help you compare the performance of different models?
3	LS	How useful is the error profiling visualization in identifying problematic subpopulations?
3	LS	How well do you understand the differences in model performance when specific features, such as laboratory data, are removed, as depicted in the visualization?
3	LS	How useful is the visualization in understanding potential biases introduced by excluding specific feature sets?
3	LS	How much does the visualization help increase your confidence in the feature selection process used in model development?
3	LS	How useful is the error profiling visualization in identifying the effect of inclusion/exclusion of specific features on model performance?
3	LS	How much does the figure help increase your confidence in understanding the reliability of different models?
3	LS	How much do you trust the information presented in the visualization?
General	LS	How likely are you to use this type of visualization for deploying and evaluating machine learning models in healthcare?
General	LS	Without this visualization, how likely are you to make assumptions about machine learning model performance that could impact clinical decisions?
General	LS	To what extent do you believe this visualization will impact your decision-making process when evaluating machine learning models for healthcare applications?

General	LS	How useful do you find this visualization in identifying the strengths and weaknesses of different machine learning models?
General	LS	How effective is this visualization in facilitating the clinical implementation of machine learning models?
General	LS	How much does this visualization improve the transparency of machine learning model performance?
General	LS	How difficult would it be to understand the limitations and biases of machine learning models without this visualization?
General	LS	Without this visualization, how challenging would it be to identify critical errors in model predictions?
General	LS	How confident would you be in deploying machine learning models in healthcare without the insights provided by this visualization?
General	LS	Would you include this visualization in your evaluation of machine learning models?
General	Text	What do you like MOST about this visualization?
General	Text	What do you like LEAST about this visualization?
General	Text	What improvements or additional features would you suggest for this visualization to make it more useful for evaluating and deploying ML models in healthcare?
General	MC	Which of the following best describes you?

Appendix G. Code

The code to generate the error profiles is at: https://github.com/uclamii/superr_venn.