STRESS-TESTING PREDICTION MODELS: A SYSTEMATIC ROBUSTNESS EVALUATION ON LARGE-SCALE ICU DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Limited robustness to real-world perturbations is a key source of generalization failure in healthcare, yet it remains underexplored in systematic evaluations. We introduce a robustness evaluation that assesses multiple model classes under systematized clinically motivated perturbations ranging from missingness, measurement noise, label errors, to distribution shifts. Using three large-scale intensive care databases (MIMIC-III, MIMIC-IV, and eICU), we evaluate linear models, tree-based ensembles, neural networks, and tabular foundation models on both in-hospital mortality classification and ICU length of stay regression. Across most perturbation settings, transformer-based models exhibit smaller performance degradation than tree-based methods. However, robustness varies substantially across perturbation types and model implementations, and tuned gradient-boosting methods remain competitive in several settings. We further characterize the computational cost of robustness, finding that the robustness gains of tabular foundation models are accompanied by substantially higher prediction time than optimized gradient-boosting models. These results highlight that robustness is multidimensional and model-dependent, and that model selection for clinical deployment should consider not only predictive performance under nominal conditions, but also sensitivity to realistic data perturbations and computational constraints. Our benchmark provides a systematic framework for evaluating these trade-offs in clinical tabular prediction.

1 INTRODUCTION

While their purpose has not changed over time, the methods used to design and evaluate prediction models (or scores) in healthcare have evolved. For instance, Le Gall et al. (1984) introduced the Simplified Acute Physiology Score (SAPS), which provides a score for the risk of mortality during an intensive care unit (ICU) stay based on 14 clinical and biological variables. The score was simple enough to be calculated by hand. Now, with the availability of large-scale health databases, prediction models in healthcare are developed using machine learning (ML) methods that can capture subtle signals from a wide range of patient characteristics (Shmatko et al., 2025).

The robustness of these models, that is, how resilient they are to perturbations, variations, that does not match what they were trained on, has been less studied, even though recent research has highlighted the importance of that concept in healthcare (Lekadir et al., 2025). However, insufficient robustness can contribute to deployment failure when differences in patient populations, measurement processes, target definition, clinical workflows, or data quality cause deployment data to depart from development conditions (Wong et al.; Obermeyer et al., 2019; Cohen et al.; Beede et al., 2020). In a recent systematic review, Balendran et al. (2025) proposed a taxonomy of robustness concepts for prediction models in healthcare.

In clinical settings, tabular data from electronic health records (EHRs) contains diverse types of data including demographics, laboratory results, diagnoses, procedures, and administrative/billing information (Johnson et al., 2023; 2016). These rich EHR databases can be leveraged to extract and describe clinically relevant patterns from patients' health. However, this results in complex datasets combining continuous, categorical, and ordinal variables, often accompanied by challenges such

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

as measurement errors, missing values, misdiagnosis/labeling errors, and distribution shifts across populations from different hospitals or at different times (Goldstein et al., 2017).

ML-based prediction models on tabular data have historically been dominated by tree-based methods, particularly approaches based on gradient-boosted decision trees (GBDT) (McElfresh et al., 2023; Shwartz-Ziv & Armon; Grinsztajn et al., 2022). Nevertheless, a small number of recent deep learning approaches have demonstrated that carefully designed architectures can be competitive with GBDTs on tabular data (Holzmüller et al., 2025; Erickson et al., 2025). At the same time, obtaining strong performance from both specialized neural networks and GBDTs can depend substantially on model configuration and hyperparameter optimization (HPO), motivating approaches that aim to reduce the need for dataset-specific model selection and tuning. For example, Hollmann et al. (2025) introduced TabPFN, a tabular foundation model that uses in-context learning, and has demonstrated performance competitive with gradient boosting on various benchmark datasets without any dataset-specific tuning (Erickson et al., 2025). New versions and variants of this model have since been proposed, slowly contributing to a growing body of tabular foundation models Grinsztajn et al. (2026); Qu et al. (2026); Ma et al. (2025).

Existing tabular benchmarks focus on evaluating different models across extensive collections of datasets, Grinsztajn et al. (2022); McElfresh et al. (2023); Erickson et al. (2025); Gardner et al. (2024); Yıldız & Kalayci (2025); Liu et al.; Purucker et al. (2026); Berta et al. (2026). While valuable for establishing general-purpose baselines, performance on static tasks rarely reflects the non-stationary, messy conditions characteristic of real-world clinical deployments Gu et al. (2025). Conversely, studies that do evaluate robustness typically focus on isolated, single-type perturbations Cheng et al. (2025); Hendrycks & Dietterich (2019); Perez-Lebel et al. (2022).

In this work, we make the following contributions:

- We introduce a systematic robustness evaluation for clinical tabular prediction that evaluates models under a broad range of clinically motivated perturbations, including label noise, input noise, missing data, class imbalance, model training, feature selection, subgroup analysis, dataset shift, and combinations of perturbations.
- Using three large-scale ICU databases and two prediction tasks, we compare linear models, tree-based ensembles, a neural network, and recent tabular foundation models under both individual and composite perturbations, and examine how HPO affects their robustness.
- We characterize the trade-off between one particular measure of robustness and computational efficiency, providing a better perspective on model suitability for clinical deployment.

2 METHODS

2.1 CASE STUDY

We conducted our study using three intensive care databases: the Medical Information Mart for Intensive Care (**MIMIC-III**, v1.4) (Johnson et al., 2016), **MIMIC-IV**, (v3.1) (Johnson et al., 2023), and the **eICU Collaborative Research Database**[1] (Pollard et al., 2019). These databases contain routinely collected clinical data from patients admitted to ICUs. MIMIC-III and MIMIC-IV comprise data from a single hospital in the United States, covering the periods 2001–2012 and 2008–2022, respectively, whereas eICU is a multicenter database covering admissions from 2014–2015. The databases include a broad range of information collected during ICU stays, including demographics, physiological measurements, laboratory tests, interventions, and diagnostic information. These datasets have been extensively used for clinical machine learning research, including studies of fairness in hypoxemia detection (Wong et al., 2021), privacy-preserving methods (Bey et al., 2020), and mortality prediction (Hou et al., 2020).

We considered two prediction tasks:

- **Classification:** predicting patients' in-hospital mortality.
- **Regression:** predicting patients' length of stay in the ICU.

[1]For conciseness, we will refer to the dataset as **eICU**

For each experiment, one database was used as the development dataset, while the other two were used as external datasets for evaluation. We considered all three possible choices of development dataset, thereby evaluating each dataset both as a development and as an external dataset.

Because some patient admissions are present in both MIMIC-III and MIMIC-IV, we took additional steps to prevent admission-level overlap between the development and external datasets. Specifically, the MIMIC-IV cohort was restricted to admission occuring 2014 or later[2] whenever MIMIC-III and MIMIC-IV were paired, which the largest non overlapping period with MIMIC-III (2001–2012).

Features used to develop the models were informed by the variables included in the Simplified Acute Physiology Score II (SAPS II) (Le Gall et al., 1993).

More information of the preprocessing and a descriptive table are provided in Appendix C.

2.2 Machine Learning Models

We considered a diverse set of modeling approaches spanning four families, selected to cover a range of inductive biases and architectural paradigms commonly used for tabular prediction and represented in recent benchmarks (Purucker et al., 2026; Erickson et al., 2025).

1. Linear and Regularized Baselines: As low-complexity and interpretable baselines Van Calster et al.; Christodoulou et al., we considered **Linear/Logistic Regression** together with its penalized variants: **LASSO** (L_1 regularization) and **Ridge** (L_2 regularization).

2. Tree-based Ensembles: We considered tree-based ensemble methods, which are strong and widely used approaches for tabular data McElfresh et al. (2023); Shwartz-Ziv & Armon. Specifically, we included **Random Forest** and gradient boosted decision trees (GBDTs). To capture differences in optimization, tree growth, regularization, and feature handling across GBDT implementations, we evaluated four variants:

- **Gradient Boosting** implementing the original gradient boosting framework.
- **XGBoost** Chen & Guestrin (2016) for scalable, regularized gradient boosting.
- **LightGBM** Ke et al. (2017) for histogram-based optimization and leaf-wise tree growth.
- **CatBoost** Prokhorenkova et al. (2018) well suited for handling categorical features with minimal preprocessing.

3. Neural Network: As a representative conventional deep learning approach for tabular data, we considered **RealMLP** (Holzmüller et al., 2025), a feed-forward neural network with architectural and optimization improvements designed for tabular prediction.

4. Tabular Foundation Models: Finally, we evaluated tabular foundation models that perform in-context inference, i.e., make prediction without gradient-based optimization of model parameters on the target dataset. We considered the following models:

- **TabPFNv3** Grinsztajn et al. (2025; 2026), a transformer-based Prior-Data Fitted Network pretrained on millions of synthetic datasets.
- **TabICLv2** Qu et al. (2026), which extends the in-context learning framework to larger context sizes and employs a different feature-embedding strategy.
- **TabDPT** Ma et al. (2025), a tabular foundation model whose pretraining procedure also incorporates real-world tabular datasets[3].

2.3 Model Development

Hyperparameters for models requiring tuning were optimized using **10-fold nested cross-validation**. Within each outer training set, we tuned models requiring HPO using a three-fold

[2]The admission year in MIMIC-IV is reported in period of three years: 2008–2010, 2011–2013, 2014–2016, 2017–2019, 2020–2022

[3]None of the three ICU datasets considered in our evaluation were part of the datasets reported by the authors for fine-tuning TabDPT.

inner cross-validation procedure with a budget of 15 randomly sampled hyperparameter configurations per model. For gradient boosted trees, candidate configurations were evaluated using early stopping using 10% of each inner training fold. The complete hyperparameter search spaces are reported in Appendix B. For foundation models, we used the default hyperparameters provided by their respective implementations.

For models without native support for missing values, missing continuous features were imputed using mean imputation and a binary missingness indicator was added for each imputed feature Perez-Lebel et al. (2022). Missing categorical features were imputed using the most frequent category, together with a corresponding binary missingness indicator. For models with native support for missing values, namely **Random Forest**, **LightGBM**, **XGBoost**, **CatBoost**, and the foundation models, we relied on the respective implementations' native missing-value handling mechanisms. Features were standardized prior to training the linear models.

For the classification task, we additionally calibrated the predicted probabilities of each model using Platt scaling Niculescu-Mizil & Caruana. More specifically, within each outer cross-validation iteration, the validation fold was further partitioned into a calibration subset and an evaluation subset. 10% of the held-out fold was used exclusively to fit Platt scaling, while the remaining was used exclusively to estimate model performance. The calibration subset was not included in the calculation of performance metrics.

2.4 Evaluation Methods

For classification, model performance was assessed using the **AUC** and **Brier score** Steyerberg et al. (2010). For regression, we considered Mean Absolute Error (**MAE**), Root Mean Squared Error (**RMSE**), and R^2.

Performance on the development-set was estimated using the outer folds of the 10-fold nested cross-validation. The two other datasets were used exclusively for external validation.

To provide an overall measure of model robustness across the diverse set of perturbations, we also considered the **mean Robustness Score (mRS)** adapted from Hendrycks & Dietterich (2019); Zhang et al. (2022) which aggregates the performance across the different levels of noise (or categories) within each scenario and normalizes it relative to a baseline model, here, logistic regression before averaging across all scenarios. We refer to Appendix E for details on how the score was derived.

In clinical practice, inference is performed repeatedly, whereas model training is typically performed only once during development. Therefore, in this work, we focus on **prediction time** as the primary measure of computational efficiency. For each model, we therefore report for each development dataset, the median prediction time across all perturbation scenarios considered in the benchmark.

Finally, we adapt the Elo rating (Erickson et al., 2025) to assess the robustness of the different models to composite perturbations presented in section 2.5. The details regarding the Elo rating methodology are presented in Appendix F.

2.5 Types of Perturbations

Building on the robustness taxonomy proposed in healthcare in Balendran et al. (2025), we considered several families of perturbations: input noise, missing data, label noise, imbalanced data, feature selection, model training, and subgroup analysis and dataset shift. For each family of perturbations, we consider different scenarios. Moreover, we also consider composite perturbations, where multiple perturbations are added successively.

We give here a brief description of each family and scenario. Details on the different methods are presented in appendix D.

2.5.1 Univariate perturbation

Label Noise. Clinical outcome labels may be affected by documentation errors, imperfect diagnostic criteria, coding inaccuracies, proxy targets, or incomplete follow-up. We considered five scenarios: (1) random label corruption, with binary labels flipped in both directions ($0 \leftrightarrow 1$) and chi-squared

noise added to continuous outcomes; (2) class-conditional noise affecting only positive binary labels ($1 \rightarrow 0$); (3) class-conditional noise affecting only negative binary labels ($0 \rightarrow 1$); (4) age-dependent corruption, where the probability or magnitude of label perturbation increased with patient age; and (5) alternative proxy targets for mortality.

The proxy targets corresponded to ICU mortality, excluding deaths occurring after ICU discharge, and mortality observed during follow-up after hospitalization. For all scenarios except the proxy-target analysis, the level of noise was progressively increased.

Input Noise. Clinical measurements are inherently subject to measurement variability, differences in acquisition procedures, and occasional recording errors, motivating the assessment of model robustness to noisy clinical inputs. We considered three settings for adding noise to the clinical features: (1) only to inputs in the training data, (2) only to the validation data, and (3) to both training and validation data. Additionally, we assessed the impact of introducing noise exclusively to continuous features versus categorical features. For each scenario, we incrementally increased the noise level to measure the degradation in model performance.

Imbalanced Data. Clinical outcomes are often unequally distributed, with events such as mortality occurring less frequently than non-events, making it important to assess whether model performance remains reliable under more extreme outcome imbalance. As the dataset was initially imbalanced with a majority of observations coming from patients who survived, we further reduced the number of samples in the minority class (patients who died) to achieve extreme levels of class imbalance. We did not consider this test for the regression task as imbalance for a continuous target cannot be defined.

Missing Data. Missing clinical information is common in healthcare because measurements may not be collected for every patient, may depend on clinical decisions, or may be unavailable across care settings, making robustness to different missingness mechanisms clinically relevant.

More specifically, we evaluated models under three settings: (1) data were missing completely at random (MCAR), (2) data were missing at random (MAR), and (3) data were missing not at random (MNAR). For each setting, we explored the effect of having missing data only in the training data, validation data, or both. For each scenario, we varied the proportion of missing values. Methods to generate the different missingness mechanism were taken from Muzellec et al. (2020).

Feature Selection. In clinical applications, models may encounter a large number of potentially irrelevant or weakly informative variables, making it important to assess whether predictions remain driven by genuinely informative clinical features rather than spurious or irrelevant inputs. To explore the model robustness to predictive clinical features in the presence of irrelevant inputs, we randomly shuffled values of some of the features in three different scenarios: (1) shuffling a subset of features only in the training set, (2) shuffling a subset of features only in the validation set, and (3) shuffling features in both. For each scenario, we incrementally increased the proportion of features for which we randomly permuted the values.

Model training. Clinical datasets can vary substantially in size, and model performance may depend on the amount of data available for learning or the choice of training configuration, motivating an assessment of robustness to limited training data and hyperparameter optimization. We evaluated models' robustness to limited training data. To do so, we trained each method by gradually decreasing the size of the training data.

We also performed a sensitivity analysis where we evaluated each method, except for foundation models, on the exact same perturbations with and without performing HPO.

Subgroup Analysis and Dataset Shift. Clinical models should generalize across patient populations, institutions, making it essential to assess whether their performance is consistent across subgroups and under changes in the data distribution. We evaluated the performance of each method across demographic subgroups and clinical strata based on patients' gender, ICU service, and age group, on both the development dataset and the two other external datasets. When eICU or MIMIC-IV was used as external dataset, we also evaluated robustness to geographical and temporal shift, using the hospital region and the admission year, respectively.

2.5.2 COMPOSITE PERTURBATION

In real-world clinical settings, multiple perturbations may occur simultaneously, making it important to assess whether model robustness is maintained when several perturbations are present jointly. We conducted an exploratory assessment of model robustness on composite perturbation, defined as the sequential application of three scenarios across different families of perturbation, sampled uniformly at random.

3 RESULTS

Foundation models show strong but task- and dataset-dependent robustness. In many scenarios, foundation models showed better performance. In particular, TabPFNv3 and TabICLv2 outperformed the best gradient boosting method (Fig 1). The methods were especially more robust on perturbations related to class imbalance, training data, and extreme level of label noise ($1 \rightarrow 0$ and $0 \rightarrow 1$). However, in some particular settings, such as when eICU was used as the development set, gradient boosted trees such as CatBoost performed better on both external datasets (Fig 43). TabICLv2 and TabDPT performed poorly both in classification and regression when perturbations were applied only on the validation data (Fig 26, 30, 36, 24, and 40) For regression, some foundation models also appeared to show higher performance degradation at high levels of noise such as missing data (Fig 33).

The same pattern was observed on composite perturbations. For classification, all three foundation models achieve the highest Global Robustness Elo (Fig 3), followed by optimized gradient boosted trees. However, for regression, only TabPFNv3 ranked first under composite perturbations, while RealMLP without and with HPO, ranked second and third, respectively.

The mean Robustness Score also offers similar interpretation: on both MIMIC-III and MIMIC-IV, TabPFNv3 achieve the highest mRS (1.49 and 1.47 on MIMIC-III and MIMIC-IV, respectively) (Fig 2 and 18), whereas on eICU, the best mRS is achieved by CatBoost (Fig 19).

Hyperparameter optimization does not uniformly improve robustness For the classification task, HPO generally improved performance under missing-data and label-noise perturbations, with particularly large gains observed for XGBoost (Fig 2, 5, and 7). Moreover, HPO also improved Brier scores in class imbalance and training size settings (Fig 4). For label noise, the benefit of HPO tended to increase as the perturbation severity increased, suggesting that model configuration became increasingly important under stronger target corruption. Moreover, for regression, the regularization strength also influenced the performance of the penalized linear models, in particular for the LASSO under different families of perturbations (Fig 10). By contrast, models such as CatBoost, LightGBM, and Ridge tended lie along the diagonal, indicating relatively limited changes in performance after tuning across the evaluated perturbations.

This is also supported with composite perturbations. HPO increased the ELO for several tree-based methods, particularly XGBoost (increased from 916 to 1445 and from 681 to 1157 for classification and regression, respectively). For some model–dataset combinations, however, HPO reduced Global Robustness ELO; for example, Gradient Boosting showed lower performance after tuning in some external validation settings, especially on eICU data (Fig 10). The effect of HPO was inconsistent for RealMLP across the evaluated perturbations and, in several classification settings, tuning was associated with lower performance (Fig 2). One possible explanation is that the default RealMLP configuration was derived from extensive meta-benchmarking and may therefore already provide a strong general-purpose configuration when the budget for HPO is limited (Holzmüller et al., 2025).

Robustness–inference time trade-offs differ across model families and development datasets While foundation models achieve the highest mRS (Fig 2 and 18), this was accompanied by substantially longer prediction times than those of tree-based methods (when MIMIC-III was used as the development set, the median prediction time for TabPFNv3 was 5.12s against 0.02 s for the tree-based method with the highest mRS, here LightGBM). Tuning some of the tree-based method such as LighGBM, XGBoost, and Random Forest improved the mRS but also increased the inference time.

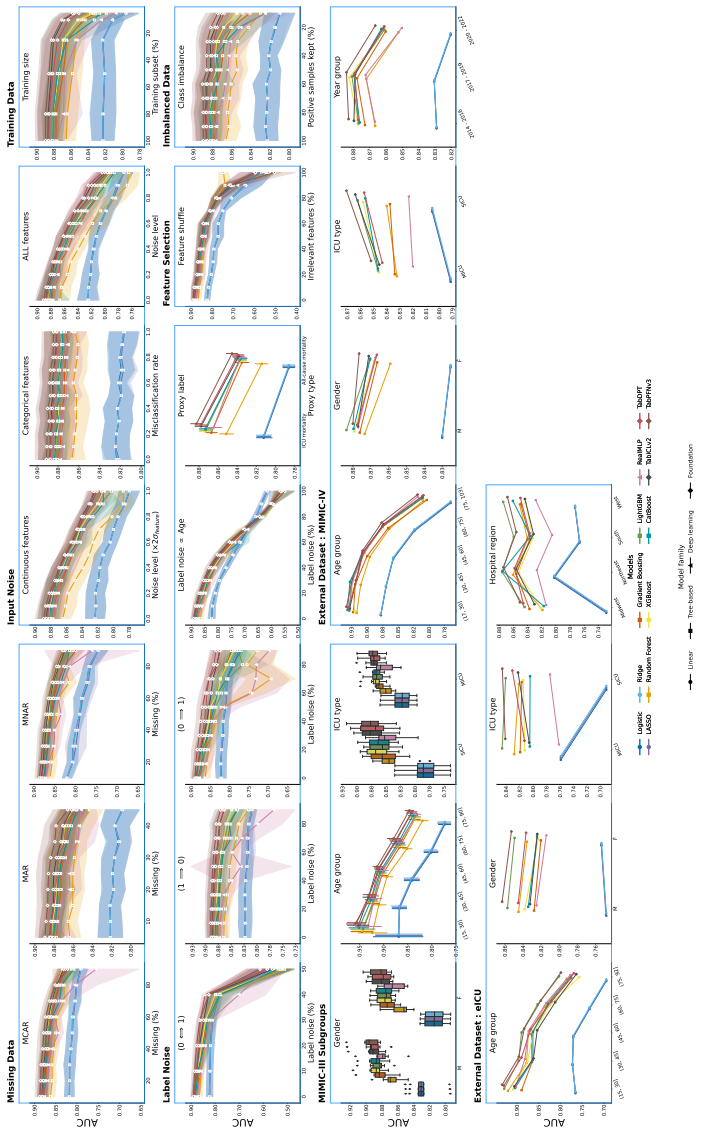

Figure 1: AUC under training-data perturbations and external validation: MIMIC-III development cohort

4 Discussion

In this study, we evaluated the robustness of various machine learning models for predicting patients' in-hospital mortality and ICU length of stay under a wide range of perturbations. Across many scenarios, foundation models performed better than gradient-boosting methods and sometimes markedly better at varying level of noise. One possible explanation is that the pretraining of foundation models is exposed to various synthetic data configuration with a broader range of feature configurations and perturbation such as missing data (Hollmann et al., 2025). These findings align with recent evidence suggesting that deep learning approaches, once outperformed by tree-based methods on tabular data, are now closing this gap Erickson et al. (2025).

Nonetheless, prediction time remains non-negligible (Fig 2). Another practical consideration is software accessibility: many (tabular) foundation models are primarily available through Python libraries and are not yet as readily integrated into clinical workflows exclusively centered on R, which may limit their immediate adoption in some institutions. These software constraints, together with the greater computational requirements, should therefore be considered when assessing practical de-

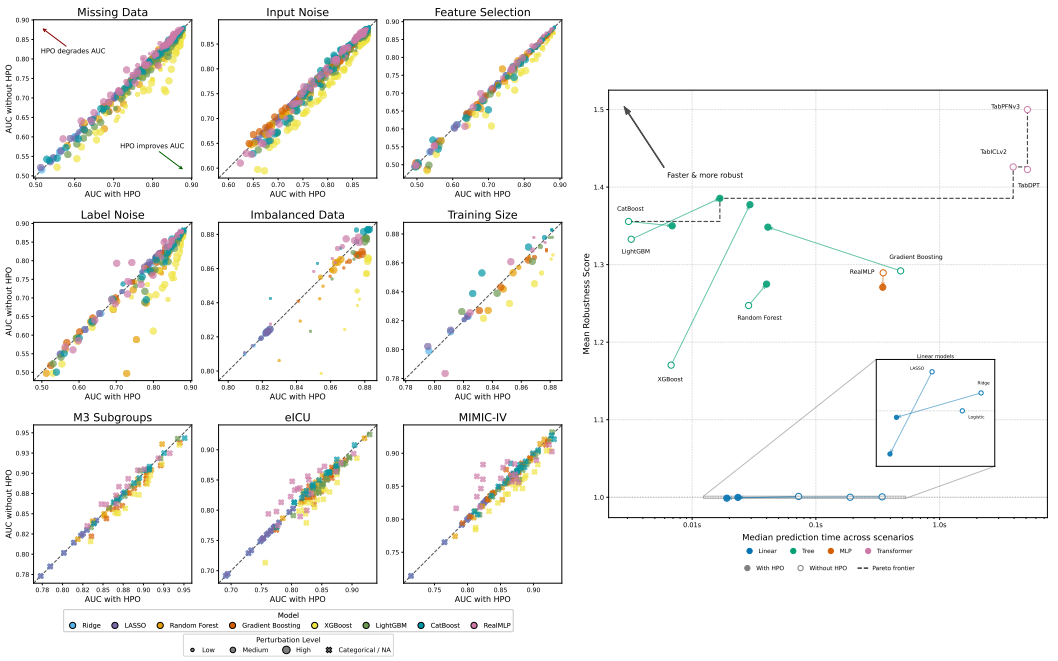

Figure 2: **Left:** Effect of HPO across perturbation families. Each point represents the mean AUC for a scenario at a specific noise level or category. **Right:** Mean Robustness Score versus median inference time across all scenarios. **MIMIC-III** is used as the development set for classification (AUC).

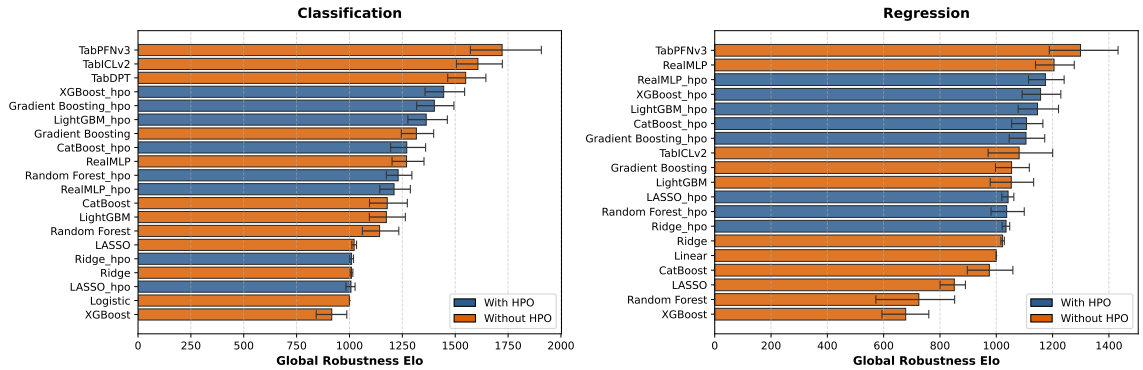

Figure 3: **Left:** Global robustness ELO computed on perturbation composites for classification. **Right:** Global robustness ELO computed on perturbation composites for regression.

ployability (Beede et al., 2020). In contrast, conventional tree-based methods generally have lower inference costs, although achieving their best performance may require extensive hyperparameter optimization, particularly for methods such as XGBoost and LightGBM.

While we focus on a model-centric perspective on robustness in this study, it is important to consider other aspects of model development. Robustness can often be improved through data quality control, expert-guided feature selection, and well-designed imputation or feature engineering strategies Lopez-Ayala et al. (2025); Nestor et al. (2019). These considerations should also inform model selection, as they may enable simpler, faster, and more interpretable models to achieve comparable performance to complex architectures, particularly in real-world healthcare settings where transparency and interpretability are crucial Rudin (2019).

Various benchmarks have been proposed, contributing to a leaderboard-oriented evaluation. However, such evaluations do not provide a clear understanding of how a model would translate from development to real-world deployment. In contrast to existing benchmarks, our benchmark offers a task-specific systematic evaluation across diverse robustness concepts at a task-specific level, better aligning with a real-world application. In practice, the relevance of a specific robustness test should reflect realistic data conditions. For example, robustness to label noise is most relevant for outcomes derived from diagnostic tests with low sensitivity or specificity, whereas it is less critical for endpoints such as in-hospital mortality, which are typically less subject to labeling errors. Likewise, evaluating models under extreme conditions, such as high levels of missingness or very low signal-to-noise ratios, may not be meaningful for all clinical applications. Thus, selection of a predictive model should be guided more by a robustness evaluation within a well-defined family of potential perturbations and within a reasonable range of noise where simpler methods may be more appropriate than more complex models.

Future work. This analysis could be extended to other data modalities (e.g., text, imaging, multimodal data), and prediction tasks Gu et al. (2025); Purucker et al. (2026). With the emergence of foundation models for tabular data, it will also be essential to broaden robustness concepts to include adversarial perturbations Finlayson et al. (2019), input preprocessing Wolff & Hulsebos (2025); Blayer et al. (2026), foundation-model-specific concerns, including pretraining-data contamination, memorization, privacy leakage, and extraction attacks Carlini et al. (2020) to retrieve data used during pretrained (TabPFNv3 and TabICLv2 only rely on synthetic data whereas real data is used pretrained TabDPT).

Limitations. In this study, we chose to assess robustness through absolute performance degradation under perturbation (illustrated with the mRS). However, robustness can also be defined as the relative degradation compared to whereas an alternative approach would be to measure the model relative degradation to the model performance on unperturbed baseline. Second, the features used to develop our prediction models were derived from the SAPS II (developed in 1993), which does not fully represent the granularity of modern clinical databases. Future studies should examine robustness when considering larger set of features and different sampling strategies (e.g., hourly measurements). Also, as with many data analysis studies, methodological choices, such as data preprocessing pipeline, imputation strategy, method hyperparameter optimization, and model evaluation protocol, can introduce additional variability in performance estimates Botvinik-Nezer et al. (2020). Finally, while we included recent foundation models, this is a fast-growing domain. These models are frequently updated and their performance may be different in the near future.

5 Conclusion

Using large-scale real-world ICU datasets, we systematically evaluated the robustness of machine learning models for in-hospital mortality and ICU length-of-stay prediction under a broad range of clinically motivated perturbations. Tabular foundation models showed strong performance across several robustness dimensions, including under combinations of perturbations, but their advantage was not consistent across all development datasets. Although electronic health records provide large amounts of data, many clinical settings have limited access to specialized hardware and software, and some applications require low-latency or real-time prediction. In such settings, conventional tree-based methods may provide a favorable trade-off between robustness, computational efficiency, and practical usability. Model selection for clinical deployment should therefore consider not only predictive performance under perturbation, but also the computational resources, latency requirements, and operational constraints of the intended use case.

AI use statement

In this work, we used generative AI tools to provide feedback on research methodology and experimental design and to assist with the implementation of methods in software code. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, assist with mathematical proofs, translate research material, clean or reformat the study datasets, support qualitative or thematic data analysis, or interpret the study results; these tasks were either performed by the authors or were not applicable to this work.

Additionally, we used generative AI tools for brainstorming research ideas, editing and optimizing software code, editing visualization code to improve the clarity of figures, editing manuscript text to improve readability, and proposing potential titles and keywords. All AI-assisted outputs were reviewed by the authors before being incorporated into the work. LLM-generated or modified code was inspected by the authors. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

Ethics statement

This study used secondary, de-identified clinical data from the MIMIC-III, MIMIC-IV, and eICU Collaborative Research Database, which are distributed through PhysioNet under controlled-access conditions. Access was obtained by credentialed researchers who completed the required training. All data were handled in accordance with the corresponding data-use requirements, and no attempt was made to re-identify individual patients.

Subgroup analyses (on age group and gender) were conducted to characterize variation in predictive performance rather than to support differential clinical treatment of demographic groups.

Reproducibility statement

Descriptions of the experimental design, model development procedure, evaluation protocol, and robustness analyses are reported throughout the Methods and Appendix. We also report the software and hardware environment used for model training and inference in the Appendix A. The code used to construct the benchmark, apply the perturbations, train and evaluate the models, and reproduce the reported analyses and figures can be found at the following link. The datasets (MIMIC-III, MIMIC-IV, and eICU) used in this study can not be shared but are available through PhysioNet subject to their respective access requirements. The code to extract the raw data are available in the repository.

References

Alan Balendran, Céline Beji, Florie Bouvier, Ottavio Khalifa, Theodoros Evgeniou, Philippe Ravaud, and Raphaël Porcher. A scoping review of robustness concepts for machine learning in healthcare. 8(1):38, 2025. ISSN 2398-6352. doi: 10.1038/s41746-024-01420-1. Publisher: Nature Publishing Group.

Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12. ACM, 2020. ISBN 978-1-4503-6708-0.

Eugène Berta, David Holzmüller, Francis Bach, and Michael I. Jordan. Calarena: A large-scale post-hoc calibration benchmark, 2026. URL https://arxiv.org/abs/2605.30188.

Romain Bey, Romain Goussault, François Grolleau, Mehdi Benchoufi, and Raphaël Porcher. Fold-stratified cross-validation for unbiased and privacy-preserving federated learning. 27(8):1244–1251, 2020. ISSN 1067-5027. doi: 10.1093/jamia/ocaa096.

Gioia Blayer, Myung Jun Kim, Félix Lefebvre, Lennart Purucker, Alan Arazi, Eilam Shapira, Roi Reichart, Frank Hutter, Marine Le Morvan, David Holzmüller, and Gaël Varoquaux. Strable: Benchmarking tabular machine learning with strings, 2026. URL https://arxiv.org/abs/2605.12292.

Rotem Botvinik-Nezer, Felix Holzmeister, Colin F. Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A. Mumford, R. Alison Adcock, Paolo Avesani, Blazej M. Baczkowski, Aahana Bajracharya, Leah Bakst, Sheryl Ball, Marco Barilari, Nadège Bault, Derek Beaton, Julia Beitner, Roland G. Benoit, Ruud M. W. J. Berkers, Jamil P. Bhanji, Bharat B. Biswal, Sebastian Bobadilla-Suarez, Tiago Bortolini, Katherine L. Bottenhorn, Alexander Bowring, Senne Braem, Hayley R. Brooks, Emily G. Brudner, Cristian B. Calderon, Julia A. Camilleri, Jaime J. Castrellon, Luca Cecchetti, Edna C. Cieslik, Zachary J. Cole, Olivier Collignon, Robert W. Cox, William A. Cunningham, Stefan Czoschke, Kamalaker

Dadi, Charles P. Davis, Alberto De Luca, Mauricio R. Delgado, Lysia Demetriou, Jeffrey B. Dennison, Xin Di, Erin W. Dickie, Ekaterina Dobryakova, Claire L. Donnat, Juergen Dukart, Niall W. Duncan, Joke Durnez, Amr Eed, Simon B. Eickhoff, Andrew Erhart, Laura Fontanesi, G. Matthew Fricke, Shiguang Fu, Adriana Galván, Remi Gau, Sarah Genon, Tristan Glatard, Enrico Glerean, Jelle J. Goeman, Sergej A. E. Golowin, Carlos González-García, Krzysztof J. Gorgolewski, Cheryl L. Grady, Mikella A. Green, João F. Guassi Moreira, Olivia Guest, Shabnam Hakimi, J. Paul Hamilton, Roeland Hancock, Giacomo Handjaras, Bronson B. Harry, Colin Hawco, Peer Herholz, Gabrielle Herman, Stephan Heunis, Felix Hoffstaedter, Jeremy Hogeveen, Susan Holmes, Chuan-Peng Hu, Scott A. Huettel, Matthew E. Hughes, Vittorio Iacovella, Alexandru D. Iordan, Peder M. Isager, Ayse I. Isik, Andrew Jahn, Matthew R. Johnson, Tom Johnstone, Michael J. E. Joseph, Anthony C. Juliano, Joseph W. Kable, Michalis Kassinopoulos, Cemal Koba, Xiang-Zhen Kong, Timothy R. Koscik, Nuri Erkut Kucukboyaci, Brice A. Kuhl, Sebastian Kupek, Angela R. Laird, Claus Lamm, Robert Langner, Nina Lauharatanahirun, Hongmi Lee, Sangil Lee, Alexander Leemans, Andrea Leo, Elise Lesage, Flora Li, Monica Y. C. Li, Phui Cheng Lim, Evan N. Lintz, Schuyler W. Liphardt, Annabel B. Losecaat Vermeer, Bradley C. Love, Michael L. Mack, Norberto Malpica, Theo Marins, Camille Maumet, Kelsey McDonald, Joseph T. McGuire, Helena Melero, Adriana S. Méndez Leal, Benjamin Meyer, Kristin N. Meyer, Glad Mihai, Georgios D. Mitsis, Jorge Moll, Dylan M. Nielson, Gustav Nilsonne, Michael P. Notter, Emanuele Olivetti, Adrian I. Onicas, Paolo Papale, Kaustubh R. Patil, Jonathan E. Peelle, Alexandre Pérez, Doris Pischedda, Jean-Baptiste Poline, Yanina Prystauka, Shruti Ray, Patricia A. Reuter-Lorenz, Richard C. Reynolds, Emiliano Ricciardi, Jenny R. Rieck, Anais M. Rodriguez-Thompson, Anthony Romyn, Taylor Salo, Gregory R. Samanez-Larkin, Emilio Sanz-Morales, Margaret L. Schlichting, Douglas H. Schultz, Qiang Shen, Margaret A. Sheridan, Jennifer A. Silvers, Kenny Skagerlund, Alec Smith, David V. Smith, Peter Sokol-Hessner, Simon R. Steinkamp, Sarah M. Tashjian, Bertrand Thirion, John N. Thorp, Gustav Tinghög, Loreen Tisdall, Steven H. Tompson, Claudio Toro-Serey, Juan Jesus Torre Tresols, Leonardo Tozzi, Vuong Truong, Luca Turella, Anna E. van 't Veer, Tom Verguts, Jean M. Vettel, Sagana Vijayarajah, Khoi Vo, Matthew B. Wall, Wouter D. Weeda, Susanne Weis, David J. White, David Wisniewski, Alba Xifra-Porxas, Emily A. Yearling, Sangsuk Yoon, Rui Yuan, Kenneth S. L. Yuen, Lei Zhang, Xu Zhang, Joshua E. Zosky, Thomas E. Nichols, Russell A. Poldrack, and Tom Schonberg. Variability in the analysis of a single neuroimaging dataset by many teams. 582(7810):84–88, 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2314-9. Publisher: Nature Publishing Group.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, D. Song, U Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. 2020.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794. Association for Computing Machinery, 2016. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785.

Zi-Jian Cheng, Zi-Yi Jia, Zhi Zhou, Yu-Feng Li, and Lan-Zhe Guo. TabFSBench: Tabular benchmark for feature shifts in open environments, 2025.

Evangelia Christodoulou, Jie Ma, Gary S. Collins, Ewout W. Steyerberg, Jan Y. Verbakel, and Ben Van Calster. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. 110:12–22. ISSN 0895-4356, 1878-5921. doi: 10.1016/j.jclinepi.2019.02.004. Publisher: Elsevier.

Joseph Paul Cohen, Tianshi Cao, Joseph D. Viviano, Chin-Wei Huang, Michael Fralick, Marzyeh Ghassemi, Muhammad Mamdani, Russell Greiner, and Yoshua Bengio. Problems in the deployment of machine-learned models in health care. 193(35):E1391–E1394. ISSN 0820-3946. doi: 10.1503/cmaj.202066.

Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. TabArena: A living benchmark for machine learning on tabular data, 2025.

Samuel G. Finlayson, John D. Bowers, Joichi Ito, Jonathan L. Zittrain, Andrew L. Beam, and Isaac S. Kohane. Adversarial attacks on medical machine learning. 363(6433):1287–1289, 2019.

doi: 10.1126/science.aaw4399. Publisher: American Association for the Advancement of Science.

Josh Gardner, Zoran Popovic, and Ludwig Schmidt. Benchmarking distribution shift in tabular data with TableShift, 2024.

Benjamin A Goldstein, Ann Marie Navar, Michael J Pencina, and John P A Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. 24(1):198–208, 2017. ISSN 1067-5027. doi: 10.1093/jamia/ocw042.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pp. 507–520. Curran Associates Inc., 2022. ISBN 978-1-71387-108-8.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Benjamin Jäger, Dominik Safaric, Simone Alessi, Adrian Hayler, Mihir Manium, Rosen Yu, Felix Jablonski, Shi Bin Hoo, Anurag Garg, Jake Robertson, Magnus Bühler, Vladyslav Moroshan, Lennart Purucker, Clara Cornu, Lilly Charlotte Wehrhahn, Alessandro Bonetto, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. TabPFN-2.5: Advancing the state of the art in tabular foundation models, 2025.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, Dominik Safaric, Jake Robertson, Benjamin Jäger, Simone Alessi, Adrian Hayler, Vladyslav Moroshan, Lennart Purucker, Philipp Singer, Alan Arazi, Julien Siems, Jan Hendrik Metzen, Georg Grab, Nick Erickson, Siyuan Guo, Eliott Kalfon, Simon Bing, David Salinas, Clara Cornu, Lilly Charlotte Wehrhahn, Diana Kriuchkova, Kursat Kaya, Lydia Sidhoum, Marie Salmon, Jerry Chen, Madelon Hulsebos, Yann LeCun, Samuel Müller, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. Tabpfn-3: Technical report, 2026. URL https://arxiv.org/abs/2605.13986.

Yu Gu, Jingjing Fu, Xiaodong Liu, Jeya Maria Jose Valanarasu, Noel CF Codella, Reuben Tan, Qianchu Liu, Ying Jin, Sheng Zhang, Jinyu Wang, Rui Wang, Lei Song, Guanghui Qin, Naoto Usuyama, Cliff Wong, Hao Cheng, Hohin Lee, Praneeth Sanapathi, Sarah Hilado, Jiang Bian, Javier Alvarez-Valle, Mu Wei, Khalil Malik, Jianfeng Gao, Eric Horvitz, Matthew P. Lungren, Hoifung Poon, and Paul Vozila. The illusion of readiness: Stress testing large frontier models on multimodal medical benchmarks, 2025.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. 637(8045):319–326, 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08328-6. Publisher: Nature Publishing Group.

David Holzmüller, Léo Grinsztajn, and Ingo Steinwart. Better by default: strong pre-tuned MLPs and boosted trees on tabular data. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37 of *NIPS '24*, pp. 26577–26658. Curran Associates Inc., 2025. ISBN 9798331314385.

Nianzong Hou, Mingzhe Li, Lu He, Bing Xie, Lin Wang, Rumin Zhang, Yong Yu, Xiaodong Sun, Zhengsheng Pan, and Kai Wang. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. 18(1):462, 2020. ISSN 1479-5876. doi: 10.1186/s12967-020-02620-5.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. 3(1):160035, 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35. Publisher: Nature Publishing Group.

Alistair E W Johnson, David J Stone, Leo A Celi, and Tom J Pollard. The mimic code repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25(1):32–39, 2018.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. 10(1):1, 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01899-x. Publisher: Nature Publishing Group.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Myung Jun Kim, Félix Lefebvre, Gaëtan Brison, Alexandre Perez-Lebel, and Gaël Varoquaux. Table foundation models: on knowledge pre-training for tabular learning, 2025. URL https://arxiv.org/abs/2505.14415.

J. R. Le Gall, P. Loirat, A. Alperovitch, P. Glaser, C. Granthil, D. Mathieu, P. Mercier, R. Thomas, and D. Villers. A simplified acute physiology score for ICU patients. 12(11):975–977, 1984. ISSN 0090-3493. doi: 10.1097/00003246-198411000-00012.

J. R. Le Gall, S. Lemeshow, and F. Saulnier. A new simplified acute physiology score (SAPS II) based on a european/north american multicenter study. 270(24):2957–2963, 1993. ISSN 0098-7484. doi: 10.1001/jama.270.24.2957.

Karim Lekadir, Alejandro F. Frangi, Antonio R. Porras, Ben Glocker, Celia Cintas, Curtis P. Langlotz, Eva Weicken, Folkert W. Asselbergs, Fred Prior, Gary S. Collins, Georgios Kaissis, Gianna Tsakou, Irène Buvat, Jayashree Kalpathy-Cramer, John Mongan, Julia A. Schnabel, Kaisar Kushibar, Katrine Riklund, Kostas Marias, Lameck M. Amugongo, Lauren A. Fromont, Lena Maier-Hein, Leonor Cerdá-Alberich, Luis Martí-Bonmatí, M. Jorge Cardoso, Maciej Bobowicz, Mahsa Shabani, Manolis Tsiknakis, Maria A. Zuluaga, Marie-Christine Fritzsche, Marina Camacho, Marius George Linguraru, Markus Wenzel, Marleen De Bruijne, Martin G. Tolsgaard, Melanie Goisauf, Mónica Cano Abadía, Nikolaos Papanikolaou, Noussair Lazrak, Oriol Pujol, Richard Osuala, Sandy Napel, Sara Colantonio, Smriti Joshi, Stefan Klein, Susanna Aussó, Wendy A. Rogers, Zohaib Salahuddin, and Martijn P. A. Starmans. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. 388: e081554, 2025. ISSN 1756-1833. doi: 10.1136/bmj-2024-081554. Publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting.

Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, Huai-Hong Yin, Tao Zhou, Jun-Peng Jiang, and Han-Jia Ye. Talent: A tabular analytics and learning toolbox. 26(226):1–16. ISSN 1533-7928.

Pedro Lopez-Ayala, Richard D Riley, Gary S Collins, and Tobias Zimmermann. Dealing with continuous variables and modelling non-linear associations in healthcare data: practical guide. 390: e082440, 2025. ISSN 1756-1833. doi: 10.1136/bmj-2024-082440. Publisher: BMJ.

Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Alex Labach, Hamidreza Kamkari, Jesse C. Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L. Caterini, and Maksims Volkovs. TabDPT: Scaling tabular foundation models on real data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=pIZxEOZCId.

Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C., Ganesh Ramakrishnan, Micah Goldblum, and Colin White. When do neural nets outperform boosted trees on tabular data? In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pp. 76336–76369. Curran Associates Inc., 2023.

Boris Muzellec, Julie Josse, Claire Boyer, and Marco Cuturi. Missing data imputation using optimal transport. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 7130–7140. PMLR, 2020. ISSN: 2640-3498.

Bret Nestor, Matthew BA McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. pp. 381–405. PMLR, 2019. ISBN 2640-3498.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pp. 625–632. Association for Computing Machinery. ISBN 978-1-59593-180-1. doi: 10.1145/1102351.1102430.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. 366(6464):447–453, 2019. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aax2342. Publisher: American Association for the Advancement of Science Section: Research Article.

Alexandre Perez-Lebel, Gaël Varoquaux, Marine Le Morvan, Julie Josse, and Jean-Baptiste Poline. Benchmarking missing-values approaches for predictive models on health databases. 11:giac013, 2022. ISSN 2047-217X. doi: 10.1093/gigascience/giac013.

Tom Pollard, Alistair Johnson, Jesse Raffa, Leo Anthony Celi, Omar Badawi, and Roger Mark. eICU Collaborative Research Database. *PhysioNet*, April 2019. doi: 10.13026/C2WM1R. URL https://doi.org/10.13026/C2WM1R. Version 2.0.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Lennart Purucker, Andrej Tschalzev, Nick Erickson, Gioia Blayer, David Holzmüller, Alan Arazi, Alexander Pfefferle, Mustafa Tajjar, Gaël Varoquaux, and Frank Hutter. Beyond iid: How general are tabular foundation models, really?, 2026. URL https://arxiv.org/abs/2606.30410.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. Tabiclv2: A better, faster, scalable, and open tabular foundation model, 2026. URL https://arxiv.org/abs/2602.11139.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. 1(5):206–215, 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x. Publisher: Nature Publishing Group.

Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, Søren Brunak, Laust Hvas Mortensen, Ewan Birney, Tom Fitzgerald, and Moritz Gerstung. Learning the natural history of human disease with generative transformers. pp. 1–9, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09529-3. Publisher: Nature Publishing Group.

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need.

Ewout W. Steyerberg, Andrew J. Vickers, Nancy R. Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J. Pencina, and Michael W. Kattan. Assessing the performance of prediction models: a framework for some traditional and novel measures. 21(1):128–138, 2010. ISSN 1044-3983. doi: 10.1097/EDE.0b013e3181c30fb2.

Ben Van Calster, David J. McLernon, Maarten van Smeden, Laure Wynants, Ewout W. Steyerberg, Patrick Bossuyt, Gary S. Collins, Petra Macaskill, David J. McLernon, Karel G. M. Moons, Ewout W. Steyerberg, Ben Van Calster, Maarten van Smeden, Andrew J. Vickers, and On behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the achilles heel of predictive analytics. 17(1):230. ISSN 1741-7015. doi: 10.1186/s12916-019-1466-7.

Cornelius Wolff and Madelon Hulsebos. How well do LLMs reason over tabular data, really?, 2025.

An-Kwok Ian Wong, Marie Charpignon, Han Kim, Christopher Josef, Anne A. H. de Hond, Jhalique Jane Fojas, Azade Tabaie, Xiaoli Liu, Eduardo Mireles-Cabodevila, Leandro Carvalho, Rishikesan Kamaleswaran, R. W. M. A. Madushani, Lasith Adhikari, Andre L. Holder, Ewout W. Steyerberg, Timothy G. Buchman, Mary E. Lough, and Leo Anthony Celi. Analysis of discrepancies between pulse oximetry and arterial oxygen saturation measurements by race and ethnicity and association with organ dysfunction and mortality. 4(11):e2131674, 2021. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2021.31674.

Andrew Wong, Erkin Otles, John P. Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, Muhammad Ghous, and Karandeep Singh. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. 181(8):1065–1070. ISSN 2168-6106. doi: 10.1001/jamainternmed.2021.2626.

A. Yarkın Yıldız and Asli Kalayci. Gradient boosting decision trees on medical diagnosis over tabular data. In *2025 IEEE International Conference on AI and Data Analytics (ICAD)*, pp. 1–8, 2025. doi: 10.1109/ICAD65464.2025.11114069.

Yunlong Zhang, Yuxuan Sun, Honglin Li, Sunyi Zheng, Chenglu Zhu, and Lin Yang. Benchmarking the robustness of deep neural networks to common corruptions in digital pathology. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pp. 242–252. Springer Nature Switzerland, 2022. ISBN 978-3-031-16434-7. doi: 10.1007/978-3-031-16434-7_24.

Chloe Qinyu Zhu, Muhang Tian, Lesia Semenova, Jiachang Liu, Jack Xu, Joseph Scarpa, and Cynthia Rudin. eicu$_e xtraction$. *arXiv preprint arXiv:2311.13015*, 2023.

A SOFTWARE AND HARDWARE SPECIFICATION

Logistic regression, Linear regression, LASSO, Ridge regression, Random Forest and default Gradient Boosting were implemented using the scikit-learn library (version 1.8.0). The remaining models were implemented using their respective libraries: xgboost (version 3.2.0), lightgbm (version 4.6.0), catboost (version 1.2.10), tabpfn (version 8.0.8), and tabicl (version 2.1.1), tabdpt (version 1.2.0).

All methods were run for hyperparameter optimization, training, and inference on a system with an Intel® Core™ Ultra 9 285K CPU (3.70 GHz) and an NVIDIA GeForce RTX 4090 GPU and 64GB of RAM.

B HYPERPARAMETER OPTIMIZATION

Table 1: Hyperparameter search spaces considered for hyperparameter optimization for **classification**. Adapted from (Erickson et al., 2025). The number of trees for each boosting methods is set to 1000.

Model	Hyperparameter	Values
LASSO	C	[0.01, 0.1, 1, 10, 100, 1000]
Ridge	C	[0.01, 0.1, 1, 10, 100, 1000]
Random Forest	n_estimators	[100, 200, 400]
	max_features	[0.4, 0.6, 0.8, 1]
	max_samples	[0.5, 0.8, 1]
	min_samples_split	[2, 3, 4]
	bootstrap	[True, False]
	min_impurity_decrease	[0.00001, 0.0001, 0.001]
Gradient Boosting	loss	['log_loss', 'exponential']
	learning_rate	[0.001, 0.005, 0.01, 0.05, 0.1]
	subsample	[0.5, 0.8, 1]
	criterion	['friedman_mse', 'squared_error']
	max_depth	[None, 2, 3, 4, 5]
	min_samples_split	[2, 3]
	min_samples_leaf	[1, 3, 8, 20, 50]
	min_impurity_decrease	[0, 0.01, 0.02, 0.05]
	max_leaf_nodes	[None, 5, 10, 15]
	validation_fraction	[0.1]
	n_iter_no_change	[50]
XGBoost	learning_rate	[0.005, 0.014, 0.037, 0.1]
	max_depth	[4, 6, 8, 10]
	min_child_weight	[0.001, 0.017, 0.292, 5.0]
	subsample	[0.6, 0.8, 1]
	colsample_bylevel	[0.6, 0.8, 1]
	colsample_bynode	[0.6, 0.8, 1]
	reg_lambda	[0.0001, 0.001, 0.1, 1, 5]
	reg_alpha	[0.0001, 0.001, 0.1, 1, 5]
	grow_policy	['depthwise', 'lossguide']
	max_cat_to_onehot	[8, 16, 32, 64, 100]
	max_leaves	[8, 32, 64, 256, 1014]
	early_stopping_rounds	[50]
LightGBM	learning_rate	[0.005, 0.014, 0.037, 0.1]
	colsample_bytree	[0.4, 0.6, 0.8, 1]
	subsample	[0.7, 0.8, 0.9, 1]
	subsample_freq	[1]
	num_leaves	[2, 25, 50, 100, 200]
	min_child_samples	[1, 4, 16, 64]

Table 1 continued from previous page

Model	Hyperparameter	Values
	extra_tree	[False, True]
	min_data_per_group	[2, 25, 50, 100]
	cat_l2	[0.005, 0.037, 0.271, 2.0]
	cat_smooth	[0.001, 0.046, 2.154, 100.0]
	max_cat_to_onehot	[8, 15, 50, 100]
	reg_alpha	[0.0001, 0.001, 0.01, 0.1, 1]
	reg_lambda	[0.0001, 0.001, 0.01, 0.1, 2]
	metric	['auc']
CatBoost	learning_rate	[0.005, 0.014, 0.037, 0.1]
	bootstrap_type	['Bernoulli']
	subsample	[0.7, 0.8, 0.9, 1]
	grow_policy	['SymmetricTree', 'Depthwise']
	depth	[4, 6, 8]
	colsample_bylevel	[0.85, 0.90, 0.95, 1]
	l2_leaf_reg	[0.0001, 0.004, 0.136, 5.0]
	leaf_estimation_iterations	[1, 5, 10, 20]
	one_hot_max_size	[8, 16, 32, 100]
	model_size_reg	[0.1, 0.247, 0.608, 1.5]
	max_ctr_complexity	[2, 3, 4, 5]
	boosting_type	['Plain']
	max_bin	[254]
	od_wait	[50]
	od_type	['Iter']
	eval_metric	['AUC']
RealMLP	num_emb_type	["none", "pbld", "plr"]
	add_front_scale	[True, False]
	lr	[0.02, 0.1, 0.3]
	p_drop	[0.0, 0.15, 0.3]
	act	["relu", "selu", "mish"]
	hidden_sizes	[[64, 64, 64, 64, 64], [512]]
	wd	[0.0, 2e-2]
	plr_sigma	[0.05, 0.1, 0.5]
	ls_eps	[0.0, 0.1]
	use_early_stopping	[True]

Table 2: Hyperparameter search spaces for regression hyperparameter optimization. Adapted from (Kim et al., 2025; Grinsztajn et al., 2022; Holzmüller et al., 2025). The number of trees for each boosting methods is set to 1000.

Model	Hyperparameter	Values
LASSO	alpha	[0.01, 0.1, 1, 10, 100]
Ridge	alpha	[0.01, 0.1, 1, 10, 100]
Random Forest	n_estimators	[50, 100, 150, 200, 250]
	max_depth	[None, 2, 3, 4]
	max_features	['sqrt', 'log2', None, 0.2, 0.4, 0.6, 0.8]
	min_samples_leaf	[1, 3, 8, 20, 50]
	bootstrap	[True, False]
	min_impurity_decrease	[0, 0.01, 0.02, 0.05]
Gradient Boosting	learning_rate	[0.00001, 0.0001, 0.001, 0.01, 0.1, 1.0]
	subsample	[0.5, 0.8, 1]
	max_depth	[None, 2, 3, 4, 5]

Table 2 continued from previous page

Model	Hyperparameter	Values
	min_samples_split	[2, 3]
	min_samples_leaf	[1, 3, 8, 20, 50]
	min_impurity_decrease	[0, 0.01, 0.02, 0.05]
	max_leaf_nodes	[None, 5, 10, 15]
	validation_fraction	[0.1]
	n_iter_no_change	[50]
XGBoost	max_depth	[2, 4, 6, 8, 10]
	learning_rate	[0.00001, 0.0001, 0.001, 0.01, 0.1, 1.0]
	min_child_weight	[1.0, 2.5, 6.3, 15.8, 39.8, 100]
	subsample	[0.5, 0.8, 1]
	colsample_bylevel	[0.5, 0.8, 1]
	colsample_bytree	[0.5, 0.8, 1]
	gamma	[0.0, 3e-5, 0.002, 0.11915, 7.0]
	lambda	[1.0, 1.587, 2.52, 4.0]
	alpha	[0.0, 3.2e-06, 0.001, 0.3162, 100]
	early_stopping_rounds	[50]
	tree_method	['hist']
LightGBM	bagging_freq	[1]
	num_leaves	[2, 10, 100, 1000, 10000]
	learning_rate	[0.00001, 0.0001, 0.001, 0.01, 0.1, 1.0]
	subsample	[0.5, 0.8, 1]
	feature_fraction	[0.5, 0.8, 1]
	min_data_in_leaf	[100, 398, 1585, 6310, 25119, 100000]
	min_sum_hessian_in_leaf	[1e-7, 1e-4, 0.1, 100]
	lambda_l1	[0, 1e-16, 1e-10, 0.0001, 100.0]
	lambda_l2	[0, 1e-16, 1e-10, 0.0001, 100.0]
	verbose	[-1]
	objective	['regression']
	metric	['rmse']
CatBoost	depth	[2, 4, 6]
	learning_rate	[0.00001, 0.0001, 0.001, 0.01, 0.1, 1.0]
	bagging_temperature	[0, 0.25, 0.5, 0.75, 1]
	l2_leaf_reg	[1.0, 2.2, 4.6, 10.0]
	one_hot_max_size	[0, 5, 15, 25]
	random_strength	[1, 10, 20]
	leaf_estimation_iterations	[1, 10, 20]
	od_wait	[50]
	od_type	['Iter']
	eval_metric	['RMSE']
RealMLP	num_emb_type	['none', 'pbld', 'plr']
	add_front_scale	[True, False]
	lr	[0.02, 0.1, 0.3]
	p_drop	[0.0, 0.15, 0.3]
	act	['relu', 'selu', 'mish']
	hidden_sizes	[[64, 64, 64, 64, 64], [512]]
	wd	[0.0, 2e-2]
	plr_sigma	[0.05, 0.1, 0.5]
	use_ls	[False]
	use_early_stopping	[True]

C DATASETS

C.1 DATA PREPROCESSING

Raw data for all three datasets were extracted in Google BigQuery using queries taken and adapted from (Johnson et al., 2018; Zhu et al., 2023).

For each of the three datasets, we included patients aged 16 years or older at the time of their first ICU admission. For these patients, we retained only their first ICU stay.

The predictor set was based on the variables used in the Simplified Acute Physiology Score II (SAPS II), which is a clinical score developed to provide an estimate of in-hospital mortality for patients admitted to the ICU (Le Gall et al., 1993). However, instead of applying the original SAPS II binning scheme and measurement thresholds, we extracted the corresponding raw physiological and laboratory measurements to preserve the granularity of the data for modern machine learning methods. More specifically, for variables for which both unusually low and unusually high measurements may be prognostic, we extracted the minimum and maximum values observed during the first 24 hours as separate predictors. A list of all included predictors are presented in Table 3.

We excluded patients with more than 50% missing values across the selected predictors.

We additionally removed observations containing negative values for continuous predictors.

Because urine output contained extremely high and implausible values, we removed patients whose values exceeded a plausible value that was defined based on the observed data.

To conduct some of the analyses presented in section 2.5, we additionally extracted variables such as gender and the specific ICU service to which the patient was admitted. For the latter, we performed analysis only on ICU services represented in all three datasets.

To evaluate robustness to label noise, when available, we also defined two proxies for hospital mortality: the first restricted to deaths occurring during the specific ICU stay. The second extended to include post-discharge mortality, ranging from 90 days up to four years after hospital discharge, depending on the patient.

For MIMIC-IV, we additionally extracted the admission-year group available in the database. For eICU patients, we retrieved the geographic region of the hospital in which the patient was admitted.

For MIMIC-III and MIMIC-IV, PaO_2/FiO_2 was defined only for patients receiving mechanical ventilation or continuous positive airway pressure; values were otherwise encoded as 0.

Finally, categorical variables were encoded using ordinal encoding prior to model fitting.

C.2 DATA DESCRIPTION

D DESCRIPTION OF DATA PERTURBATIONS

Notation Let $\mathcal{D} = \{(\mathbf{x}_i, y_i, z_i)\}_{i=1}^{N}$ denote the full dataset. Each observation consists of a feature vector $\mathbf{x}_i \in \mathbb{R}^{d_c} \times \mathcal{C}^{d_k}$ (containing d features with d_c continuous features and d_k categorical features), the total number of predictors with d ($d = d_c + d_k$) and a binary outcome $y_i \in \{0, 1\}$ (hospital mortality) and a continuous outcome $z_i \in \mathbb{R}^+$ (iICU length of stay). We partition $\mathcal{D}$ into disjoint training and validation sets, denoted as $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{val}}$ respectively.

To evaluate robustness under different scenarios, we define a target subset $\mathcal{S} \subseteq \mathcal{D}$ to which the perturbations are applied. Depending on the experimental setup, $\mathcal{S}$ may represent $\mathcal{D}_{\text{train}}$ (perturbations are added in the training data only), $\mathcal{D}_{\text{val}}$ (perturbations are applied in the validation data only), or $\mathcal{D}$ (perturbations are added in the full dataset). The perturbed dataset is denoted as $\tilde{\mathcal{D}}$, where samples inside $\mathcal{S}$ are transformed into $(\tilde{\mathbf{x}}_i, \tilde{y}_i, \tilde{z}_i)$ according to the mechanisms defined below, while samples outside $\mathcal{S}$ remain unchanged. Unless it is specified, we consider all three perturbation scenarios.

We ensured exactly the same perturbation was applied across models within each comparison.

Table 3: MIMIC-III, MIMIC-IV, and eICU descriptive table. Continuous variables are reported as median [IQR] with missingness percentages where applicable; categorical variables are reported as count (%). An asterisk indicates that the variable was used as a predictor.

Characteristic	MIMIC-III	MIMIC-IV	eICU-CRD
n	36,256	61,741	86,728
Outcomes			
Hospital mortality (n, %)	4,031 (11.1%)	6,405 (10.4%)	7,453 (8.6%)
ICU length of stay (days)	2.1 [1.2, 4.1]	1.9 [1.1, 3.7]	1.8 [1.0, 3.2]
Proxy target			
Death in ICU	2,807 (7.7%)	4,247 (6.9%)	4,693 (5.4%)
Overall death	14,602 (40.3%)	20,388 (33.0%)	—
Age* (years)	66.0 [53.0, 78.0]	67.0 [55.0, 78.0]	65.0 [52.0, 76.0]
Max. Heart rate*	101.0 [89.0, 115.0]	100.0 [88.0, 114.0]	95.0 [82.0, 110.0]
Min. Heart rate*	70.0 [60.0, 80.0]	68.0 [60.0, 79.0]	73.0 [63.0, 85.0]
Max. Systolic BP* (mmHg)	147.0 [134.0, 163.0]	146.0 [132.0, 161.0]	137.0 [122.0, 155.0] (14% miss)
Min. Systolic BP* (mmHg)	92.0 [82.0, 103.0]	92.0 [83.0, 103.0]	105.0 [92.0, 120.0] (14% miss)
Max. Temperature* (°C)	37.4 [37.0, 37.9] (2% miss)	37.2 [36.9, 37.6] (3% miss)	37.2 [36.9, 37.6] (2% miss)
Min. Temperature* (°C)	36.2 [35.7, 36.6] (2% miss)	36.4 [36.2, 36.7] (3% miss)	36.4 [36.1, 36.7] (2% miss)
Urine output* (mL/day)	1700.0 [1060.0, 2515.0]	1490.0 [910.0, 2240.0]	1350.0 [750.0, 2190.0]
Min. Blood urea nitrogen* (mmol/L)	16.0 [11.0, 25.0] (1% miss)	16.0 [11.0, 26.0] (3% miss)	16.0 [11.0, 25.0] (2% miss)
Max. Blood urea nitrogen* (mmol/L)	19.0 [14.0, 30.0] (1% miss)	18.0 [13.0, 29.0] (3% miss)	20.0 [14.0, 32.0] (2% miss)
Min. WBC*	9.5 [7.0, 12.7] (1% miss)	9.7 [7.1, 13.0] (3% miss)	8.9 [6.6, 12.2] (3% miss)
Max. WBC*	12.4 [9.1, 16.6] (1% miss)	11.7 [8.4, 16.1] (3% miss)	11.9 [8.7, 16.4] (3% miss)
Min. Potassium* (mEq/L)	3.7 [3.4, 4.0]	3.9 [3.6, 4.3] (3% miss)	3.8 [3.4, 4.1] (2% miss)
Max. Potassium* (mEq/L)	4.5 [4.1, 5.1]	4.3 [3.9, 4.7] (3% miss)	4.3 [4.0, 4.8] (2% miss)
Min. Sodium*	137.0 [134.0, 139.0] (1% miss)	138.0 [135.0, 140.0] (3% miss)	137.0 [134.0, 139.0] (2% miss)
Max. Sodium*	140.0 [138.0, 142.0] (1% miss)	139.0 [137.0, 142.0] (3% miss)	140.0 [137.0, 142.0] (2% miss)
Min. Bicarbonate* (mEq/L)	23.0 [20.0, 25.0] (1% miss)	22.0 [20.0, 25.0] (3% miss)	23.0 [20.0, 25.0] (6% miss)
Max. Bicarbonate* (mEq/L)	25.0 [23.0, 27.0] (1% miss)	24.0 [22.0, 26.0] (3% miss)	25.0 [23.0, 28.0] (6% miss)
Min. Glasgow Coma Scale*	15.0 [14.0, 15.0]	15.0 [14.0, 15.0]	15.0 [11.0, 15.0] (20% miss)
Min. PaO_2/FiO_2* (mm Hg/%)	0.0 [0.0, 184.0]	0.0 [0.0, 114.0]	192.1 [118.0, 288.3] (76% miss)
Bilirubin (min)*	0.6 [0.4, 1.2] (57% miss)	0.6 [0.4, 1.3] (58% miss)	0.5 [0.4, 0.9] (34% miss)
Bilirubin (max)*	0.7 [0.4, 1.4] (57% miss)	0.7 [0.4, 1.4] (58% miss)	0.6 [0.4, 1.0] (34% miss)
AIDS*	317 (0.9%)	295 (0.5%)	84 (0.1%)
Hematologic malignancy*	963 (2.7%)	1,839 (3.0%)	1,416 (1.6%)
Metastatic cancer*	2,273 (6.3%)	3,927 (6.4%)	1,663 (1.9%)
Admission type*			
Medical	25,182 (69.5%)	41,923 (67.9%)	69,378 (80.0%)
Unscheduled surgical	6,274 (17.3%)	17,918 (29.0%)	2,099 (2.4%)
Scheduled surgical	4,800 (13.2%)	1,900 (3.1%)	15,251 (17.6%)
ICU unit			
MICU (Medical ICU)	12,018 (33.1%)	11,569 (18.7%)	7,044 (8.1%)
SICU (Surgical ICU)	5,659 (15.6%)	8,722 (14.1%)	5,887 (6.8%)
gender: M	20,486 (56.5%)	34,810 (56.4%)	48,182 (55.6%)
Admission year			
2008–2010	—	16,785 (27.2%)	—
2011–2013	—	12,894 (20.9%)	—
2014–2016	—	12,521 (20.3%)	—
2017–2019	—	11,429 (18.5%)	—
2020–2022	—	8,112 (13.1%)	—
Region			
Midwest	—	—	27,238 (31.4%)
Northeast	—	—	8,014 (9.2%)
South	—	—	34,469 (39.7%)
West	—	—	13,226 (15.2%)
Missing	—	—	3,781 (4.4%)

D.1 Label Noise

In this section, perturbations are only applied to $\mathcal{D}_{\text{train}}$. For all samples $(\mathbf{x}_i, y_i, z_i) \in \mathcal{S}$, we apply the four following label perturbation mechanisms.

Random Label Noise Given a noise level $\varepsilon \in [0, 0.5]$, flip each binary label in the target subset independently. Namely, let ξ_i sampled i.i.d. in a Bernoulli of parameter ε, then $\tilde{y}_i = y_i \oplus \xi_i$, where $\oplus$ denotes the XOR operation.

Since ICU length of stay only takes positive values, each continuous label is perturbed using a chi-square type noise i.e. $\tilde{z}_i = z_i + \varepsilon.\widehat{\sigma}.n_i^2$ where $\{n_i\}_i$ are sampled from a standard normal distribution and $\widehat{\sigma}$ is the empirical standard deviation of $\{z_i\}_i$.

Class-Conditional Targeted Label Noise (Classification only) Given a noise level $\varepsilon \in [0, 0.9]$, each label in the target subset is flipped independently:

Type $0 \to 1$**:** Only negative labels are flipped with probability ε. Let ξ_i sampled i.i.d. in a Bernoulli of parameter ε, then

$$\tilde{y}_i = \begin{cases} y_i \oplus \xi_i & \text{if } y_i = 0 \\ y_i & \text{if } y_i = 1 \end{cases}$$

Type $1 \to 0$**:** Only positive labels are flipped with probability ε. Let ξ_i sampled i.i.d. in a Bernoulli of parameter ε, then

$$\tilde{y}_i = \begin{cases} y_i & \text{if } y_i = 0 \\ y_i \oplus \xi_i & \text{if } y_i = 1 \end{cases}$$

Covariate-Dependent Label Noise The flip probability depends on a covariate. Let $r_i = n^{-1} \cdot \text{rank}(x_i^{(\text{age})}) \in [0, 1]$ denote the percentile rank of patient i's age relative to all patients in $\mathcal{D}_{\text{train}}$ and $\varepsilon \in [0, 1]$ a noise level.

Each binary label in $\mathcal{D}_{\text{train}}$ is flipped independently conditionally on its age percentile rank i.e. let ξ_i sampled i.i.d. in a Bernoulli of parameter $r_i\varepsilon$:

$$\tilde{y}_i = y_i \oplus \xi_i$$

Each continuous label is perturbed using a chi-square type noise i.e. $\tilde{z}_i = z_i + 0.5 r_i.\varepsilon.\widehat{\sigma}.n_i^2$ where $\{n_i\}_i$ are sampled from a standard normal distribution and $\widehat{\sigma}$ is the empirical standard deviation of $\{z_i\}_i$. Namely, as the noise level ε increases, older patients have higher probability of label corruption.

Proxy Label Noise (Classification only) The original label is replaced by a proxy $\tilde{y}_i = y_i^{(\text{proxy})}$ with proxy $\in$ {ICU death, overall death} . Unlike stochastic noise, proxy labels introduce systematic bias: using only ICU death underestimates hospital mortality (missing post-ICU discharge deaths), while overall death obtained from follow-up overestimates it (includes death obtained during available follow-up collected up to 90 days to 4 years post-hospital discharge, depending on the patient).

D.2 Input Noise

For all samples $(\mathbf{x}_i, y_i, z_i) \in \mathcal{S}$, input features are perturbed as follows.

Continuous Feature Noise For each continuous feature $j \in \{1, \ldots, d_c\}$, additive Gaussian noise is applied:

$$\tilde{x}_i^{(j)} = x_i^{(j)} + \eta_{ij}, \quad \eta_{ij} \sim \mathcal{N}\left(0, (\varepsilon \cdot 2\sigma_j)^2\right)$$

where σ_j is the empirical standard deviation of feature j within $\mathcal{S}$.

Categorical Feature Noise For each categorical feature $j \in \{1, \ldots, d_k\}$ with categories $\mathcal{K}_j$ and empirical distribution p_j computed on $\mathcal{S}$:

$$\tilde{x}_i^{(j)} = \begin{cases} k' \sim p_j & \text{with probability } \varepsilon \\ x_i^{(j)} & \text{with probability } 1 - \varepsilon \end{cases}$$

Replacement categories are sampled from the marginal distribution, preserving overall category frequencies.

D.3 Missing Data Mechanisms

Let $\mathbf{M} \in \{0, 1\}^{|\mathcal{S}| \times d}$ be the missingness indicator matrix for the target subset, where $M_{ij} = 1$ indicates that $x_i^{(j)}$ is missing. The observed data becomes:

$$\tilde{x}_i^{(j)} = \begin{cases} x_i^{(j)} & \text{if } M_{ij} = 0 \\ \text{NA} & \text{if } M_{ij} = 1 \end{cases}$$

Missing Completely at Random (MCAR) Given a target level of missingness fraction $\varepsilon \in [0, 0.9]$, missingness is generated independently of the observed or unobserved value:

$$M_{ij} \sim \text{Bernoulli}(\varepsilon), \quad \text{i.i.d.}$$

Missing at Random (MAR) Following Muzellec et al. (2020), a subset of features $\mathcal{O} \subset \{1, \ldots, d\}$ is always fully observed. In our case, a randomly selected 50% of features are always observed. For features $j \notin \mathcal{O}$, missingness depends on the observed features via a logistic model:

$$P(M_{ij} = 1 \mid \mathbf{x}_i) = \sigma\left(\mathbf{w}_j^\top \mathbf{x}_i^{(\mathcal{O})} + b_j\right)$$

where $\sigma(\cdot)$ is the sigmoid function. The weights $\mathbf{w}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ are rescaled to ensure unit variance of the linear predictor, and intercepts b_j are calibrated to achieve the target missingness rate $\varepsilon \in [0, 0.9]$ (in the remaining 50 %).

Missing Not at Random (MNAR) Missingness depends on the feature's own (potentially unobserved) value through a self-masking logistic model:

$$P(M_{ij} = 1 \mid x_i^{(j)}) = \sigma\left(w_j \cdot x_i^{(j)} + b_j\right)$$

where $w_j \sim \mathcal{N}(0, 1)$ (rescaled for unit variance) and b_j is calibrated to achieve the target missingness rate $\varepsilon \in [0, 0.9]$.

D.4 Feature Selection

Given a shuffle proportion $\varepsilon \in [0, 1]$, a random subset $\mathcal{O} \subset \{1, \ldots, d\}$ with $|\mathcal{O}| = \lfloor \varepsilon \cdot d \rfloor$ is selected uniformly at random. For each feature $j \in \mathcal{O}$, a random permutation π_j over the indices of $\mathcal{S}$ is applied:

$$\tilde{x}_i^{(j)} = x_{\pi_j(i)}^{(j)}, \quad \forall i \in \mathcal{S}$$

This breaks the feature-outcome association while preserving marginal distributions.

D.5 Imbalanced Data

Class-imbalance was only considered for the classification task. Because in-hospital mortality constituted the minority class, we randomly subsampled the positive class in the outer training set while retaining all negative observations. Given $\varepsilon \in [1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1]$ The positive class was retained at fraction ε of its original size.

D.6 Training-Set Size

To assess sensitivity to the amount of available development data, the training set (or training sample used as context for tabular foundation models) was randomly subsampled. Given $\varepsilon \in [0.05, 0.10, 0.25, 0.50, 0.80, 1.00]$ the outer training-fold was subsampled ε fraction of its observations.

D.7 Composite perturbation

Three perturbation families f_1, f_2, f_3 are drawn at random without replacement from the set given by

$$\mathcal{T} = \{\text{label noise}, \text{input noise}, \text{missing data}, \text{imbalanced data}, \text{training data}, \text{feature selection}\}$$

For each perturbation type f_i one scenario s_i (e.g., $1 \to 0$) and one perturbation level ε_i are drawn uniformly at random. We did not consider imbalanced data for regression and we removed proxy label from the scenarios as one of the proxy target was not defined for eICU. Each triplet $(f_i, s_i, \varepsilon_i)$ describes a single well-defined perturbation p_i. The perturbations $\{p_1, p_2, p_3\}$ are then applied sequentially. In our experiment, we set the number of composites to 50.

E ROBUSTNESS SCORE

Following Hendrycks & Dietterich (2019); Zhang et al. (2022), we define an aggregate robustness score that summarizes model performance across heterogeneous perturbation scenarios. We use a hierarchical aggregation scheme that first averages performance across perturbation levels or categories within each scenario, then normalizes performance relative to a baseline model, averages the resulting relative scores across perturbation scenarios within each family, and finally aggregates across perturbation families using a geometric mean.

Let $\mathcal{F} = \{f_1, \ldots, f_F\}$ denote the set of perturbation families, comprising missing data, input noise, label noise, feature selection, subgroup analysis, class imbalance, training-set size, and domain shift. Domain-shift evaluations on different external datasets are treated as distinct families. For example, when MIMIC-III is used as the development dataset, external evaluation on MIMIC-IV and eICU constitutes two separate family-level components rather than being combined into a single domain-shift score. This ensures that each external dataset contributes equally to the final aggregate robustness score.

For each family $f \in \mathcal{F}$, let $\mathcal{S}_f = \{s_1, s_2, \ldots, s_{K_f}\}$, denote the set of perturbation scenarios belonging to that family. For example, within the input-noise family, separate scenarios include continuous-feature noise applied to the training data, categorical-feature noise applied to the validation data, etc.

For each scenario s and model m, let $\mathcal{L}_s$ denote the set of perturbation levels or categories associated with that scenario. We first compute the mean AUC across these levels or categories:

$$\bar{P}_{m,s} = \frac{1}{|\mathcal{L}_s|} \sum_{\ell \in \mathcal{L}_s} P_{m,s,\ell}, \tag{1}$$

where $P_{m,s,\ell}$ denotes the AUC of model m in scenario s at perturbation level or category ℓ.

We next normalize the performance of each model relative to a baseline model b. For classification, we use logistic regression as the baseline model and AUC as the performance metric. Because AUC is bounded above by 1, we define relative performance in terms of residual error:

$$R_{m,s} = \frac{1 - \bar{P}_{b,s}}{1 - \bar{P}_{m,s}}. \tag{2}$$

We then compute a family-level robustness score by averaging the relative performance scores across the unique perturbation scenarios belonging to family f:

$$\mathrm{FS}_{m,f} = \frac{1}{|\mathcal{S}_f|} \sum_{s \in \mathcal{S}_f} R_{m,s}. \tag{3}$$

Finally, we obtain the Mean Robustness Score (mRS) for model m by aggregating the family-level scores using their geometric mean:

$$\mathrm{mRS}_m = \left(\prod_{f \in \mathcal{F}} \mathrm{FS}_{m,f} \right)^{1/F}. \tag{4}$$

F ELO RATING FOR COMPOSITE PERTURBATIONS

To complement the aggregate robustness score, we evaluate models under composite perturbations using an Elo-style rating (Erickson et al., 2025). For each run, development dataset, composite perturbation, and cross-validation fold, we compare every pair of models on the same data, yielding paired win, loss, or tie outcomes. Small differences between models are treated as ties when they are not statistically significant across the folds of the corresponding composite (using a paired t-test). Model failures are treated as decisive losses against models that successfully complete the same evaluation.

We then fit a Bradley–Terry model to all pairwise outcomes to estimate a latent strength θ_m for each model:

$$P(i \succ j) = \frac{1}{1 + \exp[-(\theta_i - \theta_j)]}. \tag{5}$$

The fitted strengths are mapped to the conventional Elo scale,

$$R_m = \frac{400}{\ln 10}\theta_m + R_0, \tag{6}$$

with Logistic and Linear Regression anchored to 1000 Elo, for classification and regression, respectively. For the main analysis, we report the Elo rating obtained by pooling all composite-perturbation duels with uniform weight. This score measures relative model performance over the empirical distribution of composite perturbations in the benchmark, providing a pairwise complement to the family-balanced aggregate robustness score. We report two complementary Elo views.

Clean-performance Elo. The first rating is fitted using only duels from the unperturbed baseline evaluation. It summarizes relative model performance under standard, unperturbed conditions and provides a reference for interpreting the robustness rating.

Global robustness Elo. The second rating is fitted using all duels from the multi-perturbation benchmark. All perturbed-composite contexts are pooled into a single duel set and assigned uniform weight. The resulting rating therefore reflects a model's ability to outperform other models across the full distribution of composite perturbations used in the benchmark.

Confidence intervals for both views were obtained using 5,000 bootstrap replicates.

G ADDITIONAL FIGURES

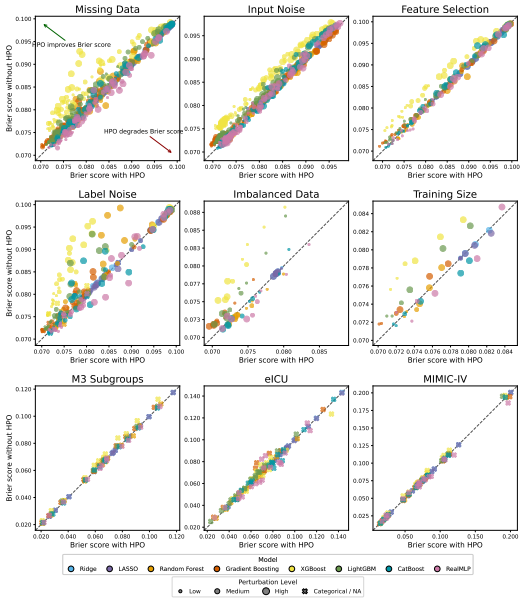

Figure 4: Classification - Effect of HPO across perturbation families when **MIMIC-III** is used as the development set. Each point represents the mean **Brier score** for a scenario at a specific noise level or category.

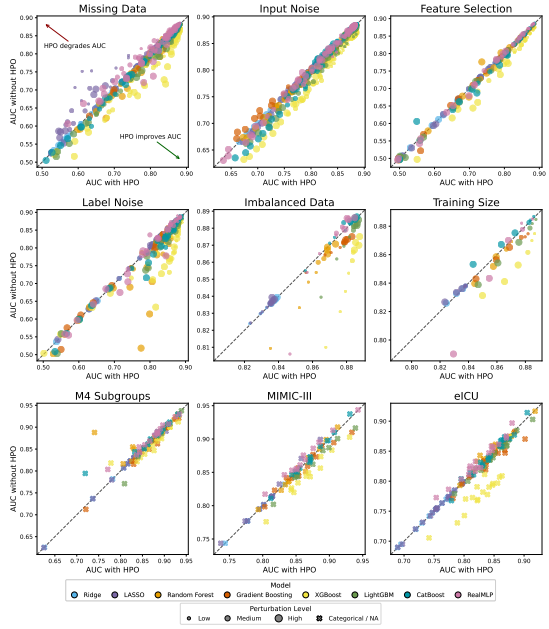

Figure 5: Classification - Effect of HPO across perturbation families when **MIMIC-IV** is used as the development set. Each point represents the mean **AUC** for a scenario at a specific noise level or category.

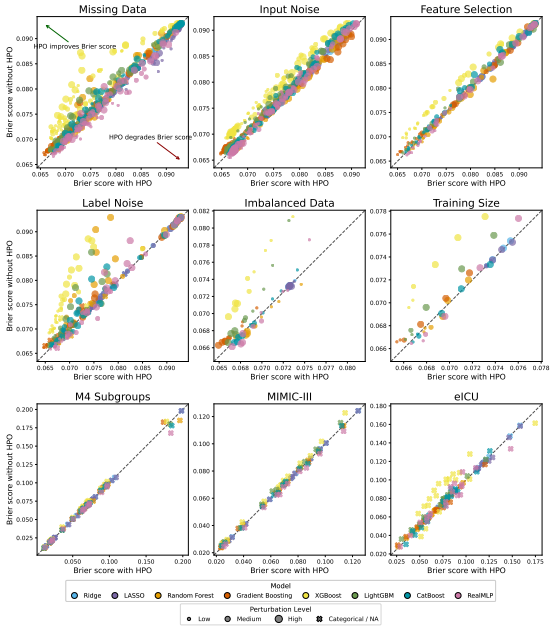

Figure 6: Classification - Effect of HPO across perturbation families when **MIMIC-IV** is used as the development set. Each point represents the mean **Brier score** for a scenario at a specific noise level or category.

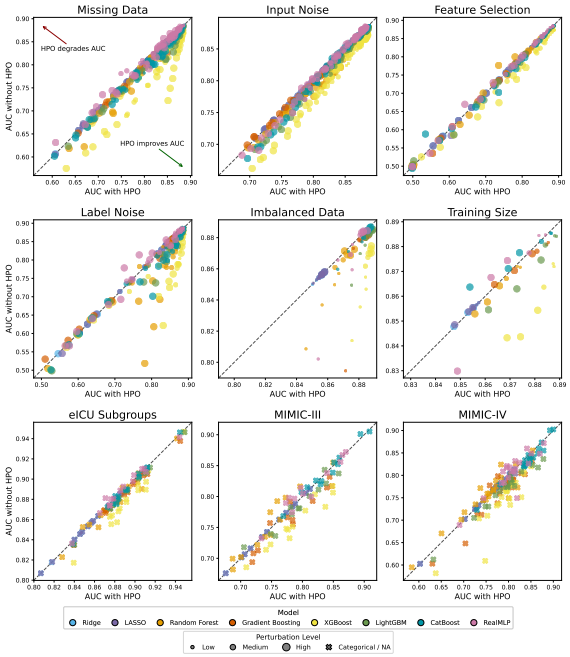

Figure 7: Classification - Effect of HPO across perturbation families when **eICU** is used as the development set. Each point represents the mean **AUC** for a scenario at a specific noise level or category.

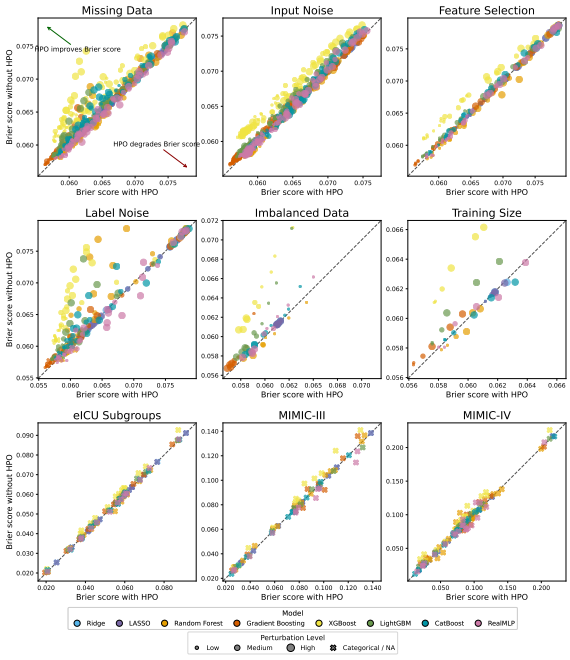

Figure 8: Classification - Effect of HPO across perturbation families when **eICU** is used as the development set. Each point represents the mean **Brier score** for a scenario at a specific noise level or category.

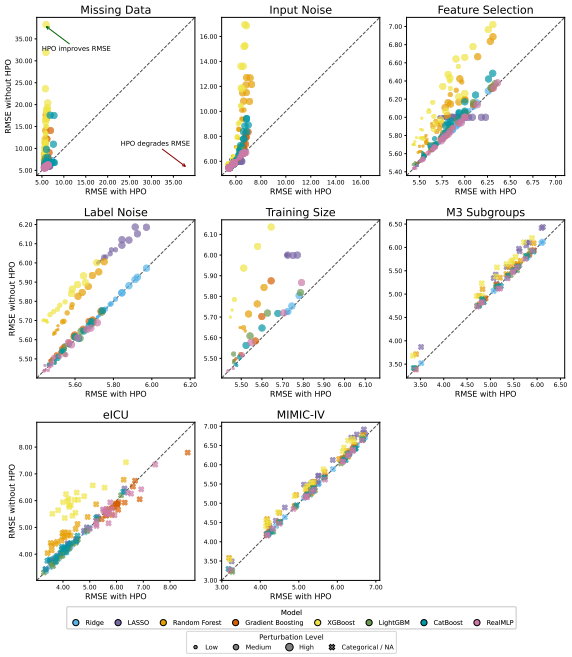

Figure 9: Regression - Effect of HPO across perturbation families when **MIMIC-III** is used as the development set. Each point represents the mean **RMSE** for a scenario at a specific noise level or category.

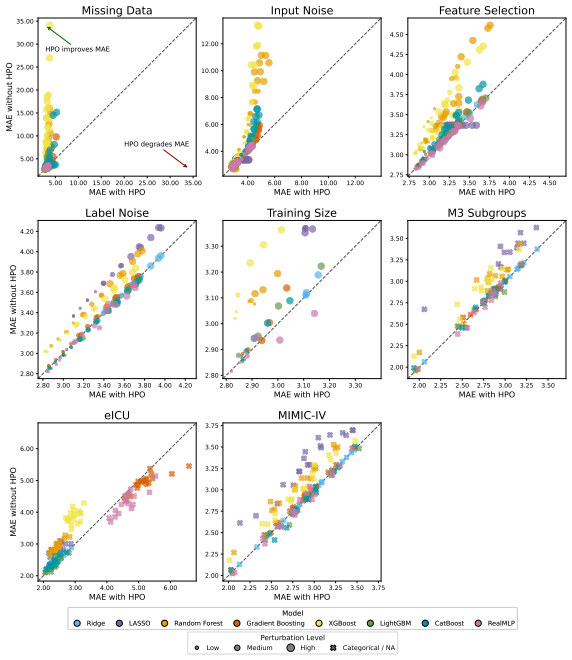

Figure 10: Regression - Effect of HPO across perturbation families when **MIMIC-III** is used as the development set. Each point represents the mean **MAE** for a scenario at a specific noise level or category.

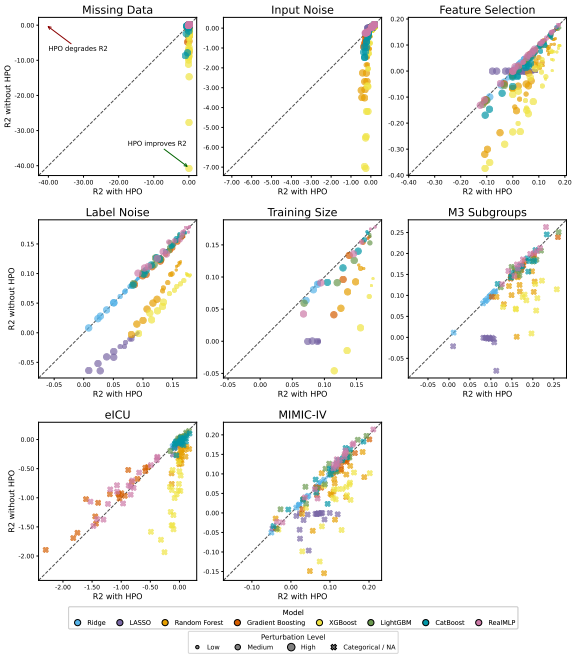

Figure 11: Regression - Effect of HPO across perturbation families when **MIMIC-III** is used as the development set. Each point represents the mean $\mathbf{R^2}$ for a scenario at a specific noise level or category.

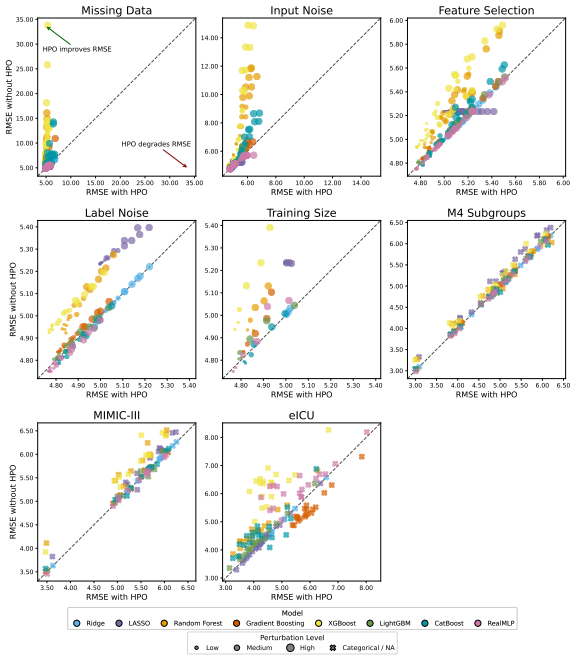

Figure 12: Regression - Effect of HPO across perturbation families when **MIMIC-IV** is used as the development set. Each point represents the mean **RMSE** for a scenario at a specific noise level or category.

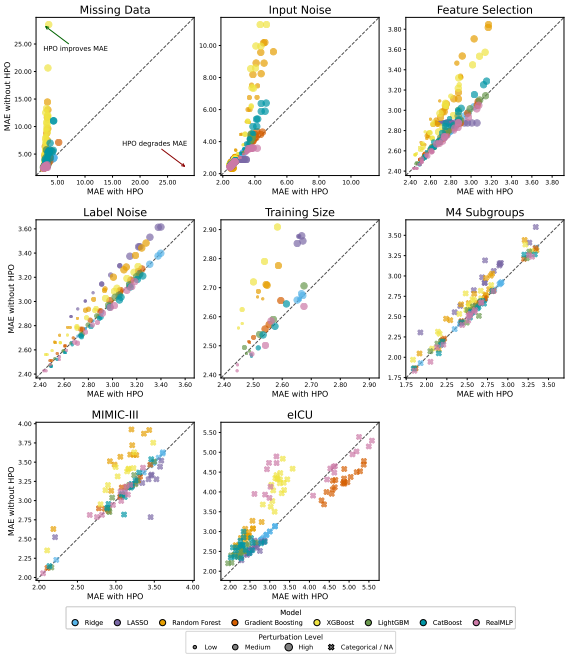

Figure 13: Regression - Effect of HPO across perturbation families when **MIMIC-IV** is used as the development set. Each point represents the mean **MAE** for a scenario at a specific noise level or category.

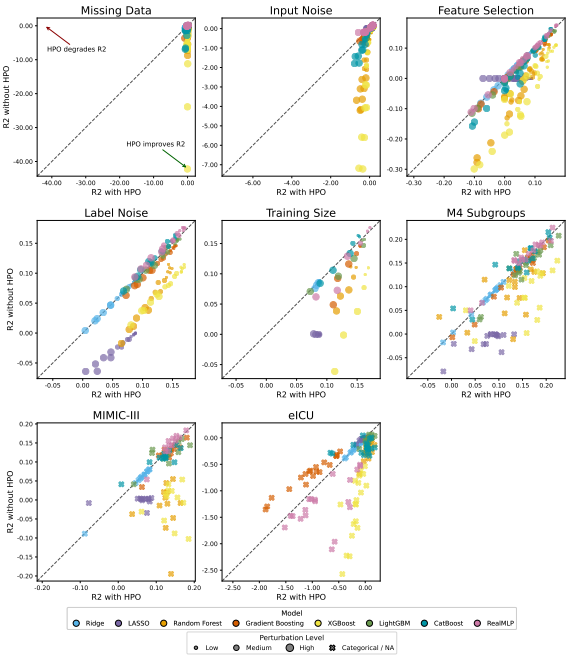

Figure 14: Regression - Effect of HPO across perturbation families when **MIMIC-IV** is used as the development set. Each point represents the mean $\mathbf{R^2}$ for a scenario at a specific noise level or category.

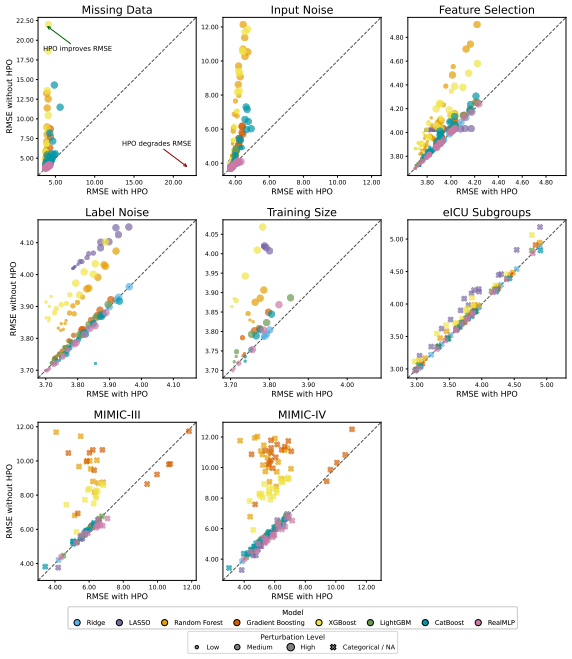

Figure 15: Regression - Effect of HPO across perturbation families when **eICU** is used as the development set. Each point represents the mean **RMSE** for a scenario at a specific noise level or category.

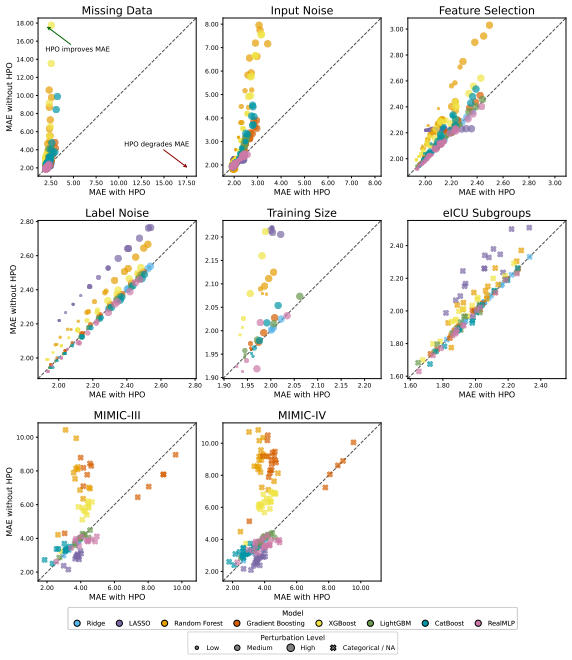

Figure 16: Regression - Effect of HPO across perturbation families when **eICU** is used as the development set. Each point represents the mean **MAE** for a scenario at a specific noise level or category.

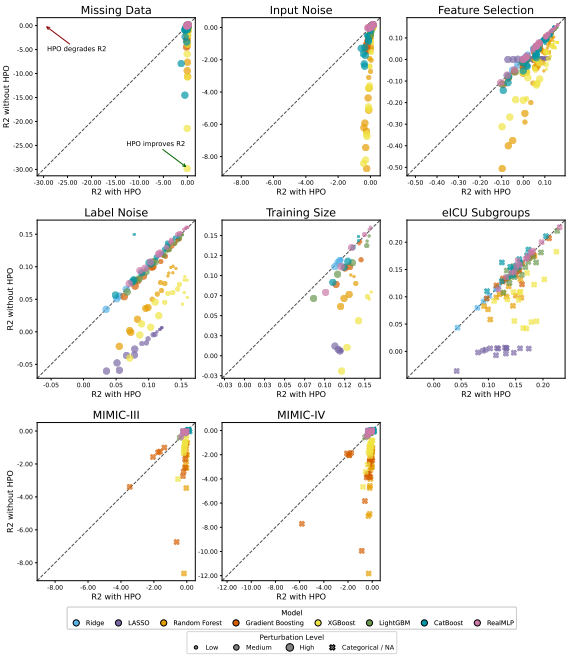

Figure 17: Regression - Effect of HPO across perturbation families when **eICU** is used as the development set. Each point represents the mean $\mathbf{R^2}$ for a scenario at a specific noise level or category.

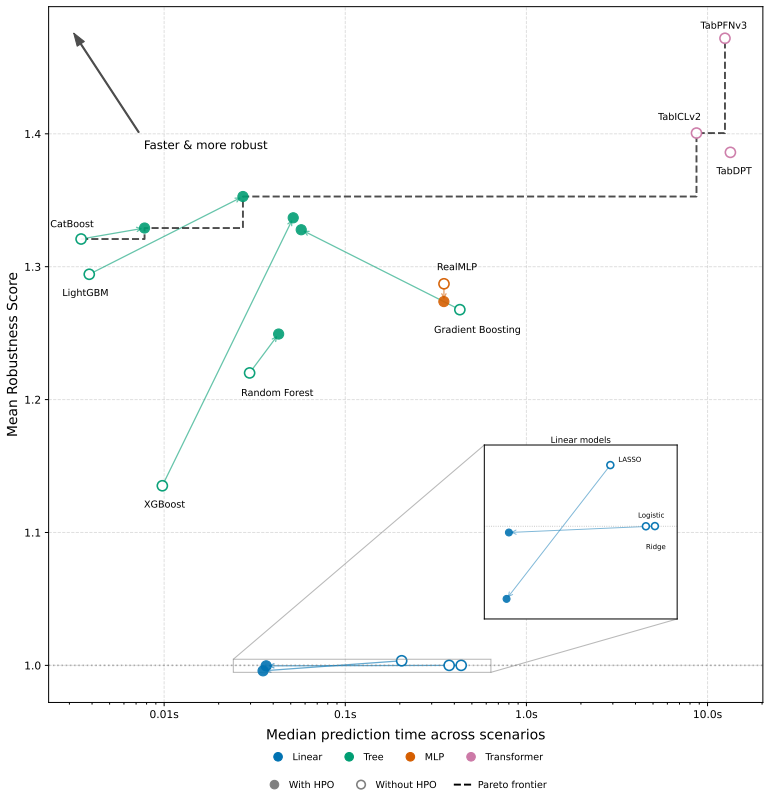

Figure 18: *Mean Robustness Score* versus median inference time across all scenarios. **MIMIC-IV** is used as the development set.

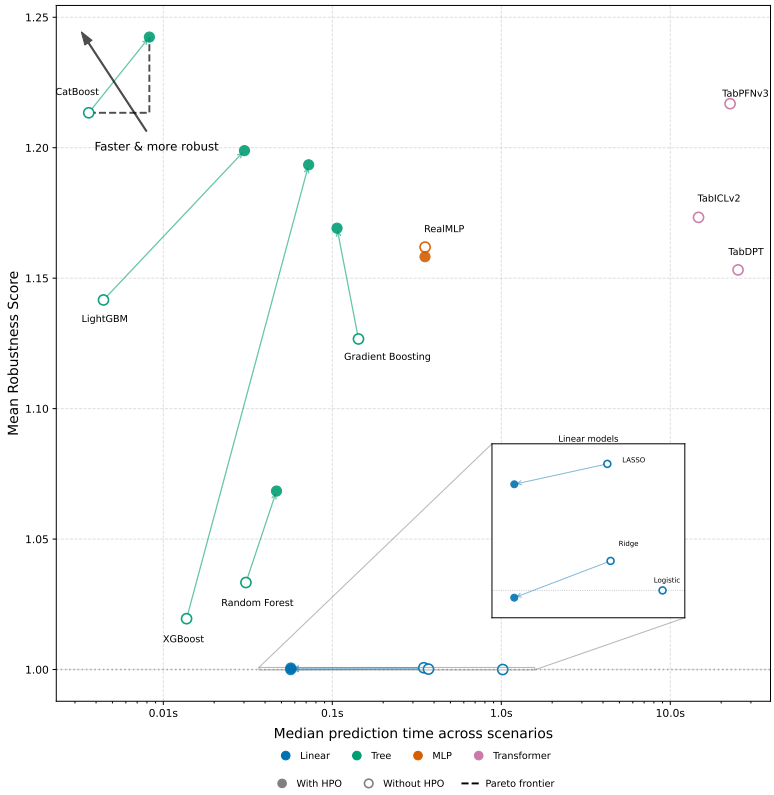

Figure 19: *Mean Robustness Score* versus median inference time across all scenarios. **eICU** is used as the development set.

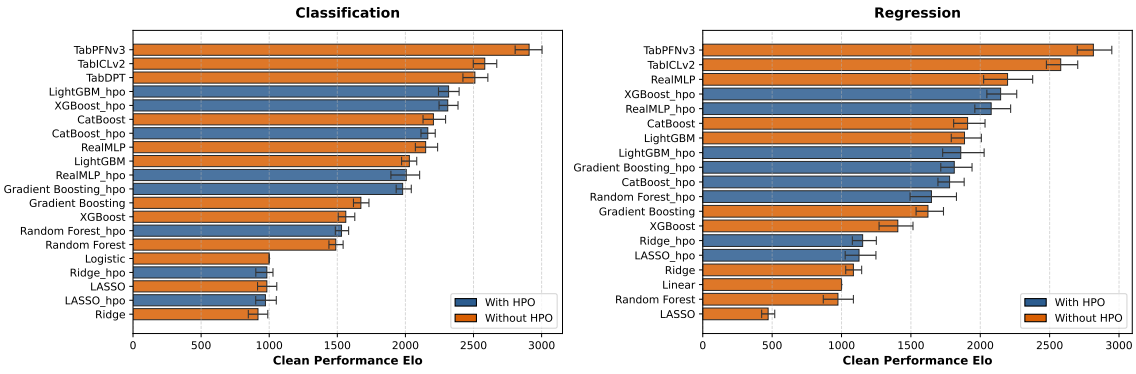

Figure 20: **Left:** Clean robustness ELO computed on unperturbed baseline for classification. **Right:** Clean robustness ELO computed on unperturbed baseline for regression.

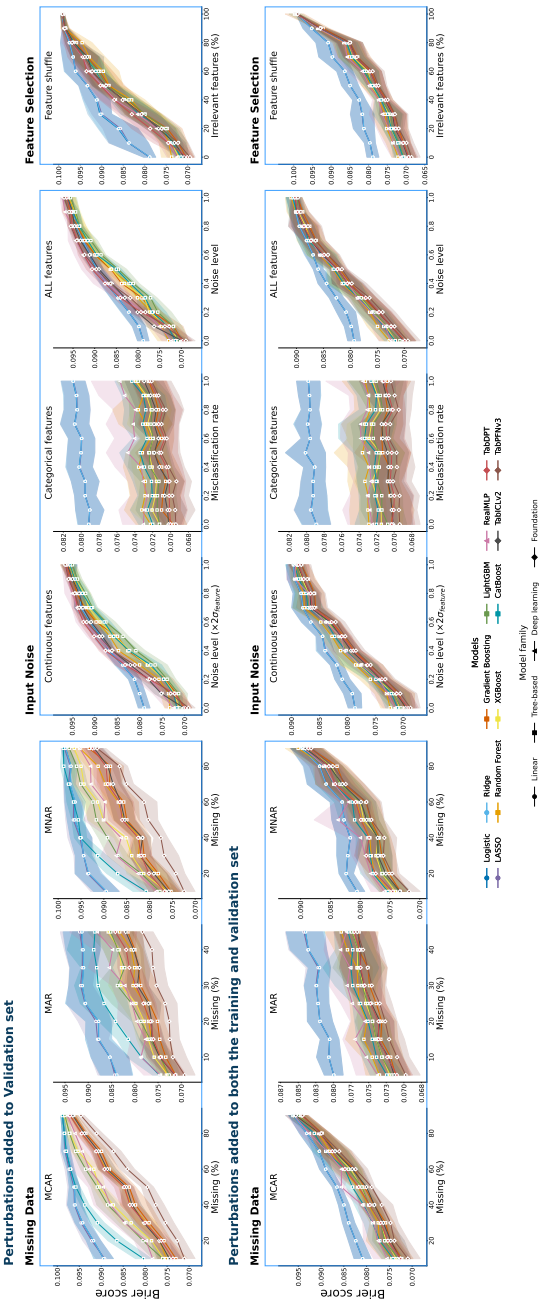

Figure 21: **Brier score** under perturbation added only in the validation-data (first row) and both training and validation (second row): **MIMIC-III** development cohort

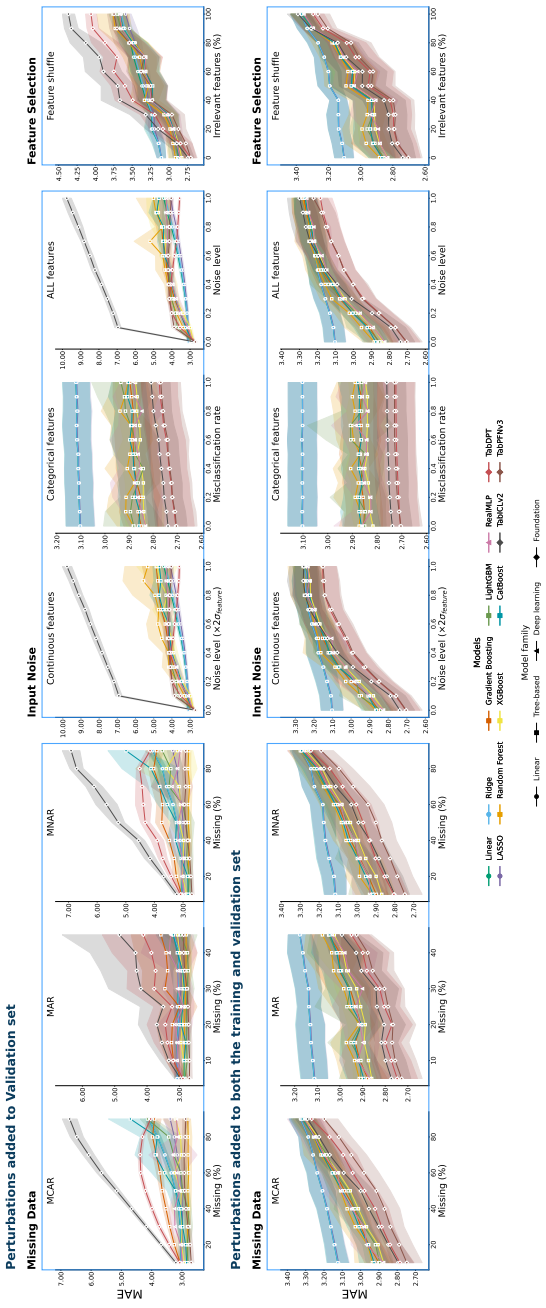

Figure 22: **MAE** under perturbation added only in the validation-data (first row) and both training and validation (second row): **MIMIC-III** development cohort

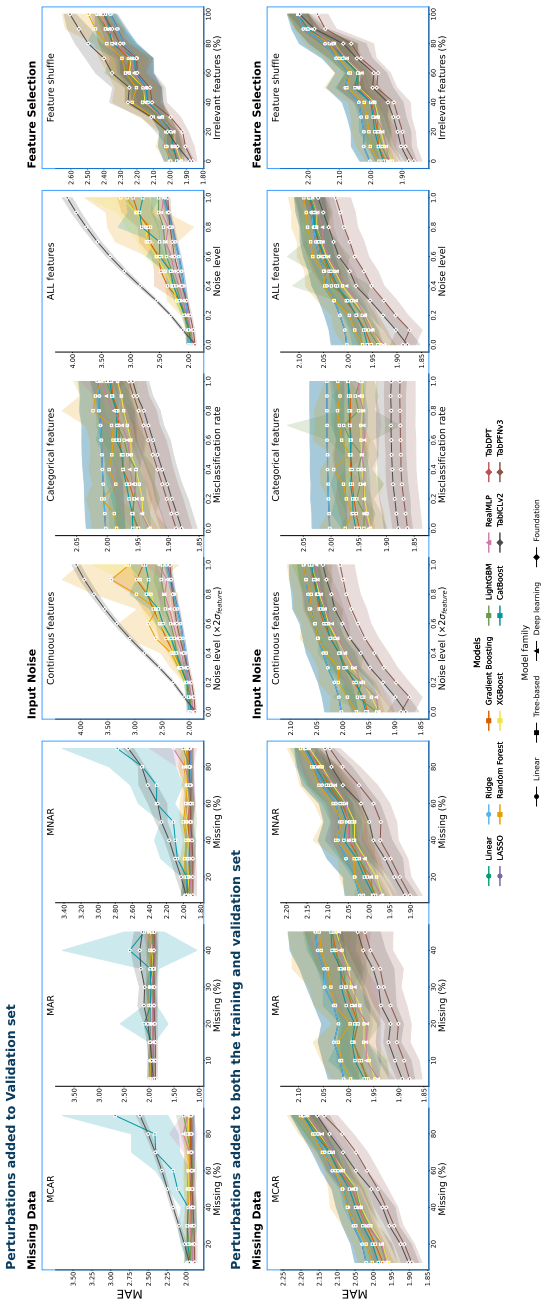

Figure 23: **MAE** under perturbation added only in the validation-data (first row) and both training and validation (second row): **eICU** development cohort

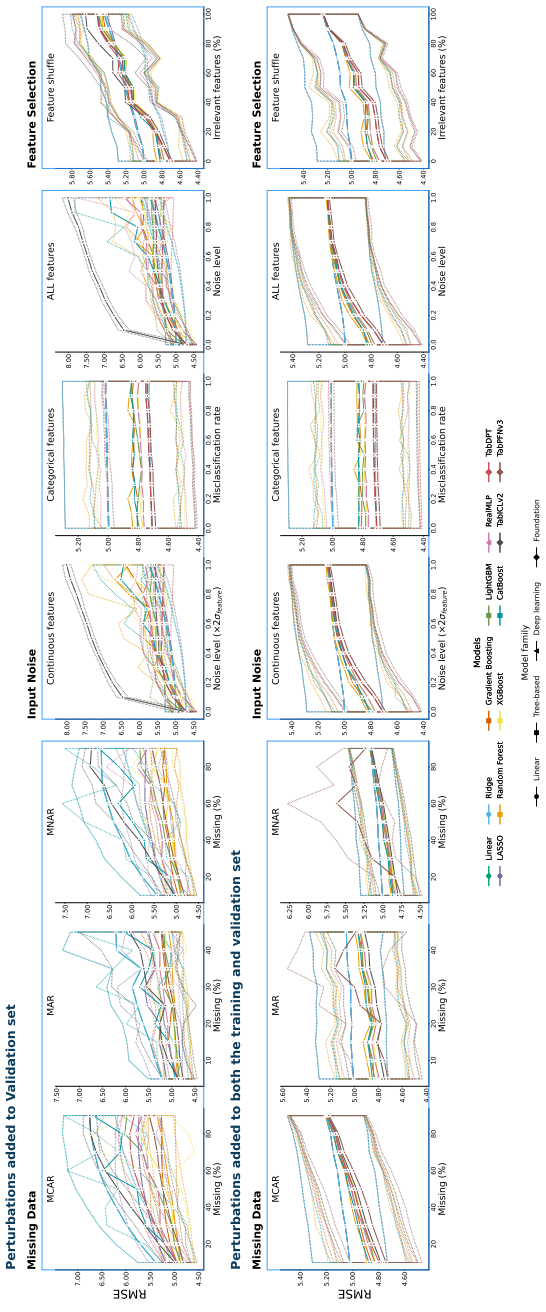

Figure 24: **RMSE** under perturbation added only in the validation-data (first row) and both training and validation (second row): **MIMIC-IV** development cohort

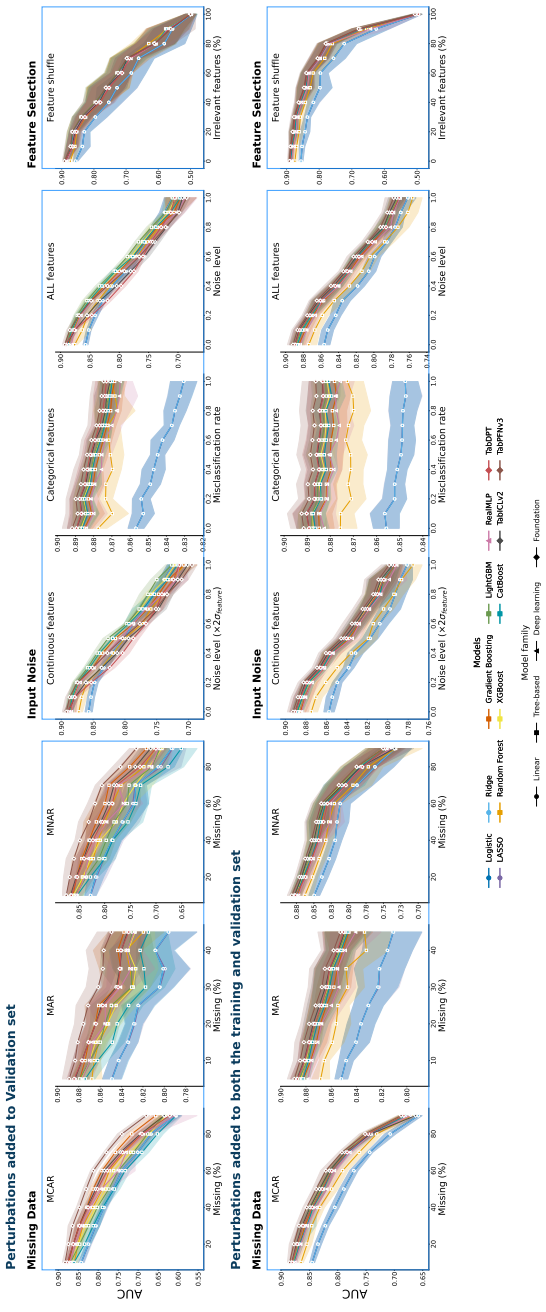

Figure 25: **AUC** under perturbation added only in the validation-data (first row) and both training and validation (second row): **eICU** development cohort

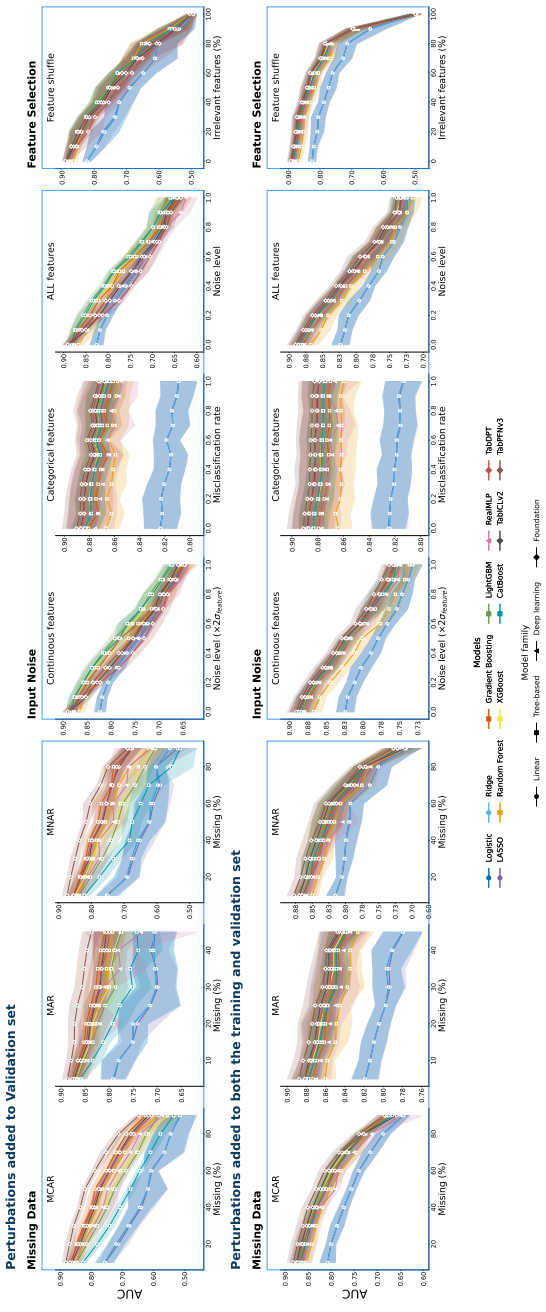

Figure 26: **AUC** under perturbation added only in the validation-data (first row) and both training and validation (second row): **MIMIC-III** development cohort

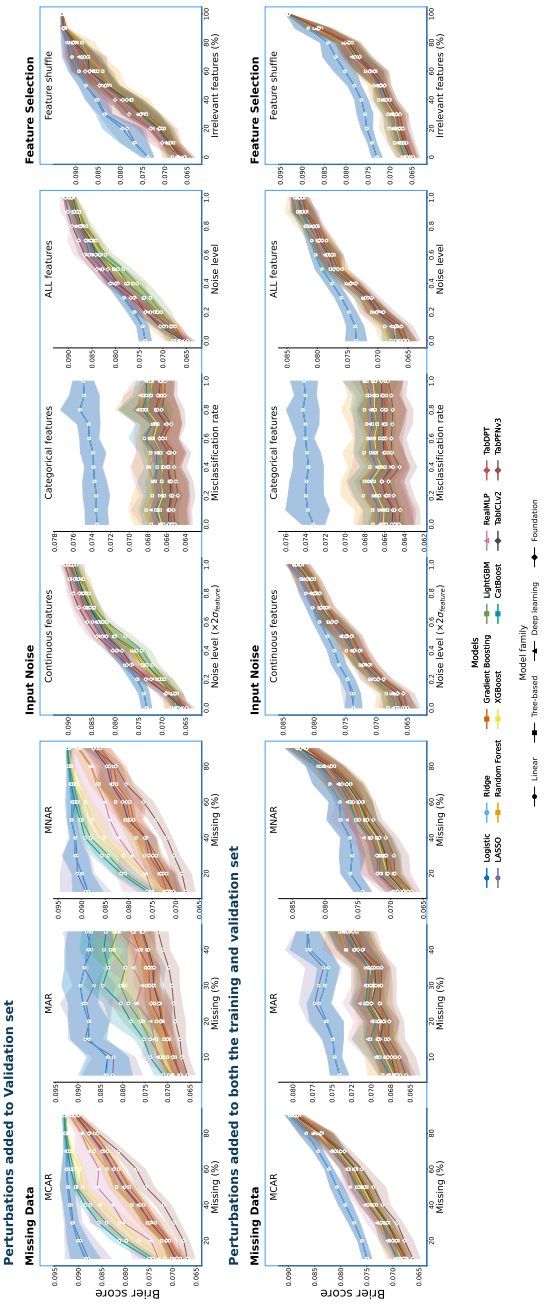

Figure 27: **Brier score** under perturbation added only in the validation-data (first row) and both training and validation (second row): **MIMIC-IV** development cohort

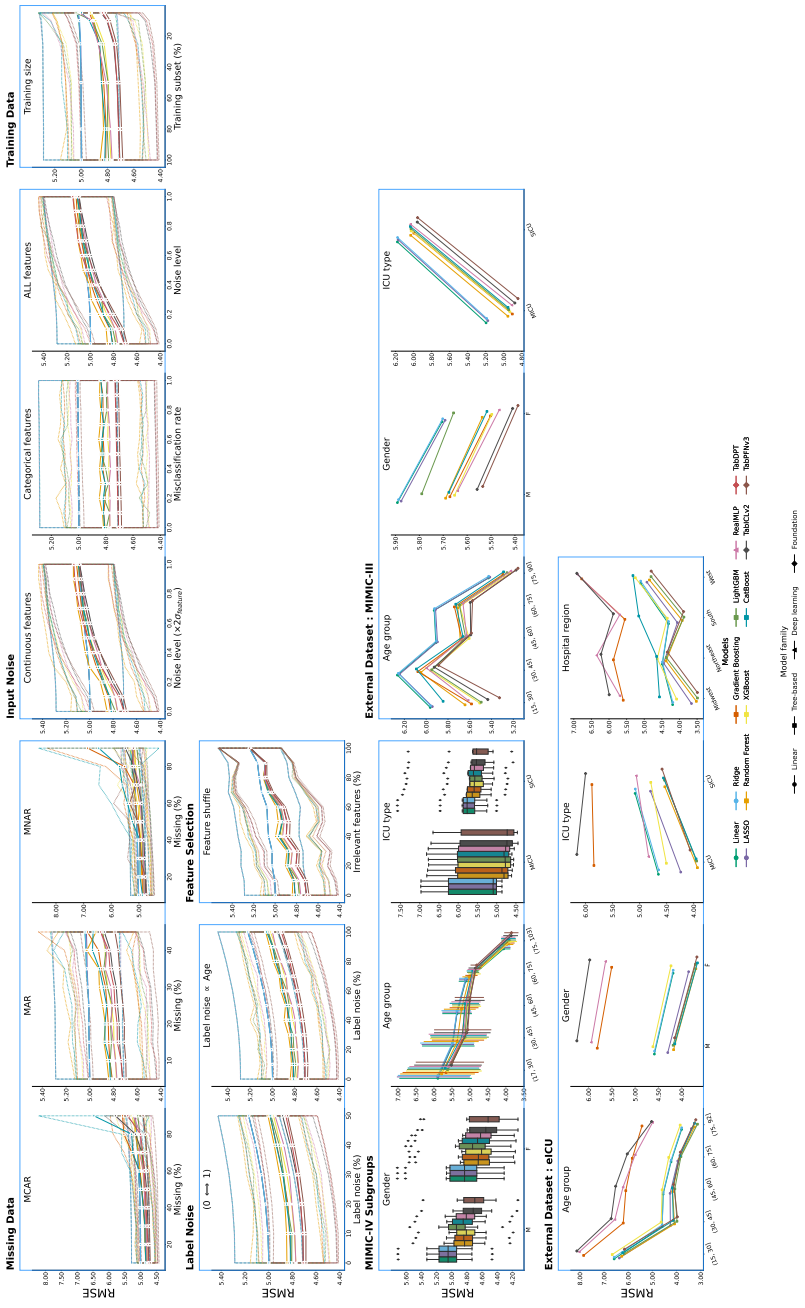

Figure 28: **RMSE** under training-data perturbations and external validation: **MIMIC-IV** development cohort

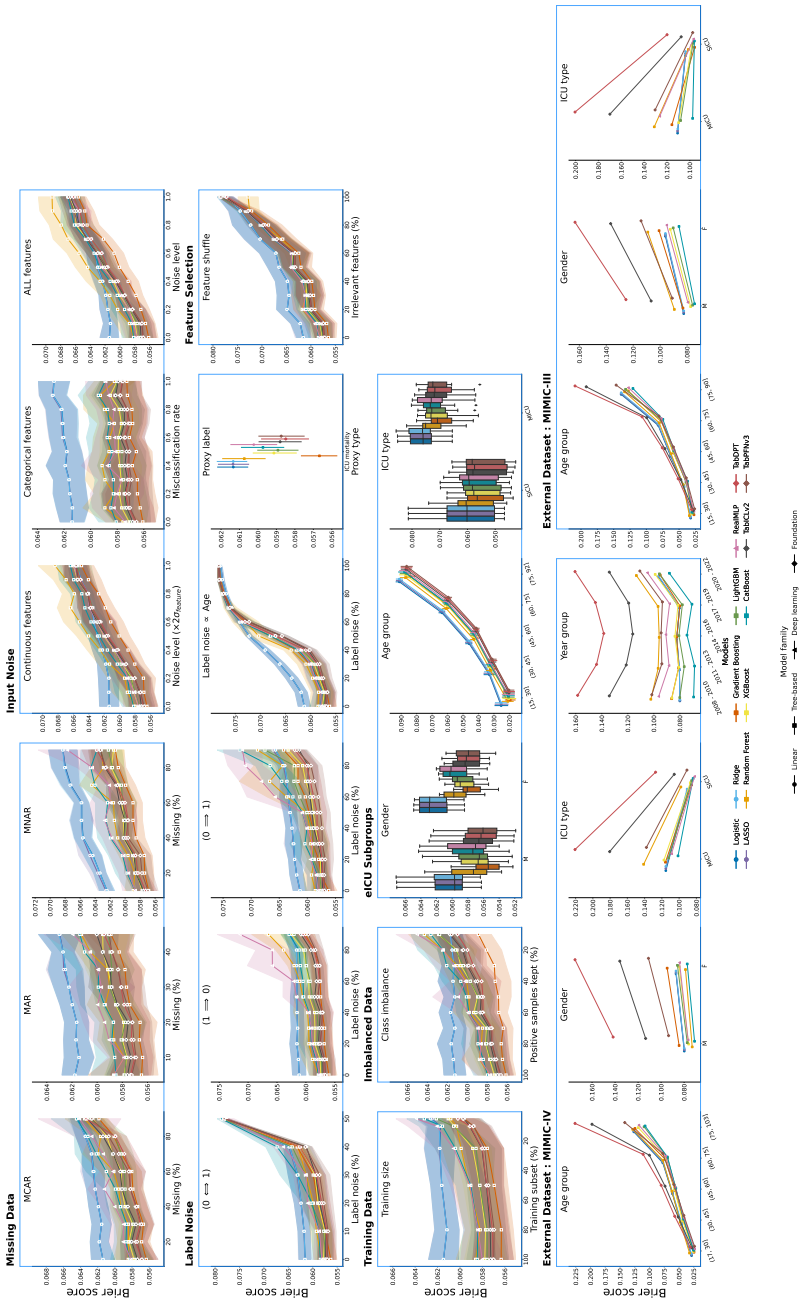

Figure 29: **Brier score** under training-data perturbations and external validation: **eICU** development cohort

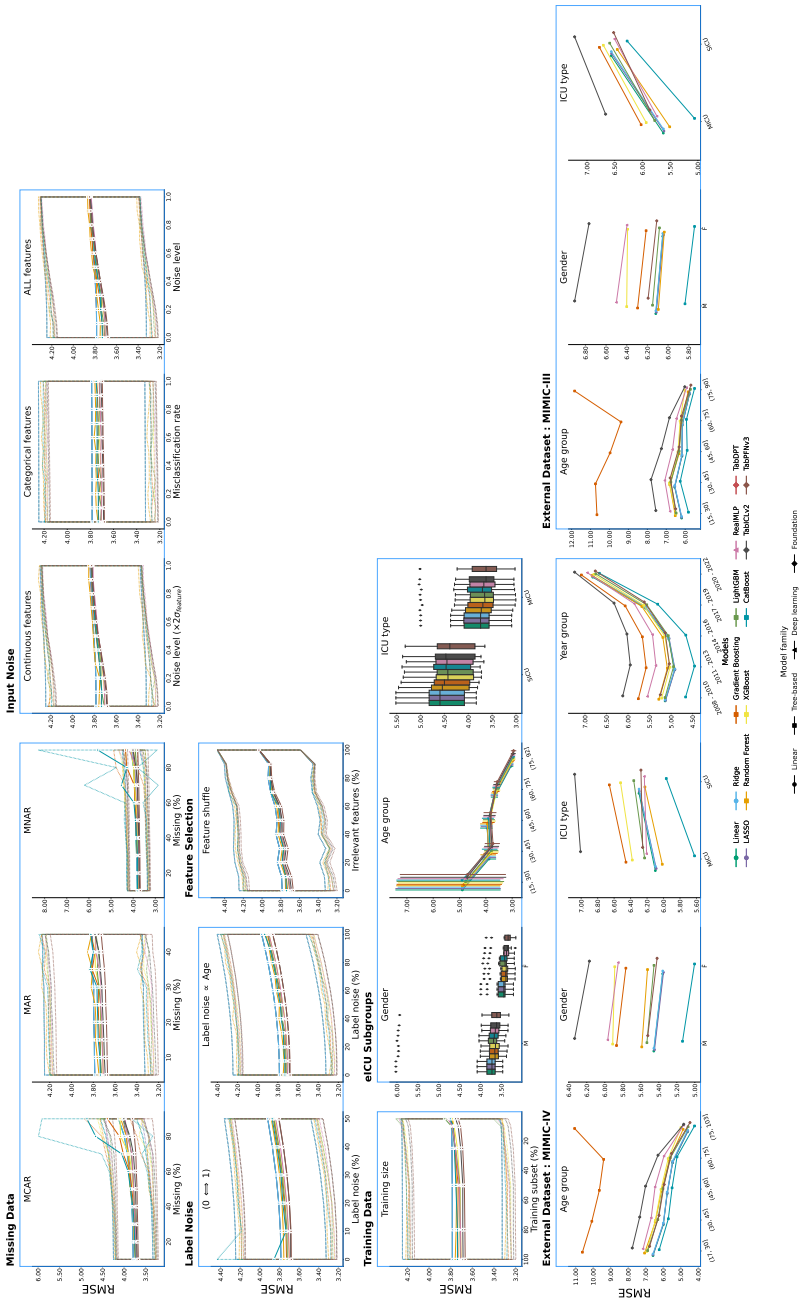

Figure 30: **RMSE** under training-data perturbations and external validation: **eICU** development cohort

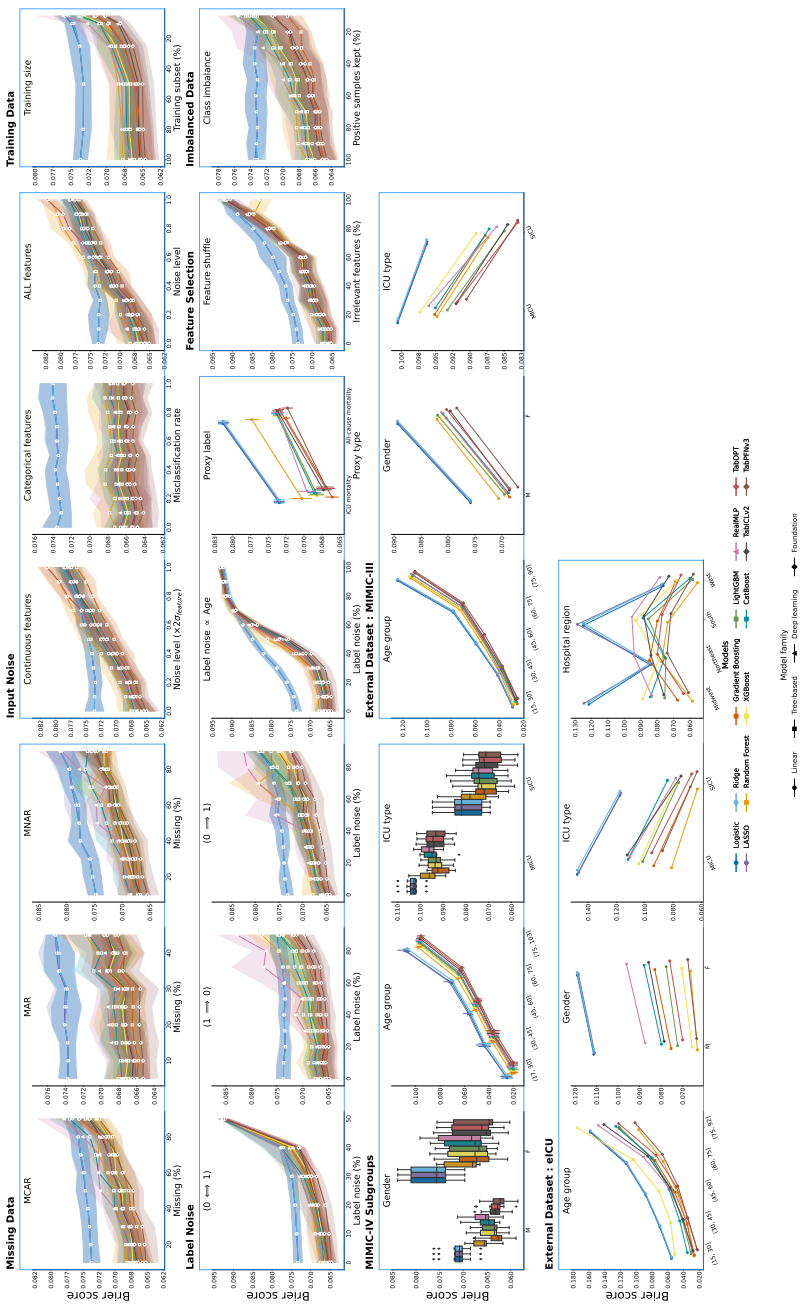

Figure 31: **Brier score** under training-data perturbations and external validation: **MIMIC-IV** development cohort

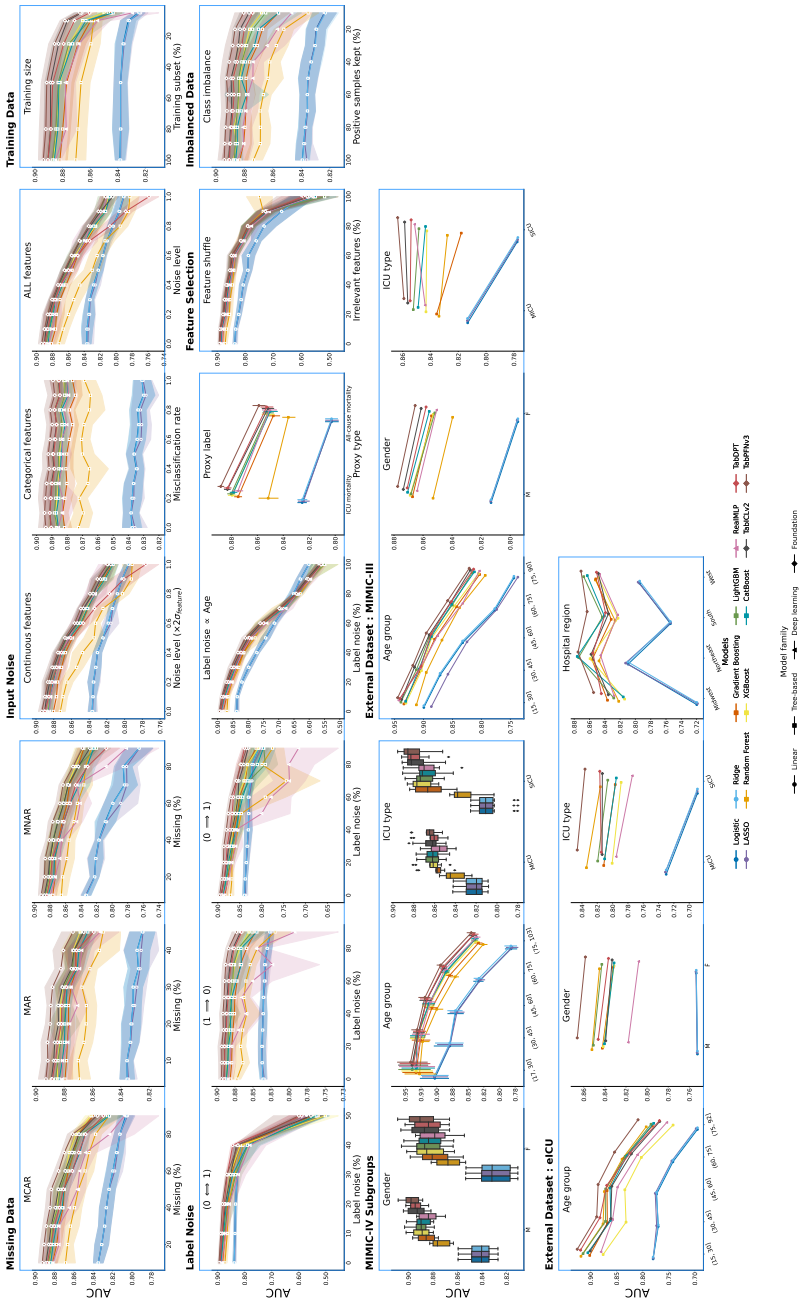

Figure 32: **AUC** under training-data perturbations and external validation: **MIMIC-IV** development cohort

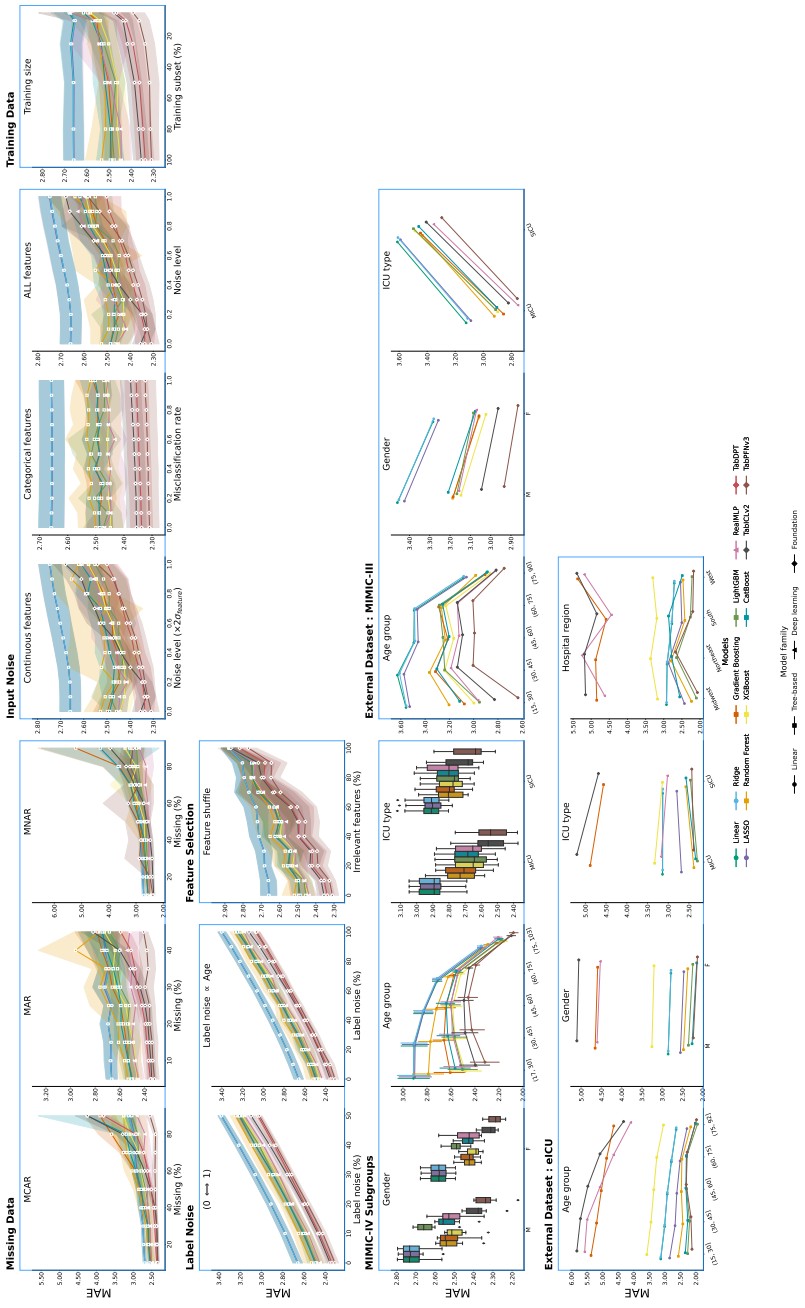

Figure 33: **MAE** under training-data perturbations and external validation: **MIMIC-IV** development cohort

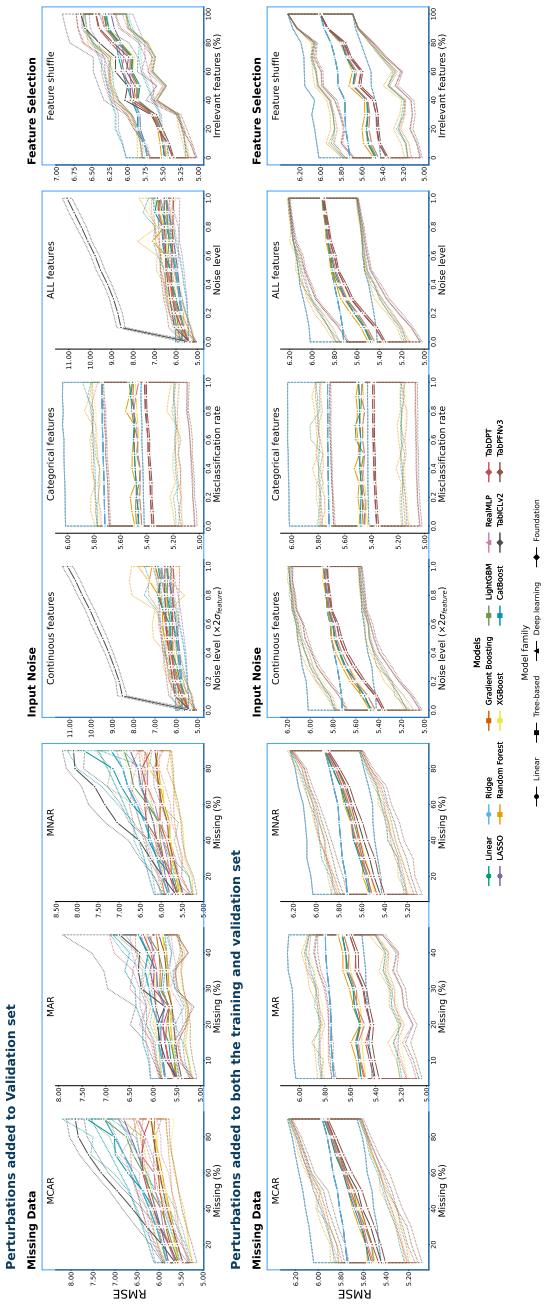

Figure 34: **RMSE** under perturbation added only in the validation-data (first row) and both training and validation (second row): **MIMIC-III** development cohort

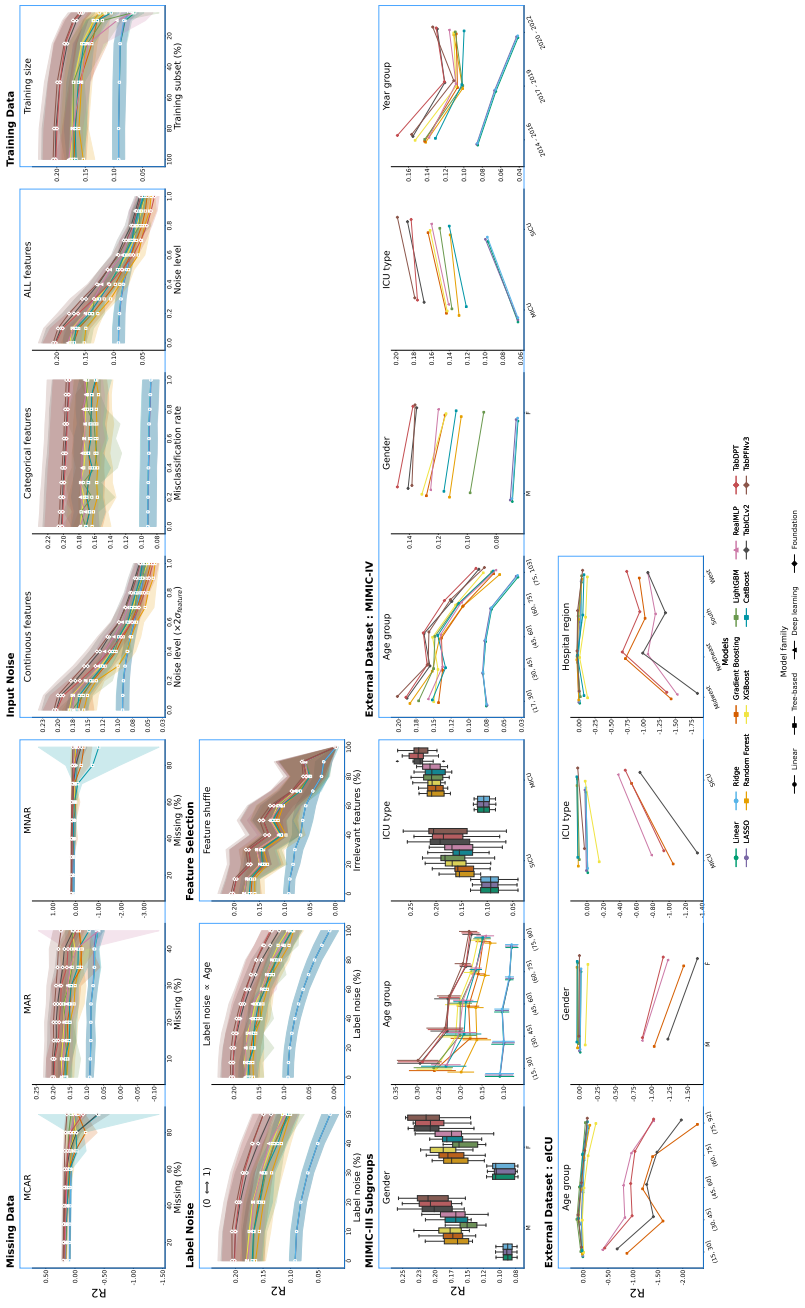

Figure 35: R^2 under training-data perturbations and external validation: **MIMIC-III** development cohort

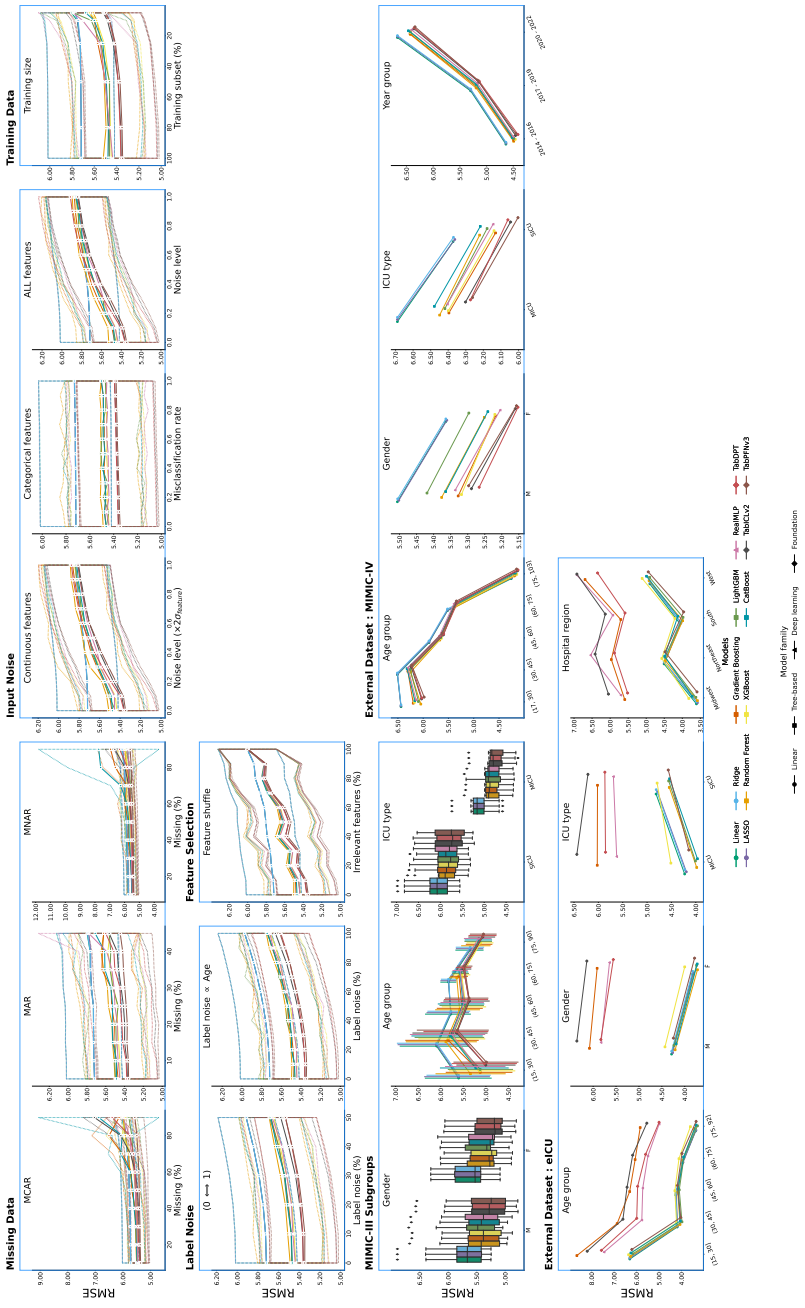

Figure 36: **RMSE** under training-data perturbations and external validation: **MIMIC-III** development cohort

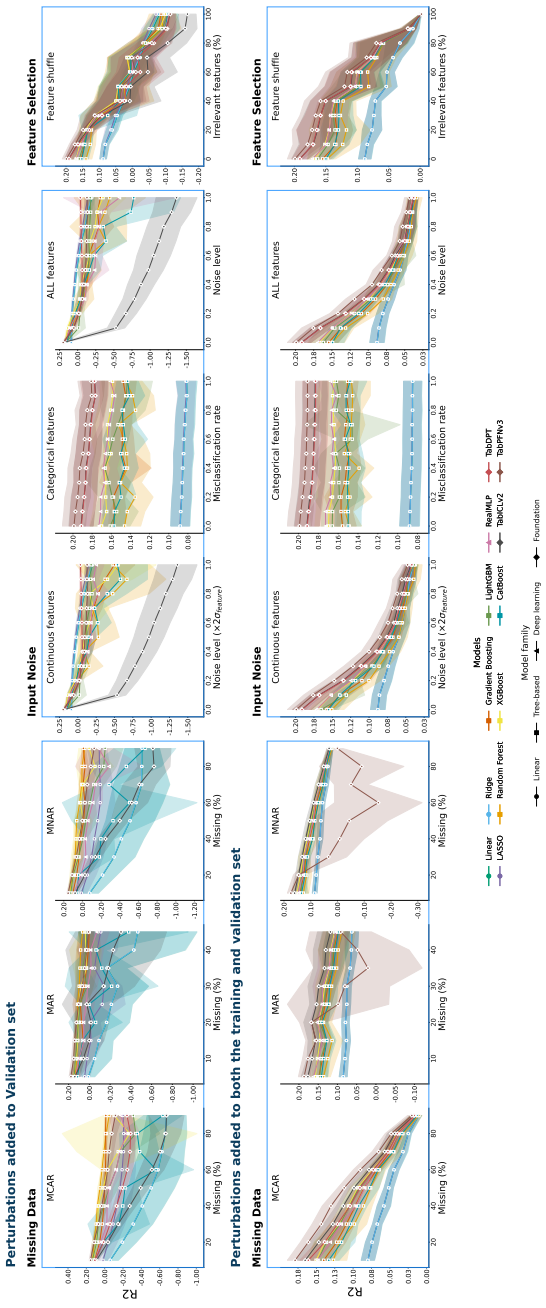

Figure 37: R^2 under perturbation added only in the validation-data (first row) and both training and validation (second row): **MIMIC-IV** development cohort

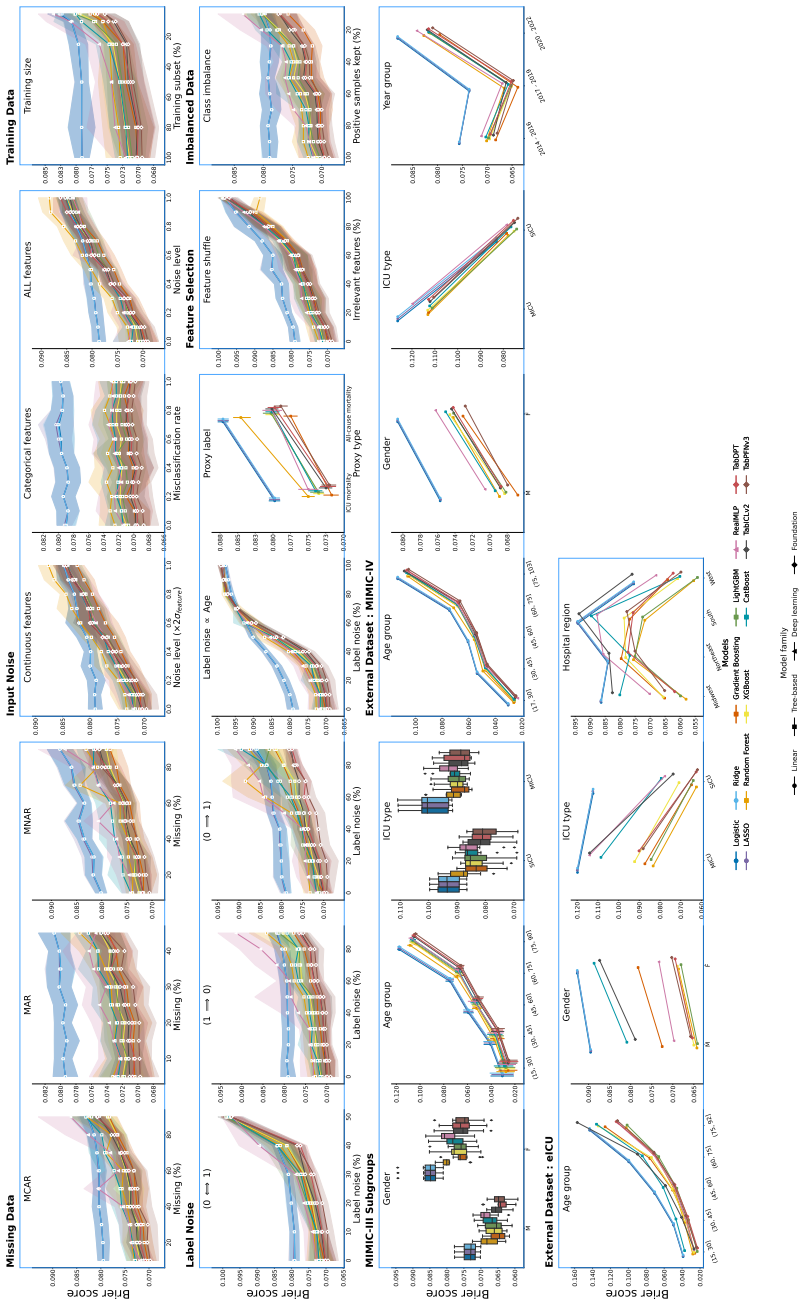

Figure 38: **Brier score** under training-data perturbations and external validation: **MIMIC-III** development cohort

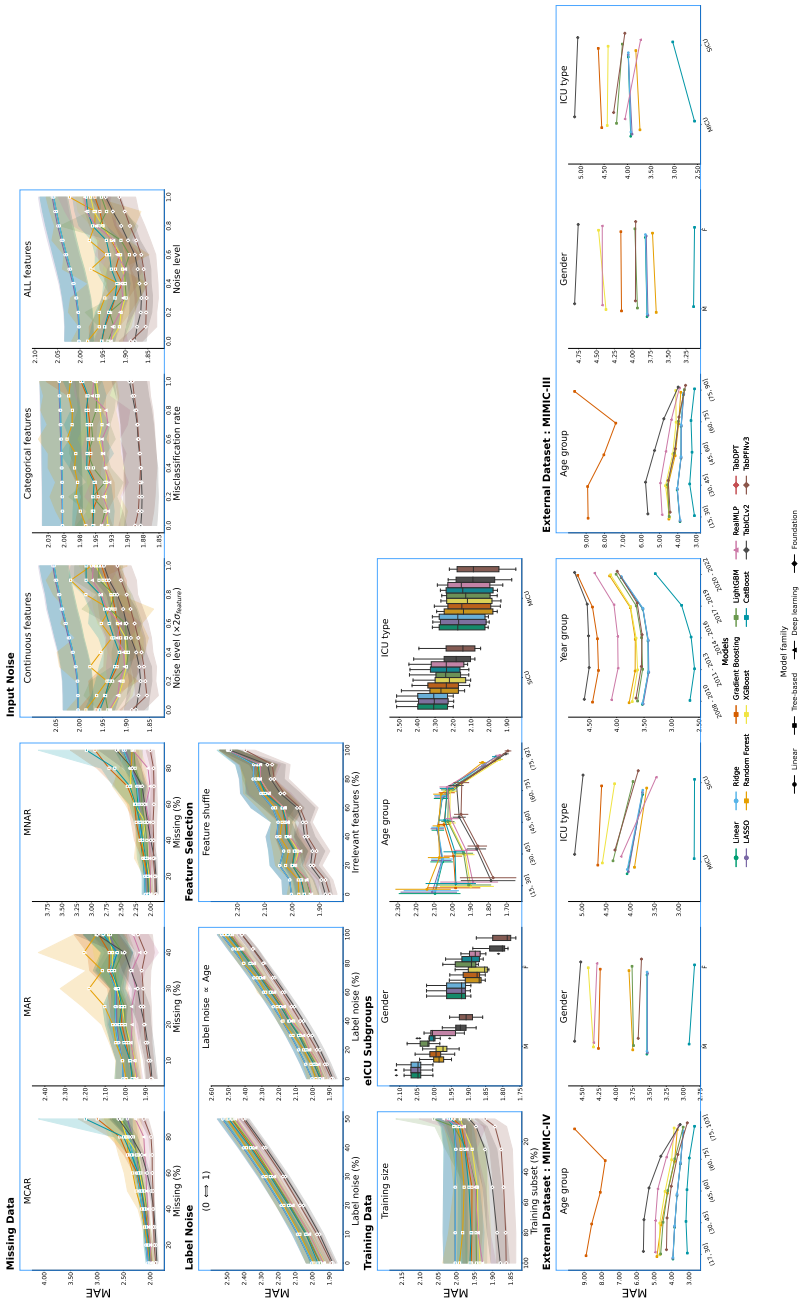

Figure 39: **MAE** under training-data perturbations and external validation: **eICU** development cohort

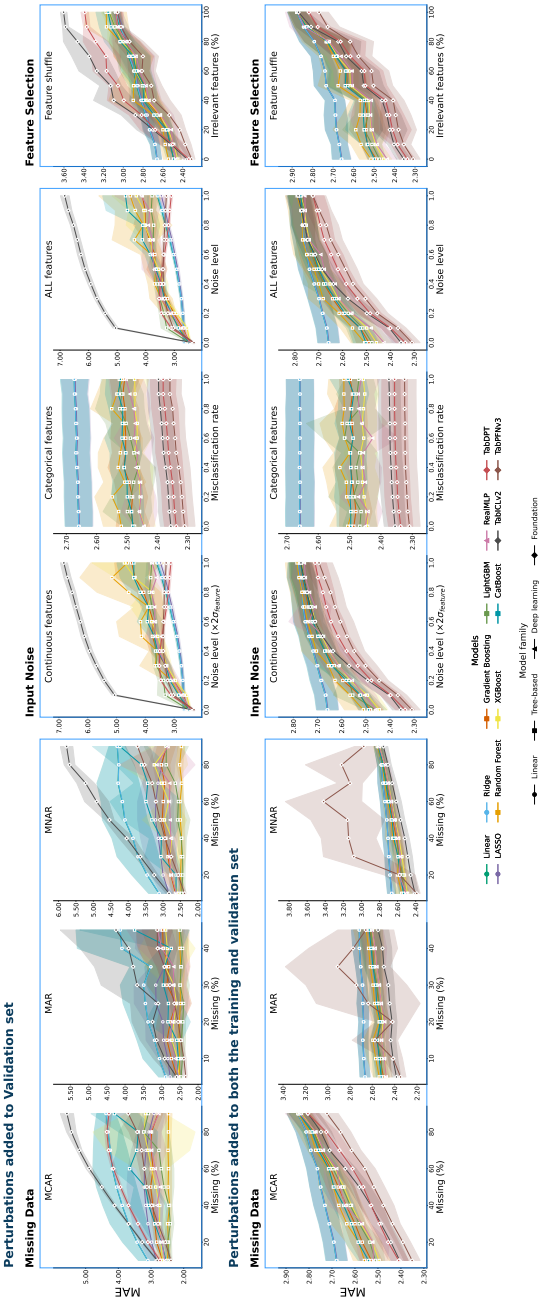

Figure 40: **MAE** under perturbation added only in the validation-data (first row) and both training and validation (second row): **MIMIC-IV** development cohort

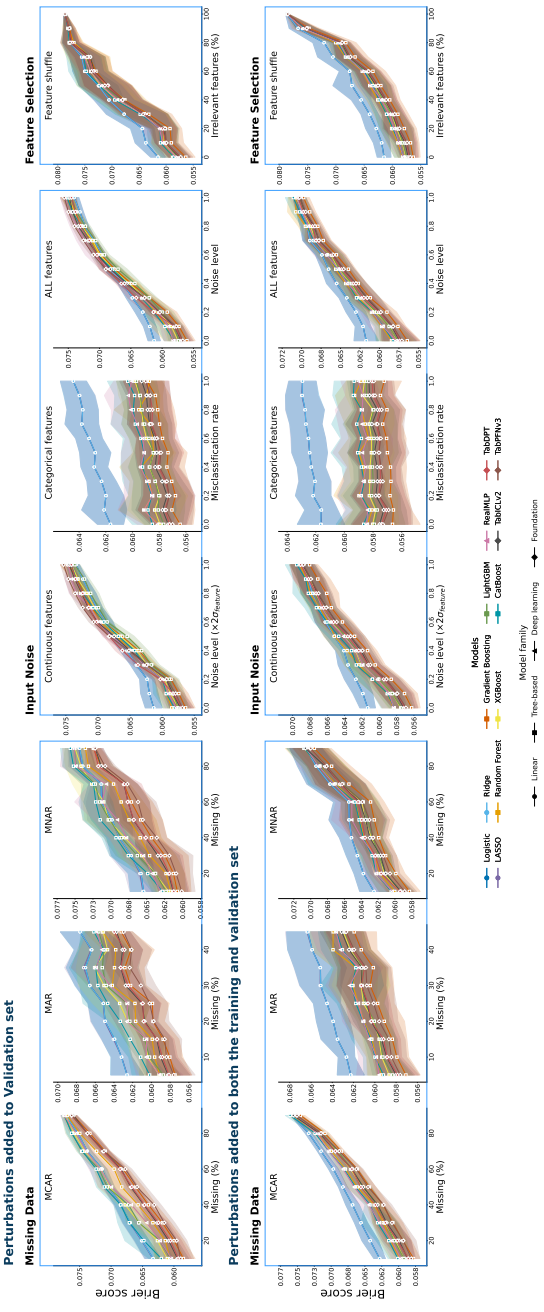

Figure 41: **Brier score** under perturbation added only in the validation-data (first row) and both training and validation (second row): **eICU** development cohort

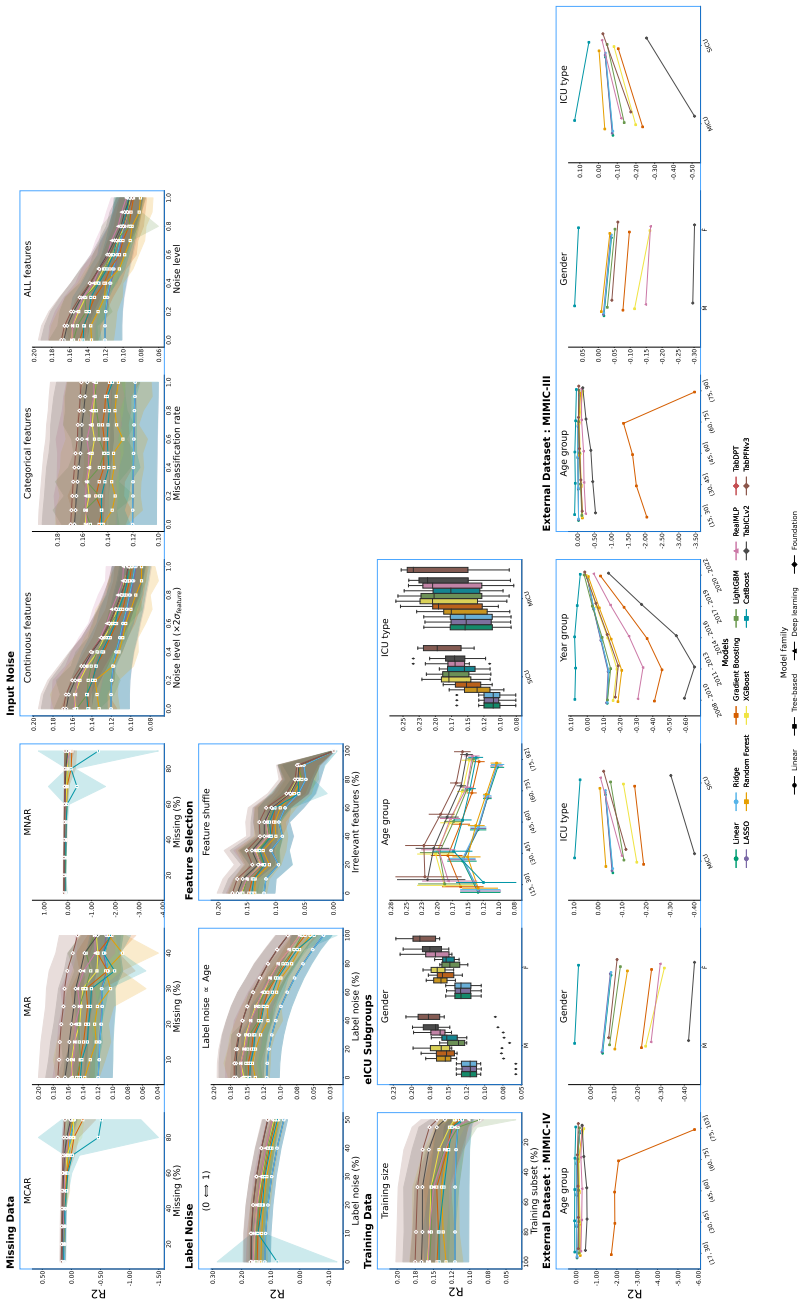

Figure 42: R^2 under training-data perturbations and external validation: **eICU** development cohort

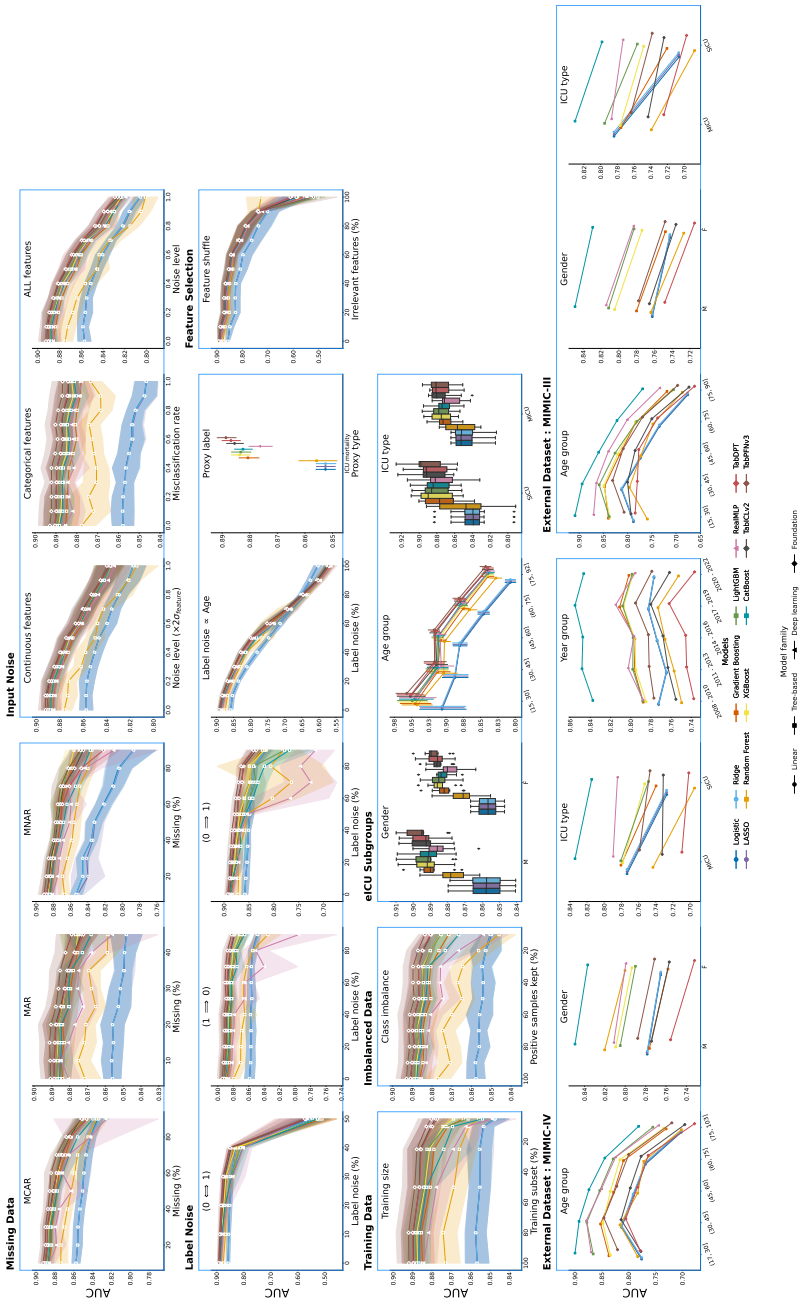

Figure 43: **AUC** under training-data perturbations and external validation: **eICU** development cohort

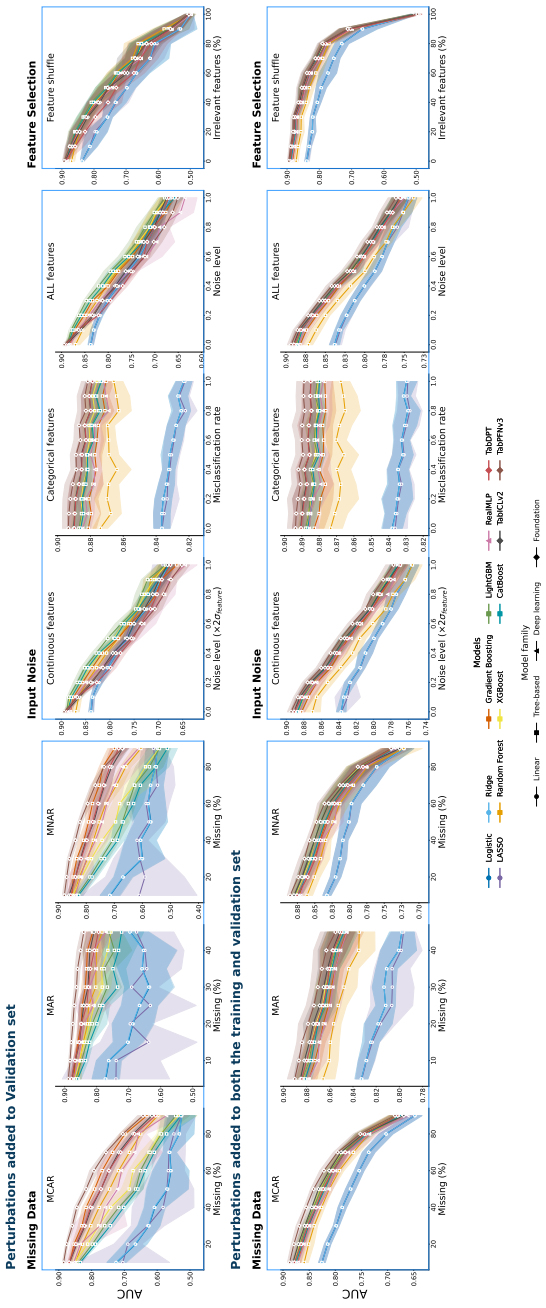

Figure 44: **AUC** under perturbation added only in the validation-data (first row) and both training and validation (second row): **MIMIC-IV** development cohort

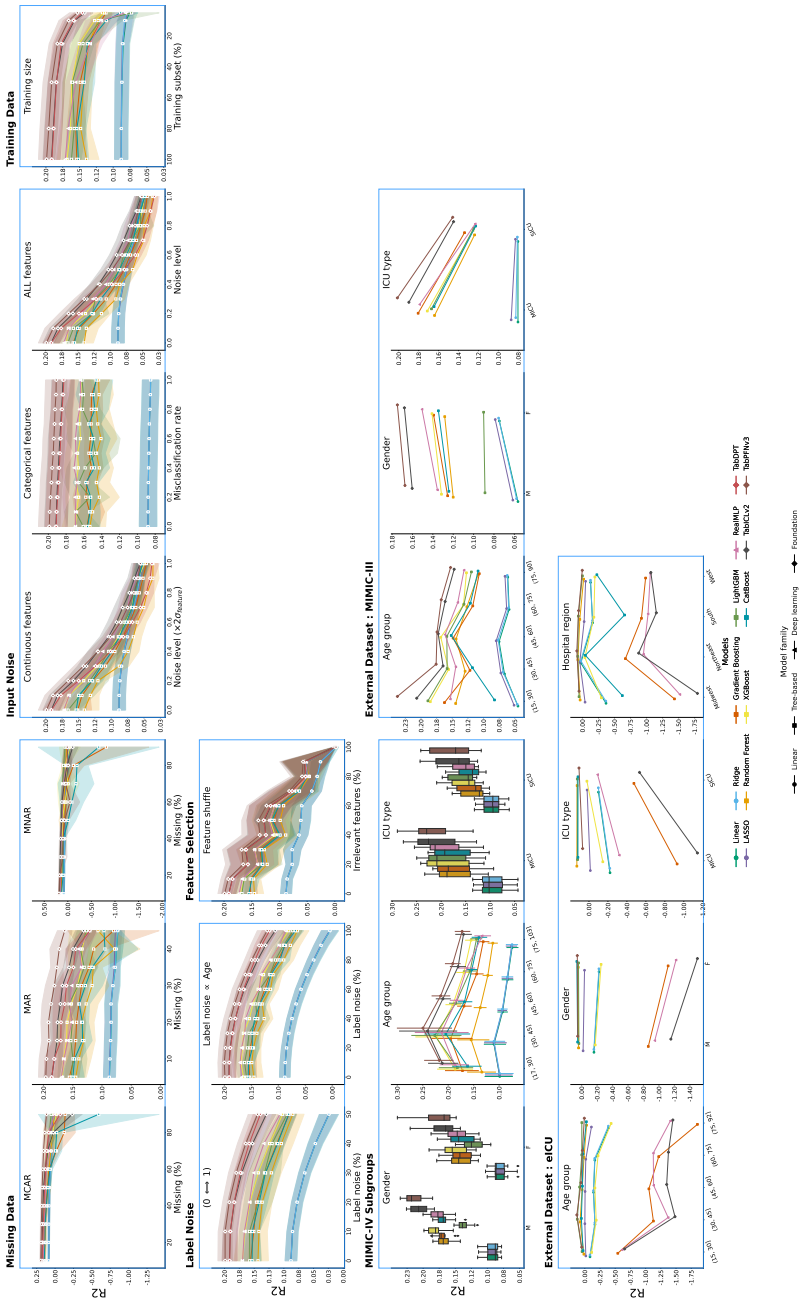

Figure 45: R^2 under training-data perturbations and external validation: **MIMIC-IV** development cohort

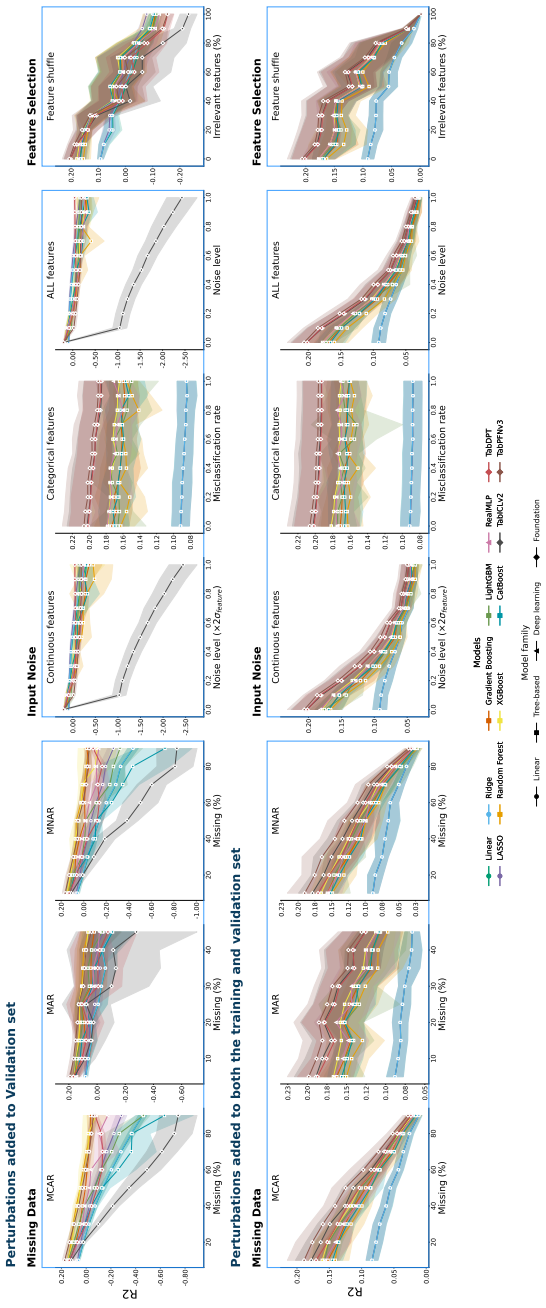

Figure 46: R^2 under perturbation added only in the validation-data (first row) and both training and validation (second row): **MIMIC-III** development cohort

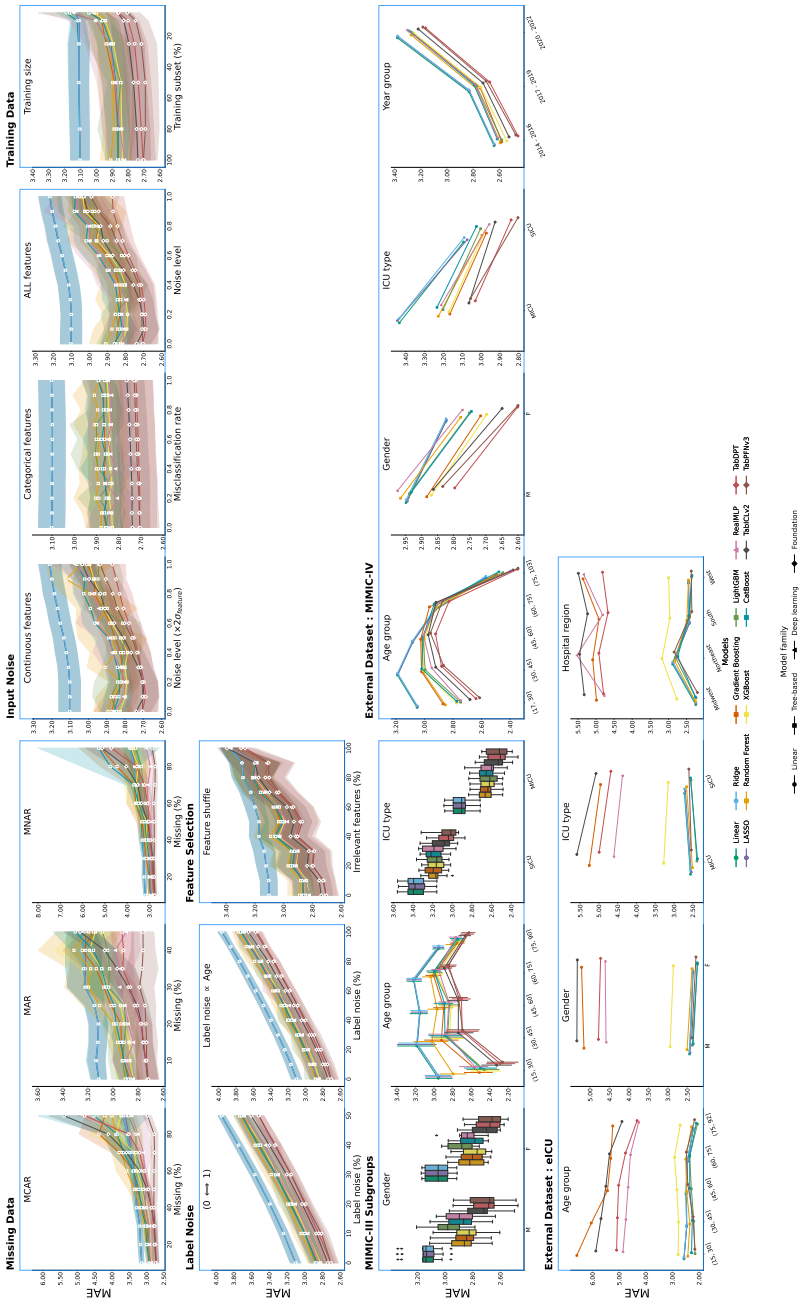

Figure 47: **MAE** under training-data perturbations and external validation: **MIMIC-III** development cohort

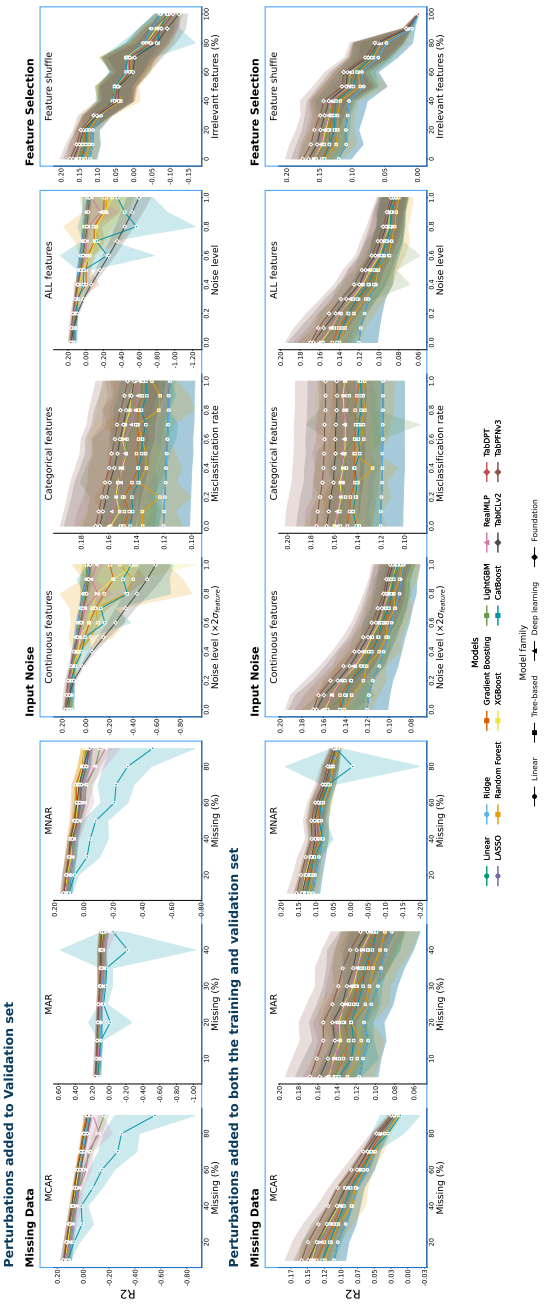

Figure 48: R^2 under perturbation added only in the validation-data (first row) and both training and validation (second row): **eICU** development cohort

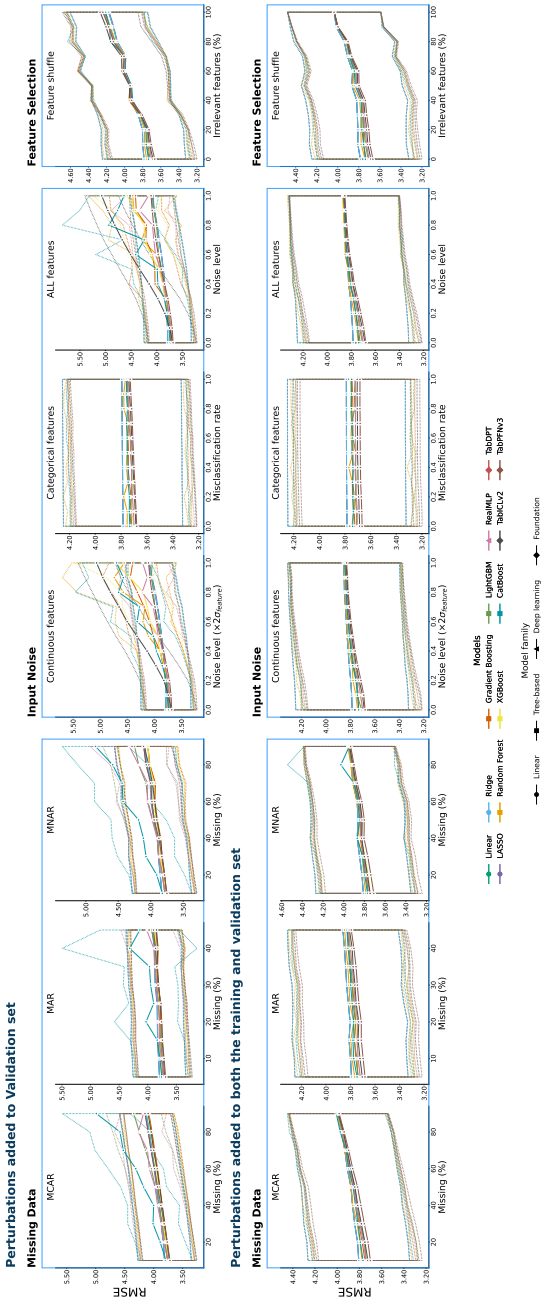

Figure 49: **RMSE** under perturbation added only in the validation-data (first row) and both training and validation (second row): **eICU** development cohort