

Effect of Features Burdened with Classical Error, Misclassifications, Berkson Error, and Group-Summary Assignments on Machine Learning Performance and Interpretability

Zeinab Mohamed

Oberlin College

Tamer Oraby

tamer.oraby@utrgv.edu

The University of Texas Rio Grande Valley

Research Article

Keywords: Classical Measurement Error, Berkson Error, Misclassifications, Group-summary Assignment, Machine Learning, Monte Carlo Experiments, Statistical Analyses

Posted Date: February 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8963493/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Effect of Features Burdened with Classical Error, Misclassifications, Berkson Error, and Group-Summary Assignments on Machine Learning Performance and Interpretability

Zeinab Mohamed* Tamer Oraby†

February 25, 2026

Abstract

Measurement error is widespread in modern data pipelines, but its influence on predictive modeling and variable importance assessments in machine learning is often underestimated. Our study examines how input measurement error affects both prediction accuracy and interpretability across four common additive error types: classical error, misclassification of predictive features, Berkson error, caused by algorithmic or model-based assignment errors, and group-summary (aggregation) assignment errors. Using controlled Monte Carlo simulation experiments that cover regression and classification, we evaluate five popular machine learning models: generalized linear models, support vector machines, random forests, XGBoost, and multilayer perceptrons, across various error levels. To distinguish the effect of the error from optimization issues, hyperparameters are tuned separately for each dataset, and performance is measured using standard predictive metrics and permutation feature importance. Results show that measurement error generally reduces predictive performance and can significantly alter the ranking and estimated contribution of highly predictive features, with different degradation patterns depending on the error type and model. These findings clarify when performance declines might be mistaken for distributional shifts, among other things, and highlight the importance of diagnosing and correcting measurement error before drawing conclusions about model robustness, generalization, or variable relevance.

Keywords: Classical Measurement Error, Berkson Error, Misclassifications, Group-summary Assignment, Machine Learning, Monte Carlo Experiments, Statistical Analyses

1 Introduction

Measurement errors (ME) are known to cause bias in statistical risk estimates and reduce statistical power. For example, classical measurement error attenuates coefficient estimates in multiple linear regression. Berkson error has a similar impact in logistic regression. Furthermore, errors such as misclassification and group-based assignment, a type of Berkson error, can also introduce bias in risk estimates. These effects are well documented in the statistical literature; see, for instance, Buonaccorsi (2010); Fuller (2009); Grace et al. (2021); Keogh et al. (2020).

Measurements taken with uncalibrated instruments are prone to classical errors and misclassification. Often, self-reporting is influenced by recall bias, which contributes to these measurement errors. Although mislabeling in responses has been widely examined in machine learning, there is a significant gap in understanding how measurement errors, especially classical and Berkson errors, in input features affect machine learning.

In modern data analysis, many statistical and machine-learning workflows involve predictors that are not directly observed. Often, features are estimated from upstream models, combined across spatial, temporal, or

*Data Science, College of Arts and Sciences, Oberlin College, Oberlin, Ohio 44074, USA.

†School of Mathematical and Statistical Sciences, University of Texas Rio Grande Valley, Edinburg, Texas 78539, USA.

Corresponding author: Dr. Tamer Oraby, 1201 W University Dr, Edinburg, Texas 78539, USA. Tel: (956) 665-3536. Fax: (956) 665-5091. Email: tamer.oraby@utrgv.edu

group levels, or derived from proxies due to limited data access. This leads to the Berkson error and group summary assignment. Berkson error can arise when a feature is generated by a deterministic function or algorithm applied to covariates, leading to identical feature values for individuals with the same covariate profile. The real feature values, however, fluctuate around these assigned values. This structure is frequently used in practice, for example, when spatial exposure models (such as kriging of air pollution) generate a single exposure estimate for an entire ZIP code, which is then assigned to all residents in that area.

Most current research in machine learning emphasizes noise in the response variable, such as mislabeled outcomes, rather than systematic errors in covariates (see, e.g., Frenay and Verleysen (2014); Natarajan et al. (2013); Patrini et al. (2017)). Feature-level measurement error fundamentally differs from label noise because it corrupts the information available to the learning algorithm itself, rather than the supervision signal. Consequently, insights from the label-noise literature do not directly apply to situations where predictors are measured with error Carroll et al. (2006); Fuller (2009).

Measurement errors are prevalent in real-world datasets, yet systematic evaluations across various ML methods remain scarce. A few recent studies have explored the effect of feature measurement errors on machine learning models. For instance, Jiang et al. (2021) shows that misclassifying categorical predictors can significantly affect both predictive accuracy and variable importance in random forests, and suggests that quantitative bias analysis can help mitigate these effects. However, these studies usually focus on a single algorithm or error type and lack a comprehensive comparison across various models, error mechanisms, and prediction tasks.

In statistics, adjustment methods for measurement error depend on the type of measurement error and vary with the statistical model. The literature offers many techniques to correct for classical measurement error, such as regression calibration Luan et al. (2023), Simulation–Extrapolation (SIMEX) Fung and Krewski (1999), likelihood-based methods Keogh and Bartlett (2021); Shaw et al. (2020), and Hierarchical Bayesian approaches Bartlett and Keogh (2018); Gray (2018); Shaw et al. (2020). Fewer options are available for other ME types, like Hierarchical Bayesian techniques and likelihood-based methods for misclassifications Shaw et al. (2020), likelihood-based methods Oraby et al. (2018, 2023), and Hierarchical Bayesian techniques Shaw et al. (2020) for Berkson error and group summaries assignments, see also Kremers (2021). These methods are generally incompatible with machine learning because they rely on likelihood or on specific data models, do not scale well to high-dimensional data, and are consequently often ignored in practice. Additionally, many of these methods depend on strong parametric assumptions or validation data, which are seldom available in contemporary applied machine learning workflows involving complex, high-dimensional predictors.

Moreover, interpretability has become a key focus in applied machine learning. Measures of feature importance, like permutation importance, are commonly used to explain model behavior, inform scientific conclusions, and aid decision-making Breiman (2001); Fisher et al. (2019); Molnar (2022). Nevertheless, little is known about how measurement error in input features biases these importance metrics, even when models are optimized for best predictive accuracy. One can contrast this with the known effects of attenuation towards zero for classical and Berkson errors in generalized linear models (GLMs). Therefore, understanding how various error mechanisms affect both accuracy and interpretability is crucial for the responsible use of machine learning on error-prone data.

This paper investigates how four types of feature measurement errors, classical, Berkson, misclassification, and group summary assignment errors, impact five widely used machine learning methods: multiple linear and logistic regression, random forests, support vector machines, XGBoost, and neural networks for regression and classification. We focus on non-differential measurement errors in the features that do not depend statistically on the response. All the ML models are tuned with hyperparameters to achieve optimal performance. Researchers often ignore these errors, assuming that large datasets can offset their effects. We conducted four Monte Carlo experiments using a four-factor design to assess how measurement error magnitude affects model performance and dataset sizes across the five ML approaches.

In particular, we use a four-factor $3 \times 2 \times 3 \times 5$ factorial design to assess how error magnitude, signal structure, sample size, and method choice interact to affect model performance and feature importance. For each ML problem type and measurement error category, we conduct both descriptive and inferential statistical analyses of the Monte Carlo factorial design simulation results. Since simulations can be time-consuming, we

also conducted a post hoc power analysis based on the number of Monte Carlo replications used.

The remainder of the paper is organized as follows. We first describe the simulation framework and evaluation metrics used to assess how ME affects ML predictive accuracy and feature importance. The results are divided into regression and classification tasks. Each section contains descriptive and inferential analyses, grouped by measurement error type. The discussion and conclusion follow the results, where we summarize and emphasize the main findings of our study and recommend future research directions.

2 Methods

In this section, we describe methods to test the impact of different types of measurement errors on machine learning performance. This study used simulated data to examine the effects of classical error (CE), misclassification (MsC), Berkson error (BE), and group-summary assignment (GSA) on both regression and classification tasks in machine learning (ML). We explored the effects of those measurement errors at two levels ($\sigma = 0.5, 0.75$) and two levels ($q = 0.1, 0.3$), each using five commonly used ML methods for regression and classification. Additionally, we employed relevant performance measurement techniques. The results will clearly show how the effects can be extrapolated and interpolated across different levels of σ and q .

2.1 Machine Learning Methods

For regression problems, we examine the effect of CE, MsC, BE, and GSA on the performance of multiple linear regression (MLR), support vector regression (SVR), random forest regression (RF-R), gradient boosting regression (XGBoost-R), and multi-layer perceptron regression (MLP-R). Model performance is evaluated using the coefficient of determination (R^2) and the Loss function, the mean squared error (MSE).

For classification problems, we examined the impact of CE, MsC, BE, and GSA on the performance of logistic regression (LogR), support vector machine (SVM), random forest classifier (RF-C), gradient-boosting classifier (XGBoost-C), and multilayer perceptron classifier (MLP-C). We evaluated their performance using Accuracy and F1-score. We calibrated the hyperparameters of each ML method at each dataset. See Tables S1 and S2 for the list of hyperparameters and the ranges of values we considered. We also studied the effects of CE, MsC, BE, and GSA on feature permutation importance relative to the NE feature.

2.2 Monte Carlo Simulation Experiments

We conducted a Monte Carlo simulation experiment (MCE) to evaluate the effects of classical error (CE), misclassification (MsC), Berkson error (BE), and group-summary assignment (GSA) relative to no-error (NE) on various machine learning methods. In the MCE, we used three simulated features. Two of these features were used to generate the response, whereas the third is non-predictive. During ML training, one of the first two variables was replaced with its error-affected version. We assessed the performance of the ML methods on both the data with the original three features and the data affected by errors. In each case, we tuned the hyperparameters to achieve optimal performance.

We used three data sizes, $n \in \{2000, 3000, 4000\}$, in the MCEs. We performed three MCEs for each type of error. In each one, we generated two variables Z_1 and Z_2 independently using a Normal(0, 1) probability distribution. In all MCEs, we assume that Z_2 is not a predictive feature; thus, we do not use it in the calculation of the response Y .

In the first MC experiment, we generated the classical error, which is randomly scattered around the actual measurements using an additive classical error model.

Experiment 1: (CE)

We generated an error-free feature X independently of Z_1 and Z_2 using a Normal(0, 1) probability distribution. We then generated the classical error-burdened feature W using

$$W_i = X_i + \epsilon_i$$

such that $\epsilon_i \sim N(0, \sigma)$, independent of X , for $i = 1, 2, \dots, n$. We examine the effect of $\sigma = 0.5$ and 0.75 on the performance of ML methods. Notice that the level $\sigma = 0$ is equivalent to using the error-free feature X .

In the second MC experiment, we studied how different levels of misclassification in a categorical feature affect the predictive performance of machine learning methods.

Experiment 2: (MsC)

We generate an error-free feature X independently of Z_1 and Z_2 using a Bernoulli(0.5) probability distribution, representing a binary feature. We then generate a misclassification feature W using

$$W_i = (1 - \delta_i)X_i + \delta_i(1 - X_i)$$

where $\delta_i \sim \text{Bernoulli}(q)$, independent of X , for $i = 1, 2, \dots, n$. We examine the effect of $q = 0.1$ and 0.3 on the performance of ML methods. Note that the level $q = 0$ corresponds to using the misclassification-free feature X .

In the third MC experiment, we investigated the effect of Berkson error (BE) on the predictive performance of machine learning methods. In the BE model, the actual measurements are randomly distributed around the Berkson-error-affected measurements. We use a normal distribution to model the additive Berkson error.

Experiment 3: (BE)

We generate the underlying Berkson error burdened feature W independently of Z_1 and Z_2 using a Normal(0, 1) probability distribution. We then generate the actual feature X using

$$X_i = W_i + \epsilon_i$$

such that $\epsilon_i \sim N(0, \sigma)$, independent of W , for $i = 1, 2, \dots, n$. We examine the effect of $\sigma = 0.5$ and 0.75 on the performance of ML methods. Notice that the level $\sigma = 0$ is equivalent to using the error-free feature X .

We also conduct an additional MCE on the Berkson error using group-summary assignments (GSA). We mainly choose the mean as the summary statistic. (In the simulation study, it is also the median and the mode since we use a normal distribution for generating the actual measurement around the summary statistic.) We assume that units belong to g distinct groups. Each group has a mean value μ_k , for $k = 1, 2, \dots, g$. The values of actual exposures are randomly scattered around the groups' means μ_k . We examine the effect of using the group-summary assignment for training ML methods.

Experiment 4: (GSA)

We generate mean values μ_k for $k = 1, 2, \dots, g$ using a Normal(0, 1) probability distribution. (We choose $g = 5$.) Then we do group assignments to the error burdened feature W by setting $W_i = \mu_G$ where $G \sim \text{Discrete Uniform}(1, 2, \dots, g)$, for $i = 1, 2, \dots, n$. We independently generate Z_1 and Z_2 using a Normal(0, 1) probability distribution. Next, we generate the actual feature X using

$$X_i = W_i + \epsilon_i$$

such that $\epsilon_i \sim N(0, \sigma)$, independent of W , for $i = 1, 2, \dots, n$. We examine the effect of $\sigma = 0.5$ and 0.75 on the performance of ML methods. Notice that the level $\sigma = 0$ is equivalent to using the error-free feature X . In this case, X could be seen as an ordinal variable.

After each one of those four types of feature generation, we apply the following linear functional

$$L(X, Z_1, Z_2) = \alpha_0 + \alpha_1 X + \alpha_2 Z_1 + \alpha_3 Z_2.$$

See Table1 for list of selected values of $\alpha_0, \alpha_1, \alpha_2$, and α_3 , called Trials, T1–T6. There are three levels of σ ($= 0, 0.5, 0.75$) and q ($= 0, 0.1, 0.3$).

In the regression problems, we produce a continuous response Y using

$$Y = L(X, Z_1, Z_2) + \epsilon$$

where $\epsilon \sim N(0, 0.1)$, independent of all of the random variables mentioned above.

In the classification problems, we produce a binary response Y by generating random binary numbers using Bernoulli(p) probability distribution such that

$$\text{logit}(p) = L(X, Z_1, Z_2).$$

We compare the performance of ML methods in the four experiments to the no-error (NE) case in which X, Z_1, Z_2 are used to predict Y .

2.3 MC Simulation Design

Different experimental trials T1–T6 for $(\alpha_0, \alpha_1, \alpha_2, \alpha_3)$ are given in Table 1. In each of the ML problem types, for all values of $n \in \{2000, 3000, 4000\}$ and for every one of the six trials in Table 1, we replicate the simulations for $m = 10$ times, see below for the post-hoc statistical power calculation for m .

Table 1: Input parameter values for the Monte Carlo simulation experiments.

Trial	α_0	α_1	α_2	α_3
T1	0	2	1	0
T2	0	1	2	0
T3	0	2	2	0
T4	1	2	1	0
T5	1	1	2	0
T6	1	2	2	0

We use a standardization transformation to normalize the variables X, W, Z_1, Z_2 , and Y , except when the variables are binary, as a usual step in most of the ML exercises. In each simulation m , we use an 80–20 training-testing data division rule across the same subjects of X, W, Z_1 , and Z_2 , for uniformity. For classification, to avoid the effects of imbalanced data on the split, we stratify the split by the binary output Y . First, we train the ML methods on the training set X, Z_1, Z_2 , and Y , and then evaluate predictive performance on the test set. Then, we repeat the same procedure using W, Z_1, Z_2 and Y .

Each time we implement an ML method, we optimize its hyperparameters to ensure the best performance for each use case X and W across all simulated datasets. Tables S1 and S2 show the hyperparameter list, which includes descriptions and value ranges used during calibration. For hyperparameter tuning, we use the Python tree search library Optuna, running 20 trials with the median pruner for efficient sampling and pruning. During each ML application, we compute permutation feature importance and compare it to the NE’s feature importance, performing 10 permutation repetitions. To compare performance between cases with Error (W) and without Error (X), we analyze the changes in predictive performance metrics.

In the case of CE and MsC, we generate feature W from feature X at two different levels of σ or q , and label these as W_1 and W_2 . For BE and GSA, we use a feature W to produce two features, X_1 and X_2 , each yielding a response (Y_1, Y_2). We then predict each response in separate runs using the same W , denoted W_1 and W_2 for simplicity, although they are identical.

To quantify changes in performance due to measurement error, we use the difference of the performance metric M . In the following, $M(X)$ denotes the performance measure when using X, Z_1 , and Z_2 to predict Y , and so on. For single-baseline settings (CE, MsC),

$$\Delta M(W_k; X) = M(W_k) - M(X), \quad k = 1, 2.$$

For the two-baseline settings (BE, GSA),

$$\Delta M(W_k; X_k) = M(W_k) - M(X_k), \quad k = 1, 2.$$

For instance, a negative ΔR^2 (or positive ΔMSE) indicates degraded performance under measurement error.

Statistical Design of Experiment and Analysis

Statistical Inference. We examined the impact of measurement error using a four-factor design with a $3 \times 2 \times 3 \times 5$ factorial structure. The factors included sample size n (3 levels), intercept α_0 (2 levels), coefficients α_i (3 levels), and machine learning methods (5). For each combination, we ran $m = 10$ Monte Carlo simulations and assessed performance under both error-free and error-prone measurement conditions. The *Measurement* factor, within-subjects, had three levels for CE and MsC (X, W_1, W_2) and two levels for BE and GSA (X_k, W_k) for each $k = 1, 2$. Regression outcomes included R^2 and MSE, while classification outcomes included accuracy and F1-score.

For inference, we fitted repeated-measures ANOVA models of the form

$$\text{Outcome} \sim n \times \text{Trial} \times \text{ML} + \text{Error}(\text{Subject}/\text{Measurement}),$$

where each unique combination of sample size $n \in \{2000, 3000, 4000\}$, Trial (T1–T6), ML method, and replication defined a *Subject*, and *Measurement* was the within-subject factor. We used Type-III sums of squares, Greenhouse–Geisser sphericity corrections, and reported generalized eta-squared (η^2_{ges}) as the effect size. To localize within-subject effects, we calculated Estimated Marginal Means (EMMs) for each measurement level within each factorial cell. We tested treatment-versus-control contrasts (e.g., W_k vs. X , or W_k vs. X_k) with Holm adjustment for multiplicity. Exploratory visualization employed line graphs of within-subject trajectories ($X \rightarrow W_1 \rightarrow W_2$) for CE and MsC as well as ($X_k \rightarrow W_k$ for $k = 1, 2$) for BE and GSA by trial, showing replicate paths and cell means.

Post-hoc Power Analysis. We used G*Power (Faul et al., 2007) to calculate the power of the statistical tests using the number of Monte Carlo replications m equals to 10 and significance level of $\alpha = 0.05$ targeting adequate power for the Measurement main effect and key interactions.

Software and Computing. Data generation, model fitting, and hyperparameter tuning were implemented in Python (Optuna (Akiba et al., 2019), scikit-learn (Pedregosa et al., 2011), multiprocessing (Foundation, 2008)) and executed on 56 CPU cores on a single Lonestar6 node (TACC, 2021). Statistical analyses (repeated-measures ANOVA, EMMs, and contrasts) were performed in R using **afex** (Singmann et al., 2023) and **emmeans** (Lenth, 2024). Reproducible scripts are available upon request.

3 Results

We divide the results into regression and classification problems. For each problem type, we present a descriptive analysis followed by statistical inference for the four experiments.

Recall that we quantify changes in performance due to measurement error using the difference of a performance measure M . For regression problems, $M \in \{R^2, \text{MSE}\}$, whereas for classification problems $M \in \{\text{Accuracy}, \text{F1 score}\}$. A negative ΔR^2 , $\Delta \text{Accuracy}$, or $\Delta \text{F1 score}$ (or a positive ΔMSE) indicates degraded performance under measurement error. For Classical Error (CE) and Misclassification (MsC), which involve a single error-free baseline X , we define $\Delta M(W_k; X) = M(W_k) - M(X)$ for $k = 1, 2$. For Berkson Error (BE) and Group-Summary Assignment (GSA), which involve two error-free baselines X_k , we define $\Delta M(W_k; X_k) = M(W_k) - M(X_k)$ for $k = 1, 2$.

3.1 Regression Problems

This section presents the results of both descriptive and inferential analyses of the four MCEs for regression tasks. In regression problems, the performance measures M are R^2 , MSE, and permutation feature importance.

3.1.1 Descriptive Analysis

Classical Error (CE). The degradation scales with the error level, with W_2 being more harmful than W_1 . We observed that increasing the sample size n primarily reduces variability between the m 's replications without altering the degradation patterns.

Figure 1 illustrates dense, mainly downward trajectories within subjects from X to W_1 (with $\sigma = 0.5$), and then to W_2 (with $\sigma = 0.75$), when the sample size n reaches 4000 across all models and trials. The bolded mean paths indicate a monotone decreasing trend, whereas model-specific differences cause slight shifts in absolute levels but do not affect the overall pattern observed across repetitions. Notably, trials T1 and T4 exhibit the largest effects, followed by T3 and T6, because X contributes twice as much to Y as in T2 and T5. The error-free feature Z_1 contributes equally in the last two trial pairs. It appears from these three pairs that both error-related and error-free contributions to Y influence performance degradation. See Figure A.1 in the Supplementary Material for similar plots and conclusions related to the MSE.

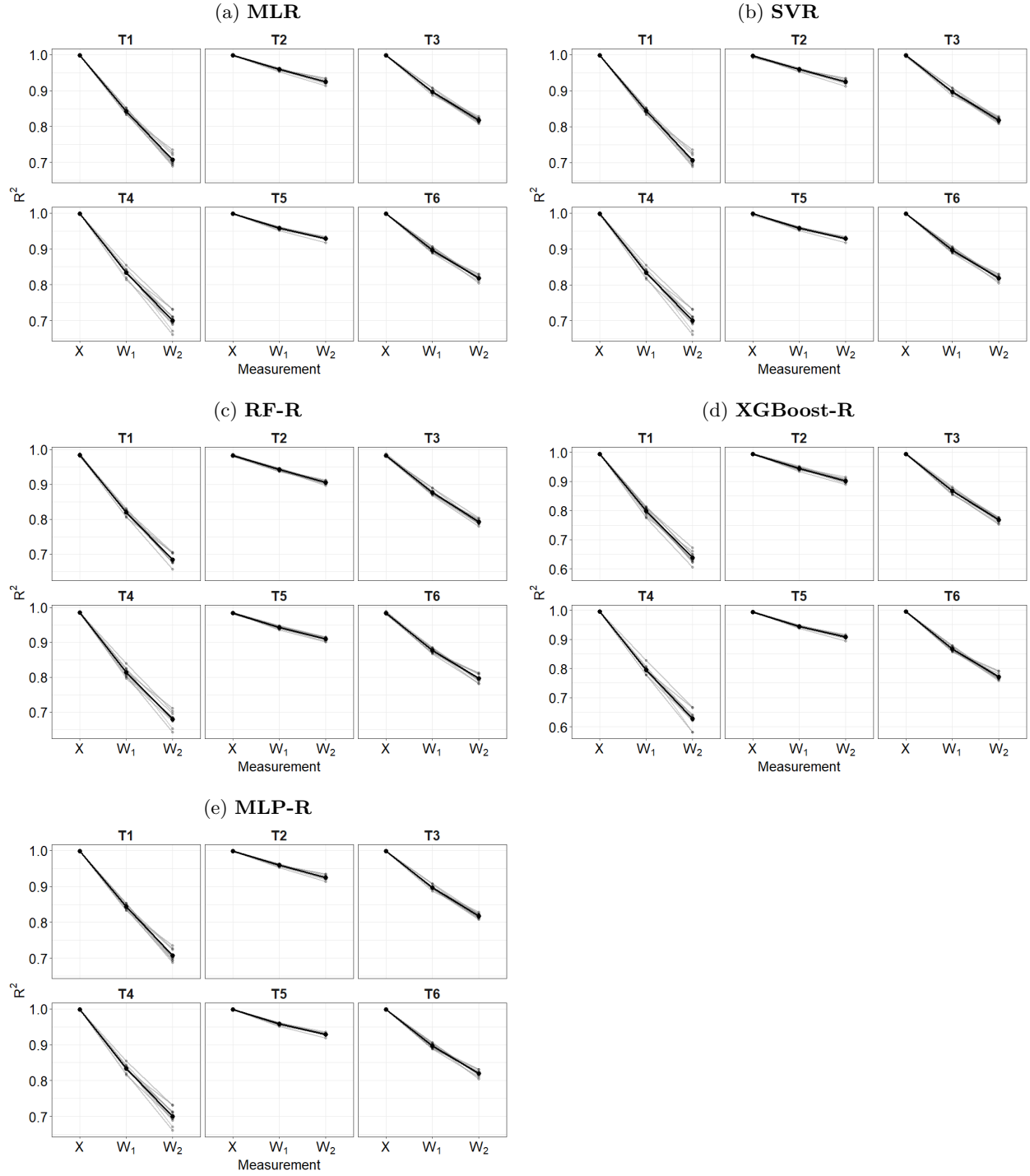


Figure 1: Line graphs of R^2 for the MCE of CE with $\sigma = 0, 0.5, 0.75$ at trials T1–T6 and for (a) MLR, (b) SVR, (c) RF-R, (d) XGBoost-R, and (e) MLP-R.

Figure 2 visualizes the permutation importance of the main predictor X , the predictive feature Z_1 , and the non-predictive feature Z_2 under CE across six trials T1–T6 and five ML methods. Generally, X 's importance

decreases steadily as the magnitude of the classical measurement error increases, with the highest importance
 in the error-free scenario (X), followed by lower importance levels at increased error levels (W_1, W_2). This
 trend is consistent across different trials and models. Meanwhile, the importance of Z_1 stays stable, and Z_2 's
 importance remains near zero in all cases. Trials T2 and T5, where Z_1 has a larger coefficient α than X ,
 Z_1 shows higher permutation importance than the primary predictor, even without error. The error bars
 in the Figure 2, shown as $\pm$ SD, where SD is the standard deviation, indicate high precision of the results.
 Overall, CE reduces the impact of the affected features, and the relative importance of features depends on
 the coefficient setup in each trial.

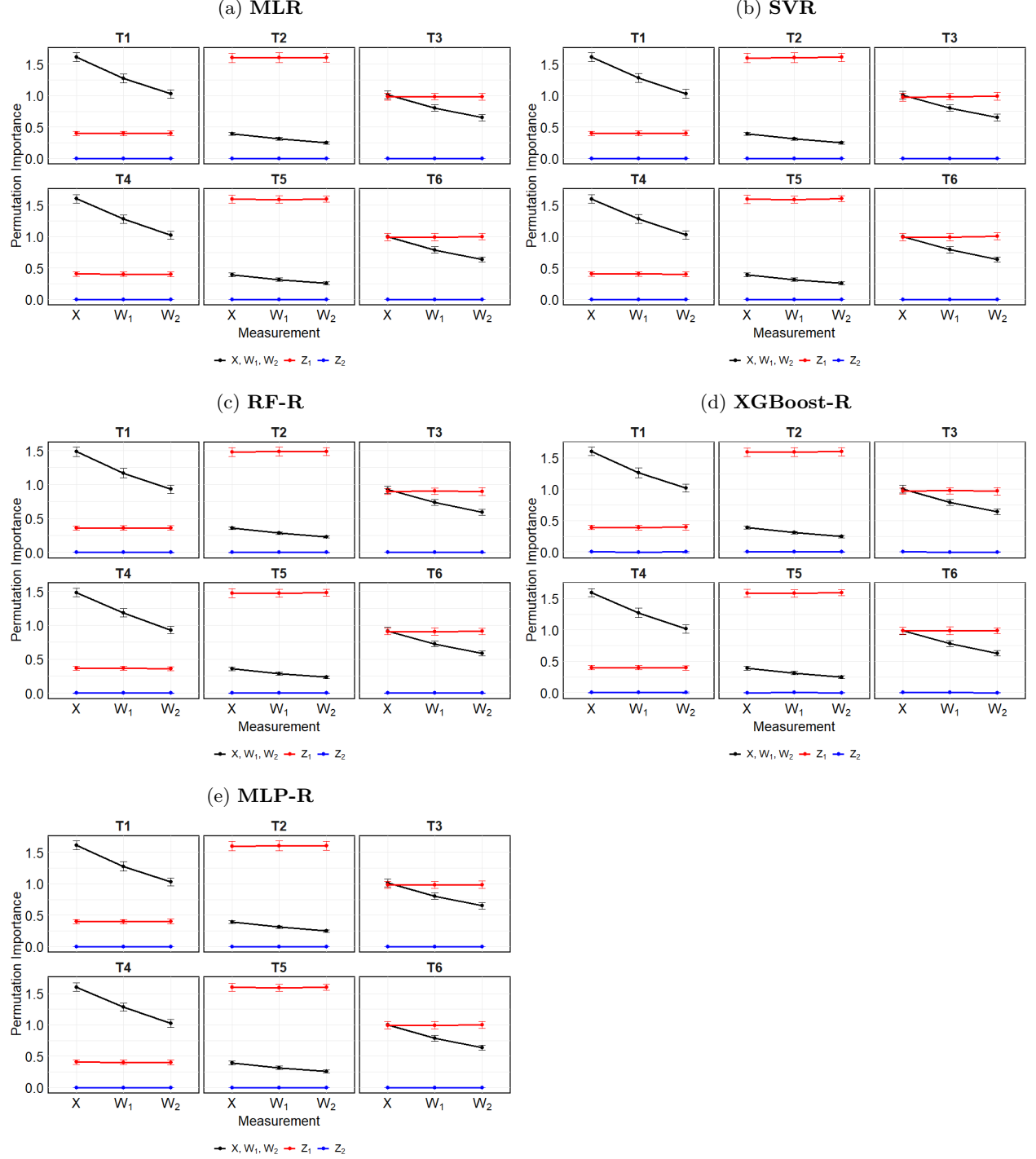


Figure 2: Line graphs of the mean permutation importance ($\pm$ SD) for the MCE of CE with $\sigma = 0, 0.5, 0.75$ at trials T1–T6 and for (a) MLR, (b) SVR, (c) RF-R, (d) XGBoost-R, and (e) MLP-R.

Misclassification (MsC) Under misclassification, the difference $\Delta R^2(W_k; X) = R^2(W_k) - R^2(X)$ were mostly negative across trials (T1–T6), models, and sample sizes, with W_2 (with $q = 0.30$) showing larger losses

than W_1 (with $q = 0.10$). Conversely, $\Delta\text{MSE}(W_k; X) = \text{MSE}(W_k) - \text{MSE}(X)$ was mainly positive and also increased from W_1 to W_2 . While the magnitudes slightly varied by model, the general order $W_2 < W_1 < X$ for R^2 (and X followed by W_1 then W_2 for MSE) remained consistent.

Increasing the sample size from $n = 2000$ to 4000 primarily reduced the spread of both metrics, R^2 and MSE, rather than altering their means. Figure 3 illustrates clear, stepwise drops from the baseline to W_1 ($q = 0.10$) and further to W_2 ($q = 0.30$) within each trial at the largest sample size $n = 4000$, with boldfaced lines between means trajectories emphasizing this consistent monotone trend.

Like CE, trials T1 and T4 show the strongest effects, followed by T3 and T6. This is because X influences Y twice as much in these trials compared to T2 and T5, while Z_1 's contribution remains consistent across the last two pairs. Nonetheless, the decline in predictive performance for T1 and T4 in the MsC seems more significant than in the CE. See Figure A.2 in the Supplementary Material for similar plots related to the MSE.

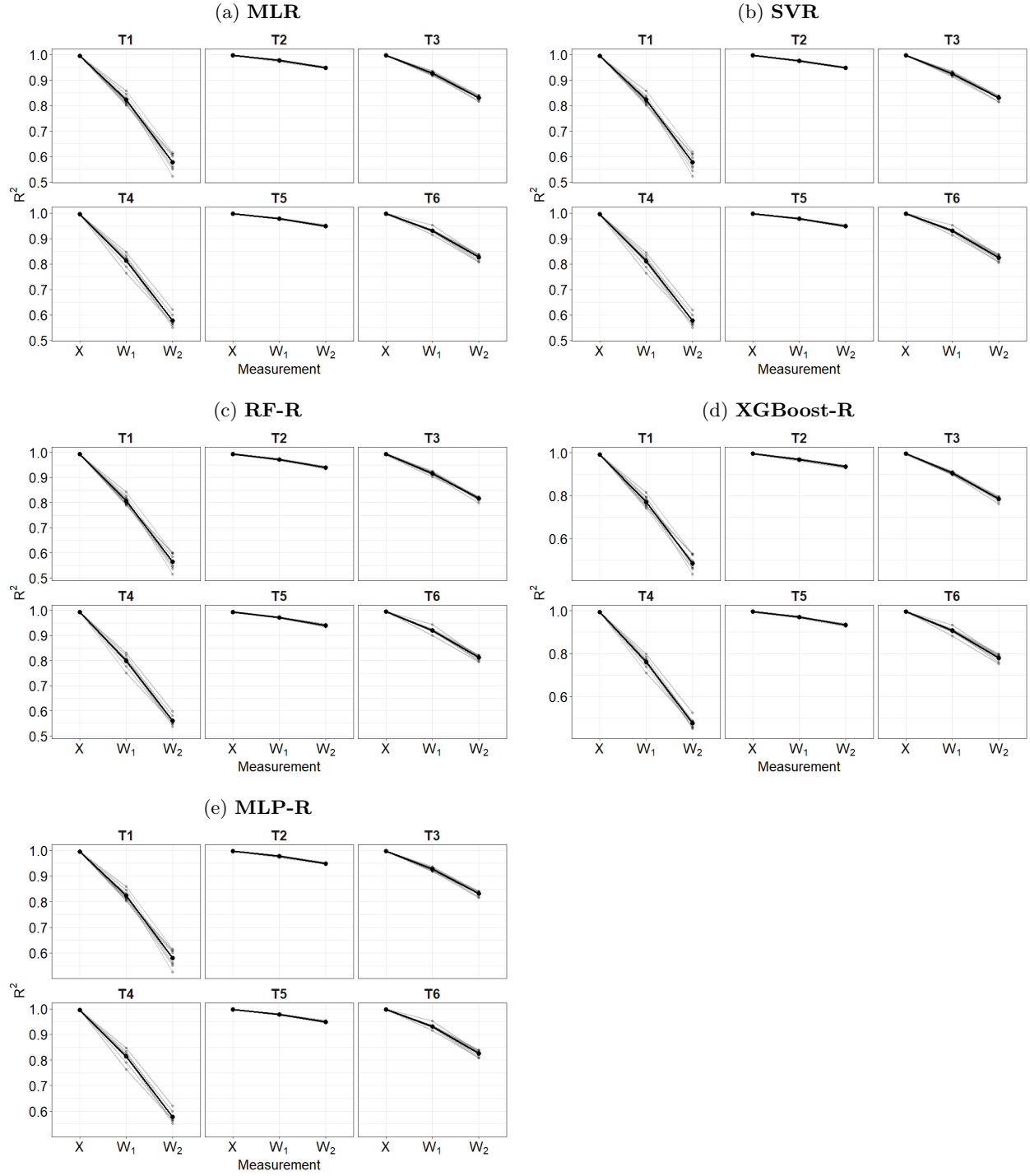


Figure 3: Line graphs of R^2 for the MCE of MsC with $q = 0, 0.1, 0.3$ at trials T1–T6 and for (a) MLR, (b) SVR, (c) RF-R, (d) XGBoost-R, and (e) MLP-R.

Figure 4 illustrates how permutation importance behaves in the MsC experiment. Across all trials and models, MsC leads to a noticeable decrease in the importance of the affected predictor as the misclassification

rate rises. In several trials, especially those where we assign a larger coefficient to Z_1 than the primary predictor, the importance of the misclassified variable nears zero at higher misclassification levels (W_2). However, the importance of the predictive feature Z_1 remains stable regardless of misclassification levels and consistently ranks higher in trials where it has a stronger true signal. Meanwhile, the non-predictive variable Z_2 stays nearly at zero throughout.

These patterns also emphasize key differences between the CE and MsC. The CE causes a slow decline in the affected predictor’s importance, whereas the MsC leads to quicker, more severe losses and a shift in the importance to other features, as illustrated in trials T1, T3, T4, and T6 in Figure 4.

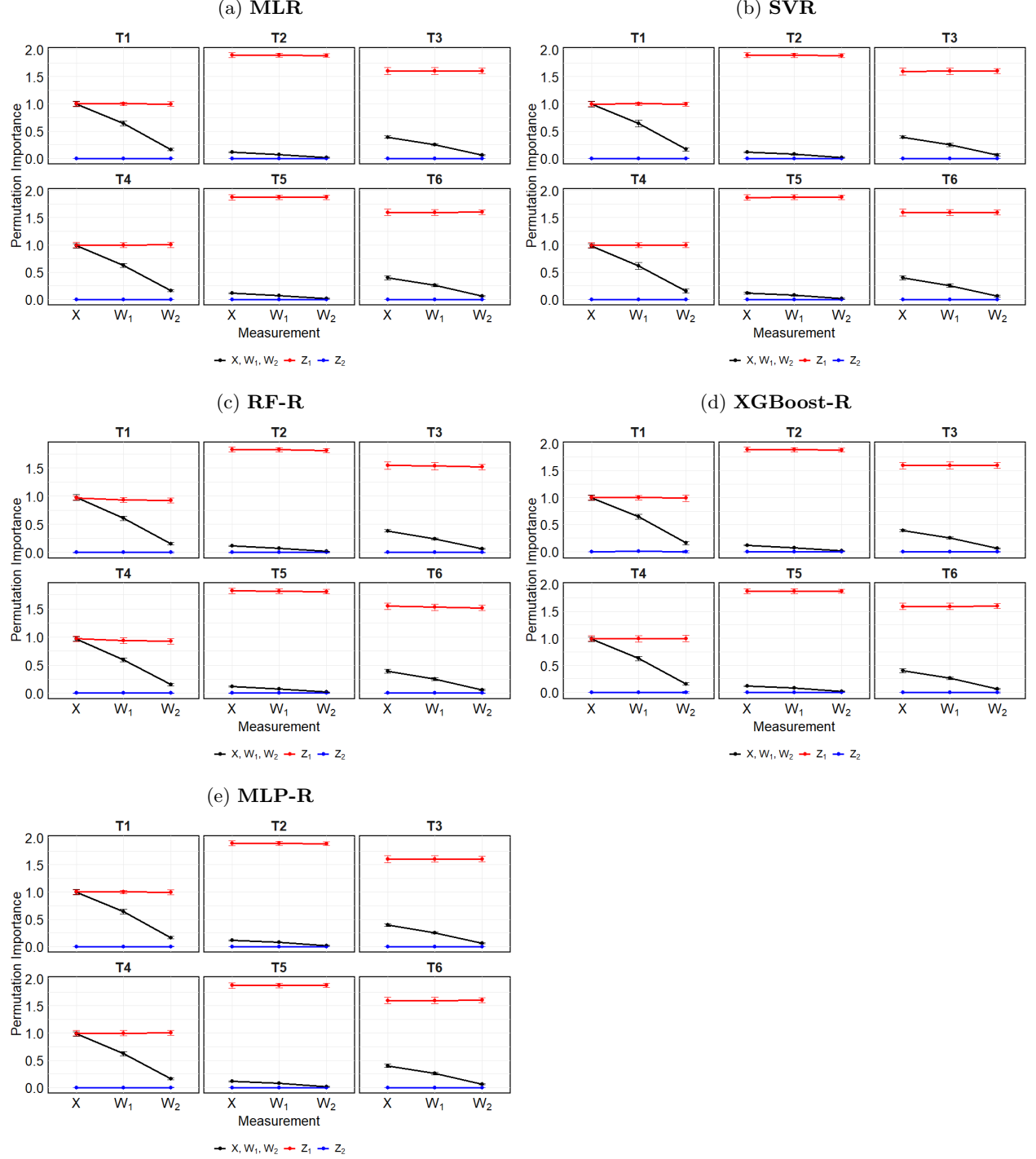


Figure 4: Line graphs of the mean permutation importance ($\pm$ SD) for the MCE of MsC with $q = 0, 0.1, 0.3$ at trials T1–T6 and for (a) MLR, (b) SVR, (c) RF-R, (d) XGBoost-R, and (e) MLP-R.

Berkson Error (BE) Under Berkson error, the difference $\Delta R^2(W_k; X_k) = R^2(W_k) - R^2(X_k)$ are consistently negative, and $\Delta \text{MSE}(W_k; X_k) = \text{MSE}(W_k) - \text{MSE}(X_k)$, are consistently positive across trials (T1–T6),

models, and sample sizes. This indicates that replacing baseline features with error-prone measurements W_1 ($\sigma = 0.5$) and W_2 ($\sigma = 0.75$) reduces the explained variance and increases the MSE. The extent of these effects depends on the error levels, with W_2 causing greater degradation than W_1 . Additionally, the choice between the baseline features (X_1 and X_2) has little to no effect.

In Figure 5, for $n = 4000$ and across all models and trials, the two baselines (X_1 , X_2) are visually overlaid, with the boldfaced lines between means and individual trajectories nearly overlapping. Any minor separations are infrequent and slight, and do not affect the main conclusion: increased Berkson error results in larger declines in R^2 and corresponds to increases in MSE. The pattern observed in trials T1–T6, similar to CE and MsC, also appears in BE; the predictive performance depends on the contributions of the error-free and error-burdened features to Y . See Figure A.3 in the Supplementary Material for similar plots related to the MSE.

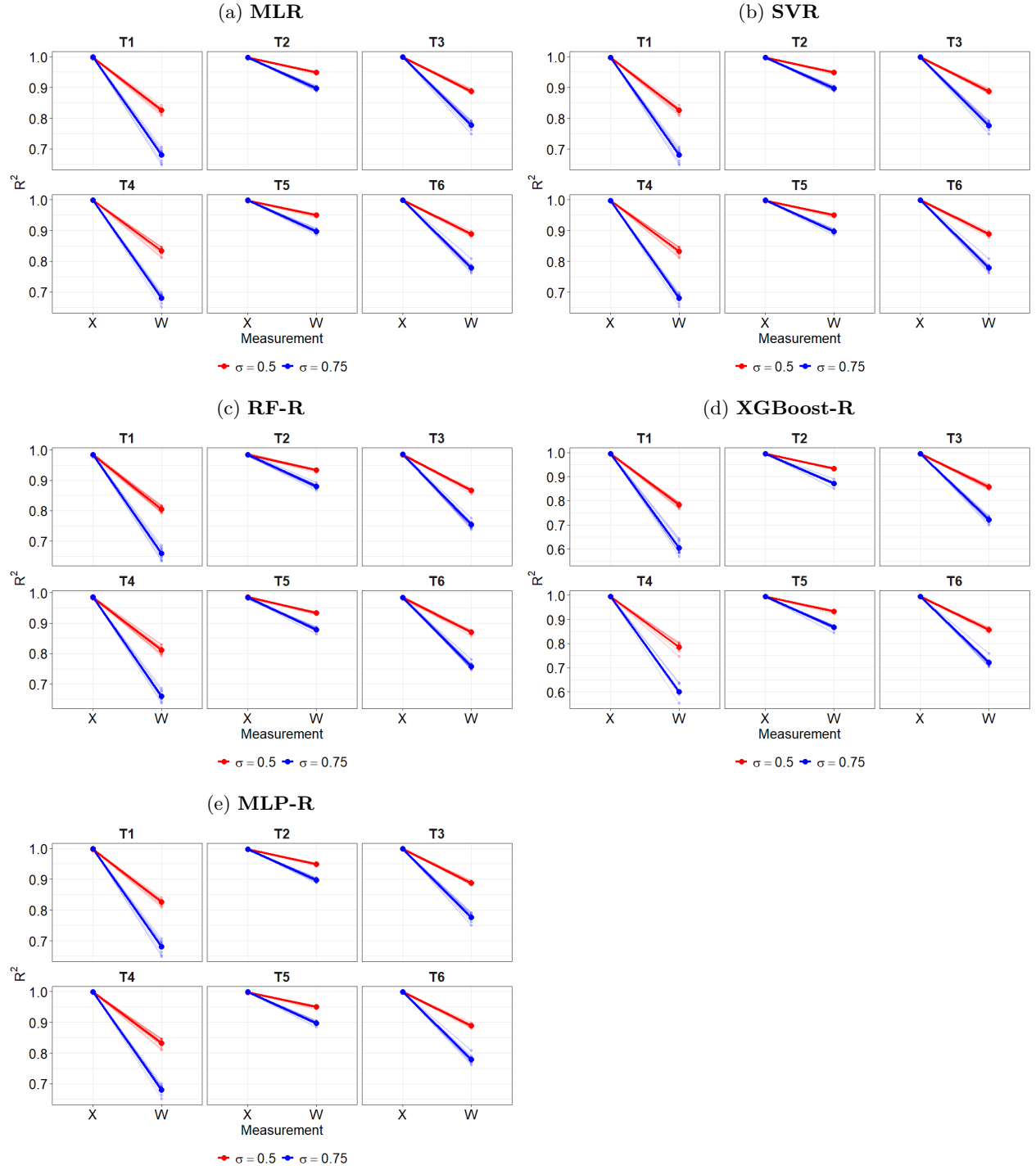


Figure 5: Line graphs of R^2 for the MCE of BE with $\sigma = 0, 0.5, 0.75$ at trials T1-T6 and for (a) MLR, (b) SVR, (c) RF-R, (d) XGBoost-R, and (e) MLP-R.

The permutation importance under BE, Figure 6, shows a high degree of stability across trials and models, with only modest changes observed when moving from the error-free measurement X to the BE-affected

281 measurement W , even at higher error levels. Trials T1 and T4 were the most affected cases. In trials where
282 the primary predictor carries a clearly weaker signal than the predictive feature Z_1 (e.g., T2 and T5), relative
283 importance rankings are largely preserved, and the separation between predictors remains similar to the
284 no-error case. However, an informative pattern emerges in Trials T3 and T6, where the primary predictor
285 and Z_1 have comparable true effects: under BE, their permutation importance remains close and tends
286 to converge slightly without pronounced reordering. This behavior contrasts with the stronger reduction
287 observed under CE and the sharper degradation under MsC, and it highlights that when predictive signals
288 are balanced, BE primarily compresses relative importance rather than redistributing it substantially.

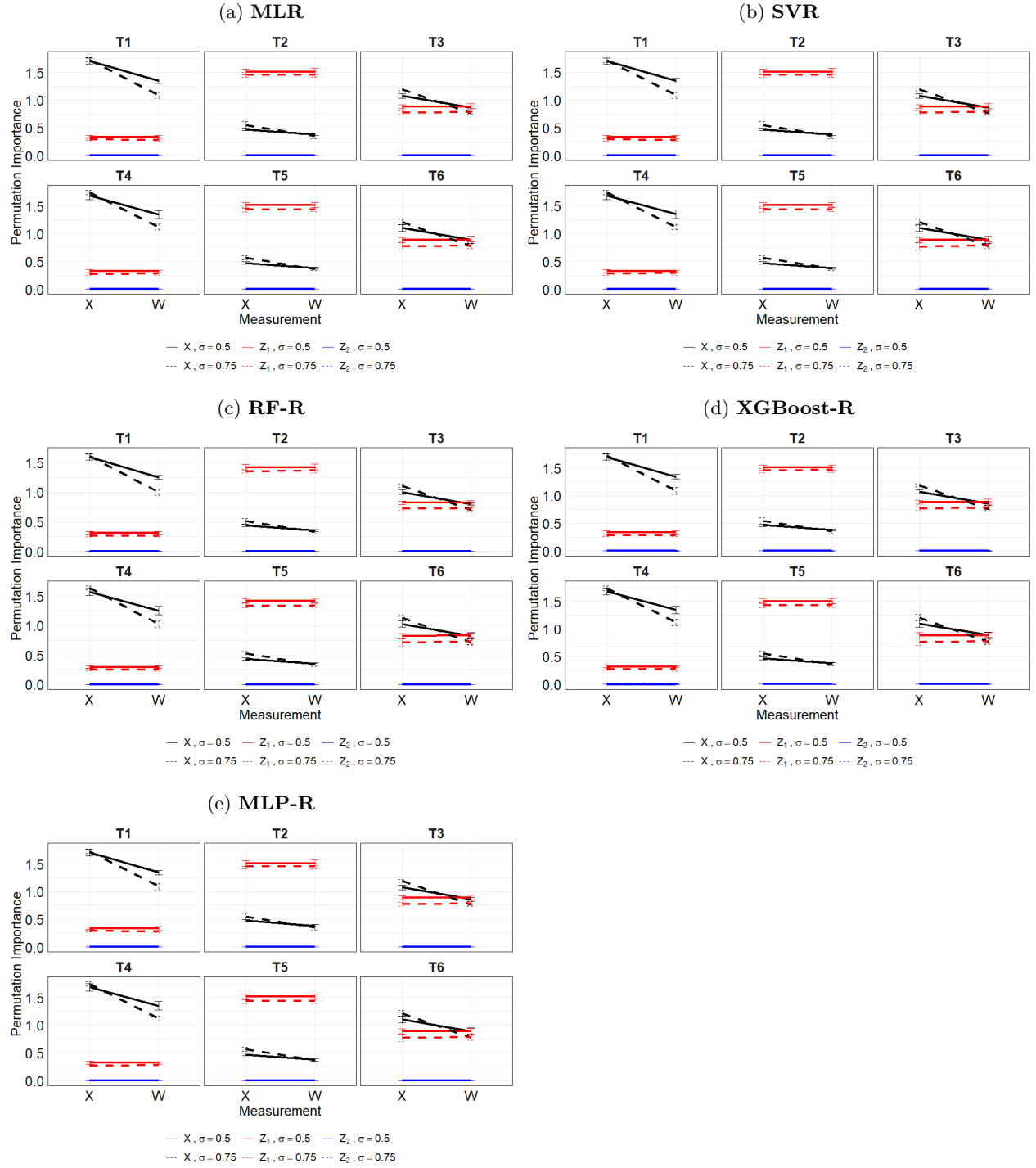


Figure 6: Line graphs of the mean permutation importance ($\pm$ SD) for the MCE of BE with $\sigma = 0, 0.5, 0.75$ at trials T1–T6 and for (a) MLR, (b) SVR, (c) RF-R, (d) XGBoost-R, and (e) MLP-R.

Group-Summary Assignment (GSA) In the group-summary assignment (GSA), the sign of the difference remains consistent across trials (T1–T6), models, and sample sizes: $\Delta R^2(W_k; X_k) = R^2(W_k) - R^2(X_k)$ are

consistently negative, while $\Delta\text{MSE}(W_k; X_k) = \text{MSE}(W_k) - \text{MSE}(X_k)$ are consistently positive, confirming that using error-burdened measurements consistently reduce model fit and worsen prediction performance compared to baselines (X_1 or X_2). The extent of this degradation increases with error level, with W_2 ($\sigma = 0.75$) causing larger changes than W_1 ($\sigma = 0.5$). The downward trend of MSE when using W_2 is more severe than that of W_1 . Increasing the sample size to $n = 4000$ reduces variability primarily, without significantly shifting the mean, although it remains larger than in CE, MsC, and BE.

In line with these summaries, Figure 7 ($n = 4000$; all models and trials) displays within-cell trajectories from X_k to W_k , for $k = 1, 2$, that decrease in R^2 , with the two baseline mean paths almost overlapping, suggesting only minor baseline differences. A similar pattern is observed for differences and similarities between trials, as seen in the CE, MsC, and BE, though potentially with greater degradation across replications. See Figure A.4 in the Supplementary Material for similar plots related to the MSE.

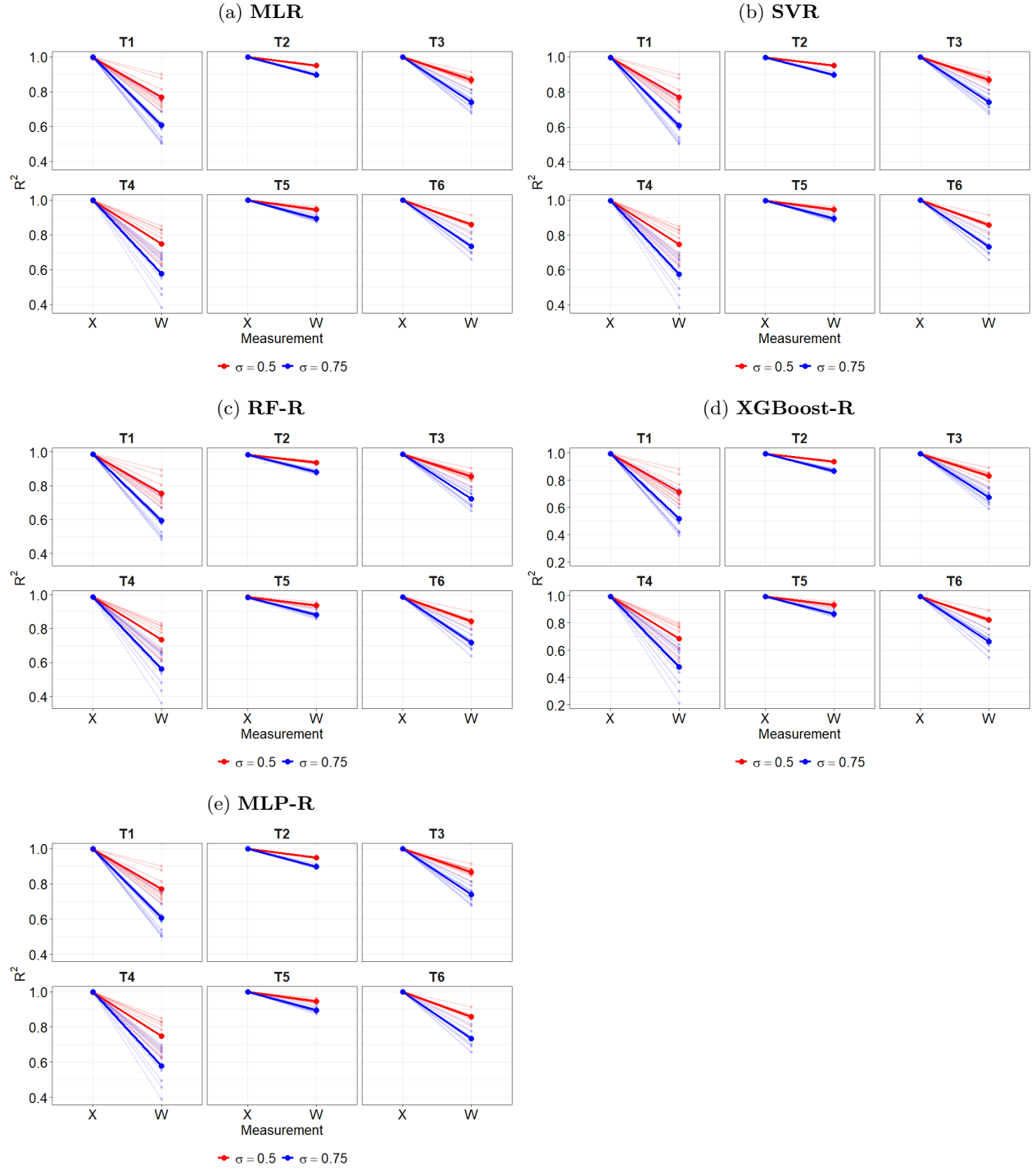


Figure 7: Line graphs of R^2 for the MCE of GSA with $\sigma = 0, 0.5, 0.75$ at trials T1–T6 and for (a) MLR, (b) SVR, (c) RF-R, (d) XGBoost-R, and (e) MLP-R.

Figure 8 illustrates the permutation importance pattern observed under GSA across all trials and models. Substituting predictor X with its group-level summary, W, results in a noticeable decrease in importance,

304 indicating the removal of within-group variability rather than just noise. As the error level increases, the
305 importance of the primary predictor diminishes further, although the magnitude of this decrease varies across
306 trials and does not follow a consistent pattern. The importance of Z_1 remains relatively unchanged, while Z_2
307 consistently shows negligible importance. Unlike CE and MsC, GSA primarily influences variable importance
308 by eliminating within-group variation by replacing individual measurements with group summaries.

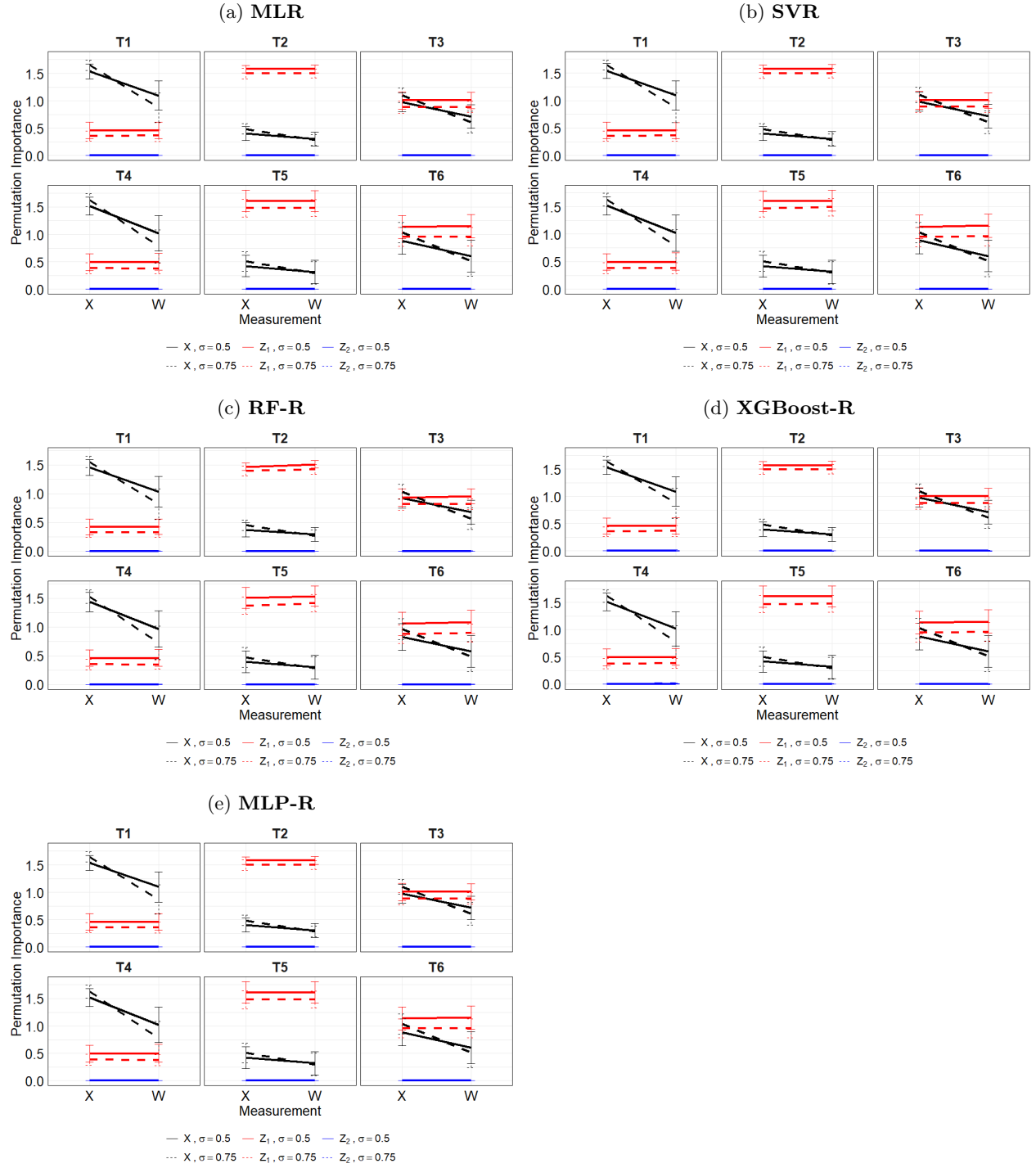


Figure 8: Line graphs of the mean permutation importance ($\pm$ SD) for the MCE of GSA with $\sigma = 0, 0.5, 0.75$ at trials T1–T6 and for (a) MLR, (b) SVR, (c) RF-R, (d) XGBoost-R, and (e) MLP-R.

3.1.2 Statistical Analysis

Classical Error (CE). The analysis indicates that the within-subject Measurement factor ($X \rightarrow W_1 \rightarrow W_2$) predominantly influences R^2 and MSE, with significant effects (R^2 : $F = 82,195.33$; MSE: $F = 102,708.26$; $p < .001$). This suggests that introducing classical error significantly reduces R^2 and increases prediction error, as reflected in higher MSE values. The extent of this impact varies with different scenarios: the interactions between Trial and Measurement are also substantial (R^2 : $F = 3,952.64$; MSE: $F = 4,934.96$; $p < .001$), indicating that deterioration levels depend heavily on the data-generating conditions. Additionally, model architecture influences results: the Model $\times$ Measurement interaction is significant for R^2 ($F = 214.43$, $p < .001$) and somewhat less for MSE ($F = 9.21$, $p < .001$), suggesting some models are more resistant to classical error than others.

The between-subject factors also play a role: Trial, Model, and sample size (n) are all significant main effects for both metrics (e.g., R^2 : Trial $F = 7,829.01$; Model: $F = 730.54$; n : $F = 15.26$, all $p < .001$), indicating baseline differences across scenarios, models, and sample sizes. However, the interactions between $n \times$ Measurement are minor (R^2 : $F = 1.92$, $p = .118$; MSE: $F = 2.84$, $p = .035$), implying that larger samples mainly reduce variability rather than change the average CE impact. Overall, CE shows a consistent, monotonic decline from X to W_1 to W_2 , with the size of this decline primarily influenced by the Trial and the learning model, with only slight effects from sample size. For further statistical analyses, see Appendix B.1.

For permutation importance of Z_1 , importance varies by Trial ($F = 15,362.70$, $p < .001$) with a clear additional effect of Model ($F = 85.11$, $p < .001$); sample size has no main effect ($p = .600$). The within-subject Measurement factor ($X \rightarrow W_1 \rightarrow W_2$) is small but reliable ($F = 4.41$, $p = .016$), indicating that classical error slightly shifts the salience of Z_1 . That shift is modestly contingent on design, with Trial $\times$ Measurement ($F = 3.94$, $p < .001$) and $n \times$ Measurement ($F = 4.06$, $p = .004$) both significant but very small in magnitude; other interactions with Measurement are negligible. In contrast, permutation importance of Z_2 shows only weak effects: small main effects of Trial ($F = 2.64$, $p = .022$) and Model ($F = 3.03$, $p = .017$), while Measurement is not significant ($p = .225$) and its interactions are non-significant or marginal (e.g., $n \times$ Measurement $p = .071$). Overall, classical error significantly influences the relative importance of Z_1 but minimally affects Z_2 , with any Measurement-related effects for Z_1 being statistically detectable but practically small.

Misclassification (MsC). The measurement factor within subjects (comparing baseline X vs. misclassified W levels) is the primary source of variance in both outcomes (R^2 : $F = 77,778.86$, $p < .001$; MSE: $F = 100,884.62$, $p < .001$), showing that misclassification leads to significant systematic changes in model fit and error. These effects vary considerably depending on the data-generating Trial (R^2 : Trial $\times$ Measurement $F = 8,106.22$, $p < .001$; MSE: $F = 10,446.55$, $p < .001$), with smaller but notable influences from Model and sample size (R^2 : Model $\times$ Measurement $F = 211.18$, $p < .001$; $n \times$ Measurement $F = 7.91$, $p < .001$; MSE: Model $\times$ Measurement $F = 10.39$, $p < .001$; $n \times$ Measurement $F = 7.13$, $p < .001$). Main effects of Trial and Model are also substantial (e.g., R^2 : Trial $F = 14,145.56$, $p < .001$; Model: $F = 438.62$, $p < .001$), indicating considerable baseline variability across different data scenarios and methods. For further statistical analyses, see Appendix B.2.

The importance of Z_1 heavily depends on the data-generating Trial ($F = 12,240.01$, $p < .001$), with additional influences from Model ($F = 69.48$, $p < .001$) and a small Trial $\times$ n interaction ($F = 6.55$, $p < .001$). The error level (Measurement: X vs. W_k) has a minor yet significant main effect on Z_1 's importance ($F = 14.69$, $p < .001$), further weakly affected by Trial and Model (Trial $\times$ Measurement $F = 2.77$, $p = .004$; Model $\times$ Measurement $F = 15.00$, $p < .001$). This suggests that misclassification can alter Z_1 's contribution, but the overall impact is small compared to the underlying scenario and method choice. Conversely, the permutation importance of Z_2 remains relatively stable: effects from Trial and Model are statistically significant but very small (Trial: $F = 2.35$, $p = .040$; Model: $F = 3.19$, $p = .013$), and the main effect of error level is not significant (Measurement $F = 2.05$, $p = .143$); interactions are generally negligible. In summary, under MsC, Z_1 's importance is mainly influenced by the generative scenario and, to a lesser extent, by the model, with only minor variations due to misclassification severity. Meanwhile, Z_2 's importance is mostly unaffected by misclassification.

Berkson Error (BE). The analysis reveals a significant within-subject effect of Measurement (X_k vs. W_k) on both metrics (R^2 : $F = 107,902.26$, $p < .001$; MSE: $F = 141,398.47$, $p < .001$). This suggests that Berkson error markedly decreases R^2 and raises MSE. Additionally, the Measurement $\times$ Trial interaction is highly significant (R^2 : $F = 3,662.76$; MSE: $F = 4,760.20$, both $p < .001$), indicating that the extent of degradation varies considerably depending on the generative trial (T1–T6).

Between-subject factors such as Trial, sample size n , and Model significant main effects for both metrics (all $p \leq .001$ for R^2 , and for MSE, Trial and Model ($p < .001$; n ($p < .001$), and they exhibit several modest interactions, like Trial $\times$ n , both metrics $p < .001$. For R^2 , the interactions Trial $\times$ Model and Model $\times$ Measurement are additionally significant (both $p < .001$) with medium effect sizes, suggesting that different model families vary in how quickly R^2 decreases from baseline (X_1, X_2) to (W_1, W_2) across trials. For MSE, Model $\times$ Measurement is significant but has a smaller ($F = 8.81$, $p < .001$). Higher-order three-way interactions involving Measurement are negligible and not significant, indicating a consistent pattern: across models and sample sizes, W_2 results in the largest performance loss, W_1 a smaller loss, and the baselines (X_1, X_2) outperform W_1, W_2 . These findings replicate the descriptive line graphs ($n = 4,000$), where trajectories decline from X_1 to W_1 and further X_2 to W_2 across trials. For further statistical analyses, see Appendix B.3.

For the variable importance, the significance of Z_1 varies strongly and systematically across data-generating trials, with a significant trial effect ($F = 32,772.81$, $p < .001$). It also shows smaller but notable modulation by sample size (n : $F = 16.82$, $p < .001$) and model class (Model: $F = 148.42$, $p < .001$). Importantly, the within-subject Measurement factor (X_1, X_2, W_1, W_2) greatly influences Z_1 (Measurement: $F = 1,610.56$, $p < .001$). There are significant trial-specific and size-specific shifts as well (Trial $\times$ Measurement: $F = 31.89$, $p < .001$; $n \times$ Measurement: $F = 7.63$, $p < .001$). This suggests that increasing Berkson error ($W_1 \rightarrow W_2$) consistently alters the importance assigned to Z_1 , with the magnitude of the change depending slightly on trial and sample size.

In contrast, Z_2 displays considerably weaker effects overall: both Trial and Model are statistically significant, but with small effect sizes (Trial: $F = 3.50$, $p = .004$; Model: $F = 14.20$, $p < .001$). Although the Measurement effect is significant, it remains modest in magnitude ($F = 19.60$, $p < .001$). Interactions involving Measurement for Z_2 are present but are minor in practical significance. Overall, these findings suggest that Berkson error significantly disrupts the estimated importance of Z_1 , with effects varying by trial and to some extent by sample size, while the importance of Z_2 remains relatively stable and is only slightly influenced by error level and modeling choices.

Group-Summary Assignment (GSA). In the GSA regression, Measurement (X_1, X_2, W_1, W_2) was the primary factor influencing both outcomes, showing very large effects (R^2 : $F = 8,032.21$; MSE: $F = 8,766.28$; $p < .001$). This aligns with expectations that the error-prone inputs W_1 and especially W_2 produce lower R^2 and higher MSE compared to the baseline variables (X_1, X_2). Significant main effects of Trial (R^2 : $F = 368.83$; MSE: $F = 391.96$; $p < .001$) and clear Trial $\times$ Measurement interactions (R^2 : $F = 382.67$; MSE: $F = 416.00$; $p < .001$) were observed, suggesting that the extent of decline from X_k to W_k varies depending on the data scenario, though not its direction.

The main effect of the Model was significant for R^2 ($F = 26.76$, $p < .001$) and less so for MSE ($F = 8.21$, $p < .001$), indicating some differences in absolute performance between models. However, the interaction between Model and Measurement was modest for R^2 ($F = 21.52$, $p < .001$) and negligible for MSE ($p = .977$), suggesting similar sensitivity to W_1, W_2 across models regarding error. Sample size had only small or marginal effects (R^2 : $p = .044$; MSE: $p = .056$), and $n \times$ Measurement was not significant for either metric ($p = .098$ for R^2 , $p = .112$ for MSE), which indicates that larger samples mainly help stabilize estimates rather than change their relative order (X followed by W_1 then W_2 for R^2 and the other way around for MSE). Higher-order interactions were insignificant. For further statistical analyses, see Appendix B.4.

The significance of Z_1 remains consistent across different trials and measurement conditions, while Z_2 remains largely stable. Permutation importance analysis reveals that the main effect of Trial is highly significant ($F = 1,674.99$, $p < .001$), demonstrating that data-generating scenarios (T1–T6) cause substantial changes in Z_1 's contribution. Additionally, a smaller but significant Trial $\times$ n interaction ($F = 4.02$, $p < .001$)

indicates that these shifts modestly scale with sample size. Measurement effects (baseline/error level) are also highly significant for Z_1 ($F = 1,928.06$, $p < .001$), and the Trial $\times$ Measurement interaction ($F = 14.74$, $p < .001$) suggests that moving from X_1, X_2 to W_1, W_2 systematically alters Z_1 importance, with the change magnitude depending on the trial.

In contrast, most effects for Z_2 are not significant; the Model’s main effect is statistically significant but quite small ($F = 4.07$, $p = .003$). Measurement shows only a marginal effect ($p = .071$), and higher-order terms are negligible. This suggests that Z_2 ’s importance remains consistent across different trials, sample sizes, and error levels.

3.2 Classification Problems

This section presents the results of both descriptive and inferential analyses of the four MCEs for classification tasks. In classification problems, the performance measures M are accuracy, F1 score, and permutation feature importance.

3.2.1 Descriptive Analysis

3.2.2 Classical Error (CE)

For CE, the differences $\Delta\text{Accuracy}(W_k; X) = \text{Accuracy}(W_k) - \text{Accuracy}(X)$ and $\Delta F1(W_k; X) = F1(W_k) - F1(X)$ were consistently negative across trials (T1–T6), models, and sample sizes, indicating that adding classical noise to X degrades performance. The magnitude reflected the noise level: W_1 ($\sigma = 0.5$) caused declines, while W_2 ($\sigma = 0.75$) resulted in larger drops, maintaining the order $M(W_2) < M(W_1) < M(X)$. The differences among classifiers were noticeable but modest and did not alter the overall pattern. At $n = 4000$, variability in the difference decreased, but remained higher than in regression problems, consistent with more stable estimates at larger n . Figure 9 shows dense, downward trajectories of Accuracy from X to W_1 , and further to W_2 , with differences between models being modest. The trajectories exhibit a consistent left-to-right decrease and a tight clustering of individual lines around the mean, though less tight than in regression, reflecting stable behavior at $n = 4000$. The results for $\Delta F1(W_k; X)$ mirror the Accuracy findings; see Figure A.5.

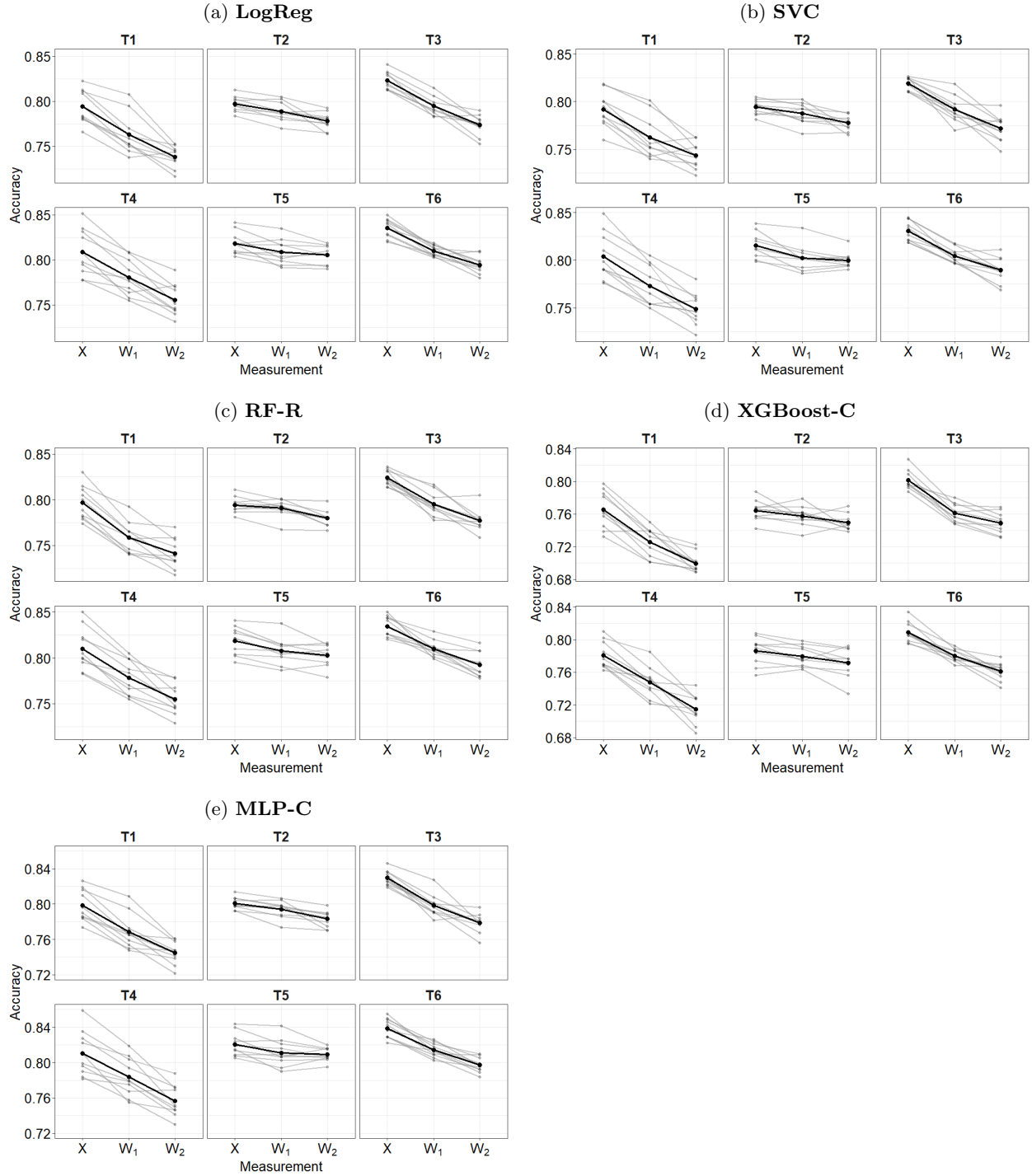


Figure 9: Line graphs of accuracy for the MCE of CE with $\sigma = 0, 0.5, 0.75$ at trials T1–T6 and for (a) LogReg, (b) SVC, (c) RF-C, (d) XGBoost-C, and (e) MLP-C.

Figure 10 shows permutation importance in the classification setting under CE across various trials and models. The patterns in classifier results under CE closely mirror those seen in the regression setting

(Figure 2). With increased measurement error, the permutation importance of the main predictor consistently decreases across trials and classifiers, while the importance of the predictive feature Z_1 stays stable, and the non-predictive feature Z_2 remains minimally important. Differences between trials, such as when Z_1 takes precedence over the primary predictor due to the underlying coefficient structure (e.g., T3 and T6), are maintained in the classification context. Overall, these results suggest that CE’s qualitative detrimental effect on variable importance is mostly unaffected by whether the task involves regression or classification.

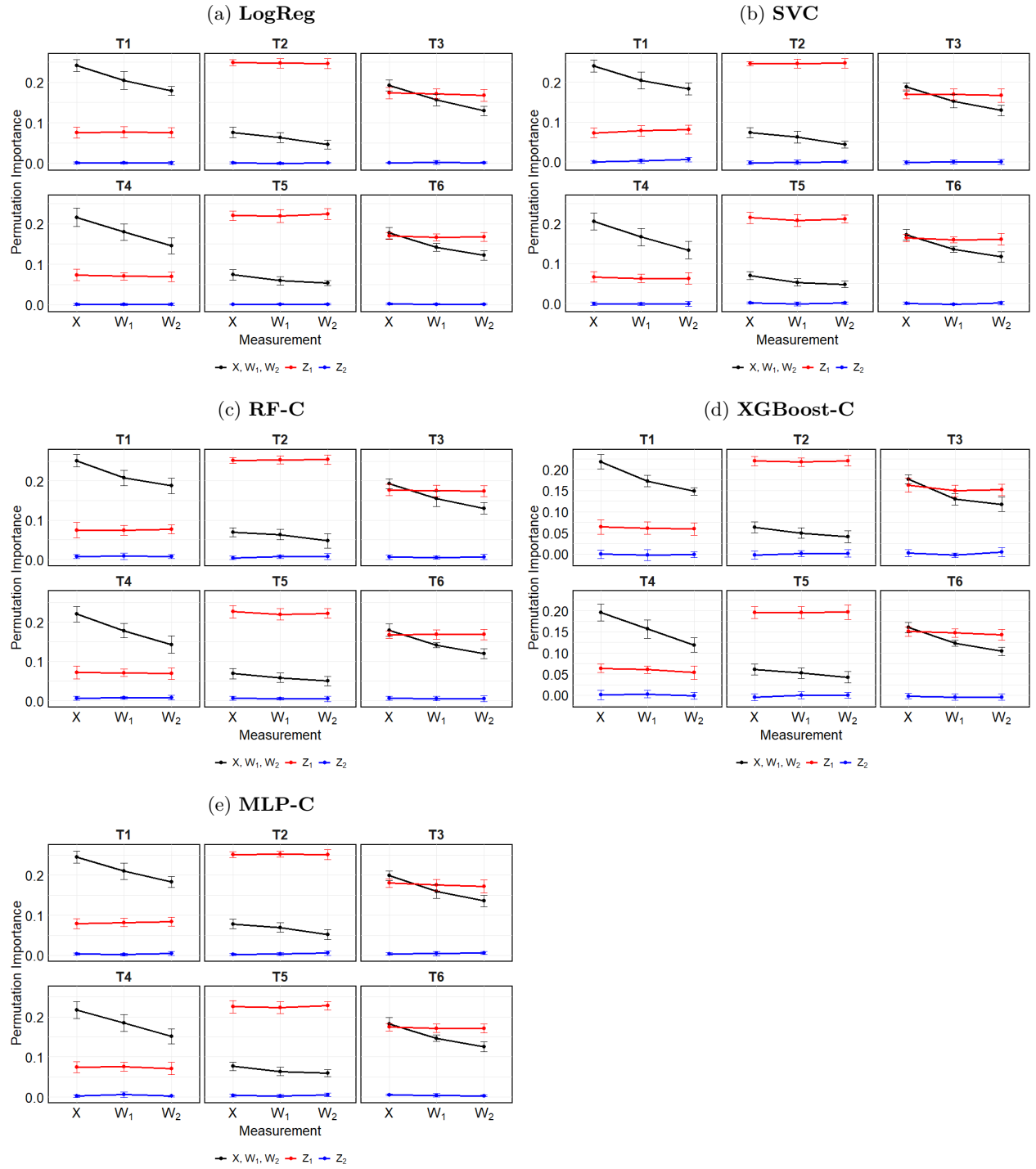


Figure 10: Line graphs of the mean permutation importance ($\pm$ SD) for the MCE of CE with $\sigma = 0, 0.5, 0.75$ at trials T1–T6 and for (a) LogReg, (b) SVC, (c) RF-C, (d) XGBoost-C, and (e) MLP-C.

3.2.3 Misclassification (MsC)

The differences were mostly negative across all trials (T1–T6), models, and sample sizes, indicating a performance decline when the misclassified predictor W_k replaces X . The magnitudes of these differences correlated with the error rate q : W_1 ($q = 0.10$) produced moderate decreases, whereas W_2 ($q = 0.30$) resulted in larger decreases, maintaining the order $M(W_2) < M(W_1) < M(X)$. Although there were minor differences among classifiers, they did not change the overall trend. At $n = 4000$, the variability of the difference decreased, but remained higher than in regression problems, which reflects more consistent estimates at larger sample sizes. Figure 11 displays dense, downward trends for Accuracy from X to W_1 , and further to W_2 , at $n = 4000$ across all models, with boldfaced lines between means paths highlighting the same monotonic pattern. However, we observed less significant differences here, particularly when an intercept is present: trials T4–T6, along with trial T2, show that X contributes less to Y than Z_1 . Even trial T3 has shown modest changes. Results for $\Delta F1(W_k; X)$ mirror the Accuracy results; see Figure A.6.

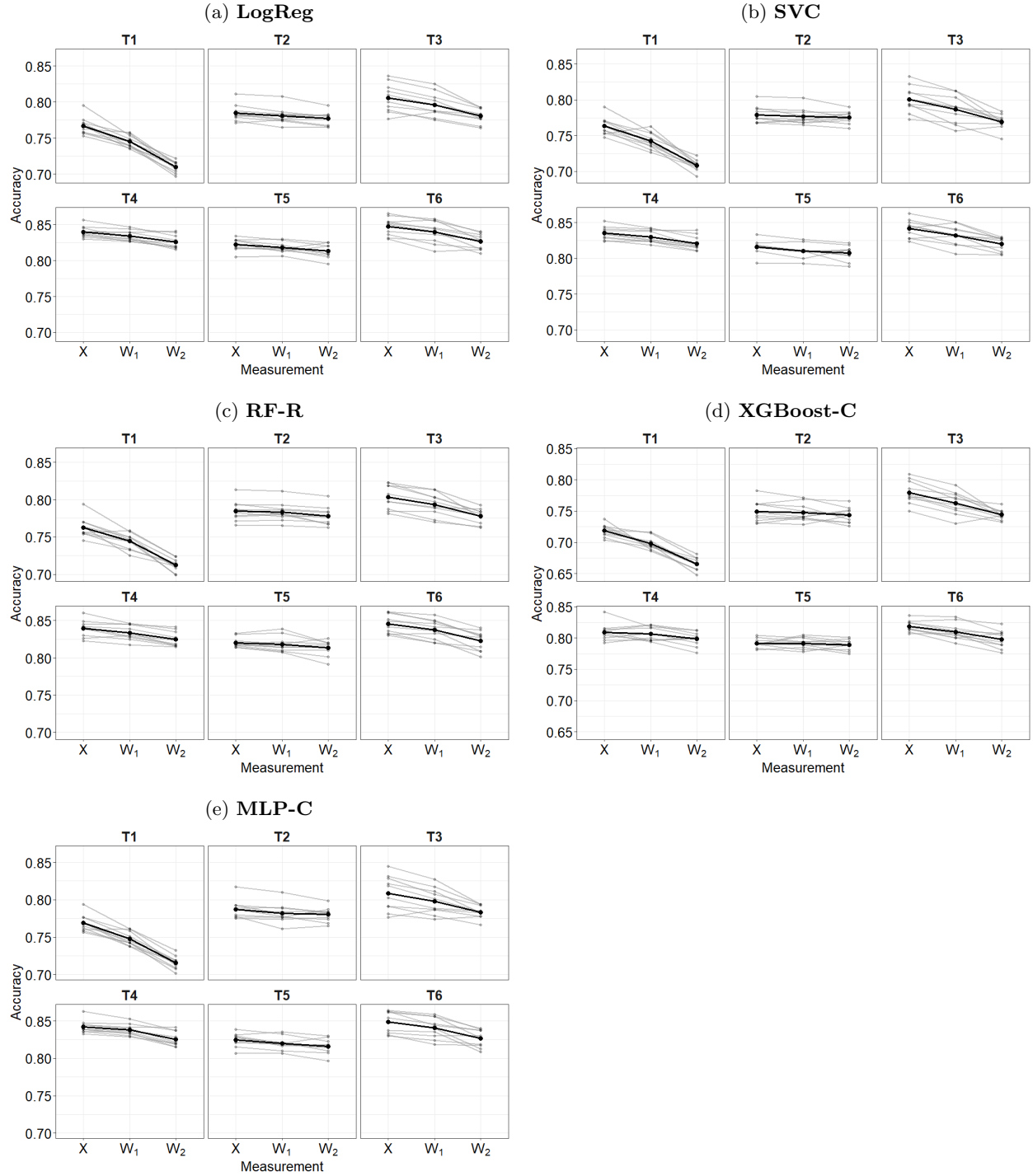


Figure 11: Line graphs of accuracy for the MCE of MsC with $q = 0, 0.1, 0.3$ at trials T1–T6 and for (a) LogReg, (b) SVR, (c) RF-R, (d) XGBoost-R, and (e) MLP-C.

In the MsC classification setting, the permutation importance of the affected predictor shows a strong trial-dependent pattern across classifiers, as shown in Figure 12. During trials (T1-T3) with a zero intercept,

MsC results in significant but fluctuating decreases in the primary predictor’s importance, with the extent of reduction depending on the relative strengths of competing predictors. Conversely, in trials (T4–T6) with a nonzero intercept, MsC causes a more pronounced decrease, often reducing the importance of the misclassified predictor to nearly zero. Throughout all trials, the importance of the feature Z_1 remains fairly stable, while the non-predictive feature Z_2 consistently shows negligible importance.

Unlike the CE, these findings suggest that in classification models, the influence of MsC on variable importance is collectively determined by the error mechanism and the intercept, which adjusts the baseline contribution of predictors to class separation.

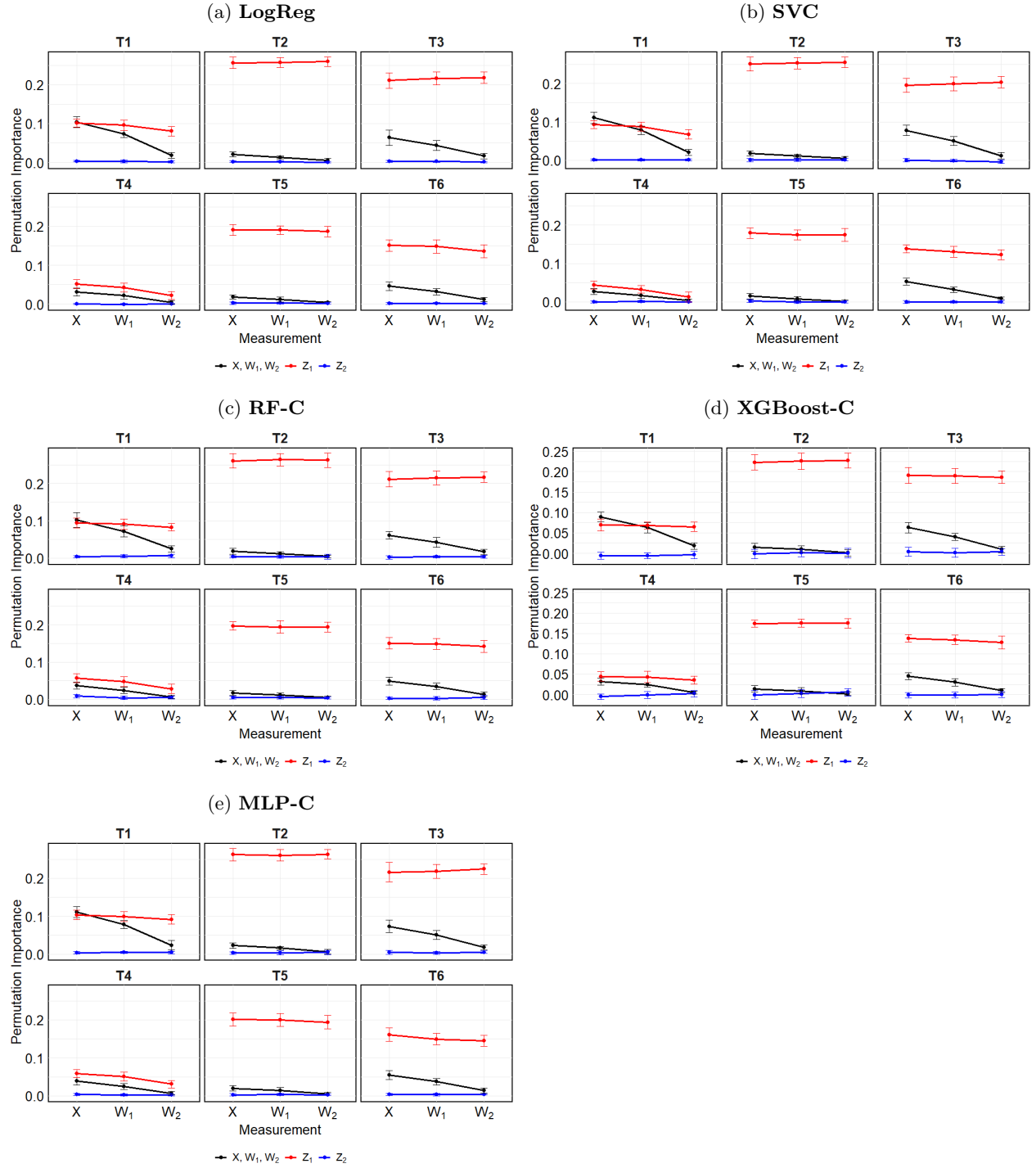


Figure 12: Line graphs of the mean permutation importance ($\pm$ SD) for the MCE of MsC with $q = 0, 0.1, 0.3$ at trials T1-T6 and for (a) LogReg, (b) SVC, (c) RF-C, (d) XGBoost-C, and (e) MLP-C.

3.2.4 Berkson Error (BE)

Under BE, the differences were mostly negative across trials (T1–T6), models, and sample sizes, with larger losses at the higher-error level (W_2) than at W_1 . For X_2 , W_2 resulted in a slightly less negative difference than X_1 , W_1 , but the pattern performance for W_2 is less than W_1 and remained consistent. For visual clarity at the largest sample size, the line plot shows Accuracy across measurements ($M(X_k) > M(W_k)$) for $n = 4000$ and all models. Figure 13 displays dense downward trends from X to W , with boldfaced lines between means confirming this pattern across trials. As the sample size increases to 4000, dispersion decreases, but the variability is larger than in regression problems, indicating more stable estimates. At the same time, model-specific differences are minor and do not change the overall conclusion. The results for $\Delta F1(W_k; X_k)$ are similar to the Accuracy outcomes, see Figure A.7.

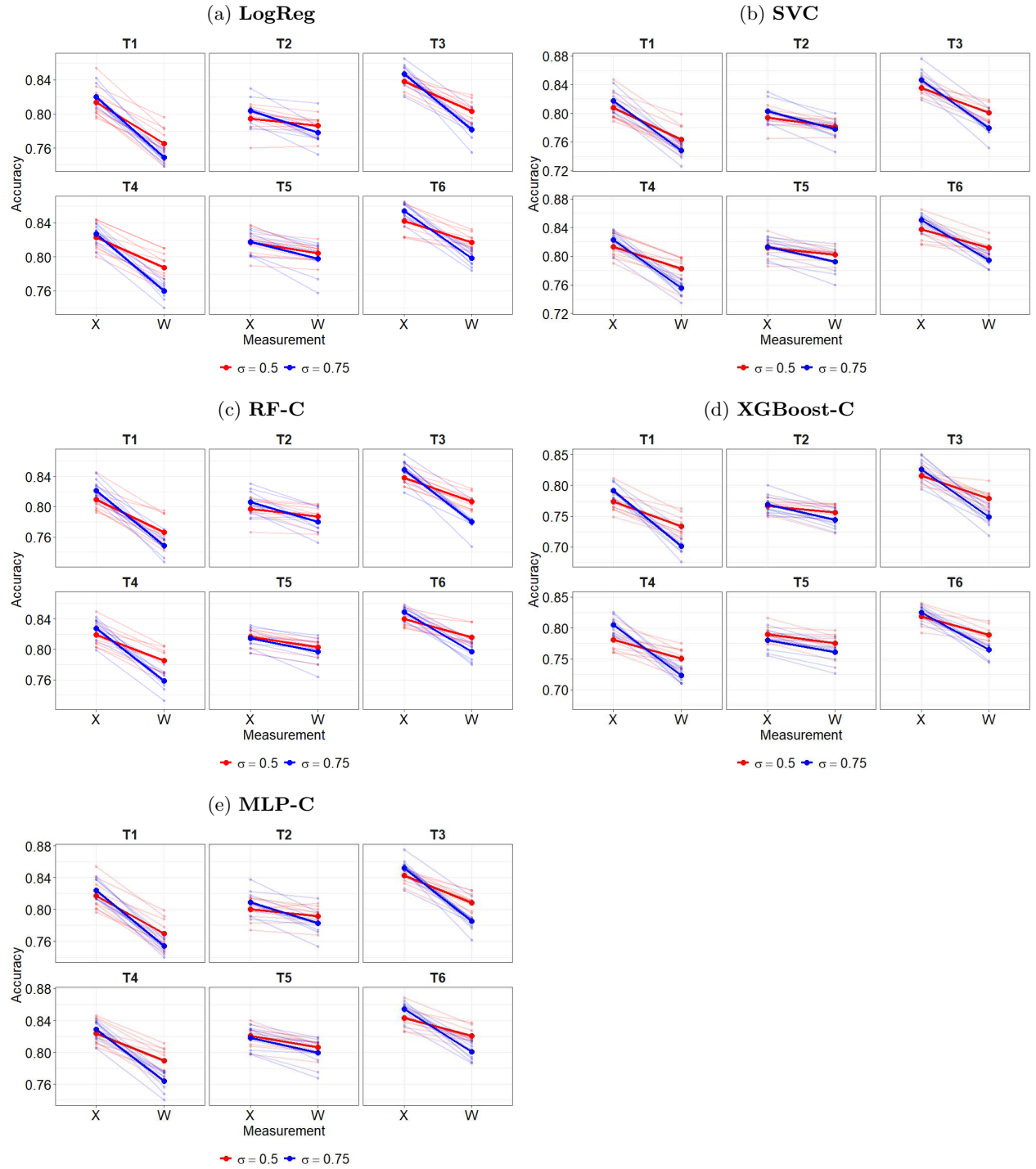


Figure 13: Line graphs of accuracy for the MCE of BE with $\sigma = 0, 0.5, 0.75$ at trials T1-T6 and for (a) LogReg, (b) SVC, (c) RF-C, (d) XGBoost-C, and (e) MLP-C.

Under BE in the classification setting, permutation importance exhibits relatively moderate changes across trials and classifiers, Figure 14. Replacing the error-free predictor X with the BE affected measurement W

476 leads to a moderate reduction in the importance of the primary predictor; however, differences between error
477 levels are minor and do not follow a consistent or monotonic pattern. The importance of Z_1 remains stable
478 across measurement conditions, and the importance of Z_2 continues to exhibit negligible importance. In
479 general, BE alters the magnitude of variable importance without substantially changing relative rankings,
480 and its effect is notably weaker and less structured than that observed under CE or MsC.

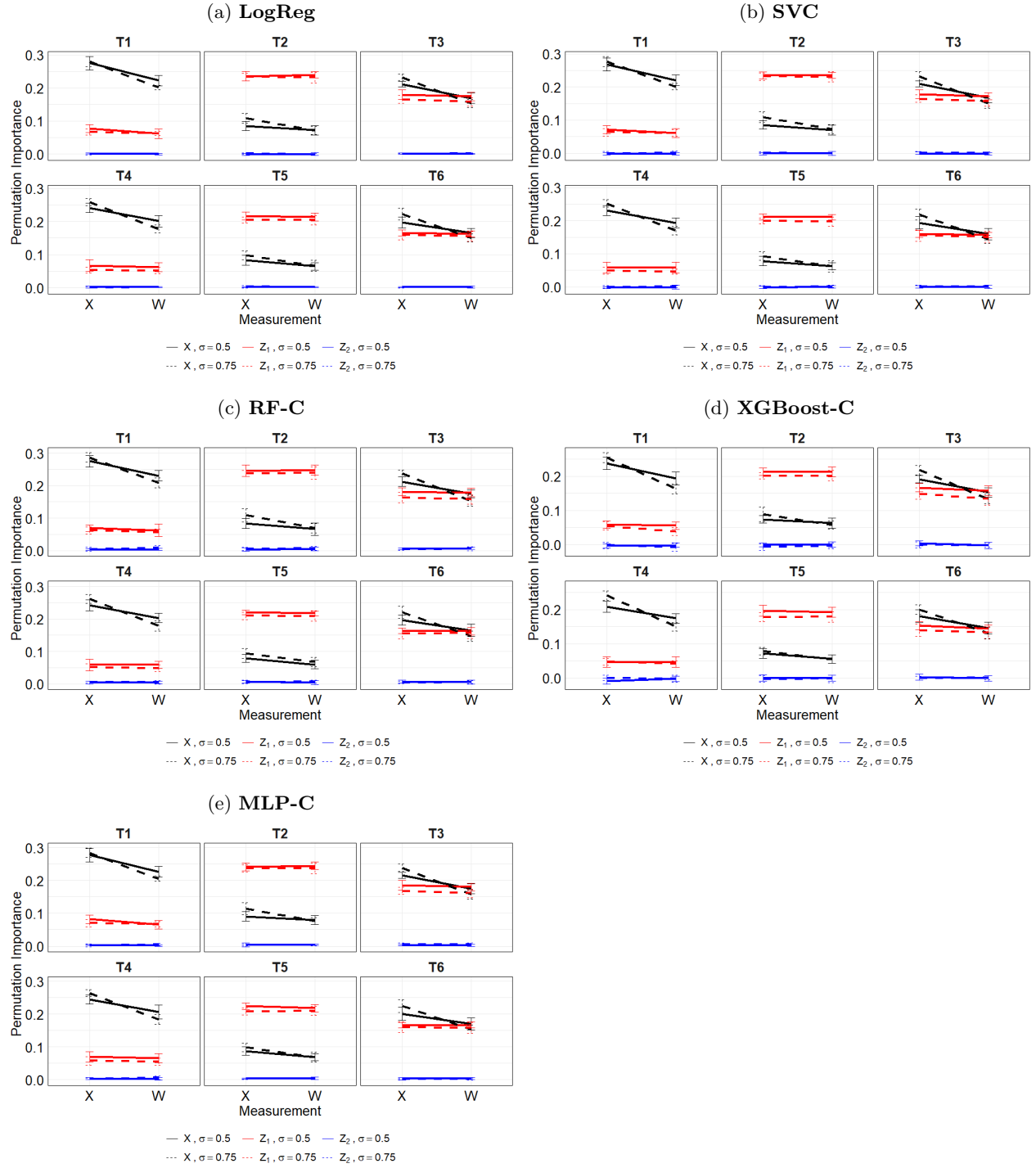


Figure 14: Line graphs of the mean permutation importance ($\pm$ SD) for the MCE of BE with $\sigma = 0, 0.5, 0.75$ at trials T1–T6 and for (a) LogReg, (b) SVC, (c) RF-C, (d) XGBoost-C, and (e) MLP-C.

3.2.5 Group-Summary Assignment (GSA)

Under GSA, differences were mostly negative across trials (T1–T6), models, and sample sizes, and their magnitude was larger for W_2 than for W_1 . Compared with X_1 , W_1 , the difference was slightly less detrimental than with X_2 , W_2 , yet the consistent monotone decline order W_1 then W_2 persisted. Figure 15 illustrates this downward trend, for $n = 4000$ and all models ($M(X_k) > M(W_k)$), with some variability. The larger sample size results also indicate more stable estimates. The $\Delta F1(W_k; X_k)$ findings follow the same pattern as Accuracy, as shown in Figure A.8.

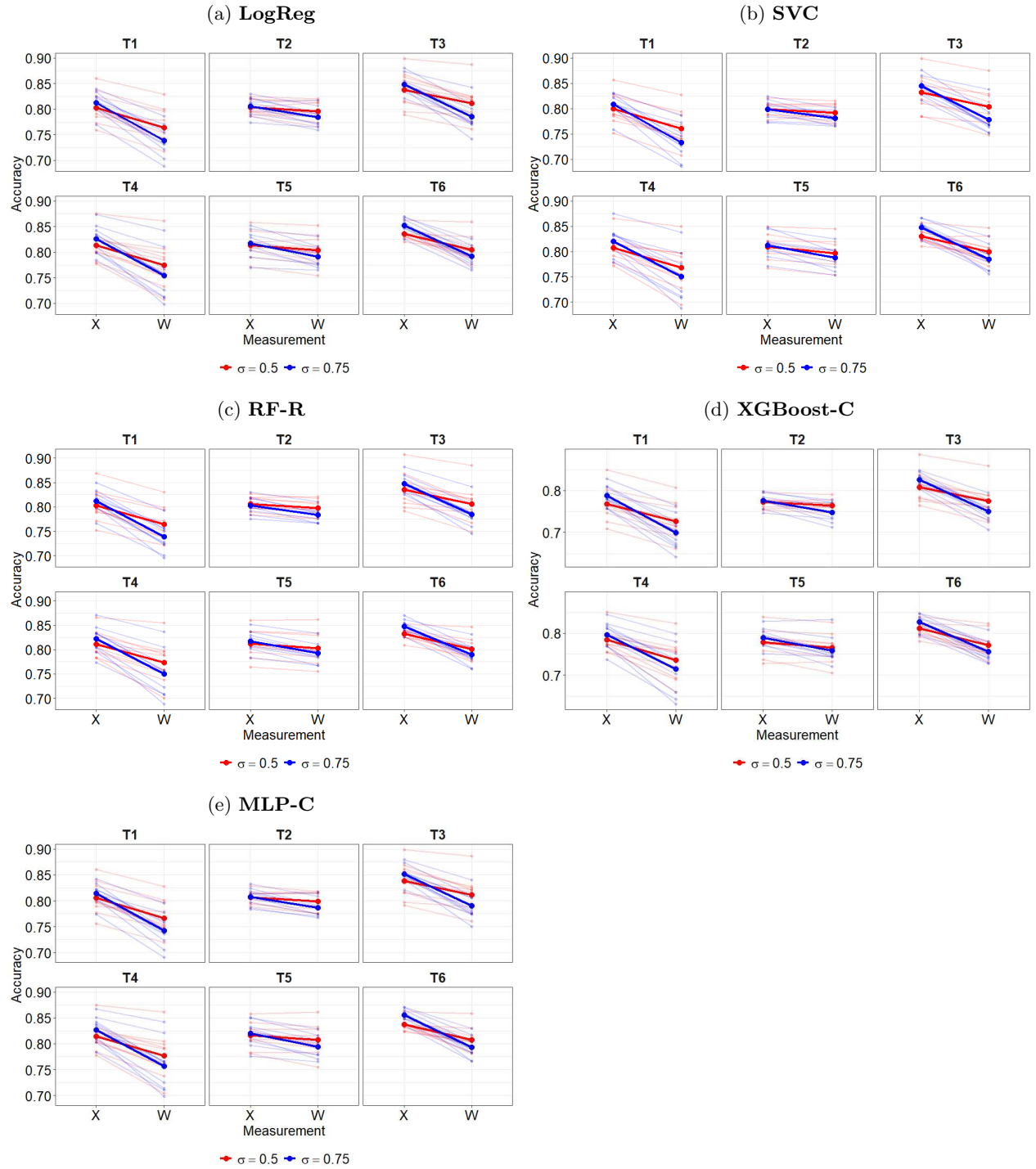


Figure 15: Line graphs of the Accuracy for the MCE of GSA with $\sigma = 0, 0.5, 0.75$ at trials T1-T6 and for (a) LogReg, (b) SVR, (c) RF-R, (d) XGBoost-R, and (e) MLP-C.

Figure 16 shows the permutation importance in the GSA classification across trials and classifiers. Replacing the individual predictor X with its group-level summary W causes a noticeable decrease in

importance, indicating the loss of within-group variation due to aggregation. Differences between error levels are minimal and do not show a consistent pattern. The importance of Z_1 remains fairly stable, while Z_2 continues to have negligible importance.

Overall, the Group-Summary Assignment primarily compresses variable importance through aggregation rather than by sensitivity to error variance. Compared to CE, GSA decreases variable importance mainly via aggregation instead of noise contamination, causing a steady decline from X to W without depending systematically on the error variance. Unlike MsC, which often causes sudden and trial-dependent drops in importance, and BE, which results in only minor changes, GSA produces a structured reduction in importance that maintains the predictors' relative order.

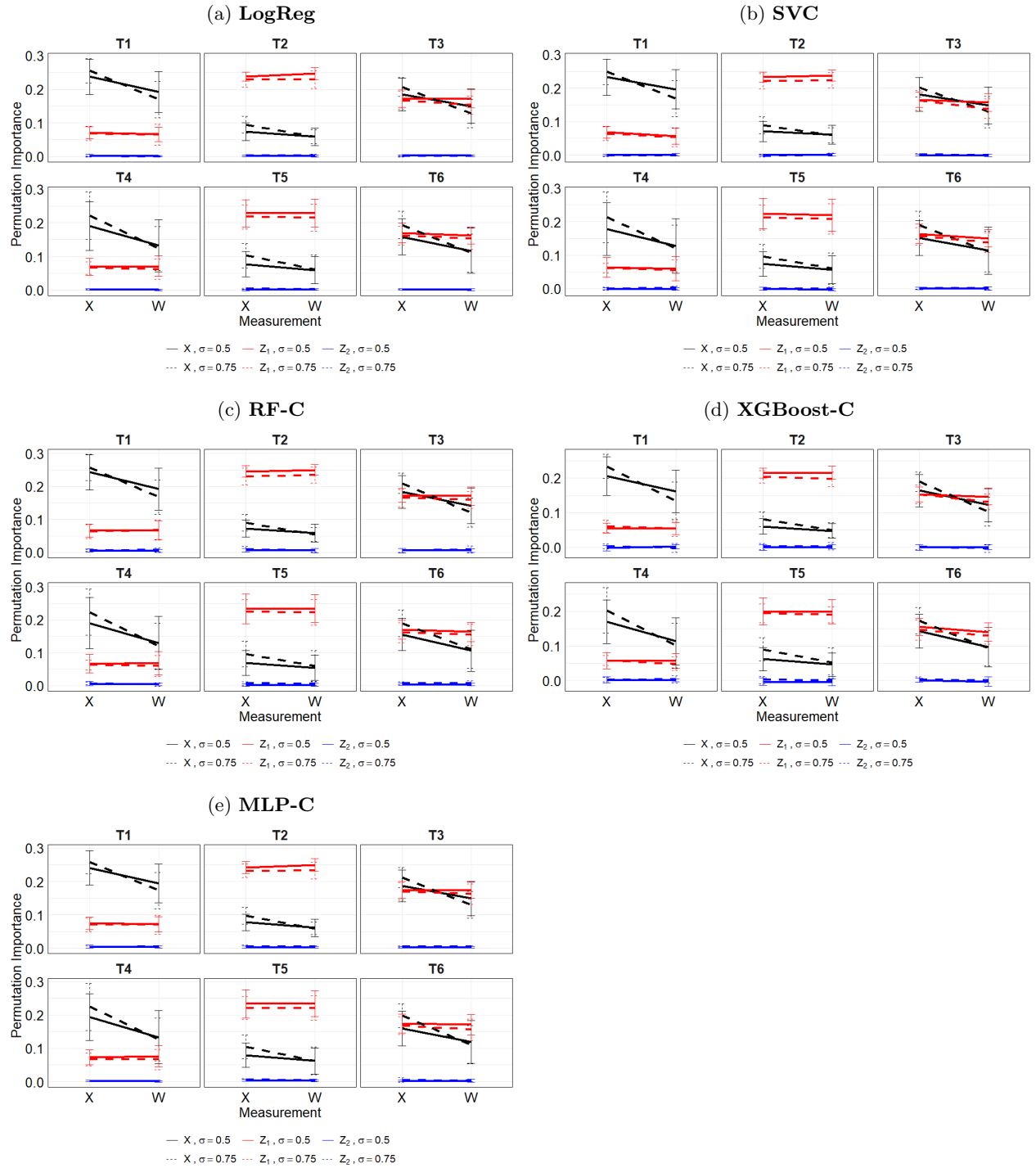


Figure 16: Line graphs of the mean permutation importance ($\pm$ SD) for the MCE of GSA with $\sigma = 0, 0.5, 0.75$ at trials T1–T6 and for (a) LogReg, (b) SVC, (c) RF-C, (d) XGBoost-C, and (e) MLP-C.

In the previous figures, as in the regression problems, trials T2 and T5 showed minimal or no change relative to trials T1, T3, T4, and T6. Unlike regression problems, classification problems exhibited greater

variability in performance across the MCEs, but were consistent with the loss pattern.

3.2.6 Statistical Analysis

Classical Error (CE). We examined Accuracy and $F1$ -score using repeated-measures ANOVA, with Measurement (X , W_1 , W_2) as the within-subject factor, and Trial (α configuration), sample size (n), and Model as between-subject factors. Both metrics showed significant main effects for Trial (Accuracy: $F = 213.13$; $F1$ -score: $F = 959.51$, both $p < .001$), Model (Accuracy: $F = 216.39$; $F1$ -score: $F = 191.54$, both $p < .001$), and notably Measurement (Accuracy: $F = 3298.33$; $F1$ -score: $F = 2667.17$, both $p < .001$). Sample size had a negligible effect on Accuracy ($p = .757$) and was marginal for $F1$ -score ($p = .065$).

Importantly, the effect of measurement error was influenced by the α setup (Trial $\times$ Measurement: Accuracy $F = 145.09$; $F1$ -score: $F = 133.56$; $p < .001$). It was also modestly affected by sample size and model ($n \times$ Measurement and Model $\times$ Measurement showed significant but small effect sizes). Post-hoc analyses revealed consistent declines from W_1 and W_2 compared to X , with W_2 (higher error) causing larger drops (typically = 0.04–0.08), and W_1 showing smaller decreases (= 0.005–0.04). Some effects of W_1 during T2/T5 or at larger sample sizes were close to zero. Higher-order interactions were negligible. Overall, measurement error reliably lowered both Accuracy and $F1$ -score, primarily depending on the α setup and secondarily on model choice and sample size. For further statistical analyses, see Appendix B.1.

We evaluated the permutation importance of the genuinely predictive covariate Z_1 using a repeated-measures ANOVA. The importance of Z_1 was predominantly affected by the α configuration (Trial: $F = 3556.09$, $p < .001$) and, to a lesser degree, by the Model (Model: $F = 94.89$, $p < .001$); sample size had a smaller impact ($F = 11.82$, $p < .001$). Error level caused a small but significant decrease in Z_1 importance (Error level: $F = 29.58$, $p < .001$): the grand mean dropped from 0.160 for X to 0.157 for W_1 and 0.156 for W_2 (= -0.003 and -0.004 compared to X). This relationship was modestly influenced by the α configuration (Trial $\times$ Error level: $F = 7.53$, $p < .001$) and was only weakly affected by the model (Model $\times$ Error level: $F = 2.08$, $p = .039$); other Measurement-related interactions were negligible. Cell-wise contrasts mostly revealed small, frequent decreases, particularly at T3 and T6 across various models, with occasional isolated increases that did not alter the overall pattern. In summary, classical measurement error slightly underestimates the learned importance of Z_1 , but this impact is much smaller than the effects of α configuration and model selection.

The permutation importance of Z_2 remained relatively stable across different conditions. Throughout the trials, the estimated importance of Z_2 did not show variation (Trial: $F = 0.37$, $p = .866$), with only a minor main effect of sample size (n : $F = 6.69$, $p = .001$) and a clear model-class effect ($F = 102.80$, $p < .001$). Additionally, a slight interaction between Trial and n was observed ($F = 3.36$, $p < .001$), as was an interaction between n and Model ($F = 4.43$, $p < .001$), with no other significant higher-order interactions. Importantly, the error level (none = X , mild = W_1 , strong = W_2) did not significantly influence Z_2 importance (main effect: $F = 1.87$, $p = .154$); only very minor interactions were detected (Model $\times$ Error level: $F = 2.70$, $p = .006$; Trial $\times n \times$ Error level: $F = 1.72$, $p = .024$). Comparisons between W_1 or W_2 and X mostly yielded null results; when differences appeared, they were small and inconsistent in sign, typically $|\Delta| \in [0.002, 0.008]$. Collapsing across factors, the grand means were $X = 0.00254$ (95% CI [0.00211, 0.00296]), $W_1 = 0.00305$ (95% CI [0.00257, 0.00352]), and $W_2 = 0.00301$ (95% CI [0.00257, 0.00345]), indicating that increased error levels did not systematically alter the importance assigned to Z_2 .

Misclassification (MsC). Under misclassification (MsC), both Accuracy and $F1$ -score follow a clear pattern: performance decreases as error increases. For Accuracy, the within-subject factor Measurement is substantial and highly significant ($F(1.69, 1372.05) = 1863.77$, $p < .001$), and for $F1$ -score, it is similarly strong ($F(1.70, 1380.64) = 1144.56$, $p < .001$). Trial exerts a very large main effect on both metrics (Accuracy: $F(5, 810) = 1040.53$; $F1$ -score: $F(5, 810) = 2077.43$; both $p < .001$), and Model also influences results (Accuracy: $F(4, 810) = 231.00$; $F1$ -score: $F(4, 810) = 262.38$; both $p < .001$). Nonetheless, models do not significantly differ in their sensitivity to misclassification (Model $\times$ Measurement: Accuracy $p = .474$; $F1$ -score $p = .299$).

Sample size has only a minor influence (Accuracy main effect: $F(2, 810) = 3.73$, $p = .024$; $F1$ -score: $F(2, 810) = 1.47$, $p = .230$), with small interactions involving Measurement (Accuracy: $F(3.39, 1372.05) = 3.37$, $p = .014$; $F1$ -score: $F(3.41, 1380.64) = 3.19$, $p = .018$). The observed degradation from X to W_1 to W_2 is consistent across models and largely stable across sample sizes, though its extent depends on the specific trial scenario. For further statistical analyses, see Appendix B.2.

Permutation importance results differ for Z_1 versus Z_2 . For Z_1 , there is a strong main effect of misclassification level (Measurement: $F = 205.5$, $p < .001$), indicating a clear decline in importance (X followed by W_1 then W_2). The importance of Z_1 also varies significantly with Trial and Model (both $p < .001$), with a smaller impact from sample size; the effect of misclassification remains fairly consistent across different sample sizes ($n \times$ Measurement) and is only modestly influenced by Trial and Model (both $p < .001$). Conversely, for Z_2 , misclassification level is not a reliable driver (Measurement $p = .178$); instead, differences are mainly driven by sample size and Model (both $p < .001$), with minor interactions such as Trial $\times$ Model ($p = .041$) and $n \times$ Model ($p = .001$). Overall, misclassification substantially reduces the importance of Z_1 , whereas Z_2 remains comparatively stable, with sample size and method choice having a larger influence on Z_2 than the error level.

Berkson Error (BE). We evaluated Accuracy and $F1$ -score using repeated-measures ANOVA. A very strong main effect of the measurement condition/error level (X_k and W_k) was observed on both metrics (Accuracy: $F = 2444.28$, $p < .001$; $F1$ -score: $F = 2120.42$, $p < .001$). Additionally, Trial and Model have significant main effects (e.g., Accuracy: Trial $F = 317.38$; Model: $F = 358.43$; $F1$ -score: Trial $F = 1437.34$; Model: $F = 317.02$; all $p < .001$). Sample size also shows modest but significant improvements.

The critical interaction between Trial $\times$ error level is significant for both metrics: Accuracy ($F = 93.96$) and $F1$ -score ($F = 97.91$; $p < .001$), indicating that the magnitude of degradation depends on the α -configuration. Interactions involving error level with sample size and with the Model are not statistically significant, indicating that the response pattern (X followed by W_1 followed by W_2) remains consistent across different models and sample sizes. Pairwise comparisons against X_1 and X_2 reveal that W_1 usually diminishes performance, while W_2 causes even greater declines, with many cell-wise confidence intervals excluding zero. In summary, Berkson error progressively worsens classification quality, and these findings are robust for both Accuracy and $F1$. For further statistical analyses, see Appendix B.3.

The patterns of feature importance clearly differ for Z_1 and Z_2 . For Z_1 , the α -configuration (Trial) is most prominent ($F = 5953$, $p < .0001$), and there are significant differences between models ($F = 140.8$, $p < .0001$). Measurement also plays a role with a moderate average effect ($F = 134.6$, $p < .001$), and this measurement effect slightly varies depending on the Trial ($F = 6.81$, $p < .001$) and, to a lesser extent, on the interaction between Trial and sample size ($F = 3.94$, $p < .001$).

By contrast, the significance of Z_2 shows no main effect of Trial ($F = 1.00$, $p = .417$), a minor yet meaningful variation across models ($F = 128.8$, $p < .001$), and only a small average measurement effect ($F = 5.11$, $p = .002$). There are minor interactions between Trial and Measurement ($F = 2.18$, $p = .007$) and between Model and Measurement ($F = 2.17$, $p = .013$). Sample size has little impact in both cases. Essentially, Z_1 provides most of the predictive signal and is moderately sensitive to using W rather than X (and to α -configuration), while Z_2 remains relatively stable. Its importance varies little with measurement choice or α -configuration, and only slightly with Model or n .

Group-Summary Assignment (GSA). In the group-summary assignment (GSA) experiment, both Accuracy and $F1$ -score strongly depend on the trial (α configuration) and show a small but clear dependence on model and sample size. For Accuracy, significant factors include Trial ($F(5, 810) = 50.31$, $p < .001$), sample size ($F(2, 810) = 5.68$, $p = .004$), and Model ($F(4, 810) = 55.15$, $p < .001$) are significant; critically, the Measurement factor (X_1, X_2 vs. W_1, W_2) is highly significant with a large effect ($F = 2147$, $p < .001$), indicating that using group-averaged exposures (W_1, W_2) reduces performance compared to error-free X . This decrease is modulated by Trial (Trial $\times$ Measurement: $F = 84.9$, $p < .001$), and to a lesser extent with sample size and model ($n \times$ Measurement: $p = .161$; Model $\times$ Measurement: $F = 2.94$, $p = .001$; Trial $\times n \times$ Measurement: $F = 5.39$, $p < .001$).

The $F1$ -score exhibits a similar qualitative pattern: significant effects are seen in Trial ($F(5, 810) = 49.86$,

$p < .001$), Model ($F(4, 810) = 13.65$, $p < .001$), and a modest effect of sample size ($F(2, 810) = 3.12$, $p = .045$). Measurement remains highly significant ($F = 1189.6$, $p < .001$), with interactions showing that the performance drop under W_1, W_2 varies notably by Trial ($F = 57.0$, $p < .001$) and only slightly by sample size ($n \times$ Measurement: $F = 2.35$, $p = .043$). In general, GSA consistently causes a loss in Accuracy/ $F1$ compared to error-free X , especially pronounced in some α settings, while differences between models and sample sizes exist, but are relatively minor. For further statistical analyses, see Appendix B.4.

For the variable importance, the significance of Z_1 is primarily influenced by the generative setup (Trial: $F = 1108.21$, $p < 0.001$), with smaller contributions from sample size ($F = 6.99$, $p < 0.001$) and the model ($F = 30.75$, $p < 0.001$). Importantly, Z_1 exhibits a modest but noticeable dependence on the error level (Measurement: X_1, X_2 versus W_1, W_2 ; $F = 156.85$, $p < 0.001$), and this relationship varies slightly across different trials (Trial $\times$ Measurement: $F = 8.34$, $p < 0.001$) and models (Model $\times$ Measurement: $F = 2.99$, $p = 0.002$).

In contrast, Z_2 remains mostly unaffected by the error level in GSA, as indicated by $p = 0.878$; its variability is mainly due to the model ($F = 145.64$, $p < 0.001$) and, to a lesser extent, the sample size ($F = 18.37$, $p < 0.001$). Only minor Trial-linked modulations are observed (Trial $\times$ Measurement: $F = 1.76$, $p = 0.038$). Overall, GSA reduces the impact of error levels on Z_2 but does not have the same effect on Z_1 .

3.3 Post-hoc Power Analysis

We performed a post-hoc power analysis using G*Power (Faul et al., 2007) within a repeated-measures ANOVA framework. The power estimates for the statistical tests were based on Monte Carlo replications with $m = 10$ and a significance level of $\alpha = 0.05$. The effect size, Cohen’s $F = 7.39$, was determined based on observed generalized η^2 estimates from the repeated-measures ANOVA. The correlation between repeated measurements was set at 0.5, and a Greenhouse–Geisser correction with $\epsilon = 0.82$ was applied to account for nonsphericity. Based on these assumptions, the study’s power was approximately 1.00, which shows that the design was more than capable of detecting the observed effects.

4 Discussion

This study systematically compares how various feature-level measurement error mechanisms affect modern machine learning outcomes across regression and classification tasks. Our experiments consistently reveal a monotonic trend: replacing error-free covariates with error-prone measurements reduces predictive accuracy, with the extent of the decline increasing with the magnitude of the error. This trend is observed across classical error, misclassification, Berkson error, and group-summary assignment, and across all considered ML methods. Notably, increasing the sample size does not eliminate these effects; instead, it primarily reduces variability of the trends around the mean trend in the results, confirming that feature measurement error causes persistent bias in predictive performance rather than merely finite-sample noise.

A post-hoc power analysis was conducted within a repeated-measures ANOVA framework, using 10 Monte Carlo replications and a significance level of $\alpha = 0.05$. The achieved power was approximately 1.00, indicating sufficient power to detect small effects. However, exploratory and inferential analyses showed large effect sizes.

In both regression and classification ML problems, a clear pattern emerges: measurement errors in predictive features consistently impair model performance, with the severity of impairment increasing as error levels rise. Tracking difference (ΔM) simplifies understanding of this trend across different metrics: ΔR^2 , Δ Accuracy, and $\Delta F1$ usually show negative values, while Δ MSE tends to be positive. This pattern persists across all four measurement-error mechanisms (CE, MsC, BE, GSA), all tested model types, and all six generative trials (T1–T6). In essence, whether errors originate from additive noise (CE and BE), discrete mislabeling (MsC), or group-level aggregation (GSA), substituting error-affected features for the error-free baseline consistently yields poorer predictions.

A second important finding is that the extent of performance loss varies significantly across scenarios, reflecting the underlying signal structure of the data-generating process. In the descriptive line plots, the

largest drops occur in trials in which the error-prone feature X exerts a greater influence on Y , notably at T1 and T4, followed by T3 and T6. Conversely, trials where X 's contribution is weaker relative to the other error-free predictor Z_1 show more minor changes, such as T2 and T5. The repeated-measures ANOVA supports this visual pattern: the Measurement $\times$ Trial interaction is consistently large and highly significant for both regression and classification metrics across different measurement error mechanisms. This suggests that measurement error does not merely add noise; instead, it interacts with the distribution of the predictive signal across features, making the impact of feature error predictable from the data's signal-to-noise ratio and coefficient structure.

Third, although all ML model performances are affected, the inferential results indicate that model choice can influence sensitivity to measurement error, particularly in regression tasks. For CE, MsC, and BE regression, measurement variability significantly impacts R^2 and MSE. However, the interactions between Model and Measurement are statistically significant, with greater effect on R^2 than on MSE, suggesting that some methods are slightly more resilient. Conversely, in the classification context, the misclassification experiment reveals that classifiers vary in baseline performance across trials but show minimal differences in sensitivity to increased feature misclassification. The interaction between Model and Measurement is not significant for MsC in Accuracy and F1 score, indicating that when the predictor is discretely corrupted, the error mechanism generally overwhelms model-specific differences.

A fourth key finding relates to sample size n . Increasing n primarily decreases the variability in ΔM by producing tighter trajectories and less variability across replications, but it generally does not significantly alter the overall degradation pattern. This is especially evident in regression analysis, where $n \times$ Measurement interactions are often minor or non-significant across several mechanisms; the degradation direction for R^2 (and the reverse for MSE) remains consistent. In classification, variability tends to be higher than in regression; however, the same conclusion applies: larger samples yield more stable estimates of performance decline without eliminating it. Collecting more data does not eliminate measurement errors in features; rather, it enables more precise estimation of the extent of damage.

Finally, the permutation-importance results underscore a key interpretability insight: measurement error can distort the estimated importance of features, but this distortion varies across predictors. For all measurement error mechanisms, the importance of the genuinely predictive feature Z_1 depends primarily on Trial and Model. In contrast, Measurement often exhibits a statistically significant effect—ranging from modest (CE and MsC) to more pronounced (BE and GSA)—suggesting that feature error can alter how models assign importance to predictors. Conversely, the non-predictive feature Z_2 remains relatively stable: its importance is typically weakly impacted (or not at all) by Measurement and is more affected by the model class and, occasionally, sample size. These findings caution against overlooked effects in applied ML: measurement error not only hampers predictive performance but can also bias post hoc interpretations, especially for predictors that truly carry the signal. Meanwhile, low-signal predictors may appear unexpectedly stable, primarily due to model artifacts rather than measurement issues.

In the regression setting, we observe that feature-level measurement error can influence not just prediction accuracy but also how models assign importance to predictors. CE typically weakens the role of the affected feature gradually, while MsC produces sharper, often more disruptive changes that can significantly impact interpretability. Conversely, BE usually preserves importance patterns, whereas GSA lowers importance by aggregating and removing individual-level variation. The strength of competing predictors also affects these patterns; in cases with balanced signals, importance tends to compress rather than reordering abruptly. Overall, these findings indicate that, in regression tasks, practitioners should be cautious when interpreting variable importance, especially when MsC or aggregated exposures are involved, even if the overall model performance remains stable.

In classification tasks, how measurement error affects feature importance is mainly influenced by the underlying decision structure, especially the intercept, which introduces greater trial-to-trial variability. MsC has the greatest impact, often reducing the importance of the affected predictor to nearly zero when baseline class probabilities reduce dependence on individual features. CE leads to more gradual changes, whereas BE causes consistent shifts in importance across classifiers. GSA falls in between, systematically decreasing importance through aggregation but without causing sudden collapse or reordering. These results emphasize

that, in classification problems, both the type of feature error and the decision boundary configuration should be considered when interpreting post hoc importance measures.

This study examines independent, nondifferential feature measurement error, but real-world data often show correlated, heteroscedastic, or differential errors across features and populations. Adapting the framework to account for these complexities would make it more widely applicable. But if measurement error degrades performance in simple settings, it is more likely to cause similar effects in the more complex settings. Additionally, we assess predictive performance and permutation-based feature importance using naive training; future research could explore explicit error-aware learning methods, such as latent-variable models, SIMEX-type corrections, or joint representation learning, and compare their effectiveness with the same simulation setup. Our current analysis does not consider end-to-end deep learning models, nor does it address calibration, uncertainty quantification, or distribution shift, all of which are especially relevant when upstream models produce features. Lastly, while simulations enable controlled causal interpretation, applying the framework to real datasets with validation samples or partial ground truth, particularly in fields such as health, environmental exposure, and social science, would enhance external validity and support practical application.

5 Broader Impact Statement

Machine learning systems are increasingly used to guide decisions across fields such as healthcare, environmental monitoring, education, the social sciences, and public policy, where data quality can be imperfect. Our research shows that measurement errors in input features consistently and significantly degrade predictive performance and interpretability across various error types and ML modeling methods. This decline in performance is not offset by using larger datasets; rather, the bias introduced by noisy or inaccurately measured variables persists as an intrinsic part of the learning process.

First, our findings have important implications for the creation and deployment of predictive systems. They emphasize that data quality can set fundamental limits on a model’s reliability, regardless of whether advanced algorithms or extensive datasets are used. This indicates that enhancing measurement processes, such as using better sensors, adopting more consistent labeling methods, or explicitly modeling measurement uncertainty, may have a greater impact than merely using big data.

Second, our findings highlight the importance of evaluation methods that mirror real-world data quality conditions. When models are trained or tested on perfect, error-free features, they might overestimate their performance on real-world data, leading to unwarranted confidence in critical fields such as healthcare, environmental forecasting, and social policy analysis.

Finally, our work encourages further investigation into methods that can withstand measurement error and into frameworks for measuring and reporting uncertainty in feature data. Although it does not present new mitigation techniques, it highlights the extent and commonality of the issue, laying the groundwork for future methodological improvements.

This research is unlikely to cause direct negative societal effects. However, neglecting the implications of measurement error in real-world applications can lead to biased or unreliable decisions. By explicitly addressing these risks, our work seeks to promote a more trustworthy, scientifically rigorous use of machine learning.

Acknowledgments and Disclosure of Funding

The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing computational resources that contributed to the research reported in this paper. URL: <http://www.tacc.utexas.edu>. Project allocation number on TACC is DMS25015. No funds were received by the authors to work on this research project.

Conflicts of Interest

The authors declare no conflicts of interest.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author(s) used ChatGPT and Grammarly to edit and correct the English language grammar and flow. After using these tools, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

- Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework, 2019. URL <https://arxiv.org/abs/1907.10902>.
- Jonathan W Bartlett and Ruth H Keogh. Bayesian correction for covariate measurement error: A frequentist evaluation and comparison with regression calibration. Statistical Methods in Medical Research, 27(6): 1695–1708, 2018.
- Leo Breiman. Random forests. Machine Learning, 45(1):5–32, 2001.
- John P Buonaccorsi. Measurement error: models, methods, and applications. Chapman and Hall/CRC, 2010.
- Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. Measurement error in nonlinear models: a modern perspective. Chapman and Hall/CRC, 2006.
- Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39(2): 175–191, 2007. doi: 10.3758/BF03193146.
- Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research, 20(177):1–81, 2019.
- Python Software Foundation. PEP 371 – Addition of the multiprocessing package to the standard library. 2008. URL <https://peps.python.org/pep-0371/>.
- Benoît Frenay and Michel Verleysen. Classification in the presence of label noise: A survey. IEEE Transactions on Neural Networks and Learning Systems, 25(5):845–869, 2014.
- Wayne A Fuller. Measurement error models. John Wiley & Sons, 2009.
- KY Fung and D Krewski. Evaluation of regression calibration and simex methods in logistic regression when one of the predictors is subject to additive measurement error. Journal of Epidemiology and Biostatistics, 4(2):65–74, 1999.
- Y Yi Grace, Aurore Delaigle, and Paul Gustafson. Handbook of measurement error models. CRC Press, 2021.
- Christen M Gray. Use of the Bayesian family of methods to correct for effects of exposure measurement error in polynomial regression models. PhD thesis, London School of Hygiene & Tropical Medicine, 2018.

776 Tammy Jiang, Jaimie L Gradus, Timothy L Lash, and Matthew P Fox. Addressing measurement error in
777 random forests using quantitative bias analysis. American journal of epidemiology, 190(9):1830–1840, 2021.

778 Ruth H Keogh and Jonathan W Bartlett. Measurement error as a missing data problem. In Handbook of
779 measurement error models, pages 429–452. Chapman and Hall/CRC, 2021.

780 Ruth H Keogh, Pamela A Shaw, Paul Gustafson, Raymond J Carroll, Veronika Deffner, Kevin W Dodd,
781 Helmut Küchenhoff, Janet A Tooze, Michael P Wallace, Victor Kipnis, et al. Stratos guidance document
782 on measurement error and misclassification of variables in observational epidemiology: part 1—basic theory
783 and simple methods of adjustment. Statistics in medicine, 39(16):2197–2231, 2020.

784 Walter K Kremers. A general, simple, robust method to account for measurement error when analyzing data
785 with an internal validation subsample. arXiv preprint arXiv:2106.14063, 2021.

786 Russell V. Lenth. emmeans: Estimated Marginal Means, aka Least-Squares Means, 2024. URL <https://CRAN.R-project.org/package=emmeans>. R package version 1.10.0.

788 Yuanyuan Luan, Roger S Zoh, Sneha Jadhav, Lan Xue, and Carmen D Tekwe. Generalized functional
789 linear regression models with a mixture of complex function-valued and scalar-valued covariates prone to
790 measurement error. arXiv preprint arXiv:2304.02651, 2023.

791 Christoph Molnar. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.
792 Online, 2 edition, 2022. URL <https://christophm.github.io/interpretable-ml-book/>.

793 Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy
794 labels. In Advances in Neural Information Processing Systems, pages 1196–1204, 2013.

795 Tamer Oraby, Siva Sivaganesan, Joseph D Bowman, Laurel Kincl, Lesley Richardson, Mary McBride, Jack
796 Siemiatycki, Elisabeth Cardis, and Daniel Krewski. Berkson error adjustment and other exposure surrogates
797 in occupational case-control studies, with application to the canadian interocc study. Journal of exposure
798 science & environmental epidemiology, 28(3):251–258, 2018.

799 Tamer Oraby, Santanu Chakraborty, Siva Sivaganesan, Laurel Kincl, Lesley Richardson, Mary McBride, Jack
800 Siemiatycki, Elisabeth Cardis, and Daniel Krewski. Adjusting for berkson error in exposure in ordinary
801 and conditional logistic regression and in poisson regression. BMC Medical Research Methodology, 23(1):
802 225, 2023.

803 Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep
804 neural networks robust to label noise: A loss correction approach. In Proceedings of the IEEE Conference
805 on Computer Vision and Pattern Recognition, pages 1944–1952, 2017.

806 Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel,
807 Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning
808 in python. Journal of machine learning research, 12(Oct):2825–2830, 2011.

809 Pamela A Shaw, Paul Gustafson, Raymond J Carroll, Veronika Deffner, Kevin W Dodd, Ruth H Keogh,
810 Victor Kipnis, Janet A Tooze, Michael P Wallace, Helmut Küchenhoff, et al. Stratos guidance document on
811 measurement error and misclassification of variables in observational epidemiology: part 2—more complex
812 methods of adjustment and advanced topics. Statistics in medicine, 39(16):2232–2263, 2020.

813 Henrik Singmann, Ben Bolker, Jake Westfall, Frederik Aust, and Mattan S. Ben-Shachar. afex: Analysis
814 of Factorial Experiments, 2023. URL <https://CRAN.R-project.org/package=afex>. R package version
815 1.3-0.

816 Texas Advanced Computing Center TACC. Lonestar6 supercomputer. [https://www.tacc.utexas.edu/](https://www.tacc.utexas.edu/systems/lonestar6/)
817 [systems/lonestar6/](https://www.tacc.utexas.edu/systems/lonestar6/), 2021. The University of Texas at Austin.

A Appendix

A.1 Supplementary Tables

Table A.1: List of Hyperparameters of Regression ML Methods and the Ranges of Values Used in Calibrations.

ML Method	Hyperparameter	Description	Range/Values
Multiple Linear Regression (MLR)	Intercept	Whether to include a y-intercept in the model	True or False
	Positive coefficients	Whether to constrain the coefficients to be positive	True or False
Support Vector Regression (SVR)	λ	Regularization parameter	0.1, 10
	Kernel	Specifies the kernel type to be used in the algorithm	Linear, Polynomial, RBF, Sigmoid
	ϵ	Specifies the ϵ -tube within which there is no penalty	0.01, 1.0
Random Forest Regression (RF-R)	Number of trees	Number of trees in the forest	10 to 200
	Max depth	Maximum depth of the trees	2, 20
	Min samples to split	Minimum number of samples required to split an internal node	2, 5
	Min samples for a leaf	Minimum samples required to be at a leaf node	1, 2
	Max number of features	Number of features considered for best split	Square root or log base 2
Gradient Boosting Regression (XGBoost-R)	Number of estimators	Number of boosting stages	100, 200
	Learning rate	Shrinks contribution of each tree	0.01, 0.1
	Max depth	Maximum depth of individual estimators	3, 5
	Subsample proportion	Fraction of samples used for fitting	0.8, 1.0
	Feature proportion	Fraction of features used for fitting	0.5, 1.0
Multi-layer Perceptron Regression (MLP-R)	Size of hidden layers	Number of neurons in each hidden layer	10 to 100
	Number of hidden layers	Number of hidden layers in the network	1 to 3
	Activation function	Activation function used in the network	Identity, logistic, tanh, relu
	Solver algorithm	Optimization algorithm	lbfgs, sgd, adam
	Learning rate	Learning rate	0.001, 0.01
	α	L2 penalty	0.0001, 0.1
	Max iterations	Maximum number of iterations	500, 2000

Table A.2: List of Hyperparameters of Classification ML Methods and the Ranges of Values Used in Calibrations.

ML Method	Hyperparameter	Description	Range/Values
Logistic Regression (LogReg)	Penalty	Type of regularization	L1 or L2
	λ	Inverse of regularization strength	0.01, 10
	Solver	Algorithm for optimization	lbfgs, saga, liblinear
	Max iterations	Maximum number of iterations	100, 500
Support Vector Machine (SVM)	λ	Regularization parameter	0.1, 10
	Kernel	Kernel type	Linear, Polynomial, RBF, Sigmoid
Random Forest Classification (RF-C)	Number of trees	Number of trees	10 to 200
	Max depth	Maximum depth	2, 20
	Min samples to split	Minimum samples to split internal node	2, 5
	Min samples for a leaf	Minimum samples at a leaf node	1, 2
	Max number of features	Features considered per split	Square root or $\log_2$
Gradient Boosting Classification (XGBoost-C)	Number of estimators	Number of boosting stages	100, 200
	Learning rate	Shrinks contribution	0.01, 0.1
	Max depth	Maximum depth	3, 5
	Subsample proportion	Fraction of samples used for fitting	0.8, 1.0
	Feature proportion	Fraction of features used for fitting	0.5, 1.0
Multi-layer Perceptron Classification (MLP-C)	Size of hidden layers	Neurons in each hidden layer	10 to 100
	Number of hidden layers	Number of hidden layers	1 to 3
	Activation function	Activation function	logistic, tanh, relu
	Solver algorithm	Optimization algorithm	lbfgs, sgd, adam
	Learning rate	Learning rate	0.001, 0.01
	α	L2 penalty	0.0001, 0.1
	Max iterations	Maximum number of iterations	500, 2000

A.2 Further Descriptive Figures

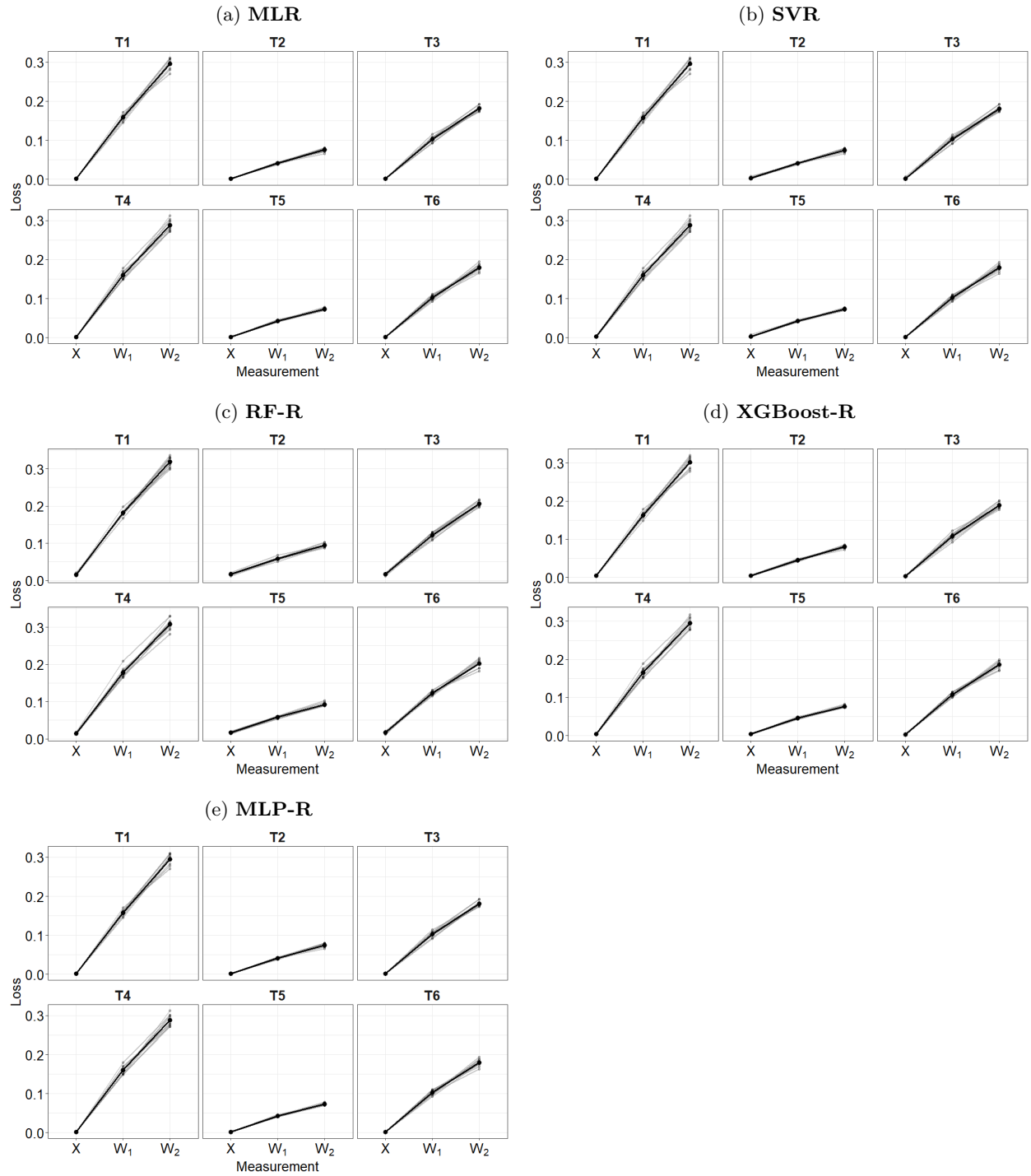


Figure A.1: Line graphs of the MSE for the MCE of CE with $\sigma = 0, 0.5, 0.75$ at trials T1–T6 and for (a) MLR, (b) SVR, (c) RF-R, (d) XGBoost-R, and (e) MLP-R.

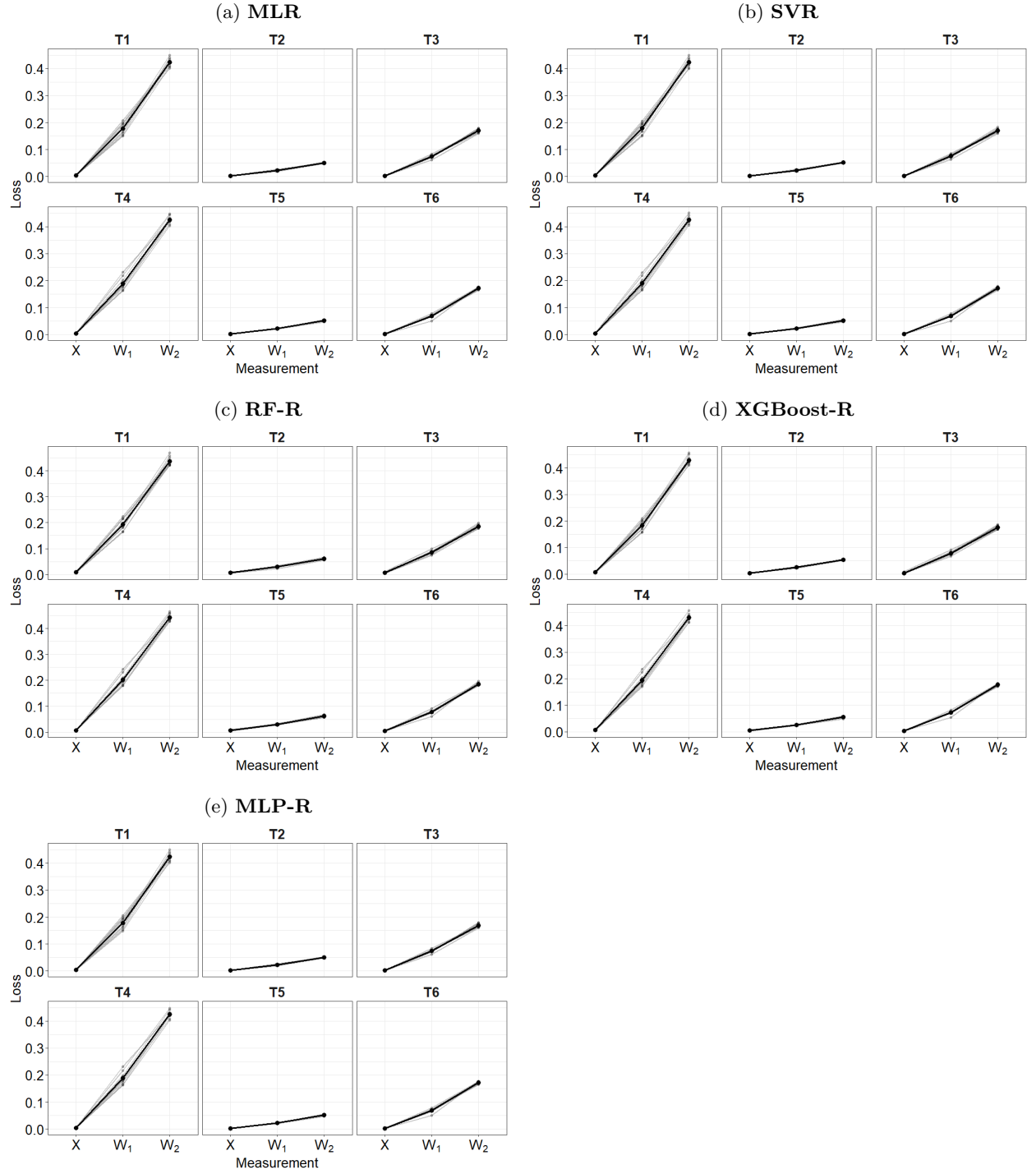


Figure A.2: Line graphs of MSE for the MCE of MsC with $q = 0, 0.1, 0.3$ at trials T1–T6 and for (a) MLR, (b) SVR, (c) RF-R, (d) XGBoost-R, and (e) MLP-R.

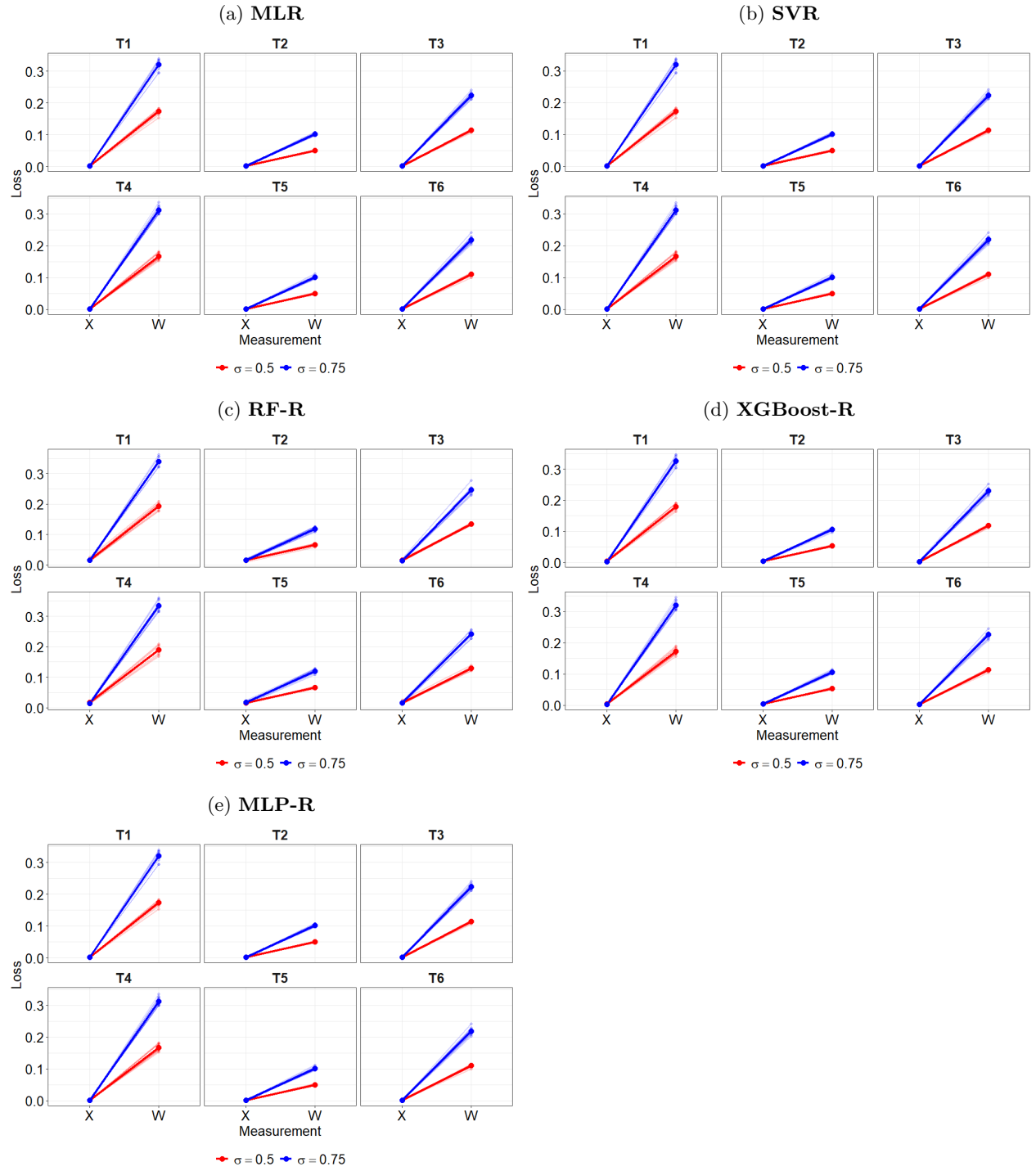


Figure A.3: Line graphs of the MSE for the MCE of BE with $\sigma = 0, 0.5, 0.75$ at trials T1-T6 and for (a) MLR, (b) SVR, (c) RF-R, (d) XGBoost-R, and (e) MLP-R.

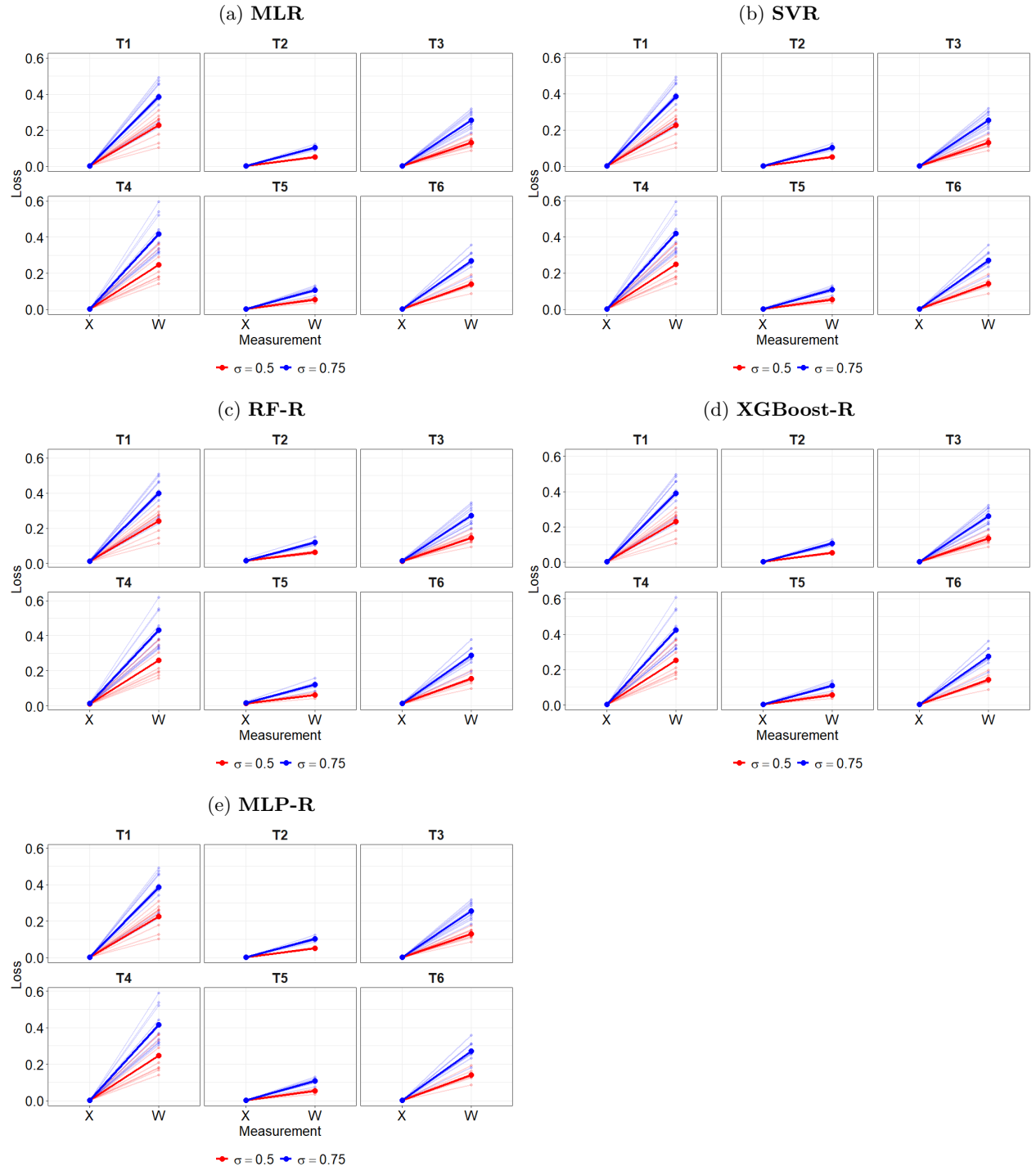


Figure A.4: Line graphs of the MSE for the MCE of GSA with $\sigma = 0, 0.5, 0.75$ at trials T1–T6 and for (a) MLR, (b) SVR, (c) RF-R, (d) XGBoost-R, and (e) MLP-R.

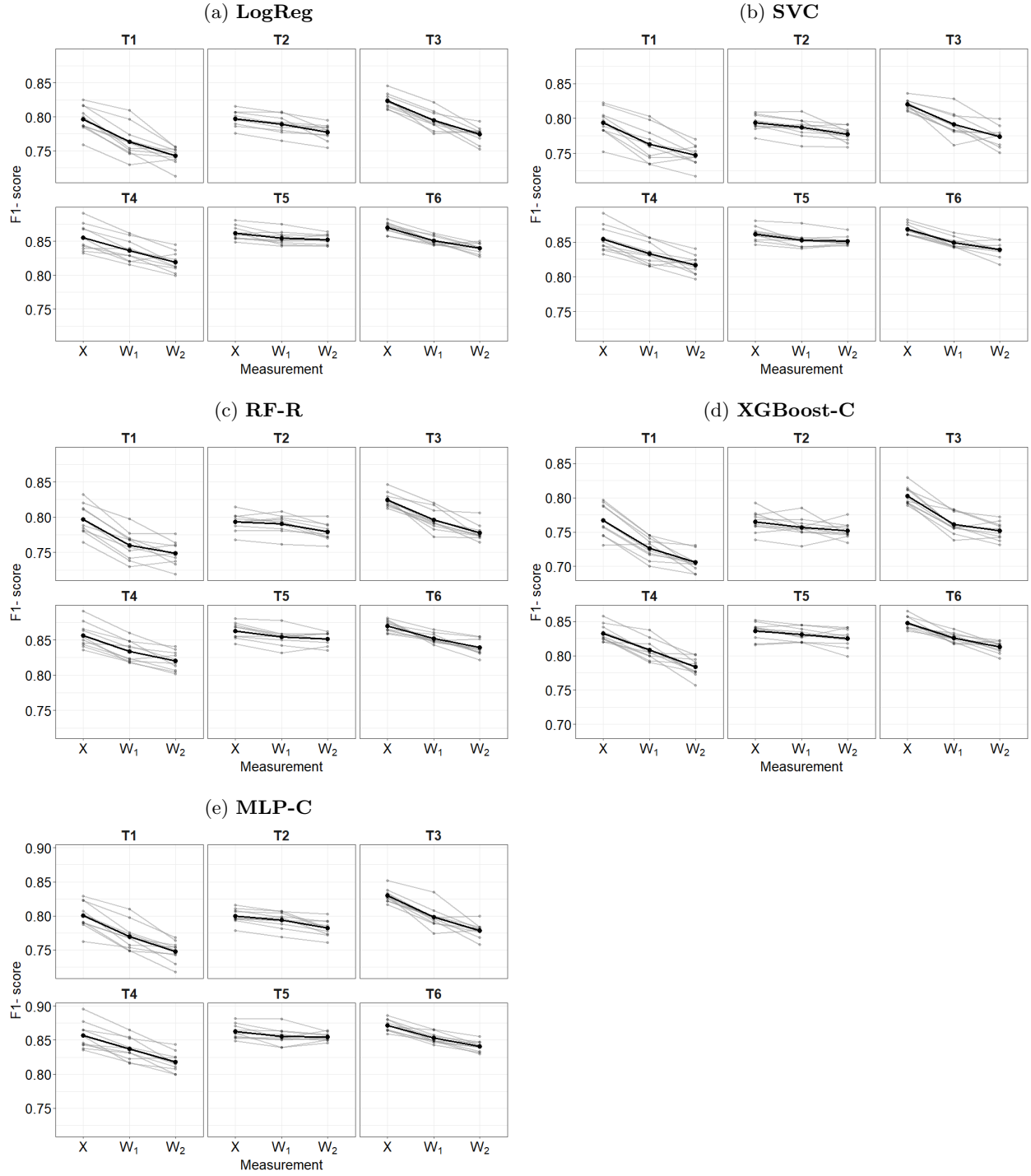


Figure A.5: Line graphs of F1-score for the MCE of CE with $\sigma = 0, 0.5, 0.75$ at trials T1–T6 and for (a) LogReg, (b) SVR, (c) RF-R, (d) XGBoost-R, and (e) MLP-C.

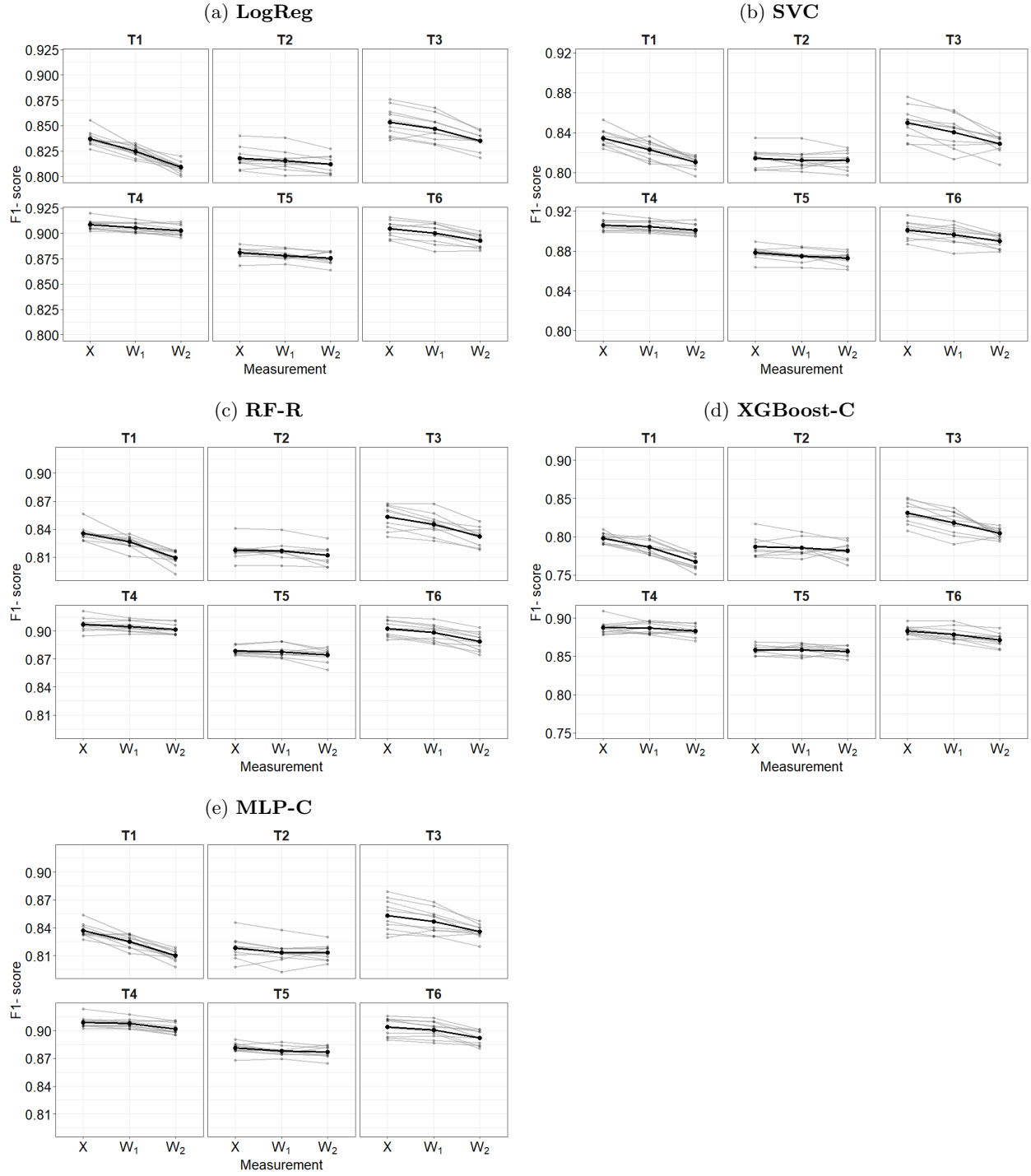


Figure A.6: Line graphs of F1-score for the MCE of MsC with $q = 0, 0.1, 0.3$ at trials T1-T6 and for (a) LogReg, (b) SVR, (c) RF-R, (d) XGBoost-R, and (e) MLP-C.

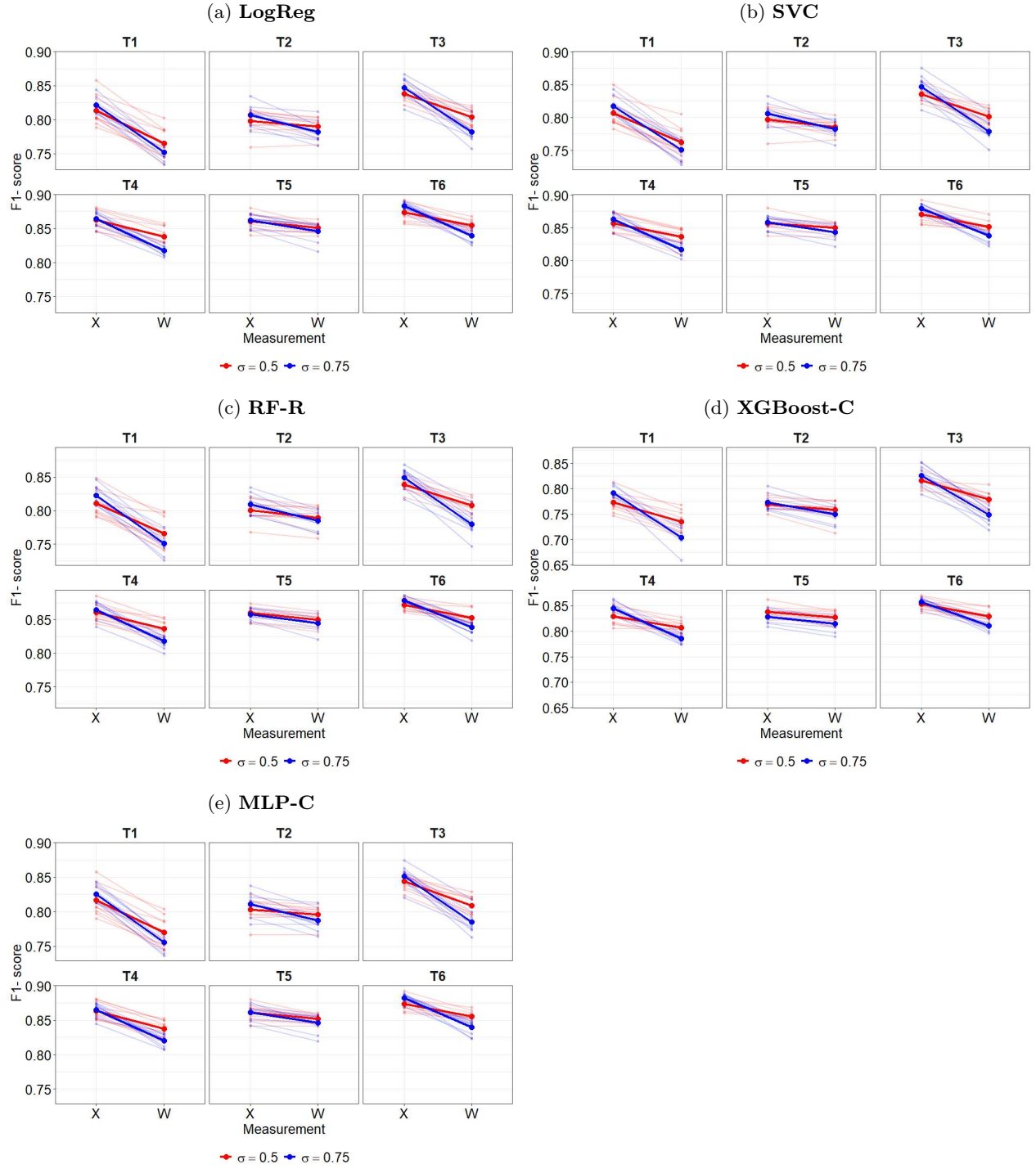


Figure A.7: Line graphs of F1-score for the MCE of BE with $\sigma = 0, 0.5, 0.75$ at trials T1–T6 and for (a) LogReg, (b) SVR, (c) RF-R, (d) XGBoost-R, and (e) MLP-C.

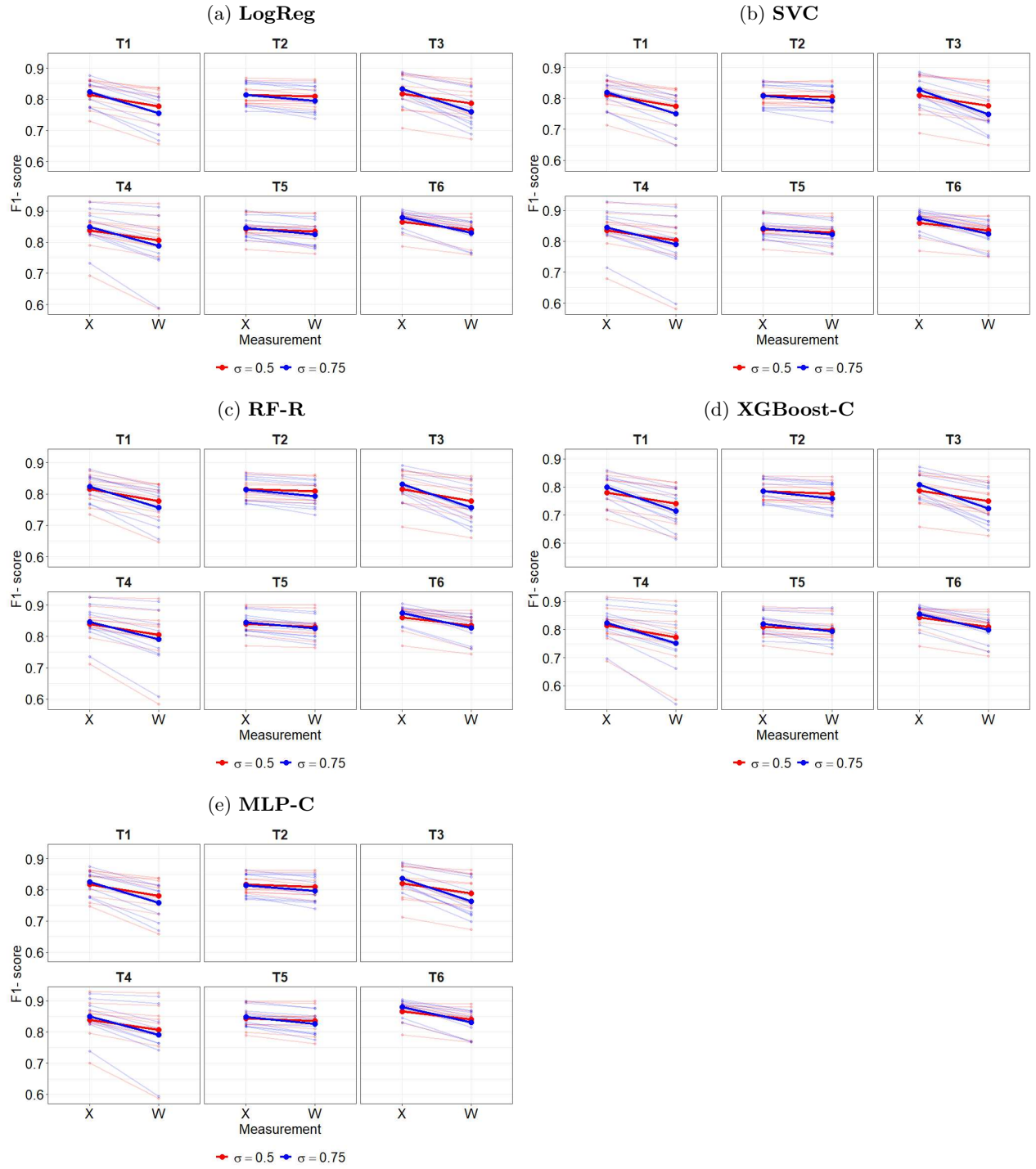


Figure A.8: Line graphs of F1-score for the MCE of GSA with $\sigma = 0, 0.5, 0.75$ at trials T1–T6 and for (a) LogReg, (b) SVR, (c) RF-R, (d) XGBoost-R, and (e) MLP-C.

B Further Analyses

B.1 Classical Error (CE).

The findings of the statistical analyses are further supported by contrasts' confidence intervals (CIs) summarizing the change in performance between measurement conditions under the classical error assumption. Figure B.1 (a) and (b) display 95% CIs for contrasts in ΔR^2 and ΔMSE for regression, and (c) and (d) display 95% CIs for contrasts in $\Delta \text{Accuracy}$ and ΔF1 for classification across trials and models. Consistent with the ANOVA results, the contrasts show a systematic degradation in performance as measurement error increases from X to W_1 and W_2 , with confidence intervals largely excluding zero across trials. The magnitude of these contrasts varies across models and trials, but their direction is stable, confirming that classical error induces a robust, practically meaningful reduction in predictive performance rather than a trial-specific artifact.

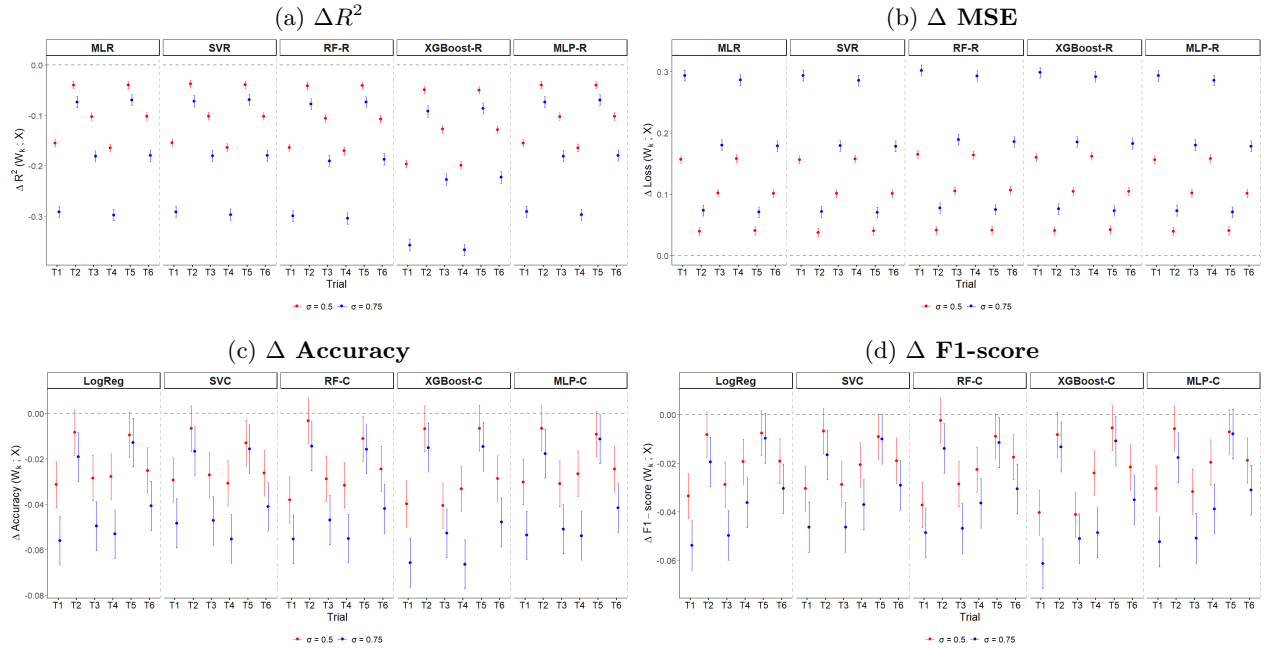


Figure B.1: Contrast confidence intervals (CIs) for model performance across trials under CE. The top row reports regression metrics, including (a) ΔR^2 and (b) ΔMSE , while the bottom row reports classification metrics, including (c) $\Delta \text{Accuracy}$ and (d) $\Delta \text{F1-score}$. Each panel corresponds to a different learning algorithm, with points indicating estimated contrasts and error bars representing 95% confidence intervals across trials. Results are shown for two error levels ($\sigma = 0.5$ and $\sigma = 0.75$).

B.2 Misclassification (MsC).

These patterns are further illustrated by contrast confidence intervals (CIs) summarizing performance differences under MsC. Figure B.2, panels (a) and (b), display 95% CIs for contrasts in ΔR^2 and ΔMSE for regression, and (c) and (d) display 95% CIs for contrasts in $\Delta \text{Accuracy}$ and ΔF1 for classification across trials and regression models. Consistent with the ANOVA results, the contrasts show substantial reductions in explanatory power and corresponding increases in prediction loss when moving from the baseline X to misclassified inputs W_k . The magnitude of these effects varies across trials and models, with greater degradations observed at higher misclassification levels, but the direction of the contrasts remains stable across experimental conditions, indicating that misclassification produces a robust and practically meaningful deterioration in regression performance.

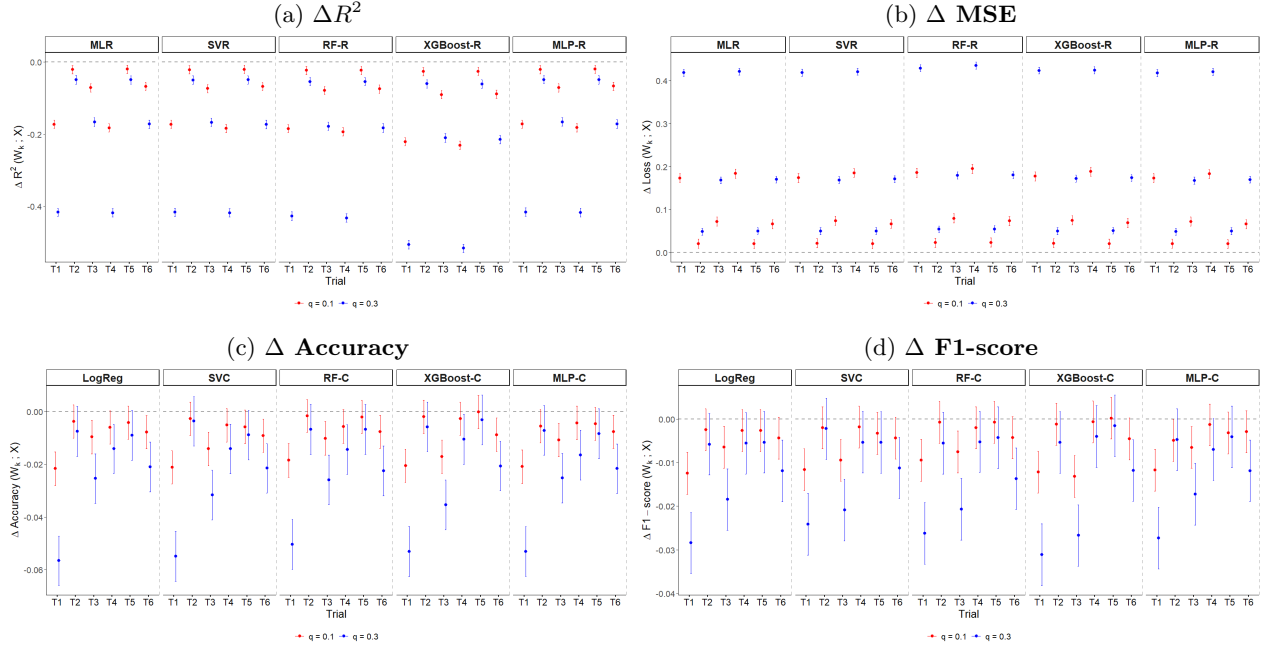


Figure B.2: Contrast confidence intervals (CIs) for model performance across trials under MsC. The top row reports regression metrics, including (a) ΔR^2 and (b) ΔMSE , while the bottom row reports classification metrics, including (c) $\Delta \text{Accuracy}$ and (d) $\Delta \text{F1-score}$. Each panel corresponds to a different learning algorithm, with points indicating estimated contrasts and error bars representing 95% confidence intervals across trials. Results are shown for two error levels ($q = 0.1$ and $q = 0.3$).

B.3 Berkson Error (BE).

The reported Measurement effect corresponds to an overall repeated-measures test across all four measurement conditions (X_1, X_2, W_1, W_2), rather than separate pairwise tests for the transitions $X_1 \rightarrow W_1$ and $X_2 \rightarrow W_2$. Accordingly, a single p -value is reported for the Measurement factor, while differences in the severity of performance degradation across specific transitions are evaluated using directional contrasts and their associated confidence intervals. Figure B.3, panels (a) and (b), displays 95% contrast confidence intervals (CIs) for ΔR^2 and ΔMSE for regression, and (c) and (d) display 95% CIs for contrasts in $\Delta \text{Accuracy}$ and ΔF1 for classification across trials and regression models. Consistent with the omnibus ANOVA results, the contrasts show pronounced reductions in explanatory power and corresponding increases in prediction loss as inputs transition from the baseline conditions (X_1, X_2) to Berkson-contaminated representations (W_1, W_2). The magnitude of degradation varies across trials and model classes, but the direction of the contrasts is stable: W_2 produces the largest performance deterioration, and W_1 exhibits a smaller yet systematic effect. Together, these CI-based contrasts confirm that Berkson error induces a robust, practically meaningful decline in regression and classification performance across experimental settings.

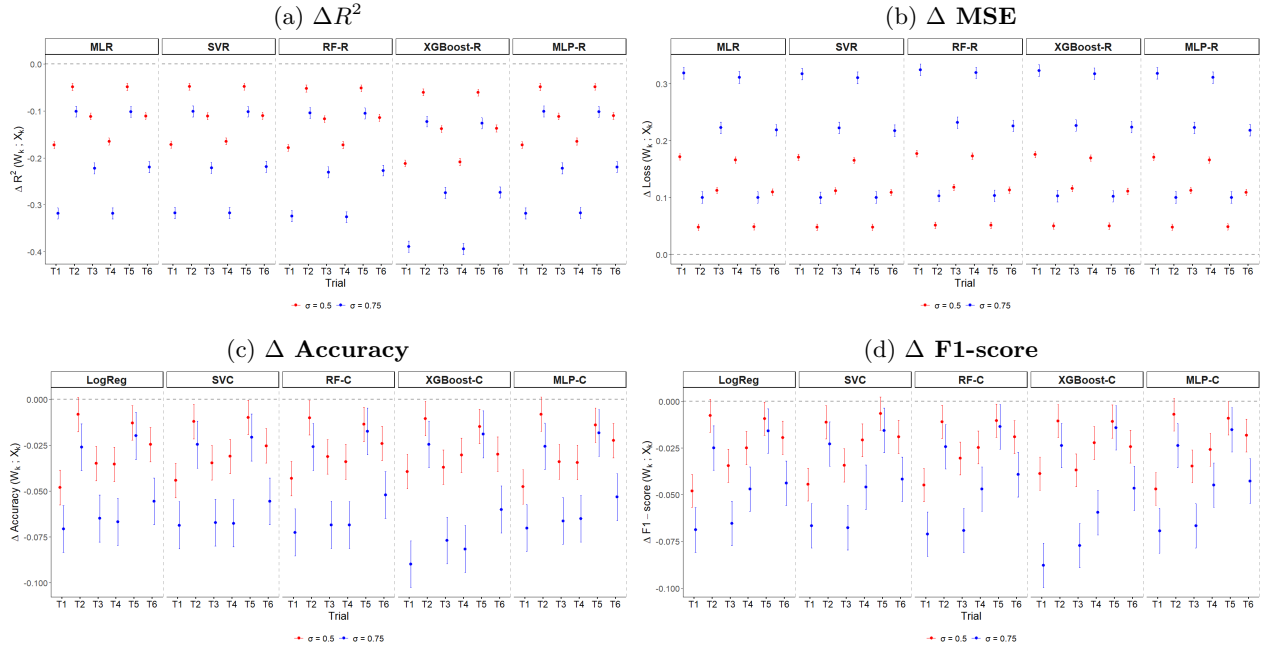


Figure B.3: Contrast confidence intervals (CIs) for model performance across trials under BE. The top row reports regression metrics, including (a) ΔR^2 and (b) ΔMSE , while the bottom row reports classification metrics, including (c) $\Delta \text{Accuracy}$ and (d) $\Delta \text{F1-score}$. Each panel corresponds to a different learning algorithm, with points indicating estimated contrasts and error bars representing 95% confidence intervals across trials. Results are shown for two error levels ($\sigma = 0.5$ and $\sigma = 0.75$).

B.4 Group-Summary Assignment (GSA).

As in the BE analysis, the reported Measurement effect corresponds to an omnibus repeated-measures test across all four measurement conditions (X_1, X_2, W_1, W_2), rather than separate pairwise tests for the transitions $X_1 \rightarrow W_1$ and $X_2 \rightarrow W_2$. Accordingly, a single p -value is reported for the Measurement factor, while differences in the severity of performance degradation across specific transitions are evaluated using directional contrasts and their associated confidence intervals. Figure B.4, panels (a) and (b), display 95% contrast confidence intervals (CIs) for ΔR^2 and ΔMSE for regression, and (c) and (d) display 95% CIs for contrasts in $\Delta \text{Accuracy}$ and ΔF1 for classification across trials and regression models. Consistent with the omnibus ANOVA results, the contrasts show pronounced reductions in explanatory power and corresponding increases in prediction loss when moving from the baseline inputs (X_1, X_2) to group-assigned representations (W_1, W_2). While the magnitude of these effects varies across trials, their direction is stable across models, indicating that GSA induces a robust and practically meaningful degradation in regression and classification performance.

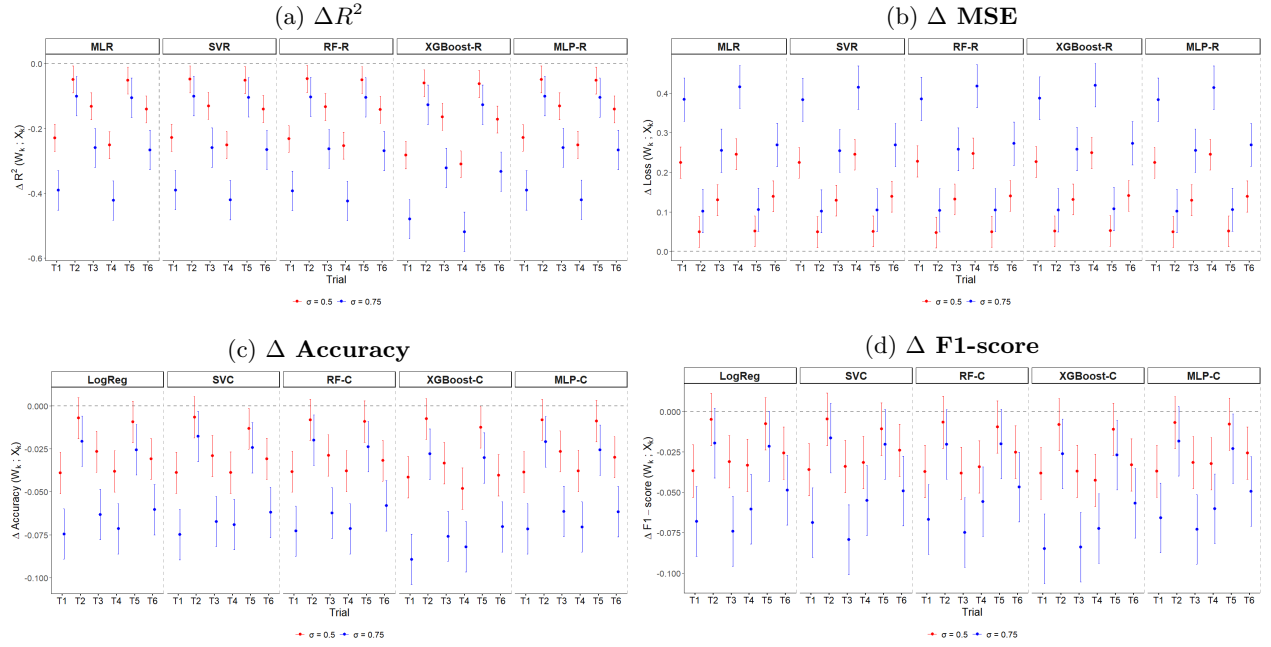


Figure B.4: Contrast confidence intervals (CIs) for model performance across trials under GSA. The top row reports regression metrics, including (a) ΔR^2 and (b) ΔMSE , while the bottom row reports classification metrics, including (c) $\Delta \text{Accuracy}$ and (d) $\Delta \text{F1-score}$. Each panel corresponds to a different learning algorithm, with points indicating estimated contrasts and error bars representing 95% confidence intervals across trials. Results are shown for two error levels ($\sigma = 0.5$ and $\sigma = 0.75$).