

ROBUST PREDICTIVE MODELING UNDER DATA HETEROGENEITY: INTEGRATING STATISTICAL INFERENCE, ENSEMBLE LEARNING, AND UNCERTAINTY QUANTIFICATION

Dr. Alexander Bennett^{1*}, Dr. Lukas Schneider², Dr. Claire Dubois³

¹*Department of Data Science, Example University, London, United Kingdom*

²*Institute of Applied Statistics, Example University, Munich, Germany*

³*Department of Quantitative Methods, Example Institute, Paris, France*

***Corresponding Author: Dr. Alexander Bennett**

Abstract

Robust predictive modeling requires simultaneous consideration of data heterogeneity, predictive performance, and uncertainty. This study developed an integrated framework combining statistical inference, ensemble learning, and uncertainty assessment using heterogeneous bank-marketing data. After duplicate removal, 41,176 observations were analyzed, with successful subscriptions accounting for 11.27% of cases. Logistic regression, Random Forest, XGBoost, and a soft-voting ensemble were evaluated using discrimination, classification, and probability-based metrics. Contact duration was excluded from the primary prospective models and subsequently included in sensitivity analysis. Random Forest achieved the highest prospective ROC-AUC (0.813) and PR-AUC (0.486), whereas XGBoost obtained the lowest Brier score (0.076). The soft-voting ensemble achieved the highest F1-score (0.526) and Matthews correlation coefficient (0.462). Bootstrap confidence intervals showed considerable overlap among model discrimination estimates, indicating uncertainty around small performance differences. Inclusion of contact duration increased ROC-AUC to 0.940–0.950, demonstrating that predictor availability affected performance more strongly than algorithm selection. These findings emphasize the importance of evaluating predictive models through multiple performance measures, uncertainty estimates, and realistic predictor specifications. The proposed framework provides a statistically grounded approach for assessing predictive robustness in heterogeneous data environments.

Keyword: - Data heterogeneity, Ensemble learning, Predictive modeling, Statistical inference, Uncertainty quantification

INTRODUCTION

In situations where observations are the result of heterogeneous demographic, behavioral, socioeconomic, operational, and contextual factors, more and more, predictive modeling is used to support decision making. This heterogeneity poses a crucial statistical problem as the functional, numerical and complexity of the relationships between predictors and outcomes can vary across the feature space. While conventional statistical models are still important for interpreting potential associations and assessing uncertainty, they often have a fixed functional form, which may not capture nonlinearities and higher order interactions. On the other hand, machine-learning models are able to model complex predictive relationships and may offer less direct inferential interpretation. Thus, when developing robust prediction, it is important to take into account not only the accuracy of the prediction but also the stability and uncertainty of prediction models and the nature of the data.

Heterogeneous response patterns may be caused, in part, by individual background characteristics. In the field of education, for instance, studies have shown that background factors of learners can be important factors in determining the motivational process and outcomes (Kashefian-Naeeni et al., 2024; Yuniar, 2021). These studies are about educational populations and not necessarily about predictive banking applications, but they are examples of an overarching analytical challenge: statistics derived from a diverse population can be systematically different and this fact should not be hidden by aggregate statistics. Hence, statistical inference has continued to play a crucial role in the identification of conditional association and quantification of uncertainty prior to drawing conclusions, which then can be translated into predictive decisions.

The ensemble learning is another approach which combines the information of several predictive learners. The idea of using heterogeneous ensembles is especially designed to make use of the differences in the component models and data representations. Salama et al. (2022) came up with a combination of heterogeneous learning and active learning, and Huang et al. (2024) showed how to do collaborative ensemble learning of heterogeneous large language models. In fraud detection, distribution-aware heterogeneous ensembles have also been shown, indicating that the diverse learners are useful when the data have complex structures that affect classification (Xie et al., 2021). These findings indicate that the diversity of models can help to deliver valuable predictive information, information which may not be available through a single statistical or machine-learning algorithm.

Additional examples of heterogeneous ensemble strategies exist in other application areas, also illustrating the flexibility. These methods have been explored in breast cancer prediction (Nanglia et al., 2022), landslide susceptibility mapping (Fang et al., 2021), missing data estimation for IoV (Internet of Vehicles) system (Zhang et al., 2021) and traffic-accident duration prediction (Zhao & Deng, 2022). Instead of using either statistical or machine-learning methods in isolation, Ahmad et al. (2023) demonstrated how these two methods could be combined in a stacking framework for crash forecasting. Likewise, the heterogeneous ensemble methods have been investigated for student-performance prediction (Evangelista & Sy, 2022), and imbalance-oriented heterogeneous ensembles have been created to enhance learning when classes of the outcomes are not balanced (Zefrehi & Altınçay, 2020).

Ensemble modeling has also been successfully applied to software-defect prediction and risk estimation in health. Chen et al. (2022) presented a balanced multiset ensemble model for heterogeneous defect prediction and Cui et al. (2021) designed a nested heterogeneous ensemble model to predict the mortality rate of COVID-19. All together, these studies indicate a significant interest in increasing the accuracy of prediction by having to do with diversity of models, balancing methods, and integrated learning architectures. But, predictive performance is not the only indicator of model robustness. These small discrepancies among the algorithms may prove to be unstable from sample to sample, may be sensitive to the predictors available at the time of the prediction, or may be linked to probability estimates that are uncertain. These factors gain additional significance if there is outcome imbalance and heterogeneous data.

A more complete predictive framework should therefore be tied with the statistical inference, ensemble learning and uncertainty quantification. The statistical models offer interpretable benchmarks and ensemble algorithms can model non-linear and interactive effects; Uncertainty assessment identifies if the apparent performance differences are still credible given sampling variation. Alternative predictor specifications also can be used to assess whether there is any added value in the modelling or in information that may not be available in a prospective application. This holistic view is important in relation to heterogeneous tabular data that can lead to different views on the behaviour of a model, such as discrimination, class-specific performance, probability accuracy and robustness.

In this context, this research aims at building a powerful predictive model for the heterogeneous data using the statistical and ensemble learning techniques along with the explicit uncertainty and sensitivity analyses. A multivariate binary-classification framework is used to test the ability of conventional statistical modeling and nonlinear ensemble approaches to predict and to analyze their stability when the information structure is different. The study had three objectives:

- □ To characterize heterogeneous data patterns and their relationships with the binary outcome.
- □ To compare statistical, machine-learning, and ensemble models across complementary predictive metrics.
- □ To quantify predictive uncertainty and assess robustness under alternative predictor specifications.

2. MATERIALS AND METHODS

2.1 Data Source and Study Design

A secondary quantitative analysis was performed using a multivariate bank-marketing data set, which included demographic, socioeconomic, credit, campaign, temporal and macroeconomic variables (Naim, 2026). The binary outcome was whether or not an individual had a term deposit. A total of 41,176 observations after removing duplicate records were retained with 4,639 positive (11.27%) and 36,537 negative (88.73%) outcomes.

2.2 Data Preparation

Duplicates, distributions of categorical data, and data type consistency were checked for variables. Entries that were marked “unknown” were kept as stand-alone categories. Categorical data were coded for modelling and numerical data were transformed as per the model requirement. The outcome distribution was similar in stratified data partitioning. Contact duration was not included in the primary models to create a setting for prospective prediction as it is only available after customer contact.

2.3 Statistical and Predictive Modeling

Data for continuous variables were presented as mean $\pm$ SD; for categorical, the values were presented as number and percentage. Logistic regression was the statistical baseline with Random Forest and XGBoost capturing interactions between predictors and nonlinear relationships. In order to assess whether the integration of models enhanced the classification performance, an ensemble of models was performed using soft voting method to combine the probability estimates given by each model.

2.4 Model Evaluation and Uncertainty Quantification

ROC-AUC, PR-AUC, Sensitivity, Specificity, F1-score, Matthews correlation coefficient and Brier score were used to assess the predictive performance. The discrimination of the models was compared using ROC curves. Uncertainties in the predictions were determined using 300 bootstrap resamples and 95% confidence intervals were computed using the empirical method. This enabled differences between models to be interpreted in the context of the sampling variability.

2.5 Robustness Analysis

A prospective specification (excluding contact duration) was compared to a full-information specification (including contact duration) via sensitivity analysis. The trained algorithms were Logistic regression, Random Forest and XGBoost for both conditions to gain insight into performance difference due to algorithm types or availability of predictors. This analysis was a direct test of the ability of this model to discriminate against other information structures.

2.6 Statistical Analysis

Data was analyzed using python packages pandas, numpy, scikit-learn, statsmodels, xgboost and matplotlib. Thresholds for classification-based measures were set at 0.50 and for ROC-AUC, PR-AUC and Brier scores, predictions were based on probabilities. Preprocessing and evaluation was done in the same way for all models to allow for consistent and repeatable comparisons.

3. RESULTS

3.1 Sample Characteristics and Data Heterogeneity

The number of observations after filtering out 12 duplicated records was 41,176. The number of successful subscription is 4,639 (11.27%) which shows significant imbalance in classes. The mean age was 40.02 ± 10.42 years. Predictors included demographic, socioeconomic, credit, campaign, temporal and macroeconomic dimensions. The major characteristics of the analytical sample are given in Table 1.

Table 1. Characteristics of the Analytical Sample

Characteristic	Value
Observations	41,176
Successful subscriptions, n (%)	4,639 (11.27%)
Non-subscriptions, n (%)	36,537 (88.73%)
Age, mean $\pm$ SD (years)	40.02 ± 10.42
Contact duration, mean $\pm$ SD (s)	258.32 ± 259.31
Categorical “unknown” entries	12,716

There was significant variation in the outcomes of the outcome rates for previous campaigns. As seen in Figure 1, the number of subscribers after a successful campaign is 65.11%, after failure is 14.23% and when no campaign was observed was 8.83%.

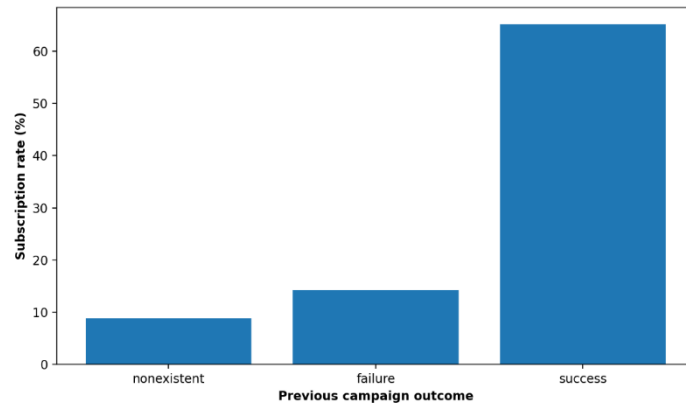


Figure 1. Response Heterogeneity by Previous Campaign Outcome

3.2 Predictive Performance

The models were evaluated mainly without contact duration in order to keep the models in a prospective prediction setting. Finally, from Table 2, it can be seen that the ROC-AUC and PR-AUC is found to be highest for Random Forest with the values 0.813 and 0.486 respectively. The lowest Brier score was obtained by XGBoost (0.076) while the ensemble had the highest F1-score (0.526) and MCC (0.462).

Table 2. Comparative Performance of Predictive Models

Model	ROC-AUC	PR-AUC	Sensitivity	Specificity	F1	MCC	Brier
Logistic regression	0.800	0.443	0.645	0.855	0.463	0.394	0.162
Random Forest	0.813	0.486	0.633	0.892	0.509	0.444	0.138
XGBoost	0.810	0.480	0.224	0.986	0.336	0.349	0.076
Ensemble	0.809	0.479	0.594	0.916	0.526	0.462	0.108

Fig. 2 reveals that there is no significant difference among the machine learning models in terms of discrimination, although Random Forest is expected to achieve the highest ROC-AUC.

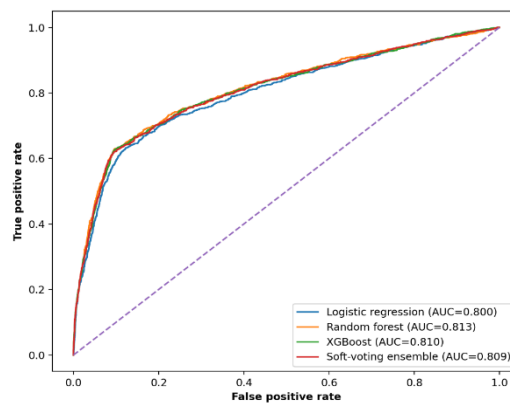


Figure 2. ROC Curves of the Prospective Predictive Models

3.3 Uncertainty and Robustness

Discrimination estimates by the bootstrap seemed fairly stable. Overlapping 95% confidence intervals (CIs) for the competing models, as shown in Table 3, indicate that small differences in ROC-AUC need to be interpreted with caution. Contact duration was a significant inclusion term for discrimination of models.

Table 3. Predictive Uncertainty and Robustness Analysis

Model	Prospective ROC-AUC (95% CI)	Full-information ROC-AUC
Logistic regression	0.800 (0.780–0.818)	0.940
Random Forest	0.813 (0.797–0.832)	0.948
XGBoost	0.810 (0.794–0.825)	0.950
Ensemble	0.809 (0.790–0.824)	—

As seen in Figure 3, the addition of contact duration led to an increase in ROC-AUC from ~0.80–0.81 in the prospective specification to 0.94–0.95. This sensitivity was much more significant than the interalgorithm differences, indicating that the availability of the predictors had a significant impact on apparent predictive performance.

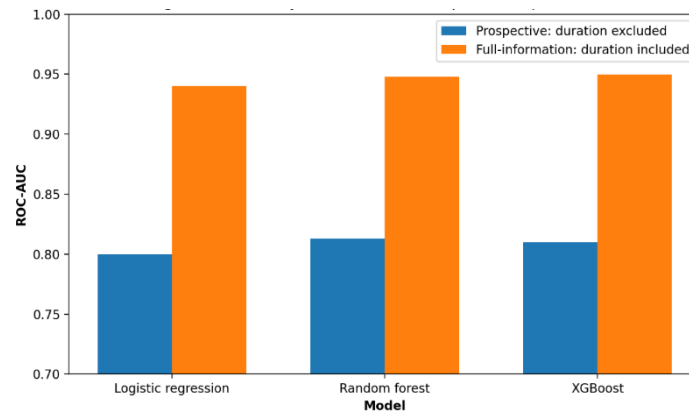


Figure 3. Robustness of Predictive Performance Across Predictor Specifications

4. DISCUSSION

In this study, robust prediction in the context of heterogeneous data conditions was investigated by combining statistical modeling, ensemble learning and uncertainty assessment. With a bank marketing dataset (UCI) (Naim, 2026), the analysis showed that the predictive performance was related to both the algorithms being used and the knowledge present when making predictions. XGBoost was the best in terms of Brier score (0.076 under the prospective specification) while Random Forest had the best discrimination (ROC-AUC = 0.813; PR-AUC = 0.486). The soft-voting ensemble got the best F1-score (0.526) and Matthews correlation coefficient (0.462). These differences suggest that none of the models performed best on all aspects of performance. Furthermore, there was overlap of the bootstrap confidence intervals, which indicated that relative differences in the point estimates should be viewed with caution.

This is all in a context of the heterogeneous structure of the data. All the demographic, socioeconomic, credit, campaign, temporal and macroeconomic features were modeled together simultaneously and the binary response was highly imbalanced. It has been shown that heterogeneous data distributions can have an impact on the stability and reliability of models (Zawad et al., 2021; Qian et al., 2020). In recent years, explicit modelling of data heterogeneity has been highlighted as a key aspect of trustworthy machine learning (Liu & Cui, 2025). There have been also developed robustness-oriented approaches for heterogeneous federated learning such as data-free robustness distillation and model-segmentation strategies (Su et al., 2025) and (Wang et al., 2023) respectively. While each of these studies is conducted in a different computational environment, they all confirm the current result that it is important to study predictive behaviour under varying data and information conditions.

Logistic regression had a prospective ROC-AUC of 0.800, and, despite the mixed predictor set, had good discriminatory power. There were comparatively modest improvements in ROC-AUC for Random Forest and XGBoost. Thus, the more complex the algorithms, the more discrimination was not the more. When the predictions from a machine learning algorithm are combined with inferential procedures, it is important to keep a statistical benchmark that can be easily interpreted. Prediction-powered inference is a good example of how the predictions generated by the machine can be used alongside classical statistical information, without changing the inferential goals (Angelopoulos et al., 2023). Conversely, active statistical inference deals with the interplay between inference and adaptively collected data (Zrníc & Candès, 2024). Based on the current results, there is a need for statistical and machine learning techniques to be used in a complementary manner, and not as mutually exclusive techniques.

The interpretation of model performance was also modified with the uncertainty assessment. Random Forest outperformed the other models in terms of the prospective ROC-AUC, whereas XGBoost was the least performing model for the Brier score while the ensemble performed best for both F1-score and MCC. Emphasis on ROC-AUC alone would therefore mask significant differences in the accuracy of probability predictions and classification behaviour. This is in line with the increasing focus on predictive models with uncertainty. The uncertainty calibration and conformalization can be handled by computational approaches that offer a way to supplement point predictions with reliability information (Mendil et al., 2023). In the realm of uncertainty-aware prediction, predictive methods based on conformal machine learning have also shown to be useful for uncertainty-aware predictions in various applications (Tamuly & Nava, 2025), and uncertainty-aware process prediction (Majlatow et al., 2025) has also been highlighted. In the current study, bootstrap confidence intervals were used to give an empirical assessment of sampling uncertainty, and to minimize the likelihood of rejection of a sample because of a small numerical change as evidence of an actual improvement.

The most transparent methodological result was from the robustness analysis. ROC-AUC was improved by the addition of contact duration from 0.800 to 0.940 for logistic regression, 0.813 to 0.948 for Random Forest and 0.810 to 0.950 for XGBoost. These increases were significantly higher than the algorithm differences with the prospective specification. So, in this case, the choice of predictors (not the choice of models) seemed to be more important for apparent predictive improvement. This observation is echoed by other findings that suggest that data structure can significantly affect model robustness (Qian et al., 2020, Zawad et al., 2021). So reporting just the full-information models may be misleading with respect to potential predictive power since the length of contact is not fully known in advance of the interaction.

Overall, the results validate a multi-dimensional evaluation of prediction. There were six different measures of model quality (discrimination, precision-recall performance, classification balance, probability error, uncertainty, and sensitivity to predictor availability). This aligns with the emerging trend of designing trustworthy models in the presence of heterogeneous data (Liu & Cui, 2025) and uncertainty-aware models for prediction (Majlatow et al., 2025;

Tamuly & Nava, 2025). Model evaluation should therefore be treated with caution to be sure that apparent improvements are not lost as a result of realistically imposed information restrictions. There are a number of constraints to take into account. The study was based on one single observatory dataset (Naim, 2026) thus restricting the generalisation of the results to other groups of people and applications. Some classification measures were found to be sensitive to the decision threshold and outcome prevalence as only 11.27% of the observations were classified as positive class. Categorical values of “unknown” were coded as such, but the reasons for this could not be determined. Bootstrap intervals quantified sampling uncertainty but were not representative of all kinds of predictive uncertainty. Future studies are needed to externally validate the framework using independent heterogeneous datasets and to further develop uncertainty estimation using formal calibration and conformal prediction. These advances may bring even stronger ties between powerful machine learning and statistically sound uncertainty quantification and inference (Angelopoulos et al., 2023, Mendil et al., 2023).

5. CONCLUSION

The results of this study show that the best predictive model with the most discriminative algorithm for a set of heterogeneous data is insufficient. Random Forest had the best prospective discrimination (ROC-AUC 0.813, PR-AUC 0.486) and the soft-voting ensemble had the highest F1-score and Matthews correlation coefficient and the lowest Brier score among the 41,176 observations. The differences suggest that the success of models were not always consistent favoring one model approach over another based on the evaluation criterion. When modelling small differences in performance, it was important that uncertainty was accounted for in the interpretation of the models, which was evident from the substantial overlap seen in the bootstrap confidence intervals. More importantly, however, the inclusion of contact duration resulted in a greater increase in ROC-AUC (from approximately 0.80 to 0.81 to 0.94 to 0.95), than did the choice of algorithm. In summary, the use of statistical baselines, ensemble learning, uncertainty estimation and sensitivity analysis gives a more solid foundation for assessing predictive models under heterogeneous data. Future research should make use of other data sets to validate the framework and further develop uncertainty assessment by external validation, probability calibration, conformal prediction.

REFERENCES

1. Ahmad, N., Wali, B., & Khattak, A. J. (2023). Heterogeneous ensemble learning for enhanced crash forecasts—A frequentist and machine learning based stacking framework. *Journal of safety research*, 84, 418-434.
2. Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., & Zrníc, T. (2023). Prediction-powered inference. *Science*, 382(6671), 669-674.
3. Chen, H., Jing, X. Y., Zhou, Y., Li, B., & Xu, B. (2022). Aligned metric representation based balanced multiset ensemble learning for heterogeneous defect prediction. *Information and Software Technology*, 147, 106892.
4. Cui, S., Wang, Y., Wang, D., Sai, Q., Huang, Z., & Cheng, T. C. E. (2021). A two-layer nested heterogeneous ensemble learning predictive method for COVID-19 mortality. *Applied Soft Computing*, 113, 107946.
5. Evangelista, E., & Sy, B. D. (2022). An approach for improved students' performance prediction using homogeneous and heterogeneous ensemble methods. *International Journal of Electrical and Computer Engineering*, 12(5), 5226.
6. Fang, Z., Wang, Y., Peng, L., & Hong, H. (2021). A comparative study of heterogeneous ensemble-learning techniques for landslide susceptibility mapping. *International Journal of Geographical Information Science*, 35(2), 321-347.
7. Huang, Y., Feng, X., Li, B., Xiang, Y., Wang, H., Liu, T., & Qin, B. (2024). Ensemble learning for heterogeneous large language models with deep parallel collaboration. *Advances in Neural Information Processing Systems*, 37, 119838-119860.
8. Kashefian-Naeeni, S., Shokrpour, N., & Pakdel, F. (2024). Optimizing EFL learning: Exploring the role of learner background factors and the nuances of their effects on intrinsic and extrinsic motivation among university students in a mixed-methods study. *BMC psychology*, 12(1), 535.
9. Liu, J., & Cui, P. (2025, August). Data heterogeneity modeling for trustworthy machine learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2* (pp. 6086-6095).
10. Majlatow, M., Shakil, F. A., Emrich, A., & Mehdiyev, N. (2025). Uncertainty-aware predictive process monitoring in healthcare: explainable insights into probability calibration for conformal prediction. *Applied Sciences*, 15(14), 7925.
11. Mendil, M., Mossina, L., & Vigouroux, D. (2023, August). Puncc: a python library for predictive uncertainty calibration and conformalization. In *Conformal and Probabilistic Prediction with Applications* (pp. 582-601). PMLR.
12. Naim, M. (2026). Bank marketing dataset (UCI) [Data set]. Kaggle
13. Nanglia, S., Ahmad, M., Khan, F. A., & Jhanjhi, N. Z. (2022). An enhanced Predictive heterogeneous ensemble model for breast cancer prediction. *Biomedical Signal Processing and Control*, 72, 103279.
14. Qian, J., Hansen, L. K., Fafoutis, X., Tiwari, P., & Pandey, H. M. (2020). Robustness analytics to data heterogeneity in edge computing. *Computer Communications*, 164, 229-239.
15. Salama, M., Abdelkader, H., & Abdelwahab, A. (2022). A novel ensemble approach for heterogeneous data with active learning. *International Journal of Engineering Business Management*, 14, 18479790221082605.
16. Su, Z., Lu, Z., Wu, Y., Shen, R., & Lu, S. (2025, October). FLSeg: Enhancing Privacy and Robustness in Federated Learning under Heterogeneous Data via Model Segmentation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 3916-3925). IEEE.

17. Tamuly, P., & Nava, V. (2025). Machine learning based conformal predictors for uncertainty-aware compressive strength estimation of concrete. *Construction and Building Materials*, 487, 141844.
18. Wang, S., Fu, Y., Li, X., Lan, Y., & Gao, M. (2023). Dfrd: Data-free robustness distillation for heterogeneous federated learning. *Advances in Neural Information Processing Systems*, 36, 17854-17866.
19. Xie, Y., Li, A., Gao, L., & Liu, Z. (2021). A heterogeneous ensemble learning model based on data distribution for credit card fraud detection. *Wireless Communications and Mobile Computing*, 2021(1), 2531210.
20. Yuniar, R. F. (2021). The impact of students' perception on their academic background on their learning motivation. *Journal of English Language Teaching and Learning (JETLE)*, 2(2), 78-84.
21. Zawad, S., Ali, A., Chen, P. Y., Anwar, A., Zhou, Y., Baracaldo, N., ... & Yan, F. (2021, May). Curse or redemption? how data heterogeneity affects the robustness of federated learning. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, No. 12, pp. 10807-10814).
22. Zefrehi, H. G., & Altınçay, H. (2020). Imbalance learning using heterogeneous ensembles. *Expert Systems with Applications*, 142, 113005.
23. Zhang, T., Zhang, D. G., Yan, H. R., Qiu, J. N., & Gao, J. X. (2021). A new method of data missing estimation with FNN-based tensor heterogeneous ensemble learning for internet of vehicle. *Neurocomputing*, 420, 98-110.
24. Zhao, Y., & Deng, W. (2022). Prediction in traffic accident duration based on heterogeneous ensemble learning. *Applied Artificial Intelligence*, 36(1), 2018643.
25. Zrnic, T., & Candès, E. J. (2024). Active statistical inference. *arXiv preprint arXiv:2403.03208*.