Missing Data Imputation: Do Advanced ML/DL Techniques Outperform Traditional Approaches?

Youran Zhou, Sunil Aryal, Mohamed Reda Bouadjenek

School of Information Technology, Deakin University, Geelong, VIC, Australia

ECML PKDD 2024

Agenda

1. Introduction
2. Challenges
3. Research Focus
4. Experiments Setting
5. Results and Discussion
6. Conclusion

Introduction:
Definition and Impact of Missing Data

What is Missing data?

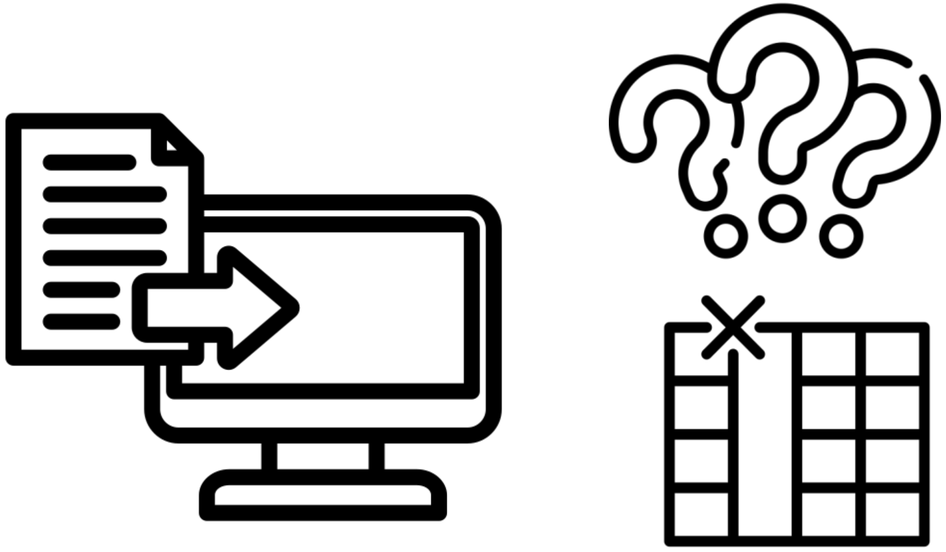
- Absence of values within a dataset
- Occurs in any types of data

How does Missing Data Occur?

- Data Entry, Transformation & storage Error
- Privacy Concern, Non-Response

Impact of Missing Data

- Preserve Data Quality
- Ensure Reliability of Analyses
- Avoid Biases in Results



Introduction:
Factors Influencing the Missing Data

Missing Rate:

- Proportion of data that is missing from a dataset

Missing Mechanism :

- Salary values are randomly missing due to impute error (MCAR)
- Salary values are missing for female employees (MAR)
- Salary values are missing for high-earning employees (MNAR)

Gender	Salary			
	Full	MCAR	MAR	MNAR
F	High	High	High	?
F	High	?	?	?
M	High	?	High	?
F	High	High	?	?
M	High	High	High	?
M	Low	Low	Low	Low
F	Low	?	?	Low
M	Low	Low	Low	Low
M	Low	?	Low	Low
F	Low	Low	?	Low

Introduction:
Solutions to dealing with Missing Data

Imputation Methods

- Statistical-based Methods
- Machine Learning-based (ML) Methods
- Deep Learning-based (DL) Methods

Evaluation Method

- Distance Similarity
  - Root Mean Square Error (RMSE)
  - Mean Absolute Error (MAE)
  - Mean Squared Error (MSE)
- Distributional Similarity
  - Kullback-Leibler (KL) Divergence
  - Wasserstein Distance
- Impact on Downstream Tasks

ECML PKDD 2024

Challenges

- **Overlooked Mechanisms in Existing Methods**
  - Limited focus on MAR and MNAR data imputation in current approaches.
- **Inadequate Evaluation Metrics**
  - Existing metrics like RMSE and MAE fail to capture the real-world utility, particularly in downstream tasks.
- **Experimental Limitations**
  - Inconsistent settings and lack of a comprehensive approach in comparing imputation methods across various missing data scenarios.

Research Focus

- **Comprehensive Evaluation**
  - Systematically evaluate statistical-based, ML-based, and DL-based imputation methods on tabular data, considering different missing mechanisms (MCAR, MAR, MNAR) and varying levels of missing data.
- **Practical Application**
  - Focus on assessing how these methods perform in real-world scenarios, particularly in downstream tasks like regression, classification, and clustering.
- **Refined Metrics Future Directions**
  - Offer insights into future directions for refining the evaluation metrics of the data imputation problem, aiming to improve the practical application of imputed data.

Experiments Setting: Dataset Selection

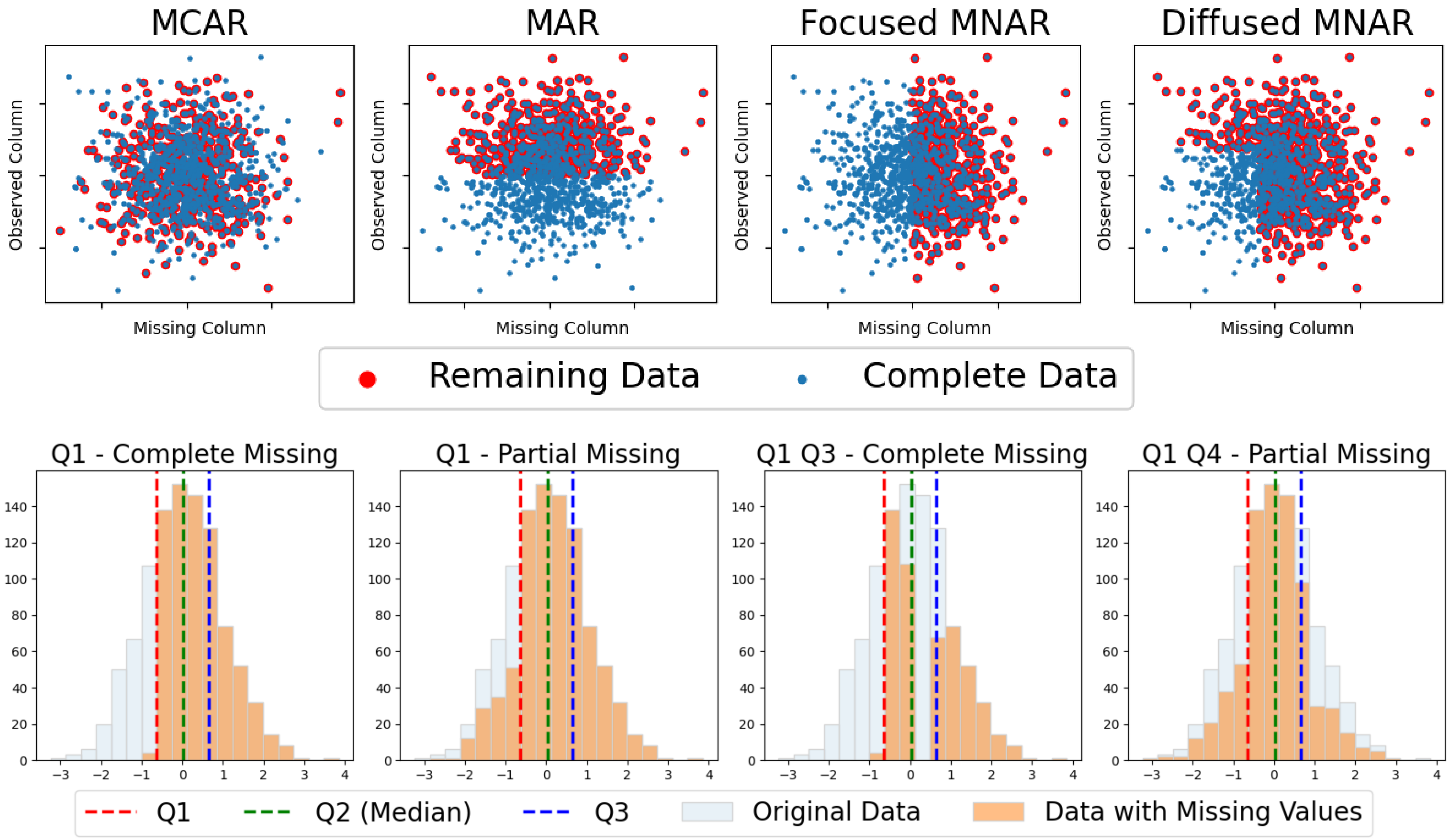
- **Dataset**
  - 10 from the UCI Machine Learning Repository
  - Features are all numerical fields
  - Applied MinMaxScaler to scale features within the range [0, 1]
  - Various tasks including Regression, Classification, and Clustering
  - Clustering methods also applied to datasets typically used for classification

Dataset	Bank	Cali	Climate	Concre	Qsar	Red	Sonar	White	Yachts	Yeast
#Inst	1372	20640	540	1030	1055	1500	208	4898	308	1484
#Dim	5	9	20	8	41	11	60	11	6	8
Task	C	R	C	C	C	R	C	R	R	C

Experiments Setting: Missing Data Generation

- **MCAR**
  - Random
- **MAR**
  - Logistic
- **Focused MNAR**
  - Percentile Rule
  - Logistic
- **Diffused MNAR**
  - Diffused

- **Missing Rate**
  - Ψ: 0.3, 0.5, 0.7



Experiments Setting:
Imputation Models (14 Methods)

Model Name	Type	Subtype
Random Imputer **(RD)**	**Statistical Based**	**Baseline**
Zero Imputer **(ZR)**		
Mean Imputer **(MEAN)**		
K-NN Imputer (2001) **(KNN)**	**Machine Learning Based**	**-**
Matrix Factorization (2001) **(MF)**		**-**
MICE (2011) **(MICE)**		**Regression Based**
XGBImputer (2014) **(XGB)**		**Tree Based**
MissForest (2012) **(MisF)**		**Tree Based**
Optimal Transport (2020) **(OT)**		**Enhanced Machine Learning**
Hyper Imputer (2022) **(HI)**		**Enhanced Machine Learning**
GAIN (2018) **(GAIN)**	**Deep Learning Based**	**GAN Based**
MiWAE (2018) **(Mi)**		**VAE Based**
Not-MiWAE (2020) **(NMi)**		**VAE Based**
Tab-CSDI (2022) **(CSDI)**		**Diffusion Based**

Experiments Setting: Evaluation Process

- **Quantitative Metrics**
  - RMSE/MAE
  - Pearson Correlation (between imputed value and ground truth)
- **Downstream Task**
  - Regression - RMSE
  - Classification – F1
  - Clustering - Adjusted Mutual Information (AMI)

Results and Discussion Baseline Methods

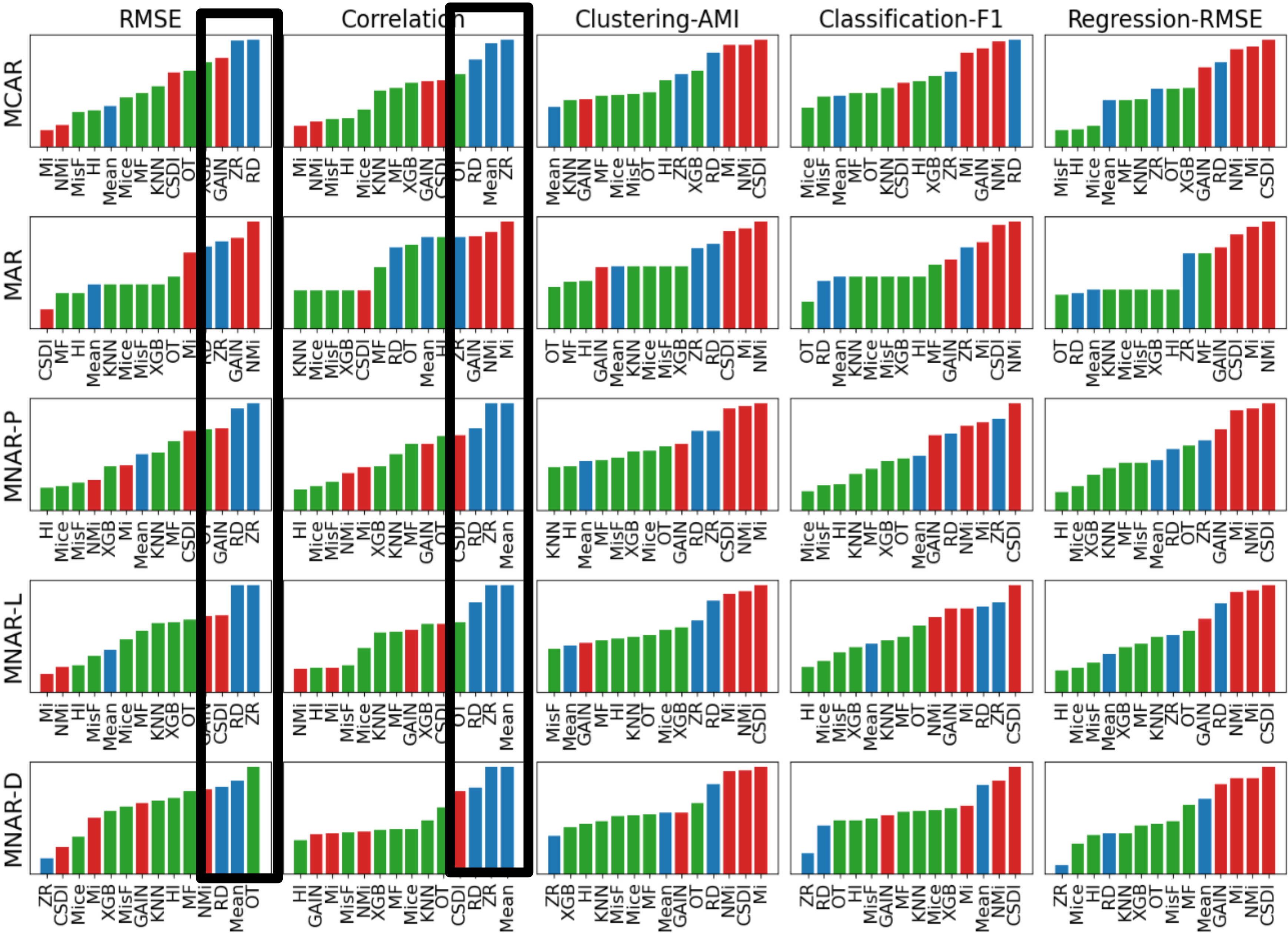
Baseline Methods

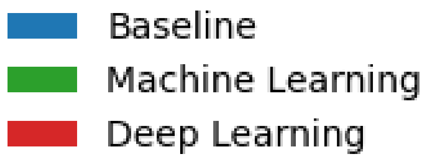
Quantitative Results:

- Not performing well.

MNAR-D Performance:

- Shows promising (Mean Imputer)





Average ranking of different imputation methods

Results and Discussion Baseline Methods

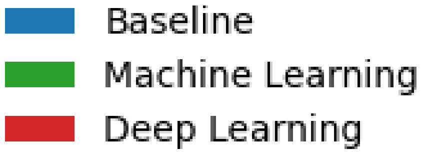
Baseline Methods

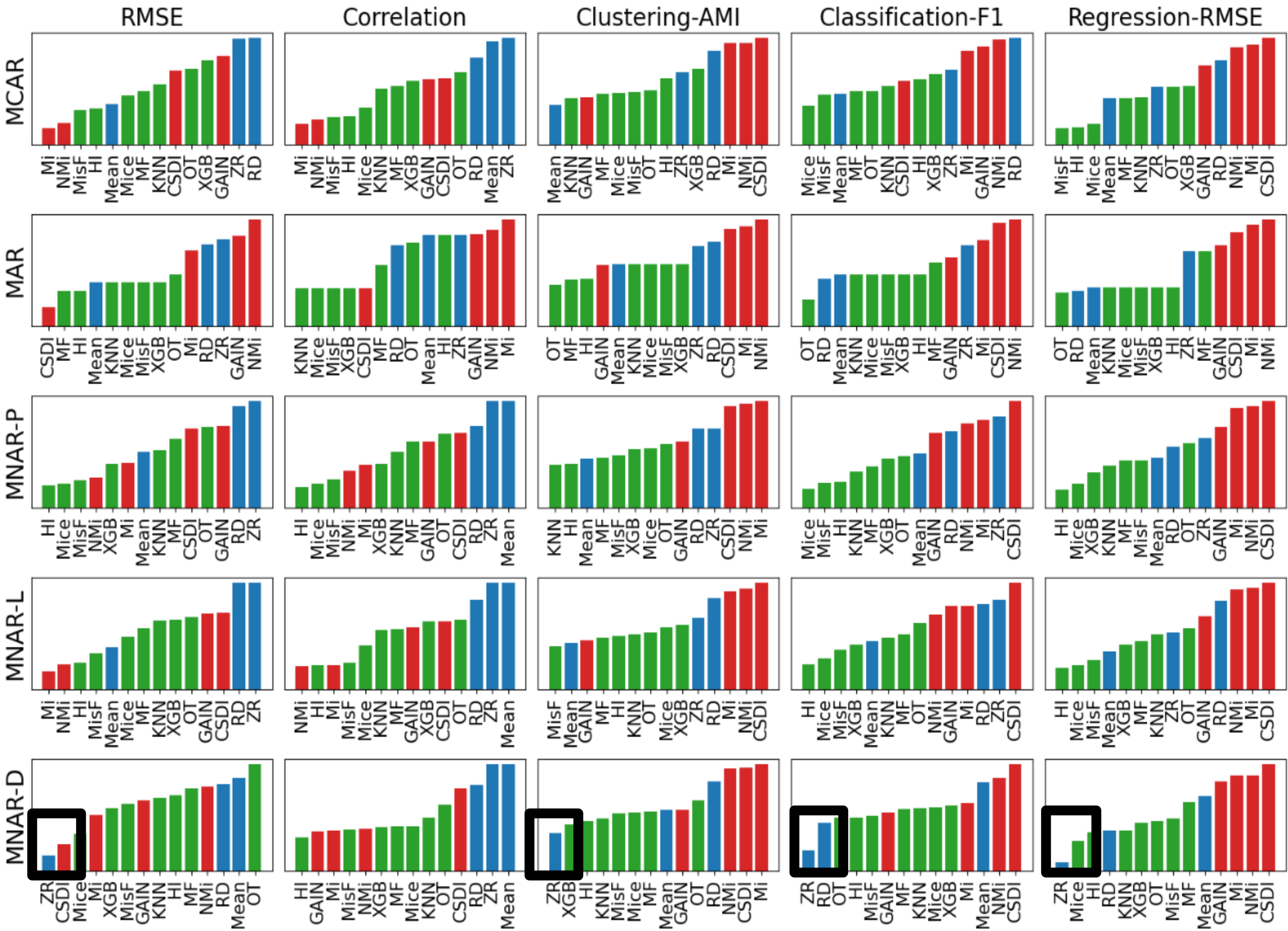
Quantitative Results:

- Not performing well.

MNAR-D Performance:

- Shows promising (Mean Imputer)





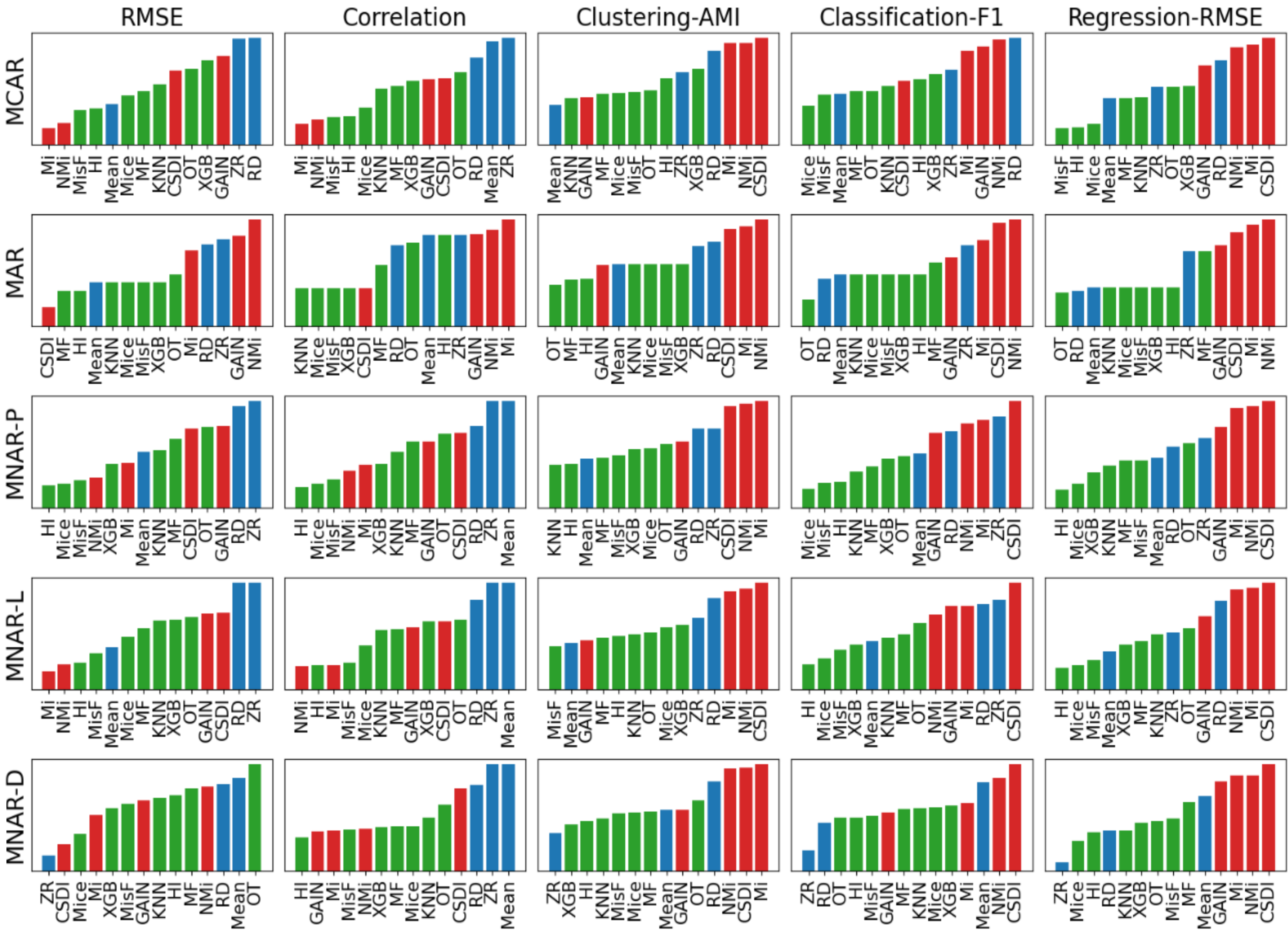
Average ranking of different imputation methods

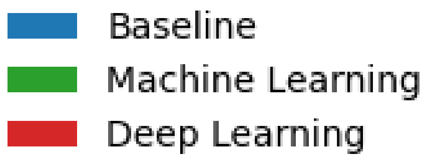
Results and Discussion
Baseline Methods

ML-Based Methods:

Quantitative & Downstream:

- Generally performs well in both, showing balanced effectiveness.





Average ranking of different imputation methods

Results and Discussion
Baseline Methods

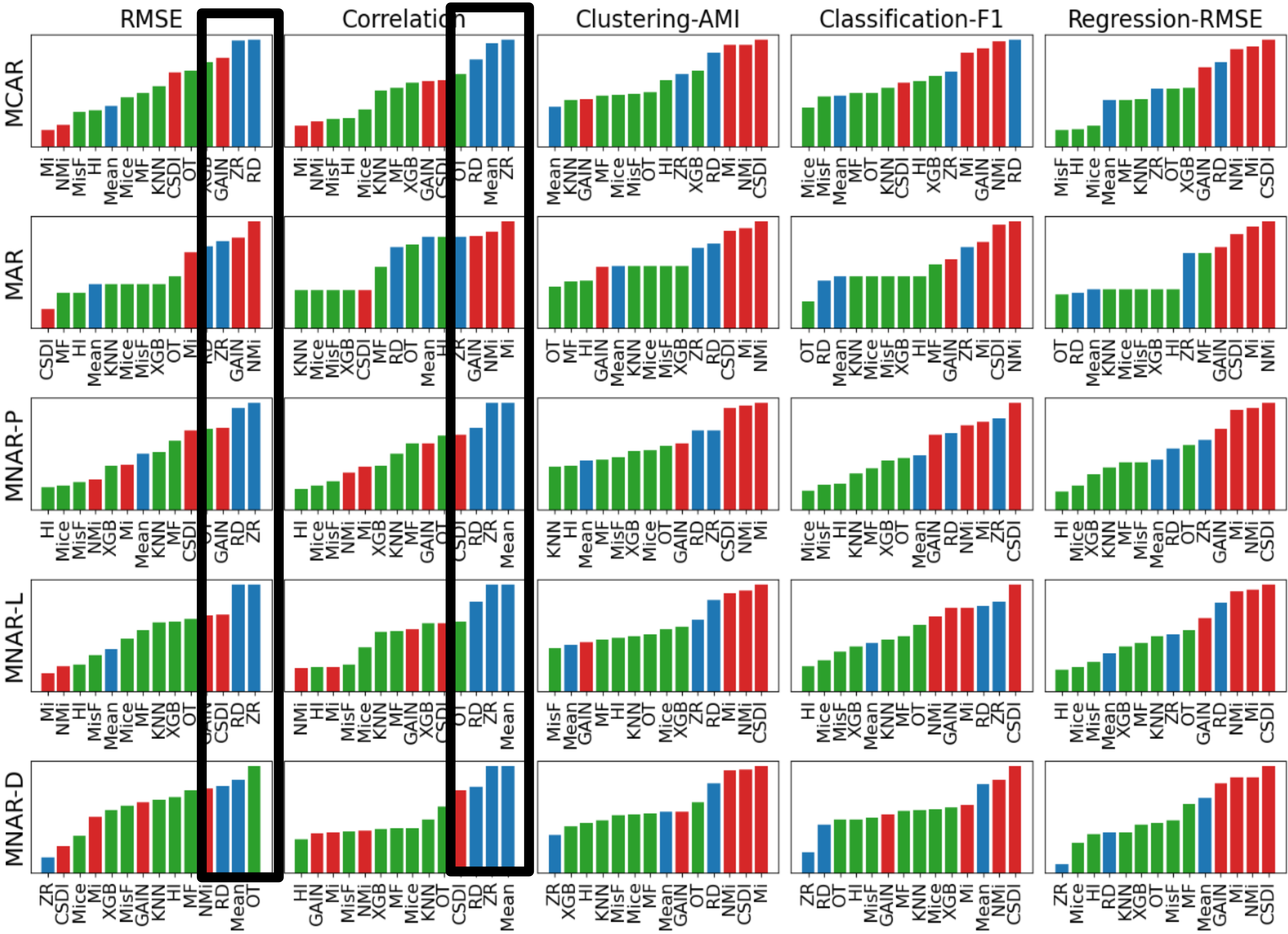
DL-Based Methods:

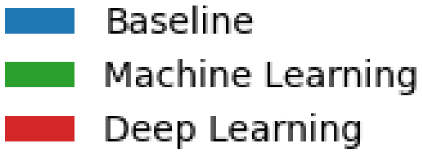
Quantitative Analysis:

- Generally, excels in quantitative analysis, yielding strong RMSE and MAE scores.

Downstream Tasks:

- Fails to perform effectively in downstream tasks.





Average ranking of different imputation methods

Results and Discussion
Baseline Methods

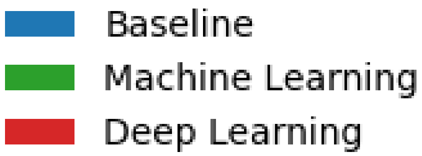
DL-Based Methods:

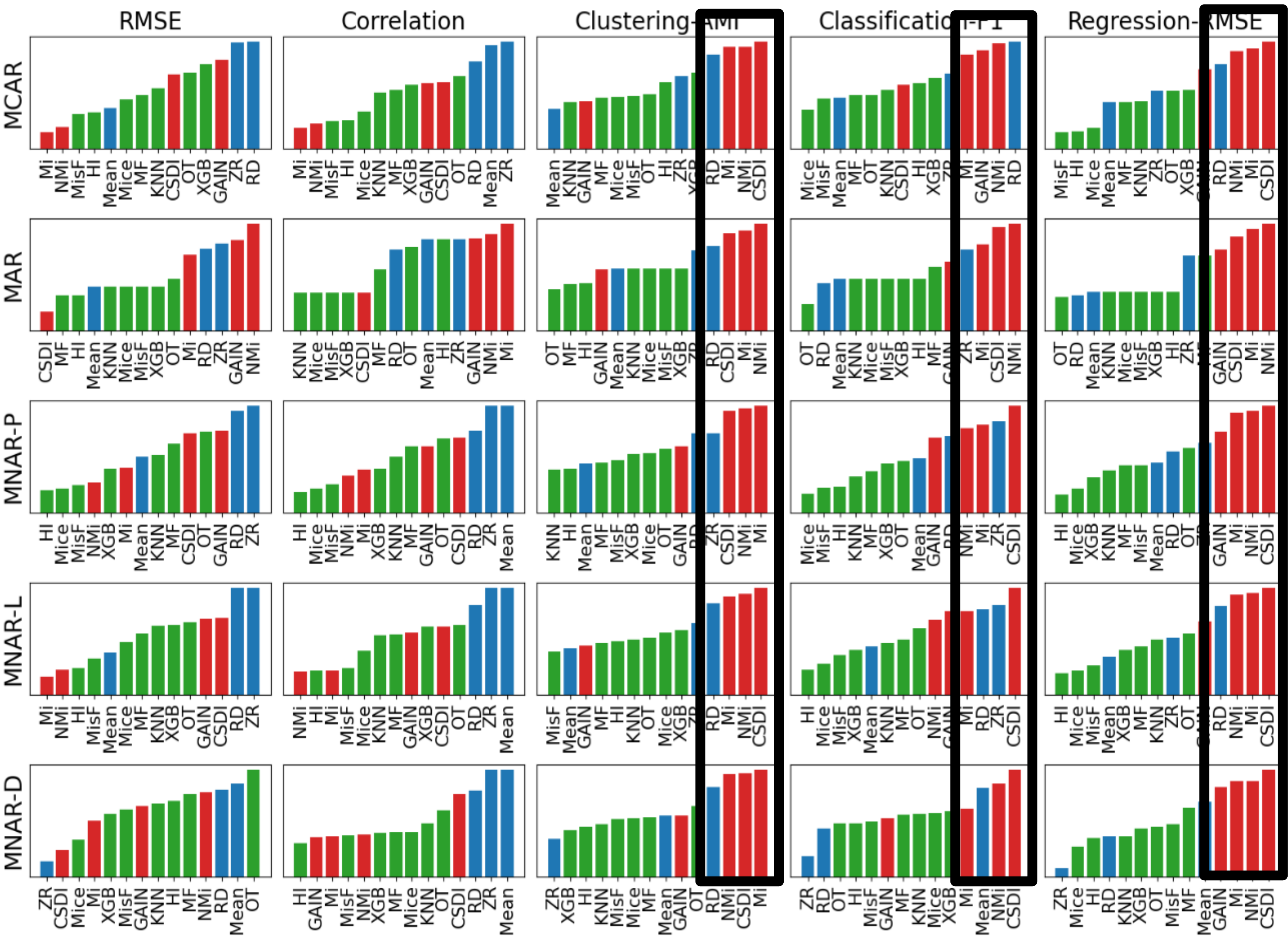
Quantitative Analysis:

- Generally, excels in quantitative analysis, yielding strong RMSE and MAE scores.

Downstream Tasks:

- Fails to perform effectively in downstream tasks.





Average ranking of different imputation methods

Conclusion

Key Findings:

- **Performance Across Missing Mechanisms:**
  - Imputation methods show strong performance under MCAR but face challenges with MAR and MNAR due to their complexity.
- **Imputation Model Insights:**
  - Statistical Methods: Effective, especially in complex missing scenarios.
  - ML-Based Methods: Robust across both quantitative metrics and downstream tasks.
  - DL-Based Methods: While promising in qualitative analysis, often fail in downstream tasks, likely due to the limited size of tabular datasets.

Conclusion

Future Directions:

- **Broader Evaluation Metrics:**
  - Beyond RMSE, explore a wider set of metrics to better assess imputation quality across various analytical tasks.
- **Focus on MAR and MNAR:**
  - Develop techniques tailored to handle MAR and MNAR mechanisms, as they are more prevalent in real-world scenarios.
- **Handling Diverse Data Types:**
  - Extend research to address missing data in discrete and categorical forms, beyond the current focus on numeric data.

Reference

1. Alabadla, M., Sidi, F., Ishak, I., Ibrahim, H., Affendey, L.S., Ani, Z.C., Jabar, M.A., Bukar, U.A., Devaraj, N.K., Muda, A.S., et al. (2022). Systematic review of using machine learning in imputing missing values. IEEE Access, 10, 44483-44502.
2. Gomer, B., & Yuan, K.H. (2021). Subtypes of the missing not at random missing data mechanism. Psychological Methods, 26(5), 559.
3. Harel, O., & Zhou, X.H. (2007). Multiple imputation: Review of theory, implementation and software. Statistics in Medicine, 26(16), 3057-3077.
4. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840-6851.
5. Ipsen, N.B., Mattei, P.A., & Frellsen, J. (2020). not-miwae: Deep generative modelling with missing not at random data. arXiv preprint arXiv:2006.12871.
6. Jäger, S., Allhorn, A., & Biessmann, F. (2021). A benchmark for data imputation methods. Frontiers in Big Data, 4, 693674.
7. Jarrett, D., Cebere, B.C., Liu, T., Curth, A., & van der Schaar, M. (2022). Hyperimpute: Generalized iterative imputation with automatic model selection. In International Conference on Machine Learning (pp. 9916-9937). PMLR.
8. Liao, S.G., Lin, Y., Kang, D.D., Chandra, D., Bon, J., Kaminski, N., Sciurba, F.C., & Tseng, G.C. (2014). Missing value imputation in high-dimensional phenomic data: Imputable or not, and how? BMC Bioinformatics, 15(1), 1-12.
9. Lin, W.C., & Tsai, C.F. (2020). Missing value imputation: A review and analysis of the literature (2006-2017). Artificial Intelligence Review, 53, 1487-1509.
10. Little, R.J., & Rubin, D.B. (2002). Bayes and multiple imputation. Statistical Analysis with Missing Data (pp. 200-220).
11. Liu, M., Li, S., Yuan, H., Ong, M.E.H., Ning, Y., Xie, F., Saffari, S.E., Shang, Y., Volovici, V., Chakraborty, B., & Liu, N. (2023). Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. Artificial Intelligence in Medicine, 142, 102587. doi: https://doi.org/10.1016/j.artmed.2023.102587.
12. Luo, Y. (2021). Evaluating the state of the art in missing data imputation for clinical data. Briefings in Bioinformatics, 23(1), bbab489. doi: 10.1093/bib/bbab489.
13. Ma, C., & Zhang, C. (2021). Identifiable generative models for missing not at random data imputation. Advances in Neural Information Processing Systems, 34, 27645-27658.
14. Madhu, G., Bharadwaj, B.L., Nagachandrika, G., & Vardhan, K.S. (2019). A novel algorithm for missing data imputation on machine learning. In 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 173-177). IEEE.
15. Mattei, P.A., & Frellsen, J. (2019). Miwae: Deep generative modelling and imputation of incomplete data sets. In International Conference on Machine Learning (pp. 4413-4423). PMLR.
16. Miao, X., Wu, Y., Chen, L., Gao, Y., & Yin, J. (2022). An experimental survey of missing data imputation algorithms. IEEE Transactions on Knowledge and Data Engineering.
17. Muzellec, B., Josse, J., Boyer, C., & Cuturi, M. (2020). Missing data imputation using optimal transport. In International Conference on Machine Learning (pp. 7130-7140). PMLR.
18. Pereira, R.C., Santos, M.S., Rodrigues, P.P., & Abreu, P.H. (2020). Reviewing autoencoders for missing data imputation: Technical trends, applications and outcomes. Journal of Artificial Intelligence Research, 69, 1255-1285.
19. Ranjbar, M., Moradi, P., Azami, M., & Jalili, M. (2015). An imputation-based matrix factorization method for improving accuracy of collaborative filtering systems. Engineering Applications of Artificial Intelligence, 46, 58-66.
20. Rubin, D.B. (1976). Inference and missing data. Biometrika, 63(3), 581-592.
21. Song, Q., & Shepperd, M. (2007). Missing data imputation techniques. International Journal of Business Intelligence and Data Mining, 2(3), 261-291.
22. Stekhoven, D.J., & Bühlmann, P. (2012). Missforest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112-118.
23. Tashiro, Y., Song, J., Song, Y., & Ermon, S. (2021). Csdi: Conditional score-based diffusion models for probabilistic time series imputation. Advances in Neural Information Processing Systems, 34, 24804-24816.
24. Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45, 1-67.
25. Yoon, J., Jordon, J., & Schaar, M. (2018). Gain: Missing data imputation using generative adversarial nets. In International Conference on Machine Learning (pp. 5689-5698). PMLR
26. Zheng, S., & Charoenphakdee, N. (2023). Diffusion models for missing value imputation in tabular data.

Thank You!

Question?

Full Paper

Connect Me on LinkedIn