1 Literature Review: Good Practice for Managing and Reporting Missing Data Not Often Followed

The aim of this project was to develop a prototype (Vaar) for evaluating data imputation efforts when applied to hospital or clinical data, enabling researchers to select the most appropriate choice based on the nature or characteristics of their data. Vaar is an Orcadian sailing term meaning to guide or direct (Orkney is well known for its wild seas); and Vaar also conveys the variable nature of data.

Health data provides important insights to improve patient care. However, data is typically not research-ready. The intrinsic data issues of sparsity, scarcity and imbalance must be addressed before training machine learning models (Mandreoli *et al.*, 2022). The most appropriate combination of pre-processing decisions has the potential to reduce bias and improve the quality of insights, and the Vaar prototype could save time. However, there are multiple steps and methods and no consistent approach across application areas.

1.1 Consensus and Challenges

There is more consensus in the missing data and machine learning literature that:

- Reason for missingness is an important first step
- Complete case analysis (CCA) is usually only appropriate if data is missing completely at random (MCAR)
- Sensitivity analyses should be conducted with different imputation methods
- In certain contexts, outliers may represent extremely valuable information that must not be discarded
- Data transformation to a more normal pattern reduces bias from skewed data in downstream models
- Scaling allows models to compare the relative relationship between data points more effectively

Despite this consensus, many of the papers reviewed for this project did not follow of any these steps. There is a lack of consensus in the literature on:

- The proportion of missing data at which imputation no longer boosts performance
- Relationship between reason for missingness and most effective imputation models, with particular challenges on handling non-ignorable missingness (MNAR – missing not at random)
- The effect of data distribution on pre-processing steps

The Vaar project will focus research in this area and consider how a prototype can support researchers through good practice steps.

1.2 State-of-the-art in Missing Data Research

In a 2023 paper, (Lee *et al.*, 2023) focus on 'recoverability' (whether missing data can be consistently estimated from the patterns and associations in the observed data) rather than missingness mechanisms. They use missingness directed acyclic graphs (m-DAGs) to display causal assumptions and indicators of missingness. The authors suggest this simpler alternative due to their observation that many authors state that MAR (missing at random) is the assumed missingness mechanism without justifying this assumption. However, they acknowledge that further investigation is required for practical applications. As such, this project focused on more established conventions to test imputation approaches.

(Chan and Meng, 2021) propose an innovation for multiple imputation inference, a new likelihood ratio test that is simpler to compute than the one currently used in the MICE R package pool function. It compares well in the experiments presented in their updated 2021 paper, however more research and validation is needed to evaluate the ratio test more thoroughly.

(Du *et al.*, 2022) present a Bayesian latent variable selection model (BLVSM) to impute missing data due to MNAR (missing not at random) which current approaches, such as MICE, cannot always do as there is not enough information available to make a prediction. Their analysis shows that BLVSM works well on large MNAR datasets but not small ones. It would be interesting to see further research in this area, as non-ignorable missingness is one of the most challenging missing data problems. (Lüdtke, Robitzsch and West, 2020) developed the R package mdmb, to facilitate a factored regression modelling approach which can estimate some selection models.

1.3 Reason for Missingness is an Important First Step

There is broad consensus in the literature that considering reason for missingness is an important first step before applying imputation methods (Hassler *et al.*, 2019) and (van Buuren, 2018) as this can bias analysis results and influence imputation technique choices (Schober and Vetter, 2020). Data can be:

- MCAR (missing completely at random): any piece of data has the same chance of being missing and is not related to any other characteristics, loss could be accidental
- MAR (missing at random): missing conditional on another variable
- MNAR (missing not at random): missing data for a specific variable is systematically related to the values of this variable itself.

(Schober and Vetter, 2020) argue that whilst complete case analysis (CCA) is the most common approach to managing missing clinical data, it is usually only appropriate if data is MCAR particularly when missingness is greater than 5%. However, they do not explain why or evidence the claims. (van Buuren, 2018) provides a strong rationale for why the missing mechanism is important:

- As the probability that any piece of data could be missing with MCAR is the same, CCA produces unbiased results (the standard error and significance levels are correct for the data subset).
- CCA under MAR or MNAR severely biases mean and regression coefficient estimates. As regression coefficient estimates are used to predict unknown variables (using known variables) this can lead to bias in model results.
- CCA can also produce nonsensical results for time series data.

(Hughes *et al.*, 2019) argue that there are circumstances in which CCA is appropriate using causal diagrams. They provide a good rationale, using two datasets and scenarios, to show that in addition to MCAR, CCA can also be unbiased where missingness is dependent upon independent variables. (Hughes *et al.*, 2019) also argue that multiple imputation (MI) generally gives biased results for MNAR as most implementations are based on a MAR assumption. This view is supported by (Li *et al.*, 2018) who argue that applying imputation methods designed for MAR to MNAR data can lead to bias.

Using an experimental method that assumed meaningful missingness patterns, (Li *et al.*, 2018) demonstrated that regardless of whether missingness in Electronic Health Records (EHRs) is MAR or MNAR, a per-pattern model and CM (causal matching) method can outperform basic imputation and most other proposed methods. (Li *et al.*, 2018) evidence this through experiments, and (Kahale *et al.*, 2020) clearly explain why this is a robust approach. By considering risk, pattern mixture models increase the uncertainty within data to account for the fact that data is imputed. Single imputation, by contrast, can falsely increase precision. Robustness and reliability are particularly important for health data.

(Nijman *et al.*, 2022) identified 152 machine learning clinical prediction model studies published in 2018-19. Of these, they found that whilst a majority 96 (63%) reported missing data only eight of these discussed missing data mechanisms. One of the challenges is that whilst the MCAR assumption can be rejected (Schober and Vetter, 2020) it is not possible to test whether data is MAR or MNAR. MNAR is very complex. Moreover, as (Lee *et al.*, 2023) write, the missing data mechanism is not the only factor in determining the best handling method.

1.4 No Consensus on Proportion of Missing Data at Which Imputation No Longer Boosts Performance

While there is broad consensus that imputation boosts model performance, there is no agreement on how much missing data is too much to improve results. (Madley-Dowd *et al.*, 2019) effectively highlight the varying guidance that exists in the literature – with limited evidence to support the varying recommendations – on what proportion of missing data warrants MI:

- 5% lower threshold below which MI provides negligible benefit. (Schafer JL, 1999 cited by Madley-Dowd, P.et al. 2019)
- 5% maximum threshold for large data sets. (Alice, M. 2015 cited by Madley-Dowd, P.et al. 2019)
- >40% in important variables, results should be considered as hypothesis-generating. (Dong and Peng 2013 and Jakobsen et al 2017 cited by Madley-Dowd, P.et al. 2019)

(Pfob, Lu and Sidey-Gibbons, 2022) recommend removing any variable with more than 50% of data points missing. However, (Wu *et al.*, 2019) found that most methods gain strong robustness and discriminant power even where a dataset had high missing rates (> 50%). A very thorough comparison of MI effectiveness and bias to CCA was undertaken by (Hyuk Lee and Huber Jr., 2021) using MCAR, MAR and MNAR assumptions; with 20%, 40%, 60% and 80% missingness, comparing regression, predictive mean matching, and Markov Chain Monte Carlo as MI mechanisms. (Hyuk

Lee and Huber Jr., 2021) showed that whilst the Root Mean Square Error (RMSE) increased as missingness increased with MI and CCA under all mechanisms, CCA estimates were more seriously biased. However, MI with MNAR data produced biased results even at relatively low levels of missingness.

1.5 Sensitivity Analyses Should Be Conducted With Different Imputation Methods

There is also consensus that sensitivity analyses should be conducted with different imputation methods, as the missingness mechanism cannot be concretely determined, the relationship between variables may have a strong influence on effectiveness also and the reliability of results may be improved. (van Buuren, 2018), (Kahale *et al.*, 2020), (Dong *et al.*, 2021) and (Lee *et al.*, 2023).

The key rationale for undertaking sensitivity analyses is to assess the potential impact that MNAR may have on the estimated results, as multiple imputation assumes MAR for example. (van Buuren, 2018) explains sensitivity analysis very clearly and advocates simple adjustments to imputed data under the δ-adjustment with a CCA comparison also, ensuring the changes are reasonable to the assumption being tested. In a thorough sensitivity analysis case study, to test the assumption that the degree of departure from MAR varied according to a self-reported HIV status variable, δ-adjustment presented a flexible and transparent solution. (Leacy *et al.*, 2017)

1.6 Complete Case Analysis is Potentially Wasteful

There is broad consensus that data subsets could "seriously degrade the ability to detect the effects of interest" (van Buuren, 2018) and "introduce bias, as missingness itself can be associated with outcomes" (Li *et al.*, 2018). MI can use information from auxiliary variables that explain missingness reasons or provide information about missing values. (Austin *et al.*, 2021) make the point that even if data are MCAR, reducing sample size correspondingly reduces the precision with which statistics and regression coefficients are estimated. Estimated confidence intervals will be wider when using CCA than if all the data were used.

(von Hippel, 2007) showed that for linear regression models, if the dependent variable is missing values then MI followed by exclusion of the missing values produces better estimates. The advantage is that all cases are used for imputation providing information for the regression of interest, then deletion removes noise to improve analysis accuracy. (von Hippel, 2007) was challenged, and concedes, that deleting imputed dependents increases the standard error within the imputed datasets but also reduces the variation between them thus increasing accuracy. MI can outperform this approach where the dependent variable benefits from auxiliary information.

1.7 Outliers May Represent Extremely Valuable Information That Must Not Be Discarded

(Kantardzic, 2011) stresses that automatic elimination of outliers is risky, and deletion may be counter-productive, if data are correct it could result in the loss of important hidden information. If none of the outlier removal techniques improve the performance of a classification model it suggests extreme values are data variation rather than errors (Salgado *et al.*, 2016). Outliers are therefore likely to contain useful information in their extreme values and automatically excluding them results in a loss of information.

1.8 Data Transformation Reduces Bias from Skewed Data

Most parametric tests to determine whether observed differences are statistically significant – like ANOVA – assume a normal distribution of data and therefore require the mean and standard deviation to be reliable statistics. Non-parametric tests can be used for skewed data but are less powerful and dependable, especially for small datasets. Data can be transformed to reduce bias from skewed data in the downstream model, and scaling allows models to compare the relative relationship between data points more effectively. (Felix and Lee, 2019) Normalisation changes the distribution shape of data and scaling changes the data range, although both terms are used interchangeably within the literature.

1.9 Scaling Allows Models to Compare the Relative Data Relationships More Effectively

(Izonin *et al.*, 2022) demonstrated that scaling methods significantly affect the performance of classifiers. They investigated the effectiveness of five methods on short, unbalanced medical datasets against three different machine learning models for a binary classification task.

(Singh and Singh, 2021) conducted experiments on 20 publicly available medical datasets with normal distribution, testing the effects of min-max, z-score, median and median absolute deviation on classification models. They saw a positive difference in 14 out of 20 datasets concluding that scaling was superior in a majority of 70% datasets. Their results indicate that data distribution may also be important in selecting the most effective approach.

1.10 Literature Comparing Imputation Approaches

The table below highlights several papers that compare different imputation approaches which influenced the project's experimental design.

Citation	Missingness and Datasets	Key Discussion Points in Paper	Evaluation	Project Considerations
(Dong *et al.*, 2021)	• Real world data • 141,516 patients with diabetes, 14/21 independent baseline variables had missing data, 12 <20%. Ranged from 0.50% (systolic blood pressure) to 48.99% (Urine ACR). • Hypertension data. 10,000 subjects without any missing values for 10 independent variables randomly selected. • (MAR) simulated at different missingness rates (20 and 50%) for both datasets. • Missing values not simulated in dependent variables.	• MICE has limited ability to handle non-linear relationships • missForest overcomes non-linearity but computation high • Imputation accuracy measured by normalised root mean square error (NRMSE) for continuous variables and proportion falsely classified for categorical variables • Distributions tested by Shapiro-Wilk normality test • Imputation accuracy differences tested by one-way ANOVA or non-parametric test • GAIN and missForest outperformed MICE • GAIN outperformed missForest at 50% missingness	• Authors stress limited to imputation accuracy only which is not the aim of imputation. • Good consideration of data skew in choosing statistical analysis methods. • Missing values not simulated in dependent variables.	• The aim of the Vaar project was to evaluate the impact of imputation on downstream data models. • The first step, to gain confidence in the approach, was to evaluate the accuracy of imputation as this paper has done.
(Chowdhury, Islam, and Khan, 2017)	65,000 patient records with missing values in the gender attribute of patients.	Amelia, FURIA, and MICE tested against synthetically generated missingness (20%). MICE produced the best results.	Missingness percentage calculated based on data given in paper, but neither this nor missingness mechanism explicitly stated.	MICE for Vaar experiments.

Citation	Missingness and Datasets	Key Discussion Points in Paper	Evaluation	Project Considerations
(Li *et al.*, 2018)	Synthetic MNAR data, and electronic health records (EHR) data from a tertiary care provider.	Pattern-Wise Analysis performs modelling without imputation There may be a combination of mechanisms in a dataset, consider on a variable by variable basis They recommend Causal Matching (PPM+CM) for MNAR missingness in EHR data analysis Impute MAR/MCAR missing data as pre-processing	Thorough analysis, detailed experiments, clear consideration, and explanation of missing mechanisms.	Evidence to support PPM+CM for MNAR missingness.
(Orczyk and Porwick, 2013)	HEPA, BREA, and HEART datasets. 5-25% missingness synthetically created. MAR stated.	All tested imputation methods were based on arithmetic means. Naïve Bayes, Random Trees and Random Forest models then run to classify datasets. Imputation method can affect classification accuracy by 10%. Random Forest produced the best result.	More information on missingness patterns and justification for mechanism would have been useful. Different experiments clearly explained, and results visualised.	Random Forest as effective classification model to choose for some experiments.
(Perez-Lebel *et al.*, 2022)	Traumabase, UK Biobank, MIMIC III, NHIS	13 different prediction tasks across the four datasets. When using imputation, concatenating the missingness indicator with the input features significantly improves predictions. Adding an indicator to express which values have been imputed is important.	Very thorough paper, with good consideration of missing mechanisms and testing imputation on the effectiveness of downstream models.	Focus on prediction quality which is more in line with Vaar project aims.

1.11 Benchmarking Papers for Experiments

The research questions, design and methodology compared different approaches, based on consensus, and explored challenges in more detail. UCI healthcare datasets were chosen as a basis for the experiments for reproducibility. The table below summarises a comparative paper considered for various UCIs dataset and considers how it has dealt with missing data. Not all datasets have missing data, as the first stage in the project is to create missing data levels with different missingness methods for the data experiments.

Paper Name and Citation	Dataset Name	Missing Data	Reason for missingness	Sensitivity Analyses	Consideration
A Fuzzy-Rough based Binary Shuffled Frog Leaping Algorithm for Feature Selection (Anaraki *et al.*, 2018)	Arrhythmia	Not mentioned	Not mentioned	Not mentioned	Feature selection was the focus of this paper, and several datasets were compared using different classification models. There is missing data in this dataset (Guvenir *et al.*, 1998). There was no mention in the paper of any pre-processing steps undertaken or assumptions made. It is difficult, therefore, to have confidence that the results are unbiased or robust.
Application of CART Algorithm in Blood Donors Classification (Santhanam and Sundaram, 2010)	Blood Transfusion Service Center	No	N/A	N/A	The paper gives a good description of the dataset and assumptions made. There is no missing data, and this dataset was used in the missing data experiments.
RBF KERNEL OPTIMIZATION METHOD WITH PARTICLE SWARM OPTIMIZATION ON SVM USING THE ANALYSIS OF INPUT DATA'S MOVEMENT (Indraswari and Arifin, 2017)	Breast Cancer Wisconsin (Original)	Not mentioned	Not mentioned	Not mentioned	The focus of this paper was parameter optimisation rather than data optimisation. However, there is missing data in this dataset (Wolberg, 1992) and missing data can bias results.

Paper Name and Citation	Dataset Name	Missing Data	Reason for missingness	Sensitivity Analyses	Consideration
The Wisconsin breast cancer problem: Diagnosis and TTR/DFS time prognosis using probabilistic and generalised regression information classifiers (Anagnostopoulos *et al.*, 2006)	Breast Cancer Wisconsin (Prognostic)	Deleted	Not mentioned	Not mentioned	There is missing data in this dataset. (Wolberg, Street and Mangasarian, 1995) The four instances were excluded from the training and test data.
Curvature-based Feature Selection with Application in Classifying Electronic Health Records(Zuo *et al.*, 2021)	Cervical cancer (Risk Factors)	Deleted	Not mentioned	Not mentioned	There is missing data in this dataset. (Fernandes, Cardoso and Fernandes, 2017) and the authors acknowledge in the discussion that results may have been improved through imputation.
Prediction of Chronic Kidney Disease - A Machine Learning Perspective(Chittora *et al.*, 2021)	Chronic Kidney Disease	Mentioned	Not mentioned	Not mentioned	Data is missing (Rubini, Soundarapandian and Eswaran, 2015). Missing data is mentioned but no details are provided on how it was managed. The focus of the paper was to compare feature selection techniques and their impact on different machine learning models. Feature selection took place after pre-processing. Assumptions made at the pre-processing stage could influence model selection and performance so it would be good practice to provide these details.
SECRET: Semantically Enhanced Classification of Real-World Tasks (Akmandor *et al.*, 2021)	Contraceptive Method Choice	No	N/A	N/A	No data is missing. (Lim, 1997). The authors acknowledge the importance of data pre-processing in the paper which is handled by the SECRET algorithm.

Paper Name and Citation	Dataset Name	Missing Data	Reason for missingness	Sensitivity Analyses	Consideration
ESTIMATION OF CONTINUOUS BLOOD PRESSURE FROM PPG VIA A FEDERATED LEARNING APPROACH.(Brophy *et al.*, 2021)	Cuff-Less Blood Pressure Estimation	Not mentioned	Not mentioned	Not mentioned	Missing data present. (Kachuee *et al.*, 2015). However, the authors used the first 5 (part1.mat − part5.mat) records only and segmented them into 8-second intervals of 144,000 training records which was then tested against a different dataset. A good level of information was provided on pre-processing.
Differential Diagnosis of Erythmato-Squamous Diseases Using Classification and Regression Tree (Maghooli *et al.*, 2016)	Dermatology	Mean imputation	Not mentioned	Not mentioned	Missing data is present. (Ilter and Guvenir, 1998) The authors used histograms to assess data distribution prior to imputation. However, no discussion of reason for missingness or sensitivity analysis.
Diabetic Retinopathy Detection via Deep Convolutional Networks for Discriminative Localization and Visual Explanation (Wang and Yang, 2017)	Diabetic Retinopathy Debrecen Data Set	No	N/A	N/A	There is no data missing (Antal and Hajdu, 2014). The authors provide a good amount of detail on the image pre-processing, augmentation, and architecture design for their neural network model.
Training cost-sensitive neural networks with methods addressing the class imbalance problem (Zhi-Hua Zhou and Xu-Ying Liu, 2006)	Echocardiogra m	Continuous attributes set to the average; binary/ nominal set to the majority value	Not mentioned	Not mentioned	There is missing data (Salzberg, 1988). Mean imputation used on continuous attributes, which cannot be used on categorical features, so the authors have used majority value (it is assumed this means the mode – most frequent value). No other pre-processing methods or assumptions are discussed.

Paper Name and Citation	Dataset Name	Missing Data	Reason for missingness	Sensitivity Analyses	Consideration
Better Multi-class Probability Estimates for Small Data Sets (Alasalmi *et al.*, 2020)	Ecoli	No	N/A	N/A	There is no missing data. (Nakai, 1996) There was no detail in paper about pre-processing and whether any had been undertaken.
Classification of epileptic seizure dataset using different machine learning algorithms (Almustafa, 2020)	Epileptic Seizure Recognition	No	N/A	N/A	No data missing. (Wu and Fokoue, 2017) Sensitivity analyses were performed on parameter changes but as no data was missing, missingness assumptions did not need to be tested. Information was provided on the dataset used, such as distribution, and their analysis took a binary shape for epileptic seizure and not (1 class vs 4 classes).
Toward Efficient Breast Cancer Diagnosis and Survival Prediction Using L-Perceptron (Mansourifar and Shi, 2018)	Haberman's Survival	No	N/A	N/A	No missing data. (Haberman, 1999) Paper explicitly states that no pre-processing or feature selection took place.
A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients (Santos *et al.*, 2015)	HCC Survival	Nearest neighbour imputation	Not mentioned	Not mentioned	Data is missing (Santos *et al.*, 2017). A thorough discussion of missingness rates and patterns. Mean and mode imputation were also tested but rejected. CCA was rejected due to the high missingness rate.
Enhancing Simple Models by Exploiting What They Already Know (Dhurandhar, Shanmugam and Luss, 2019)	Heart Disease	Not mentioned	Not mentioned	Not mentioned	There is missing data in dataset. (Janosi *et al.*, 1988) No data pre-processing, or statistical description of data, included in the paper.

Paper Name and Citation	Dataset Name	Missing Data	Reason for missingness	Sensitivity Analyses	Consideration
Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone (Chicco and Jurman, 2020)	Heart failure clinical records	No	N/A	N/A	No missing data (Ahmad *et al.*, 2020). There is a good description of the dataset in the paper, including suggested improvements or descriptors which would have improved its quality. There is no discussion of data pre-processing.
Analyzing performance of classifiers for medical datasets (Rosly *et al.*, 2018)	Hepatitis	Deleted	Not mentioned	Not mentioned	There is missing data (Gong, 1988). Paper references Naïve Bayes and decision tree classification as models that can handle missing data. All instances with missing data were removed.
Chickenpox Cases in Hungary: a Benchmark Dataset for Spatiotemporal Signal Processing with Graph Neural Networks (Rozemberczki *et al.*, 2021)	Hungarian Chickenpox Cases	No	N/A	N/A	None missing (Rozemberczki, 2021). The authors describe dataset and its characteristics in detail. Pre-processing is not mentioned.
Evaluation of the Performance of the Markov Blanket Bayesian Classifier Algorithm (Madden, 2002)	Lymphography	No	N/A	N/A	No missing data (Zwitter and Soklic, 1988). The authors confirm there are no missing variables in this dataset and plan to extend the algorithm to support missing values also. No data pre-processing steps are mentioned.
Reliable Probabilistic Prediction for Medical Decision Support (Papadopoulos, 2011)	Mammographic Mass	Deleted	Not mentioned	Not mentioned	Data is missing (Elter, 2007). Assume data are independently and identically distributed. Data pre-processing described, including removal of missing data and feature selection.
Performance, Transparency and Time. Feature selection to	Parkinson Speech	No	N/A	N/A	No missing data (Kursun *et al.*, 2014) There is a comprehensive description of the

Paper Name and Citation	Dataset Name	Missing Data	Reason for missingness	Sensitivity Analyses	Consideration
speed up the diagnosis of Parkinson's disease **(Costanzo and Orphanou, 2022)**	Dataset with Multiple Types of Sound Recordings				dataset. A power transformation and standardisation was applied in pre-processing.
A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform **(Sakar *et al.*, 2019)**	Parkinson's Disease Classification	No	N/A	N/A	No data missing (Sakar *et al.*, 2018) Broad description of dataset, minimum redundancy-maximum relevance used for feature selection discussed as pre-processing step.
Risk Factor Prediction of Chronic Kidney Disease based on Machine Learning Algorithms **(Islam *et al.*, 2020)**	Risk Factor prediction of Chronic Kidney Disease	Mean imputation	Not mentioned	Not mentioned	There is data missing (Islam *et al.*, 2020). Good overview of pre-processing techniques, such as Z-score normalisation, and data description and collation. However, missingness was not discussed in any detail.
SCADI: A standard dataset for self-care problems classification of children with physical and motor disability (Zarchi, Fatemi Bushehri and Dehghanizadeh, 2018)	SCADI	None	N/A	N/A	No data is missing. (Bushehri *et al.*, 2018) There is a detailed description of the dataset and a note that data is normalised in data pre-processing.

Table 1: An Evaluation of How the Benchmark Papers for Experiments Have Managed Missing Data

Just over half of the datasets (13 out of 25) considered have missing data, and the approaches used to manage this in the UCI comparative papers is shown in the figure below.

- Four papers make no mention of the missing data at all
- One paper mentions that data is missing but does not explain how it was managed
- Four papers use complete case analysis (delete missing data)
- Two papers used mean imputation
- Nearest neighbour was used by one paper and a combination of mean and most frequent by another

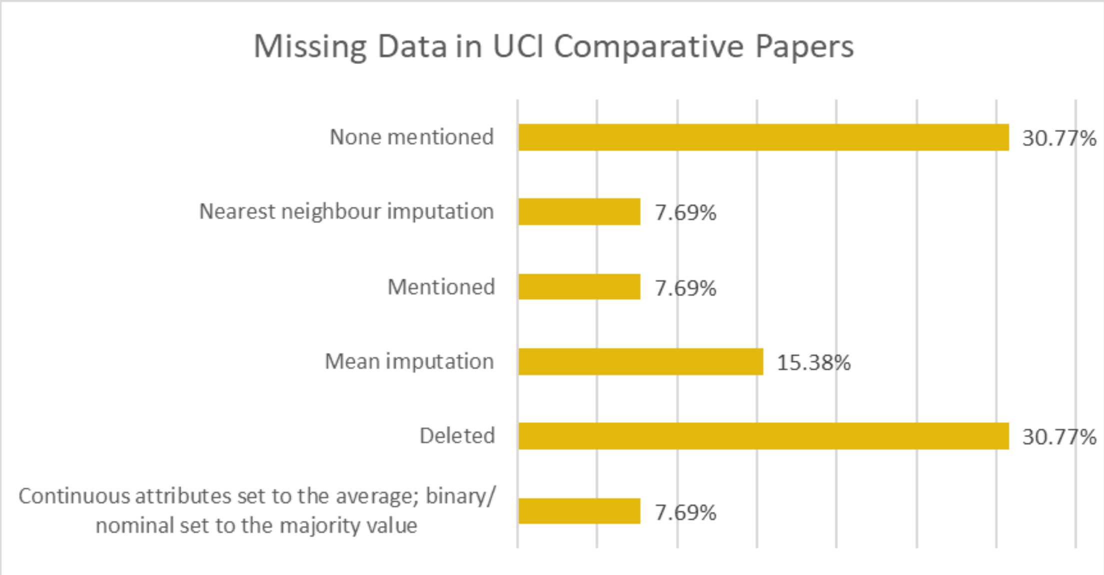

Figure 1: How Comparative UCI Dataset Papers Have Managed Missing Data

None of the papers discussed reasons for missingness and none undertook sensitivity analyses of their imputation methods. This small sample of papers shows that good practice for managing and reporting missing data used in models is not often followed, as shown by (Nijman *et al.*, 2022). The Vaar project focused on those areas where there is a lack of consensus or the biggest challenges – missingness levels, non-ignorable missingness, effect on downstream models – and considered how a prototype can support researchers through good practice steps with robust analysis.

1.12 Literature Review References

Ahmad, T. *et al.* (2020) *Heart failure clinical records*, *UCI Machine Learning Repository*. Available at: https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records (Accessed: 28 April 2023).

Akmandor, A.O. *et al.* (2021) 'SECRET: Semantically Enhanced Classification of Real-World Tasks', *IEEE Transactions on Computers*, 70(3), pp. 440–456. Available at: https://doi.org/10.1109/TC.2020.2989642 .

Alasalmi, T. *et al.* (2020) ‘Better Multi-class Probability Estimates for Small Data Sets’.

Almustafa, K.M. (2020) ‘Classification of epileptic seizure dataset using different machine learning algorithms’, *Informatics in Medicine Unlocked*, 21, p. 100444. Available at: https://doi.org/10.1016/j.imu.2020.100444 .

Anagnostopoulos, I. *et al.* (2006) ‘The Wisconsin breast cancer problem: Diagnosis and TTR/DFS time prognosis using probabilistic and generalised regression information classifiers’, *Oncology Reports* [Preprint]. Available at: https://doi.org/10.3892/or.15.4.975 .

Anaraki, J.R. *et al.* (2018) ‘A Fuzzy-Rough based Binary Shuffled Frog Leaping Algorithm for Feature Selection’. Available at: https://doi.org/10.5281/zenodo.1474575 .

Antal, B. and Hajdu, A. (2014) *Diabetic Retinopathy Debrecen Data Set*, *UCI Machine Learning Repository*. Available at: https://doi.org/10.24432/C5XP4P .

Austin, P.C. *et al.* (2021) ‘Missing Data in Clinical Research: A Tutorial on Multiple Imputation’, *Canadian Journal of Cardiology*, 37(9), pp. 1322–1331. Available at: https://doi.org/10.1016/j.cjca.2020.11.010 .

Brophy, E. *et al.* (2021) ‘Estimation of Continuous Blood Pressure from PPG via a Federated Learning Approach’.

Bushehri, S.M.M. *et al.* (2018) *SCADI*, *UCI Machine Learning Repository*. Available at: https://doi.org/10.24432/C5C89G .

van Buuren, S. (2018) *Flexible Imputation of Missing Data*. Second Edition. Boca Raton, FL.: CRC Press.

Chan, K.W. and Meng, X.-L. (2021) ‘Multiple Improvements of Multiple Imputation Likelihood Ratio Tests’. Available at: https://doi.org/doi.org/10.48550/arXiv.1711.08822 .

Chicco, D. and Jurman, G. (2020) ‘Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone’, *BMC Medical Informatics and Decision Making*, 20(1), p. 16. Available at: https://doi.org/10.1186/s12911-020-1023-5 .

Chittora, P. *et al.* (2021) ‘Prediction of Chronic Kidney Disease - A Machine Learning Perspective’, *IEEE Access*, 9, pp. 17312–17334. Available at: https://doi.org/10.1109/ACCESS.2021.3053763 .

Chowdhury, M.H., Islam, M.K. and Khan, S.I. (2017) ‘Imputation of missing healthcare data’, in *2017 20th International Conference of Computer and Information Technology (ICCIT)*. IEEE, pp. 1–6. Available at: https://doi.org/10.1109/ICCITECHN.2017.8281805 .

Costanzo, P. and Orphanou, K. (2022) ‘Performance, Transparency and Time. Feature selection to speed up the diagnosis of Parkinson’s disease’.

Dhurandhar, A., Shanmugam, K. and Luss, R. (2019) ‘Enhancing Simple Models by Exploiting What They Already Know’.

Dong, W. *et al.* (2021) ‘Generative adversarial networks for imputing missing data for big data clinical research’, *BMC Medical Research Methodology*, 21(1), p. 78. Available at: https://doi.org/10.1186/s12874-021-01272-3 .

Du, H. *et al.* (2022) ‘A Bayesian Latent Variable Selection Model for Nonignorable Missingness’, *Multivariate Behavioral Research*, 57(2–3), pp. 478–512. Available at: https://doi.org/10.1080/00273171.2021.1874259 .

Elter, M. (2007) *Mammographic Mass*, *UCI Machine Learning Repository*. Available at: https://doi.org/https://doi.org/10.24432/C53K6Z .

Felix, E.A. and Lee, S.P. (2019) 'Systematic literature review of preprocessing techniques for imbalanced data', *IET Software*, 13(6), pp. 479–496. Available at: https://doi.org/10.1049/iet-sen.2018.5193 .

Fernandes, K., Cardoso, J. and Fernandes, J. (2017) *Cervical cancer (Risk Factors)*, *UCI Machine Learning Repository*.

Gong, G. (1988) *Hepatitis*, *UCI Machine Learning Repository*. Available at: https://doi.org/10.24432/C5Q59J .

Guvenir, H. *et al.* (1998) *Arrhythmia*, *UCI Machine Learning Repository*. Available at: https://doi.org/10.24432/C5BS32 .

Haberman, S. (1999) *Haberman's Survival*, *UCI Machine Learning Repository*. Available at: https://doi.org/10.24432/C5XK51 .

Hassler, A.P. *et al.* (2019) 'Importance of medical data preprocessing in predictive modeling and risk factor discovery for the frailty syndrome', *BMC Medical Informatics and Decision Making*, 19(1), p. 33. Available at: https://doi.org/10.1186/s12911-019-0747-6 .

von Hippel, P.T. (2007) 'Regression with Missing Ys: An Improved Strategy for Analyzing Multiply Imputed Data', *Sociological Methodology*, 37(1), pp. 83–117. Available at: https://doi.org/10.1111/j.1467-9531.2007.00180.x .

Hughes, R.A. *et al.* (2019) 'Accounting for missing data in statistical analyses: multiple imputation is not always the answer', *International Journal of Epidemiology*, 48(4), pp. 1294–1304. Available at: https://doi.org/10.1093/ije/dyz032 .

Hyuk Lee, J. and Huber Jr., J.C. (2021) 'Evaluation of Multiple Imputation with Large Proportions of Missing Data: How Much Is Too Much?', *Iranian Journal of Public Health* [Preprint]. Available at: https://doi.org/10.18502/ijph.v50i7.6626 .

Ilter, N. and Guvenir, H. (1998) *Dermatology*, *UCI Machine Learning Repository*. Available at: https://doi.org/10.24432/C5FK5P .

Indraswari, R. and Arifin, A.Z. (2017) 'RBF KERNEL OPTIMIZATION METHOD WITH PARTICLE SWARM OPTIMIZATION ON SVM USING THE ANALYSIS OF INPUT DATA'S MOVEMENT', *Jurnal Ilmu Komputer dan Informasi*, 10(1), p. 36. Available at: https://doi.org/10.21609/jiki.v10i1.410 .

Islam, Md.A. *et al.* (2020) 'Risk Factor Prediction of Chronic Kidney Disease based on Machine Learning Algorithms', in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*. IEEE, pp. 952–957. Available at: https://doi.org/10.1109/ICISS49785.2020.9315878 .

Izonin, I. *et al.* (2022) 'Towards Data Normalization Task for the Efficient Mining of Medical Data', in *2022 12th International Conference on Advanced Computer Information Technologies (ACIT)*. IEEE, pp. 480–484. Available at: https://doi.org/10.1109/ACIT54803.2022.9913112 .

Janosi, A. *et al.* (1988) *Heart Disease*, *UCI Machine Learning Repository*. Available at: https://doi.org/10.24432/C52P4X .

Kachuee, M. *et al.* (2015) *Cuff-Less Blood Pressure Estimation*, *UCI Machine Learning Repository*. Available at: https://doi.org/10.24432/C5B602 .

Kahale, L.A. *et al.* (2020) 'Potential impact of missing outcome data on treatment effects in systematic reviews: imputation study', *BMJ*, p. m2898. Available at: https://doi.org/10.1136/bmj.m2898 .

Kantardzic, M. (2011) *Data Mining*. Hoboken, NJ, USA: John Wiley & Sons, Inc. Available at: https://doi.org/10.1002/9781118029145 .

Kursun, O. *et al.* (2014) *Parkinson Speech Dataset with Multiple Types of Sound Recordings*, *UCI Machine Learning Repository*. Available at: https://doi.org/10.24432/C5NC8M .

Leacy, F.P. *et al.* (2017) ‘Analyses of Sensitivity to the Missing-at-Random Assumption Using Multiple Imputation With Delta Adjustment: Application to a Tuberculosis/HIV Prevalence Survey With Incomplete HIV-Status Data’, *American Journal of Epidemiology* [Preprint]. Available at: https://doi.org/10.1093/aje/kww107 .

Lee, K.J. *et al.* (2023) ‘Assumptions and analysis planning in studies with missing data in multiple variables: moving beyond the MCAR/MAR/MNAR classification’, *International Journal of Epidemiology* [Preprint]. Available at: https://doi.org/10.1093/ije/dyad008 .

Li, J. *et al.* (2018) ‘Don’t Do Imputation: Dealing with Informative Missing Values in EHR Data Analysis’, in *2018 IEEE International Conference on Big Knowledge (ICBK)*. IEEE, pp. 415–422. Available at: https://doi.org/10.1109/ICBK.2018.00062 .

Lim, T.-S. (1997) *Contraceptive Method Choice*, *UCI Machine Learning Repository*. Available at: https://doi.org/10.24432/C59W2D .

Lüdtke, O., Robitzsch, A. and West, S.G. (2020) ‘Analysis of Interactions and Nonlinear Effects with Missing Data: A Factored Regression Modeling Approach Using Maximum Likelihood Estimation’, *Multivariate Behavioral Research*, 55(3), pp. 361–381. Available at: https://doi.org/10.1080/00273171.2019.1640104 .

Madden, M.G. (2002) ‘Evaluation of the Performance of the Markov Blanket Bayesian Classifier Algorithm’.

Madley-Dowd, P. *et al.* (2019) ‘The proportion of missing data should not be used to guide decisions on multiple imputation’, *Journal of Clinical Epidemiology*, 110, pp. 63–73. Available at: https://doi.org/10.1016/j.jclinepi.2019.02.016 .

Maghooli, K. *et al.* (2016) ‘Differential Diagnosis of Erythmato-Squamous Diseases Using Classification and Regression Tree’, *Acta Informatica Medica*, 24(5), p. 338. Available at: https://doi.org/10.5455/aim.2016.24.338-342 .

Mandreoli, F. *et al.* (2022) ‘Real-world data mining meets clinical practice: Research challenges and perspective’, *Frontiers in Big Data*, 5. Available at: https://doi.org/10.3389/fdata.2022.1021621 .

Mansourifar, H. and Shi, W. (2018) ‘Toward Efficient Breast Cancer Diagnosis and Survival Prediction Using L-Perceptron’.

Nakai, K. (1996) *Ecoli*, *UCI Machine Learning Repository*.

Nijman, S. *et al.* (2022) ‘Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review’, *Journal of Clinical Epidemiology*, 142, pp. 218–229. Available at: https://doi.org/10.1016/j.jclinepi.2021.11.023 .

Orczyk, T. and Porwick, P. (2013) ‘Influence of missing data imputation method on the classification accuracy of the medical data’, *Journal of Medical Informatics and Technologies*, 22.

Papadopoulos, H. (2011) ‘Reliable Probabilistic Prediction for Medical Decision Support’, in, pp. 265–274. Available at: https://doi.org/10.1007/978-3-642-23960-1_32 .

Perez-Lebel, A. *et al.* (2022) ‘Benchmarking missing-values approaches for predictive models on health databases’.

Pfob, A., Lu, S.-C. and Sidey-Gibbons, C. (2022) ‘Machine learning in medicine: a practical introduction to techniques for data pre-processing, hyperparameter tuning, and model comparison', *BMC Medical Research Methodology*, 22(1), p. 282. Available at: https://doi.org/10.1186/s12874-022-01758-8 .

Rosly, R. *et al.* (2018) 'Analyzing performance of classifiers for medical datasets', *International Journal of Engineering & Technology*, 7(2.15), p. 136. Available at: https://doi.org/10.14419/ijet.v7i2.15.11370 .

Rozemberczki, B. *et al.* (2021) 'Chickenpox Cases in Hungary: a Benchmark Dataset for Spatiotemporal Signal Processing with Graph Neural Networks'.

Rozemberczki, B. (2021) *Hungarian Chickenpox Cases*, *UCI Machine Learning Repository*. Available at: https://archive.ics.uci.edu/ml/datasets/Hungarian+Chickenpox+Cases (Accessed: 28 April 2023).

Rubini, L., Soundarapandian, P. and Eswaran, P. (2015) *Chronic Kidney Disease*, *UCI Machine Learning Repository*. Available at: https://doi.org/10.24432/C5G020 .

Sakar, C. *et al.* (2018) *Parkinson's Disease Classification*, *UCI Machine Learning Repository*. Available at: https://doi.org/10.24432/C5MS4X .

Sakar, C.O. *et al.* (2019) 'A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform', *Applied Soft Computing*, 74, pp. 255–263. Available at: https://doi.org/10.1016/j.asoc.2018.10.022 .

Salgado, C.M. *et al.* (2016) 'Missing Data', in *Secondary Analysis of Electronic Health Records*. Cham: Springer International Publishing, pp. 143–162. Available at: https://doi.org/10.1007/978-3-319-43742-2_13 .

Salzberg, S. (1988) *Echocardiogram Data Set*, *UCI Machine Learning Repository*. Available at: https://archive.ics.uci.edu/ml/datasets/echocardiogram (Accessed: 2 June 2023).

Santhanam, T. and Sundaram, S. (2010) 'Application of CART Algorithm in Blood Donors Classification ', *Journal of Computer Science*, 6(5), pp. 548–552.

Santos, M. *et al.* (2017) *HCC Survival*, *UCI Machine Learning Repository*. Available at: https://doi.org/10.24432/C5TS4S .

Santos, M.S. *et al.* (2015) 'A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients', *Journal of Biomedical Informatics*, 58, pp. 49–59. Available at: https://doi.org/10.1016/j.jbi.2015.09.012 .

Schober, P. and Vetter, T.R. (2020) 'Missing Data and Imputation Methods', *Anesthesia & Analgesia*, 131(5), pp. 1419–1420. Available at: https://doi.org/10.1213/ANE.0000000000005068 .

Singh, N. and Singh, P. (2021) 'Exploring the effect of normalization on medical data classification', in *2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)*. IEEE, pp. 1–5. Available at: https://doi.org/10.1109/AIMV53313.2021.9670938 .

Wang, Z. and Yang, J. (2017) 'Diabetic Retinopathy Detection via Deep Convolutional Networks for Discriminative Localization and Visual Explanation'.

Wolberg, W., Street, W. and Mangasarian, O. (1995) *Breast Cancer Wisconsin (Prognostic)*, *UCI Machine Learning Repository*. Available at: https://doi.org/10.24432/C5GK50 .

Wolberg, Wi. (1992) *Breast Cancer Wisconsin (Original)*, *UCI Machine Learning Repository*. Available at: https://doi.org/10.24432/C5HP4Z .

Wu, Q. and Fokoue, E. (2017) *Epileptic Seizure Recognition*, *UCI Machine Learning Repository*. Available at: https://doi.org/10.24432/C5G308 .

Wu, X. *et al.* (2019) 'Imputation techniques on missing values in breast cancer treatment and fertility data', *Health Information Science and Systems*, 7(1), p. 19. Available at: https://doi.org/10.1007/s13755-019-0082-4 .

Zarchi, M.S., Fatemi Bushehri, S.M.M. and Dehghanizadeh, M. (2018) 'SCADI: A standard dataset for self-care problems classification of children with physical and motor disability', *International Journal of Medical Informatics*, 114, pp. 81–87. Available at: https://doi.org/10.1016/j.ijmedinf.2018.03.003 .

Zhi-Hua Zhou and Xu-Ying Liu (2006) 'Training cost-sensitive neural networks with methods addressing the class imbalance problem', *IEEE Transactions on Knowledge and Data Engineering*, 18(1), pp. 63–77. Available at: https://doi.org/10.1109/TKDE.2006.17 .

Zuo, Z. *et al.* (2021) 'Curvature-based Feature Selection with Application in Classifying Electronic Health Records'. Available at: https://doi.org/10.1016/j.techfore.2021.121127 .

Zwitter, M. and Soklic, M. (1988) *Lymphography*, *UCI Machine Learning Repository*. Available at: https://doi.org/10.24432/C54598 .