Assessing the impact of distance functions on k-nearest neighbours imputation of biomedical datasets*

Miriam S. Santos[1,3], Pedro H. Abreu[1], Szymon Wilk[2], and João Santos[3]

[1] CISUC, Department of Informatics Engineering, University of Coimbra, Portugal
miriams@dei.uc.pt, pha@dei.uc.pt
[2] Institute of Computing Science, Poznan University of Technology, Poznan, Poland
szymon.wilk@cs.put.poznan.pl
[3] IPO-Porto Research Centre, Porto, Portugal joao.santos@ipoporto.min-saude.pt

Abstract. In healthcare domains, dealing with missing data is crucial since absent observations compromise the reliability of decision support models. k-nearest neighbours imputation has proven beneficial since it takes advantage of the similarity between patients to replace missing values. Nevertheless, its performance largely depends on the distance function used to evaluate such similarity. In the literature, k-nearest neighbours imputation frequently neglects the nature of data or performs feature transformation, whereas in this work, we study the impact of different heterogeneous distance functions on k-nearest neighbour imputation for biomedical datasets. Our results show that distance functions considerably impact the performance results of classifiers learned from the imputed data, especially when data is complex.

Keywords: Missing Data · Heterogeneous Data · Data Imputation · Distance Functions · K-Nearest Neighbours · Biomedical Data

1 Introduction

A common data quality problem in healthcare domains is the presence of Missing Data, which consists of absent observations in patients' medical records. Dealing with missing data is of outstanding importance, since absent observations may jeopardise algorithms' predictions, compromising the reliability of patient-oriented models for decision making. k-nearest neighbours (KNN) imputation is a popular imputation technique in healthcare domains, since it takes advantage of the similarity between patients to produce accurate estimates for imputation. Furthermore, it is a nonparametric method which does not require any assumptions on the data [13], has proven to preserve the data distribution [12] and allows for a great interpretability and explainability, crucial in healthcare domains [2].

* This work was supported in part by the project NORTE-01-0145-FEDER-000027 (Norte Portugal Regional Operational Programme – Norte 2020) and in part by the FCT Research Grant SFRH/BD/138749/2018.

Nevertheless, KNN performance largely depends on the distance function used to evaluate such similarity. For heterogeneous data, typical solutions include feature transformation, although leading to lost of information (e.g., discretisation of continuous features) or increased dimensionality (e.g., one-hot encoding) and the use of heterogeneous distance functions that handle different scales [15]. However, besides their heterogeneous nature and susceptibility to missing data, biomedical data is also prone to other difficulty factors, such as data imbalance, the presence of subconcepts in data (small disjuncts), class overlap, and noisy data [11], which make them especially complex domains where choosing suitable distance functions becomes a more strenuous and critical task.

In related work, KNN imputation frequently neglects the nature of data or performs feature transformation [13]. Considering KNN classification, either the studies consider only complete datasets (or derisory amounts of missing values) or the nature of data is ignored [5]. This work studies the impact of different heterogeneous distance functions on KNN imputation, evaluating their effect on the classification performance of biomedical datasets with different characteristics. The purpose of this research is two-fold: ***1) Determining if distance functions impact KNN imputation of biomedical datasets and whether the type of features affected by missing data influences the results*** and ***2) Determining whether obtained results are related to the characteristics of biomedical datasets***, i.e., if there are scenarios where the choice of distance function considerably influences the obtained results. To that end, a benchmark of biomedical datasets with different characteristics was collected and missing data was generated following 4 different variants and percentages (5 to 30%). Then, data imputation is performed using 7 different distance functions and imputation results are evaluated through the analysis of a classifier learned from the imputed data.

To the authors knowledge, no study has yet investigated the impact of different distance functions on the imputation of biomedical data with different characteristics and its effect on classification performance, which constitutes the main contribution of this work. Furthermore, we explore recent distance measures never before studied for imputation purposes, such as SIMDIST and MDE (Section 2), extend MDE to handle categorical data, study redefinitions of popular distance functions (HEOM and HVDM), often overlooked in related work, and propose yet another redefinition of HVDM, which constitute additional contributions.

2 Heterogeneous Distance Functions for Missing Data

All distance functions measure the distance between two patterns $\mathbf{x}_A$ and $\mathbf{x}_B$ through a sum of their individual distances in each j-th feature, $d_j(x_{Aj}, x_{Bj})$, as $D(\mathbf{x}_A, \mathbf{x}_B) = \sqrt{\sum_{j=1}^{p} d_j(x_{Aj}, x_{Bj})^2}$; yet they differ on the computation of individual d_j distances and treatment of missing values (p represents the total number of features and x_{Aj}, x_{Bj} are two values of feature j). The mathematical formulation of all distance functions may be found in the Appendix.

HEOM and HVDM: The definition of $d_j(x_{Aj}, x_{Bj})$ for Heterogeneous Euclidean-Overlap Metric (HEOM) and Heterogeneous Value Difference Metric (HVDM) depends on the type of feature j (equations 1 and 2) [15]. For categorical features, HEOM defines d_j as an overlap metric, d_O (equation 3) whereas HVDM uses d_{vdm} (equation 4). For continuous features, HEOM uses the normalised euclidean distance d_N (equation 5), whereas HVDM considers d_{diff} (equation 6). However, d_O, d_{diff}, d_N and d_{vdm} are only computed if both x_{Aj} and x_{Bj} are observed; otherwise, $d_j(x_{Aj}, x_{Bj}) = 1$.

HEOM-R, HVDM-R and HVDM-S: HEOM-R and HVDM-R [6] consider missing values as "special values": if both x_{Aj} and x_{Bj} are missing, then $d_j(x_{Aj}, x_{Bj}) = 0$ (equation 7). In addition, we propose another redefinition of HVDM: missing values are considered an "special" category and d_{vdm} is applied when only x_{Aj} or x_{Bj} are missing and j is categorical, referred to as HVDM-S (equation 9).

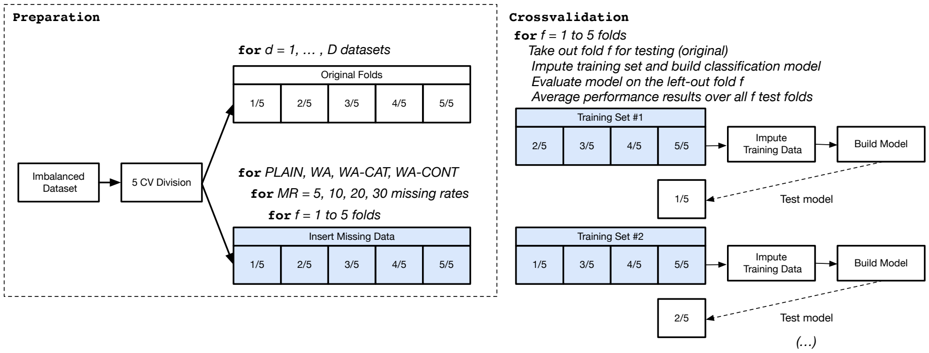
SIMDIST: SIMDIST defines a similarity measure S (equation 8), where s_{ABj} is an intermediate similarity between patterns according to j, s_j represents the mean similarity among all patterns according to j and z is a normalisation function: $z(a) = \frac{a}{a+1}$ [3]. For categorical features, s_{ABj} is defined by equation 10, whereas for continuous features, s_{ABj} is determined by equation 11. $S_{ABj} = \frac{1}{2}$ when x_{Aj} or x_{Bj} are missing which is the equivalent of replacing the missing similarity by the mean similarities of all patterns according to j (equation 8). The individual similarities S_{ABj} are then transformed to distances $D_{ABj} = 1 - S_{ABj}$ and aggregated to produce $D(\mathbf{x}_A, \mathbf{x}_B)$.

MDE: When both values are observed, Mean Euclidean Distance (MD_E) [1] is defined as the euclidean distance (equation 12). When either x_{Aj} or x_{Bj} are missing, MD_E is approximated as the mean distance of each value of x_j to the observed value (equation 14) and when both values are missing, MD_E is approximated as the mean distance between all values of x_j (equation 16). To allow a proper weighting of continuous features with different ranges, a min-max normalisation, $z_i = \frac{x_i - min(x)}{max(x) - min(x)}$ is applied before the euclidean distance is computed. When first proposed, MD_E considered only continuous features. Therefore, starting from the overlap distance, d_O (equation 3) we extended MD_E for categorical features, MD_O: when both values are known, MD_O is the same as d_O; when one value is missing, MD_O is computed as the mean distance between all elements in x_j and the observed value (equation 15); when both values are missing MD_O is determined as the mean distance between all elements in x_j (equation 17). After the individual distances are computed, their aggregation is performed as for the remaining distances.

3 Experimental Setup

We started by collecting 31 complete and binary-classification datasets from open-source repositories (UCI Machine Learning, KEEL, KAGGLE, OPENML), comprising different biomedical contexts, sample sizes, number of features, type of features (continuous and categorical), imbalance ratios (IR) and complexity characteristics. Each dataset is divided into 5 folds following a stratified crossvalidation (SCV) approach (as some datasets have a lower number of minority examples, using 10 folds would result in test sets with a very small amount of minority examples or the need to repeat minority examples across folds). Then, missing data is introduced in the training set, following 4 different variants, herein referred to as Weighted-Plain (PLAIN), Weighted-All (WA), Weighted-Continuous (WA-CONT) and Weighted-Categorical (WA-CAT). The same missing rate was inserted in both classes according to the IR of each dataset (hence the "weighted" designation), to guarantee that missing data is affecting both classes proportionally to their distribution. However, the features affected by missing data differ for each type. PLAIN generation does not control for the number or type of features where missing values are placed. In this case, missing data is generated over the entire dataset with no restrictions, simulating a scenario more likely to be found in real-world domains. Nevertheless, WA approach generates the same percentage of missing values for each feature (all features are equally affected by missing data), whereas WA-CONT and WA-CAT approaches generate the same amount of missing data for all continuous and categorical features, respectively. The goal of comparing different generations variants of missing data is to determine if the type of features (continuous or categorical) affected by missing data influenced the choice of a proper distance function for imputation. Also, missing data is generated at 4 different rates (5, 10, 20 and 30%) under a Missing Completely At Random (MCAR) mechanism. MCAR mechanism was considered for a rigorous control of the missing generation. As missing data is synthetically generated in a multivariate scheme (several features, if not all, are affected by missing data), for each variant and rate, and the considered datasets comprise heterogeneous features, other mechanisms could be compromised due to existing limitations of generation approaches for categorical data [10]. After the injection of missing data, the datasets with missing values are either directly classified with Classification and Regression Trees (CART) model (BASELINE approach) or first imputed with KNN ($k = 1, 3, 5, 7$) and then classified with CART. CART model is also a non-parametric classifier and relatively fast to construct and to provide classification results. Furthermore, it handles missing data directly through the use of surrogate splits (without discarding any patterns from the dataset or assuming a particular missing mechanism) thus allowing a comparison between learning a model with missing data or complete data (via imputation) [14]. Regarding the chosen distance functions, we started by considering distance functions that handled the three components of the problem (continuous, categorical and missing data) or required minimal adjustments, as described in Section 2. On that note, MDE, although lacking the treatment of categorical data, was extended and included given its similarity with HEOM and SIMDIST, yet using the data distribution to handle missing data (as is performed for continuous data). Similarly, we propose HVDM-S as a modification of HVDM, only altering the treatment of missing data and maintaining the remaining aspects regarding continuous and categorical data [9]. Finally, classification performance is evaluated using Accuracy, Sensitivity, Specificity,

Fig. 1: Stratified crossvalidation and missing data generation: missing data is injected after the splitting of the data into training and test sets, for each fold. The same splits are used for all methods (both for training and testing stages).



Precision, F-measure, G-mean and AUC (although for simplicity we present the most relevant metrics for imbalanced domains: Sensitivity, F-measure and G-mean) [8,11]. For each dataset, 10 versions of the crossvalidation procedure were performed, resulting in a 10 × 5 SCV approach (Fig. 1). Overall, 31 *datasets* × 10 *SCV Versions* × 4 *missing variants* × 4 *missing rates* × 7 *distance functions* × 4 *k values* (imputed datasets) + 31 *datasets* × 10 *SCV Versions* × 4 *missing variants* × 4 *missing rates* (for BASELINE approach) sums up to an equivalent of 143,840 datasets evaluated.

4 Results and Discussion

Tables 1 and 2 report on the average sensitivity ranks obtained for each approach, considering training sets with missing values (BASELINE) and training sets imputed with each of the 7 considered distances. Furthermore, results are grouped by missing data variant (PLAIN, WA, WA-CAT and WA-CONT) and missing rate (5% to 30%). Overall, for both $k = 1$ and $k = 3$, HVDM-S is globally the top performing approach, independently of the generation variant. For $k = 1$, where KNN imputation has a more local behaviour, HVDM-S is consistently the best approach for most missing rates ($> 5\%$) in all variants, only surpassed by SIMDIST when missing data is generated exclusively on continuous features. This suggests that although HVDM-S handles efficiently both continuous, categorical and missing values, the strategy used by SIMDIST to handle continuous values might be superior. For $k = 3$, HVDM-S surpasses the remaining approaches for higher missing rates ($> 10\%$), with MDE showing competitive results for WA datasets in lower rates (5 and 10%). Furthermore, for $k = 3$, HVDM-S presents a lower average rank for higher missing rates than for $k = 1$; whereas as the value of k increases further, $k = \{5, 7\}$, KNN imputation shows a more global behaviour, and differences among the approaches

Table 1: CART average sensitivity ranks per missing rate, and variant ($k = 1$).

	MR	*B*	*HEOM*	*HEOM-R*	*HVDM*	*HVDM-R*	*HVDM-S*	*MDE*	*SIMDIST*
PLAIN Datasets	***5%***	5.31	4.50	4.58	3.95	5.18	4.24	**3.79**	4.45
	10%	5.52	4.35	5.31	5.03	4.63	**3.48**	3.73	3.95
	20%	4.32	4.66	5.06	4.77	5.37	**3.32**	4.10	4.39
	30%	4.55	4.47	4.94	4.44	5.35	**3.10**	3.56	5.60
WA Datasets	***5%***	**3.55**	4.98	4.89	4.63	4.61	4.65	**3.55**	5.15
	10%	5.24	4.81	4.87	4.42	5.06	**3.16**	4.18	4.26
	20%	5.10	5.05	4.87	4.34	4.21	**3.63**	3.77	5.03
	30%	5.23	4.27	5.08	3.97	5.21	**3.60**	3.71	4.94
WA-CAT Datasets	***5%***	5.57	4.93	4.12	3.91	**3.86**	4.10	5.03	4.47
	10%	5.02	4.59	5.00	4.64	5.21	**3.12**	3.64	4.79
	20%	4.91	4.38	5.14	4.00	4.78	**3.60**	4.83	4.36
	30%	4.97	4.71	5.02	4.45	4.81	**3.52**	4.29	4.24
WA-CONT Datasets	***5%***	5.31	4.26	4.63	4.08	4.39	4.39	5.03	**3.92**
	10%	4.92	4.79	4.73	4.56	4.37	4.37	4.24	**4.02**
	20%	5.31	4.18	4.71	5.10	**4.08**	**4.08**	4.19	4.35
	30%	**3.92**	4.56	4.71	4.97	4.63	4.63	4.19	4.39

B: *BASELINE;* ***HEOM-R***: *HEOM-REDEF;* ***HVDM-R***: *HVDM-REDEF;* ***HVDM-S***: *HVDM-SPECIAL*

Table 2: CART average sensitivity ranks per missing rate, and variant ($k = 3$).

	MR	*B*	*HEOM*	*HEOM-R*	*HVDM*	*HVDM-R*	*HVDM-S*	*MDE*	*SIMDIST*
PLAIN Datasets	***5%***	5.76	4.56	4.52	4.26	4.61	3.76	**3.74**	4.79
	10%	6.40	4.66	3.94	5.18	3.87	**3.44**	4.31	4.21
	20%	5.40	5.26	4.65	4.29	4.52	**3.48**	3.89	4.52
	30%	6.39	5.02	4.02	4.24	5.00	**2.95**	3.56	4.82
WA Datasets	***5%***	4.44	4.90	4.47	5.06	5.31	3.82	**3.60**	4.40
	10%	5.60	4.06	4.76	4.50	4.44	4.47	**3.94**	4.24
	20%	5.50	4.61	5.11	4.89	4.56	**3.34**	3.76	4.23
	30%	6.21	5.16	3.84	4.15	4.21	**3.53**	4.32	4.58
WA-CAT Datasets	***5%***	5.34	4.28	4.83	3.59	**3.52**	4.31	5.07	5.07
	10%	4.79	4.53	4.93	4.60	4.86	4.00	4.64	**3.64**
	20%	5.72	3.93	4.76	4.29	4.67	**3.76**	5.00	3.86
	30%	4.86	3.98	4.66	5.50	5.12	**3.34**	4.45	4.09
WA-CONT Datasets	***5%***	5.29	4.85	**3.89**	4.48	4.35	4.35	4.23	4.55
	10%	5.87	4.35	4.32	4.42	4.37	4.37	**3.66**	4.63
	20%	5.27	5.00	4.35	4.27	**3.98**	**3.98**	4.19	4.94
	30%	4.98	4.66	4.37	4.82	**4.16**	**4.16**	4.29	4.55

B: *BASELINE;* ***HEOM-R***: *HEOM-REDEF;* ***HVDM-R***: *HVDM-REDEF;* ***HVDM-S***: *HVDM-SPECIAL*

become less clear, although HVDM-S remains in the top performing approaches, especially when missing data is generated across the entire dataset (PLAIN and WA) approaches. This behaviour of k (differences between approaches becoming more smoothed) is expected given that with a larger k-neighborhood, the local properties of KNN which grant it its greatest advantage (taking advantage of the similarity between patterns) become more and more lost. As the analysis of ranks does not provide information of the classification results directly, we analyse also several important performance metrics for complex, imbalanced data, such as Sensitivity, F-measure and G-mean, as shown in Table 3. As follows, HVDM-S is the top performing approach across all metrics and missing rates, and its superiority becomes more evident for higher missing rates (20%

Table 3: CART performance results (mean ± standard deviation) on *PLAIN Datasets* without imputation (BASELINE) and with KNN ($k = 3$) imputation using several distances.

Distance	*MR*	*Sens*	*F-measure*	*G-mean*	*MR*	*Sens*	*F-measure*	*G-mean*
BASELINE	*5%*	0.468 ± 0.331	0.472 ± 0.326	0.536 ± 0.300	*10%*	0.460 ± 0.334	0.463 ± 0.331	0.524 ± 0.306
HEOM		0.482 ± 0.324	0.483 ± 0.317	0.553 ± 0.283		0.479 ± 0.332	0.475 ± 0.319	0.541 ± 0.292
HEOM-R		0.482 ± 0.324	0.483 ± 0.317	0.554 ± 0.283		0.483 ± 0.329	0.480 ± 0.316	0.548 ± 0.288
HVDM		0.480 ± 0.328	0.480 ± 0.320	0.549 ± 0.288		0.475 ± 0.331	0.472 ± 0.320	0.539 ± 0.295
HVDM-R		0.480 ± 0.327	0.479 ± 0.320	0.549 ± 0.286		0.482 ± 0.325	0.478 ± 0.314	0.547 ± 0.285
HVDM-S		**0.485** ± 0.328	**0.485** ± 0.320	**0.556** ± 0.286		**0.487** ± 0.329	**0.481** ± 0.316	**0.550** ± 0.288
MDE		**0.485** ± 0.329	0.484 ± 0.319	0.552 ± 0.288		0.480 ± 0.332	0.474 ± 0.319	0.544 ± 0.290
SIMDIST		0.481 ± 0.328	0.482 ± 0.320	0.551 ± 0.286		0.483 ± 0.332	0.478 ± 0.320	0.544 ± 0.294
BASELINE	*20%*	0.461 ± 0.337	0.459 ± 0.332	0.516 ± 0.311	*30%*	0.436 ± 0.334	0.437 ± 0.334	0.489 ± 0.317
HEOM		0.463 ± 0.326	0.454 ± 0.315	0.519 ± 0.293		0.450 ± 0.320	0.437 ± 0.309	0.505 ± 0.288
HEOM-R		0.469 ± 0.320	0.463 ± 0.311	0.529 ± 0.289		0.461 ± 0.314	0.445 ± 0.302	0.514 ± 0.279
HVDM		0.470 ± 0.327	0.460 ± 0.314	0.526 ± 0.292		0.462 ± 0.321	0.444 ± 0.307	0.512 ± 0.285
HVDM-R		0.466 ± 0.329	0.455 ± 0.315	0.522 ± 0.292		0.456 ± 0.321	0.441 ± 0.309	0.507 ± 0.289
HVDM-S		**0.479** ± 0.323	**0.468** ± 0.310	**0.539** ± 0.284		**0.476** ± 0.321	**0.456** ± 0.303	**0.532** ± 0.276
MDE		0.476 ± 0.328	0.465 ± 0.314	0.534 ± 0.291		0.470 ± 0.324	0.452 ± 0.307	0.523 ± 0.284
SIMDIST		0.468 ± 0.331	0.458 ± 0.319	0.523 ± 0.296		0.456 ± 0.325	0.440 ± 0.311	0.509 ± 0.289

and 30%). However, despite HVDM-S presents the highest performance results, it becomes clear from the analysis of Table 3 that the classification performance is overall poor, even if data is imputed. As we focus solely on the analysis of the effect of data imputation, the datasets did not suffer any pre-processing, such as data oversampling, outlier removal or cleaning approaches. As biomedical datasets are often complex by nature, presenting a considerable imbalance ratio and associated problems such as small disjuncts, overlap and outliers, among others, we moved to a more detailed analysis of the characteristics of the collected datasets, with the objective to determine if some datasets were in fact complex and whether that complexity could be related to differences in performance for some distance functions. To that end, several data complexity measures where computed for each dataset. These metrics regard key properties of datasets such as geometry/topology (L3, N4), class overlap (F1, F2, F3) and class separability (L1, L2, N1, N2 and N3) and have proved to accurately provide important meta-information on the learning abilities of classifiers, especially in imbalanced domains [11]. We found the most informative features to be related to class overlap (F1) and class separability (L2 and N1), as presented in Fig. 2. F1 measures the highest discriminative power of all features in data and lower values indicate more complex problems. In turn, L2 and N1 focus on the characteristics of the decision boundary between classes, where L2 measures the error rate of a support vector machine with linear kernel and N1 measures the fraction of data points connected to the opposite class by an edge in a minimum spanning tree. On contrary to F1, higher values of L2 and N1 indicate more complex problems. Accordingly, the top most complex datasets are *caesarian*, *dmft-health*, *pharynx-1year*, *pharynx-status*, *plasma-retinol*, *schizo*, and *veteran* (Fig. 2), which were further analysed. Fig. 3 compares the mean performance ($k = 1$ and 3, for a more local behaviour of distances) of each dataset for PLAIN variant and a missing rate of 30%, where differences are more relevant (PLAIN variant is also the most likely to encounter in real-world domains where missing data is scattered throughout

Fig. 2: Data complexity measures of the considered datasets: F1, L2 and N1.

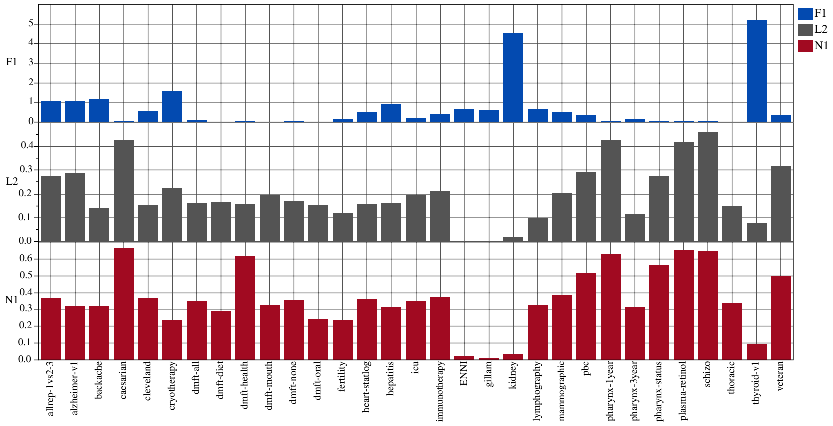
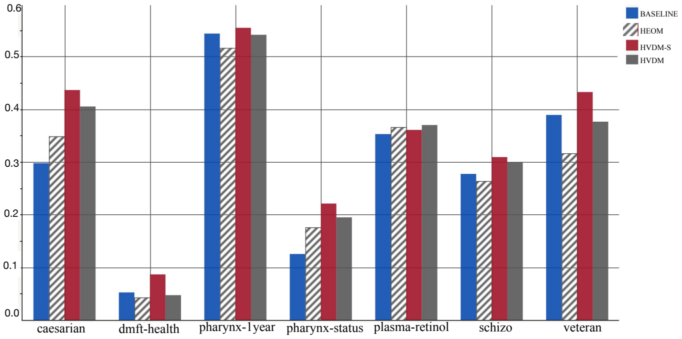


Fig. 3: CART sensitivity results for most complex datasets, considering a PLAIN variant and 30% of missing data (considering $k = 1$ and 3 for a more local analysis).



the entire dataset). For simplicity, and to determine clinical relevance, we focus on a direct comparison of HVDM-S with the BASELINE and the most common used approaches in the literature for healthcare data, i.e., HEOM and HVDM [4,8,7]. However, results for the remaining distances follow a similar trend (for MDE, some datasets obtain similarly performances to HVDM-S, as expected from Table 3). An analysis of Fig. 3 reveals that HVDM-S provides a substantial improvement in sensitivity results for more complex datasets (especially in comparison to HEOM). This suggests that choosing a proper distance function for imputation is important to produce quality training sets and that choice is even more important when data is complex, as determining the most similar patterns becomes a more crucial task to obtain better classification results. To confirm our hypothesis, we have also analysed the impact of increasing missing rate, showing that, as expected, the performance results decrease considerably for more complex datasets (*caesarian*, *dmft-health*, *pharynx-status*, *schizo*, and *veteran*). As a basis for comparison we added *kidney* and *thyroid-v1* datasets, the two datasets with highest discriminative power, i.e., highest F1 values (Fig. 2), which confirms that for easier classification problems, the increase of missing rate has a low impact on classification performance and that, in such cases, differences in performance of models learned from data imputed with different distances are negligible. As a final remark, future work will be focused on further analysing other meta-features to fully characterise the data domains and investigate their relation with obtained imputation and performance results.

5 Conclusions and Clinical Relevance

From the experimental results, the following conclusions may be derived:

- Distance functions impact KNN imputation, where HVDM-S has proved to be a feasible and robust approach for the imputation of heterogeneous biomedical data, independently of the type of features affected by missing data (generation variant);
- HVDM-S shows a particular good behaviour when compared to more common distance function (HEOM and HVDM) for more complex datasets, indicating that choosing a proper distance function becomes crucial when data is complex;
- Missing Data should be considered as yet another data difficulty factor for imbalanced domains, as it influences the computation of distances and assignment of nearest neighbours, becoming specially critical when other factors are present in data;

In an era where the biomedical community is shifting its attention towards the paradigms of Personalised Medicine, where machine learning algorithms play an instrumental role, guaranteeing the quality of data to develop decision-support models is of extreme importance. In that sense, to improve the quality of biomedical data, we explore KNN imputation on the performance of CART models in different missing data scenarios. These are both non-parametric, interpretable and explainable supervised models, which is a critical aspect in healthcare domains. Our results show that distance functions considerably impact data imputation and that for complex data, such as biomedical data, the choice of a proper distance function is of utmost importance. Furthermore, we also show that HEOM, although widely used across medical domains may not be the go-to approach, as others have shown to be more beneficial, especially HVDM-S. Despite the work is focused on data imputation, distance functions are also *i)* crucial for the success of other approaches on clinical domains, such as clustering medical data and subgroup analysis and *ii)* critical for the development and improvement of accurate decision-support models operating with distances among patterns, such as neural networks with radial basis functions or self-organising maps, often used in healthcare domains. Finally, as discussed throughout this work, biomedical data is also subjected to other difficulty factors, which often require some preprocessing prior to developing the classification model. For the case of biomedical imbalanced data with additional data irregularities, popular strategies are based on data oversampling, where a plethora of approaches also rely on the computation of distances to generate new synthetic data.

References

1. AbdAllah, L., Shimshoni, I.: k-means over incomplete datasets using mean euclidean distance. In: Machine Learning and Data Mining in Pattern Recognition, pp. 113–127. Springer (2016)
2. Amorim, J.P., Domingues, I., Abreu, P.H., Santos, J.: Interpreting deep learning models for ordinal problems. In: ESANN (2018)
3. Belanche Muñoz, L.A., Hernández González, J.: Similarity networks for heterogeneous data. In: ESANN 2012. pp. 215–220 (2012)
4. García-Laencina, P., Abreu, P.H., Abreu, M.H., Afonoso, N.: Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. Computers in biology and medicine **59**, 125–133 (2015)
5. Hu, L.Y., Huang, M.W., Ke, S.W., Tsai, C.F.: The distance function effect on k-nearest neighbor classification for medical datasets. SpringerPlus **5**(1), 1304 (2016)
6. Juhola, M., Laurikkala, J.: On metricity of two heterogeneous measures in the presence of missing values. Artificial Intelligence Review **28**(2), 163–178 (2007)
7. Sáez, J.A., Krawczyk, B., Woźniak, M.: Handling class label noise in medical pattern classification systems. Journal of Medical Informatics & Technologies **24** (2015)
8. Santos, M.S., Abreu, P.H., García-Laencina, P., Simão, A., Carvalho, A.: A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. Journal of biomedical informatics **58**, 49–59 (2015)
9. Santos, M.S., Abreu, P.H., Wilk, S., Santos, J.: How distance metrics influence missing data imputation with k-nearest neighbours. Pattern Recognition Letters **136**, 111–119 (2020)
10. Santos, M.S., Pereira, R.C., Costa, A., Soares, J., Santos, J., Abreu, P.H.: Generating synthetic missing data: A review by missing mechanism. IEEE Access **1**(1), 1–18 (2019)
11. Santos, M.S., Soares, J.P., Abreu, P.H., Araújo, H., Santos, J.: Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [research frontier]. IEEE Computational Intelligence Magazine **13**(4), 59–76 (2018)
12. Santos, M.S., Soares, J.P., Abreu, P.H., Araújo, H., Santos, J.: Influence of data distribution in missing data imputation. In: Conference on Artificial Intelligence in Medicine in Europe. pp. 285–294. Springer (2017)
13. Tutz, G., Ramzan, S.: Improved methods for the imputation of missing data by nearest neighbor methods. Computational Statistics & Data Analysis **90**, 84–99 (2015)
14. Twala, B., Cartwright, M.: Ensemble missing data techniques for software effort prediction. Intelligent Data Analysis **14**(3), 299–331 (2010)
15. Wilson, R., Martinez, T.: Improved heterogeneous distance functions. Journal of artificial intelligence research **6**, 1–34 (1997)

Table 1: Mathematical formulation of heterogeneous distance functions that handle missing data.

HEOM	**HVDM**
$\mathrm{d}_j(x_{Aj}, x_{Bj}) = \begin{cases} 1, & \text{if } j \text{ is missing in } x_{Aj} \text{ or } x_{Bj}, \\ d_O(x_{Aj}, x_{Bj}), & \text{if } j \text{ is a categorical feature,} \\ d_N(x_{Aj}, x_{Bj}), & \text{if } j \text{ is a continuous feature} \end{cases} \quad (1)$	$\mathrm{d}_j(x_{Aj}, x_{Bj}) = \begin{cases} 1, & \text{if } j \text{ is missing in } x_{Aj} \text{ or } x_{Bj}, \\ d_{vdm}(x_{Aj}, x_{Bj}), & \text{if } j \text{ is a categorical feature,} \\ d_{diff}(x_{Aj}, x_{Bj}), & \text{if } j \text{ is a continuous feature} \end{cases} \quad (2)$
$\mathrm{d}_O(x_{Aj}, x_{Bj}) = \begin{cases} 0, & \text{if } x_{Aj} = x_{Bj} \\ 1, & otherwise \end{cases} (3)$	$\mathrm{d}_{vdm}(x_{Aj}, x_{Bj}) = \sqrt{\sum_{c=1}^{C} \left\| \frac{N_{x_{Aj},c}}{N_{x_{Aj}}} - \frac{N_{x_{Bj},c}}{N_{x_{Bj}}} \right\|^2} (4)$
$\mathrm{d}_N(x_{Aj}, x_{Bj}) = \frac{\left\|x_{Aj} - x_{Bj}\right\|}{max(x_j) - min(x_j)} (5)$	$\mathrm{d}_{diff}(x_{Aj}, x_{Bj}) = \frac{\left\|x_{Aj} - x_{Bj}\right\|}{4\sigma_{x_j}} (6)$
HEOM-R/HVDM-R/HVDM-S	**SIMDIST**
$\mathrm{d}_j(x_{Aj}, x_{Bj}) = \begin{cases} 1, & \text{if } j \text{ is missing only on } x_{Aj} \text{ or } x_{Bj}, \\ 0, & \text{if } j \text{ is missing in both } x_{Aj} \text{ and } x_{Bj} \end{cases} (7)$	$\mathrm{S}_{ABj} = \begin{cases} \frac{1}{2}, & \text{if either } x_{Aj} \text{ or } x_{Bj} \text{ are missing,} \\ z\left(\frac{s_{ABj}}{s_j}\right), & \text{if both } x_{Aj} \text{ and } x_{Bj} \text{ are observed} \end{cases} (8)$
$\mathrm{d}_j(x_{Aj}, x_{Bj}) = \begin{cases} 0, & \text{if } x_{Aj} \text{ and } x_{Bj} \text{ are both missing,} \\ 1, & \text{if } x_{Aj} \text{ or } x_{Bj} \text{ are missing and } j \text{ is continuous,} \\ d_{vdm}(x_{Aj}, x_{Bj}), & \text{if } x_{Aj} \text{ or } x_{Bj} \text{ are missing and } j \text{ is categorical} \end{cases} (9)$	$\mathrm{s}_{ABj} = \begin{cases} 0, & \text{if } x_{Aj} \neq x_{Bj}, \\ 1 - P_{lj}, & \text{if } x_{Aj} = x_{Bj} \end{cases} (10)$
	$\mathrm{s}_{ABj} = 1 - \frac{\|x_{Aj} - x_{Bj}\|}{max(x_j) - min(x_j)} (11)$

MDE
$\mathrm{MD}_E(x_{Aj}, x_{Bj}) = (x_{Aj} - x_{Bj})^2 (12)$
$\mathrm{MD}_O(x_{Aj}, x_{Bj}) = \begin{cases} 0, & \text{if } x_{Aj} = x_{Bj} \\ 1, & otherwise \end{cases} (13)$
$\mathrm{MD}_E(x_{Aj}, x_{Bj}) = E\Big((x - x_{Bj})^2\Big) = \int p(x)(x - x_{Bj})^2 dx = (x_{Bj} - \mu_x)^2 + \sigma_x^2 (14)$
$\mathrm{MD}_O(x_{Aj}, x_{Bj}) = \sum_x p(x)\ d_O(x, x_{Bj}) = \sum_{x \neq x_{Bj}} p(x) = 1 - p(x_{Bj}) (15)$
$\mathrm{MD}_E(x_{Aj}, x_{Bj}) = \int\int p(x)p(y)(x - y)^2 dxdy = \Big(E(x) - E(y)\Big)^2 + \sigma_x^2 + \sigma_y^2 = 2\sigma_x^2 (16)$
$\mathrm{MD}_O(x_{Aj}, x_{Bj}) = \sum_x \sum_y p(x)p(y)\ d_O(x, y) = \sum_x \sum_{x \neq y} p(x)p(y) = 1 - \sum_x p^2(x) (17)$

Notes: In the formulae of MD_E and MD_O, we consider $x = x_j$ and μ_x and σ_x are equivalent to μ_{x_j} and σ_{x_j} (mean and standard deviation of all the observed values of x_j.)
Similarly, we consider the auxiliary variable $y = x_j$. $E(x)$ corresponds to the expected value of x and $p(x)$ is the probability distribution of x.
For SIMDIST, P_{lj} is the fraction of patterns that takes value x_{lj} for x_j.
In practice, P_{lj} is the fraction of examples that assume value x_{Aj} or x_{Bj} for x_j, since for this computation they are equal.