Computer Science and Information Systems 17(2):379–401 https://doi.org/10.2298/CSIS190710003V

Missing Data Imputation in Cardiometabolic Risk Assessment: A Solution Based on Artificial Neural Networks

Dunja Vrbaški[1], Aleksandar Kupusinac[1⋆], Rade Doroslovački[1], Edita Stokić[2], and Dragan Ivetić[1]

[1] University of Novi Sad, Faculty of Technical Sciences
Trg Dositeja Obradovića 6, 21000 Novi Sad, Republic of Serbia
{dunja.vrbaski, sasak, rade.doroslovacki, ivetic}@uns.ac.rs
[2] University of Novi Sad, Faculty of Medicine
Hajduk Veljkova 3, 21000 Novi Sad, Republic of Serbia
edith@sezampro.rs

Abstract. A common problem when working with medical records is that some measurements are missing. The simplest and the most common solution, especially in machine learning domain, is to exclude records with incomplete data. This approach produces datasets with reduced statistical power and can even lead to biased or erroneous final results. There are, however, many proposed imputing methods for missing data. Although some of them, such as multiple imputation, are mature and well researched, they can be prone to misuse and are not always suitable for building complex frameworks. This paper explores neural networks as a potential tool for imputing univariate missing laboratory data during cardiometabolic risk assessment, comparing it to other simple methods that could be easily set up and used further in building predictive models. We have found that neural networks outperform other algorithms for diverse fraction of missing data and different mechanisms causing their missingness.

Keywords: missing data, cardiometabolic risk, artficial neural networks

1. Introduction

Missing data is a well known and commonly present problem in both research and industry. Many datasets contain information that is incomplete, due to a variety of reasons: data could have been unavailable, recorded incorrectly or not collected at all, damaged or lost. Health data is not an exception. Actually, those datasets are very prone to having different types of missing data according to its structure, volume and relation with observed data . Not dealing with missing data properly could represent a significant problem for further analysis or building predictive models [23] [33] [28] [12] [1] [41].

The subject of this study is to explore whether a machine learning method, such as an artificial neural network (ANN), could serve as an imputation tool for missing data in cardiometabolic risk assessment (CMR), comparing it to several other single imputation statistics methods. More specifically, we are interested in imputation of missing values

⋆ Corresponding Author: Aleksandar Kupusinac (sasak@uns.ac.rs)

that constitute outcome CMR values and which would not be further used in building final predictive models, rather than missing values for predictors or outcomes on their own.

1.1. Cardio-Metabolic Risk Assessment

There is a number of factors that participate in development of cardiovascular diseases and which are related to CMR [37] [32] [16]. In a clinical routine for obese patients, an evaluation of CMR starts with an anamnesis, an estimation of nutritional status and an adipose tissue distribution. These steps are usually performed by simple anthropometric measurements. The following steps include laboratory analyses of lipids and lipoproteins levels, glycemia, insulinemia and other indicators of obesity comorbidities. Since those procedures have a higher level of invasiveness and they induce additional costs, there is an interest in building predictive models for risk scores that rely only upon inexpensive, commonly available and non-laboratory data.

But, in order to build that cost-effective, prognostic machine learning model, such as the one based on simple parameters and neural network algorithm [19], we need this laboratory data, since it is used for computing the outcome (CMR) values in the data preprocessing phase. If all such subjects, for which only some laboratory data is missing, are omitted, then the number of outcomes, and accordingly, the dataset used for training, could be significantly reduced, making learning more challenging and the results potentially erroneous.

1.2. Missing Data and Machine Learning

In statistics, analysis of the missing data itself is an important task. Data can be missing as an univariate or a multivariate, missingness can follow some pattern and an origin of the missingness could be explained by the observed or the missing data itself. We are using standard notion when differentiating origin of missingness, in literature usually called the missingness mechanism [34] [24]:

- When missing data does not depend neither on observed nor on unobserved values (missing completely at random; MCAR).
- When missing data does not depend on the unobserved, but may depend on the observed values (missing at random; MAR).
- When missing data depends on the unobserved values themselves (missing not at random; MNAR).

Based on the analysis of missing data itself, an appropriate statistic method could be performed for dealing with missing data. Removing the data that contains the missing information is the easiest and, very often, a method of choice in machine learning application domain. But, this approach, where only the complete cases are used, could lead to flawed or unreliable results. Even in the case of MCAR data, where deletion produces unbiased results, we could end up reducing analysis power and weakening some of our tests [10]. In the case of MAR or MNAR data, when there is an underlying reason for missingness, and especially when proportion of missing data is larger, we must exercise

caution. In contrast to ignoring data with missing values, one could choose some imputation method if assumptions for the selected method are met [26]. There are many proposed imputation methods. Properly chosen and used method can significantly reduce the impact which missing data has on the research result. That means that an approach to a missing data problem should be careful, in order to make an educated decision whether the data could safely be deleted, or how and why some imputation method is chosen.

On the other hand, in an engineering and machine learning domain, such analysis is usually not performed, mostly because engineers are not very interested in explaining the data, but more in building and validating their models using that data [4] [40]. If an imputation method that is not sensitive to the nature or volume of missing data and that does not require previous analysis could be developed, that would enable automation of the data preprocessing and feature engineering task, which many recognize as one of the holy grails of machine learning [5].

Machine learning methods themselves are good candidates for such a task, since in their essence is to learn complex relations and learn them from the data without additional instructions. Therefore, we choose to explore how one machine learning algorithm handles imputation in different scenarios with missing data in CMR assessment. We have chosen ANN, amongst other possible algorithms, since it is already shown that it can successfully predict CRM from non-laboratory values, and we presume that there will also exist a dependency function between CMR and laboratory data which some ANNs can approximate. Accordingly, in this paper, we have considered different amount with different missingness mechanism of univariate missing laboratory data, hypothesizing that there are ANNs which could successfully deal with it, regardless of the nature and volume of the data that is missing.

There are some previous results that explore comparison between methods for imputation [25] [11] [14] and even study ANNs within missing data problem [3] [22] [21] [29] [36]. However, all of the research which has been pointed out is somewhat different than our goal. These papers either: explore imputation for predictor values; handle missing data to explore and describe the data, observe the estimates such as regression coefficients and standard errors; use multiple imputation; use a large number of predictors and big data to train machine learning models. On the other side, our subject is imputation through ANN that explores imputing values that are used to calculate outcome values, further leading to building predictive, machine learning, models and one that uses small, simple structured data and explore all three missingness mechanisms. Moreover, although ANNs are used to predict CMR, they are not researched as imputation tool in this domain and in this manner.

2. Methodology

Through simulation and analysis, we have compared neural networks with other single imputation methods in univariate missing data for laboratory values through different settings that reflect different missingness scenarios.

Firstly, we have established structures of ANNs that will be used in comparison. Then we have simulated different scenarios of missing data occurrence and compared performance of ANNs with other methods using several measurements.

2.1. Data

Dataset was produced as a result of the study at the Department of Endocrinology, Diabetes and Metabolic Disorders of the Clinical Center of Vojvodina in Novi Sad, Serbia. The inquired group consisted of 2985 individual respondents, 1980 women and 1005 men, aged 18 to 69 years. The study was conducted in accordance with the Declaration of Helsinki and approved by Ethical Committee of the Clinical Center of Vojvodina (No. 0020/649).

Dataset contains the following CMR risk factors:

- non-laboratory: gender (GEN), age (AGE), body mass index (BMI), waist-to-height ratio (WHtR)
- laboratory: triglycerides (TG), total cholesterol (TCH), low-density lipoprotein (LDL), high-density lipoprotein (HDL) and glycemia (GLY).

Descriptive statistics for the data is shown in the Table 1.

Table 1. Descriptive statistics for risk factors in the CMR dataset

	Mean	St.Dev.	Min	Max
AGE	43.413	10.615	18	69
BMI	29.732	6.472	16.600	50.440
WHtR	0.565	0.091	0.338	0.899
LDL	3.762	0.950	2.030	10.140
HDL	1.124	0.262	0.460	2.090
TCH	5.952	1.376	2.770	13.240
TG	2.057	1.819	0.350	27.320
GLY	5.145	1.321	2.800	13.800

Since adipose tissue for men shows a tendency towards central or abdominal accumulation, male gender bears higher potential risk of cardiovascular diseases. With women, the risk increases with aging, due to the adipose tissue centralization. It is known that acceleration of atherosclerosis increases with age and that the cardiometabolic risk increases with age. Beside the gender and genetic predisposition, this is an additional risk factor that cannot be controlled.

BMI is an indication of nutritional state and is used to quantify the level of obesity. Despite of a lot of controversy about its reliability in fat mass prediction, it shows high efficiency in cardiovascular risk prediction. Values of BMI over $25\ kg/m^2$ correspond to being overweight, and values over $30\ kg/m^2$ correspond to obesity [27]. BMI is calculated as a ratio of body mass and body height squared. Body weight is measured with a balance beam scale. Body height is measured with Harpenden anthropometer (Holtain Ltd, Croswell, UK) with precision of $0.1\ cm$.

Waist circumference is correlated with the amount of visceral abdominal adipose tissue, but also with the level of lipids, lipoproteins and insulin and it is a significant predictor of the obesity comorbidity. An index calculated as a ratio of waist circumference and body height (WHtR) has been shown to be a better risk indicator and the values WHtR≥ 0.5

are considered to indicate increased risk [2] [18]. Waist circumference is measured with a measurement tape with precision of $0.1\ cm$. It is measured at half the distance between the lowest point of the costal arch and the highest point of the lilac crest.

Disturbances of lipid and lipoprotein metabolism are present in 30% of obese persons. They are manifested as one or more of the following disruptions: hypercholesterolemia, hypertriglyceridemia, protective HDL-cholesterol level drop off, raised level of LDL-cholesterol and increased fraction of small, dense, aterogenous LDL-particles. In our study, cholesterol and triglycerides levels are determined by the standard enzyme procedure. The values of HDL-cholesterol were determined by precipitation procedure with sodium-phosphor-wolframate. The values of LDL-cholesterol were calculated using Friedewald's formula [8].

Hyperglycemia is also a risk factor for cardiovascular diseases. Increased level of glucose accelerates the process of atherosclerosis by increasing the oxidative stress and protein glycolization [20]. In this research, the glycemia values were determined using Dialab glucose GOD-PAP method. All inquiries were taken during the morning hours (after fasting overnight).

In this research, we have observed missing data imputation for: high-density cholesterol (HDL), low-density cholesterol (LDL), total cholesterol (TCH), triglycerides (TG) and glycemia (GLY) using following cut off values as indication of cardiometabolic risk [7] [13]:

- HDL $<$ 1.29 (woman) and $<$ 1.03 (man)
- LDL $\geq$ 3.3
- TCH $\geq$ 5.2
- TG $\geq$ 1.71
- GLY $\geq$ 6.1

Distribution of HDL, LDL, TCH, TG and GLY is shown in Figure 1 with highlighted CMR threshold values. For HDL, since different values are used for man and woman, plot has two lines annotated with *M* (male) and *F* (female) respectively.

2.2. Structure of the Neural Networks

For each variable of interest, we have tested several networks in order to find their optimal structure. Input vectors for ANNs are all variables from dataset except variable of interest which represent output value. Since the data is simple, and not massive, we have opted for single layered feedforward artificial network. For each variable, ANNs with 2 to 10 neurons in the hidden layer are tested. Data split for training and testing is performed through bootstrap resampling [6]. For each number of hidden neurons, 100 simulations of ANN testing are performed.

Experiments were performed using *R* software v3.5.0 [30] with packages *neuralnet* v1.33 [9] and *caret* v6.0 [17] with following settings for ANN:

- training algorithm: resilient backpropagation
- starting weights: random initialization
- activation function: tanh (tangent hyperbolicus transfer function)
- output neuron: linear function

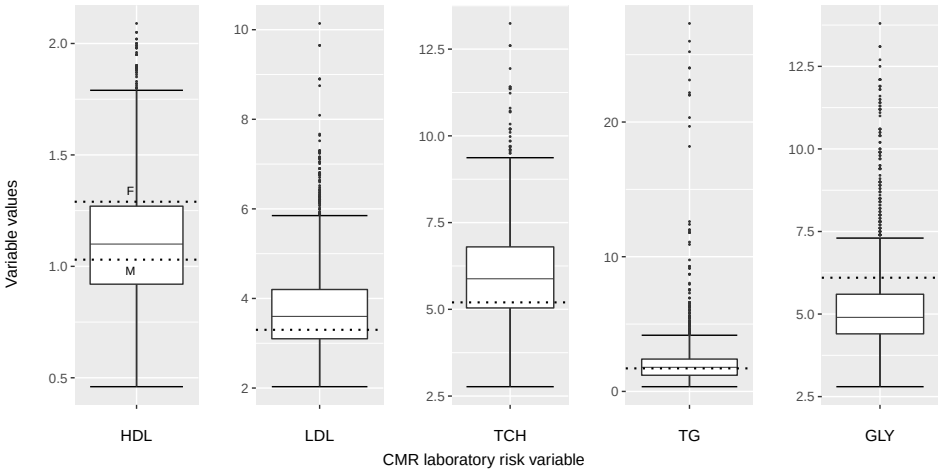

Fig. 1. Boxplots representing distribution of laboratory CMR risk values in the dataset. Cutoff values, used as indication of CMR, are highlighted with dashed lines. For HDL, two distinctive cutoff lines are shown corresponding to different genders

– stopping criteria: threshold for the partial derivatives of the error function or maximum number of steps, whichever is reached before.

Instead of the regular back propagation algorithm, resilient backpropagation (rprop+) is used [15] [31], as a faster method that does not require learning rate as a parameter for its training. Since the study covers a number of simulations, improving speed and reducing grid search for hyperparameter tuning in this phase of work was beneficial.

An optimal number of neurons for each variable was selected according to the number of neurons which, on average, produced the smallest root mean squared error (RMSE). According to this criteria, networks trained in the second part of the research for HDL, LDL, TCH, TG and GLY were networks with single hidden layer and number of neurons: 6, 9, 5, 2 and 2, respectively. Results are shown in Table 2.

2.3. Comparison of Imputation Methods

In the second part of the study, missing data is artificially introduced in the dataset. Accuracy of imputation for neural networks and several other methods is addressed. For each variable, the missing data is introduced by varying percentage (10%, 20%, 30% and 50%) and according to different missingness mechanism (MCAR, MAR and MNAR). For each combination of settings (60 in total), comparison of 5 imputation methods was performed through 100 simulations.

Depending on the selected mechanism, percentage of the original data is removed from the dataset. Data that is missing under MCAR assumption is randomly selected based solely on the percentage of missing data. For both MAR and MNAR mechanism, logistic function for probability that data is missing was used. To simulate MAR mechanism, values for AGE and BMI were used to introduce missingness. Observations with

Table 2. Average **RMSE** errors produced through 100 simulations during testing of neural networks with different number of hidden neurons. Smallest values are highlighted. For each variable minimum and maximum values are displayed

Variable	LDL	HDL	TCH	TG	GLY
(min - max)	*(2.03-10.14)*	*(0.46 - 2.09)*	*(2.77 - 13.24)*	*(0.35 - 27.32)*	*(2.80 - 13.80)*
Hidden neurons					
2	0.74019	0.24994	0.93934	**1.78078**	**1.25208**
3	0.72843	0.24584	0.92250	1.84684	1.26054
4	0.71176	0.24282	0.91913	1.86890	1.27322
5	0.70994	0.24334	**0.91842**	1.93557	1.28708
6	0.70764	**0.24189**	0.92108	1.94169	1.28131
7	0.70391	0.24341	0.92424	1.82648	1.28666
8	0.70358	0.24198	0.92970	1.84911	1.27962
9	**0.70343**	0.24529	0.92365	1.83521	1.27707
10	0.70500	0.24631	0.92339	1.84555	1.28290

lower AGE and BMI are considered to have bigger probability of missing laboratory data. For MNAR mechanism, the same variable which was investigated was used as a cause for missingness. Observations with lower values, except for HDL, are considered to have greater probability of missing data. For HDL, lower values had lower probability of missing data. Function *ampute* from *R* library *mice* v3.3.0 was used to carry out those deletions (amputations) [38] [39] [35].

For all data sets produced in the described manner, 5 different imputation methods are performed: neural network (NN), predictive mean matching (PMM), stochastic linear regression (SLR), random forest (RF) and mean imputation (MEAN). Neural network training and imputation is performed using the settings and architectures obtained in first set of experiments. Other methods are implemented using *R* package *mice*. SLR and MEAN methods did not require special parameters. For RF, training number of trees was set to 10. For PMM, all variables except the one of interest, are used for finding five possible donors and distance between predicted and drawn values was used as a matching distance.

Comparison between methods was based on imputation performance which is evaluated by: root mean squared error (RMSE), mean absolute percentage error (MAPE) and classification accuracy (CA) between imputed and original values. RMSE measures deviation of this difference and is used as common metric for model comparison. MAPE explain accuracy as percentage error and is given due to its intuitive interpretation. CA measures what portion of imputed values will be correctly classified as risk factor for CRM according to their threshold values. Lower values for RMSE and MAPE denote better performance, and higher CA values indicate better results.

3. Results

Comparison results for imputation of HDL, LDL, TCH, GLY and TG are respectively shown in tables 3, 4, 5, 6 and 7. For each variable, missingness mechanism and percent of missing data average results for RMSE, MAPE and CA are shown where best performance values are highlighted. Also, due to space constraints, graphical illustration of obtained results is given in Appendix (figures: 2, 3, 4, 5, 6).

Table 3. Algorithm comparison for variable **HDL** considering three types of missing data mechanisms (MAR, MCAR, MNAR) and different levels of missingness (10, 20, 30 and 50%) based on imputation accuracy (root mean squared error, mean absolute percentage error and classification accuracy)

HDL	RMSE	MAPE	CA	RMSE	MAPE	CA
MCAR						
	10%			20%		
NN	0.2802	20.6955	**0.7316**	0.2785	20.5654	**0.7202**
PMM	0.3608	26.1986	0.6050	0.3617	26.3017	0.5969
SLR	0.3667	27.1279	0.5854	0.3660	27.1464	0.5831
RF	0.2946	**18.8455**	0.7010	0.2953	**19.2711**	0.6936
MEAN	**0.2610**	19.2869	0.6953	**0.2612**	19.3118	0.6903
	30%			50%		
NN	0.2794	20.7115	0.6382	0.2797	20.6359	0.6398
PMM	0.3617	26.4378	0.5953	0.3628	26.4785	0.5941
SLR	0.3649	27.0686	0.5857	0.3659	27.1615	0.5856
RF	0.3035	20.0461	0.6826	0.3143	21.3192	0.6623
MEAN	**0.2617**	**19.4449**	**0.6902**	**0.2615**	**19.3626**	**0.6924**
MAR						
	10%			20%		
NN	**0.2389**	**16.9590**	**0.7391**	**0.2406**	**17.2297**	**0.7290**
PMM	0.3598	26.2777	0.6071	0.3587	26.2944	0.6063
SLR	0.3694	27.7332	0.5881	0.3646	27.4427	0.5913
RF	0.2933	19.3031	0.7019	0.2982	20.0395	0.6914
MEAN	0.2601	19.7928	0.6932	0.2592	19.8083	0.6914
	30%			50%		
NN	**0.2437**	**17.4571**	**0.7261**	**0.2453**	**17.5117**	**0.7200**
PMM	0.3597	26.2950	0.6077	0.3616	26.3412	0.5996
SLR	0.3651	27.3646	0.5928	0.3663	27.2623	0.5881
RF	0.3083	21.0124	0.6747	0.3168	21.7676	0.6611
MEAN	0.2597	19.8737	0.6903	0.2609	19.6912	0.6901
MNAR						
	10%			20%		
NN	**0.2394**	**23.3216**	0.6742	**0.2581**	**25.5564**	0.6472
PMM	0.3582	32.0925	0.5974	0.3724	33.7851	0.5734
SLR	0.3635	33.6794	0.5693	0.3772	35.2030	0.5442
RF	0.2856	23.5159	**0.6986**	0.3080	26.2282	**0.6604**
MEAN	0.2575	26.5496	0.6231	0.2773	28.7606	0.6227
	30%			50%		
NN	**0.2818**	**28.3264**	**0.6240**	**0.2675**	**23.1169**	**0.6560**
PMM	0.3889	35.7334	0.5439	0.3762	30.7928	0.5674
SLR	0.3928	37.0190	0.5188	0.3783	31.4889	0.5472
RF	0.3305	29.1086	0.6188	0.3238	25.3834	0.6331
MEAN	0.3025	31.5800	0.6235	0.2835	25.5359	0.6444

Table 4. Algorithm comparison for variable **LDL** considering three types of missing data mechanisms (MAR, MCAR, MNAR) and different levels of missingness (10, 20, 30 and 50%) based on imputation accuracy (root mean squared error, mean absolute percentage error and classification accuracy)

LDL	RMSE	MAPE	CA	RMSE	MAPE	CA
MCAR						
	10%			20%		
NN	**0.6517**	**13.5230**	**0.7495**	**0.6649**	**13.6613**	**0.7549**
PMM	0.9998	19.5170	0.6591	1.0105	19.5738	0.6638
SLR	1.0302	22.1732	0.6555	1.0307	22.1106	0.6581
RF	0.8660	14.4301	0.7272	0.8870	15.0236	0.7171
MEAN	0.9451	19.1504	0.6505	0.9451	19.1504	0.6552
	30%			50%		
NN	**0.6788**	**13.8307**	**0.7500**	**0.7112**	**14.3998**	**0.7425**
PMM	1.0119	19.6088	0.6607	1.0171	19.6594	0.6610
SLR	1.0301	22.0974	0.6589	1.0318	22.1538	0.6560
RF	0.9087	15.5668	0.7101	0.9508	16.8752	0.6884
MEAN	0.9495	19.2054	0.6522	0.9521	19.2136	0.6507
MAR						
	10%			20%		
NN	**0.6387**	**13.5412**	**0.7508**	**0.6704**	**14.0882**	**0.7409**
PMM	0.9611	19.2735	0.6520	0.9655	19.2723	0.6549
SLR	1.0151	22.4694	0.6440	1.0253	22.7197	0.6429
RF	0.8646	15.0221	0.7086	0.8904	15.7338	0.6998
MEAN	0.9612	20.7286	0.6022	0.9659	20.9315	0.6040
	30%			50%		
NN	**0.6903**	**14.3724**	**0.7374**	**0.7274**	**14.7458**	**0.7349**
PMM	0.9738	19.4508	0.6519	1.0069	19.6371	0.6562
SLR	1.0262	22.6180	0.6442	1.0343	22.3459	0.6507
RF	0.9206	16.6055	0.6841	0.9551	17.1912	0.6780
MEAN	0.9696	21.1739	0.6037	0.9659	20.1857	0.6341
MNAR						
	10%			20%		
NN	**0.6149**	**15.7534**	0.6277	**0.6518**	**16.7199**	0.6057
PMM	0.8784	20.9437	0.5788	0.9152	22.0134	0.5657
SLR	0.9903	25.8651	0.5446	1.0144	26.4461	0.5435
RF	0.7268	15.8575	**0.6670**	0.7731	17.3816	**0.6369**
MEAN	0.8324	25.4867	0.3621	0.9006	27.7790	0.3618
	30%			50%		
NN	**0.6947**	**18.0946**	0.5807	**0.7176**	**16.3713**	**0.6638**
PMM	0.9458	22.9883	0.5546	0.9969	21.4488	0.6240
SLR	1.0457	27.2633	0.5358	1.0464	24.5275	0.6051
RF	0.8347	19.4666	**0.6006**	0.9259	18.9541	0.6466
MEAN	0.9902	30.8539	0.3622	0.9869	25.4807	0.5285

Table 5. Algorithm comparison for variable **TCH** considering three types of missing data mechanisms (MAR, MCAR, MNAR) and different levels of missingness (10, 20, 30 and 50%) based on imputation accuracy (root mean squared error, mean absolute percentage error and classification accuracy)

TCH	RMSE	MAPE	CA	RMSE	MAPE	CA
MCAR						
	10%			20%		
NN	**0.9044**	**11.2777**	**0.8465**	**0.9096**	**11.3255**	**0.8447**
PMM	1.3849	17.9032	0.7646	1.3858	17.8030	0.7667
SLR	1.4125	19.3143	0.7096	1.4071	19.3211	0.7167
RF	1.1452	13.2419	0.8244	1.1760	13.6968	0.8172
MEAN	1.3765	19.6452	0.7120	1.3768	19.5699	0.7130
	30%			50%		
NN	**0.9159**	**11.3713**	**0.8477**	**0.9254**	**11.5042**	**0.8448**
PMM	1.3831	17.7930	0.7669	1.3798	17.7741	0.7672
SLR	1.4101	19.3680	0.7142	1.4124	19.3726	0.7138
RF	1.2203	14.5303	0.8060	1.2504	15.1500	0.7976
MEAN	1.3772	19.6186	0.7127	1.3788	19.5867	0.7135
MAR						
	10%			20%		
NN	**0.8916**	**11.6872**	**0.8259**	**0.9012**	**11.8159**	**0.8257**
PMM	1.3595	18.3046	0.7359	1.3640	18.2839	0.7381
SLR	1.4035	20.3694	0.6782	1.4078	20.4852	0.6761
RF	1.1616	14.3625	0.7973	1.2017	15.1341	0.7838
MEAN	1.4254	22.5705	0.6222	1.4333	22.8869	0.6230
	30%			50%		
NN	**0.9152**	**11.9812**	**0.8235**	**0.9405**	**11.8581**	**0.8342**
PMM	1.3671	18.3581	0.7371	1.3727	17.9005	0.7556
SLR	1.4080	20.4700	0.6749	1.4083	19.7614	0.6981
RF	1.2422	16.0273	0.7751	1.2683	15.9046	0.7849
MEAN	1.4480	23.3124	0.6225	1.4142	21.3220	0.6764
MNAR						
	10%			20%		
NN	**0.8253**	**13.4357**	**0.7565**	**0.8647**	**14.2280**	**0.7442**
PMM	1.3150	20.6076	0.6800	1.3530	21.4554	0.6689
SLR	1.3586	23.1945	0.6075	1.3939	23.7496	0.6021
RF	1.1121	16.0774	0.7388	1.1956	17.8133	0.7139
MEAN	1.3765	26.6161	0.4466	1.4726	28.6865	0.4482
	30%			50%		
NN	**0.9302**	**15.5389**	**0.7213**	**0.9445**	**13.6014**	**0.7896**
PMM	1.4125	22.4742	0.6621	1.4210	20.3798	0.7268
SLR	1.4522	25.0296	0.5859	1.4406	21.9095	0.6678
RF	1.2946	19.8007	0.6870	1.3073	18.0855	0.7481
MEAN	1.6098	31.7311	0.4478	1.4896	25.4515	0.6143

Table 6. Algorithm comparison for variable **TG** considering three types of missing data mechanisms (MAR, MCAR, MNAR) and different levels of missingness (10, 20, 30 and 50%) based on imputation accuracy (root mean squared error, mean absolute percentage error and classification accuracy)

TG	RMSE	MAPE	CA	RMSE	MAPE	CA
MCAR						
	10%			20%		
NN	**1.6781**	45.8713	0.6947	**1.7419**	47.0029	0.6794
PMM	2.4287	63.7588	0.6149	2.4450	63.1532	0.6062
SLR	2.4359	103.3195	0.5578	2.4596	102.7914	0.5577
RF	2.0806	**44.0431**	**0.7074**	2.0547	**44.7400**	**0.7023**
MEAN	1.7646	57.4358	0.5290	1.8152	57.5840	0.5279
	30%			50%		
NN	**1.7228**	47.7940	0.6762	**1.7834**	**48.2514**	**0.6755**
PMM	2.3925	63.0147	0.6056	2.4565	64.0044	0.6048
SLR	2.4405	104.0732	0.5564	2.4431	103.1582	0.5599
RF	2.1222	**47.7587**	**0.6877**	2.1983	51.0667	0.6650
MEAN	1.7804	57.7058	0.5268	1.8126	57.6760	0.5264
MAR						
	10%			20%		
NN	**1.7303**	49.2283	0.6911	**1.8401**	**49.4026**	**0.6982**
PMM	2.3680	65.6342	0.6064	2.4370	65.0048	0.6053
SLR	2.4674	109.4875	0.5521	2.5070	109.1327	0.5512
RF	2.0619	**46.5159**	**0.7054**	2.2181	49.7677	0.6896
MEAN	1.8133	65.2662	0.4868	1.8857	65.7658	0.4868
	30%			50%		
NN	**1.8198**	**51.1575**	**0.6783**	**1.8659**	**50.0006**	**0.6637**
PMM	2.4042	65.0420	0.6074	2.4305	63.5184	0.6034
SLR	2.4843	108.4079	0.5566	2.4544	102.3056	0.5595
RF	2.2407	53.1501	0.6720	2.2509	52.7012	0.6629
MEAN	1.8649	66.6036	0.4857	1.8821	61.2814	0.5108
MNAR						
	10%			20%		
NN	**0.8652**	68.9414	0.5467	**1.0042**	80.4539	0.4532
PMM	1.7352	82.8811	0.5752	1.7816	88.5791	0.5513
SLR	1.9790	145.7663	0.4997	2.0904	153.9056	0.4926
RF	1.3241	**59.3654**	**0.6693**	1.5014	**67.5492**	**0.6345**
MEAN	0.9872	95.4777	0.1981	1.0842	104.9737	0.1980
	30%			50%		
NN	**1.1993**	96.6311	0.3699	1.7054	74.3427	0.4908
PMM	1.9182	98.8680	0.5078	2.3500	81.8861	0.5489
SLR	2.2228	164.3346	0.4758	2.4714	131.9337	0.5173
RF	1.6640	**79.4840**	**0.5810**	2.1402	**68.1128**	**0.6036**
MEAN	1.2115	117.6751	0.1988	**1.6757**	87.7802	0.3676

Table 7. Algorithm comparison for variable **GLY** considering three types of missing data mechanisms (MAR, MCAR, MNAR) and different levels of missingness (10, 20, 30 and 50%) based on imputation accuracy (root mean squared error, mean absolute percentage error and classification accuracy)

GLY	RMSE	MAPE	CA	RMSE	MAPE	CA
MCAR						
	10%			20%		
NN	**1.2561**	**16.1328**	0.8387	**1.2590**	**16.2744**	0.8405
PMM	1.7727	23.3703	0.7566	1.7698	23.4156	0.7550
SLR	1.7831	26.4223	0.6979	1.7701	26.3160	0.7001
RF	1.4650	16.5648	0.8279	1.4737	17.1581	0.8244
MEAN	1.3308	17.3669	**0.8480**	1.3223	17.3917	**0.8499**
	30%			50%		
NN	**1.2636**	**16.2883**	0.8384	**1.2759**	**16.4178**	0.8375
PMM	1.7739	23.4765	0.7552	1.7703	23.4657	0.7549
SLR	1.7712	26.3031	0.7000	1.7812	26.4517	0.6970
RF	1.4894	17.5696	0.8186	1.5417	18.7614	0.8089
MEAN	1.3230	17.3702	**0.8474**	1.3253	17.4029	**0.8484**
MAR						
	10%			20%		
NN	**1.1772**	**16.4996**	0.8693	**1.1963**	**16.7634**	0.8712
PMM	1.6304	23.0727	0.8011	1.6436	23.0982	0.8001
SLR	1.7147	26.8912	0.7427	1.7331	27.2023	0.7401
RF	1.3543	17.0068	0.8581	1.3818	17.7424	0.8502
MEAN	1.2193	18.5149	**0.8802**	1.2283	18.8077	**0.8822**
	30%			50%		
NN	**1.2200**	**17.2983**	0.8735	**1.2587**	**17.0065**	0.8539
PMM	1.6478	23.1569	0.8019	1.7262	23.3995	0.7719
SLR	1.7483	27.4624	0.7383	1.7811	27.0636	0.7138
RF	1.4170	18.6302	0.8426	1.5060	19.1082	0.8203
MEAN	1.2356	19.1823	**0.8826**	1.2910	18.3311	**0.8644**
MNAR						
	10%			20%		
NN	**1.0146**	20.9776	0.9733	**1.0910**	22.8274	0.9642
PMM	1.5630	26.4797	0.8574	1.6378	27.9884	0.8395
SLR	1.6236	31.2847	0.7685	1.6987	32.8102	0.7441
RF	1.1721	**18.8606**	0.9312	1.2621	**20.9540**	0.9162
MEAN	1.0738	23.5457	**1.0000**	1.1629	25.7221	**1.0000**
	30%			50%		
NN	**1.2012**	25.6262	0.9562	**1.2858**	**21.3637**	0.8708
PMM	1.7052	29.6050	0.8210	1.7721	26.6352	0.7625
SLR	1.7859	34.6600	0.7103	1.8149	30.4577	0.6779
RF	1.3510	**23.2115**	0.8989	1.4924	21.4904	0.8303
MEAN	1.2823	28.7493	**1.0000**	1.3336	23.2268	**0.8966**

Performances from tables 3, 4, 5, 6 and 7 are summarized in Table 8. The table shows the number of cases where the selected method has the best performance values. Only NN, MEAN and RF methods are shown as competing algorithms with best results overall. One case represents combination of: performance measure, variable, missingness mechanism and volume of missing data. Since there are: 3 metrics, 5 variables, 3 mechanisms and 4 percentages, there are 180 cases in total. For MEAN and RF there are also, in parentheses, values which indicate how many cases have NN as next best performance. In addition to the total winning scores, also given in the table are scores grouped by different context: by variables that are missing, by missingness mechanism and by percentage of missing data.

Table 8. Overview of the number of different tested cases where selected algorithm shows best performance, in total and grouped by: variable, mechanism and volume of missing data. Number of cases where NN has next best performance is shown in parentheses

Method	**Total number of winning cases**				
NN	133				
MEAN	21 (18)				
RF	26 (20)				
by variable	**HDL**	**LDL**	**TCH**	**TG**	**GLY**
NN	24	33	36	19	21
MEAN	8 (5)	0	0	1 (1)	12 (12)
RF	4 (1)	3 (3)	0	16 (13)	3 (3)
by mechanism	**MCAR**	**MAR**	**MNAR**		
NN	40	54	39		
MEAN	12 (9)	4 (4)	5 (5)		
RF	8 (6)	2 (2)	16 (12)		
by volume	**10%**	**20%**	**30%**	**50%**	
NN	31	33	33	36	
MEAN	4 (4)	4 (4)	6 (4)	7 (6)	
RF	10 (9)	8 (5)	6 (5)	2 (1)	

3.1. Discussion

Observing all obtained performance values, ANNs prevail as the best method for imputation, considering different missing mechanisms and proportion of missing data.

For HDL, MEAN imputation shows the best results for all missing frequencies but only for MCAR data. Also, all those MEAN results are closely followed with ANN and RF algorithm. As missing mechanism changes to mechanisms that enclose dependency in missing data, ANNs emerge as a method with the best performance.

For LDL, there are only three cases for CA metric where RF shows better results than ANN and only in case where data is missing according to MNAR. Even in those cases, ANNs are closely behind.

For TCH, ANNs display the best performance results, considering all three measures across all missing pattern cases.

For TG, mixed results can be observed. MAPE errors are very large for each algorithm in all scenarios, which gives fairly inaccurate imputation overall. In MCAR and MAR case, ANN and RF are competing with similar imputation accuracy (both MAPE and CA) while the other methods perform notably worse. In an MNAR setting, the disparity between performance of ANN and RF results is bigger. Still, ANNs demonstrate the smallest RMSE errors through all settings.

For GLY, it can be observed that MEAN shows best results according to classification, especially in the MNAR case where it shows 100% accuracy. From data distribution (Figure 1), it can be observed that GLY has CMR cutoff value higher than the upper quartile with lots of outliers. MNAR mechanism is simulated by removing lower GLY values with higher probability. When the volume of missing data is smaller (10, 20, 30%) mean of sample set with complete cases stays lower than the cutoff value, therefore imputation gives values that are, as original data, lower than the cutoff, which explains very high classification accuracy. Similarly as for previous variables, ANNs have the smallest RMSE error through all simulation settings.

Regarding the overview of performance given in the table 8, it is noticeable that ANNs are winning in most scenarios. Even for cases where other method is the winner, for a large number of them, ANN is the next best imputation method.

Here should be noted that the obtained performance results (RMSE, MAPE, CA), considering the distribution of values for each variable (Table 1, Figure 1), except maybe for TCH, indicate that neither explored imputation method should be used as a final prediction model for variables separately. Nevertheless, in this research, we are not interested in prediction models for those values as separate models, but rather in imputation in preprocessing phase of building CMR models. That is why we are solely interested in relative, comparison values. If a goal of some future research would be to build prediction models for HDL, LDL,TCH, TG and GLY by and of itself, one can use this research as a supporting ground and explore other sets of predictors, as well as other models for each variable independently.

Limitations and Further Research What should be examined is how ANN and RF with different architectures compare solely, especially for TG, since RFs exhibit good performance for some scenarios. In that case, fine tuning of RFs hyperparameters should be performed since RFs with fixed number of trees is used in this research. Also, although this work provides promising results, it should be explored how distribution around CMR cutoff values is correlated with imputation accuracy, especially for GLY, and should be determined if performance is a result of this specific clinical dataset. Additionally, the used MAR and MNAR settings demonstrate just some examples of these mechanisms. Simulations with different, more sophisticated MAR and MNAR mechanisms could be performed and ascertain if the results are agreeable. Lastly, it could be tested whether introducing additional hidden layers or tuning process and parameters of ANNs could further improve accuracy and performance of neural networks as an imputation method. At the end, final networks could be ensembled in order to broaden the proposed methodology for multivariate imputation, along with exploring how those imputations affect final results in development of new and enhanced CMR prediction models.

4. Conclusions

Prognostic CMR models require simple anthropometric measures as well as some laboratory values. In order to build machine learning models that solely use small, low-cost set of predictors - laboratory values are necessary, but only in the preprocessing phase when outcome CMR values are produced. If some of those laboratory values are missing, a dataset used for learning could be fairly reduced and even produce flawed end results, which can then make process of building final CMR model more difficult. In statistics, missingness is often analyzed separately. In engineering, during machine learning research, this step is sometimes skipped or overlooked. Therefore, we have explored how one machine learning model (ANN) behaves as an imputation method in CMR risk assessment to enable fully independent algorithmic building process for CMR model, diminishing the need for separate analysis of missingness mechanisms.

To explore how neural networks perform as an imputation method for laboratory data used for calculating outcome CMR values which could further be used in building predictive model for CMR, we have explored a number of ANN structures and compared their imputation performance with other simple single imputation methods. First, we have built and tested single layered neural networks with different numbers of hidden neurons to propose optimal settings for univariate imputation of missing values for each variable. Those settings have afterwards been used for comparison of ANNs with other imputation methods. We have simulated three missingness mechanisms (MCAR, MAR and MNAR) and performed simulations for different volume of missing data (10, 20, 30 and 50%). Through all scenarios, ANNs showed strong performance according to different measures of imputation accuracy. They outperformed or were closely behind other methods in almost all the cases, considering both proportion of missing data and missingness mechanism.

Considering the results, we propose that an ANN should be considered and used in imputation of laboratory values, in preprocessing phase, as a step in pipeline framework which could lead to development of more robust and precise CMR prediction models.

Although this work calls for next steps in the future research such as: ensembling obtained networks or development of new types of ANNs which will deal with multivariate missing data and analysis of impact of those imputations in the final model development, this step was necessary in order to explore versatile settings of missing data, systemize results and conclusions and prepare the basis for final CMR assessment.

It is worth noting that this research does not exclude other ML algorithms as potential imputation tools. On the contrary, we also propose that ML algorithms, in general, should be considered, researched and used as imputation methods for both predictors and outcomes values, since they could enable automatic integration of imputation in model development process without a need for separate analysis of data distribution and missingness mechanisms, leaving datasets used for learning complete and final results less prone to error due to missingness mechanisms and volume. It is self-evident that any method for dealing with missing data cannot substitute real data, but machine learning could provide us tools for imputation that can be automatic, self-sufficient and domain independent. In that context, the proposed comparison methodology for this specific problem justifies the effort and could be used as a guideline for further research.

Acknowledgments. This work was partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia within the project TR-32044.

References

1. Altman, D.G., Bland, J.M.: Missing data. Bmj 334(7590), 424–424 (2007)
2. Ashwell, M., Gunn, P., Gibson, S.: Waist-to-height ratio is a better screening tool than waist circumference and BMI for adult cardiometabolic risk factors: systematic review and meta-analysis. Obes Res 13(3), 275–286 (Mar 2012)
3. Beaulieu-Jones, B.K., Moore, J.H.: Missing Data Imputation in the Electronic Health Records Using Deeply Learned Autoencoders. Pacific Symposium on Biocomputing 22, 207–218 (2016)
4. Breiman, L.: Statistical modeling: The two cultures. Statistical science 16(3), 199–215 (2001)
5. Domingos, P.M.: A few useful things to know about machine learning. Commun. acm 55(10), 78–87 (2012)
6. Efron, B.: Estimating the error rate of a prediction rule: improvement on cross-validation. Journal of the American statistical association 78(382), 316–331 (1983)
7. Expert Panel on Detection Evaluation and Treatment of High Blood Cholesterol in Adults: Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) (Adult Treatment Panel III). JAMA: The Journal of the American Medical Association 285, 2486–2497 (2001)
8. Friedewald, W.T., Levy, R.I., Fredrickson, D.S.: Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. Clinical chemistry 18(6), 499–502 (1972)
9. Fritsch, S., Guenther, F.: neuralnet: Training of Neural Networks (2016), https://CRAN.R-project.org/package=neuralnet, r package version 1.33
10. Graham, J.W., Cumsille, P.E., Elek-Fisk, E.: Methods for Handling Missing Data. In: Handbook of Psychology, pp. 87–114. John Wiley & Sons, Inc. (2003)
11. Greenland, S., Finkle, W.D.: A critical look at methods for handling missing covariates in epidemiologic regression analyses. American journal of epidemiology 142(12), 1255–1264 (1995)
12. Groenwold, R.H., White, I.R., Donders, A.R.T., Carpenter, J.R., Altman, D.G., Moons, K.G.: Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. CMAJ 184(11), 1265–1269 (2012)
13. Grundy, S.M., Cleeman, J.I., Daniels, S.R., Donato, K.A., Eckel, R.H., Franklin, B.A., Gordon, D.J., Krauss, R.M., Savage, P.J., Smith, S.C., Spertus, J.A., Costa, F.: Diagnosis and Management of the Metabolic Syndrome. Circulation 112, 2735–2752 (2005)
14. Hughes, R.A., Heron, J., Sterne, J.A., Tilling, K.: Accounting for missing data in statistical analyses: multiple imputation is not always the answer. International journal of epidemiology 1, 11 (2019)
15. Igel, C., Toussaint, M., Weishui, W.: Rprop using the natural gradient. In: Trends and applications in constructive approximation, pp. 259–272. Springer (2005)
16. Kannel, W.B., McGee, D., Gordon, T.: A general cardiovascular risk profile: The framingham study. The American Journal of Cardiology 38(1), 46 – 51 (1976)
17. Kuhn, M.: Building predictive models in r using the caret package. Journal of Statistical Software, Articles 28(5), 1–26 (2008)
18. Kupusinac, A., Stokić, E., Srdić, B.: Determination of WHtR Limit for Predicting Hyperglycemia in Obese Persons by Using Artificial Neural Networks. TEM J 1(4), 270–272 (2012)
19. Kupusinac, A., Doroslovački, R., Malbaški, D., Srdić, B., Stokić, E.E.: A primary estimation of the cardiometabolic risk by using artificial neural networks. Computers in Biology and Medicine 43(6), 751–757 (2013)
20. Laakso, M.: Hyperglycemia and cardiovascular disease in type 2 diabetes. Diabetes 48(5), 937–942 (1999)
21. Leke, C., Marwala, T.: Missing data estimation in high-dimensional datasets: A swarm intelligence-deep neural network approach. In: Advances in Swarm Intelligence. pp. 259–270. Springer International Publishing (2016)

22. Leke, C., Marwala, T., Paul, S.: Proposition of a Theoretical Model for Missing Data Imputation using Deep Learning and Evolutionary Algorithms. arXiv (2015)
23. Little, R.J., D'Agostino, R., Cohen, M.L., Dickersin, K., Emerson, S.S., Farrar, J.T., Frangakis, C., Hogan, J.W., Molenberghs, G., Murphy, S.A., Neaton, J.D., Rotnitzky, A., Scharfstein, D., Shih, W.J., Siegel, J.P., Stern, H.: The Prevention and Treatment of Missing Data in Clinical Trials. New England Journal of Medicine 367(14), 1355–1360 (oct 2012)
24. Little, R.J.A., Rubin, D.B.: Statistical analysis with missing data. Wiley (2019)
25. Marshall, A., Altman, D.G., Royston, P., Holder, R.L.: Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. BMC medical research methodology 10(1), 7 (2010)
26. Masconi, K.L., Matsha, T.E., Echouffo-Tcheugui, J.B., Erasmus, R.T., Kengne, A.P.: Reporting and handling of missing data in predictive research for prevalent undiagnosed type 2 diabetes mellitus: a systematic review. The EPMA Journal (2015)
27. Organization, W.H.: Obesity: preventing and managing the global epidemic. report of a world health organization consultation. Tech. Rep. 894, WHO Obesity Technical Report Series (2000)
28. Papageorgiou, G., Grant, S.W., Takkenberg, J.J.M., Mokhles, M.M.: Statistical primer: how to deal with missing data in scientific research? Interactive CardioVascular and Thoracic Surgery 27(2), 153–158 (05 2018)
29. Pesonen, E., Eskelinen, M., Juhola, M.: Treatment of missing data values in a neural network based decision support system for acute abdominal pain. Artificial Intelligence in Medicine 13(3), 139 – 146 (1998)
30. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2018), https://www.R-project.org/
31. Riedmiller, M.: Advanced supervised learning in multi-layer perceptrons — from backpropagation to adaptive learning algorithms. Computer Standards and Interfaces 16(3), 265 – 278 (1994)
32. Rosolova, H., Nussbaumerova, B.: Cardio-metabolic risk prediction should be superior to cardiovascular risk assessment in primary prevention of cardiovascular diseases. The EPMA journal 2, 15–26 (2011)
33. Salgado, C.M., Azevedo, C., Proença, H., Vieira, S.M.: Missing Data, pp. 143–162. Springer International Publishing, Cham (2016)
34. Schafer, J.L., Graham, J.W.: Missing data: our view of the state of the art. Psychological methods 7(2), 147–77 (2002)
35. Schouten, R.M., Lugtig, P., Vink, G.: Generating missing values for simulation purposes: a multivariate amputation procedure. Journal of Statistical Computation and Simulation (2018)
36. Silva-Ramírez, E.L., Pino-Mejías, R., López-Coello, M., de-la Vega, M.D.C.: Missing value imputation on missing completely at random data using multilayer perceptrons. Neural Networks 24(1), 121 – 129 (2011)
37. Stokić, E., Tomić-Naglić, D., Derić, M., Jorga, J.: Therapeutic options for treatment of cardiometabolic risk. Medicinski pregled 62 Suppl 3, 54–8 (2009)
38. Van Buuren, S.: Flexible imputation of missing data, vol. 20125245. Chapman and Hall/CRC (2012)
39. Van Buuren, S., Groothuis-Oudshoorn, K.: mice: Multivariate imputation by chained equations in r. Journal of Statistical Software 45(3), 1–67 (2011)
40. Waljee, A.K., Higgins, P.D., Singal, A.G.: A primer on predictive models. Clinical and translational gastroenterology 5(1), e44 (2014)
41. Ware, J.H., Harrington, D., Hunter, D.J., D'Agostino, R.B.: Missing data. New England Journal of Medicine 367(14), 1353–1354 (2012)

Dunja Vrbaški is a PhD student at the Faculty of Technical Sciences at the University of Novi Sad, Serbia. She received her BSc and MSc degrees in Computer Science from

the Faculty of Sciences at the University of Novi Sad in 2003 and 2013, respectively. Prior to her enrolment to doctoral studies she has been working as a software developer and project manager in the IT industry. Her research interests are algorithms, machine learning and data visualization.

Aleksandar Kupusinac is an Associate Professor at the Department of Computing and Control Engineering at the Faculty of Technical Sciences, University of Novi Sad, Serbia. He received the Dipl. Ing. degree in Electrical Engineering in 2005, and the MSc and PhD degrees in Computer Science, in 2008 and 2010 from the Faculty of Technical Science, University of Novi Sad. His research interests include data science and artificial intelligence.

Rade Doroslovački is a Professor of Mathematics at the Faculty of Technical Sciences, University of Novi Sad. He received his BSc, MSc and PhD in Mathematics from the Faculty of Sciences at the University of Novi Sad in 1976, 1984 and 1989, respectively. His research interests include combinatorics, graph theory and discrete mathematics. He is currently serving as the Dean of the Faculty of Technical Sciences, University of Novi Sad.

Edita Stokić holds a PhD in Internal Medicine and is a full professor at University of Novi Sad, Faculty of Medicine. She is also employed in the Clinic of Endocrinology, Diabetes and Metabolic Disorders of the Clinical Centre of Vojvodina.

Dragan Ivetić was born in Subotica and received the Dipl. Ing. degree in Electrical Engineering in 1990, and the MSc and PhD degrees in computer science, in 1994 and 1999 respectively, from the Faculty of Technical Science, University of Novi Sad. Since 2000 he has been Professor of Computer Science at the Department for Computing and Automatics, Faculty of Technical Sciences, University of Novi Sad. His research interests include HCI, medical informatics, and computer graphics. Professor Ivetić received DAAD scholarship in 1997 and ACM scholarship in 1998.

Received: July 10, 2019; Accepted: January 20, 2020.

5. Appendix

Graphical representations of algorithm comparisons for different percent and mechanism of missingness are given on the following figures.

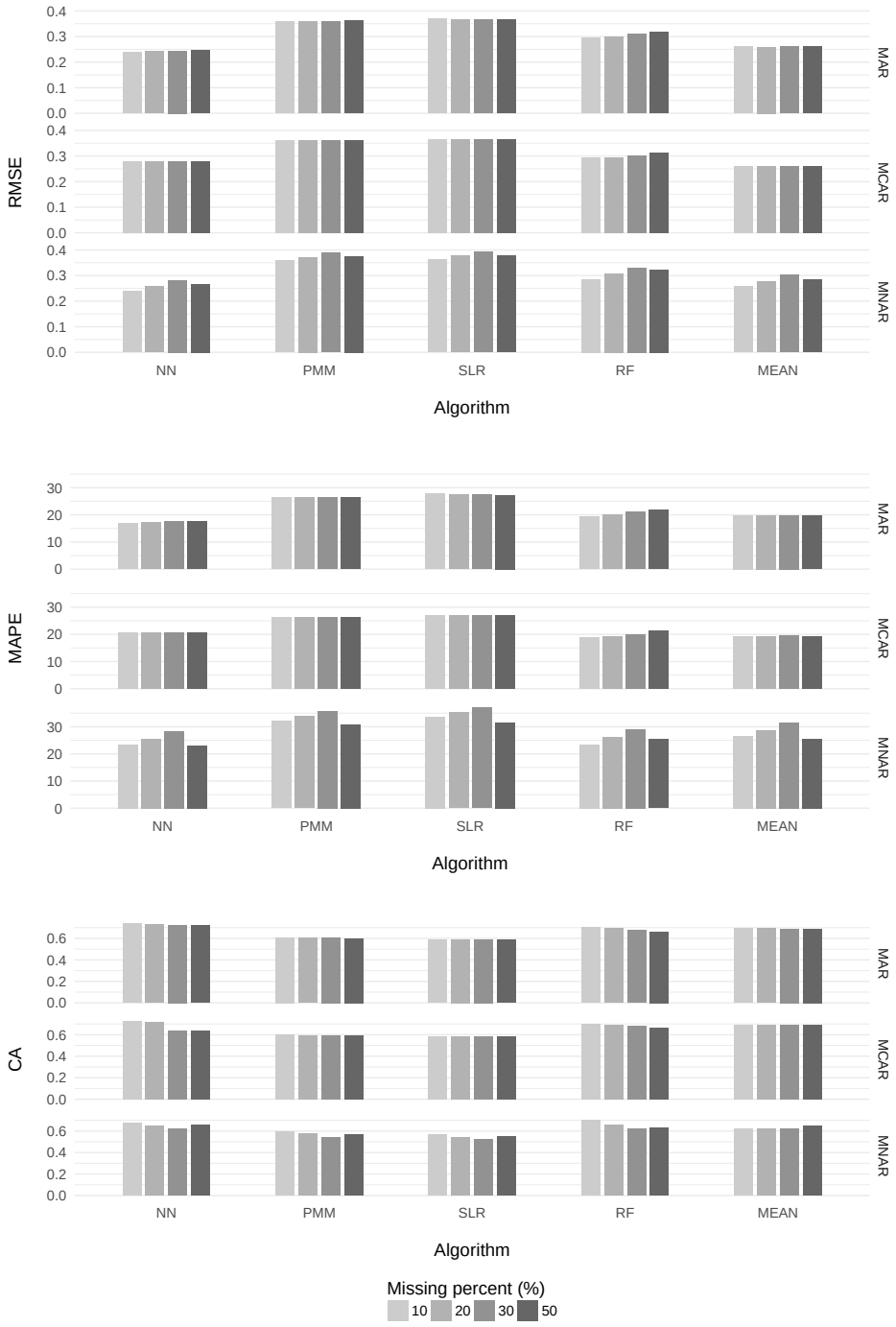

Fig. 2. Results of algorithms comparison for variable **HDL**. Results are grouped by accuracy performance (RMSE, MAPE, CA) shown on the right and missing data mechanisms (MAR, MCAR, MNAR) shown on the left. Every volume of missingness (10, 20, 30 and 50%) is displayed by separate color

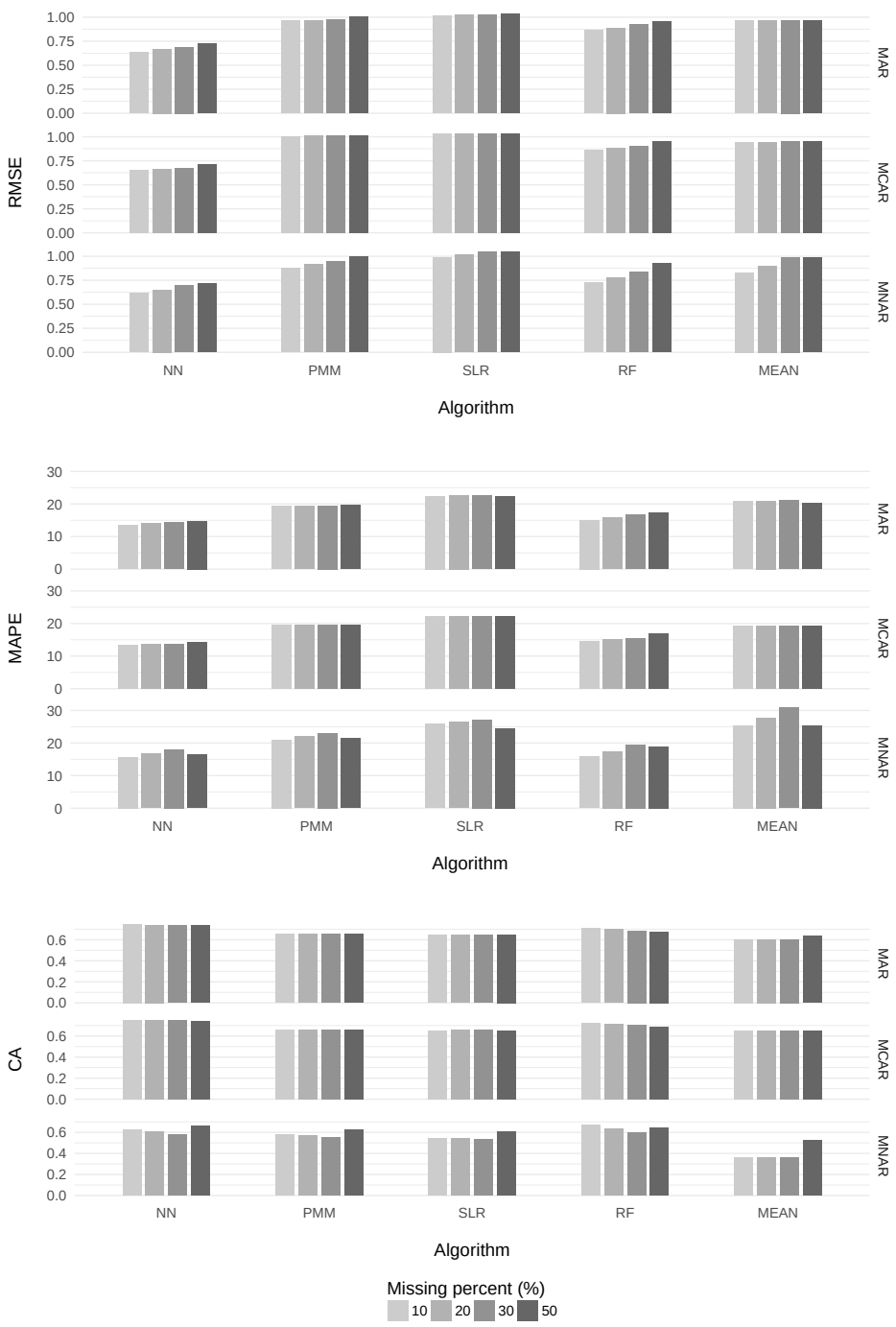

Fig. 3. Results of algorithms comparison for variable **LDL**. Results are grouped by accuracy performance (RMSE, MAPE, CA) shown on the right and missing data mechanisms (MAR, MCAR, MNAR) shown on the left. Every volume of missingness (10, 20, 30 and 50%) is displayed by separate color

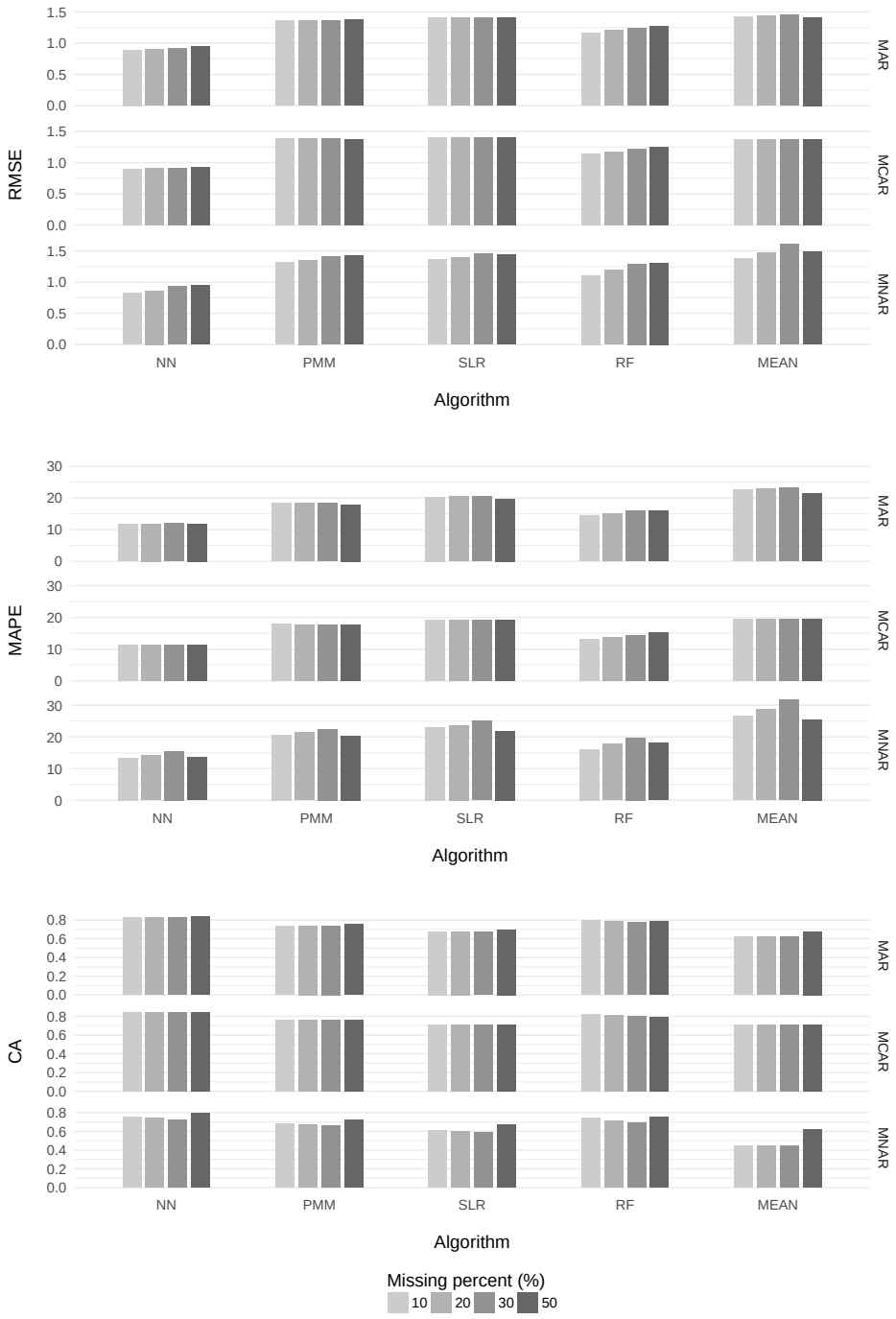

Fig. 4. Results of algorithms comparison for variable **TCH**. Results are grouped by accuracy performance (RMSE, MAPE, CA) shown on the right and missing data mechanisms (MAR, MCAR, MNAR) shown on the left. Every volume of missingness (10, 20, 30 and 50%) is displayed by separate color

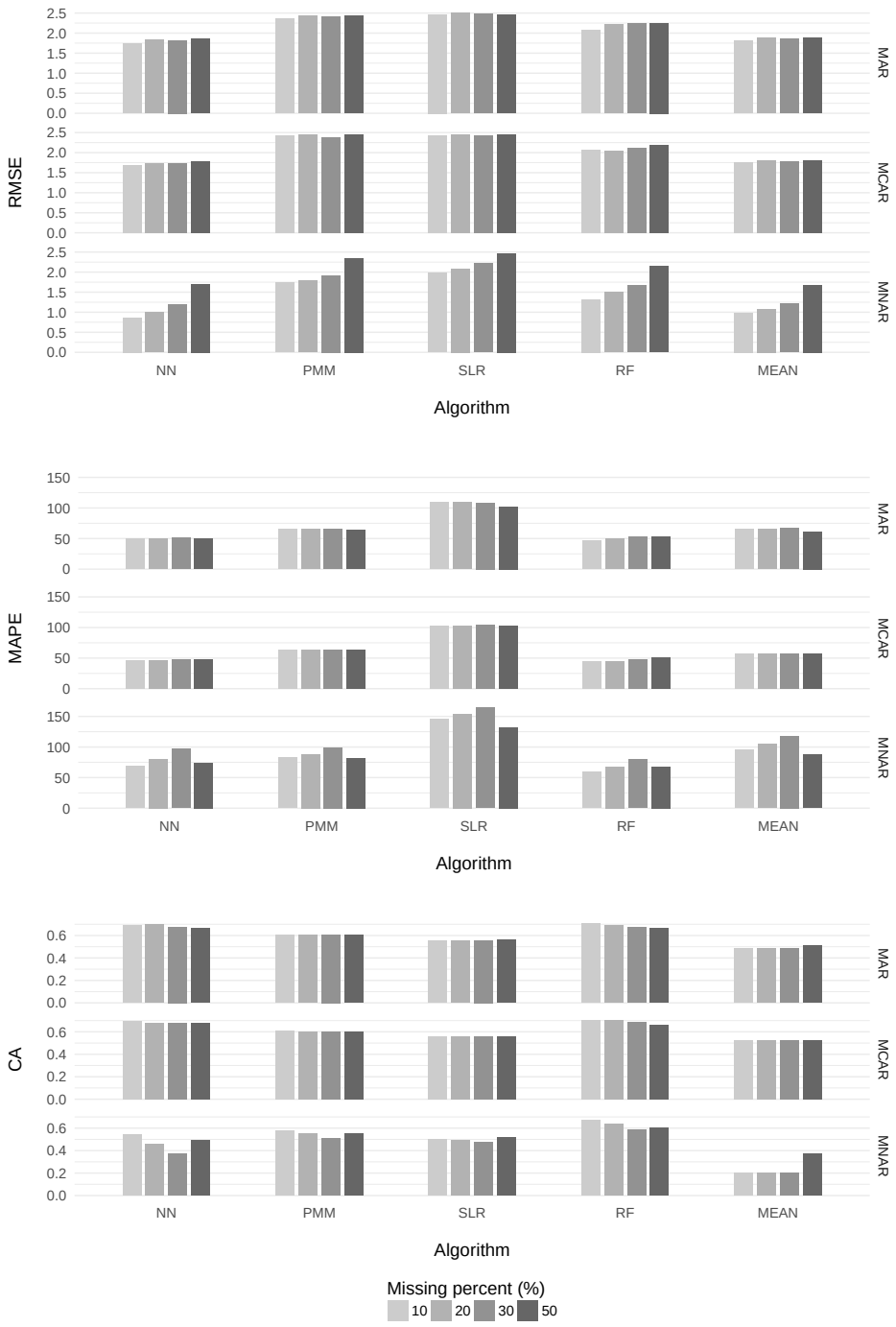

Fig. 5. Results of algorithms comparison for variable **TG**. Results are grouped by accuracy performance (RMSE, MAPE, CA) shown on the right and missing data mechanisms (MAR, MCAR, MNAR) shown on the left. Every volume of missingness (10, 20, 30 and 50%) is displayed by separate color

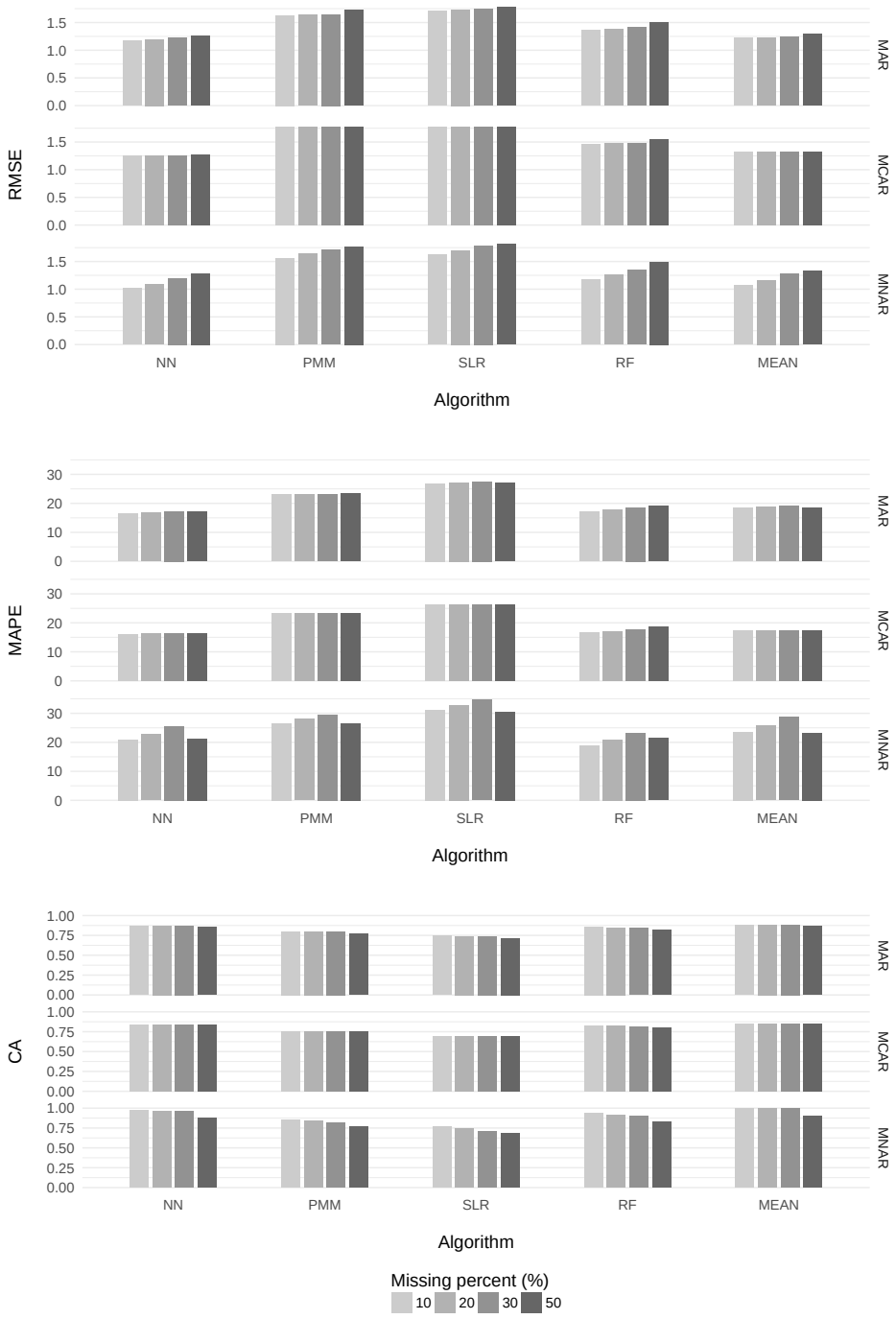

Fig. 6. Results of algorithms comparison for variable **GLY**. Results are grouped by accuracy performance (RMSE, MAPE, CA) shown on the right and missing data mechanisms (MAR, MCAR, MNAR) shown on the left. Every volume of missingness (10, 20, 30 and 50%) is displayed by separate color