

Sets with Incomplete and Missing Data – NN Radar Signal Classification

Ivan Jordanov
School of Computing,
University of Portsmouth,
Portsmouth, UK
Ivan.Jordanov@port.ac.uk

Nedyalko Petrov
School of Computing,
University of Portsmouth,
Portsmouth, UK
Nedyalko.Petrov@port.ac.uk

Abstract— We investigate further the problem of radar signal classification and source identification with neural networks. The available large dataset includes pulse train characteristics such as signal frequencies, type of modulation, pulse repetition intervals, scanning type, scan period, etc., represented as a mixture of continuous, discrete and categorical data. Typically, considerable part of the data samples is with missing values. In our previous work we used only part of the radar dataset, applying listwise deletion to get rid of the samples with missing values and processed relatively small subset of complete data. In this work we apply multiple imputation (MI) method, which is a model based approach of dealing with missing data, by producing confidence intervals for unbiased estimates without loss of statistical power (using both complete and incomplete cases). We employ MI to all data samples with up to 60% missingness, this way increasing more than twice the size of the initially used data subset. We apply feedforward backpropagation neural network (NN) supervised learning for solving the classification and identification problem and investigate and critically compare the same three case studies, researched in the previous paper and report improved, superior results, which is a consequence of the implemented MI and improved NN training.

Keywords— missing data; multiple imputation; NN learning; radar signal classification.

I. INTRODUCTION

Statistical analysis and research classify the nature of missing data in three groups [1-3]: missing completely at random (MCAR), where the probability that an observation is missing is unrelated to its value or to the value of any other variables; missing at random (MAR) – that missingness does not depend on the value of the observed variable, but on the extent that the missingness is correlated with other variables that are included in the analysis (in other words, the cause of missingness is considered); and missing not at random (MNAR) – when the data are not MCAR or MAR (missingness still depends on unobserved data). The problem associated with MNAR is that it yields biased parameter estimates while MCAR and MAR analyses yield unbiased ones (at the same time the main MCAR consequence is loss of statistical power).

That's why dealing with missingness requires analysis strategy that leads to least biased estimates while not loosing

statistical power. The problem is that those are contradictory criteria and in order to use the information from the partial data in missing data samples (keeping the statistical power), and substituting the missing data samples with estimates, inevitably brings biases.

The most popular approaches in dealing with missing data generally fall in three groups: Deletion methods; Single imputation methods; and Model-based methods [1, 4, 5].

Deletion methods include pairwise and listwise deletion. The pairwise deletion (also called "unwise" deletion) keeps as many cases as possible for each analysis (this way uses all available information for each of it), but the problem is that the analyses are incomparable because each of them is based on different subsets of data, with different sample sizes and different standard errors. The listwise deletion (also known as complete case analysis) is a simple approach in which all cases with missing data are omitted. The advantages of this approach include comparability across the analyses and it leads to unbiased parameter estimates (assuming the data is MCAR), while the disadvantage is that there may be substantial loss of statistical power (because not all information is used in the analysis, especially if a large number of cases is excluded).

The single imputation methods include mean/mode substitution, dummy variable method, and single regression. Mean/mode substitution is an old procedure, currently rejected because of its intrinsic problems, e. g., it adds no new information (the overall mean stays the same), reduces the variability, and weakens the covariance and correlation estimates (it ignores relationship between variables). The dummy variable technique uses all available information about missing observation, but produces biased estimates. In the regression approach, the linear regression is used to predict what the missing value should be (on the basis of the available other variables) and then implement it as an actual value. The advantage of this technique is that it uses information from the observed data, but it overestimates the model fit and the correlation estimates, and weakens the variance.

Most popular, "modern" model-based approaches (that require MAR data), fall into two categories: multiple

imputation (MI) methods and maximum likelihood (ML) methods (often referred to as full-information maximum likelihood) [3]. Their advantage is that they model the missingness and give confidence intervals for estimates, rather than relying on a single imputation. If the assumption for MAR missingness holds, both groups of methods result in unbiased estimates (i.e., tend to "preserve" means, variances, covariances, correlations and linear regression coefficients) without loss of statistical power.

ML identifies a set of parameter values that produces the highest (log) likelihood and estimates the most likely value that would result in the observed data. It has the advantage that both complete and incomplete cases are used, in other words, it makes use of the full information and produces unbiased parameter estimates (with MCAR/MAR data). The disadvantage of this approach is that statistical estimates can be biased downward (but this can be adjusted by using observed information matrix).

We will consider the MI estimation approach in more detail, as it is used in this paper for pre-processing the datasets used for solving the radar signal identification and classification problem with neural network learning.

The MI approach involves three distinct steps: first, sets of plausible data for the missing observations are created and these sets are filled in separately to create many 'completed' datasets; second, each of these datasets is analysed using standard procedures for complete datasets; and thirdly, the results from previous step are combined and pooled into one estimate for the inference. The aim of the MI process is not just to fill in the missing values with plausible estimates, but to plug in multiple times these values by preserving important characteristics of the whole dataset. As with most (multiple regression) prediction models, the danger of overfitting the data is real and this can lead to less generalizable results than the original data would have [6].

The advantage of MI is that it provides more accurate variability with multiple imputations for each missing value (and considers variability due to sampling and variability due to imputation) and the disadvantage that it depends on the correctly specified model and also includes cumbersome coding, but this has been greatly simplified by the existence of easy to use off-shelf software packages. For the purpose of this investigation we used R open source statistical software, which is free to download from internet.

II. RADAR SIGNAL CLASSIFICATION

Radars are considered "active" sensors, as they use their own transmitter for locating targets. Although initially developed for military surveillance, navigation and weapon guidance (detecting, locating, tracing, and identifying air, marine, and terrestrial located objects, at small to medium and large distances), nowadays they are widely used for civil purposes as well (traffic control, navigation, weather forecast, pollution control, space observation, and sport systems) [7]. The radar range, resolution and sensitivity are generally

determined by its transmitter and waveform generator. Most radar systems operate in the microwave region of the electromagnetic spectrum with frequency range of about 200 MHz to about 95 GHz, but there are also radars that operate at frequencies as low as 2 MHz and as high as 300 GHz. The lower frequency bands are usually preferred for longer range surveillance, whereas the higher frequencies tend to be used for shorter range applications with higher resolution [7].

Radar detection, classification and tracking of targets against a background of clutter and interference are considered as 'general radar problem'. For military purposes 'the general radar problem' includes searching for, interception, localisation, analysis and identification of radiated electromagnetic energy which is commonly known as radar Electronic Support Measures (ESM). They are considered reliable source of valuable information regarding threat detection, threat avoidance, and in general, situation awareness for timely deployment of counter-measures [8].

A real-time identification of the radar emitter associated with each intercepted pulse train is a very important function of the radar ESM. Typical approaches include sorting incoming radar pulses into individual pulse trains, then comparing their characteristics with a library of parametric descriptions, in order to get list of likely radar types. This can be very difficult task as there may be radar modes for which there is no record in ESM library; overlaps of different radar type parameters; increases in environment density (e.g., Doppler spectrum radars transmitting hundreds of thousands of pulses per second); agility of radar features such as radio frequency and scan, pulse repetition interval etc.; multiplication and dispersion of the modes for military radars; noise and propagation distortion that lead to incomplete or erroneous signals [9].

Intercepted and collected pulse train characteristics typically include signal frequencies, type of modulation, pulse repetition intervals, etc., and usually consist of mixture of continuous, discrete and categorical data, and also frequently include missing values. Missing values are imminent part of real world datasets and radar datasets make no exception. There is a variety of reasons why data may be missing, but common ones are related to human factor, equipment malfunction, or coarse environment conditions that result in noise and propagation distortion leading to incomplete, erroneous, or not intercepted signals. Table 1 shows several samples of radar data that consist of continuous, discrete, and categorical data and also includes missing values.

III. NEURAL NETWORKS IN RADAR IDENTIFICATION

Various approaches and methods have been investigated for radar emitter recognition and identification, where considerable part of the research in the area incorporates Neural Networks (NN), because of their massively parallel architecture, fault tolerance and ability to handle incomplete radar type descriptions and inconsistent and noisy data. NN techniques have previously been applied to several aspects of radar ESM processing, including PDW sorting and radar type recognition [10]. More recently, many new radar recognition systems include neural networks as a key classifier [10, 11]. Examples

of a variety of NN architectures and topologies used for radar identification recognition and classification based on ESM data include popular Multilayer Perceptron (MLP), Radial Basis Function (RBF), a vector neural network, single parameter dynamic search neural network, and others [12].

For example, in [9] the authors use initial clustering algorithm to separate pulses from different emitters according to position-specific parameters of the input pulse stream when implementing their “What-and-Where fusion strategy” and then apply fuzzy ARTMAP neural network to classify streams of pulses according to radar type, using their functional parameters. They also do simulations with data set that has missing input pattern components and missing training classes and incorporate a bank of Kalman filters to demonstrate high level performance of their system on incomplete, overlapping and complex radar data. In [10] higher order spectral analysis (HOSA) techniques are used to extract information from LPI (low probability of intercept) radar signals and to produce 2D signatures, which are then fed to a NN for detecting and identifying the LPI radar signal. The work presented in [13] investigates the potential of NN (MLP) when used in Forward Scattering Radar (FSR) applications for target classification. The authors analyze collected radar signal data and extract features, which are then used to train NN for target classification. They also apply K-Nearest Neighbor classifier to compare the results from the two approaches and conclude that the NN one is superior. In [14] an approach combining rough sets (for data reduction) and NN as a classifier is proposed for radar emitter recognition problem, while [15] combines wavelet packets and neural networks for target classification.

In many cases the NN are hybridized with fuzzy systems, clustering algorithms, wavelet packets, Kalman filters, etc., which in turn leads to recognition systems with increased accuracy and improved efficiency [9, 11, 16].

IV. PROBLEM STATEMENT AND DATA SET ANALYSIS

Reliable and real-time identification of radar signals is of crucial importance for timely threat detection, threat avoidance, general situation awareness and timely deployment of countermeasures. In this context, this paper investigates the potential application of a NN-based approach for timely and trustworthy identification of radar types, associated with intercepted pulse trains.

In our previous work [17], we used listwise deletion to obtain 7693 samples with no missing data from a total of 29094 intercepted generic data samples. Each of the captured signals is pre-classified by experts in one of 125 categories, based on the main functions the radar emitter performs (surveillance, air defence, air traffic control, weather tracking, etc.). In this work we use 15656 samples after employing multiple imputation. From the samples with missing data (example given in Table I) we excluded only those with above 60% of missing data. As it can be seen from Table II we used MI for substituting data samples with up to 60% missingness. Each data entry represents a list of 12 recorded pulse train characteristics (signal frequencies, type of modulation, pulse repetition intervals, etc. that will be considered as input parameters). The first column of Table I is data sample

identifier, the second is a category label (specifying the radar function and being considered as system output) and the rest are radar signal characteristics.

TABLE I. SAMPLE RADAR DATA SUBSET. MISSING VALUES (I.E., VALUES THAT COULD NOT HAVE BEEN INTERCEPTED OR RECOGNIZED) ARE DENOTED WITH ‘-’. THE REST OF THE ACRONYMS ARE DEFINED IN TABLE II.

ID	FN	RFC	RFmn	RFmx	PRC	PRImn	PRImx	PDC	PDmn	PDmx	ST	SPmn	SPmx
84	SS	B	5300	5800	K	-	-	S	-	-	A	5.9	6.1
4354	AT	F	2700	2900	F	1351.3	1428.6	S	-	-	A	9.5	10.5
7488	3D	B	8800	9300	K	100	125	S	13	21	B	1.4	1.6
9632	WT	F	137	139	T	-	-	V	-	-	D	-	-
9839	3D	S	2900	3100	J	-	-	V	99	101	A	9.5	10.5

A more comprehensive summary of the data distribution is presented in Table II, where an overview of the type, range and percentage of missing values for the parameters in the data set is given. The data considered consists of both numerical (integer and float) and categorical values, therefore, coding of the categorical fields will be required during the data pre-processing stage, in order to convert them to numerical representations. The table also shows that the percentage of missing data varies from 11% for the *RF* variable to 59% for the *SP* variable.

TABLE II. DATA DESCRIPTION AND PERCENTAGE OF MISSING VALUES. IN COLUMN ‘TYPE’: I – INTEGER; C – CATEGORICAL; R – REAL VALUES.

Field	Field Description	Type	Levels	% Missing
ID	Reference for the line of data	I	-	-
FN	Function performed by the radar (‘3D’ – 3D surveillance, ‘AT’ – airtraffic control, ‘SS’ – surface search, ‘WT’ – weather tracker, etc.)	C	142	1.4
RFC	Type of modulation used by the radar to change the frequency from pulse to pulse (‘A’ – agile, ‘F’ – fixed, etc.)	C	12	20.7
RFmn	Min frequency used by the radar	R	-	11.2
RFmx	Max frequency used by the radar	R	-	11.2
PRC	Type of modulation used by the radar to change the Pulse Repetition Interval (PRI), (‘F’ – fixed, etc.)	C	15	15
PRImn	Min PRI that used by the radar	R	-	46.7
PRImx	Max PRI that used by the radar	R	-	46.7
PDC	Type of modulation used by the radar to change the pulse duration (‘S’ – stable)	C	5	12.9
PDmn	Min pulse duration used by the radar	R	-	46.1
PDmx	Max pulse duration used by the radar	R	-	46.1
ST	Scanning type – used method by the radar to move the antenna beam (‘A’ – circular, ‘B’ – bidirectional, ‘W’ – electronically scanned, etc.)	C	28	11.3
SPmn	Min scan period used by the radar	R	-	59.4
SPmx	Max scan period used by the radar	R	-	59.4

V. DATA IMPUTATION AND PRE-PROCESSING

The pre-processing of the available data is of a great importance for the subsequent machine learning stage and usually can affect significantly the overall success or failure of the application of a given classification algorithm. In this context, the main objective at this stage is to analyse the

available data of inconsistencies, outliers and irrelevant entries and to transform it in a form that could facilitate the underlying mathematical apparatus of the machine learning algorithm and lead to an overall improvement of the classifier’s performance.

A. Data Imputation

For imputing the missing multivariate data we use sequential imputation algorithm [18] implemented in *impSeq* function from **R** package (we also tried two other R functions, *impNorm* and *impSeqRob*, but they didn’t produce better results when tested on complete dataset). If the available data set is denoted with Y and the complete subset with Y_c , the procedure will start with the complete subset to estimate sequentially the missing values of an incomplete observation y^* , by minimizing the covariance of the augmented data matrix $Y^* = [Y_c, x^*]$. Subsequently the data sample x^* is added to the complete data subset and the algorithm continues with the estimate of next data sample with missing values.

Because *impSeq* uses the sample mean and covariance matrix, it is vulnerable to the presence of outliers, but this can be enhanced by including robust estimators of location and scatter (which is realised in *impSeqRob* function). Because the outlygness metric can be computed for a complete dataset only, firstly the sequential imputation of the missing data is done and then the outlygness measure is computed and used to define whether the observation is an outlier or not. If the measure doesn’t exceed a predefined threshold, the observation is included in the next steps of the algorithm (nevertheless, the use of *impSeqRob* in our case didn’t produce better results when tested on complete dataset, which may be simply because of lack of outliers, so we stuck to the *impSeq* function).

As we mentioned before, in the available radar dataset of 29094 samples, there are 7693 fully intercepted and recognised radar signals that constitutes the complete subset (received after listwise deletion of the original dataset). Subsequently, employing the MI on the missing data samples with less than 60% missingness led to dataset of 15656 observations, which more than doubled the size of the initial data subset

TABLE III. SAMPLE RADAR DATA SUBSET WITH IMPUTED VALUES FOR THE MISSING CONTINUOUS VALUES.

ID	FN	RF	RF	RF	PRC	PRI	PRI	PD	PD	PD	ST	SP	SP
		C	mn	mx		mn	mx	C	mn	mx		mn	mx
84	SS	B	5300	5800	K	963.2	5625	S	5.8	17	A	5.9	6.1
4354	AT	F	2700	2900	F	1351	1428	S	4	6.3	A	9.5	10.5
7488	3D	B	8800	9300	K	100	125	S	13	21	B	1.4	1.6
9632	WT	F	137	139	T	622.6	31312	V	61.1	93.1	D	12	47.8
9839	3D	S	2900	3100	J	2058	48128	V	99	101	A	9.5	10.5

For the identification and classification of the radar signals, the applied NN supervised learning uses from two to eleven output classes: in the first set of simulations we use 2 classes – civil and military (defined by experts in the field from a total of 125 functional categories); and in the second set of simulations, four civil and seven military classes, which gives eleven output classes for the NN to classify. Processing the missing subset with MI more than doubled the complete data subset investigated in [17], enabling us to include valuable

information and statistical power of the data contained in the samples with missing values.

Table III shows the samples from Table I with the inputted values produced by the implemented MI.

B. Data Coding and Transformation

This stage of the pre-processing aims to transform the data into a form that is appropriate for feeding to the selected classifier and would facilitate faster and more accurate machine learning. In particular, a transformation known as coding is applied to convert the categorical values presented in the data set to numerical ones. Three of the most broadly applied coding techniques are investigated and evaluated – continuous, binary and introduction of dummy variables.

TABLE IV. SAMPLE SUBSET WITH IMPUTED RADAR DATA AND NATURAL NUMBER CODING OF ‘RFC’, ‘PRC’, ‘PDC’, AND ‘ST’.

ID	RFC	RF	RF	PRC	PRI	PRI	PD	PD	PD	ST	SP	SP
		mn	mx		mn	mx	C	n	mx		mn	mx
84	2	5300	5800	7	963.2	5625	1	5.8	17	1	5.9	6.1
4354	4	2700	2900	4	1351	1428	1	4	6.3	1	9.5	10.5
7488	2	8800	9300	7	100	125	1	13	21	2	1.4	1.6
9632	4	137	139	11	622.6	31312	2	61.1	93.1	4	12	47.8
9839	9	2900	3100	6	2058	48128	2	99	101	1	9.5	10.5

For the first type of coding, each of the categorical values is substituted by a natural number, e.g., the 12 categories for the RFC input are encoded with 12 ordinal numbers, the 15 PRC categories – with 15 ordinal numbers, etc. A sample of data subset coded with continuous values is given in Table IV. Binary coding, wherein each non-numerical value is substituted by $\log_2 N$ (where N is the number of categories taken by that variable) new binary variables (i.e., taking value of either 0 or 1), is illustrated in Table V for 32 categories.

TABLE V. EXAMPLE OF BINARY CODING FOR 32-LEVEL CATEGORICAL VARIABLE.

Original Category		Encoded Variables				
Index	Label	B1	B2	B3	B4	B5
1	‘2D’	0	0	0	0	0
2	‘3D’	0	0	0	0	1
3	‘AA’	0	0	0	1	0
...						
16	‘CS’	0	1	1	1	1
...						
32	‘ME’	1	1	1	1	1

Finally, the non-numerical attributes are coded using dummy variables. In particular, every p levels of a categorical variable are represented by introducing p dummy variables. An example of dummy coding for 32 categorical levels is shown in Table VI.

Taking into account the large number of categories presented for the categorical attributes in the input data set (Table I), continuous and binary codings are considered for transforming the input variables. On the other hand, binary and dummy variable codings are chosen for representing the output parameters.

TABLE VI. EXAMPLE OF DUMMY CODING FOR 32-LEVEL CATEGORICAL VARIABLE.

Original Category		Encoded Variables										
Index	Label	D1	D2	D3	D4	D5	...	D16	...	D32		
1	'2D'	1	0	0	0	0	...	0	...	0	...	
2	'3D'	0	1	0	0	0	...	0	...	0	...	
3	'AA'	0	0	1	0	0	...	0	...	0	...	
...												
16	'CS'	0	0	0	0	0	...	1	...	0	...	
...												
32	'ME'	0	0	0	0	0	...	0	...	1	...	

Finally, in order to balance the impact of the different input parameters on the training algorithm, data scaling is used. Correspondingly, each of the conducted experiments in the next section is evaluated using 3 forms of the input data set: the original data (with no scaling); normalized data (i.e. scaling the attribute values within (0, 1) interval); and standardized data (i.e. scaling the attribute values to a zero mean and unit variance). A sample binary coded and normalised data subset is given in Table VII.

TABLE VII. SAMPLE SUBSET WITH IMPUTED RADAR DATA AND BINARY CODING.

ID	RFC	RF mn	RF mx	PRC	PRI mn	PRI mx	PD C	PD mn	PD mx	ST	SP mn	SP mx
84	0001	5300	5800	0110	963	5625	0	5.8	17	00000	5.9	6.1
4354	0011	2700	2900	0011	1351	1428	0	4	6.3	00000	9.5	10.5
7488	0001	8800	9300	0110	100	125	0	13	21	00001	1.4	1.6
9632	0011	137	139	1010	622	31312	1	61.1	93.1	00011	12	47.8
9839	1000	2900	3100	0101	2058	48128	1	99	101	00000	9.5	10.5

VI. NN TRAINING AND RESULTS

Three broader experiments are conducted for investigating the application of neural network classifiers for solving the radar emitter recognition problem. The investigated neural network topologies include one hidden layer, with fully connected neurons in the adjacent layers and batch-mode training. For a given experiment with P learning samples, the error function is given as:

$$E_p = \frac{1}{2} \sum_{p=1}^P \sum_{i=1}^L (x_i^p - t_i^p)^2, \quad (1)$$

where for each sample $p=1, \dots, P$ and each neuron of the output layer $i=1, \dots, L$, a pair (x_i, t_i) of NN output and target values, respectively, is defined.

For all of the studies, NN learning with *Levenberg-Marquardt* algorithm (we also used *Conjugate Gradient* technique, but it produced inferior results) and tangent sigmoid transfer functions are used (Matlab *nprtool* and *mtool*). A split-sample technique with randomly selected 70% of the available data for training, 15% for validation and 15% for testing is employed and mean squared error (MSE) is used for evaluating the learning performance. The stopping criterion is set to 500 training epochs, or gradient reaching value less than $1.0e-06$, or 6 consequent failed validation checks, whichever occurs first.



Fig. 1. No imputed data classification results for 12-10-2 NN with normalised input data. The values in green specify the correctly classified samples for each class (10 - Civil, 01 - Military).

For the purposes of the first study, the categorical attributes of the input data are coded with consecutive integers. In this way a total of 12 input variables are received (Table IV). Two neural network topologies are examined – 12-10-1 (10 neurons in the hidden layer) and 12-10-2 (two output neurons), where the output parameter is coded as one binary neuron taking values 0 (“Civil”) and 1 (“Military”) for the first topology and 2 binary neurons, taking values 10 (“Civil”) and 01 (“Military”) for the second topology.



Fig. 2. Classification results for imputed data case for 12-10-2 NN with normalised input data. The values in green specify the correctly classified samples for each class (10 - Civil, 01 - Military).

The performance of each of the topologies is investigated, evaluated and compared for training with the original data (no pre-processing), and after normalisation and standardisation. The results are summarised in Table VIII, showing up to 5% accuracy improvement for the case with imputation.

TABLE VIII. NN CLASSIFICATION ACCURACY (TESTING) FOR CONTINUOUS INPUT CODING AND 12-10-N TOPOLOGIES WITH NO DATA SCALING, AFTER NORMALISATION AND AFTER STANDARDISATION.

Topology	Input data	% Accuracy	
		No imputation	With imputation
12-10-1	no scaling	78.1	83.3
	normalised	80.8	84.5
	standardised	80.8	85.2
12-10-2	no scaling	80.1	82.1
	normalised	81.6	83.6
	standardised	82.1	84.5

Sample confusion matrices for 12-10-2 NN classifiers after training (validation step) with normalised input data, for the two cases: no imputed data; and with imputed data are given in Fig. 1 and Fig. 2 respectively. They show the NN performance on the training, validation, and testing subsets, and the overall performance. The network outputs demonstrate high accuracy, as it can be seen by the high number of correct responses in the green squares and the low number of incorrect responses in the red squares. The lower right squares illustrate the overall classifier accuracies. Fig. 2 illustrates improved overall results for the imputed data, and the test confusion matrix which is of course of most interest (NN generalization abilities), shows increased accuracy, especially for the *Military* class.

TABLE IX. NN CLASSIFICATION ACCURACY (TESTING) FOR BINARY INPUT CODING AND 22-22-N TOPOLOGIES WITH NO DATA SCALING, AFTER NORMALISATION AND AFTER STANDARDISATION.

Topology	Input data	% Accuracy	
		No imputation	With imputation
22-22-1	no scaling	81.9	85.6
	normalised	83.3	87.3
	standardised	83.1	87.2
22-22-2	no scaling	81.77	84.8
	normalised	83.90	85.0
	standardised	84.30	86.8

The second case study investigates two additional NN topologies – 22-22-1 and 22-22-2, where the output parameter is again coded by one binary neuron (0 for “Civil” and 1 for “Military”) for the first topology and by two binary neurons for the second one (10 for “Civil” and 01 for “Military”). The NN performance for each of the topologies is investigated, evaluated and compared using the original data, after normalisation and after standardisation for the two cases with and without imputed data. The performance results are summarised in Table IX, showing again improved NN performances for the cases with imputed data.

The final case study investigates a broader output space of 11 classes (4 civil and 7 military) and evaluates a 22-22-11 NN classifier with the original, normalised and standardised training data, with dummy variable coded outputs. Summary of

the obtained results from training on data subsets with and without imputation is presented in Table X.

TABLE X. NN CLASSIFICATION ACCURACY (TESTING) FOR BINARY INPUT CODING AND 22-22-11 TOPOLOGY WITH NO DATA SCALING, AFTER NORMALISATION AND AFTER STANDARDISATION.

Topology	Input data	% Accuracy	
		No imputation	With imputation
22-22-11	no scaling	61.94	66.1
	normalised	66.70	66.4
	standardised	67.49	66.7

Sample test confusion matrices for the two cases of training NN on standardised input dataset without imputation and with imputation are given in Fig. 3 and Fig.4 respectively. Although the results from Fig.4 seem slightly inferior, they give higher statistical confidence with the largely increased number of ‘hits’.

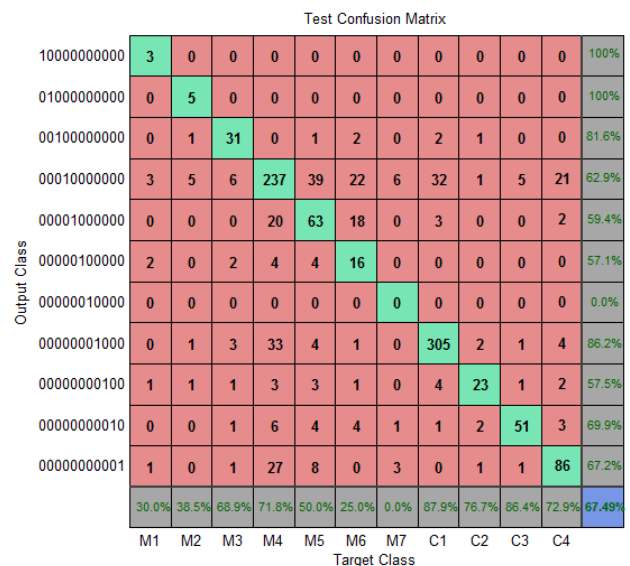


Fig. 3. No imputed data classification results of 22-22-11 NN classifier with standardized data on 7 military (‘M1’ – *Multi-function*, ‘M2’ – *Battlefield*, ‘M3’ – *Aircraft*, ‘M4’ – *Search*, ‘M5’ – *Air Defense*, ‘M6’ – *Weapon*, and ‘M7’ – *Info*) and 4 civil classes (‘C1’ – *Maritime*, ‘C2’ – *Airborne Navigation*, ‘C3’ – *Meteorological*, and ‘C4’ – *Air Traffic Control*).

It can be seen from Figure 4 that the accuracy of the classifier trained on dataset with imputed data has increased by x% on average. Some may argue that this is not a big improvement, but if we consider the results for each class separately, one can notice that, e.g., for the first class in the case with no missing data, there were only three testing samples, which the NN hit. In the second case for training with imputed data, there are now xx testing samples ... the presence of imputed data actually increases the statistical power of the data, as there are more samples available for training as well as for testing the generalization abilities of the NN.

Test Confusion Matrix

Output Class	M1	M2	M3	M4	M5	M6	M7	C1	C2	C3	C4	Accuracy
10000000000	59	1	14	1	4	4	0	0	2	1	2	67.0%
01000000000	4	34	6	2	0	6	0	1	4	1	1	57.6%
00100000000	18	9	88	9	6	6	0	4	9	3	7	55.3%
00010000000	11	27	11	372	112	35	5	40	4	15	48	54.7%
00001000000	1	2	6	17	45	16	0	3	0	4	2	46.9%
00000100000	7	5	7	8	12	132	5	3	4	9	1	68.4%
00000010000	1	0	2	4	2	0	107	1	0	3	4	86.3%
00000001000	0	2	12	70	2	2	0	411	9	8	13	77.7%
00000000100	1	1	9	2	3	5	3	5	76	3	9	65.0%
00000000010	0	1	0	11	2	2	2	2	0	141	1	87.0%
00000000001	0	0	0	25	11	3	1	0	0	0	101	71.6%
10000000000	57.8%	41.5%	56.8%	71.4%	22.6%	62.6%	87.0%	87.4%	70.4%	75.0%	53.4%	66.7%

Fig. 4. Data classification results for imputed data of 22-22-11 NN classifier with standardised data on 7 military ('M1' – *Multi-function*, 'M2' – *Battlefield*, 'M3' – *Aircraft*, 'M4' – *Search*, 'M5' – *Air Defence*, 'M6' – *Weapon*, and 'M7' – *Info*) and 4 civil classes ('C1' – *Maritime*, 'C2' – *Airborne Navigation*, 'C3' – *Meteorological*, and 'C4' – *Air Traffic Control*).

VII. CONCLUSION

Multiple imputation model is employed for dealing with the large number of missing data in the available radar signal dataset. The use of MI on samples with up to 60% of missingness enabled us to double the data subset used in [17]. Subsequently, this larger data subset is used for feedforward backpropagation supervised learning, when solving the radar signal classification and identification problem.

The application of neural network classifiers for recognition of generic radar data signal train pulse sources is investigated, implemented, tested and validated for the two cases with and without imputation.

The results are compared and critically analysed, showing overall improved accuracy when the NN are trained on the larger subset with imputed values.

Future work that can further improve the classification accuracy will include linear and principal component analysis

for dimensionality reduction and increased separability. When preprocessing the available large dataset we will also investigate the applicability of MI to samples with more than 60% missingness.

REFERENCES

- [1] C. K. Enders, *Applied missing data analysis*. Guilford Press, New York, 2010.
- [2] A. N. Baraldi, and C. K. Enders, "An introduction to modern missing data analyses," *Journal of School Psychology*, vol. 48, no. 1, pp. 5-37, 2010.
- [3] J. W. Graham, "Missing data analysis: Making it work in the real world," *Annual review of psychology*, vol. 60, pp. 549-576, 2009.
- [4] N. J. Horton, and S. R. Lipsitz, "Multiple imputation in practice: comparison of software packages for regression models with missing variables," *The American Statistician*, vol. 55, no. 3, pp. 244-254, 2001.
- [5] R. J. Little and D.B. Rubin, *Statistical analysis with missing data*. Wiley, New York, 2002.
- [6] J. W. Osborne, *Best Practices in Data Cleaning*. SAGE, 2013.
- [7] M. A. Richards, *Fundamentals of radar signal processing*. Tata McGraw-Hill Education, 2005.
- [8] D. C. Schleher, *Electronic warfare in the information age*. Artech House Inc, 1999.
- [9] E. Granger, M. Rubin, S. Grossberg, and P. Lavoie, "A What-and-Where fusion neural network for recognition and tracking of multiple radar emitters," *Neural Networks*, vol.14 (3), pp. 325-344, 2001.
- [10] N. Ibrahim, R. R. Abdullah, and M. Saripan, "Artificial neural network approach in radar target classification," *Journal of Computer Science*, vol. 5, no. 1, pp. 23, 2009.
- [11] M. R. Azimi-Sadjadi, D. Yao, Q. Huang, and G. J. Dobeck, "Underwater target classification using wavelet packets and neural networks," *IEEE Trans. on Neural Networks*, vol. 11, no. 3, pp. 784-794, 2000.
- [12] L. Anjaneyulu, N. Sarma, and N. Murthy, "Identification of LPI radar signals by higher order spectra and neural network techniques," *Int. J of Inf. and Comm. Technology*, vol. 2, no. 1, pp. 142-155, 2009.
- [13] Z. Yin, W. Yang, Z. Yang, L. Zuo, and H. Gao, "A study on radar emitter recognition based on SPDS neural network," *Information Technology Journal*, vol. 10, no. 4, pp. 883-888, 2011.
- [14] Z. C. Zhang, X. Guan and Y. He, "Study on radar emitter recognition signal based on rough sets and RBF neural network," *IEEE Int. Conf. on Machine Learning and Cybernetics*, vol. 2, pp. 1225-1230, 2009.