



**HAL**  
open science

## Autoencoder-based Attribute Noise Handling Method for Medical Data

Thomas Ranvier, Haytham Elgazel, Emmanuel Coquery, Khalid Benabdeslem

► **To cite this version:**

Thomas Ranvier, Haytham Elgazel, Emmanuel Coquery, Khalid Benabdeslem. Autoencoder-based Attribute Noise Handling Method for Medical Data. 1793, Springer Nature Singapore, pp.212-223, 2023, Communications in Computer and Information Science, 10.1007/978-981-99-1645-0\_18 . hal-03696250

**HAL Id: hal-03696250**

**<https://hal.science/hal-03696250v1>**

Submitted on 15 Jun 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Autoencoder-based Attribute Noise Handling Method for Medical Data

Thomas Ranvier  $\Gamma^{1,2}$ [0000 – 0001 – 9250 – 9530], Haytham Elgazel<sup>1,3</sup>,  
Emmanuel Coquery<sup>1,4</sup>, and Khalid Benabdeslem<sup>1,5</sup>

<sup>1</sup> Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205  
43 bd du 11 Novembre 1918, 69622 Villeurbanne, France

<sup>2</sup> [thomas.ranvier@univ-lyon1.fr](mailto:thomas.ranvier@univ-lyon1.fr)

<sup>3</sup> [haytham.elgazel@univ-lyon1.fr](mailto:haytham.elgazel@univ-lyon1.fr)

<sup>4</sup> [emmanuel.coquery@liris.cnrs.fr](mailto:emmanuel.coquery@liris.cnrs.fr)

<sup>5</sup> [khalid.benabdeslem@univ-lyon1.fr](mailto:khalid.benabdeslem@univ-lyon1.fr)

**Abstract.** Medical datasets are particularly subject to attribute noise, that is, missing and erroneous values. Attribute noise is known to be largely detrimental to learning performances. To maximize future learning performances it is primordial to deal with attribute noise before any inference. We propose a simple autoencoder-based preprocessing method that can correct mixed-type tabular data corrupted by attribute noise. No other method currently exists to handle attribute noise in tabular data. We experimentally demonstrate that our method outperforms both state-of-the-art imputation methods and noise correction methods on several real-world medical datasets.

**Keywords:** Data Denoising · Data Imputation · Attribute Noise · Machine Learning · Deep Learning

## 1 Introduction

Medical studies are particularly subject to outliers, erroneous, meaningless, or missing values. In most real-life studies, not solely limited to the medical field, the problem of incomplete data and erroneous data is unavoidable. Those corruptions can occur at any data collection step. They can be a natural part of the data (patient noncompliance, irrelevant measurement, etc.) or appear from corruption during a later data manipulation phase [15]. Regardless of their origin, those corruptions are referred as “noise” in the following work. Noise negatively impacts the interpretation of the data, be it for a manual data analysis or training an inference model on the data. The goal of a machine learning model is to learn inferences and generalizations from training data and use the acquired knowledge to perform predictions on unseen test data later on. Thus, the quality of training data on which a model is based is of critical importance, the less noisy the data is, the better results we can expect from the model.

Noise can be divided into two categories, namely class noise and attribute noise [17]. Class noise corresponds to noise in the labels, e.g. when data points are

labeled with the wrong class, etc. Attribute noise on the other hand corresponds to erroneous and missing values in the attribute data, that is, the features of the instances. Attribute noise tends to occur more often than class noise in real-world data [15][17][13]. Despite this fact, compared to class noise, very limited attention has been given to attribute noise [17]. In real-world medical data, the probability of mislabeled data in a survival outcome context is quite low, we focused our work on attribute noise to maximize prediction performance while trying to compensate for a lack of appropriate methods within the literature.

The problem of imputing missing values has been vastly addressed in the literature, one can choose from many imputation methods to complete its data depending on its specific needs [9]. Imputation methods only address part of the attribute noise problem, they can handle missing values but do not handle erroneous values, which can be highly detrimental to imputation results. Those methods have been widely researched, but methods able to deal with erroneous values have been less researched and can be considered incomplete at the moment [15].

Handling erroneous values can be done in three main ways: using robust learners that can learn directly from noisy data and naturally compensate or partially ignore the noise, filtering methods that remove data points that are classified as noisy, and polishing methods that aim to correct noisy instances. Robust learners are models that are less sensitive to noise in the data than classic models but they present several disadvantages [13]. They usually have limited learning potential compared to other learners. Using robust learners is not useful if we aim to perform anything else than the task the learner will solve. Filtering methods aim to detect which instances are noisy to delete them from the training set [17][13]. By training a learner on this cleaned set it can learn inferences without being disturbed by erroneous values and outliers which eventually leads to better prediction performances on test data. The third way to deal with erroneous values is the polishing method [11], which corrects instances detected as noisy. Such a method can correct erroneous values on small datasets but lacks scalability for larger datasets containing more features [13]. Those three methods are able to deal with erroneous values and outliers but are not able to deal with incomplete data, they only address part of the attribute noise problem.

At the moment the only way to handle attribute noise in its entirety is to use a combination of an imputation method followed by a noise correction method, to the best of our knowledge the literature lacks a method that would be able to perform both those tasks at once. Real-world data and especially medical data are subject to attribute noise in its entirety, it is important to conceive an approach able to handle the totality of attribute noise and not just subpart of it.

In this paper, we propose a preprocessing method based on autoencoders that deals with attribute noise in its entirety in real-world tabular and mixed-type medical data. Our method is able to learn from incomplete and noisy data to produce a corrected version of the dataset. It does not require any complete instance in the dataset and can truly handle attribute noise by performing both

missing values completion and correction of erroneous values at the same time. We conduct extensive experiments on an imputation task on real-world medical data to compare our method to other state-of-the-art methods and obtain competitive and even significantly better results on classification tasks performed on the corrected data. We extend our experiments to show that our method can both complete missing data while correcting erroneous values, which further improves the obtained results.

The complete source code used to conduct the experiments is available at the following github repository<sup>6</sup>.

The rest of the paper is organized as follows: we first present related work of data imputation and noise correction in both tabular and image data, especially in the medical field, in section 2. Then, we present and explain our proposed approach in section 3. Section 4 shows our experimental results compared to both data imputation and noise correction state-of-the-art methods. Finally, we conclude with a summary of our contributions.

## 2 Related work

Denosing is vastly researched in the image field, in the image medical domain it is easy to find recent reviews and methods to correct medical images [6]. Correction of tabular data on the other hand is less researched, only the imputation part seems to attract lots of attention. In this paper, we are especially focused on methods that can be applied to mixed-type tabular data.

Recently lots of autoencoder-based imputation methods have been researched [8]. An autoencoder is a machine learning algorithm that takes an input  $x \in \mathbb{R}^d$  with  $d$  the number of features and learns an intermediate representation of the data noted  $z \in \mathbb{R}^h$  with  $h$  the size of the newly constructed latent space. Then, from the intermediate representation  $z$  the model reconstructs the original data  $x$ , we note the model output  $\hat{x}$ . During its training, the reconstruction error between  $x$  and  $\hat{x}$  is minimized.

One of those autoencoder-based imputation methods is MIDA: Multiple Imputation using Denoising Autoencoders, introduced in 2018 [4]. Unlike most autoencoder-based methods which are usually applied to images, MIDA has been successfully applied to tabular data. This imputation method learns from a complete training dataset and can then be applied to unseen incomplete test data to impute the missing values. The authors assume that in order to learn how to impute missing values MIDA must learn from complete data. However, in this paper our experimental protocol does not provide a clean dataset to train on, therefore we show that MIDA obtains satisfactory results when properly parameterized, even when learning on incomplete data.

We want to show that autoencoders can not only be used to impute missing values, but also to correct erroneous values that are part of the observed values. It is easier to correct erroneous values in an image than in tabular data, since

<sup>6</sup> [https://github.com/ThomasRanvier/Autoencoder-based\\_Attribute\\_Noise\\_Handling\\_Method\\_for\\_Medical\\_Data](https://github.com/ThomasRanvier/Autoencoder-based_Attribute_Noise_Handling_Method_for_Medical_Data)

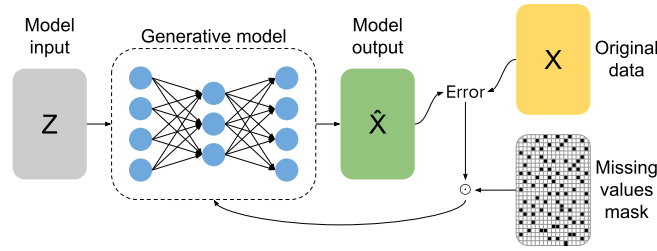
in images pixels in a close neighborhood are related to each other, which might not be true for arbitrarily ordered features in tabular data. As stated earlier correction of images is a very active research domain. Recently, Ulyanov et al. introduced a new approach called “Deep Image Prior [12].” That innovative approach uses autoencoders to restore images but does not use the original data  $X$  as model input, instead, the autoencoder is given pure noise as input and is trained to reconstruct the original corrupted data  $X$  from the noise. In that way, the model is no longer considered an autoencoder but a generative model, however, in practice the model keeps the same architecture. Therefore, the only information required to correct the input image is already contained in the image itself. By stopping the training before complete convergence it is possible to obtain a cleaner image than the original corrupted image. Ulyanov et al. showed that their approach outperforms other state-of-the-art methods on a large span of different tasks.

In this paper, we aim to conceive a method that would be able to correct mixed-type tabular data, we aim to use the lessons from [4] and [12] to conceive a method able to handle attribute noise as a whole as a preprocessing method.

### 3 A Method to Truly Handle Attribute Noise

Our method is based on a deep neural architecture that is trained to reconstruct the original data from a random noise input. We note the original data with its attribute noise  $X \in \mathbb{R}^{n \times d}$  with  $n$  the number of instances in the dataset and  $d$  the number of features. We note the deep generative model  $\hat{X} = f_{\Theta}(\cdot)$  with  $\Theta$  the model parameters that are learned during training and  $\hat{X}$  is the model output, in our case the model output is a reconstruction of  $X$ . The input of the model is noted  $Z \in \mathbb{R}^{n \times d}$  and has the same dimension as  $X$ , which keeps our model a kind of autoencoder. The model is trained to reconstruct  $X$  using the following loss term:  $L(X, \hat{X}) = \|(\hat{X} - X) \odot M\|^2$ , where  $\odot$  is the Hadamard product and  $M \in \mathbb{R}^{n \times d}$  is a binary mask that stores the locations of missing values in  $X$ ,  $M_{ij} = 1$  if  $X_{ij}$  is observed and  $M_{ij} = 0$  if  $X_{ij}$  is missing. By applying the mask  $M$  to the loss ensure that the loss is only computed on observed values, in this way the reconstruction  $\hat{X}$  will fit the observed values in  $X$ , while missing values will naturally converge to values that are statistically consistent given the learned data distribution. Figure 1 shows how the model is fitted to the original data  $X$  with the application of the binary mask  $M$  during training.

To determine the right step at which to stop training we define a stopping condition based on the evolution of a given metric. In a supervised setting, for example, we regularly compute the AUC on a prediction task performed on the reconstructed data  $\hat{X}$ , which gives us the evolution of the quality of the reconstruction. We stop the training when the AUC degrades for a set number of iterations, then we obtain a reconstruction with consistent imputations and noise correction, which provides better data quality than the original data. If no supervision is possible the only metric that can be used to determine when



**Fig. 1.** The model parameters are trained so that the model learns to reconstruct the original data. The training must be stopped at the right moment for the reconstruction  $\hat{X}$  to be cleaner than the original data  $X$ .

to stop training is the loss value, which gives correct results but is quite limited since it is harder to determine when overfitting starts.

In practice, we set the input of the model either as random noise or as the original data depending on the obtained results. We note that on datasets containing large amounts of features, using 1D convolutions instead of classic fully-connected layers tends to give better results, it helps our method to scale on datasets with large amounts of data and features.

What makes this method different from most other autoencoder-based methods and able to handle attribute noise as a whole is that we determine an early-stop condition to stop training at the proper moment. Stopping at the right moment allows the method to reach a point where the reconstructed data  $\hat{X}$  contains less noise than the original data  $X$  while containing all important information from  $X$ .

## 4 Experimental Results

### 4.1 Used Datasets

We ran our experiments on three real-life medical mixed-type tabular datasets naturally containing missing values. We evaluated our method and compared our results to other state-of-the-art methods on those medical datasets.

- NHANES, US National Health and Nutrition Examination Surveys: Those are surveys conducted periodically by the US NHCS to assess the health and nutritional status of the US population [1]. We used data from studies spanning from 2000 to 2008, with 95 features and about 33% missing values. We selected the “diabete” feature as a class and randomly selected 1000 samples from both outcomes to evaluate the quality of the data correction on a classification task on this class.
- COVID19: This dataset was publicly released with the paper [14], it contains medical information collected between in early 2020 on pregnant and breastfeeding women. We based our data preprocessing on the one realized

in the original paper, we selected only the measurements from the last medical appointment for each patient. After preprocessing, we obtain a dataset composed of 361 patients with 76 features, with about 20% missing data. We evaluate the quality of the data correction on a classification task on the survival outcome, 195 patients have survived and 166 are deceased.

- Myocardial infarction complications: This medical dataset is available on the UCI machine learning repository, it was publicly released with the paper [3]. It is composed of 1700 patients with 107 features, with about 5% missing values. We evaluate the quality of the data correction on a classification task on the survival outcome, 1429 patients have survived and 271 are deceased.

## 4.2 Used Metrics

We evaluate the quality of the obtained correction on classification tasks. As can be seen from the previous section, the medical datasets we used are not all balanced, the Myocardial dataset is especially imbalanced. In such a context it is important to choose metrics that are not sensitive to imbalance.

In a medical context where we aim to predict the outcome between sane and sick, it is extremely important not to classify sick patients as sane since it would be very detrimental for them not to get an appropriate medical response. In machine learning terms we are in cases where false positives on the negative class would be less detrimental than false negatives, therefore we should aim to minimize false negatives.

Appropriate metrics, in this case, are the AUC: Area Under the Receiver Operating Characteristic (ROC) Curve, and the balanced accuracy. The AUC corresponds to the area under the ROC curve obtained by plotting the true positive rate (recall) against the false positive rate (1-specificity). An AUC score of 1 would mean that the classifier gives true positives 100% of the time, whereas a value of 0.5 means that the classifier is no better than a random prediction. The balanced accuracy is defined as the average of recall obtained on each class, which is simply the average of the true positive rate between all the classes.

## 4.3 Experimental Protocol

Our experiments aim to compare our method to other state-of-the-art methods for both imputation and noise correction tasks. All our experiments are evaluated using the balanced accuracy and the AUC metrics. We repeated each experiment 10 times with 10 different stochastic seeds to set up the random state of non-deterministic methods. For each experiment we compare the performances of each method to ours using t-tests. We use the results from those statistical tests to determine if our method is significantly better, even, or significantly worse than each other method, based on a  $p$ -value set at 0.05.

We first evaluate our method on an imputation task on the three medical datasets previously described. Each of those datasets is missing part of its data, we compare the quality of the data imputation by training a decision tree on a classification task on each dataset after imputation.

The capacity of our method to impute missing values on incomplete and noisy data is assessed by introducing artificial noise in the datasets. Noise is artificially added to the data by randomly replacing attribute values with a random number at a certain rate, as described by Zhu et al. [17]. We compare the results at the following noise rates: 0/5/10/15/20/40/60%.

Finally, to assess the effectiveness of our method to both complete missing values while correcting erroneous values, we introduce artificial noise in the naturally incomplete datasets. We apply our method and compare its results to those obtained by a sequential execution (*i.e.* pipeline) of an imputation method followed by a noise correction method.

The results from our method are compared to those of other state-of-the-art imputation methods:

- MEAN, MEDIAN and KNN: We used the “SimpleImputer” and “KNNImputer” classes from the python library “scikit-learn”<sup>7</sup>.
- MICE: Multivariate Imputation by Chained Equations has been introduced in 2011 in [2]. This is a very popular method of imputation because it provides fast, robust, and good results in most cases. We used the implementation from the experimental “IterativeImputer” class from “scikit-learn”.
- GAIN: Generative Adversarial Imputation Nets, introduced recently in [16], two models are trained in an adversarial manner to achieve good imputation. We used the implementation from the original authors<sup>8</sup>.
- SINKHORN: An optimal transport based method for data imputation introduced in [7] We used the implementation from the original authors<sup>9</sup>.
- SOFTIMPUTE: The SOFTIMPUTE algorithm has been proposed in 2010 [5], it iteratively imputes missing values using an SVD. We used the public re-implementation by Travis Brady of the Mazumder and Hastie’s package<sup>10</sup>.
- MISSFOREST: An iterative imputation method based on random forests introduced in 2012 in [10]. We used the “MissForest” class from the python library “missingpy”<sup>11</sup>.
- MIDA: Multiple Imputation Using Denoising Autoencoders has been recently proposed in [4]. We implemented MIDA using the author description from the original paper and the code template supplied in this public gist<sup>12</sup>

The following noise correction methods are used as comparison:

- SFIL: Standard Filtering, which we implemented such as described in [11].
- SPOL: Standard Polishing, which we also implemented such as described in [11].
- PFIL, PPOL: Improved versions of SFIL and SPOL where the noisy instances to filter or polish are identified using the noise detection method Panda [13].

<sup>7</sup> <https://scikit-learn.org>

<sup>8</sup> <https://github.com/jsyoon0823/GAIN>

<sup>9</sup> <https://github.com/BorisMuzellec/MissingDataOT>

<sup>10</sup> <https://github.com/travisbrady/py-soft-impute>

<sup>11</sup> <https://pypi.org/project/missingpy/>

<sup>12</sup> <https://gist.github.com/lgonlara/18387c5f4d745673e9ca8e23f3d7ebd3>



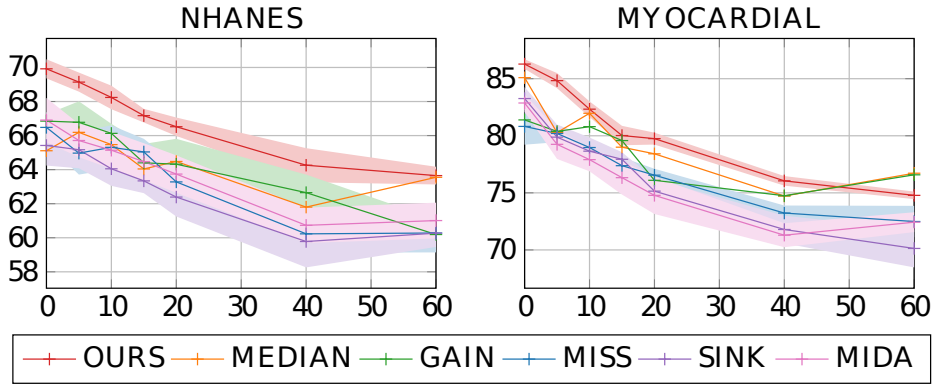
#### 4.4 Results

In this subsection, we present and analyze the most important comparative results between our proposed method and state-of-the-art methods. The entire experimental results with statistical significance can be found in our code.

**Imputation on Incomplete Medical Data** With our first experiment, we show that our method can impute missing values in real-world medical datasets.

Table 1 shows the results on the three real-world medical datasets. We can see that our method obtains very competitive results on all datasets. We obtain significantly better results than other state-of-the-art methods in most cases for both metrics. The only cases in which our method performs significantly worse are against KNN and MICE on COVID data on the balanced accuracy metric. This shows that our method is able to impute missing values on incomplete real-world medical mixed-type tabular data with results as good as other state-of-the-art imputation methods and even better in most cases.

**Imputation on Incomplete and Noisy Medical Data** Our second experiment shows that our method can impute missing values in real-world medical datasets in a noisy context. We artificially add noise to the data at various rates: 0/5/10/15/20/40/60%, and evaluate each imputation method at each noise level.



**Fig. 2.** AUC results on imputation on incomplete and noisy medical data

Figure 2 shows AUC results obtained on NHANES and MYOCARDIAL data at each noise rate against several imputation methods. In both cases, we note that our method globally obtains significantly better results than other methods. The performance of all methods drops when the noise level increases, which is expected. On NHANES data our method performs largely better than others until a noise rate of 60% where the MEDIAN imputation gets similar results to ours. This can probably be explained by the fact that with a noise level that high

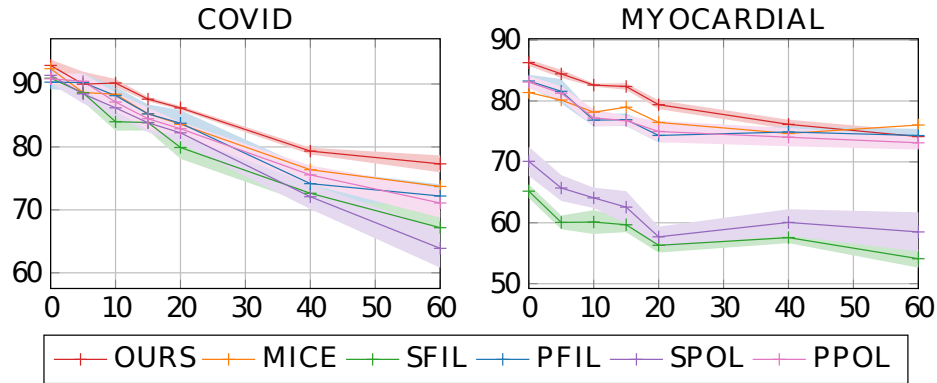
**Table 1.** Comparative study between our method and other methods. BalACC corresponds to the balanced accuracy, AUC is the area under the ROC curve. Our method is compared to each other using t-tests with a  $p$ -value of 0.05, when our method is significantly better it is indicated by  $\bullet$ , even by  $\equiv$ , and significantly worse by  $\circ$ .

Model	Metric	MYOCARDIAL	NHANES	COVID
OURS	BalACC	<b>77.91%</b> $\pm 1.12\%$	<b>64.17%</b> $\pm 0.36\%$	86.84% $\pm 1.23\%$
	AUC	<b>86.28%</b> $\pm 0.42\%$	<b>69.92%</b> $\pm 0.56\%$	<b>92.95%</b> $\pm 0.93\%$
MEAN	BalACC	77.30% $\pm 0.00\%$ $\equiv$	60.35% $\pm 0.00\%$ $\bullet$	85.91% $\pm 0.00\%$ $\bullet$
	AUC	85.09% $\pm 0.00\%$ $\bullet$	66.10% $\pm 0.00\%$ $\bullet$	91.20% $\pm 0.00\%$ $\bullet$
KNN	BalACC	68.83% $\pm 0.00\%$ $\bullet$	63.00% $\pm 0.00\%$ $\bullet$	<b>88.08%</b> $\pm 0.00\%$ $\circ$
	AUC	78.94% $\pm 0.00\%$ $\bullet$	67.78% $\pm 0.00\%$ $\bullet$	91.53% $\pm 0.00\%$ $\bullet$
GAIN	BalACC	63.89% $\pm 2.21\%$ $\bullet$	61.36% $\pm 0.53\%$ $\bullet$	85.14% $\pm 0.91\%$ $\bullet$
	AUC	74.22% $\pm 1.11\%$ $\bullet$	66.85% $\pm 0.40\%$ $\bullet$	91.36% $\pm 0.73\%$ $\bullet$
MICE	BalACC	76.55% $\pm 0.00\%$ $\bullet$	61.70% $\pm 0.00\%$ $\bullet$	87.98% $\pm 0.00\%$ $\circ$
	AUC	81.39% $\pm 0.00\%$ $\bullet$	67.30% $\pm 0.00\%$ $\bullet$	92.43% $\pm 0.00\%$ $\equiv$
MISSFOREST	BalACC	73.00% $\pm 0.87\%$ $\bullet$	61.40% $\pm 1.03\%$ $\bullet$	85.15% $\pm 1.67\%$ $\bullet$
	AUC	80.82% $\pm 1.60\%$ $\bullet$	66.48% $\pm 0.90\%$ $\bullet$	91.30% $\pm 1.20\%$ $\bullet$
SOFTIMPUTE	BalACC	77.24% $\pm 0.99\%$ $\equiv$	61.70% $\pm 0.93\%$ $\bullet$	84.48% $\pm 0.78\%$ $\bullet$
	AUC	84.88% $\pm 0.77\%$ $\bullet$	66.93% $\pm 1.08\%$ $\bullet$	91.12% $\pm 0.85\%$ $\bullet$
SINKHORN	BalACC	75.66% $\pm 1.22\%$ $\bullet$	60.77% $\pm 0.98\%$ $\bullet$	86.82% $\pm 1.49\%$ $\equiv$
	AUC	83.26% $\pm 1.01\%$ $\bullet$	65.42% $\pm 1.18\%$ $\bullet$	91.48% $\pm 1.13\%$ $\bullet$
MIDA	BalACC	75.09% $\pm 0.70\%$ $\bullet$	62.15% $\pm 1.26\%$ $\bullet$	85.55% $\pm 1.12\%$ $\bullet$
	AUC	82.87% $\pm 0.78\%$ $\bullet$	66.91% $\pm 1.30\%$ $\bullet$	91.67% $\pm 0.62\%$ $\bullet$

it is nearly impossible to impute coherent values other than the median or mean value for each feature. We can observe the same pattern on MYOCARDIAL data, with the difference that GAIN seems to have learned how to adapt to such an amount of noise in this case. Those results show that on low to high noise rates, our method can impute missing values while correcting erroneous values.

It provides better data correction than most other methods. At extreme noise rates naive methods might provide better results.

**Comparison with the Combination of Imputation and Noise Correction Methods** The last experiment compares our method results to those obtained from the combination of an imputation method followed by a noise correction method. We chose MICE as the state-of-the-art imputation method since it obtains competitive results against ours in a not noisy context. We then apply the four noise correction methods SFIL, PFIL, SPOL, and PPOL.



**Fig. 3.** AUC results on combination of imputation and noise correction

Figure 3 show AUC results obtained on COVID and MYOCARDIAL data at each noise rate. We note that SFIL and SPOL perform worse than the Panda alternative of both those methods at all noise rates. We also note that for both datasets the other state-of-the-art noise correction methods give very poor results as soon as the noise level reaches more than 5%, at higher noise rates the data quality is better before noise correction than before. For COVID data all methods yield similar results at low noise levels, with our method on top with a very slight advantage. At high rates, however, our method gives very significantly better results than all other methods. For MYOCARDIAL data the opposite pattern can be observed, our method gives significantly better results up until a noise rate of 40%, after which MICE imputation is slightly better. This experiment completes the conclusions drawn from the second experiment, our method provides very good data correction, up until the noise rate becomes too extreme, at that point, simpler methods achieve slightly better results. The fact that the opposite is observed on COVID data is probably due to a remarkable original data quality, which would explain why our method becomes significantly better only at higher noise levels.

## 5 Conclusion

Handling attribute noise means imputing missing values while correcting erroneous values and outliers. This phenomenon is of critical importance in medical data, where attribute noise is especially present and detrimental to analysis and learning tasks on the data. No method in the literature is capable of handling attribute noise in its entirety in mixed-type tabular data. Many methods exist to impute missing values while other methods can correct erroneous values.

In this paper, we propose an autoencoder-based preprocessing approach to truly handle attribute noise. Our method imputes missing values while correcting erroneous values without requiring any complete or clean instance in the dataset to correct. Our experiments show that our method competes against and even outperforms other imputation methods on real-world medical mixed-type tabular data. Our method is less sensitive to noise on an imputation task.

Finally, as autoencoder approaches are amenable to an empirical tuning phase, we plan to implement in the future an algorithm able to automatically define an adapted architecture depending on the dataset dimensions.

## References

1. Barnard, J., Meng, X.L.: Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research* **8**(1), 17–36 (Feb 1999). <https://doi.org/10.1177/096228029900800103>
2. Buuren, S.v., Groothuis-Oudshoorn, K.: mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* **45**(3), 1–67 (Dec 2011). <https://doi.org/10.18637/jss.v045.i03>
3. Golovenkin, S.E., Bac, J., Chervov, A., Mirkes, E.M., Orlova, Y.V., Barillot, E., Gorban, A.N., Zinovyev, A.: Trajectories, bifurcations, and pseudo-time in large clinical datasets: applications to myocardial infarction and diabetes data. *Giga-Science* **9**(11), g1aa128 (Nov 2020). <https://doi.org/10.1093/gigascience/g1aa128>
4. Gondara, L., Wang, K.: MIDA: Multiple Imputation Using Denoising Autoencoders. *Pacific-Asia Conference on Knowledge Discovery and Data Mining 2018* pp. 260–272 (2018). [https://doi.org/10.1007/978-3-319-93040-4\\_21](https://doi.org/10.1007/978-3-319-93040-4_21)
5. Mazumder, R., Hastie, T., Tibshirani, R.: Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *Journal of machine learning research : JMLR* **11**, 2287–2322 (Mar 2010)
6. Mohd Sagheer, S.V., George, S.N.: A review on medical image denoising algorithms. *Biomedical Signal Processing and Control* **61**, 102036 (Aug 2020). <https://doi.org/10.1016/j.bspc.2020.102036>
7. Muzellec, B., Josse, J., Boyer, C., Cuturi, M.: Missing Data Imputation using Optimal Transport. In: *Proceedings of the 37th International Conference on Machine Learning*. pp. 7130–7140. PMLR (Nov 2020), iISSN: 2640-3498
8. Pereira, R.C., Santos, M., Rodrigues, P., Henriques Abreu, P.: Reviewing Autoencoders for Missing Data Imputation: Technical Trends, Applications and Outcomes. *Journal of Artificial Intelligence Research* **69**, 1255–1285 (Dec 2020). <https://doi.org/10.1613/jair.1.12312>
9. Stef, V.B.: *Flexible Imputation of missing Data*. Chapman & Hall, second edition edn. (2018)

10. Stekhoven, D.J., Bühlmann, P.: MissForest—Non-Parametric Missing Value Imputation for Mixed-Type Data. *Bioinformatics* **28**(1), 112–118 (Jan 2012). <https://doi.org/10.1093/bioinformatics/btr597>
11. Teng, C.M.: Polishing Blemishes: Issues in Data Correction. *IEEE Intelligent Systems* **19**(2), 34–39 (Mar 2004). <https://doi.org/10.1109/MIS.2004.1274909>, conference Name: IEEE Intelligent Systems
12. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep Image Prior. *International Journal of Computer Vision* **128**(7), 1867–1888 (Jul 2020). <https://doi.org/10.1007/s11263-020-01303-4>
13. Van Hulse, J.D., Khoshgoftaar, T.M., Huang, H.: The Pairwise Attribute Noise Detection Algorithm. *Knowledge and Information Systems* **11**(2), 171–190 (Feb 2007). <https://doi.org/10.1007/s10115-006-0022-x>
14. Yan, L., Zhang, H.T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., Sun, C., Tang, X., Jing, L., Zhang, M., Huang, X., Xiao, Y., Cao, H., Chen, Y., Ren, T., Wang, F., Xiao, Y., Huang, S., Tan, X., Huang, N., Jiao, B., Cheng, C., Zhang, Y., Luo, A., Mombaerts, L., Jin, J., Cao, Z., Li, S., Xu, H., Yuan, Y.: An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence* **2**(5), 283–288 (May 2020). <https://doi.org/10.1038/s42256-020-0180-7>
15. Yang, Y., Wu, X., Zhu, X.: Dealing with Predictive-but-Unpredictable Attributes in Noisy Data Sources. In: Boulicaut, J.F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) *Knowledge Discovery in Databases: PKDD 2004*. pp. 471–483. *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg (2004). [https://doi.org/10.1007/978-3-540-30116-5\\_43](https://doi.org/10.1007/978-3-540-30116-5_43)
16. Yoon, J., Jordon, J., Schaar, M.: GAIN: Missing Data Imputation using Generative Adversarial Nets. In: *Proceedings of the 35th International Conference on Machine Learning*. pp. 5689–5698. PMLR (Jul 2018), iSSN: 2640-3498
17. Zhu, X., Wu, X.: Class Noise vs. Attribute Noise: A Quantitative Study. *Artificial Intelligence Review* **22**(3), 177–210 (Nov 2004). <https://doi.org/10.1007/s10462-004-0751-8>

## Acknowledgments

This research is supported by the European Union’s Horizon 2020 research and innovation program under grant agreement No 875171, project QUALI-TOP (Monitoring multidimensional aspects of QUALity of Life after cancer ImmunoTherapy - an Open smart digital Platform for personalized prevention and patient management).