Data Imputation with an Autoencoder and MAGIC

Devin Eddington
Department of Mathematics and Statistics
Utah State University
Logan, Utah, USA
devineddington@yahoo.com

Andres F. Duque Correa
Department of Mathematics and Statistics
Utah State University
Logan, Utah, USA
andresduquecorrea1990@gmail.com

Guy Wolf
Department of Mathematics and Statistics
University of Montreal
Montreal, QC, Canada
wolfguy@mila.quebec

Kevin R. Moon
Department of Mathematics and Statistics
Utah State University
Logan, Utah, USA
kevin.moon@usu.edu

***Abstract*—Missing data is a common problem in many applications. Imputing missing values is a challenging task, as the imputations need to be accurate and robust to avoid introducing bias in downstream analysis. In this paper, we propose an ensemble method that combines the strengths of a manifold learning-based imputation method called MAGIC and an autoencoder deep learning model. We call our method Deep MAGIC. Deep MAGIC is trained on a linear combination of the mean squared error of the original data and the mean squared error of the MAGIC-imputed data. Experimental results on three benchmark datasets show that Deep MAGIC outperforms several state-of-the-art imputation methods, demonstrating its effectiveness and robustness in handling large amounts of missing data.**

I. Introduction

Missing data is a common problem in many real-world applications, including healthcare, finance, and social sciences [1], [2], [3]. In particular, single-cell RNA-sequencing (scRNA-seq) data are often characterized by high levels of technical noise, missing values, and batch effects, which can compromise downstream analyses [4]. Therefore, developing effective imputation methods to recover missing values in scRNA-seq data has become an important research topic.

Imputing missing values is a challenging task, as the imputations need to be accurate and robust to avoid introducing bias in downstream analysis. MAGIC was designed to use the underlying data structure to create a smoothed imputation of the original data [5]. Points are smoothed by replacing each point with a weighted average of its neighbors, where the weights are determined by a manifold-based diffusion process.

The main advantage of MAGIC is that it can handle nonlinear relationships and non-normal data distributions. MAGIC assumes that all data are subject to noise and the smoothing is applied everywhere. Thus MAGIC is well-suited for undersampling-based missing values such as in scRNA-seq, and works less well for missing values that follow a zero-inflated model. The MAGIC results are also sensitive to the choice of hyperparameters, and as a nonparametric kernel method, can not be easily applied to new points (i.e. out of sample extension).

Autoencoders are deep learning models that consist of an encoder and a decoder. The encoder maps the input data to a low-dimensional latent space and the decoder maps from the latent space back to the original inputs [6]. Choosing the dimension of the latent space to be lower than that of the input space forces the autoencoder to learn a representation that captures the most salient features of the data. Denoising autoencoders denoise the data by adding noise to the inputs during training and learning to undo the noise in the output of the decoder. Denoising autoencoders have been successful on denoising medical images [7], [8]. As parametric models, denoising autoencoders can easily perform out of sample extension. However, deep neural networks can be difficult to train. Also, autoencoders specifically have been shown to fail at capturing the true structure of the data, which may lead to less accurate reconstructions [9], [10].

In this paper, we develop an ensemble method called Deep MAGIC that combines the strengths of MAGIC with those of a denosing autoencoder. Deep MAGIC generates initial imputations using MAGIC. A denoising autoencoder is then trained on the original data with a regularization term added to the final layer that encourages the autoencoder output to match the MAGIC output. Experimental results on three benchmark datasets show that Deep MAGIC outperforms both MAGIC and the autoencoder alone as well as other state-of-the-art imputation methods, demonstrating its effectiveness and robustness in handling missing data in extreme dropout settings.

The remainder of the paper is organized as follows. Section II provides an overview of related work on missing data imputation methods. Section III presents the proposed ensemble method in detail, including the MAGIC imputation method, the Autoencoder model, and the ensemble approach. Section IV describes the experimental setup, including the datasets used and the evaluation metrics. Section V presents the experimental results and compares the proposed method with

This research was supported in part by the NSF under Grant 2212325 [K.M].

several state-of-the-art imputation methods. Finally, Section VI concludes the paper and discusses directions for future work.

II. Related Work

Various imputation methods have been proposed in the literature for handling missing data [11]. Single imputation methods such as mean imputation and regression imputation replace missing values with the mean or a regression prediction of the non-missing values. Multiple imputation methods generate several imputations and combine them to obtain a final estimate, such as MICE [12] and missForest [13].

Recently, deep learning models such as Autoencoders and Generative Adversarial Networks (GANs) have shown promise in handling missing data. GANs learn to generate imputations that are indistinguishable from the original data distribution. An autoencoder forces the data into a low-dimensional representation, thus retaining the most salient features of the data while decreasing the amount of noise in the data. Denoising autoencoders take this further by adding noise directly to the inputs, and then learn a mapping back to the original inputs. Thus the autoencoder directly learns a denoising function. By applying a missing data model, the denoising autoencoder can learn to impute missing data [14].

Random Forests have been widely used for data imputation tasks due to their ability to handle non-linear and complex relationships between variables. One way to impute missing data using a random forest is to extract proximities from it and replace the missing values with its proximity weighted sum [15]. Another way to impute missing data using Random Forests is by using the "missForest" algorithm [13]. Unlike the proximity weighted sum method, missForest directly predicts the missing values using a Random Forest trained on the observed parts of the dataset. The method sorts the variables according to the amount of missing values and imputes missing values starting with the variable that has the lowest amount of missing values. For each variable, the missing values are imputed by fitting a Random Forest with the observed values of that variable as the response and the observed values of the other variables as predictors. The trained Random Forest is then used to predict the missing values for that variable. This process is repeated until convergence or until a stopping criterion is met. Compared to other imputation methods, missForest has been shown to perform well on various types of missing data.

DrImpute uses a clustering approach to impute missing values. The expected value of a dropout event is obtained by averaging the entries in the same cell group in each clustering result. The expected value is then computed for each clustering result, and the final imputation for the putative dropout events is computed as a simple averaging [16].

Ensemble methods have also been proposed for missing data imputation [17]. Deep MAGIC builds on the strengths of both MAGIC and Autoencoder methods to generate accurate and robust imputations.

III. Deep MAGIC

Deep MAGIC is an ensemble imputation approach that combines the strengths of MAGIC and an autoencoder. MAGIC is used to generate initial imputations, and the autoencoder is trained using a loss function that incorporates the error of the network output with the original non-missing data as well as the error of the network output with the smoothed representation of the data after applying MAGIC.

Let $\mathbf{x}_i$ with $i = 1, \ldots, N$ be the training data with $\mathbf{x}_i \in \mathbb{R}^d$. MAGIC first learns a pairwise similarity between all points, which is collected into an $N \times N$ similarity matrix K. Each row in K is then normalized to obtain a probability transition matrix P. Powering P with t simulates a t-step random walk on the graph represented by K. The denoised $N \times d$ data matrix $\hat{X}$ is obtained by averaging each point with its t-step random walk neighbors: i.e. $\hat{X} = P^t X$, where X is the original data matrix.

Autoencoders are deep learning models that learn a mapping between the input data and the output data by encoding the input data into a low-dimensional latent space and decoding it back into the output space. The model is trained to minimize a loss function that measures the difference between the input data and the reconstructed output data. Let $f(\mathbf{x})$ be the output of the autoencoder when $\mathbf{x}$ is the input. The loss function for a standard autoencoder is then

$$\mathcal{L}_{AE} = \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{x}_i - f(\mathbf{x}_i)\right)^2 .$$

In a denoising autoencoder, noise is added to each of the inputs during the training process to create corrupted versions of the inputs: $\tilde{\mathbf{x}}$. Examples of commonly-used noise models include Gaussian noise, salt and pepper noise, and dropout or undersampling. The loss function for the denoising autoencoder is then

$$\mathcal{L}_{DAE} = \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{x}_i - f(\tilde{\mathbf{x}}_i)\right)^2 .$$

The input to the autoencoder during training is the noisy version of the inputs, but the loss function compares the output of the autoencoder with the original inputs. Thus the denoising autoencoder learns to "undo" the noise.

In Deep MAGIC, we regularize a denoising autoencoder to include the outputs obtained from MAGIC. Thus the loss function is

$$\mathcal{L}_{DM} = \frac{1}{N} \sum_{i=1}^{N} \left[\lambda \left(\mathbf{x}_i - f(\tilde{\mathbf{x}}_i)\right)^2 + (1-\lambda) \left(\hat{\mathbf{x}}_i - f(\tilde{\mathbf{x}}_i)\right)^2\right] ,$$

where $\lambda \in [0, 1]$ is a regularization parameter that determines the degree to which the autoencoder is constrained by the MAGIC representation of $\mathbf{x}_i$, $\hat{\mathbf{x}}_i$.

Regularizing the output of the autoencoder with the output of MAGIC is reminiscent of geometry regularized autoencoders (GRAE) [10]. In GRAE, various parts of the autoencoder are regularized to produce output similar to state of the art manifold learning methods. This approach was shown to

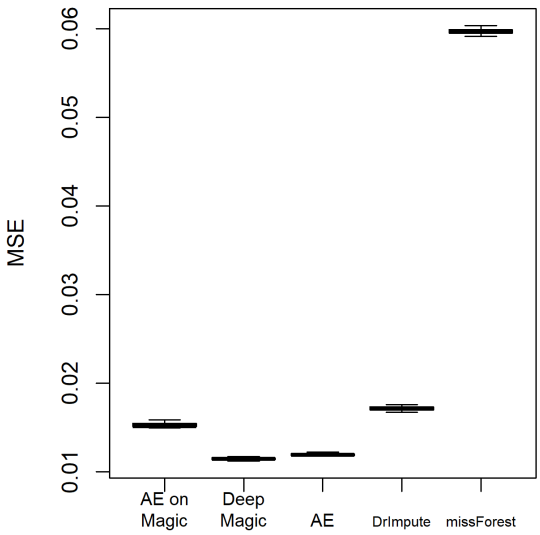

Fig. 1. Boxplot showing the distribution of mean squared error (MSE) values obtained from 10 runs of an autoencoder trained on MAGIC (AE on MAGIC), Deep MAGIC, an autoencoder (AE), DrImpute, and MissForest on the sine-cosine dataset with 40% dropout. The results show that Deep MAGIC outperforms the other methods in terms of median MSE and IQR at this dropout rate.

generally improve the reconstruction error of the autoencoder while providing a dimensionality reduction method that more faithfully captures the underlying geometry of the data, when compared to the original autoencoder. We are performing a similar task here with Deep MAGIC, where regularizing the denoising autoencoder output with the MAGIC output forces the autoencoder to pay more attention to the underlying geometry of the data. Deep MAGIC can then be easily used for out of sample extension by inputting new data into the trained architecture.

IV. Experimental Setup

The performance of Deep MAGIC was evaluated on three datasets: 1) a simulated scRNA-seq dataset generated using the Splatter package [18] with 3000 samples of 200 cells, batchCell = 3000, dp.prob = 0.5, bcv.common = 0.9, dropout.mid = 4, and 5 paths; 2) A sample of 2000 images from the MNIST database; and 3) 2000 simulated samples from a sine-cosine manifold represented in 200 dimensions with each dimension a linear combination of 3 out of 15 underlying intrinsic dimensions. Missing values are introduced by randomly removing values from the datasets according to a Bernoulli dropout distribution. The rates of dropout we used were $20\%, 40\%, 60\%, 80\%$, and 90%. The evaluation was done by comparing the mean squared error (MSE) of our proposed method against four other state-of-the-art methods: 1) an autoencoder trained to learn the MAGIC reconstruction, 2) a denoising autoencoder, 3) DrImpute, and 4) missForest.

Simulations were run using the default parameters of MAGIC in addition to a decay of 10 and 15 and a t of 2 and 3. To rescale the MAGIC imputation a Min-Max scaling function was applied which also allowed us to use the sigmoid function on the output layer of the autoencoder.

In general, the specific architecture for Deep MAGIC can be data dependent. For our experiments, we used an autoencoder with three hidden encoding layers of size 150, 100, and 75 nodes respectively and two hidden decoding layers of size 100, and 150 nodes respectively with an output layer the same size as the input data. The ReLU activation function was applied after each layer except for the output layer where a sigmoid activation function was used. We used the Adam optimizer [19] with a learning rate of $1e-3$ and a batch size of 50. The number of epochs and rules for early stopping are dependent on the dataset. After conducting experiments with $\lambda = 0.3, 0.5, 0.7$, we consistently observed superior performance with $\lambda = 0.7$ across all three datasets. Henceforth, when referring to Deep MAGIC, we use the parameter value of $\lambda = 0.7$

V. Experimental Results

The results from our experiments are given in Tables I, II, and III. In Table I we see that for the sine-cosine manifold dataset Deep MAGIC performs the best at 40% and 60% dropout and is a close runner up to the autoencoder at 80% and 90% dropout. As an example, Figure 1 shows that Deep MAGIC is consistently outperforming all other methods when the dropout rate is 40%. Table II shows that for the MNIST dataset, Deep MAGIC performs the best at all levels of dropout except 20% in which case Deep MAGIC is within 0.004 of the winner, DrImpute. Figure 2 highlights the fact that Deep MAGIC is outperforming the other four imputation methods for every single digit in the MNIST dataset at 60% dropout.

Table III shows that DrImpute is the clear winner when the dropout rate is 60% or less on the Splatter dataset, but Deep MAGIC outperforms the other methods at 80% dropout and isn't far removed from the winner at 90% dropout. Figure 3 shows that Deep MAGIC is consistently better at imputation for the splatter simulated data than the other methods at 80% dropout. Overall Deep MAGIC tends to perform better then the other methods as the rate of dropout increases and often performs comparably at lower dropout rates.

VI. Conclusion

In this paper, we presented an ensemble method called Deep MAGIC that combines the strengths of the MAGIC imputation method and a denoising autoencoder deep learning model to handle missing data. Our experimental results on benchmark datasets demonstrate that the proposed method outperforms several state-of-the-art imputation methods. Specifically, Deep MAGIC achieves superior imputation performance compared to a standard denoising autoencoder in many scenarios, suggesting that the MAGIC regularization provides valuable information to the autoencoder that is not easily learned otherwise.

Our study also reveals that Deep MAGIC is especially effective in handling high levels of dropout, where it consistently outperforms other methods. Moreover, we show that our

TABLE I

MEDIAN MEAN SQUARED ERROR AND INTERQUARTILE RANGE FROM THE SINE-COSINE SIMULATION FROM 10 RUNS. THE TABLE COMPARES THE PERFORMANCE OF AN AUTOENCODER TRAINED TO LEARN THE MAGIC RECONSTRUCTION (AE ON MAGIC), DEEP MAGIC, A STANDARD AUTOENCODER (AE), DRIMPUTE, AND MISSFOREST ON THE SINE-COSINE DATASET WITH MISSING VALUES AT DROPOUT RATES OF 20%, 40%, 60%, 80%, AND 90%. DEEP MAGIC OUTPERFORMS THE OTHER METHODS AT 40% AND 60% DROPOUT RATES AND IS NEAR THE TOP FOR THE OTHER PERCENTAGES.

	20%	40%	60%	80%	90%
AE on MAGIC	0.014 (0.014,0.014)	0.0152 (0.0151,0.0155)	0.0221 (0.0217,0.0237)	0.0462 (0.0402,0.0493)	0.083 (0.079,0.089)
Deep Magic	0.007 (0.007,0.007)	**0.0115 (0.0114,0.0116)**	**0.0163 (0.0162,0.0164)**	0.0225 (0.0219,0.0236)	0.031 (0.03,0.033)
AE	0.007 (0.007,0.007)	0.0119 (0.0118,0.012)	0.0164 (0.0163,0.0165)	**0.0214 (0.0208,0.0218)**	**0.027 (0.027,0.027)**
DrImpute	**0.005 (0.005,0.005)**	0.0172 (0.017,0.0173)	0.0418 (0.0416,0.0422)	0.0841 (0.0837,0.0842)	0.113 (0.113,0.114)
missForest	0.03 (0.0298,0.03)	0.0597 (0.0595,0.0599)	0.0897 (0.0895,0.0899)	0.1196 (0.1195,0.1197)	0.135 (0.134,0.135)

TABLE II

MEDIAN MEAN SQUARED ERROR AND INTERQUARTILE RANGE FROM THE MNIST SIMULATION FROM 800 DIGITS. THE TABLE COMPARES THE PERFORMANCE OF AN AUTOENCODER TRAINED TO LEARN THE MAGIC RECONSTRUCTION (AE ON MAGIC), DEEP MAGIC, A STANDARD AUTOENCODER (AE), DRIMPUTE, AND MISSFOREST ON MNIST DATASET WITH MISSING VALUES AT DROPOUT RATES OF 20%, 40%, 60%, 80%, AND 90%. DEEP MAGIC OUTPERFORMS ALL OF THE OTHER METHODS IN NEARLY ALL OF THE PERCENTAGES.

	20%	40%	60%	80%	90%
AE on MAGIC	0.032 (0.024,0.042)	0.04 (0.029,0.052)	0.048 (0.035,0.062)	0.073 (0.053,0.098)	0.085 (0.061,0.109)
Deep Magic	0.02 (0.014,0.028)	**0.024 (0.017,0.031)**	**0.031 (0.022,0.04)**	**0.044 (0.033,0.056)**	**0.06 (0.046,0.077)**
AE	0.02 (0.014,0.027)	0.026 (0.018,0.034)	0.035 (0.026,0.045)	0.046 (0.036,0.058)	0.065 (0.051,0.081)
DrImpute	**0.016 (0.012,0.02)**	0.024 (0.018,0.032)	0.042 (0.031,0.055)	0.071 (0.052,0.091)	0.089 (0.064,0.112)
missForest	0.021 (0.015,0.028)	0.041 (0.031,0.054)	0.063 (0.047,0.08)	0.085 (0.062,0.107)	0.096 (0.069,0.12)

TABLE III

MEDIAN MEAN SQUARED ERROR AND INTERQUARTILE RANGE FROM THE SPLATTER SIMULATION OVER 20 RUNS. THE TABLE COMPARES THE PERFORMANCE OF AN AUTOENCODER TRAINED TO LEARN THE MAGIC RECONSTRUCTION (AE ON MAGIC), DEEP MAGIC, A STANDARD AUTOENCODER (AE), DRIMPUTE, AND MISSFOREST ON THE SPLATTER DATASET WITH MISSING VALUES AT DROPOUT RATES OF 20%, 40%, 60%, 80%, AND 90%. DEEP MAGIC IS THE BEST AT 80% DROPOUT AND IS CLOSE TO THE BEST WITH 90% DROPOUT.

	20%	40%	60%	80%	90%
AE on MAGIC	8e-04 (8e-04,8e-04)	9e-04 (8e-04,9e-04)	9e-04 (9e-04,9e-04)	0.0011 (0.001,0.0011)	0.0016 (0.0015,0.0018)
Deep Magic	8e-04 (8e-04,8e-04)	8e-04 (8e-04,8e-04)	8e-04 (8e-04,9e-04)	**9e-04 (9e-04,9e-04)**	0.0012 (0.0011,0.0014)
AE	8e-04 (8e-04,8e-04)	8e-04 (8e-04,8e-04)	8e-04 (8e-04,8e-04)	9e-04 (9e-04,0.001)	**0.0011 (0.001,0.0017)**
DrImpute	**2e-04 (2e-04,2e-04)**	**4e-04 (4e-04,4e-04)**	**7e-04 (7e-04,7e-04)**	0.001 (0.001,0.001)	0.0012 (0.0012,0.0012)
missForest	3e-04 (3e-04,3e-04)	6e-04 (6e-04,6e-04)	9e-04 (9e-04,9e-04)	0.0012 (0.0012,0.0012)	0.0014 (0.0013,0.0014)

proposed method is robust to varying dropout rates, making it a versatile solution for a wide range of real-world scenarios.

Future research can extend the application of the proposed method to other types of datasets, including those with more complex and structured missing patterns. Additionally, other deep learning models such as generative adversarial networks (GANs) and variational autoencoders (VAEs) could be explored for missing data imputation using the MAGIC regularization. Overall, our study highlights the effectiveness and potential of Deep MAGIC as a powerful tool for data imputation in a variety of real-world applications.

REFERENCES

[1] A. R. Ismail, N. Z. Abidin, and M. K. Maen, "Systematic review on missing data imputation techniques with machine learning algorithms for healthcare," *Journal of Robotics and Control (JRC)*, vol. 3, no. 2, pp. 143–152, 2022.

[2] L. Yu, R. Zhou, R. Chen, and K. K. Lai, "Missing data preprocessing in credit classification: One-hot encoding or imputation?" *Emerging Markets Finance and Trade*, vol. 58, no. 2, pp. 472–482, 2022.

[3] A. Berchtold, "Treatment and reporting of item-level missing data in social science research," *International Journal of Social Research Methodology*, vol. 22, no. 5, pp. 431–439, 2019.

[4] Y. Wu and K. Zhang, "Tools for the analysis of high-dimensional single-cell rna sequencing data," *Nature Reviews Nephrology*, vol. 16, no. 7, pp. 408–421, 2020.

[5] D. Van Dijk, R. Sharma, J. Nainys, K. Yim, P. Kathail, A. J. Carr, C. Burdziak, K. R. Moon, C. L. Chaffer, D. Pattabiraman *et al.*, "Recovering gene interactions from single-cell data using data diffusion," *Cell*, vol. 174, no. 3, pp. 716–729, 2018.

[6] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[7] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.

[8] L. Gondara, "Medical image denoising using convolutional denoising autoencoders," in *2016 IEEE 16th international conference on data mining workshops (ICDMW)*. IEEE, 2016, pp. 241–246.

[9] A. F. Duque, S. Morin, G. Wolf, and K. Moon, "Extendable and invertible manifold learning with geometry regularized autoencoders," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 5027–5036.

[10] ——, "Geometry regularized autoencoders," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7381–7394, 2022.

[11] A. R. T. Donders, G. J. Van Der Heijden, T. Stijnen, and K. G. Moons, "A gentle introduction to imputation of missing values," *Journal of clinical epidemiology*, vol. 59, no. 10, pp. 1087–1091, 2006.

[12] S. Van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *Journal of statistical software*, vol. 45, pp. 1–67, 2011.

[13] D. J. Stekhoven and P. Bühlmann, "Missforest—non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.

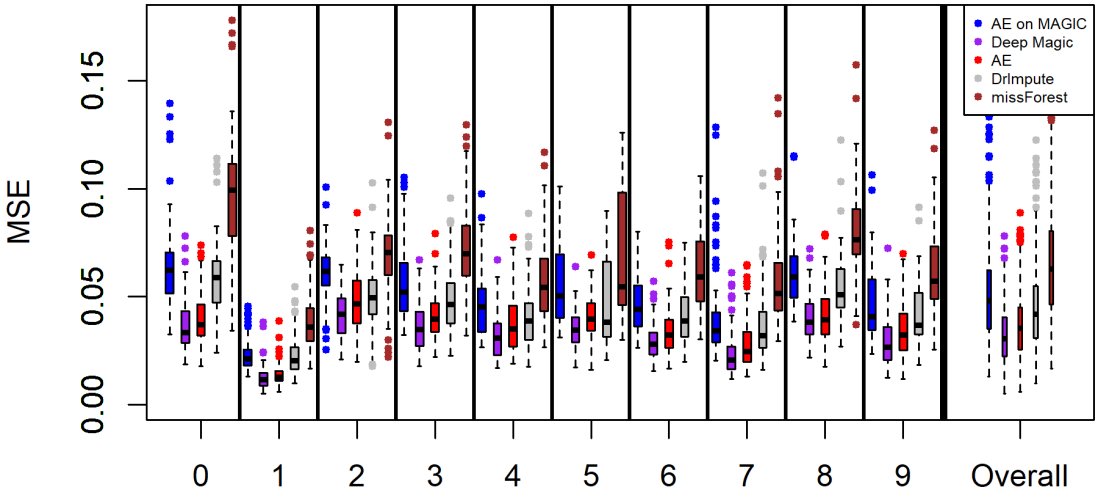

Fig. 2. Boxplot showing the distribution of mean squared error (MSE) abtained from 800 digits obtained from an autoencoder trained on MAGIC (AE on MAGIC), Deep MAGIC, an autoencoder (AE), DrImpute, and MissForest on the MNIST dataset with 60% dropout separated by digits. The results show that Deep MAGIC outperforms the other methods in terms of median MSE and IQR for all digits 0-9 and overall at this dropout rate.

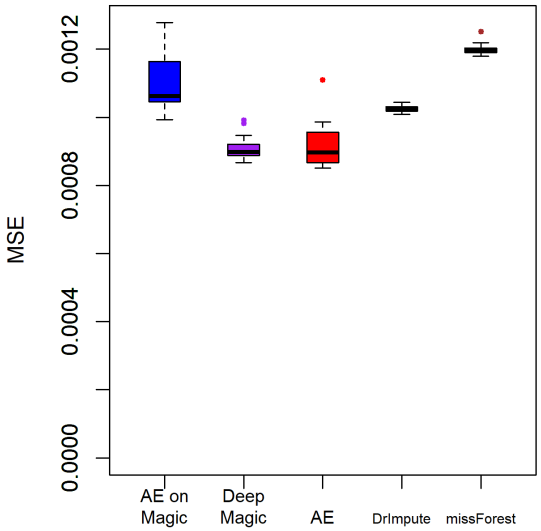

Fig. 3. Boxplot showing the distribution of mean squared error (MSE) values obtained from 20 runs of autoencoder trained on MAGIC (AE on MAGIC), Deep MAGIC, autoencoder (AE), DrImpute, and MissForest on the splatter dataset with 80% dropout. The results show that Deep MAGIC and the AE outperform the other methods in terms of median MSE and IQR at this level of dropout.

[14] Y. Bengio, L. Yao, G. Alain, and P. Vincent, "Generalized denoising auto-encoders as generative models," *Advances in neural information processing systems*, vol. 26, 2013.

[15] J. S. Rhodes, A. Cutler, and K. R. Moon, "Geometry-and accuracy-preserving random forest proximities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[16] W. Gong, I.-Y. Kwak, P. Pota, N. Koyano-Nakagawa, and D. J. Garry, "Drimpute: imputing dropout events in single cell rna sequencing data," *BMC bioinformatics*, vol. 19, pp. 1–10, 2018.

[17] G. Chao, S. Wang, S. Yang, C. Li, and D. Chu, "Incomplete multi-view clustering with multiple imputation and ensemble clustering," *Applied Intelligence*, vol. 52, no. 13, pp. 14 811–14 821, 2022.

[18] L. Zappia, B. Phipson, and A. Oshlack, "Splatter: simulation of single-cell rna sequencing data," *Genome biology*, vol. 18, no. 1, p. 174, 2017.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.