JOURNAL OF OPTOELECTRONICS LASER ISSN:1005-0086

Volume 41 Issue 5, 2022

# Avoiding Blind Spots Of Missing Data With Deep Learning

**Geeta Chhabra**
Dy. Director, Data Informatics & Innovation Division, Ministry of Statistics & PI, Govt. of India, India

**ABSTRACT** Missing data occurs even in a well-designed and controlled research. It not only lowers the statistical power, but also leads to erroneous findings due to biased estimates. This article examines the different types of missing data, the issues that arise because of missing data, and the mechanisms that cause missing data. The paper also review deep learning techniques for handling missing data and compares them with traditional and machine learning algorithms. The paper concludes with recommendations that deep learning methods outperform statistical and machine learning methods.

**Keywords:** Missing Data; Deep Learning Methods; Statistical Methods.

## I. INTRODUCTION

Missing data as the name implies that it is not captured, for a variable of observation. It's a rather prevalent issue in practically all kinds of study. It significantly effects, conclusions that needs, to be drawn from the study. As a result, some study has concentrated solely on dealing with missing data and the challenges it causes. The assumption of a complete data underlies the majority of statistical and other procedures.

Various problems are caused, by missing data. First, the incomplete data will decrease the statistical power, implying that the null hypothesis will be rejected, when it is untrue. Second, there will be biasness in the data due to loss of information, if one drops the incomplete data. Third, it has the potential to reduce the sample's representativeness. Fourth, analysis will be complicated as most of the statistical techniques work on complete dataset. As a result, distorted data may compromise the validity of experiments and lead to inaccurate results (Chhabra et al, 2019).

### Missing Data Patterns & Mechanisms

The missing data pattern describes the data matrix's missing and observed values. In general, “missing data patterns” can be grouped such as univariate, monotone and general. When "missing data" occurs in a single variable, it is referred to as a “univariate pattern”. If a particular variable is missing, it is referred to as a "monotone missing pattern" if that variable is missing forever and a “general pattern” is a data set having an arbitrary missing pattern.

The relation between missing and observed variables in a dataset is defined by missing data mechanisms. It is mainly divided into three categories: “missing completely at random (MCAR)”, “missing at random (MAR)”, and “missing not at random (MNAR)”. If, missingness of a variable is independent of itself, and observed variables, then it is MCAR, for example, the number of fireplaces in a house is independent of itself. If missingness of a variable is related to itself, then is MNAR. For example, if you are going to rent or buy a house, one of the important criteria may be the proximity to the market because it is a great convenience to reach there by walking. If missingness of a variable is dependent on another variable, then it is MAR. For example, if a house does not have a garage, then the garage capacity or quality will naturally be missing (Alruhaymi & Kim, 2021).

## II. LITERATURE REVIEW

The following is brief description of the past research and algorithms used in the field of missing value imputation using deep learning and their implications on the current paper.

**Table 1.** Summary of Various Missing Data Techniques in Deep Learning

| References | Algorithm | Summary | Implications |
|---|---|---|---|
| Mohsen, H., El-Dahshan, E. A., El-Horbaty, E. M., & Salem, A. M. (2018). “Classification using deep learning neural networks for brain tumors”. | DNN with discrete wavelet transformation | Results compared with KNN, LDA and SVM. Classification of 66 MRI into 4 classes | DNN algorithm has better accuracy |
| Cihan, P.(2020). “Deep Learning-Based Approach For Missing Data Imputation”. | Denoising Autoencoder | Compared with kNN and MICE | Deep Learning methods outperforms statistical methods |

| | | | |
|---|---|---|---|
| Lin, W., Tsai, C., & Zhong, J. R. (2022). "Deep learning for missing value imputation of continuous data and the effect of data discretization". | MLP and Deep Belief Network | Compared with svm, cart, kNN and mean. | MLP and DBN performs better than svm, cart, kNN and mean. |
| Phung, S., Kumar, A., & Kim, J. (2019). "A deep learning technique for imputing missing healthcare data". | Denoising Autoencoder | Compared with matrix factorization, KNN, SVD mean & median | Deep Learning method has better accuracy. |
| Jang, J., Choi, J., Roh, H. W., Son, S. J., Hong, C. H., Kim, E. Y., Yoon, D. (2020). "Deep Learning Approach for Imputation of Missing Values in Actigraphy Data". | Denoising Convolutional Autoencoder | mean, Bayesian regression & zero-inflated Poisson regression | Deep learning performed better |
| Wang, W., Yu, H., & Miao, C. (2017). "Deep Model for Dropout Prediction in MOOCs". | CNN and RNN hybrid algorithm | Compared with SVM(with RBF kernel), Decision Tree, Random Forest and Gaussian Naive | CNN & RNN has better accuracy |
| Pereira, R. C., Santos, M. S., Rodrigues, P. P., & Abreu, P. H. (2020). "Reviewing Autoencoders for Missing Data Imputation: Technical Trends, Applications and Outcomes". | Denoising Autoencoders | Compared with statistical methods | Denoising Autoencoders outperform statistical methods |
| Xu, D., Hu, P. J., Huang, T., Fang, X., & Hsu, C. (2020). "A deep learning–based, unsupervised method to impute missing values in electronic health records for improved patient management". | RBM | Compared with KNN, SoftImpute, MICE, and DAE | Proposed algorithm exhibits consistently superior performance |
| Gad, I., Hosahalli, D., Manjunatha, B. R., & Ghoneim, O. A. (2020). "A robust deep learning model for missing value imputation in big NCDC dataset". | Deep learning imputation on a temporal basis for a weather station. | Comparison was performed optimizers "Rmsprop", "Adam", "Nadam", "Stochastic Gradient Descent" and "Adagrad" | "Stochastic Gradient Descent" optimizer is more accurate in predicting missing numbers. |
| Doleck, T., Lemay, D. J., Basnet, R. B., & Bazelais, P. (2019). "Predictive analytics in education: A comparison of deep learning frameworks". | Keras, Theano, Tensorflow, fast.ai, and Pytorch | Compared with k-nearest neighbors, support vector machines, logistic regression and naive bayes classifier. | Predictive accuracy depends on the optimizer. Deep learning displays comparable better performance |

Table 1 (Continued)

| | | | |
|---|---|---|---|
| Gurupur, V. P., Kulkarni, S. A., Liu, X., Desai, U., & Nasir, A. (2018). "Analysing the power of deep learning techniques over the | DLT with two-hidden layer and three-hidden layer network | simple and multiple Linear Regression | DLT has better accuracy. |

| | | | |
|---|---|---|---|
| traditional methods using medicare utilisation and provider data". | | | |
| Yoon, J., Jordon, J., and Van Der Schaar, M. (2018). "Gain: Missing data imputation using generative adversarial nets". | Generative Adversarial Imputation Nets (GAIN). | MICE, MissForest, Auto-encoder,EM | GAIN outperforms state-of-the-art other imputation techniques. |
| Khare, P., Wadhvani, R., & Shukla, S. (2021). "Missing Data Imputation for Solar Radiation Using Generative Adversarial Networks". | Generative Adversarial Networks | Compared with traditional imputation method | GAN has better accuracy. |
| Yang, Y., Wu, Z., Tresp, V., & Fasching, P. A. (2019). "Categorical EHR Imputation with Generative Adversarial Nets". | Generative Adversarial Nets | Compared with traditional imputation method | GAN has better accuracy. |
| Dong, W., Fong, D. Y., Yoon, J., Wan, E. Y., Bedford, L. E., Tang, E. H., & Lam, C. L. (2021). "Generative adversarial networks for imputing missing data for big data clinical research". | Generative adversarial imputation nets (GAIN) | MICE and missForest | GAIN has better accuracy. |

## III. DEEP LEARNING TECHNIQUES FOR MISSING VALUES

Deep learning is a sort of "machine learning" that focuses on the "human brain's" structure and function. Deep learning methods use observations to train machines. Industries such as health, advertising, transport and ecommerce commonly use deep learning. It is based on "neural network", that's constructed like a "human brain," with "artificial neurons" called nodes placed in three layers namely "input layer", "hidden layer(s)" and "output layer" close to each other.

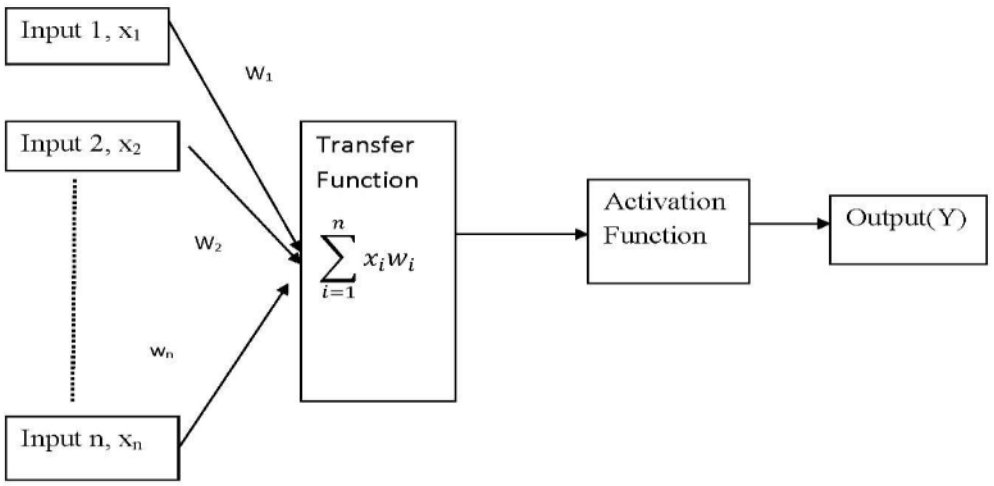


**Figure 1:** Deep Learning Neural Network

Each node receives data as inputs and multiplies them with random weights to generate bias. Finally, the activation function, which is nonlinear, applied to determine output. During the training phase, algorithms use unknown elements in the input distribution to extract features, organise objects, and find relevant data patterns. This happens at various levels, employing the algorithms to develop the models, similar to how self-learning machines are trained. Some of the popular deep learning algorithms for missing value imputations are;

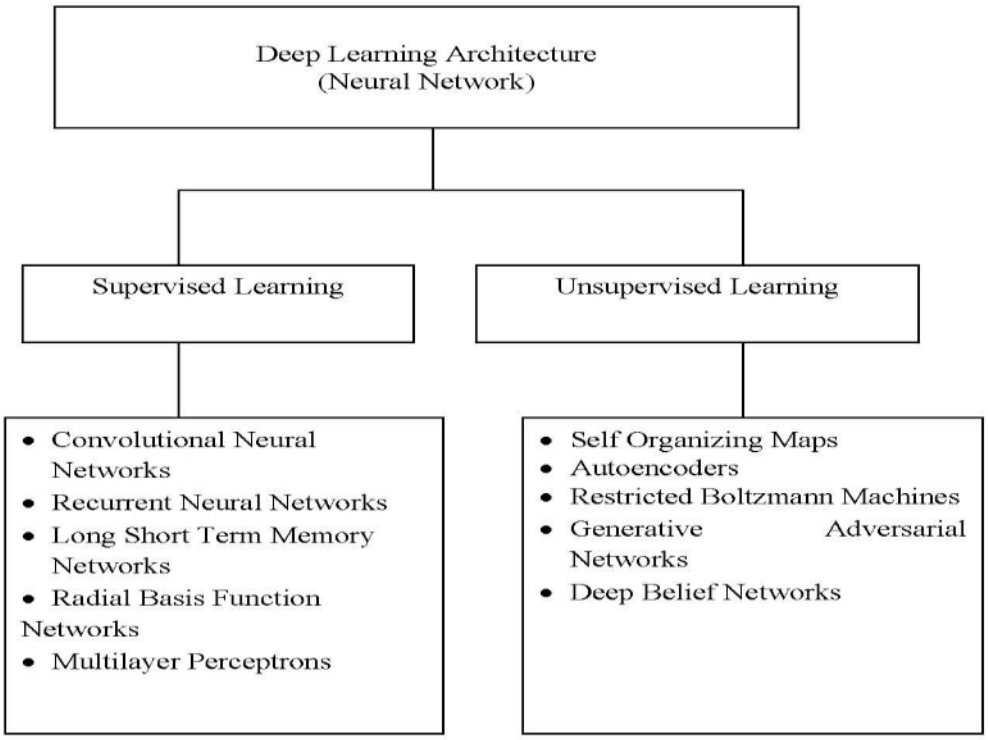


**Figure 2:** Deep Learning Algorithm

**Convolutional Neural Networks (CNNs)** are multilayer neural networks used for “image processing”, “facial recognition” and “object detection”. Miller et al. (2018) used patient-specific patterns contained in the non-missing regions of the waveform to reconstruct missing data in their study "Physiological Waveform Imputation of Missing Data Using Convolutional Autoencoders". They created a generalizable model to assess and extract information from any physiological waveforms using “convolutional neural network (CNN) autoencoders”.

**Recurrent Neural Networks (RNNs)** contain connected cycles that allow the LSTM outputs to be supplied as inputs to the current phase. The LSTM's output becomes an input to the current phase, and its internal memory allows it to remember prior inputs. “Time series analysis”, “natural language processing”, “handwriting identification” and “machine translation” are all common applications of RNNs. In their study "COVID-19: Detailed Analytics & Predictive Modelling using Deep Learning," Dutta et al (2020) used a Recurrent Neural Network to try to predict confirmed cases and fatalities all across the world.

**Long Short Term Memory (LSTMs)** is RNNs that can learn and recall long-term dependencies. The default behaviour is to recall information from the past across lengthy periods of time. LSTMs maintain track of data throughout time. They are useful in “time-series prediction” because they recall previous inputs. In addition to “time-series predictions”, it is extensively used for “speech recognition”, “music production” and “pharmaceutical research”. In their paper "A transfer Learning-Based LSTM strategy for imputing Large-Scale Consecutive Missing Data and its application in a water quality prediction system," Chen et al (2021) addressed the issue that traditional methods aren't suitable for large-scale consecutive missing data problems by introducing the “TrAdaBoost-LSTM” algorithm, which combines machine learning and artificial intelligence.

***Radial Basis Function Networks (RBFNs)*** are a sort of "feed forward neural net work" that uses radial basis functions as activation functions. They have a "input layer," a "hidden layer," and a "output layer" and are used for "classification," "regression," and "time-series prediction". It is used to analyze stock market prices and also forecast sales prices in retail industries because of their ability to work on time series based data. Other applications include speech recognition, time-series analysis, image recognition, medical diagnosis, etc. Smieja et al (2018) in their study “Processing of missing data by neural networks” has used (RBFN) for binary classification and compared with kNN, mean, dropout on UCI dataset.

***Multilayers Perceptrons*** are a sort of "feed forward neural network" that is made up of multiple layers of "perceptrons" with "activation functions." MLPs are composed of two fully interconnected layers: an "input layer" and a

"output layer." They have the same number of "input" and "output" layers, but they can contain a number of "hidden layers" and can be used to develop “voice”, “image”, and “machine translation” software. Cheng et al (2020) used MLP to impute missing data in “attention-deficit/hyperactivity disorder (ADHD)” and evaluated the ability of the imputed dataset in their publication "A Deep Learning Approach for Missing Data Imputation of Rating Scales Assessing Attention-Deficit Hyperactivity Disorder.”

***Self-Organizing Maps (SOMs)*** is used for data visualisation that uses self-organizing artificial neural networks to minimise the dimensions of data. The problem of humans being unable to visualise high-dimensional data is addressed through data visualisation. SOMs are intended to help individuals understand this “multidimensional data”.Nkiaka et al. (2016) used “Self-Organizing Maps (SOMs)” to fill gaps in a “hydro-meteorological” time series data from “the Logone Catchment”, “Lake Chad basin”, in their work "Using self-organizing maps to infill missing data in hydro-meteorological time series from the Logone watershed, Lake Chad basin."

***Autoencoders*** are unsupervised neural networks with identical input and output are They're neural networks that have been trained to repeat data from the “input layer” to the “output layer”. It's utilised for things like drug development, popularity forecasting, and picture processing. In their paper "Effect of Missing Data Imputation on Deep Learning Prediction Performance for Vesicoureteral Reflux and Recurrent Urinary Tract Infection Clinical Study," Kose et al. (2020) used autoencoders to predict vesicoureteral reflux and recurrent urinary tract infection and compared the results to “multiple imputation technique (MICE)” and “Factor analysis of mixed data (FAMD)”. They discovered improved predictive accuracy when they paired the deep learning method with appropriate missing imputation techniques.

***Restricted Boltzmann Machines (RBMs)*** are “stochastic neural networks” that can learn from a “probability distribution” across a set of inputs. Dimensionality reduction, classification, regression, collaborative filtering, feature learning, and modelling are all possible with this deep learning technique. RBMs are the fundamental components of DBNs. In their study "Optimal Predictive Analytics of Pima Diabetics Using Deep Learning," Balaji et al (2017) used “Restricted Boltzmann Machines” and compared them to “rough set theory” on the PIMA diabetic’s dataset, concluding that deep learning models are clearly more effective in terms of precision than “rough set theory”.

***Generative Adversarial Networks(GANs)*** are "deep learning" algorithms that create new data instances that are comparable to the training data. GAN is composed of two parts: a "generator" that learns to produce fictitious data and a "discriminator" that learns from that data. The use of GANs has grown over time. Andrews & Gorell (2020) employed a Generative Adversarial Imputation Network (GAIN) on well production data to impute water or gas rates that are sometimes missing in their study "Generating Missing Unconventional Oilfield Data using a Generative Adversarial Imputation Network (GAIN)".

***Deep Belief Networks (DBNs)*** are generative models composed of several layers of "stochastic" and "latent variables." Binary values are assigned to "latent variables," sometimes known as hidden units. DBNs are a stack of "boltzmann machines" with connections between the levels, and each "RBM" layer communicates with both the previous and subsequent layers. DBNs are utilised for "image identification," "video recognition" and "motion capture data". In their paper "Abnormality Detecting Deep Belief Network," Sharma et al (2016) investigated “Deep Belief Network” for “abnormality detection” and compared its performance in terms of features learnt and abnormality detection accuracy to that of a typical "neural network".

## IV. CONCLUSION & FUTURE WORK

Deep learning is the fastest growing application of “machine learning”. In just few years, its various algorithms have quickly become popular for missing value imputation. Implementation of these algorithms in various domains of research for imputation of missing values shows its utility. The analysis of various publications has been performed in this study which clearly indicates the relevance and growth of deep learning in missing value imputation & its tendency for future research in this field. The deep learning methods for missing value imputation has been compared with various machine learning and statistical method and have been found that deep learning method out performs in terms of accuracy & other metrics.

Deep learning makes use of several models. It has been observed in the literature that no single method is sufficient in all the cases. Some algorithms are better for particular tasks. Their performance depends on the characteristics of the dataset and missing pattern mechanism. They require large amount of data and computing power to solve complicated issues. It has seen that most analysts prefer simple and efficient methods. So, efforts can be made in developing such methods that can handle all kinds of missingness. Hybrid deep learning techniques are also a potential area that needs to be explored.

## REFERENCES

[1] Alruhaymi, A. Z., & Kim, C. J. (2021). Study on the Missing Data Mechanisms and Imputation Methods. Open Journal of Statistics, 11(04), 477–492. https://doi.org/10.4236/ojs.2021.114030.

[2] Andrews, J., & Gorell, S. (2020). Generating Missing Unconventional Oilfield Data using a Generative Adversarial Imputation Network (GAIN). Proceedings of the 8th Unconventional Resources Technology Conference. doi:10.15530/urtec-2020-3014.

[3] Balaji, H., Iyengar, N., & Caytiles, R. D. (2017). Optimal Predictive analytics of Pima Diabetics using Deep Learning. International Journal of Database Theory and Application, 10(9), 47-62. doi:10.14257/ijdta.2017.10.9.05.

[4] Cihan, P. (2020). Deep Learning-Based Approach for Missing Data Imputation. Eskişehir Technical University Journal of Science and Technology B- Theoretical Sciences. 8(2), 336 - 343, DOI: 10.20290/estubtdb.747821.

[5] Cheng, C., Tseng, W., Chang, C., Chang, C., & Gau, S. S. (2020). A Deep Learning Approach for Missing Data Imputation of Rating Scales Assessing Attention-Deficit Hyperactivity Disorder. Frontiers in Psychiatry, 11. doi:10.3389/fpsyt.2020.00673.

[6] Chen, Z., Xu, H., Jiang, P., Yu, S., Lin, G., Bychkov, I., Hmelnov, A., Ruzhnikov, G., Zhu, N., & Liu, Z. (2021). A transfer Learning-Based LSTM strategy for imputing Large-Scale consecutive missing data and its application in a water quality prediction system. Journal of Hydrology, 602, 126573. https://doi.org/10.1016/j.jhydrol.2021.126573.

[7] Chhabra, G., Vashisht, V. Ranjan, J. (2019). A Review on Missing Data Value Estimation Using Imputation Algorithm. Journal of Advance Research in Dynamical & Control Systems, 11(07-special issue), 312–318.

[8] Doleck, T., Lemay, D. J., Basnet, R. B., & Bazelais, P. (2019). Predictive analytics in education: A comparison of deep learning frameworks. Education and Information Technologies, 25(3), 1951-1963. doi:10.1007/s10639-019-10068-4.

[9] Dutta, A., Gupta, A., & Khan, F. H. (2020). COVID-19: Detailed Analytics & Predictive Modelling using Deep Learning. International Journal of Scientific Research in Science and Technology, 95-104. doi:10.32628/ijsrst207517.

[10] Dong, W., Fong, D. Y., Yoon, J., Wan, E. Y., Bedford, L. E., Tang, E. H., & Lam, C. L. (2021). Generative adversarial networks for imputing missing data for big data clinical research. BMC Medical Research Methodology, 21(1). doi:10.1186/s12874-021-01272-3.

[11] Gad, I., Hosahalli, D., Manjunatha, B. R., & Ghoneim, O. A. (2020). A robust deep learning model for missing value imputation in big NCDC dataset. Iran Journal of Computer Science, 4(2), 67-84. doi:10.1007/s42044-020-00065-z.

[12] Gurupur, V. P., Kulkarni, S. A., Liu, X., Desai, U., & Nasir, A. (2018). Analysing the power of deep learning techniques over the traditional methods using medicare utilisation and provider data. Journal of Experimental & Theoretical Artificial Intelligence, 31(1), 99-115. doi:10.1080/0952813x.2018.1518999.

[13] Jang, J., Choi, J., Roh, H. W., Son, S. J., Hong, C. H., Kim, E. Y., Yoon, D. (2020). Deep Learning Approach for Imputation of Missing Values in Actigraphy Data: Algorithm Development Study. JMIR MHealth and UHealth, 8(7). doi:10.2196/16113.

[14] Khare, P., Wadhvani, R., & Shukla, S. (2021). Missing Data Imputation for Solar Radiation Using Generative Adversarial Networks. Proceedings of International Conference on Computational Intelligence Algorithms for Intelligent Systems, 1-14. doi:10.1007/978-981-16-3802-2_1.

[15] Köse, T., Özgür, S., Coşgun, E., Keskinoğlu, A., & Keskinoğlu, P. (2020). Effect of Missing Data Imputation on Deep Learning Prediction Performance for Vesicoureteral Reflux and Recurrent Urinary Tract Infection Clinical Study. BioMed Research International, 2020, 1-15. doi:10.1155/2020/1895076.

[16] Lin, W., Tsai, C., & Zhong, J. R. (2022). Deep learning for missing value imputation of continuous data and the effect of data discretization. Knowledge-Based Systems, 239. doi:10.1016/j.knosys.2021.108079.

[17] Mohsen, H., El-Dahshan, E. A., El-Horbaty, E. M., & Salem, A. M. (2018). Classification using deep learning neural networks for brain tumors. Future Computing and Informatics Journal, 3(1), 68-71. doi:10.1016/j.fcij.2017.12.001.

[18] Miller, D., Ward, A., Bambos, N., Scheinker, D., & Shin, A. (2018). Physiological Waveform Imputation of Missing Data using Convolutional Autoencoders. IEEE 20th International Conference on E-Health Networking, Applications and Services (Healthcom). doi:10.1109/healthcom.2018.8531094.

[19] Nkiaka, E., Nawaz, N.R. & Lovett, J.C. (2016). Using self-organizing maps to infill missing data in hydro-meteorological time series from the Logone catchment, Lake Chad basin. Environmental Monitoring and Assessment. doi: 10.1007/s10661-016-5385-1.

[20] Phung, S., Kumar, A., & Kim, J. (2019). A deep learning technique for imputing missing healthcare data. 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). doi:10.1109/embc.2019.8856760.

[21] Pereira, R. C., Santos, M. S., Rodrigues, P. P., & Abreu, P. H. (2020). Reviewing Autoencoders for Missing Data Imputation:Technical Trends, Applications and Outcomes. Journal of Artificial Intelligence Research, 69, 1255-1285. doi:10.1613/jair.1.12312.

[22] Smieja, M., Struski, Ł., Tabor, J., Zielinski, B., and Spurek, P. Processing of missing data by neural networks(2018). 32nd Conference on Neural Information Processing Systems (NeurIPS), Montréal, Canada.

[23] Sharma, M. K., Sheet, D., & Biswas, P. K. (2016). Abnormality Detecting Deep Belief Network. Proceedings of the International Conference on Advances in Information Communication Technology & Computing - AICTC 16. doi:10.1145/2979779.2979790.

[24] Wang, W., Yu, H., & Miao, C. (2017). Deep Model for Dropout Prediction in MOOCs. Proceedings of the 2nd International Conference on Crowd Science and Engineering - ICCSE17. doi:10.1145/3126973.3126990.

[25] Xu, D., Hu, P. J., Huang, T., Fang, X., & Hsu, C. (2020). A deep learning–based, unsupervised method to impute missing values in electronic health records for improved patient management. Journal of Biomedical Informatics, 111, 103576. doi:10.1016/j.jbi.2020.103576.

[26] Yoon, J., Jordon, J., and Van Der Schaar, M. (2018). Gain: Missing data imputation using generative adversarial nets. Proceedings of the 35th International Conference on Machine Learning. https://proceedings.mlr.press/v80/yoon18a.html.

[27] Yang, Y., Wu, Z., Tresp, V., & Fasching, P. A. (2019). Categorical EHR Imputation with Generative Adversarial Nets. IEEE International Conference on Healthcare Informatics (ICHI). doi:10.1109/ichi.2019.8904717.