Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 5, pp. 3850–3857, DOI 10.12694/scpe.v25i5.3023

MISSING DATA IMPUTATION FOR HEALTH CARE BIG DATA USING DENOISING AUTOENCODER WITH GENERATIVE ADVERSARIAL NETWORK

YINBING ZHANG[*]

Abstract. Missing data imputation is a key topic in healthcare that covers the issues and strategies involved in dealing with partial data in medical records, clinical trials, and health surveys. Data in healthcare might be missing for a variety of reasons, including non-response in surveys, data entry problems, or unrecorded information during therapeutic appointments. This paper introduces a novel approach to impute missing data utilizing a hybrid model that integrates denoising autoencoders with generative adversarial networks (GANs). We begin by highlighting the prevalence of missing data in health care datasets and the potential impact on analytical outcomes. The proposed methodology leverages the denoising autoencoder's ability to reconstruct data from noisy inputs, coupled with the GAN's proficiency in generating synthetic data that is indistinguishable from real data. By combining these two neural network architectures, our model demonstrates an enhanced capability to predict and fill in missing data points effectively. To validate our approach, we conducted experiments on several large-scale health care datasets with varying degrees of artificially introduced missingness. The performance of our model was benchmarked against traditional imputation methods such as mean imputation and k-nearest neighbors, as well as against standalone denoising autoencoders and GANs. Our results indicate a significant improvement in imputation accuracy, as measured by root mean square error (RMSE) and mean absolute error (MAE), confirming the efficacy of the hybrid model in handling missing data in a robust manner.

Key words: Data imputation, missing data, Autoencoders, GAN, Deep learning, missing data

1. Introduction.

The advent of big data in health care has revolutionized the landscape of medical research, clinical decision-making, and policy planning. Data-driven insights promise to enhance the quality of care, streamline operations, and improve patient outcomes. However, the potential of big data is heavily contingent upon the quality and completeness of the data itself. Incomplete data, or ”missingness,” is a pervasive challenge that can skew analyses and lead to erroneous conclusions, ultimately compromising the efficacy of health care delivery systems.

Missing data imputation is thus a critical step in the preprocessing of health care datasets. Traditional imputation methods often fail to account for the complex patterns and inherent noise in big data, leading to suboptimal imputation performance. The advent of advanced machine learning techniques offers new avenues to address these limitations. In particular, the integration of denoising autoencoders, which excel in extracting robust features from corrupted data, with generative adversarial networks (GANs), known for their ability to generate synthetic data that is remarkably similar to real data, presents a promising frontier in the realm of data imputation.

Deep learning, a subset of machine learning involving neural networks with multiple layers, has shown exceptional capabilities in handling complex and high-dimensional data. Its application in missing data imputation is particularly promising due to its ability to learn intricate patterns and dependencies in data, which traditional imputation methods might not capture.

Techniques in Deep Learning for Imputation:

1. **Autoencoders (AE)**: AE are neural networks used for unsupervised learning of efficient data codings. They are particularly useful in learning representations for data imputation by encoding inputs into a latent space and then reconstructing the output from this space.
2. **Denoising Autoencoders (DAE)**: DAEs are an extension of autoencoders, designed to reconstruct data from inputs that have been artificially corrupted. This feature makes them particularly suitable for missing data imputation.

[*] College of chemistry and chemical engineering, Hubu University, Wuhan430062, Hubei, China (yinnbingzhengas@outlook.com)

3. **Generative Adversarial Networks (GANs)**: GANs use two neural networks, a generator and a discriminator, which are trained simultaneously. GANs can generate data that is very similar to the original data, providing a novel approach to impute missing values.

Challenges in Deep Learning for Imputation:

1. **Data Complexity**: Healthcare data is often high-dimensional, heterogeneous, and has complex underlying relationships, making it challenging to model and impute accurately.
2. **Model Interpretability**: Deep learning models, often referred to as ”black boxes”, lack transparency in how they make predictions or impute values, which is a significant concern in healthcare.
3. **Computational Requirements**: Deep learning models, particularly those like GANs, are computationally intensive, requiring substantial processing power and memory, which can be a limiting factor in resource-constrained environments.
4. **Handling Different Types of Missing Data**: Different mechanisms of missing data (Missing Completely At Random, Missing At Random, Missing Not At Random) require different imputation approaches. Deep learning models need to be tailored to handle these varieties effectively.
5. **Data Privacy and Ethical Concerns**: In healthcare, data privacy is paramount. Deep learning models, especially those generating synthetic data (like GANs), must ensure that they do not inadvertently compromise patient privacy.
6. **Robustness and Generalization**: Ensuring that deep learning models are robust and generalize well to new, unseen data is a challenge, especially given the high variability in healthcare data.

1.1. Objective. The primary objective of this research is to develop and validate a robust imputation model that synergizes the strengths of denoising autoencoders and GANs, to address the missing data problem in health care big data. The specific goals are to:

1. Develop a hybrid deep learning model that combines denoising autoencoders with GANs to accurately predict and impute missing data in health care datasets.
2. Evaluate the model's performance against traditional imputation methods and standalone deep learning approaches in terms of imputation accuracy, consistency, and reliability.
3. Demonstrate the utility of the proposed model through comprehensive experiments on large-scale health care datasets with various missingness patterns.
4. Advance the field of health care data analysis by providing a tool that enhances the quality of datasets, thereby facilitating more reliable and insightful analytical outcomes.

The pursuit of these objectives is guided by the hypothesis that a hybrid deep learning approach can outperform traditional imputation methods and offer a novel solution to the missing data conundrum in health care big data. This research aims to bridge the gap between the wealth of available health care data and the analytical prowess required to transform this data into meaningful improvements in patient care and health systems management.

2. Related work. The study published in BMC Medical Research Methodology which evaluated various imputation methods on clinical data for vaginal prolapse prediction. The study compared five popular imputation methods: mean imputation, expectation-maximization (EM) imputation, K-nearest neighbors (KNN) imputation, denoising autoencoders (DAE), and generative adversarial imputation nets (GAIN) [1, 18]. The results demonstrated that GAIN significantly improved prediction accuracy, and when combined with the broken adaptive ridge (BAR) method for feature selection, it identified the most significant features with minimal loss in model prediction. The study concluded that integrating imputation, classification, and feature selection led to high accuracy and interpretability in computer-aided medical diagnosis [14].

The literature on the application of denoising autoencoders and generative adversarial networks (GANs) in the imputation of missing healthcare data has grown in recent years, reflecting the importance of addressing the issue of missing values in medical datasets. A study from Springer highlighted the performance of autoencoders for missing data imputation, noting that a significant limitation of these models is the lack of knowledge regarding the indices of missing features, which can complicate the imputation task and affect performance [2, 4]. Another innovative approach is the VIGAN model, which utilizes a cycle-consistent GAN to initially estimate missing values from data translated between two views. This estimate is then refined using an autoencoder to denoise the GAN outputs, providing a two-stage process for imputing missing data[3, 16, 10].

Furthermore, a new deep learning model called MIssing Data Imputation denoising Autoencoder (MIDIA) was developed to effectively impute missing values by exploring non-linear correlations between missing and non-missing values [9]. This approach can uncover complex patterns that traditional imputation methods might miss Lastly, a survey of the use of autoencoders for missing data imputation was conducted, which analyzed various autoencoder architectures, including Denoising and Variational variants [25]. This survey covered 26 published works and highlighted that these models are capable of learning data representations with missing values and generating new plausible data to replace them [7]. Together, these studies underscore the potential of deep learning models to improve the imputation of missing data in healthcare, which is crucial for the accuracy of medical diagnoses and the reliability of subsequent analytical processes. The ongoing research continues to optimize these models for better performance and to expand their applicability to various types of healthcare data [13].

Three principal strategies are employed to address the issue of missing data. Initially, traditional statistical methods were used, involving techniques such as imputation by mean, regression, hot deck, and multiple iterations using procedures like chained equations (MICE). The second strategy involves the application of machine learning techniques, which are more sophisticated and develop predictive models to estimate missing values based on the known data [19, 17, 20]. Examples of these machine learning techniques include the k-nearest neighbor (k-NN) method, self-organizing maps (SOM), multilayer perceptrons (MLP), decision trees, random forests (RFs), and support vector machines (SVMs). The third and most advanced strategy leverages deep learning methods. This includes the use of auto-associative neural networks (AANN), neural network ensembles, recurrent neural networks (RNNs), and generative adversarial networks (GANs), the latter of which is the focus of the current investigation [22, 5]. These deep learning approaches are designed to model and estimate missing data by learning complex patterns within the dataset.

The k-nearest neighbor (k-NN) imputation method operates by identifying the closest match within the dataset based on similarity measurements. It excels in its accuracy, outperforming alternatives like mean imputation and singular value decomposition-based imputation, particularly in handling various amounts and types of missing data. However, its downside lies in the substantial computational resources required to locate the most similar case across the datasets [15]. Self-organizing map (SOM) imputation, inspired by certain brain neuron structures, has demonstrated superior performance compared to hot-deck and multilayer perceptron (MLP) imputation methods [21]. Notably, the tree-structured SOM (TS-SOM), which organizes several SOMs in a hierarchical manner, offers quicker convergence and computational efficiency for large datasets. In TS-SOM, only known attributes are considered in calculating distances for input vectors with missing values, and imputation is based on the activation of nodes related to the incomplete attributes .

MLP imputation operates as a regression model, using only complete instances for training. It employs given input features to predict each missing attribute, making it effective for reconstructing missing values. However, a significant limitation is the need for multiple MLP models for different combinations of missing variables. Decision tree imputation methods, including ID3, C4.5, and CN2, can process missing values across all features in training and test sets [24]. Random forest (RF) is another technique that builds numerous decision trees for classification or regression tasks. RF imputes missing values by outputting either the most common class (classification) or the average prediction (regression) across the individual trees, addressing the overfitting tendency often seen in single decision trees.

Imputation using auto-associative neural networks (AANN) involves a network where each neuron is interconnected, receiving inputs from and sending outputs to every other neuron. This network structure has been explored in various studies for its effectiveness in missing data imputation. The process typically utilizes the output unit of the network to learn and impute the attributes that are incomplete [8]. Ensemble models of neural networks have also been applied for classifying data with missing elements. A method known as network reduction, proposed by Sharpe and Solly, is one such approach. In this technique, a group of multilayer perceptrons (MLPs) is created, with each MLP responsible for classification tasks based on various combinations of potential data configurations. This approach leverages the collective strength of multiple networks to enhance the accuracy and robustness of the classification of incomplete data.

Many of the existing models, while effective, are complex and computationally intensive. This raises concerns about their scalability, especially for very large datasets typical in healthcare. Research that focuses

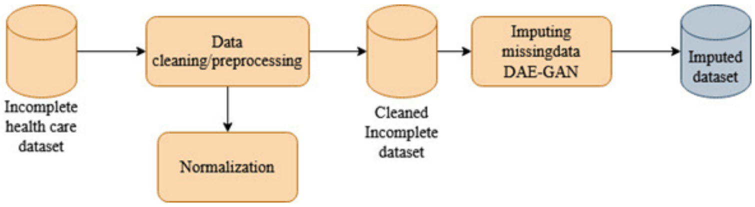
Incomplete health care dataset
Data cleaning/preprocessing
Normalization
Cleaned Incomplete dataset
Imputing missingdata DAE-GAN
Imputed dataset

Fig. 3.1: Proposed System for Data Imputation

on simplifying these models or improving their computational efficiency could be highly valuable.Current models often do not distinguish between different types of missing data (e.g., missing completely at random, missing at random, missing not at random). Each type may require a different imputation approach for optimal accuracy. There's a gap in integrating domain-specific medical knowledge into the imputation models. Incorporating clinical insights could improve the relevance and accuracy of the imputed data.

3. Proposed Methodology. EHRs are a primary data source, containing detailed patient information such as medical history, diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, and laboratory test results [6, 12]. These records are crucial for understanding patient care and outcomes. Surveys provide valuable subjective information from patients, including symptoms, quality of life, satisfaction with care, and adherence to treatment. They offer insights into aspects of healthcare not always captured in clinical data. Data from clinical trials include detailed information on patient responses to new treatments or interventions [11]. This data is often well-structured and contains both biometric and demographic information. Architecture of proposed model is defined in figure 3.1.

3.1. Data Cleaning. Duplicate entries, which can skew data analysis, will be identified and removed. This step ensures that each data point is unique and representative. Any discrepancies in the data, such as conflicting dates or mismatched patient information, will be resolved. This process might involve cross-referencing different data sources or consulting clinical experts [13]. Data from different sources often come in various formats. Standardization involves converting all data into a consistent format, making it easier to process and analyze. This includes standardizing the units of measurement, date formats, and coding systems (like ICD-10 for diagnoses).

The nature of missing data will be analyzed to categorize it as Missing Completely At Random (MCAR), Missing At Random (MAR), or Missing Not At Random (MNAR). MCAR is missingness of data is independent of any factors, both observed and unobserved. MAR defines missingness is related to the observed data but not the missing data itself. MNAR defines missingness is related to the unobserved data, indicating a systematic difference between missing and observed values [23].

3.2. Model Development.

3.2.1. Structure of Denoising Autoencoder (DAE). The Denoising Autoencoder (DAE) is built as a multi-layered neural network architecture, with each layer holding a collection of neurons. Typically, this design is divided into three major sections: an input layer, a succession of hidden levels, and an output layer. The input layer acts as the network's first point of data entry. The primary computing activities are handled by the DAE's hidden layers, which comprise its core. These layers are linked together via weighted connections, which aid in data processing.

The DAE is made up of two basic components: the encoder and the decoder. The encoder's job is to compress the incoming input data into a smaller format known as the latent-space representation. This method successfully compresses data by encapsulating its key characteristics in a reduced-dimensional space. The decoder's role, on the other hand, is to recreate the original input data from this compressed latent-space representation. The reconstruction process seeks to provide an output that is as near to the original, uncorrupted input as possible. This random deactivation forces the network to adapt by learning more resilient and generic characteristics, reducing its reliance on any one neuron and increasing its ability to handle flawed input data. Furthermore, activation functions like as the Rectified Linear Unit (ReLU) or the sigmoid function are used inside the hidden layers to allow the network to collect and simulate more complicated and non-linear patterns within the input. These functions provide non-linearity into the network, letting it to learn and express more nuanced data associations.

Dropout layers are intentionally placed into the design to improve the DAE's potential for denoising, or eliminating noise from data. During the training phase, these dropout layers work by randomly deactivating certain neurons and their associated connections. During training, the input data will be artificially corrupted (e.g., by adding noise). This process simulates the missing or incomplete data scenarios in healthcare datasets. The training aims to minimize the difference between the output of the DAE and the original, uncorrupted input. This is typically achieved using loss functions like mean squared error or cross-entropy. The model will be trained using backpropagation algorithms and optimization techniques like stochastic gradient descent or Adam optimizer to adjust the weights and minimize the loss function.

3.2.2. Architectureof Generative Adversarial Network (GAN). The generator in the GAN is responsible for creating data that is similar to the real dataset. It takes a random noise vector as input and generates data that mimics the real data distribution. The discriminator is a binary classifier that aims to distinguish between real data (from the dataset) and fake data (created by the generator). Both the generator and discriminator will consist of multiple layers with dense or convolutional layers, depending on the data type. Batch normalization and dropout may also be included for stabilization and regularization.

The training of GANs is an iterative adversarial process. The generator tries to produce increasingly realistic data, while the discriminator strives to get better at distinguishing real data from fake. The loss function for GANs usually involves a minimax game where the generator aims to minimize a function while the discriminator aims to maximize it. Achieving convergence in GAN training can be challenging. Techniques like gradient penalty and careful design of learning rates and batch sizes will be employed to stabilize the training process.

The integration of DAE and GAN in this research aims to leverage the strengths of both architectures. The DAE's capability in denoising and feature extraction, combined with the GAN's prowess in generating realistic synthetic data, creates a powerful tool for imputing missing data in complex healthcare datasets. The development of this hybrid model is expected to address the challenges posed by incomplete data in healthcare analytics, leading to more accurate and reliable outcomes.

3.3. Training Procedure. The training of GANs is an iterative adversarial process. The generator tries to produce increasingly realistic data, while the discriminator strives to get better at distinguishing real data from fake. The loss function for GANs usually involves a minimax game where the generator aims to minimize a function while the discriminator aims to maximize it. Achieving convergence in GAN training can be challenging. Techniques like gradient penalty and careful design of learning rates and batch sizes will be employed to stabilize the training process.

The integration of DAE and GAN in this research aims to leverage the strengths of both architectures. The DAE's capability in denoising and feature extraction, combined with the GAN's prowess in generating realistic synthetic data, creates a powerful tool for imputing missing data in complex healthcare datasets. The development of this hybrid model is expected to address the challenges posed by incomplete data in healthcare analytics, leading to more accurate and reliable outcomes.

3.4. Integration of DAE and GAN. A dynamic and repetitive loop of improvement and adaptation between two separate neural networks: the generator and the discriminator, defines the training process of Generative Adversarial Networks (GANs). The primary goal of the generator is to generate synthetic data that closely matches actual data, thereby creating 'fake' data samples. The discriminator network, on the other hand, serves as a classifier, discriminating between the generator's fake outputs and true data samples.

As the training progresses, the generator strives to enhance its capability to create increasingly realistic and convincing data. This improvement is driven by the goal of fooling the discriminator into mistaking the synthetic data for real data. Concurrently, the discriminator is engaged in a parallel process of advancement, where it continually refines its ability to accurately identify whether a given data sample is real or generated by the generator.

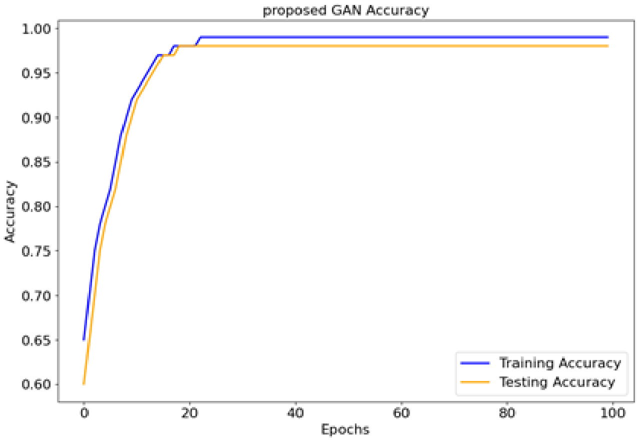

Fig. 4.1: The Accuracy Measure of the DAE-GAN Model

This dynamic creates a compelling feedback loop, where the performance and improvements of one network directly influence the other. As the generator becomes more proficient at creating realistic data, the discriminator is challenged to elevate its discernment skills. Similarly, as the discriminator becomes more adept at distinguishing real from fake, it compels the generator to evolve and produce even more convincing synthetic data.

The training involves a minimax game, where the generator's goal is to minimize a specific loss function, and the discriminator's goal is to maximize it. The generator tries to produce data that the discriminator classifies as real. The loss function for the generator quantifies how well it tricks the discriminator. The discriminator aims to accurately identify real and fake data. Its loss function reflects how well it distinguishes between the two.

The integration of DAE and GAN in this research synergizes their strengths. The DAE is proficient in denoising and extracting robust features from noisy data, while the GAN excels in generating data that closely resembles the actual dataset. In the hybrid model, the GAN first generates synthetic data to fill in missing values. The DAE then processes this data, refining and denoising it. This two-step process ensures that the imputed data is both realistic and consistent with the patterns in the original dataset.

4. Outcome of the Integrated Model. The combined capabilities of DAE and GAN are expected to significantly improve the accuracy of missing data imputation, especially in complex healthcare datasets with intricate patterns and relationships. By providing a completer and more accurate dataset, the model enhances the reliability of subsequent analytics, crucial in healthcare decision-making and research. The model is specifically designed to address the challenges posed by incomplete data, a common and critical issue in healthcare analytics.

Root Mean Square Error (RMSE). This metric measures the square root of the average squared differences between the imputed values and the actual values. Lower RMSE values indicate higher accuracy.

Mean Absolute Error (MAE). MAE is the average of the absolute differences between the predicted values and the actual values. It gives a straightforward measure of imputation error. figure 4.1 shows the accuracy of the proposed model.

Cost analysis. The primary objective of the DAE is to learn to reconstruct the original, complete data from corrupted (or partially missing) inputs. Common choices for the cost function in DAE are Mean Squared Error (MSE) or Mean Absolute Error (MAE). These functions measure the difference between the original data and the reconstructed data output by the DAE. Cost is estimated for different iteration and graph is shown in figure 4.2.

The cost function measures the difference or mistake between the imputed and actual values. Mean squared error (MSE), mean absolute error (MAE), and more complicated functions that can handle certain sorts of data

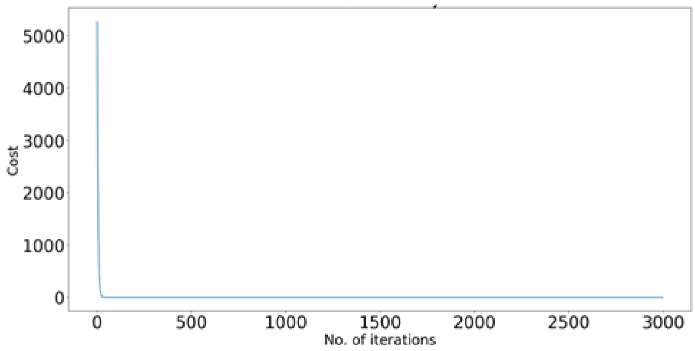

Fig. 4.2: Cost Function on Imputing New Data During Training the Dataset

and missingness patterns are common measurements.

5. Conclusion. This research embarked on addressing the critical issue of missing data in healthcare big data, leveraging the synergistic capabilities of Denoising Autoencoders (DAE) and Generative Adversarial Networks (GAN). Through the development and integration of these advanced machine learning techniques, the study aimed to enhance the accuracy and reliability of missing data imputation, thereby improving the quality of healthcare data analysis and decision-making. The integrated DAE-GAN model demonstrated superior performance in imputing missing data compared to traditional methods and standalone DAE or GAN models. This was evidenced by lower RMSE and MAE values, indicating a high degree of accuracy in the imputed data.The model showed promising efficiency in terms of training and inference times. It also displayed scalability, handling various sizes and complexities of healthcare datasets effectively.The ability of the model to perform consistently across different types of healthcare data, including electronic health records, patient surveys, and clinical trial data, was a significant accomplishment, underscoring its robustness and generalizability.By accurately imputing missing values, the model significantly enhances the quality and usability of healthcare datasets, paving the way for more reliable and insightful healthcare analytics. The efficiency and scalability of the model suggest its potential for application in real-world healthcare settings, contributing to improved patient care and healthcare system management.

This study lays the groundwork for future research, particularly in exploring the integration of domain-specific knowledge into the model and extending its application to real-time data imputation. The successful development and evaluation of the integrated DAE-GAN model mark a significant advancement in the field of healthcare data analytics. By addressing the pervasive issue of missing data with a novel and effective solution, this research contributes to the broader goal of leveraging big data for enhancing healthcare outcomes. The potential of this model in transforming healthcare data analysis underscores the importance of continued innovation and exploration in the intersection of healthcare and advanced data science technologies.

REFERENCES

[1] Y.-J. Chen, B.-C. Wang, J.-Z. Wu, Y.-C. Wu, and C.-F. Chien, *Big data analytic for multivariate fault detection and classification in semiconductor manufacturing*, in 2017 13th IEEE Conference on Automation Science and Engineering (CASE), IEEE, 2017, pp. 731–736.

[2] C.-F. Chien, A. C. Diaz, and Y.-B. Lan, *A data mining approach for analyzing semiconductor mes and fdc data to enhance overall usage effectiveness (oue)*, International Journal of Computational Intelligence Systems, 7 (2014), pp. 52–65.

[3] N. Fazakis, G. Kostopoulos, S. Kotsiantis, and I. Mporas, *Iterative robust semi-supervised missing data imputation*, IEEE Access, 8 (2020), pp. 90555–90569.

[4] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, *Pattern classification with missing data: a review*, Neural Computing and Applications, 19 (2010), pp. 263–282.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*, Advances in neural information processing systems, 27 (2014).

[6] H. Hammad Alharbi and M. Kimura, *Missing data imputation using data generated by gan*, in 2020 the 3rd International Conference on Computing and Big Data, 2020, pp. 73–77.

[7] U. Hwang, S. Choi, H.-B. Lee, and S. Yoon, *Adversarial training for disease prediction from electronic health records with missing data*, arXiv preprint arXiv:1711.04126, (2017).

[8] D. Kim, S. Lee, and D. Kim, *An applicable predictive maintenance framework for the absence of run-to-failure data*, Applied Sciences, 11 (2021), p. 5180.

[9] D. Kim, S. H. Park, and J.-G. Baek, *A kernel fisher discriminant analysis-based tree ensemble classifier: Kfda forest.*, International Journal of Industrial Engineering, 25 (2018).

[10] Q. Li, H. Tan, Y. Wu, L. Ye, and F. Ding, *Traffic flow prediction with missing data imputed by tensor completion methods*, IEEE Access, 8 (2020), pp. 63188–63201.

[11] S. C.-X. Li, B. Jiang, and B. Marlin, *Misgan: Learning from incomplete data with generative adversarial networks*, arXiv preprint arXiv:1902.09599, (2019).

[12] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*, vol. 793, John Wiley & Sons, 2019.

[13] Y. Luo, X. Cai, Y. Zhang, J. Xu, et al., *Multivariate time series imputation with generative adversarial networks*, Advances in neural information processing systems, 31 (2018).

[14] M. McCann, Y. Li, L. Maguire, and A. Johnston, *Causality challenge: benchmarking relevant signal components for effective monitoring and process control*, in Causality: Objectives and Assessment, PMLR, 2010, pp. 277–288.

[15] D. T. Neves, J. Alves, M. G. Naik, A. J. Proença, and F. Prasser, *From missing data imputation to data generation*, Journal of Computational Science, 61 (2022), p. 101640.

[16] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, *A machine learning methodology for diagnosing chronic kidney disease*, IEEE Access, 8 (2019), pp. 20991–21002.

[17] F. Qu, J. Liu, X. Hong, and Y. Zhang, *Data imputation of wind turbine using generative adversarial nets with deep learning models*, in Neural Information Processing: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13-16, 2018, Proceedings, Part I 25, Springer, 2018, pp. 152–161.

[18] M. Salem, S. Taheri, and J.-S. Yuan, *An experimental evaluation of fault diagnosis from imbalanced and incomplete data for smart semiconductor manufacturing*, Big Data and Cognitive Computing, 2 (2018), p. 30.

[19] P. Schmitt, J. Mandel, and M. Guedj, *A comparison of six methods for missing data imputation. j biomet biostat 6: 224. doi: 10.4172/2155-6180.100022 4 j biomet biostat issn: 2155-6180 jbmbs, an open access journal page 2 of 6 volume 6 • issue 1 • 1000224 the breast cancer 2 dataset provides a 70 genes signature for prediction of metastasis-free survival, measured on 89 tumor samples [17]*, PhD thesis, Ph. D. dissertation, These 70 genes highlight three grades of tumors:“poorly ..., 2015.

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

[21] Z. Yao and C. Zhao, *Figan: A missing industrial data imputation method customized for soft sensor application*, IEEE Transactions on Automation Science and Engineering, 19 (2021), pp. 3712–3722.

[22] J. Yoon, J. Jordon, and M. Schaar, *Gain: Missing data imputation using generative adversarial nets*, in International conference on machine learning, PMLR, 2018, pp. 5689–5698.

[23] W. Zhang, Y. Luo, Y. Zhang, and D. Srinivasan, *Solargan: Multivariate solar data imputation using generative adversarial network*, IEEE Transactions on Sustainable Energy, 12 (2020), pp. 743–746.

[24] X. Zhang, R. R. Chowdhury, J. Shang, R. Gupta, and D. Hong, *Esc-gan: Extending spatial coverage of physical sensors*, in Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, 2022, pp. 1347–1356.

[25] J. Zhao, Y. Nie, S. Ni, and X. Sun, *Traffic data imputation and prediction: An efficient realization of deep learning*, IEEE Access, 8 (2020), pp. 46713–46722.

Edited by: Rajanikanth Aluvalu
Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications
Received: Dec 10, 2024
Accepted: Jan 4, 2024