JOURNAL OF DATA SCIENCE | Vol.2026, Issue 3, No.15
eISSN:2805-5160

Missing Data Imputation Using Hybrid Generative Models for Improving Classification Performance

Wahyu Purnama Sari[1]*, Shah Khadafi[2], Engelbert Q. Cruz[3]

[1]Universitas Langlangbuana, Indonesia
[2]Information System Department, Institut Teknologi Adhi Tama Surabaya, East Java, Indonesia
[3]College of Information Technology Education, PHINMA – University of Pangasinan, Philippines

Email: wahyu.alyapurnama@gmail.com[1]*, khadafi@itats.ac.id[2], eqcruz.up@phinmaed.com[3]

Abstract

Missing values remain a major challenge in data preprocessing because they can distort feature distributions and reduce the reliability of downstream classification models. Conventional imputation methods, such as mean substitution and K-nearest neighbors, are simple to apply but often fail to preserve complex nonlinear relationships among variables. This study proposes a hybrid generative imputation framework that combines a Variational Autoencoder and a Generative Adversarial Imputation Network to reconstruct missing values in incomplete tabular datasets. The experiment was conducted using the Breast Cancer Wisconsin (Diagnostic) dataset, which contains 569 instances, 30 real-valued features, and two classification classes. Missing values were simulated using a missing completely at random mechanism at 10%, 20%, 30%, and 40% missingness rates. Imputation quality was evaluated using root mean square error, while downstream classification performance was assessed using accuracy, precision, recall, and F1-score. The results show that regression-based imputation achieved the lowest RMSE across all missingness rates, while the proposed VAE-GAIN framework achieved competitive classification performance and produced the highest F1-score at the 40% missingness rate. These findings indicate that hybrid generative imputation can support classification reliability under severe incomplete data conditions, although its reconstruction accuracy may vary across missingness levels. This study contributes to data preprocessing research by evaluating missing data imputation from both reconstruction quality and downstream classification perspectives.

Keywords

Missing data imputation; Variational Autoencoder; Generative Adversarial Imputation Network; Classification performance; process innovation.

Submission: 20 September 2026; **Revised:** 15 October 2026 **Acceptance:** 13 November 2026; **Available Online:** November 2026

Copyright: © 2026. All the authors listed in this paper. The distribution, reproduction, and any other usage of the content of this paper is permitted, with credit given to all the author(s) and copyright owner(s) in accordance to common academic practice. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license, as stated in the website: https://creativecommons.org/licenses/by/4.0/

Introduction

The absence of data is a key challenge in data preprocessing and in many machine learning tasks. It is common to encounter incomplete data in real-world data sets due to a number of reasons including: personal data not being fully entered; an anomaly in data collecting systems and devices; errors in data processing and merging; and variations in data collection methods. Missing values can introduce biases in parameter estimates and interpretation of findings, and when not addressed appropriately, they can lead to misrepresentation of the phenomena under study.ion methods. Missing values are often brushed aside as technical issues, but they have a non-negligible impact on data sets as they affect their completeness and representativeness.

The impact of missing values is greater in classification tasks. This is because classification models rely on an entire pattern of features in order to differentiate classification boundaries. Missing values can sever the relationship of input features with their respective target labels, affect the integrity of the training process, and yield skewed results. (Ciná et al., 2023) Incomplete data sets can result in classification models experiencing poor performance in accuracy, recall, and issues with a stable and reliable F1-score. (Palanivinayagam & Damaševičius, 2023) This is evident when the most relevant features contain missing values. Thus, improving the quality of data and the overall performance of classification tasks relies on effective strategies for addressing missing values.

Classical methods to manage missing data employ a few commonly used strategies. Mean, or median imputation, can be quick and easy, though it can work against data variability and be harmful to the original feature distribution. K-nearest neighbors imputation considers surrounding data points to fill in missing values, but it can be sensitive to the defined distance measure and can be less reliable when missingness is pronounced. Regression-based imputation can attempt to model the relationships between variables to fill in missing values. However, it can be limited by datasets that contain nonlinear relationships, complex interactions, and a variety of heterogeneous data. The limitations of the conventional methods of imputation show they can be insufficient to account for the complex relationships of data in classification-based datasets.

An advantage of using generative models for imputation is that the goal of these models is understanding the data distribution of the underlying data and generating values based on the learned distribution. Variational Autoencoders can learn hidden patterns and complete missing data. On the other hand, Generative Adversarial Imputation Networks can be used to improve imputation by learning to differentiate between observed and generated data. (Ghojogh & Ghodsi, 2026) While latent space learning with generative models can provide high-quality imputation, the use of adversarial learning can provide an improvement in the quality of imputation. It can be the case that the generative and adversarial learning techniques can provide a better quality of imputation when compared to other techniques, such as statistical techniques. According to Randhawa et al. (2026), the use of CTGAN to augment a constrained CKD dataset was able to synthesize realistic samples, preserve the statistical distribution and the feature distribution, allow the models to generalize better, and improve the stability of the models, thus providing a good alternative to alleviate overfitting in the data-scarce medical domain of chronic kidney disease. (Linville et al., 2019)

There is a notable lack of research focused on hybrid generative imputation and its effect on classification within the context of generative models and data reconstruction. (Berahmand et al., 2024). Existing studies on imputation largely focus on reconstructions that minimize error, with classification performance often receiving scant focus. (Nguyen, 2023) However, in the field

of data science, the quality of imputation should be assessed in relation to the extent that the completed dataset offers improved predictive performance rather than how close the imputed values are to the original values. For this reason, it is necessary to assess imputation and classification within a single framework.

The goal of this research is to present hybrid generative imputation with an emphasis on the classification of incomplete tabular datasets. (Salmi et al., 2024). This approach integrates the Variational Autoencoder (VAE) and the Generative Adversarial Imputation Network (GAIN). In this context, the role of the VAE is to capture latent data representations and provide initial imputed values, while the generative adversarial framework of the GAIN is used to maintain the current patterns in the data. In the completed datasets, classification models are built to understand the merit of this approach in enhancing classification performance.

This study presents three advancements. First, we present a hybrid model for generative adversarial missing data imputation. Second, we examine the robustness of the proposed model under a variety of missingness conditions. Finally, we analyze the effect of imputation quality on classification performance using accuracy, precision, recall, and F1 scores. The proposed framework is designed to support the development of robust preprocessing methods for classification-centric data science.

Methodology

This study evaluated the effectiveness of hybrid generative models for missing data imputation and their impact on the downstream classification performance. An iterative and experimental data science workflow was adopted, including data preparation, missing value imputation using hybrid generative models, establishing baseline models, building classification models, and evaluating classification performance. The focus of this study was to evaluate the impact of hybrid generative models on missing data reconstruction and classification performance as compared to traditional imputation models.

Dataset preparation

The experiments utilized the UCI Machine Learning Repository's Breast Cancer Wisconsin (Diagnostic) dataset, which includes 30 attribute fields, 569 records, and two target classes for benign and malignant instances (Thabit, 2023). Since the dataset is free of missing values, it is ideal for controlled missing data simulation, as the complete original dataset may serve as a reference to gauge the imputation error.

Prior to model training, the target variable was removed from the data features. To allow for comparison across all feature variables, all of the numeric features were min-max normalized. The data was stratified so that 80% of the data was set aside for training and 20% was set aside for testing while maintaining the original class distribution. The training set was then used to build the imputation and classification models, while the testing set was used to assess the quality of reconstruction and classification provided by the models.

The details of the experimental setup, including the dataset, setting of missingness, imputation methods, classifiers, and evaluation metrics, used in this study are provided in Table 1.

Table 1. Experimental configuration

Component	Description
Dataset	Breast Cancer Wisconsin (Diagnostic)
Source	UCI Machine Learning Repository / scikit-learn benchmark dataset
Task	Binary classification
Number of instances	569
Number of features	30 real-valued features
Number of classes	2
Class labels	Malignant and benign
Original missing values	None
Missingness mechanism	Missing completely at random
Missingness rates	10%, 20%, 30%, and 40%
Data split	80% training and 20% testing
Scaling method	Min-max normalization
Baseline imputation methods	Mean imputation, KNN imputation, and regression-based imputation
Proposed method	VAE-GAIN hybrid generative imputation
Classifier evaluator	Random Forest
Imputation metric	Root mean square error
Classification metrics	Accuracy, precision, recall, and F1-score

For baseline imputation methods, as shown in Table 1, all imputation methods were tested in the same dataset, missingness, train-test split, and classification setting (Shadbahr et al., 2023).

Missing data simulation

In order to assess the flexibility of the framework that was developed, several different conditions of missing data were assessed by incorporating missing values at different missingness levels of 10%, 20%, 30%, and 40%. The missing values were incorporated using a missing completely at random method, in which there was an equal chance for all feature values to be removed. This provided a controlled environment in which the imputed values could be tested against the original observed values (Shadbahr et al., 2023).

Simulated missingness offered a consistent framework within which to evaluate the proposed method and established baseline imputation techniques. Because the values were removed in a controlled manner, it became straightforward to assess the imputation error. This approach also facilitated the evaluation of the performance of the various imputation techniques as the missing data increased.

Proposed hybrid generative model

The proposed framework integrates two components of generative learning: Variational Autoencoders and Generative Adversarial Imputation Networks. The first component employs Variational Autoencoders to learn latent data representations and generate preliminary values for

missing data. The adversarial imputation component then modifies the imputed values by learning the observed patterns of features and the generated gap. The framework's two-step design focuses on maintaining the intended global structure of the data and consistency at the feature level (Kasar et al., 2024).

Figure 1 shows the workflow of the proposed hybrid generative imputation framework. The flow shows that latent representation of the data and adversarial imputation are used to iteratively refine incomplete data for the purpose of classification.

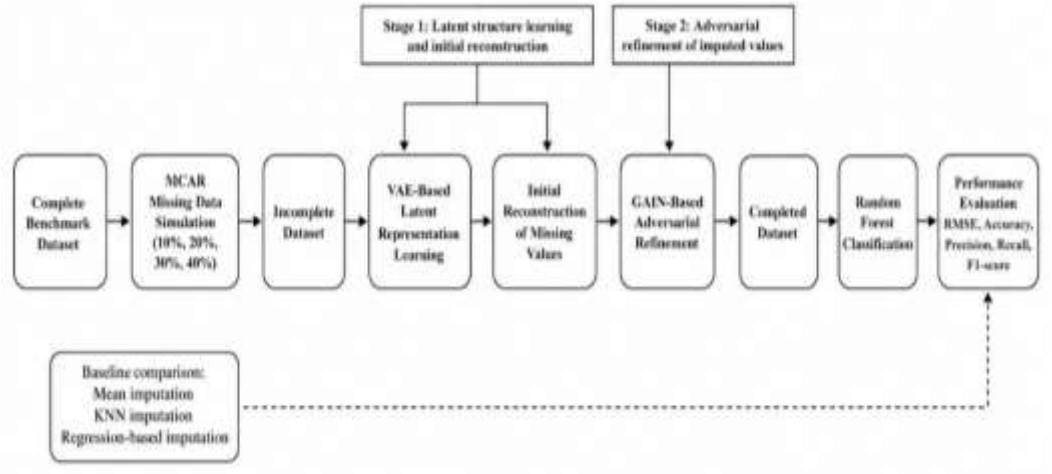


Figure 1. Proposed VAE-GAIN hybrid generative imputation framework.

The proposed framework combines the two generative stages of the imputation process. As shown in Figure 1, the first stage, represented by the Variational Autoencoder, is aimed at the latent structure of the data (Ahmed et al., 2023). The second stage in the proposed framework is the adversarial imputation component that enhances the imputed data for classification purposes.

VAE component

The proposed imputation framework employs the Variational Autoencoder as the first stage. The encoder of the Variational Autoencoder learns to represent the data in a lower dimension, and the decoder reconstructs the feature space. For the model to process the data, a temporary value is assigned to the missing data. During the imputation process, a missingness mask is applied to differentiate between the observed and missing data (Mumuni & Mumuni, 2024).

By training the VAE to minimize reconstruction loss and to control the latent distribution, the model discovered relationships among the features and generated preliminary estimates for missing values. The decoder's output was the first-stage imputed dataset, which was then subjected to further processing by the adversarial imputation module.

GAIN/adversarial component

The adversarial imputation module was designed to enhance the initial reconstruction from the VAE. Based on the principles of adversarial learning, in this module, the generator's goal was to produce realistic imputed values, while the discriminator was tasked with identifying observed

values and imputed values (Darban et al., 2024). The missingness mask directed the adversarial learning to distinguish between the observed values and generated values (Berahmand et al., 2024).

The aim of this stage was to improve the realism of the imputed values and to minimize the distributional gap between the observed data and the reconstructed data (Oquab et al., 2023). While the latent structure was captured by the VAE, the adversarial component was designed to align the imputed values with the original data distribution. The output of this stage was a filled dataset for further classification.

Implementation settings of the proposed model

To implement the proposed VAE-GAIN framework, a consistent experimental configuration was developed. The input dimension of the model was determined by the number of normalized features in the Breast Cancer Wisconsin (Diagnostic) dataset, which has 30 input features. The Variational Autoencoder had an encoder, a latent space, and a decoder. The encoder compressed the input feature vector to a lower-dimensional latent space, and the decoder performed the opposite to reconstruct the feature values. To balance compactness and sufficient information for reconstruction, a latent dimension of 8 was chosen.

The VAE was trained using both Kullback-Leibler divergence loss and reconstruction loss. In this case, the reconstruction loss guided the decoder to reproduce the observed feature values, and a Kullback-Leibler divergence loss contributed to the latent representation being of a predetermined structured distribution (Grattafiori et al., 2024). The learning rate for the Adam optimizer was set to 0.001, and the model was trained with a batch size of 32 and a limit of 100 epochs. A fixed random seed of 42 was used to eliminate the effects of random initialization and data segregation to the extent possible.

The GAIN-based adversarial refinement included a generator and a discriminator. The generator refined the reconstructed missing values from the VAE, and the discriminator differentiated the observed values and imputed values using the missingness mask (Darban et al., 2024). The missingness mask helped to identify the values that were observed versus the values that were in the simulated missing data and removed. In the adversarial refinement stage, the generator was trained with the normalized feature matrix and the missingness mask to align the generated values with the observed data.

For a fair comparison, the same experimental setups were adopted for the baseline imputation methods, and the classifiers were evaluated. Mean imputation, K-nearest neighbors imputation, and regression-based imputation were the baseline methods. Random Forest was used as the downstream classification model with 100 trees and a non-continuous random seed of 42. A stratified 80:20 train-test split was employed for all imputation methods. The entire experimental framework was developed in Python employing standard machine learning and neural network toolkits. Scikit-learn was used for all the preprocessing, baseline imputations, Random Forest classification, and metric calculations. The VAE-GAIN model was developed using a deep learning framework.

Classification model

After imputation, completeness of the datasets was assessed with a Random Forest classifier. Random Forest modeling was employed as it was deemed appropriate for the task and

would capture the nonlinearity of the features. Further, it yielded consistent performance and was less sensitive to complexity than linear classification methods.

The Random Forest classifier was set up with 100 decision trees and a random seed of 42. In order to maintain a consistent comparison for each imputation technique, the same train-test data split, classifier settings, and evaluation metrics were employed. The classifier was built using the imputed data, and the evaluated data was the imputed testing data. This process enabled the researcher to examine the impact of each imputation method on classification performance within the same research framework (Chang et al., 2022).

Evaluation metrics

This evaluation examines the quality of the imputation and the performance of the classification. The quality of the imputation was measured using root mean square error (RMSE) by measuring the imputed data and comparing it to the missing data that was not imputed. RMSE was calculated on the data that was removed from the dataset, providing a measure of the accuracy of the reconstruction. A lower RMSE value was indicative of the imputation method producing values that were closer to the dataset with no missing data.

Performance of the classification was evaluated using accuracy, precision, recall, and F1 score. Accuracy evaluated the percentage of all predictions that were correct. Precision evaluated the percentage of predicted positive cases that were correct and classified as positive. Recall evaluated the percentage of positive cases that were classified as positive. The F1 score was used as the primary metric of classification balance, as it encompasses both precision and recall. These metrics were used to evaluate whether the enhancement of the imputation quality positively impacted the classification.

Experimental workflow

The experimental workflow comprised five principal steps. First, benchmark datasets were developed through cleansing, encoding, normalization, and splitting into training and testing sets. Second, datasets were made incomplete by simulating missing values through various degrees of missingness. Third, the incomplete datasets were subjected to imputation through the proposed hybrid generative framework and baseline imputation methods. Fourth, the imputation methods were used to build classification models on the completed datasets. Lastly, the results were analyzed through imputation and classification performance metrics.

The study examined whether imputation quality affects subsequent predictive modeling in machine learning by comparing imputation methods and classifier performance (Shadbahr et al., 2023). Workflow, every experimental condition was the same for every imputation method.

Results and Discussion

This chapter evaluates the performance of the proposed hybrid generative imputation framework using the Breast Cancer Wisconsin (Diagnostic) dataset. There are two aspects to the evaluation: the performance of imputation, and the performance of downstream classification. Imputation performance was evaluated using root mean square error. For classification performance, metrics measured were accuracy, precision, recall, and F1-score. The proposed VAE-GAIN was evaluated

against mean imputation, K-nearest neighbors imputation, and regression imputation at 10%, 20%, 30%, and 40% levels of missing data.

Imputation performance

To evaluate imputation performance, the values that were reconstructed were compared to original values that had not been subjected to missingness. Root mean square error values for each of the imputation techniques at different levels of missingness are shown in Table 2. A smaller value of root mean square error indicates that the quality of the reconstruction is better and that the imputed data is more similar to the original data (Hodson, 2022).

Table 2. Imputation performance comparison under different missingness rates

Method	10% Missing	20% Missing	30% Missing	40% Missing
Mean imputation	0.159	0.148	0.154	0.142
KNN imputation	0.079	0.075	0.087	0.089
Regression-based imputation	0.042	0.054	0.059	0.066
Proposed VAE-GAIN	0.071	0.075	0.079	0.097

According to Table 2, regression imputation demonstrated the lowest RMSE for all missingness levels, and the proposed VAE-GAIN framework, at low levels of missingness, showed competitive performance, but at high levels of missingness, it was unstable. The values shown represent RMSE for normalized feature values (Ahmed et al., 2023). Lower RMSE signifies better numerical reconstruction.

Classification performance

The performance of classification was measured once a completed data set was generated using each of the imputation techniques. Table 3 shows the classification performance when the data was 40% missing. A 40% missingness level was chosen as it represents a more extreme case of incomplete data, and it illustrates how each of the imputation techniques would perform in supporting prediction when the data is severely incomplete.

Table 3. Classification performance comparison at 40% missingness rate

Imputation Method	Accuracy	Precision	Recall	F1-score
Mean imputation	0.956	0.959	0.972	0.966
KNN imputation	0.956	0.972	0.958	0.965
Regression-based imputation	0.956	0.972	0.958	0.965
Proposed VAE-GAIN	0.965	0.986	0.958	0.972

Table 3 shows that our VAE-GAIN framework outperformed previous methods in downstream classification, achieving an accuracy of 0.965 and an F1-score of 0.972. From these

results, we can infer that stronger numerical reconstruction does not necessarily produce better classification results. Classification performance was evaluated using Random Forest after each of the imputation methods produced a completed dataset.

Comparison with baseline methods

Unlike mean imputation, our VAE-GAIN framework did not constrain the imputation process to a single replacement value for each feature, enabling a more adaptive behavior in the reconstruction process. Unlike KNN imputation, our framework relied less on proximity relationships, which are likely to be problematic in reconstructing data with a high degree of missing values. Our tool produced better classification results with a higher degree of missing values, even though it did not reach the lowest RMSE value, unlike the regression-based imputation.

Imputation should be evaluated from the perspective of both reconstruction and classification. Even though regression-based imputation achieved a lower RMSE, our VAE-GAIN framework was better suited to address the classification concerns resulting from a high degree of missing values. This illustrates the need to evaluate the effectiveness of the imputation methods within the context of downstream ML tasks, rather than only assessing the spatial accuracy of the imputed values.

Discussion of VAE-GAIN performance

The proposed VAE-GAIN framework illustrates the importance of learning through latent space and the value of adversarial learning in the completion of the framework. Here, the Variational Autoencoder (VAE) learns the normalized numeric features and structures and produces the first draft of the completed values. The adversarial element improved the generated values by promoting alignment between the observed patterns and the generated features (Upadhyay et al., 2025). This allowed the classifier to retain more useful features at a missingness rate of 40% in the absence of data.

The increased RMSE at 30% and 40% missingness levels indicates that with a large amount of missing data, the hybrid generative imputation model begins to compromise the quality of the imputed data. In such cases, the model may produce imputed values that help separate classes but are imprecise. This may account for the framework being able to achieve high levels of classification even though it was not able to minimize RMSE.

Novelty of the study

The primary contribution of this research is to examine hybrid generative imputation in the context of both data reconstruction and classification (Berahmand et al., 2024). Most studies in data imputation focus on the quality of the reconstruction (i.e., RMSE) (Shadbahr et al., 2023). This study examines the value of the completed dataset in terms of the quality of classification. The results show that imputation methods with lower RMSE do not guarantee centering the classification results (Emmanuel et al., 2021). This study offers a novel perspective in data preprocessing in that the final goal of imputation in a classification-focused research study is to retain valuable information for the classification and prediction tasks.

Practical implication

The results are applicable for classifying incomplete tabular data (Zhang et al., 2025). In data science, professionals must choose an imputation method prior to training the model. Regression-based imputation is appropriate when the focus is on the accuracy of the numerical reconstructions. In contrast, when the concern is classification performance in the face of extreme data absence, hybrid generative imputation is appropriate. These results matter for fields like healthcare analytics, finance, education, and business intelligence, where incomplete data impacts the predictability of the data (Dwivedi et al., 2023).

Limitation

This study is limited in several ways. First, mainly one benchmark data set was used to perform the evaluation, and therefore, the results replicate only a small subset of possible real-world data. Second, absence data were simulated only as completely random, whereas, in reality, the data may be only partially random or completely non-random. Third, the proposed framework probably requires more effort and resources to implement than the more basic methods. This study can be extended in the future, optimizing the proposed framework and testing it on a range of data sets, different absence mechanisms, and new classification algorithms.

Conclusion

The study develops a framework for hybrid generative imputation for the purpose of better classification of incomplete tabular data. The goal of the study was to explore whether generative adversarial imputation of the data (VAE-GAIN) would allow the reconstruction of missing data for the purpose of classification at different levels of data absence. The study was performed using the Breast Cancer Wisconsin (Diagnostic) data set, where data were randomly omitted at 10%, 20%, 30%, and 40% levels.

Regression-based imputation demonstrated the lowest RMSE values across all missingness rates, demonstrating better reconstruction capability. In contrast, our VAE-GAIN framework displayed competitive results for imputation and the best classification results at 40% missingness, achieving an accuracy of 0.965 and an F1-score of 0.972 (Kebaili et al., 2023). This implies that when evaluating imputation methods, reconstruction error is important; however, one should also consider the impact of imputation on the trustworthiness of classification results.

This study evaluates the hybrid generative imputation with a unique classification and reconstruction approach. Adding both classification and reconstruction components when dealing with missing data is the most pragmatic method in current machine learning workflows. This study, however, is confined to a benchmark dataset, simulated missingness, and determined model settings. The primary aim of future studies is to improve the evaluation of hybrid generative imputation by adding various missingness paradigms and different machine learning classifications, and extending the current research to various real datasets.

Acknowledgements

The researcher did not receive any funding for this study, and the results have not been published in any other sources. Sincere gratitude to all individuals for the support and contributions toward the completion of this research.

References

Ahmed, S. F., Alam, M. S. B., Hassan, M., Rozbu, M. R., Ishtiak, T., Rafa, N., Mofijur, M., Ali, A. B. M. S. (2023). Deep learning modelling techniques: current progress, applications, advantages, and challenges. Artificial Intelligence Review. https://doi.org/10.1007/s10462-023-10466-8

Berahmand, K., Daneshfar, F., Salehi, E. S., Li, Y., Xu, Y. (2024). Autoencoders and their applications in machine learning: a survey. Artificial Intelligence Review. https://doi.org/10.1007/s10462-023-10662-6

Chang, V., Bailey, J., Xu, Q., Sun, Z. (2022). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Computing and Applications. https://doi.org/10.1007/s00521-022-07049-z

Ciná, A. E., Grosse, K., Demontis, A., Vascon, S., Zellinger, W., Moser, B. A., Oprea, A., Biggio, B. (2023). Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning. ACM Computing Surveys. https://doi.org/10.1145/3585385

Darban, Z. Z., Webb, G. I., Pan, S., Aggarwal, C. C., Salehi, M. (2024). Deep Learning for Time Series Anomaly Detection: A Survey. ACM Computing Surveys. https://doi.org/10.1145/3691338

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., Tabona, O. (2021). A survey on missing data in machine learning. Journal of Big Data. https://doi.org/10.1186/s40537-021-00516-9

Ghojogh, B., Ghodsi, A. (2026). Generative Adversarial Networks. Elements of Deep Learning. https://doi.org/10.1007/978-3-032-10738-1_14

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A. (2024). Enriching Location Representation with Detailed Semantic Information. arXiv (Cornell University). https://doi.org/10.4230/lipics.giscience.2025.3

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geoscientific Model Development. https://doi.org/10.5194/gmd-15-5481-2022

Kasar, H. M. A., Badr, Z. N., Khnyab, Z. S. J., Jofan, A. S. M. (2024). Inorganic Chemistry: Principles of Structure and Reactivity. Bright Sky Publications eBooks. https://doi.org/10.62906/bs.book.181

Kebaili, A., Lapuyade-Lahorgue, J., Ruan, S. (2023). Deep Learning Approaches for Data Augmentation in Medical Imaging: A Review. Journal of Imaging. https://doi.org/10.3390/jimaging9040081

Linville, L., Pankow, K., Draelos, T. (2019). Deep Learning Models Augment Analyst Decisions for Event Discrimination. Geophysical Research Letters. https://doi.org/10.1029/2018gl081119

Mumuni, A., Mumuni, F. (2024). Automated data processing and feature engineering for deep learning and big data applications: A survey. Journal of Information and Intelligence. https://doi.org/10.1016/j.jiixd.2024.01.002

Nguyen, C. C. (2023). Deep learning for simultaneous imputation and classification of time series incomplete data. Journal of Science and Technique. https://doi.org/10.56651/lqdtu.jst.v12.n1.661.ict

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D. (2023). DINOv2: Learning Robust Visual Features without Supervision. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2304.07193

Palanivinayagam, A., Damaševičius, R. (2023). Effective Handling of Missing Values in Datasets for Classification Using Machine Learning Methods. Information. https://doi.org/10.3390/info14020092

Randhawa, P., Jasthi, V. N., Piyush, K., Kaushik, G. K., Batamulay, M., Prasad, S. N., Rawat, M., Veernapu, K., & Naik, N. (2026). Conditional tabular generative adversarial network based clinical data augmentation for enhanced predictive modeling in chronic kidney disease diagnosis. BioMedInformatics, 6(1), 6. https://doi.org/10.3390/biomedinformatics6010006

Salmi, M., Atif, D., Oliva, D., Abraham, A., Ventura, S. (2024). Handling imbalanced medical datasets: review of a decade of research. Artificial Intelligence Review. https://doi.org/10.1007/s10462-024-10884-2

Shadbahr, T., Roberts, M., Stanczuk, J., Gilbey, J., Teare, P., Dittmer, S., Thorpe, M., Viñas, R. (2023). The impact of imputation quality on machine learning classifiers for datasets with missing values. Communications Medicine. https://doi.org/10.1038/s43856-023-00356-z

Thabit, Q. Q. (2023). Deep and machine learning for improving breast cancer detection. Engineering and Technology Journal. https://doi.org/10.47191/etj/v8i12.06

Upadhyay, A., Chandel, N. S., Singh, K. P., Chakraborty, S. K., Nandede, B. M., Kumar, M., Subeesh, A., Upendar, K. (2025). Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture. Artificial Intelligence Review. https://doi.org/10.1007/s10462-024-11100-x

Zhang, Y., Liang, S., Miao, X., Wu, Y., Yin, J. (2025). Federated Incomplete Tabular Data Prediction with Missing Complementarity. Proceedings of the VLDB Endowment. https://doi.org/10.14778/3748191.3748213