JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

An Innovative Model for Enhancing Data Detection and Recovery within Digital Communication Systems

[1*]Luu Van Dai, [1]Huynh Thanh Hoa**

[1]Faculty of Electrical and Electronics Engineering, Cao Thang Technical College, Ho Chi Minh City

Abstract: Digital communication systems face growing challenges from data corruption and missing information, which can severely reduce system reliability and performance—particularly in real-time, noise-heavy environments. Traditional error-correction and data-imputation methods often fall short when encountering burst-type errors, structured missingness, or strict latency requirements. To address these limitations, this study introduces a new two-stage framework that integrates CatBoost-driven anomaly detection with MICE-based data reconstruction, offering a unified solution for both erroneous and absent data. The proposed model is validated using real-world data from Smart EV Charging infrastructure, under conditions that emulate practical error patterns. Our experimental evaluation demonstrates significant improvements over leading baseline techniques in detection precision, imputation quality, and computational efficiency. These results highlight the framework's suitability for deployment in time-critical, resource-limited environments such as edge devices and 5G communication systems.

Index Terms **- Machine Learning, Anomaly Detection, Data Recovery, Data Imputation, Digital Communication**

I. INTRODUCTION

The worldwide surge in electric vehicle (EV) adoption has accelerated the need for advanced charging infrastructure, with Smart EV Charging stations becoming essential for real-time energy allocation, grid coordination, and user interaction. These systems depend on reliable digital communication to transmit key information—including power requirements, battery state-of-charge, grid status, and fluctuating electricity prices. In real-world environments, however, the data produced by these stations is frequently affected by noise, inaccuracies, and missing entries arising from sensor malfunctions, environmental disturbances, or unstable communication links. These data quality challenges can undermine crucial operations such as load prediction, charging optimization, and grid stability. This underscores the pressing demand for effective error detection and data recovery solutions that can cope with the fast-changing, time-sensitive conditions of modern EV charging networks.

Digital communication systems often encounter data corruption issues that can substantially impact their reliability and operational performance across a wide range of use cases [1]. For instance, contemporary wireless communication networks typically report bit error rates between 10−3 and 10−6, and packet loss can surpass 5% in challenging environments. In high-stakes industrial domains—such as automated manufacturing or autonomous driving—any unnoticed data error may result in critical system failures. To ensure safety and proper functionality, these systems often demand ultra-low error rates, approaching 10−9 per hour [2].

Conventional error correction and data recovery strategies no longer meet the demands of modern communication systems. Standard Error Correction Codes (ECCs) exhibit sharp performance degradation under low signal-to-noise conditions, struggle to handle burst-type corruption, and are incapable of correcting insertion or deletion errors. Likewise, Automatic Repeat Request (ARQ) schemes add transmission latency and perform poorly when delays are high or when errors occur in bursts [3]. These limitations become even more critical in next-generation environments—such as 5G networks, IoT deployments, and edge-computing platforms—where extremely low latency and large-scale device connectivity are fundamental requirements.

Similarly, current data imputation methods face critical shortcomings in communication contexts. K-Nearest Neighbors (KNN) and MissForest are sensitive to noise and struggle with complex patterns, while matrix factorization and tensor completion methods become slow and ineffective when data loss is extensive [4], [5]. Generative Adversarial Imputation Networks (GAIN) demand substantial computational resources and fine-tuning, making them impractical for real-time applications [6]. Furthermore, these methods often fail in the presence of burst errors, non-random missing data, or sensor drift—conditions commonly found in real-world communication systems.

To overcome these limitations, we propose a novel two- stage framework that integrates anomaly detection and imputation tailored for noisy and dynamic digital communication environments. Our method is evaluated on a real-world dataset from electric vehicle smart charging stations, containing critical parameters such as energy consumption, power, voltage, and battery state-of-charge. We simulate realistic error and missing data conditions to validate the framework's robustness. The results demonstrate that our approach can reliably detect and recover corrupted data, making it highly suitable for the demands of modern, real-time communication systems.

The main contributions of this work are as follows:

• To the best of our knowledge, this is the first work that proposes a unified two-stage framework combining anomaly detection and missing data imputation for im- proving data accuracy in digital communication systems. Our method specifically addresses both faulty and missing data simultaneously, which is crucial in noisy, real- time environments.

• We successfully implement our system on a real-world dataset collected from electric vehicle smart charging stations, which includes key communication parameters such as energy consumption, voltage, power, and battery state-of-charge. Our implementation simulates realistic error patterns and missing data conditions to evaluate system robustness.

• We conduct a comprehensive comparison with state-of- the-art data imputation and error correction methods, including KNN, MissForest, matrix factorization, and GAIN. Our results demonstrate significant improvements in detection accuracy and imputation quality, especially under burst error and non-random missing data scenarios. The rest of this paper is organized as follows: Section 2 reviews related work on error correction and data imputation; Section 3 explains our two-stage approach; Section 4 shows detailed experimental results; and Section 5 concludes with findings and future research directions.

II. Related Works

A. Anomaly Detection in Digital Communication

Digital communication networks need strong anomaly detection to spot data errors in real time. Modern deep learning methods improve on older statistical approaches but face issues in communication systems [9]. Autoencoders and their variants [10], which learn normal data patterns and flag errors, work well in high-data settings. However, they struggle in communication systems with fast-changing conditions and diverse errors. Their high computing needs also make them unsuitable for low delay 5G or edge computing systems [2].

Prediction-based methods, like recurrent neural networks and transformers [11], are good for time-series data but fail in communication systems with bursty traffic, unstable connections, or sudden changes. Their high memory and computing demand also do not fit resource-limited IoT devices or edge nodes. Graph-based methods [12], useful for social or financial data, do not suit communication systems because they assume structures that do not match real signal patterns and are too slow for large networks.

B. Limitations of Data Imputation methods

Traditional statistical imputation, like mean or regression methods, assumes data patterns that do not hold in communication systems. Deep learning methods, like autoencoders [10] and Generative Adversarial Imputation Nets (GAINs) [6], capture complex data patterns but need too much computing power for real-time use. GAINs also produce inconsistent results, which is risky for critical systems. SimpDM [13], a data imputation technique leveraging a Multi-Layer Perceptron (MLP)-based Diffusion Model, exhibits prolonged training and inference durations and demonstrates suboptimal performance when trained on limited datasets.

C. Combining Anomaly Detection and Faulty Data Recovery

Recent work combines anomaly detection and imputation to handle datasets with missing and incorrect data. Robust autoencoders try to do both tasks at once but face issues. Their goals conflict, leading to poor results in both detecting errors and filling in missing data. They also lack reliability for critical communication systems. In industrial IoT, these combined methods assume stable data patterns, which does not match the changing conditions of communication systems. Current anomaly detection methods lack the speed, clarity, and reliability needed for real-time communication monitoring [2]. Deep learning methods are too slow and resource-heavy for modern systems, especially for edge devices. Communication systems need clear, explainable methods. Tree-based methods like CatBoost offer clear results and work well with mixed data types, making them suitable for communication systems [7].

For imputation, communication systems need fast, reliable, and clear methods. MICE's iterative approach and strong theoretical base make it a good fit for handling missing data in critical applications [8].

Existing methods do not fully meet the needs of digital communication systems, which require speed, clarity, and reliability. A new two-stage framework is proposed, using CatBoost for anomaly detection and MICE for imputation. This approach avoids conflicts in combined methods by separating tasks, leveraging each method's strengths. Its modular design ensures reliability and ease of maintenance for real- world communication systems.

III. Related Works

A. Data augmentation

Figure 1 illustrates the development process of a two-phase framework for recovering faulty data and imputing missing data. The dataset comprises approximately 3 million real-world data samples collected from the logs of Smart EV Charging stations, including features such as initial state of charge (SOC), final SOC, power, voltage, energy, and more.

The energy feature is selected for anomaly simulation, as it can be affected in real-world scenarios by environmental factors like weather changes, electrical grid errors, or cybersecurity attacks. To simulate these anomalies, 10% of the energy values are randomly selected and amplified by a factor ranging from 1.5 to 5.0 times their original value. Modified data points are labeled as 1, while unmodified ones are labeled as 0. These labeled data are then used to train both the anomaly detection and data imputation models.

B. Phase 1: Anomaly Detection

The initial phase of the proposed framework employs gradient boosting decision trees (GBDT) through the CatBoost algorithm to perform binary classification for anomaly detection. CatBoost [7], developed by Yandex, represents a state- of-the-art implementation of gradient boosting that addresses fundamental limitations of traditional GBDT approaches, particularly in handling categorical features and mitigating pre- diction shift and biased pointwise gradient estimates.

1) Mathematical Foundation of CatBoost: The CatBoost algorithm constructs an ensemble of decision trees through additive

modelling, where the final prediction is expressed as:

$$F(\boldsymbol{x}) = \sum_{k=0}^{K} f_k(x) \tag{1}$$

where $F(\boldsymbol{x})$ represents the ensemble prediction, $f_k(\boldsymbol{x})$ denotes the $f_k(x)$ base learner (decision tree), and $\boldsymbol{K}$ is the total number of trees in the ensemble.

The optimization objective follows the general gradient boosting framework, minimizing the loss function:

$$L = \sum_{i=1}^{n} l\big(y_i, F(x_i)\big) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

where $l\big(y_i, F(x_i)\big)$ represents the loss function for the i-th sample with true label y_i and prediction $F(x_i)$, and $\Omega(f_k)$ denotes the regularization term for the k-th tree to prevent overfitting.

For binary classification, CatBoost employs the logistic loss function:

$$l\big(y, F(x)\big) = log\left(1 + exp\big(-y.F(x)\big)\right) \tag{3}$$

where $y \in \{-1, +1\}$ represents the binary class labels, transformed from the original $\{0, 1\}$ encoding.

2) Ordered Boosting and Prediction Shift Mitigation: CatBoost addresses the fundamental issue of prediction shift inherent in traditional gradient boosting through ordered boosting. The algorithm maintains multiple models, each trained on different subsets of the data to ensure unbiased gradient estimation. For each sample $\boldsymbol{x_i}$, the gradient is computed using a model $M_{\sigma(i)}$ trained only on samples $\{\boldsymbol{x_j} : \sigma(j) < \sigma(i)\}$, where σ represents a random permutation of the training indices.

The ordered boosting procedure can be formalized as:

$$g_i^{(t)} = \nabla_F l\left(y_i, F_{\sigma(i)}^{(t-1)}(x_i)\right) \tag{4}$$

where $g_i^{(t)}$ represents the gradient for sample i at iteration t, and $F_{\sigma(i)}^{(t-1)}$ the model trained on the first $\sigma(i) - 1$ samples according to the permutation σ.

3) Cross-Validation: The model performance is evaluated using multiple metrics appropriate for binary classification tasks:

Precision:

$$P = \frac{TP}{TP + FP} \tag{5}$$

Recall (Sensitivity):

$$R = \frac{TP}{TP + FN} \tag{6}$$

F1-Score:

$$F_1 = \frac{2PR}{P + R} \tag{7}$$

AUC-ROC:

$$Area\ Under\ ROC\ Curve \tag{8}$$

where TP, FP, and FN represent true positives, false positives, and false negatives, respectively.

4) Threshold Optimization: The binary classification threshold is optimized to maximize the F1-score on the validation set, balancing precision and recall for anomaly detection. Given the critical nature of anomaly detection in communication systems, the threshold selection prioritizes minimizing false negatives (missed anomalies) while maintaining acceptable false positive rates. The optimal threshold τ^* is determined as:

$$\tau^* = \arg max F_1(\tau) = \arg max \frac{2.Precision(\tau).Recall(\tau)}{Precision(\tau).Recall(\tau)} \tag{9}$$

The selected threshold is subsequently applied to classify data points in the test set, with predictions $\hat{y}_i = 1$ if $P(y_i = 1|x_i) \geq \tau^*$ and $\hat{y}_i = 0$ otherwise, where $P(y_i = 1|x_i)$ represents the predicted probability of anomaly for sample i.

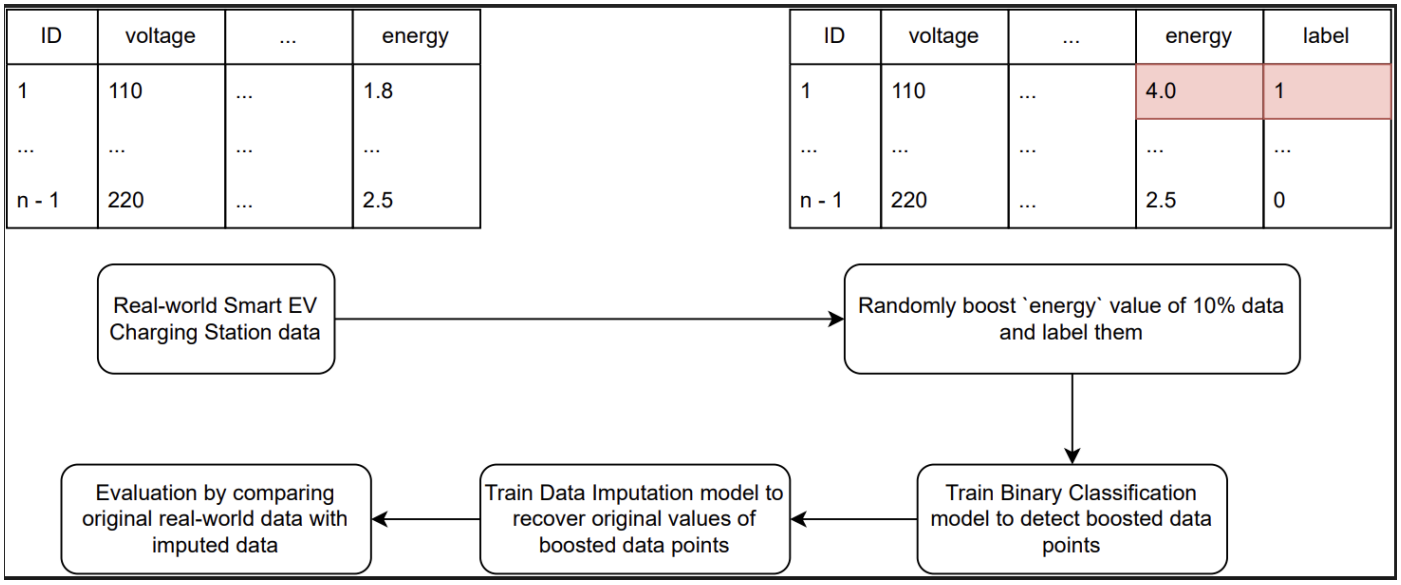

Figure 1. Development pipeline for the faulty data recovery framework

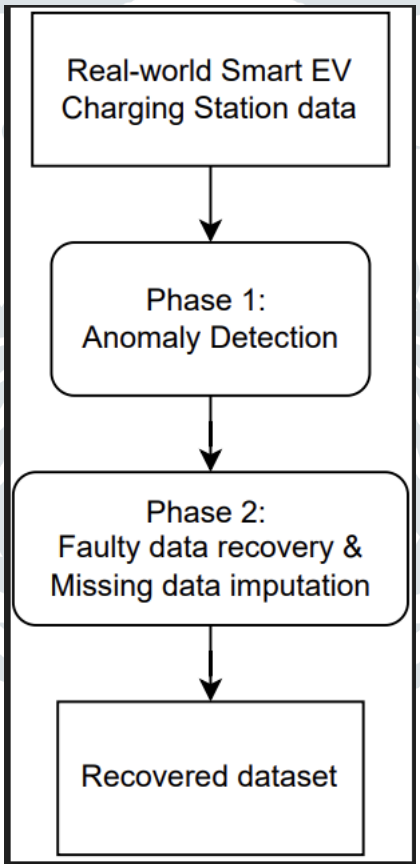

Figure 2. Two-phase data recovery & imputation framework

A. *Phase 1.5: Anomaly Values Removal*

Following the binary classification phase, detected anomalous data points undergo systematic removal through value nullification to prepare the dataset for subsequent imputation procedures. This intermediate phase serves as a critical bridge between anomaly identification and data reconstruction, ensuring that corrupted information does not propagate through the imputation algorithm and compromise the integrity of thenrecovered dataset.

The anomaly removal process operates on the binary classification outputs generated by the CatBoost model in Phase 1. For each data sample xi in the dataset $D = \{x_1, x_2, \ldots, x_n\}$, the corresponding binary prediction $\hat{y}_i$ determines whether the sample contains anomalous information. The removal operation is formally defined as:

$$x_i^{modified} = \begin{cases} x_i & if\ \hat{y}_i = 0(normal) \\ x_i^{NaN} & if\ \hat{y}_i = 0(normalous) \end{cases} \quad (10)$$

where x_i^{NaN} represents the modified feature vector with anomalous values replaced by Not-a-Number (NaN) indicators, specifically targeting the features identified as containing corrupted information.

B. Phase 2: Faulty Data Recovery and Missing Data Imputation

The second phase employs Multiple Imputation by Chained Equations (MICE) [8] to reconstruct the nullified anomalous values through iterative conditional modeling. MICE addresses single imputation limitations by generating multiple plausible values for each missing observation, preserving imputation uncertainty inherent in the reconstruction process.

1) Mathematical Formulation: Given the dataset $Y = (Y_{obs}, Y_{mis})$ with observed and missing components, MICE generates m imputed datasets $\{Y^{(1)}, Y^{(2)}, \ldots, Y^{(m)}\}$ where imputed values are drawn from the posterior predictive distribution:

$$Y_{mis}^{l} \sim P\left(Y_{mis} \middle| Y_{obs}, \theta^{(l)}\right) \tag{11}$$

The algorithm factorizes the joint distribution through conditional dependencies:

$$P(Y_1, Y_2, \ldots, Y_p) = P(Y_1) \prod_{j=2}^{p} P\left(Y_j \middle| Y_1, \ldots, Y_{j-1}\right) \tag{12}$$

2) Iterative Conditional Modeling: The chained equations procedure cycles through each variable with missing values, updating imputations based on current states of all other variables. For variable Y_j at iteration t:

$$Y_j^{(t)} | Y_{-j}^{(t-1)} \sim P\left(Y_j \middle| Y_{-j}^{(t-1)}, \theta_j^{(t)}\right) \tag{13}$$

where $Y_{-j}^{(t-1)}$ represents all variables excluding Y_j at iteration $t-1$.

For continuous variables, conditional models employ linear regression with Bayesian parameter estimation:

$$Y_{j,i} = \beta_j^T X_{i,-j} + \epsilon_{j,i}, \quad \epsilon_{j,i} \sim N\left(0, \sigma_j^2\right) \tag{14}$$

Parameter uncertainty is incorporated through posterior sampling:

$$\beta_j \middle| Y_{obs}, \sigma_j^2 \sim N\left(\hat{\beta}_j, \sigma_j^2 (X^T X)^{-1}\right) \tag{15}$$

3) Validation Metrics: Imputation quality is quantified through Mean Squared Error, providing direct measurement of reconstruction fidelity:

$$MSE_j = \frac{1}{n_{mis,j}} \sum_{i \in M_j} \left(Y_{j,i}^{true} - Y_{j,i}^{imp}\right)^2 \tag{16}$$

This framework ensures principled reconstruction of anomalous energy values while maintaining statistical consistency with the underlying data generation process.

IV. EXPERIMENT EVALUATIONS

A. Experimental Settings

1) Dataset Preparation and Anomaly Injection: The experimental evaluation employs a real-world dataset comprising 2.7 million samples collected from Smart EV Charging station operational logs.

To simulate realistic fault conditions encountered in digital communication systems, synthetic anomalies are systematically introduced into the energy feature. A random subset representing 10% of the total samples ($n_{anomaly} \approx 270{,}000$) is selected for anomaly injection. Each selected energy value $E_{original}$ is transformed according to:

$$E_{anomalous} = E_{original} \times \alpha, \quad \alpha \sim U(1.5, 5.0) \tag{17}$$

where αrepresents a multiplicative amplification factor uniformly sampled from the interval [1.5,5.0], simulating energy measurement corruption due to sensor drift, environmental interference, or communication channel errors.

2) Ground Truth Labeling and Dataset Partitioning: The augmented dataset receives binary labels $y_i \in \{0,1\}$ where $y_i = 1$ indicates anomalous samples and $y_i = 0$ denotes normal observations. This labeling scheme enables supervised training of the CatBoost anomaly detection model and provides ground truth for quantitative performance evaluation.

The labeled dataset undergoes stratified partitioning with 80% allocated for training and 20% for testing. Stratification ensures balanced anomaly representation across all partitions, maintaining the 10% anomaly ratio in each subset.

3) Model Training Configuration: The binary classification model uses 5-fold stratified cross-validation on the training set with early stopping based on validation AUC-ROC. The approach tests different synthetic data augmentation ratios to find the best balance between observed and generated data. The synthetic data ratios are 0.25, 0.5, 0.75, and 1.0 relative to the observed training data size. Two GAN architectures are used: TGAN (Tabular GAN) and CTGAN (Conditional Tabular GAN). TGAN is a Gaussian Copula-based generative model with automatic metadata detection. CTGAN is a conditional tabular GAN trained for 300 epochs with a batch size of 64, using CUDA acceleration when available and handling mixed-type data automatically.

Three imputation models are applied. MICE (Multiple Imputation by Chained Equations) runs for a maximum of 10 iterations, using Bayesian ridge regression for numerical features and mode imputation as a fallback for categorical features. MissForest uses a Random Forest Regressor with 100 estimators for numerical features and a Random Forest Classifier with 100 estimators for categorical features, applying a feature-wise iterative imputation strategy with a maximum of 10 iterations. The Denoising Autoencoder (DAE) has an encoder-decoder architecture with an intermediate bottleneck, a hidden dimension of 128 neurons, and a latent dimension of 64 neurons. It is trained for 50 epochs with a 15% noise injection rate during training, using the Adam optimizer with the default learning rate, Mean Squared Error (MSE) as the loss function, and ReLU activation for hidden layers.

B. Results

1) Anomaly Detection models evaluation: The empirical evaluation in Table I demonstrates CatBoost's decisive superiority across all performance metrics, exhibiting optimal convergence in the precision-recall trade-off space while achieving computational efficiency that surpasses deep learning architectures by an order of magnitude.

TABLE I
COMPARISON OF ANOMALY DETECTION MODELS

Model	Accuracy	Precision	Recall	F1 score	Pred. Time (s)
CatBoost	**0.98**	**0.97**	**0.82**	**0.89**	**0.11**
Autoencoder [14]	0.76	0.88	0.76	0.80	1.78
VAE [10]	0.55	0.86	0.55	0.64	2.59

TABLE II
COMPARISON OF IMPUTATION MODELS

Model	Mean Squared Error (MSE)
MICE	**0.0032**
GAIN [6]	14691.36
MissForest [5]	5283.16
SimpDM [13]	112531.32

CatBoost attains an F1-score of 0.89, representing a 11.25% and 39.06% improvement over the Autoencoder and Variational Autoencoder baselines, respectively, while maintaining exceptional precision (0.97) that minimizes false positive rates—a critical consideration in anomaly detection where classification errors impose significant operational costs.

The model's inference latency of 0.11 seconds establishes a 16.18 time and 23.55 time speedup relative to the neural architectures, indicating superior scalability for high-throughput production environments where real-time anomaly identification is paramount.

This performance differential suggests that gradient boosting frameworks with categorical feature optimization exhibit superior generalization capacity for structured anomaly detection tasks compared to reconstruction-based deep learning approaches, particularly when computational constraints and deployment latency requirements necessitate efficient inference pipelines.

2) Imputation models evaluation: Results from Table II indicates that MICE achieves far superior predictive accuracy, with errors orders of magnitude smaller than the other models.

The extremely low MSE suggests that MICE is highly effective at imputing missing data or fitting the dataset, making it the optimal choice for applications where minimizing prediction error is critical.

Additionally, the vast differences in MSE values highlight MICE's robustness and reliability over GAIN [6], MissForest [5], and SimpDM [13], which exhibit much higher errors, potentially indicating overfitting or poor generalization on the dataset.

V. CONCLUSIONS

This paper presents a novel two-phase framework for detecting and recovering faulty data in digital communication systems. By combining CatBoost for anomaly detection and MICE for data imputation, the method effectively addresses both erroneous and missing data. Experiments on a real-world Smart EV Charging dataset demonstrate superior accuracy and efficiency compared to existing approaches. Its modular design and low inference latency make it suitable for real-time edgeapplications. results show that the framework significantly reduces false positives in anomaly detection and improves imputation accuracy under burst errors and non-random missing conditions. These findings highlight the potential of combining lightweight machine learning with iterative imputation to improve data reliability in modern networks. Future work will focus on scalability and deployment in broader domains such as IoT, smart grids, and autonomous systems.

REFERENCES

[1] W. Elleuch, P. Sondi, A. Meddahi, and S. Lecomte, "Evaluation of 5G Relay-Empowered and Device-to-Device Communications for Rescue Mission," IEEE Access, vol.13, pp.104614–104629, 2025, doi: 10.1109/ACCESS.2025.3579872.

[2] Petersen, Stig, and Niels Aakvaag. "Wireless instrumentation for safety critical systems. technology, standards, solutions and future trends." (2015).

[3] Dutt, Arinjoy, et al. "Error-Correcting Codes in 5G and Beyond." International Journal of Engineering Research 10.07.

[4] Peterson, Leif E. "K-nearest neighbor." Scholarpedia 4.2 (2009): 1883.

[5] Stekhoven, Daniel J., and Peter B¨ uhlmann. "MissForest—non- parametric missing value imputation for mixed-type data." Bioinformatics 28.1 (2012): 112-118.

[6] Yoon, Jinsung, James Jordon, and Mihaela Schaar. "Gain: Missing data imputation using generative adversarial nets." International conference on machine learning. PMLR, 2018.

[7] Prokhorenkova, Liudmila, et al. ”CatBoost: unbiased boosting with categorical features.” Advances in neural information processing systems 31 (2018).

[8] White, Ian R., Patrick Royston, and Angela M. Wood. ”Multiple imputation using chained equations: issues and guidance for practice.” Statistics in medicine 30.4 (2011): 377-399.

[9] Huang, Haoqi, et al. ”Deep Learning Advancements in Anomaly Detection: A Comprehensive Survey.” arXiv preprint arXiv:2503.13195 (2025).

[10] Pinheiro Cinelli, Lucas, et al. ”Variational autoencoder.” Variational M ethods for machine learning with applications to deep networks. Cham: Springer International Publishing, 2021. 111-149.

[11] Casella, Monica, et al. ”Transformers deep learning models for missing data imputation: an application of the ReMasker model on a psychometric scale.” Frontiers in Psychology 15 (2024): 1449272.

[12] Akoglu, Leman, Hanghang Tong, and Danai Koutra. ”Graph based anomaly detection and description: a survey.” Data mining and knowledge discovery 29 (2015): 626-688.

[13] Liu, Yixin, et al. ”Self-supervision improves diffusion models for tabular data imputation.” Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. 2024.

[14] Yamanaka, Yuki, et al. ”Autoencoding binary classifiers for supervised anomaly detection.” PRICAI 2019: Trends in Artificial Intelligence: 16th Pacific Rim International Conference on Artificial Intelligence, Cuvu, Yanuca Island, Fiji, August 26–30, 2019, Proceedings, Part II 16. Springer International Publishing, 2019.