ISSN: 2755-6190

Review Article

Open Access Journal of Artificial Intelligence and Technology

Optimizing Predictive Maintenance of Machines with Innovative Data Augmentation Strategies

Hemanth Kumar Kuppan*, **Nigamanth Rajagopalan and Jeswin John John**

Student, Department of Computer Science and Engineering, SRMIST, Chennai, Tamil Nadu, India

***Corresponding author**
Hemanth Kumar K, Student, Department of Computer Science and Engineering, SRMIST, Chennai, Tamil Nadu, India

Received: May 14, 2025; **Accepted:** May 21, 2025; **Published:** May 26, 2025

ABSTRACT

This work explores the use of Wasserstein Generative Adversarial Networks (WGANs) to generate synthetic sensor data for the prediction of Remaining Useful Life (RUL). The C-MAPSS FD001 dataset was first reduced using Principal Component Analysis (PCA) to retain key features. Synthetic data was created using WGAN and Conditional WGAN models and validated through statistical tests like the Kolmogorov-Smirnov test and Wasserstein distance. Machine learning models trained on the synthetic data showed performance close to those trained on real data. The results highlight that WGAN-based data generation can effectively support predictive maintenance by addressing data shortages.

Keywords: Wasserstein GAN, Synthetic Data Generation, Remaining Useful Life, Predictive Maintenance, Principal Component Analysis, C-MAPSS Dataset

Introduction

Predicting the Remaining Useful Life (RUL) of industrial machinery is critical for reducing downtime and maintenance costs. However, the availability of labeled operational data is often limited, making it difficult to train reliable machine learning models. Synthetic data generation offers a promising solution to address this gap. In this work, we use Wasserstein Generative Adversarial Networks (WGANs) to generate highquality synthetic sensor data based on the C-MAPSS FD001 dataset. To enhance data quality and training efficiency, Principal Component Analysis (PCA) is applied to reduce the dimensionality of the original sensor readings. Both WGAN and Conditional WGAN (CWGAN) models are trained to generate realistic sensor outputs, which are then evaluated using statistical measures such as the Kolmogorov-Smirnov test and Wasserstein distance. Finally, machine learning models are trained on both real and synthetic datasets to predict RUL, demonstrating that synthetic data can achieve comparable performance to real-world data. This study highlights the potential of WGANs to support predictive maintenance applications by generating reliable synthetic datasets when real data is scarce.

Literature Review

The use of synthetic data generation in predictive maintenance has picked up tremendous pace over the last few years. Thompson proposed a diffusion-based data augmentation method for multivariate time-series data and showed how it can enhance the robustness of predictive maintenance models by making sensor inputs more diverse [1]. Likewise, Khosravi applied CTGAN to smart manufacturing, where they tackled data sparsity in pulp-and-paper production environments [2]. Their strategic augmentation method improved prediction performance in industrial anomaly detection tasks.

GANs have been successful in generating synthetic data in a variety of fields. Tanaka and Aranha laid down a basic framework for applying GANs in industry, which was further enhanced by Bowles, who used GAN-based augmentation in medical imaging to rectify imbalanced datasets and enhance model training [3,4]. Arjovsky suggested the Wasserstein GAN (WGAN), a significant advancement over traditional GANs, with stable convergence and an interpretable loss function via Wasserstein distance [5].

Citation: Hemanth Kumar Kuppan, Nigamanth Rajagopalan, Jeswin John. Optimizing Predictive Maintenance of Machines with Innovative Data Augmentation Strategies. Open Access J Artif Intel Tech. 2025. 1(1): 1-4. DOI: doi.org/10.61440/OAJPSD.2025.v1.01

In addition, Mirza and Osindero developed Conditional GANs (CGANs), allowing class-conditional generation, providing stronger control over data generation a key capability in multi-sensor contexts [6].

The data used here is based on the work of Saxena and Goebel, who constructed the C-MAPSS dataset as a benchmark for simulation of turbofan engine degradation [7]. The dataset has since become a reference for RUL prediction. Fernandez addressed the challenge of the domain gap by using domain adaptation with GANs as a way of enhancing transferability between operating conditions, an idea very similar to our intended future work [8].

Meta-learning and few shot learning techniques have also been investigated. Schwendemann introduced metalearning models with generative methods for few-shot RUL prediction and demonstrated that models can generalize well even with few data [9]. Likewise, Coble and Goebel introduced a model-based prognostic method with concurrent damage progression, combining physical modeling with datadriven techniques [10].

Anomaly detection by GANs was explored by Zimmermann and Rupprecht, specifically within the aviation sector, where generative models aided in detecting discrepancies from normal operation in time-series data [11]. Tornede put strong focus on multi-task learning for RUL prediction under noisy labeling conditions, which is a parallel to our problem of producing informative labels for synthetic sequences [12].

A thorough review by Zhao discussed deep learning techniques for RUL estimation, emphasizing the prevalence of architectures such as LSTM and CNNs [13]. Zhang investigated ensemble learning and data augmentation methods, suggesting hybrid models to enhance prediction accuracy in prognostics [14]. Zhu lastly created a deep generative model with attention mechanisms for RUL estimation, highlighting the advantages of integrating temporal modeling with generative synthesis [15].

Together, these studies indicate a distinct trend in predictive maintenance research from basic models to sophisticated generative techniques justifying the incorporation of WGANs, CWGANs, and PCA in our suggested methodology for enhancing RUL prediction in industrial systems.

Dataset

The Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) dataset was developed by NASA to support research in prognostics and health management (PHM). It simulates the degradation of aircraft engines under various operating conditions and fault modes. Each engine operates until failure, providing data suitable for Remaining Useful Life (RUL) prediction tasks.

A. Dataset Variants

C-MAPSS is divided into four subsets: FD001, FD002, FD003, and FD004. Each subset differs in terms of the number of operating conditions and fault modes.

- FD001: Single operating condition, single fault mode (simplest scenario).
- FD002: Multiple operating conditions, one fault mode.
- FD003: Single operating condition, multiple fault modes.
- FD004: Multiple operating conditions, multiple fault modes.

This study focuses on FD001, making it easier to design and validate models for early-stage RUL prediction.

B. Data Structure

Each record in the FD001 subset includes:

- Unit Number: Identifier for each engine.
- Time in Cycles: Operational cycles from the start until failure.
- Operational Settings: Three operational settings that capture different flight conditions.
- Sensor Measurements: 21 sensor readings monitoring various physical and operational parameters of the engine.

C. Rul Labeling

The Remaining Useful Life (RUL) for each cycle can be computed by subtracting the current cycle number from the maximum cycle number for that engine. Accurate RUL labels allow the dataset to be used in supervised learning tasks.

D. Importance in Research

The C-MAPSS dataset has become a widely used benchmark in predictive maintenance. Its realistic simulation of engine wear, availability of rich sensor data, and clear failure labels make it ideal for developing, testing, and comparing machine learning models for health monitoring and RUL estimation.

Methodology

This study follows a structured approach to generate synthetic data and predict the Remaining Useful Life (RUL) of aircraft engines using the C-MAPSS FD001 dataset. The complete process is divided into several key steps as given below.

A. Data Preprocessing

The raw C-MAPSS FD001 dataset is first cleaned by removing unnecessary columns and handling missing values. Since sensor readings vary widely in scale, StandardScaler is applied to normalize the data. Principal Component Analysis (PCA) is then used to reduce the dimensionality while preserving 95.

B. Synthetic Data Generation using WGAN

A Wasserstein Generative Adversarial Network (WGAN) is used to generate realistic synthetic sensor data. The WGAN consists of two networks:

- Generator: Takes random noise as input and tries to produce realistic sensor data.
- Critic: Distinguishes between real and synthetic data, guiding the generator to improve.

During training, the critic and the generator are optimized alternately. Conditional WGAN (CWGAN) is also implemented, allowing sensor-specific data generation by conditioning sensor indices. RMSprop optimizers and weight clipping are used to stabilize training and enforce the Lipschitz constraint.

C. Synthetic Data Validation

To ensure that the synthetic data closely resemble the real sensor data, statistical validation techniques are applied:

- Kolmogorov-Smirnov (KS) test: Measures the similarity of distributions.

- Wasserstein distance: Quantifies the distance between real and synthetic data distributions.
- Anonymity score: Assesses how difficult it is to distinguish synthetic data from real data.

Additionally, visual validation is performed using kernel density estimate (KDE) plots and t-SNE (t-Distributed Stochastic Neighbor Embedding) for low-dimensional visualization.

D. Remaining Useful Life (RUL) Labeling

Both real and synthetic datasets are labeled with RUL values. The RUL is calculated as the difference between the maximum cycle count for each engine and the current cycle number. This transforms the problem into a supervised regression task.

E. Model Training and Evaluation

Two machine learning models are trained separately on both real and synthetic datasets to predict the Remaining Useful Life (RUL) of engines. Linear Regression is used as a baseline model due to its simplicity and interpretability, while the Random Forest Regressor, an ensemble learning method, is employed to better capture complex, non-linear relationships in the data. The models are evaluated using two key metrics: Root Mean Squared Error (RMSE), which measures the average prediction error, and R2 Score, which indicates how much of the variance in the RUL values the model is able to explain. By comparing the performance of models trained on synthetic data with those trained on real data, the quality and reliability of the synthetic dataset generated are assessed.

F. Final Analysis

The performance of models trained in synthetic data is compared to those trained in real data. KDE plots and t-SNE visualizations confirm that synthetic samples closely mimic real-world sensor behavior. The results show that models trained on synthetic data achieve performance comparable to those trained on real data, demonstrating the effectiveness of WGAN-based data generation for predictive maintenance tasks.

V. Results

This section evaluates the performance of three regression models—Linear Regression, Random Forest, and Gradient Boosting—trained on real, synthetic, and combined datasets for Remaining Useful Life (RUL) prediction. Performance is assessed using Root Mean Squared Error (RMSE) and the Coefficient of Determination (R^2 score).

Table 1: Model Performance on Real, Synthetic, and Combined Data

Model & Dataset	RMSE	R^2 Score
Real — Linear Regression	48.11	0.49
Real — Random Forest	39.94	0.65
Real — Gradient Boosting	42.09	0.61
Synthetic — Linear Regression	15.53	0.66
Synthetic — Random Forest	15.85	0.65
Synthetic — Gradient Boosting	15.65	0.66
Fine-Tuned RF (Combined Data)	40.00	0.65

A. Performance on Real Data

Models trained exclusively on real-world sensor data showed moderate performance. Random Forest yielded the best results with an RMSE of 39.94 and an R^2 score of 0.65, indicating a reasonably good fit. Linear Regression underperformed with an RMSE of 48.11 and an R^2 of 0.49, suggesting limited capacity to capture the non-linear dynamics in the data. Gradient Boosting also performed comparably with an RMSE of 42.09 and R^2 of 0.61.

B. Performance on Synthetic Data

Training on synthetic data generated via a generative model (e.g., WGAN/CWGAN) led to significantly better performance across all models. RMSE values hovered around 15.5, and R^2 scores reached up to 0.66, suggesting that the synthetic dataset was easier to model and exhibited lower variance or noise compared to the real-world dataset. Both Random Forest and Gradient Boosting showed strong generalization capabilities under these conditions.

C. Fine-Tuning on Combined Data

To leverage the strengths of both datasets, a fine-tuned Random Forest model was trained on the combined dataset (real + synthetic). This model achieved an RMSE of 40.00 and an R^2 score of 0.65, matching the best performance on real data. This suggests that synthetic data contributed positively to the learning process without degrading performance, thus validating the potential of data augmentation in RUL prediction.

D. Observations

- Synthetic data enhanced model performance, especially during initial training phases.
- Combining synthetic and real data maintained or slightly improved real-world generalization performance.
- Tree-based models (Random Forest, Gradient Boosting) consistently outperformed Linear Regression, confirming the importance of capturing non-linear relationships.

These results reinforce the idea that synthetic data can be a valuable tool in predictive maintenance applications, particularly in scenarios where labeled real-world data is limited.

Future Scope

- Domain Adaptation Techniques: Implement domain adaptation methods to minimize the domain gap between synthetic and real data, enhancing model transferability and robustness.
- Real-Time Predictive Maintenance Systems: Extend the current framework into a real-time deployment pipeline by integrating it with industrial IoT platforms and CMMS (Computerized Maintenance Management Systems).
- Future work can explore the use of advanced deep learning architectures such as LSTMs, Transformers, or Temporal Convolutional Networks (TCNs) for capturing complex temporal dependencies in sensor data for RUL prediction.

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: All authors have reviewed and approved the final manuscript for publication.

Availability of data and materials: The dataset used in this study (C-MAPSS FD001) is publicly available at NASA's Prognostics Data Repository: https://www.nasa.gov/content/prognostics-center-of-excellence-data-set-repository.

Competing interests: The authors declare that they have no competing interests.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions: Nigamanth Rajagopalan conducted model training, handled data preprocessing, and contributed to writing the methodology. Hemanth Kumar K developed the WGAN and CWGAN models and performed validation analysis. Jeswin John J helped with dataset preparation, statistical evaluation, and visualization (KDE, t-SNE).

Acknowledgements: The authors would like to thank the Department of Computer Science and Engineering, SRMIST, for the support and encouragement provided during the project.

References

1. Thompson A, Sommers A, Russell-Gilbert A, Cummins L, Mittal S, et al. Multivariate Data Augmentation for Predictive Maintenance using Diffusion. In2024 IEEE International Conference on Big Data. 2024. 4240-4247.
2. Khosravi H, Farhadpour S, Grandhi M, Raihan AS, Das S, et al. Strategic Data Augmentation with CTGAN for Smart Manufacturing: Enhancing Machine Learning Predictions of Paper Breaks in Pulp-and-Paper Production. arXiv preprint arXiv. 2023. 2311: 09333.
3. Tanaka FH, Aranha C. Data augmentation using GANs. arXiv preprint arXiv. 2019. 1904: 09135.
4. Bowles C, Chen L, Guerrero R, Bentley P, Gunn R, et al. Gan augmentation: Augmenting training data using generative adversarial networks. arXiv preprint arXiv. 2018. 1810: 10863.
5. Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. InInternational conference on machine learning. 2017. 214-223.
6. Mirza M, Osindero S. Conditional generative adversarial nets. arXiv preprint arXiv. 2014. 1411: 1784.
7. Saxena A, Goebel K. Turbofan engine degradation simulation data set. NASA Ames Prognostics Data repository, NASA Ames Research Center, Moffett Field. NASA Ames Prognostics Data repository NASA Ames Research Center Moffett Field. 2008.
8. Fernandez D, Lundstrom J, Karlsson R. "Domain adaptation for¨ predictive maintenance using generative adversarial networks," Engineering Applications of Artificial Intelligence. 2022. 114: 105042.
9. Schwendemann J, Thoma M, Burgard W. "Few-shot RUL prediction via meta-learning with deep generative models," in Proc. IEEE Int. Conf. on Prognostics and Health Management (PHM). 2021. 1-8.
10. Coble J, Goebel K. "Model-based prognostics with concurrent damage progression processes," IEEE Transactions on Systems, Man, and Cybernetics. 2009. 39: 878-887.
11. Zimmermann T, Rupprecht C, "Anomaly detection and data generation with GANs for predictive maintenance in the aviation industry," in Proc. European Conference on Machine Learning (ECML). 2019. 275-286.
12. Tornede A, Schmidt-Thieme L, Grabocka J. "Multi-task learning for remaining useful life prediction with labeling uncertainty," in Proc. ACM SIGKDD. 2020. 2367-2375.
13. Zhao A, Liu H, Li X, He Y. "Remaining useful life prediction of machinery: A review of deep learning approaches," Measurement. 2022. 186: 110164.
14. Zhang C, Wang P, Zhou B. "Data augmentation and ensemble learning for remaining useful life prediction in prognostics: A review," IEEE Access. 2019. 7: 158490-158507.
15. Zhu Z, Wang W, Gao Y, Zhou Y. "A deep generative model with attention for remaining useful life prediction," IEEE Transactions on Industrial Informatics. 2021. 17: 2729-2737.