Overview of Data Augmentation Techniques in Time Series Analysis

Ihababdelbasset ANNAKI, Mohammed RAHMOUNE, Mohammed BOURHALEB
Université Mohammed Premier, National School of Applied Sciences,
Laboratory of Research in Applied Sciences (LARSA),
Oujda, Morocco

***Abstract*—Time series data analysis is vital in numerous fields, driven by advancements in deep learning and machine learning. This paper presents a comprehensive overview of data augmentation techniques in time series analysis, with a specific focus on their applications within deep learning and machine learning. We commence with a systematic methodology for literature selection, curating 757 articles from prominent databases. Subsequent sections delve into various data augmentation techniques, encompassing traditional approaches like interpolation and advanced methods like Synthetic Data Generation, Generative Adversarial Networks (GANs), and Variational Autoencoders (VAEs). These techniques address complexities inherent in time series data. Moreover, we scrutinize limitations, including computational costs and overfitting risks. However, it's essential to note that our analysis does not end with limitations. We also comprehensively analyzed the advantages and applicability of the techniques under consideration. This holistic evaluation allows us to provide a balanced perspective. In summary, this overview illuminates data augmentation's role in time series analysis within deep and machine-learning contexts. It provides valuable insights for researchers and practitioners, advancing these fields and charting paths for future exploration.**

***Keywords*—Time series; data augmentation; machine learning; deep learning; synthetic data generation**

I. Introduction

The concept of data augmentation has become indispensable in modern machine learning, serving as a key technique to enhance the diversity and volume of training data [1]. Its roots can be traced back to the early stages of machine learning, where the challenge of limited data first emerged. Augmentation techniques, through methods such as image rotation, flipping, or text paraphrasing, enable models to learn from a varied set of inputs, thereby increasing their generalization capabilities [2]. This is especially crucial in preventing overfitting, a common challenge in machine learning models trained on limited datasets [3].

Data augmentation transcends various learning paradigms, playing a significant role in both supervised and unsupervised learning contexts. In supervised learning, it addresses challenges like class imbalance and enriches small datasets, enhancing model accuracy and reliability [4]. In unsupervised learning, augmentation techniques help in extracting more robust features and patterns from unlabeled data, a vital aspect in domains such as natural language processing and computer vision [5]. The versatility of these techniques is also evident in their adaptability to different data types, including images, text, and audio [6], [7].

Time series data, with its sequential and often periodic nature, introduces unique augmentation challenges. Standard augmentation methods may not be directly applicable due to the temporal dependencies inherent in time series data. Techniques like time warping [8], window slicing, or injecting synthetic anomalies [9] are tailored to maintain these temporal relationships. Such methods have been shown to significantly improve the performance of models in various time series applications, from stock market predictions and weather forecasting to electrocardiogram analysis in healthcare [10].

Beyond improving model performance, data augmentation has broader impacts on the field of machine learning. It contributes to more efficient use of available data, reducing the need for extensive data collection, which can be costly and time-consuming. However, it also raises ethical considerations, particularly in ensuring that augmented data does not introduce or perpetuate biases. This is a critical aspect in applications involving human-centric data [11], [12], where fairness and representativeness are paramount.

This review provides a comprehensive analysis of data augmentation techniques with key contributions as follows:

- Holistic Overview: Showcases a wide array of data augmentation methods, presenting a broad perspective rather than focusing on a specific scope, thus providing a more inclusive understanding of the field.
- Comprehensive Analysis: Compared to earlier reviews, this approach stands out by offering a more thorough examination of data augmentation techniques across various machine learning and deep learning domains.
- Emphasis on Time Series Analysis: Particular attention is given to the applications and implications of these techniques in time series analysis, highlighting their relevance and utility in this specific area.
- Methodological Advancements: Covers the latest methodological advancements in data augmentation, providing insights into the evolving nature of these techniques.
- Real-World Applications and Cross-Domain Applicability: This review explores the practical applications and broad applicability of data augmentation techniques across various fields, highlighting their significant impact in real-world scenarios and their versatility in diverse contexts and domains.

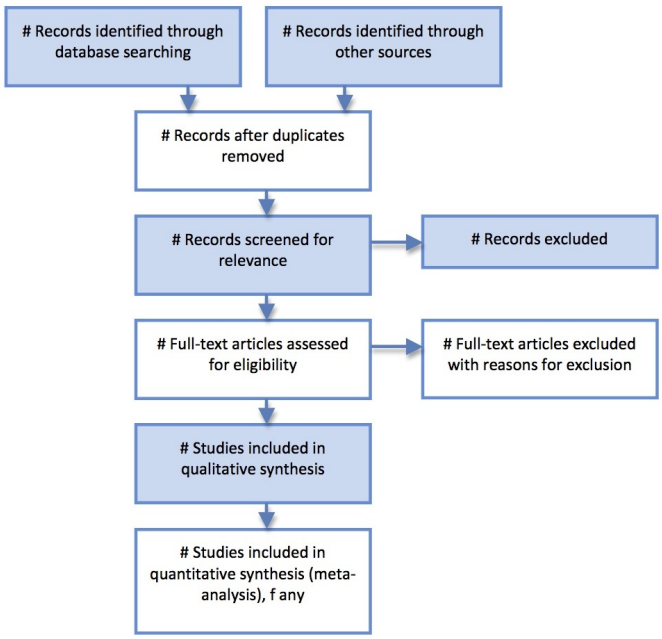

Fig. 1. PRISMA (Preferred reporting items for systematic reviews and meta-analyses) [13].

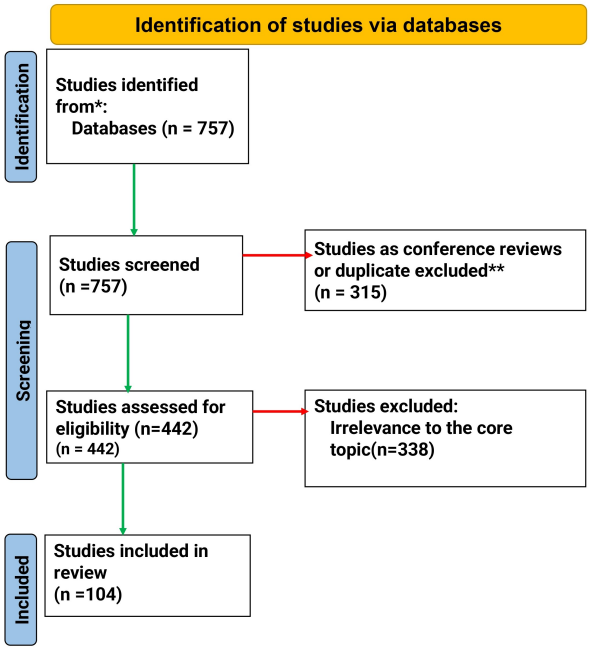

Fig. 2. PRISMA (Preferred reporting items for systematic reviews and meta-analyses) for data augmentation in time series analysis.

- Pros and Effectiveness: Highlights the advantages and effectiveness of different data augmentation techniques, demonstrating their contribution to enhancing model performance and reliability.
- Limitations and Challenges: Addresses the limitations and challenges associated with data augmentation, offering a balanced view of their capabilities and constraints.
- Future Research Directions: Outlines potential future research directions, encouraging further exploration and development in the field of data augmentation.

The review is grounded in a systematic examination of a wide range of peer-reviewed literature, adhering to the PRISMA guidelines [13] (see Fig. 1).

The paper is structured to enhance comprehension, beginning with a methodology section that details the systematic approach to literature selection and analysis. Following that, subsequent sections delve into the specifics of data augmentation techniques, their applications in various real-world scenarios, their limitations and challenges, and conclude with a discussion on future research directions.

II. Research Methodology Framework

This overview was conducted adhering to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. While a formal pre-registered protocol was not established, the methodology was meticulously developed and documented prior to initiating the review, ensuring a structured and transparent approach.

The initial dataset for this review comprised a total of 757 peer-reviewed articles and preprints, identified using the specific research query "Data Augmentation" AND "Time Series" in major academic databases including preprints. This query was designed to capture studies published between 2019 and 2024 that specifically addressed the intersection of data augmentation techniques and time series analysis in the field of machine learning. To refine the dataset for relevance and accessibility, the articles were further screened based on language and access. The final selection criteria included articles published in English and available as open access. This filtering process narrowed the dataset down to 108 articles, ensuring a focused review of studies directly relevant to the core topic and broadly accessible to the research community. Articles that did not directly respond to the research query, and publications outside the specified time frame were excluded (see Fig. 2).

The selection process entailed a rigorous screening based on titles and abstracts to assess relevance, followed by a full-text review against the inclusion criteria. The study selection process was documented using a PRISMA flow diagram, which details the number of articles screened, assessed for eligibility, and included in the final review.

Data extraction was systematically conducted, focusing on extracting key information such as study objectives, methodologies, key findings, and specific techniques related to data augmentation. The extraction process was carried out by multiple reviewers to enhance accuracy, with any discrepancies resolved through consensus. A standardized data extraction template was employed to maintain consistency across all studies.

A bias assessment was performed using established criteria to evaluate the quality and reliability of each study. This assessment considered factors such as study design, methodology, data analysis, and reporting transparency.

Given the qualitative and diverse nature of the studies, a narrative synthesis approach was utilized. This involved identifying common themes, methodologies, and findings across the studies while considering the heterogeneity of the data and study designs.

The review was based on publicly available, published academic articles; therefore, it did not involve primary data collection or require ethical approval. The analysis was conducted with respect to the intellectual property of the original authors.

III. Statistical and Machine Learning Data Augmentation Techniques

This section serves as an introduction to the diverse range of techniques encompassed by Statistical and Machine Learning Data Augmentation (see Fig. 3). It establishes the fundamental importance of data augmentation within the context of Time Series Analysis. By artificially expanding datasets and introducing variations, these techniques play a pivotal role in improving the robustness of models and the quality of insights drawn from time series data [14].

Within this subsection, we delve into the realm of statistical techniques used for data augmentation in time series analysis. Techniques such as Linear Interpolation enable the filling of gaps in data by estimating values between observed points, thus expanding datasets. Seasonal Decomposition separates time series into fundamental components, facilitating the generation of new samples by manipulating these constituent parts. Exponential Smoothing, on the other hand, focuses on forecasting future segments of time series data, effectively augmenting it with forward-looking information [15].

In this subsection, we shift our attention to Machine Learning-driven data augmentation approaches. Bootstrap Resampling enables the generation of multiple samples by randomly selecting data points with replacement, contributing to the diversification of datasets. K-Means Clustering partitions time series data into clusters based on similarity, allowing for the creation of new samples that exhibit different patterns [16]. Data Inpainting, a machine learning-based technique, aids in filling missing values by predicting them based on available data [17].

As we conclude this section, it's important to underscore the pivotal role that data augmentation plays in Time Series Analysis. By expanding datasets, improving data quality, and enabling the creation of synthetic samples, these techniques empower researchers and practitioners to extract more accurate insights from time series data [18]. The applicability of both statistical and machine learning methods underscores their relevance in a wide range of time series analysis tasks. Looking ahead, the continued development of data augmentation techniques promises to further advance the field, making it an area of ongoing interest and exploration (Table I).

IV. Deep Learning Data Augmentation Techniques

In this section, we explore advanced data augmentation techniques driven by Deep Learning models. These techniques are particularly effective in capturing complex patterns and dependencies within time series data, enabling the generation of high-quality synthetic samples.

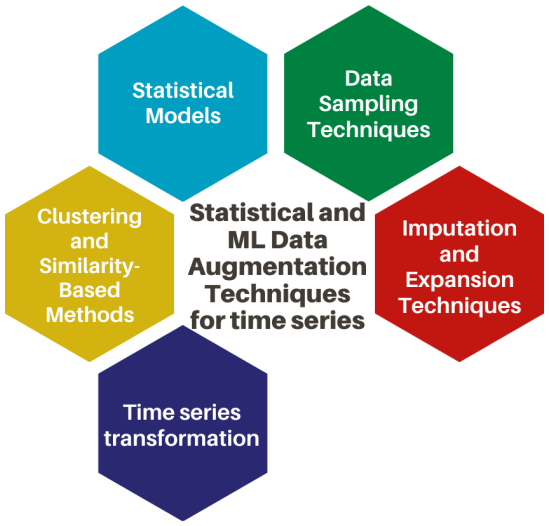

Fig. 3. Statistical and machine learning data augmentation techniques.

TABLE I. Summary of Data Augmentation Techniques in Machine Learning

Technique	Description	Reference
Imputation Techniques	Explores the use of imputation methods for augmenting incomplete time series data, including techniques like Mean, Median, KNN-based imputation, Linear Regression, Miss Forest, and MICE to fill missing values.	[14], [15], [16], [17], [18]
Data Expansion Techniques	Discusses methods for augmenting datasets by expanding time series data, including techniques for urban expansion monitoring and forecasting using remote sensing data.	[19], [20], [21], [22], [23]
Time Series Transformation	Focuses on transforming time series data using machine learning techniques for augmentation, including methods for forecasting and analysis that enhance the richness of the dataset.	[24], [25], [26], [27], [28]
Statistical Models	Examines the use of statistical models for data augmentation in time series, comparing their performance with machine learning models in applications like heart failure event prediction.	[29], [30], [31], [32], [33]
Clustering and Similarity-Based Methods	Explores the application of clustering algorithms and similarity-based methods for augmenting datasets in machine learning, including use cases like customer segmentation and data analysis.	[34], [35], [36], [37], [38]
Data Sampling Techniques	Investigates various data sampling strategies for augmenting datasets in machine learning, especially for addressing imbalanced datasets in different domains.	[39], [40], [41], [42], [43]

A. Generative Models

1) TimeGAN: TimeGAN, a generative model designed for time series data, leverages a Generative Adversarial Network (GAN) framework to generate synthetic time series data that closely resembles the original data's statistical properties and dependencies [44], [45]. It comprises two main components: the generator and the discriminator. The generator aims to produce synthetic time series data, while the discriminator tries to distinguish between real and synthetic data [46], [47].

The loss function for TimeGAN is defined as:

$$\mathcal{L}_{\text{TimeGAN}} = \lambda \cdot \mathcal{L}_{\text{AdvD}} + (1 - \lambda) \cdot \mathcal{L}_{\text{AdvG}}$$

Here, $\mathcal{L}_{\text{AdvD}}$ represents the adversarial loss for the discriminator, $\mathcal{L}_{\text{AdvG}}$ is the adversarial loss for the generator, and λ is a hyperparameter that balances the two losses [48].

2) Variational Autoencoders (VAEs): Variational Autoencoders (VAEs) are deep generative models that learn latent representations of time series data, used to generate new time series samples by sampling from the learned latent space [49], [50]. In a VAE, the encoder network maps the input time series data to a latent space where each point represents a potential data point, and the decoder network generates time series samples from points in the latent space [51], [52].

The loss function for VAEs consists of two terms: a reconstruction loss ($\mathcal{L}_{\text{rec}}$) that measures how well the generated data matches the original data and a regularization term ($\mathcal{L}_{\text{reg}}$) [53], [54]. This encourages the latent space to follow a predefined distribution, typically a Gaussian distribution. The loss is defined as:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{reg}}$$

3) Generative Adversarial Networks (GANs): Generative Adversarial Networks (GANs) consist of a generator and a discriminator network that compete during training, and they are applied to generate synthetic time series data by training the generator to produce realistic samples. In a GAN, the generator aims to produce data that is indistinguishable from real data, while the discriminator tries to distinguish between real and generated data [55], [56].

The loss function for GANs is given by:

$$\mathcal{L}_{\text{GAN}} = E_{\text{real}}[\log(D(x))] + E_{\text{fake}}[\log(1 - D(G(z)))]$$

Here, $D(x)$ represents the discriminator's output for real data, $D(G(z))$ is the discriminator's output for generated data, and z is a random noise vector [57], [58].

4) LSTM Variational Autoencoders (LSTM-VAEs): LSTM Variational Autoencoders (LSTM-VAEs) combine Long Short-Term Memory (LSTM) networks with VAEs for modeling and generating time series data, effectively capturing temporal dependencies [54], [49]. LSTM-VAEs consist of an encoder network that maps input time series data into a latent space and a decoder network that generates time series samples from points in the latent space [59], [60].

The loss function for LSTM-VAEs combines a reconstruction loss ($\mathcal{L}_{\text{rec}}$), similar to traditional VAEs, and a regularization term ($\mathcal{L}_{\text{reg}}$) that encourages the latent space to follow a predefined distribution [61]. The total loss is defined as:

$$\mathcal{L}_{\text{LSTM-VAE}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{reg}}$$

5) Temporal Generative Adversarial Networks (Temporal GANs): Temporal Generative Adversarial Networks (Temporal GANs) specialize in generating time series data while considering the temporal nature of the data. Temporal GANs extend the traditional GAN framework to handle time series data. They use recurrent layers to capture temporal dependencies and ensure that the generated data maintains the time sequence [55], [56].

The loss function for Temporal GANs is similar to the GAN loss but takes into account the sequential nature of the data, encouraging the generator to produce time-consistent samples.

6) Wasserstein Generative Models: Wasserstein Generative Models use the Wasserstein distance to measure data distribution similarity, aiming to create stable and high-quality synthetic time series data. The Wasserstein distance, also known as the Earth Mover's distance, quantifies the minimum amount of "work" required to transform one distribution into another. In the context of GANs, it provides a more stable and informative measure of the difference between real and generated data distributions [62], [63].

The loss function for Wasserstein GANs is defined as:

$$\mathcal{L}_{\text{WGAN}} = \sup_{\|D\|_L \leq 1} E_{\text{real}}[D(x)] - E_{\text{fake}}[D(G(z))]$$

Here, $D(x)$ represents the discriminator's output for real data, $D(G(z))$ is the discriminator's output for generated data, and $\|D\|_L \leq 1$ enforces a Lipschitz constraint on the discriminator.

7) Recurrent Variational Autoencoders (RNN-VAE): Recurrent Variational Autoencoders (RNN-VAE) employ recurrent neural networks (RNNs) and VAEs for modeling and generating sequential data, including time series.

RNN-VAEs incorporate RNN layers to handle sequential data and capture temporal dependencies. The encoder network maps input time series data to a latent space, and the decoder generates sequential data from points in the latent space.

The loss function for RNN-VAEs is similar to traditional VAEs, consisting of a reconstruction loss ($\mathcal{L}_{\text{rec}}$) and a regularization term ($\mathcal{L}_{\text{reg}}$) to encourage a predefined distribution in the latent space [64], [65], [66], [67], [68].

8) Conditional Generative Models: Conditional Generative Models allow for controlled generation based on specific conditions or input features.

In a conditional generative model, additional input information, known as conditions or context, is provided to the generator to influence the generation process. For example, conditions can include class labels or specific attributes that guide the generation of time series data.

The loss function for conditional generative models depends on the specific architecture and conditions used but typically involves both the reconstruction loss and a term related to the conditions used for generation [69], [70], [64], [71], [72], [73], [74], [75], [76], [77] (Table II).

B. Sequence Modeling Techniques

1) Sequence-to-Sequence Models: Sequence-to-sequence models are employed to generate new sequences based on the patterns learned from input sequences. They are widely used for time series data generation tasks [78].

TABLE II. GENERATIVE MODELS

Technique	Description	References
TimeGAN	TimeGAN is a generative model designed for time series data. It leverages a Generative Adversarial Network (GAN) framework to generate synthetic time series data that closely resembles the original data's statistical properties and dependencies.	[44], [45], [46], [47], [48].
Variational Autoencoders (VAEs)	Variational Autoencoders (VAEs) are deep generative models that can learn latent representations of time series data. They are used to generate new time series samples by sampling from the learned latent space.	[49], [50], [51], [52], [53], [54].
Generative Adversarial Networks (GANs)	Generative Adversarial Networks (GANs) consist of a generator and a discriminator network that compete during training. They can be applied to generate synthetic time series data by training the generator to produce realistic samples.	[55], [56], [57], [58].
LSTM Variational Autoencoders (LSTM-VAEs)	LSTM Variational Autoencoders (LSTM-VAEs) combine Long Short-Term Memory (LSTM) networks with VAEs for modeling and generating time series data. They are effective in capturing temporal dependencies.	[54], [49], [59], [60], [61]
Temporal Generative Adversarial Networks (Temporal GANs)	Temporal GANs are specialized GANs for generating time series data. They consider the temporal nature of the data during the generation process.	[55], [56]
Wasserstein Generative Models	Wasserstein Generative Models use the Wasserstein distance to measure the similarity between real and generated data distributions. They aim to create more stable and high-quality synthetic time series data.	[62], [63]
Recurrent Variational Autoencoders (RNN-VAE)	Recurrent Variational Autoencoders (RNN-VAE) employ recurrent neural networks (RNNs) and VAEs to model and generate sequential data, including time series.	[64], [65], [66], [67], [68]
Conditional Generative Models	Conditional generative models generate data samples based on specific conditions or input features, allowing for the controlled generation of time series data.	[69], [70], [64], [71], [72], [73], [74], [75], [76], [77]

2) Data Augmentation through Noise Addition: Data Augmentation through Noise Addition involves injecting controlled noise into the time series data to generate variations and enhance the training dataset. This approach can be represented as follows: Given an original time series $\mathbf{X} = [x_1, x_2, \ldots, x_T]$, where x_t represents the value at time t, and a noise signal $\mathbf{N} = [n_1, n_2, \ldots, n_T]$, where n_t is sampled from a predefined noise distribution, the augmented time series is obtained as $\mathbf{X}_{\text{aug}} = \mathbf{X} + \mathbf{N}$ [79].

3) Transformer Models: Transformer Models, known for their effectiveness in sequence modeling tasks, can be used to generate time series data by modeling long-range dependencies. The Transformer architecture includes self-attention mechanisms, which can capture relationships between distant time steps [80].

4) Temporal Convolutional Networks: Temporal Convolutional Networks (TCNs) utilize convolutional layers to capture temporal patterns in time series data and generate new sequences. A 1D convolutional layer with kernel size K is used to capture local patterns in TCNs [82].

V. REAL-WORLD APPLICATIONS AND USE CASES OF DATA AUGMENTATION IN TIME SERIES ANALYSIS

Data augmentation techniques have found invaluable applications in various real-world scenarios within the field of time series analysis. These methods are employed to tackle specific challenges, enhance predictive models, and enable more accurate forecasts across diverse domains.

In the realm of finance, data augmentation plays a pivotal role in generating synthetic financial time series data. This synthetic data supplements genuine financial records and is particularly useful in training predictive models for stock market analysis and portfolio management. For instance, the effectiveness of LSTM-GAN in generating synthetic time series data, achieving a close resemblance to real data with similar silhouette scores and low Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values, was demonstrated by Chen et al. [81]. Furthermore, S. Crepey et al. [82] proposed an approach to improve anomaly detection in financial time series, showing that value-at-risk estimation errors are reduced when using the proposed model. By introducing simulated market conditions and variations, data augmentation contributes to the development of robust financial models."

In the healthcare and medical research sectors, privacy regulations and limited access to patient data can pose significant hurdles. Data augmentation techniques come to the rescue by creating synthetic patient time series data. Yang et al. developed TS-GAN, a Time-series GAN based on LSTM networks, to augment sensor-based health data in healthcare. This approach significantly enhances the performance of classification models, achieving classification accuracies of 97.50% on ECG_200, 94.12% on NonInvasiveFatalECG_Thorax1, and 98.12% on mHealth datasets [83]. Furthermore, the improvement of SAX representation for time series using wavelet packet decomposition and FastDTW by Guo et al. [84] has the highest classification accuracy in 11 of 20 datasets. These artificial datasets empower the development of predictive models for disease diagnosis, patient monitoring, and drug discovery, all while safeguarding patient privacy and complying with data regulations.

Within the manufacturing and industrial domains, data augmentation strategies involve generating synthetic sensor data and introducing anomalies into existing datasets. This augmented data enhances the resilience of predictive maintenance

models, resulting in improved equipment uptime and operational efficiency. For instance, the application of simulation-based data augmentation for the quality inspection of structural adhesive with deep learning improved the performance of models in a scarce manufacturing data context with imbalanced training sets by 3.1% (mAP@0.50) [85]. Additionally, strategic data augmentation with CTGAN for smart manufacturing significantly enhanced machine learning predictions of paper breaks in pulp-and-paper production. The models' detection of machine breaks improved by over 30% for Decision Trees, 20% for Random Forest, and nearly 90% for Logistic Regression [86]. These advancements underscore data augmentation as a critical component of predictive maintenance and process optimization in industrial settings.

The energy and utilities industry leverages data augmentation to simulate energy consumption and production variations. This synthetic data aids in forecasting energy demand, optimizing grid operations, and ensuring a stable energy supply [87]Data augmentation appears to have significantly improved the forecasting accuracy in both the univariable and multivariable models. This is evident from the lower RMSE and MAPE values across all regions when comparing the augmented columns to their non-augmented counterparts. For instance, looking at the Busan region: The RMSE for the univariable model without augmentation is 0.2345, and with augmentation is 0.0853, showing a marked improvement. The RMSE for the multivariable model without augmentation is 0.1722 and with augmentation is 0.0132, which is a significant decrease. Augmented time series data contributes to effective resource management and reduced disruptions in the energy sector.

Environmental monitoring relies on data augmentation to replicate variations in environmental factors and weather conditions. Specifically, in the case of crack detection in AGR and CFD data as discussed by Branikas et al. in 2023 [88], the augmentation demonstrates a noticeable enhancement in recall and F1 score when applying a small pixel relaxation radius. Importantly, this dataset was not annotated using specialized tools or assessed by human experts. These synthetic time series datasets complement real-world observations, thereby contributing to more precise weather predictions, air quality assessments, and early detection of natural disasters. Augmentation remains a vital component in proactive environmental management and disaster preparedness.

In summary, data augmentation techniques are indispensable in time series analysis across a wide array of real-world applications and use cases. Whether in finance, healthcare, manufacturing, energy, environmental monitoring, or IoT, these methods empower the development of predictive models, improve operational efficiency, and support critical decision-making processes.

VI. Challenges and Limitations of Time Series Data Augmentation Techniques

While time series data augmentation techniques offer significant advantages in various applications, they are not without their challenges and limitations. Understanding these constraints is essential for making informed decisions when employing these methods.

A. Preservation of Temporal Dependencies

One of the primary challenges in time series data augmentation is the preservation of temporal dependencies. Many real-world time series exhibit complex dependencies and patterns over time. Data augmentation techniques must ensure that synthetic data maintains these dependencies accurately [89]. In cases where temporal structures are not adequately preserved, the performance of predictive models may degrade [90].

B. Quality of Synthetic Data

The quality of synthetic data generated through augmentation techniques is a critical concern [91]. The synthetic data should closely resemble real-world observations to ensure that predictive models trained on augmented data generalize effectively. Poorly generated synthetic data can introduce biases and inaccuracies, leading to unreliable model outcomes [92].

C. Generalization to Unseen Scenarios

Data augmentation should enable predictive models to generalize well to unseen scenarios [93]. However, there is a risk that the augmented data may be too tailored to specific training conditions, limiting the model's ability to handle novel situations [94]. Striking a balance between augmentation and maintaining generalization capabilities is a challenging task.

D. Data Privacy and Ethical Considerations

In certain domains, such as healthcare and finance, data privacy and ethical concerns pose limitations on the use of data augmentation techniques [95]. Creating synthetic patient or financial data must adhere to strict privacy regulations and ethical guidelines, which can be a complex and resource-intensive process.

E. Computational Complexity

Some advanced data augmentation techniques, particularly those involving generative models can be computationally intensive and time-consuming [96]. The computational complexity of generating large volumes of synthetic data may limit the scalability of augmentation methods.

F. Availability of Domain-Specific Augmentation Tools

The availability of domain-specific data augmentation tools and expertise can be limited [89]. Applying augmentation techniques effectively often requires domain knowledge and specialized software, which may not be readily accessible in all applications.

G. Evaluation and Validation

Evaluating the effectiveness of data augmentation methods and validating the performance of predictive models trained on augmented data can be challenging [90]. Developing appropriate evaluation metrics and conducting rigorous testing are essential but can be time and resource-intensive.

In conclusion, while time series data augmentation techniques offer numerous advantages, they also come with challenges and limitations that must be carefully considered. Addressing these limitations and understanding the constraints of each technique is crucial to ensure the successful application of data augmentation in time series analysis.

VII. Comprehensive Analysis of Data Augmentation Techniques: Advantages, Limitations, and Applicability

In the evolving landscape of machine learning and data science, data augmentation techniques play a pivotal role in enhancing model performance and reliability. These techniques are instrumental in addressing challenges such as data scarcity, imbalanced datasets, and overfitting. This section provides a thorough analysis of various data augmentation techniques, exploring their advantages, limitations, and ideal use cases.

Table III presents a comprehensive examination of both traditional and advanced data augmentation techniques, encompassing methods ranging from Imputation Techniques to cutting-edge approaches like TimeGAN, Variational Autoencoders (VAEs), and Transformer Models. The table assesses each technique's effectiveness, potential drawbacks, and the scenarios where they are most beneficial. This includes an exploration of traditional data augmentation methods as well as advanced generative models and sequence modeling techniques.

These comprehensive tables serve as a guide for researchers and practitioners to select the most appropriate data augmentation strategies, tailored to the specific needs and constraints of their machine-learning projects.

VIII. Conclusion

Time series analysis is a fundamental component of various domains, including finance, healthcare, environmental science, and more. The success of predictive models in these fields often hinges on the availability of diverse and high-quality time series data. However, obtaining such data can be challenging due to limited samples, data privacy concerns, or resource constraints. To address these challenges, data augmentation techniques have emerged as valuable tools in the time series analyst's toolkit.

In this paper, we provided an in-depth overview of data augmentation techniques in time series analysis. We explored various categories of augmentation methods, from statistical techniques to machine learning and deep learning approaches. Each category offers unique advantages and is applicable to different use cases.

Statistical techniques, such as linear interpolation, seasonal decomposition, and rolling window aggregation, provide simple and interpretable ways to augment time series data. Machine learning methods, like bootstrapping, semi-supervised learning, and time series embeddings, offer more sophisticated approaches for generating synthetic data. Deep learning techniques, including GANs, VAEs, and sequence-to-sequence models, push the boundaries of data augmentation by creating highly realistic and complex synthetic time series.

We delved into the mathematical foundations and practical applications of these techniques, showcasing their utility in tasks such as forecasting, anomaly detection, and trend analysis. Moreover, we discussed real-world use cases in finance, healthcare, and environmental monitoring, highlighting the impact of data augmentation on improving model performance and decision-making.

However, it is crucial to acknowledge that data augmentation in time series analysis is not without its challenges and limitations. Preserving temporal dependencies, ensuring data quality, and addressing computational complexity are ongoing concerns. Ethical considerations and domain-specific requirements further complicate the adoption of these techniques.

In conclusion, data augmentation techniques in time series analysis offer a promising avenue to tackle data scarcity and enhance the capabilities of predictive models. Researchers and practitioners should carefully assess the suitability of these techniques for their specific applications while being mindful of their limitations. The ever-evolving landscape of data augmentation continues to expand, opening doors to new possibilities in time series analysis and beyond.

IX. Future Research Directions

As data augmentation techniques in time series analysis continue to evolve and gain prominence, several promising avenues for future research emerge. These directions are expected to shape the field and address existing challenges while opening up new possibilities for innovation. In this section, we outline some key areas for future exploration:

- One critical area of research is the development of data augmentation methods that better preserve temporal dependencies within time series data [97].
- As data augmentation becomes more prevalent, ethical considerations surrounding the generation and use of synthetic data warrant careful examination [98].
- Expanding the applicability of data augmentation techniques to cross-domain scenarios is an exciting direction for research [99].
- Hybrid data augmentation approaches that combine statistical, machine learning, and deep learning methods offer a promising avenue for exploration [100].
- Integrating data augmentation into automated machine learning (AutoML) pipelines can streamline the model development process [101].
- Interpretable and explainable data augmentation methods are essential for building trust in augmented data and the models trained on them [102].
- Establishing standardized benchmark datasets and evaluation metrics for assessing the quality and performance of data augmentation techniques is crucial [103].
- Efforts to design resource-efficient data augmentation techniques, especially for scenarios with limited computational resources, are essential [104].

In summary, the field of data augmentation in time series analysis offers abundant opportunities for future research and innovation. Researchers and practitioners can delve into areas such as preserving temporal dependencies, addressing ethical concerns, exploring cross-domain applications, and seamlessly integrating data augmentation into AutoML processes. As data augmentation remains pivotal in enhancing time series analysis, staying at the forefront of these research directions becomes imperative to unleash its full potential.

TABLE III. ADVANTAGES, LIMITATIONS, AND APPLICABILITY OF DATA AUGMENTATION TECHNIQUES IN MACHINE AND DEEP LEARNING

Technique	Advantages	Limitations	Applicability
Imputation Techniques	- Can effectively handle missing data, improving dataset completeness. - Offers a variety of methods suitable for different data types and patterns.	- Risk of introducing bias or inaccuracies, especially if the imputation model doesn't align well with the data's nature. - Might oversimplify complex data relationships.	- Best used when dealing with datasets having missing values, especially in cases where the data is crucial and cannot be discarded.
Data Expansion Techniques	- Allows for the creation of larger and more diverse datasets. - Particularly useful in fields like remote sensing where data can be scarce.	- Expanded data might not always represent real-world scenarios accurately. - Risk of introducing artificial patterns not present in the original dataset.	- Ideal for situations where the available dataset is too small or lacks diversity, such as in certain types of research or specialized applications.
Time Series Transformation	- Enhances the diversity and richness of data, leading to potentially better model performance. - Useful for both forecasting and deeper data analysis.	- Transformation techniques can distort the original time series properties. - Requires careful selection to ensure relevance and accuracy.	- Suitable for time series forecasting, especially when the goal is to reveal hidden patterns or to adapt data to specific analytical needs.
Statistical Models	- Provides a more traditional and often simpler approach to data augmentation. - Good for understanding underlying data distributions.	- May not capture complex nonlinear relationships as effectively as more advanced machine learning models. - Limited flexibility in handling diverse data types.	- Recommended for scenarios where a straightforward, interpretable approach is needed, particularly in fields with well-understood data distributions.
Clustering and Similarity-Based Methods	- Useful for discovering natural groupings and patterns in data. - Can improve data organization and segmentation.	- Performance is heavily dependent on the choice of similarity measures. - Can be sensitive to outliers and noise in the data.	- Best applied in data segmentation, customer profiling, or any scenario requiring the identification of inherent groupings in the data.
Data Sampling Techniques	- Effective in addressing imbalanced datasets, and enhancing model training. - Various strategies available to suit different data scenarios.	- Risks include overfitting, underfitting, or introducing sampling bias. - This may lead to loss of important information if not carefully implemented.	- Particularly useful in cases of imbalanced datasets, such as in fraud detection or rare event prediction, where certain classes are underrepresented.
TimeGAN	- Excellent for capturing temporal dynamics in time series. - Generates data that closely resembles real statistical properties.	- Computationally intensive. - Requires large amounts of training data for accuracy.	- Ideal for scenarios where authentic-like time series data generation is needed, such as financial market analysis.
Variational Autoencoders (VAEs)	- Good at learning complex distributions. - Capable of generating diverse data samples.	- Can struggle with generating high-quality reconstructions. - Somewhat complex to train and tune.	- Suitable for tasks requiring the generation of new samples from complex data distributions, like image or speech synthesis.
Generative Adversarial Networks (GANs)	- Can produce highly realistic synthetic data. - Versatile for various data types.	- Training can be unstable. - Prone to mode collapse.	- Best for applications where realistic data generation is crucial, such as art creation or data augmentation.
LSTM Variational Autoencoders (LSTM-VAEs)	- Effective in modeling time dependencies. - Combines LSTM's sequence handling with VAE's generative capabilities.	- Risk of overfitting on smaller datasets. - Complex model architecture.	- Useful in sequential data applications like anomaly detection in time series.
Temporal Generative Adversarial Networks (Temporal GANs)	- Specifically designed for time series data. - Addresses temporal aspects effectively.	- Can be computationally demanding. - Requires careful tuning and training.	- Ideal for generating time-dependent synthetic data, such as in healthcare or stock market prediction.
Wasserstein Generative Models	- Offers more stable training than traditional GANs. - Better at handling data distribution.	- More challenging to implement. - Can be computationally more intensive.	- Recommended for scenarios where stable training of generative models is a priority, like in large-scale data generation.
Recurrent Variational Autoencoders (RNN-VAE)	- Good for sequential data representation. - Combines RNN's temporal modeling with VAE's generative properties.	- Training can be time-consuming. - Susceptible to vanishing gradients problem.	- Suitable for generating complex time series or sequential data, such as in natural language processing.
Conditional Generative Models	- Allows control over generated data features. - Highly versatile in data generation.	- Requires additional conditioning data. - Increased model complexity.	- Best used when specific conditions or features need to be included in the generated data, like in targeted marketing campaigns.
Sequence-to-Sequence Models	- Effective for generating sequences based on learned patterns. - Widely applicable in time series generation.	- Requires large amounts of data for accuracy. - Can be complex to tune and optimize.	- Ideal for applications like machine translation, speech recognition, and time series forecasting.
Data Augmentation through Noise Addition	- a simple and effective way to create data variations. - Enhances the robustness of models.	- Risk of distorting the original data too much. - Noise parameters need to be carefully chosen.	- Useful in scenarios where minor variations in the dataset can lead to significant improvements, such as in image or signal processing.
Transformer Models	- Excellent at capturing long-range dependencies. - Self-attention mechanism provides dynamic focus.	- Can be resource-intensive. - Requires significant amounts of training data.	- Suitable for complex sequence modeling tasks like natural language understanding and time series analysis.
Temporal Convolutional Networks (TCNs)	- Effective in capturing local and global temporal patterns. - Efficient in terms of computational resources.	- May miss intricate long-term dependencies. - Architecture needs a careful design for specific tasks.	- Recommended for tasks like audio synthesis and real-time anomaly detection in time series data.

However, it's crucial to acknowledge certain limitations in our comprehensive overview. Our scope may not cover all existing techniques, and the diverse nature of time series data, along with the choice of evaluation metrics, may limit generalizability. Overfitting risks, the ever-evolving research landscape, interdisciplinary variations, and data accessibility issues are additional factors that deserve attention. Despite these challenges, our goal was to furnish a balanced and informative overview, serving as a valuable guide for both researchers and practitioners in the field.

REFERENCES

[1] M. Rahman, M. Rivolta, F. Badilini, R. Sassi, "A Systematic Survey of Data Augmentation of ECG Signals for AI Applications," *Sensors*, vol. 23, no. 11, 2023. doi:10.3390/s23115237

[2] N. A. Andriyanov, D. Andriyanov, "The using of data augmentation in machine learning in image processing tasks in the face of data scarcity," *Journal of Physics: Conference Series*, vol. 1661, no. 1, 2020. doi:10.1088/1742-6596/1661/1/012018

[3] S. Aleem, T. Kumar, S. Little, M. Bendechache, R. Brennan, K. McGuinness, "Random Data Augmentation based Enhancement: A Generalized Enhancement Approach for Medical Datasets," *Frontiers in Medicine*, 2022. doi:10.56541/fumf3414

[4] C. M. Burlacu, A. Burlacu, M. Praisler, C. Paraschiv, "Harnessing Deep Convolutional Neural Networks Detecting Synthetic Cannabinoids: A Hybrid Learning Strategy for Handling Class Imbalances in Limited Datasets," *Inventions*, vol. 8, no. 5, 2023. doi:10.3390/inventions8050129

[5] C.-Y. Hsu, P.-Y. Chen, S. Lu, S. Liu, C.-M. Yu, "Adversarial Examples Can Be Effective Data Augmentation for Unsupervised Machine Learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 6, 2021. doi:10.1609/aaai.v36i6.20650

[6] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, Q. V. Le, "Unsupervised Data Augmentation for Consistency Training," *arXiv*, 2019. [Online]. Available: https://arxiv.org/abs/1904.12848

[7] J. Yoo, T. Zhao, L. Akoglu, "Understanding the Effect of Data Augmentation in Self-supervised Anomaly Detection," *arXiv*, 2022. doi:10.48550/arXiv.2208.07734

[8] B. K. Iwana, S. Uchida, "Time Series Data Augmentation for Neural Networks by Time Warping with a Discriminative Teacher," *IEEE International Conference on Pattern Recognition*, 2020. doi:10.1109/ICPR48806.2021.9412812

[9] A. Aboussalah, M. Kwon, R. G. Patel, C. Chi, C.-G. Lee, "Don't overfit the history - Recursive time series data augmentation," *arXiv*, 2022. doi:10.48550/arXiv.2207.02891

[10] X. Yang, Z. Zhang, X. Cui, R.-y. Cui, "A Time Series Data Augmentation Method Based on Dynamic Time Warping," *IEEE Conference*, 2021. doi:10.1109/CCAI50917.2021.9447507

[11] I. Pastaltzidis, N. Dimitriou, K. Quezada-Tavárez, S. Aidinlis, T. Marquenie, A. Gurzawska, D. Tzovaras, "Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems," *ACM Conference*, 2022. doi:10.1145/3531146.3534644

[12] J. Yuan, R. Tang, X. Jiang, X. Hu, "LLM for Patient-Trial Matching: Privacy-Aware Data Augmentation Towards Better Performance and Generalizability," *arXiv*, 2023. doi:10.48550/arXiv.2303.16756

[13] M. Zuccon, E. Topino, A. Musetti, A. Gori, "Psychodynamic Therapies for the Treatment of Substance Addictions: A PRISMA Meta-Analysis," *Journal of Personalized Medicine*, vol. 13, no. 10, 2023. doi:10.3390/jpm13101469

[14] E. A. Christobel, "Imputation Techniques in Machine Learning – A Survey," *International Journal of Recent Technology and Engineering*, vol. 11, no. 10, 2023. doi:10.17762/ijritcc.v11i10.8662

[15] H. Wang et al., "Application of machine learning missing data imputation techniques in clinical decision making," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, 2022. doi:10.1186/s12911-022-01752-6

[16] A. R. Ismail, N. Z. Abidin, and M. Maen, "Systematic Review on Missing Data Imputation Techniques with Machine Learning Algorithms for Healthcare," *Journal of Robotics and Control (JRC)*, vol. 3, no. 2, 2022. doi:10.18196/jrc.v3i2.13133

[17] V. P. C. Magboo et al., "Imputation Techniques and Recursive Feature Elimination in Machine Learning Applied to Type II Diabetes Classification," *ACM International Conference Proceeding Series*, 2021. doi:10.1145/3508259.3508288

[18] T. Thomas and E. Rajabi, "A systematic review of machine learning-based missing value imputation techniques," *Data Technologies and Applications*, vol. 55, no. 3, 2021. doi:10.1108/DTA-12-2020-0298

[19] E. Mostafa et al., "Monitoring and Forecasting of Urban Expansion Using Machine Learning-Based Techniques and Remotely Sensed Data: A Case Study of Gharbia Governorate, Egypt," *Remote Sensing*, vol. 13, no. 22, 2021. doi:10.3390/rs13224498

[20] A. Agnihotri et al., "Role of data mining and machine learning techniques in medical imaging," *International Journal of Advanced Intelligence Paradigms*, vol. 16, no. 1/2, 2020. doi:10.1504/IJAIP.2018.10017086

[21] S. Fha et al., "Development of an Efficient Method to Detect Mixed Social Media Data with Tamil-English Code Using Machine Learning Techniques," *ACM International Conference Proceeding Series*, 2022. doi:10.1145/3563775

[22] A. Nafees et al., "Forecasting the Mechanical Properties of Plastic Concrete Employing Experimental Data Using Machine Learning Algorithms: DT, MLPNN, SVM, and RF," *Polymers*, vol. 14, no. 8, 2022. doi:10.3390/polym14081583

[23] E. Aharoni et al., "HE-PEx: Efficient Machine Learning under Homomorphic Encryption using Pruning, Permutation and Expansion," *arXiv preprint arXiv:2207.03384*, 2022. doi:10.48550/arXiv.2207.03384

[24] C. Kubik et al., "Knowledge discovery from time series in engineering applications using machine learning techniques," *Journal of Manufacturing Science and Engineering, Transactions of the ASME*, vol. 144, no. 3, 2022. doi:10.1115/1.4054158

[25] D. Salwala et al., "Distributed Incremental Machine Learning for Big Time Series Data," *IEEE International Conference on Big Data (Big Data)*, 2022. doi:10.1109/BigData55660.2022.10020361

[26] F. Ott et al., "Domain Adaptation for Time-Series Classification to Mitigate Covariate Shift," *ACM International Conference Proceeding Series*, 2022. doi:10.1145/3503161.3548167

[27] "The composition of time-series images and using the technique SMOTE ENN for balancing datasets in land use/cover mapping," *Applied Mathematics and Sciences: An International Journal*, vol. 27, no. 2, 2022. doi:10.46544/ams.v27i2.05

[28] S. Kuili et al., "A holistic machine learning approach to identify performance anomalies in enterprise WiFi deployments," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 12118, 2022. doi:10.1117/12.2621087

[29] Z. Sun et al., "Comparing Machine Learning Models and Statistical Models for Predicting Heart Failure Events: A Systematic Review and Meta-Analysis," *Frontiers in Cardiovascular Medicine*, vol. 9, 2022. doi:10.3389/fcvm.2022.812276

[30] Y. Li et al., "Consistency of variety of machine learning and statistical models in predicting clinical risks of individual patients: longitudinal cohort study using cardiovascular disease as exemplar," *BMJ (Clinical research ed.)*, vol. 371, 2020. doi:10.1136/bmj.m3919

[31] D. Zhou et al., "Integration of machine learning and statistical models for crash frequency modeling," *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, 2022. doi:10.1080/19427867.2022.2158257

[32] M. S. Jaafarzadeh et al., "Groundwater recharge potential zonation using an ensemble of machine learning and bivariate statistical models," *Scientific Reports*, vol. 11, no. 1, 2021. doi:10.1038/s41598-021-85205-6

[33] X. Dastile et al., "Statistical and machine learning models in credit scoring: A systematic literature survey," *Applied Soft Computing*, vol. 96, 2020. doi:10.1016/j.asoc.2020.106263

[34] L. Kwuida and D. Ignatov, "On Interpretability and Similarity in Concept-Based Machine Learning," *Advances in Intelligent Systems and Computing*, vol. 1260, 2021. doi:10.1007/978-3-030-72610-2_3

[35] K. Shahina and T. P. Kumar, "Similarity-based clustering and data aggregation with independent component analysis in wireless sensor networks," *Transactions on Emerging Telecommunications Technologies*, 2022. doi:10.1002/ett.4462

[36] H. Zhu et al., "Assessment of the Generalization Abilities of Machine-Learning Scoring Functions for Structure-Based Virtual Screening," *Journal of Chemical Information and Modeling*, 2022. doi:10.1021/acs.jcim.2c01149

[37] D. M. S. Shakoor et al., "A Machine Learning Recommender System Based on Collaborative Filtering Using Gaussian Mixture Model Clustering," *Authorea Preprints*, 2020. doi:10.22541/au.160897179.93005705/v1

[38] K. Polat, "Similarity-based attribute weighting methods via clustering algorithms in the classification of imbalanced medical datasets," *Neural Computing and Applications*, 2018. doi:10.1007/s00521-018-3471-8

[39] M. Imani and H. Arabnia, "Hyperparameter Optimization and Combined Data Sampling Techniques in Machine Learning for Customer Churn Prediction: A Comparative Analysis," *Preprints*, 2023. doi:10.20944/preprints202308.1478.v1

[40] H. M. Abdelghany and K. Hooker, "Effective data sampling techniques for machine learning OPC in full chip production," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 11611, 2021. doi:10.1117/12.2586176

[41] C. Xie et al., "Effect of machine learning re-sampling techniques for imbalanced datasets in 18F-FDG PET-based radiomics model on prognostication performance in cohorts of head and neck cancer patients," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 47, no. 12, 2020. doi:10.1007/s00259-020-04756-4

[42] R. Gupta et al., "Diagnosis of Breast Cancer on Imbalanced Dataset Using Various Sampling Techniques and Machine Learning Models," *2021 International Conference on Digital Ecosystems and Technologies (DEST)*, 2021. doi:10.1109/DeSE54285.2021.9719398

[43] A. S. Dina et al., "Effect of Balancing Data Using Synthetic Data on the Performance of Machine Learning Classifiers for Intrusion Detection in Computer Networks," *IEEE Access*, vol. 10, 2022. doi:10.1109/ACCESS.2022.3205337

[44] H. Shi, Y. Xu, B. Ding, J. Zhou, and P. Zhang, "Long-Term Solar Power Time-Series Data Generation Method Based on Generative Adversarial Networks and Sunrise–Sunset Time Correction," *Sustainability*, vol. 15, no. 20, 2023. doi:10.3390/su152014920

[45] L. Mushunje, D. Allen, and S. Peiris, "Volatility and irregularity Capturing in stock price indices using time series Generative adversarial networks (TimeGAN)," *arXiv preprint arXiv:2311.12987*, 2023. doi:10.48550/arXiv.2311.12987

[46] C.-Y. Tai, W.-J. Wang, and Y.-M. Huang, "Using Time-Series Generative Adversarial Networks to Synthesize Sensing Data for Pest Incidence Forecasting on Sustainable Agriculture," *Sustainability*, vol. 15, no. 10, 2023. doi:10.3390/su15107834

[47] A. A. Purwita, A. Yesilkaya, and H. Haas, "Synthetic LiFi Channel Model Using Generative Adversarial Networks," *IEEE International Conference on Communications (ICC)*, 2022. doi:10.1109/ICC45855.2022.9838481

[48] S. Chattoraj, S. Pratiher, S. Pratiher, and H. Konik, "Improving Stability of Adversarial Li-ion Cell Usage Data Generation using Generative Latent Space Modelling," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021. doi:10.1109/ICASSP39728.2021.9413892

[49] C. Zhang and Y. Chen, "Time Series Anomaly Detection with Variational Autoencoders," 2019. doi:10.1109/ICMLA.2018.00207

[50] V. Fortuin, G. Rätsch, and S. Mandt, "Multivariate Time Series Imputation with Variational Autoencoders," 2019. doi:10.1109/ICMLA.2018.00207

[51] J. Li, W. Ren, and M. Han, "Mutual Information Variational Autoencoders and Its Application to Feature Extraction of Multivariate Time Series," 2022. doi:10.1142/s0218001422550059

[52] W. Todo, B. Laurent, J.-M. Loubes, and M. Selmani, "Dimension Reduction for time series with Variational AutoEncoders," 2022. doi:10.48550/arXiv.2204.11060

[53] G. G. González, P. Casas, A. Fernández, and G. Gómez, "Steps towards continual learning in multivariate time-series anomaly detection using variational autoencoders," 2022. doi:10.1145/3517745.3563033

[54] M. L. Garsdal, V. Sogaard, and S. M. Sørensen, "Generative time series models using Neural ODE in Variational Autoencoders," 2022. doi:Not available

[55] D. Li, D. Chen, L. Shi, B. Jin, J. Goh, and S.-K. Ng, "MAD-GAN: Multivariate Anomaly Detection for Time Series Data with Generative Adversarial Networks," 2019. doi:10.1007/978-3-030-30490-4_56

[56] W. Cheng, T. Ma, X. Wang, and G. Wang, "Anomaly Detection for Internet of Things Time Series Data Using Generative Adversarial Networks With Attention Mechanism in Smart Agriculture," 2022. doi:10.3389/fpls.2022.890563

[57] Z. Thompson, A. Downey, J. D. Bakos, and J. Wei, "Synthesizing Dynamic Time-series Data for Structures Under Shock Using Generative Adversarial Networks," 2022.

[58] K. Sarda, A. Yerudkar, and C. D. Vecchio, "Missing Data Imputation for Real Time-series Data in a Steel Industry using Generative Adversarial Networks," 2021. doi:10.1109/IECON48115.2021.9589716

[59] A. Takiddin, M. Ismail, U. Zafar, and E. Serpedin, "Deep Autoencoder-Based Anomaly Detection of Electricity Theft Cyberattacks in Smart Grids," 2022. doi:10.1109/JSYST.2021.3136683

[60] X. Jin, W. Gong, J. Kong, Y.-t. Bai, and T. Su, "PFVAE: A Planar Flow-Based Variational Auto-Encoder Prediction Model for Time Series Data," 2022. doi:10.3390/math10040610

[61] T. Kieu, B. Yang, C. Guo, R.-G. Cirstea, Y. Zhao, Y.-h. Song, and C. S. Jensen, "Anomaly Detection in Time Series with Robust Variational Quasi-Recurrent Autoencoders," 2022. doi:10.1109/icde53745.2022.00105

[62] A. Bouteska, M. Lavazza Seranto, p. hajek, and M. Z. Abedin, "Data-driven decadal climate forecasting using Wasserstein time-series generative adversarial networks," 2023. doi:10.1007/s10479-023-05722-7

[63] X. Hu, H. Zhang, D. Ma, and R. Wang, "Hierarchical Pressure Data Recovery for Pipeline Network via Generative Adversarial Networks," 2022. doi:10.1109/TASE.2021.3069003

[64] D. Wang, Y. Yan, R. Qiu, Y. Zhu, K. Guan, A. Margenot, and H. Tong, "Networked Time Series Imputation via Position-aware Graph Enhanced Variational Autoencoders," 2023. doi:10.1145/3580305.3599444

[65] F. Romanelli and F. Martinelli, "Synthetic Sensor Measurement Generation With Noise Learning and Multi-Modal Information," 2023. doi:10.1109/ACCESS.2023.3323038

[66] H. Qin, L. Su, C. Jiang, C. Zhang, G. Wu, and Y. Zhang, "Time Series Data Augmentation Algorithm Combining Deep Metric Learning and Variational Encoder," 2023. doi:10.1109/ICPSAsia58343.2023.10294708

[67] H. Li, S. Yu, and J. Príncipe, "Causal Recurrent Variational Autoencoder for Medical Time Series Generation," 2023. doi:10.48550/arXiv.2301.06574

[68] A. Siahkoohi, R. Morel, R. Balestriero, E. Allys, G. Sainton, T. Kawamura, and M. V. de Hoop, "Martian time-series unraveled: A multi-scale nested approach with factorial variational autoencoders," 2023. doi:10.48550/arXiv.2305.16189

[69] F. Altekrüger, P. Hagemann, and G. Steidl, "Conditional Generative Models are Provably Robust: Pointwise Guarantees for Bayesian Inverse Problems," 2023. doi:10.48550/arXiv.2303.15845

[70] P. Kashyap, C. Cheng, Y. Choi, and P. D. Franzon, "Generative Multi-Physics Models for System Power and Thermal Analysis Using Conditional Generative Adversarial Networks," 2023. doi:10.1109/EPEPS58208.2023.10314864

[71] X. Yuan, K. Chen, J. Zhang, W. Zhang, N. H. Yu, and Y. Zhang, "Pseudo Label-Guided Model Inversion Attack via Conditional Generative Adversarial Network," 2023. doi:10.48550/arXiv.2302.09814

[72] A. Elhagry, "Text-to-Metaverse: Towards a Digital Twin-Enabled Multimodal Conditional Generative Metaverse," 2023. doi:10.1145/3581783.3613432

[73] N. Vyas, S. Kakade, and B. Barak, "On Provable Copyright Protection for Generative Models," 2023. doi:10.48550/arXiv.2302.10870

[74] A. Heng and H. Soh, "Selective Amnesia: A Continual Learning Approach to Forgetting in Deep Generative Models," 2023. doi:10.48550/arXiv.2305.10120

[75] A. Tong, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, K. Fatras, G. Wolf, and Y. Bengio, "Improving and generalizing flow-based generative models with minibatch optimal transport," 2023. doi:10.48550/arXiv.2302.00482

[76] D. Ye, X. Wang, and X. Chen, "Lightweight Generative Joint Source-Channel Coding for Semantic Image Transmission with Compressed Conditional GANs," 2023. doi:10.1109/ICCCWorkshops57813.2023.10233814

[77] A. A. Xu, S. Han, X. Ju, and H. Wang, "Generative Machine Learning for Detector Response Modeling with a Conditional Normalizing Flow," 2023. doi:10.48550/arXiv.2303.10148

[78] S. Liao, H. Ni, M. Sabaté-Vidales, L. Szpruch, M. Wiese, and B. Xiao, "Sig-Wasserstein GANs for conditional time series generation," 2023. doi:10.1111/mafi.12423

[79] H. Zhang, Z. Pang, J. Wang, and T. Li, "Few-shot Learning using Data Augmentation and Time-Frequency Transformation for Time Series Classification," 2023. doi:10.48550/arXiv.2311.03194

[80] M. F. Sikder, R. Ramachandranpillai, and F. Heintz, "TransFusion: Generating Long, High Fidelity Time Series using Diffusion Models with Transformers," 2023. doi:10.48550/arXiv.2307.12667

[81] Chen et al., "Data Augmentation for Pseudo-Time Series Using Generative Adversarial Networks," 2023. [Online]. Available: https://dblp.org/rec/conf/itat/SalmiJ23

[82] S. Crepey et al., "Anomaly Detection in Financial Time Series by Principal Component Analysis and Neural Networks," *Algorithms*, vol. 15, no. 10, p. 335, 2022. [Online]. Available: https://www.mdpi.com/1999-4893/15/10/335

[83] Z. Yang, Y. Li, G. Zhou, "TS-GAN: Time-series GAN for Sensor-based Health Data Augmentation," 2023. doi:10.1145/3583593

[84] P. Guo, H. Yang, A. Sano, "Empirical Study of Mix-based Data Augmentation Methods in Physiological Time Series Data," 2023. doi:10.1109/ICHI57859.2023.00037

[85] R. Peres, A. Gomes, J. Mendes, and S. Bento, "Simulation-Based Data Augmentation for the Quality Inspection of Structural Adhesive With Deep Learning," *IEEE Access*, vol. 9, pp. 44326-44335, 2021. doi:10.1109/ACCESS.2021.3069452

[86] H. Khosravi, S. Farhadpour, M. Grandhi, A. S. Raihan, S. Das, and I. Ahmed, "Strategic Data Augmentation with CTGAN for Smart Manufacturing: Enhancing Machine Learning Predictions of Paper Breaks in Pulp-and-Paper Production," *CoRR*, vol. abs/2311.09333, 2023. doi:10.48550/ARXIV.2311.09333

[87] J. Chung, B. Jang, "Accurate prediction of electricity consumption using a hybrid CNN-LSTM model based on multivariable data," 2022. doi:10.1371/journal.pone.0278071

[88] E. Branikas, P. Murray, G. West, "A Novel Data Augmentation Method for Improved Visual Crack Detection Using Generative Adversarial Networks," 2023. doi:10.1109/ACCESS.2023.3251988

[89] H. Qin, L. Su, C. Jiang, C. Zhang, G. Wu, and Y. Zhang, "Time Series Data Augmentation Algorithm Combining Deep Metric Learning and Variational Encoder," *Proc. ICPSAsia*, 2023. doi:10.1109/ICPSAsia58343.2023.10294708

[90] Z. Cai, W. Ma, X. Wang, H. Wang, and Z. Feng, "The Performance Analysis of Time Series Data Augmentation Technology for Small Sample Communication Device Recognition," *IEEE Transactions on Robotics*, vol. 39, no. 1, pp. 5-15, 2023. doi:10.1109/TR.2022.3178707

[91] B. Shen, L. Yao, X. Jiang, Z. Yang, and J.-s. Zeng, "Time Series Data Augmentation Classifier for Industrial Process Imbalanced Fault Diagnosis," *Proc. DDCLS*, 2023. doi:10.1109/DDCLS58216.2023.10166336

[92] Y. Gao, C. A. Ellis, V. Calhoun, and R. L. Miller, "Improving age prediction: Utilizing LSTM-based dynamic forecasting for data augmentation in multivariate time series analysis," *arXiv preprint arXiv:2312.08383*, 2023.

[93] A. Wilf, A. T. Xu, P. Liang, A. Obolenskiy, D. Fried, and L.-P. Morency, "Comparative Knowledge Distillation," *arXiv preprint arXiv:2311.02253*, 2023.

[94] K. Rath, D. Rügamer, B. Bischl, U. von Toussaint, C. Rea, A. D. Maris, R. Granetz, and C. Albert, "Data augmentation for disruption prediction via robust surrogate models," *Journal of Plasma Physics*, vol. 88, no. 4, 2022. doi:10.1017/S0022377822000769

[95] Y. Kwak and J.-g. Huh, "Random Augmentation Technique for Mitigating Overfitting in Neural Networks for Financial Time Series Forecasting," *Journal of Korean Data Analysis Society*, vol. 25, no. 5, pp. 1653-1664, 2023. doi:10.37727/jkdas.2023.25.5.1653

[96] M. Li and B. C. Lovell, "End to End Generative Meta Curriculum Learning for Medical Data Augmentation," in *Proc. ICIP*, 2022. doi:10.1109/ICIP49359.2023.10222093

[97] J. C. Lin and F. Yang, "Data Augmentation for Industrial Multivariate Time Series via a Spatial and Frequency Domain Knowledge GAN," in *Proc. AdCONIP*, 2022. doi:10.1109/AdCONIP55568.2022.9894177

[98] J. Yuan, R. Tang, X. Jiang, and X. Hu, "LLM for Patient-Trial Matching: Privacy-Aware Data Augmentation Towards Better Performance and Generalizability," *arXiv preprint arXiv:2303.16756*, 2023.

[99] M. Rahul and S. Chiddarwar, "A causality-inspired data augmentation approach to cross-domain burr detection using randomly weighted shallow networks," *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 2, pp. 569-582, 2023. doi:10.1007/s13042-023-01891-w

[100] A. Gong, X. Zhang, Y. Wang, Y. Zhang, and M. Li, "Hybrid Data Augmentation and Dual-Stream Spatiotemporal Fusion Neural Network for Automatic Modulation Classification in Drone Communications," *Drones*, vol. 7, no. 6, 2023. doi:10.3390/drones7060346

[101] T. Döhmen, M. Hulsebos, C. Beecks, and S. Schelter, "GitSchemas: A Dataset for Automating Relational Data Preparation Tasks," in *Proc. ICDEW*, 2022. doi:10.1109/icdew55742.2022.00016

[102] H. Chen and Y. Ji, "Improving the Interpretability of Neural Sentiment Classifiers via Data Augmentation," in *Proc. EMNLP-IJCNLP*, 2019.

[103] P. Katiyar and A. Khoreva, "Improving Augmentation and Evaluation Schemes for Semantic Image Synthesis," *arXiv preprint arXiv:2011.12636*, 2020.

[104] F. Xie, H. Wen, J. Wu, W. Hou, H.-h. Song, T. Zhang, R. Liao, and Y. Jiang, "Data Augmentation for Radio Frequency Fingerprinting via Pseudo-Random Integration," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, no. 5, pp. 594-604, 2020. doi:10.1109/TETCI.2019.2907740