SC-β-VAE-GAN: A SHIFT CORRECTION VAE-GAN MODEL FOR IMPUTATION AND AUGMENTATION OF HANDWRITING MULTIVARIATE TIME SERIES DATA

A Thesis
Presented to the Faculty of
College of Computer and Information Sciences
Polytechnic University of the Philippines
Sta. Mesa, Manila

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

by

Alpapara, Nichole
Lagatuz, John Patrick
Peroche, John Mark
Torreda, Kurt Denver

June 2024

APPROVAL SHEET

The thesis ***"SC-β-VAE-GAN: A SHIFT CORRECTION VAE-GAN MODEL FOR IMPUTATION AND AUGMENTATION OF HANDWRITING MULTIVARIATE TIME SERIES DATA"*** submitted and presented by Nichole N. Alpapara, John Patrick D. Lagatuz, John Mark P. Peroche, and Kurt Denver P. Torreda, in partial fulfillment for the degree of Bachelor in Science in Computer Science has been

Recommended for Approval and Acceptance: Date:

Michael B. Dela Fuente, MSGITS

Faculty-in-Charge

Approved by the Thesis Defense Panel:

Prof. Lydinar D. Dastas

Prof. Illuminada Vivien R. Domingo

Prof. Angelica P. Payne

Accepted for the Department of Computer Science:

Montaigne G. Molejon, MSIT

Department Chairperson

Accepted for the College of Computer and Information Science:

Melvin C. Roxas, MSGITS

Dean of College

ACKNOWLEDGEMENT

We would like to express our gratitude to everyone who contributed in making our thesis possible. Their time and effort truly mean so much to us.

First and foremost, we are very grateful to our thesis adviser, **Prof. Michael B. Dela Fuente, MSGITS**. His guidance and patience in answering our questions made it a lot more manageable to progress when doing our thesis. We are really fortunate to have him as our adviser.

We also want to thank our panelists, **Prof. Lydinar D. Dastas, MBE, MSCS, Prof. Illuminada Vivien R. Domingo,** and **Prof. Angelica P. Payne** for their comments and recommendations which helped improve our study and made us better researchers. We are grateful for them sharing their time and expertise with us.

Lastly, we express our deepest thanks to our families, friends, and colleagues for believing in us. Their support gave us the motivation to reach this milestone. This achievement would not have been possible without all of them. We truly appreciate their support.

N.N.A

J.P.D.L

J.M.P.P

K.D.P.T

CERTIFICATION OF ORIGINALITY

This is to certify that the research work presented in this thesis, *SC-β-VAE-GAN: A SHIFT CORRECTION VAE-GAN MODEL FOR IMPUTATION AND AUGMENTATION OF HANDWRITING MULTIVARIATE TIME SERIES DATA* for the degree Bachelor of Science in Computer Science at the Polytechnic University of the Philippines embodies the result of original and scholarly work carried out by the undersigned. This thesis does not contain words or ideas taken from published sources or written works that have been accepted as basis for the award of a degree from any other higher education institution, except where proper referencing and acknowledgement were made.

NICHOLE N. ALPAPARA

Researcher

JOHN PATRICK D. LAGATUZ

Researcher

JOHN MARK P. PEROCHE

Researcher

KURT DENVER P. TORREDA

Researcher

____

Date Signed

ABSTRACT

Title	:	SC-β-VAE-GAN: A Shift Correction VAE-GAN Model for Imputation and Augmentation of Handwriting Multivariate Time Series Data
Researchers	:	Nichole N. Alpapara, John Patrick D. Lagatuz, John Mark P. Peroche, Kurt Denver P. Torreda
Degree	:	Bachelor of Science in Computer Science
Institution	:	Polytechnic University of the Philippines
Year	:	2025
Adviser	:	Michael B. dela Fuente, MSGITS

Handwriting analysis offers a cost-effective and non-intrusive method for emotion recognition, but its adoption is limited because of challenges like insufficient data and missing in-air handwriting data. This study addresses these problems by introducing SC-β-VAE-GAN, a hybrid generative model combining Variational Autoencoder (VAE) and Generative Adversarial Network (GAN) with shift correction and beta regularization. This model is designed for imputation and augmentation of multivariate handwriting time series data. A quantitative quasi-experimental design was used to evaluate SC-β-VAE-GAN against the baseline models (VAEGAN, TimeGAN, and VRNNGAN) using the EMOTHAW handwriting dataset. Performance metrics included Normalized Root Mean Square Error (NRMSE), Post-Hoc Discriminative Score, and Post-Hoc Predictive Score. The findings revealed that SC-β-VAE-GAN significantly outperformed other models, achieving an NRMSE of 0.92%, a Post-Hoc Discriminative Score of 50.79%, and a Post-Hoc Predictive Score of 5.35%, showing high reconstruction accuracy, realism, and temporal consistency. These demonstrated the model's effectiveness in augmenting and imputing missing values while preserving temporal patterns Further testing on GPS data confirmed that SC-β-VAE-GAN can generalize to other time-series datasets, particularly those with missing data. Additionally, emotion classification models trained with SC-β-VAE-GAN-generated data demonstrated improved precision, recall, and F1-scores. These results show that SC-β-VAE-GAN offers a good solution for addressing limited and missing data in handwriting analysis, particularly for emotional state recognition.

Keywords: Data Augmentation, Multivariate Time Series, VAE-GAN, Shift Correction, Missing Data Imputation, Hybrid Neural Networks, Beta Regularization, Handwriting Analysis, Emotion Recognition

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

THE PROBLEM AND ITS SETTING

Introduction

The application of deep learning models has seen remarkable success across various fields, such as healthcare, climate science, industrial automation, and emotional state recognition. However, these fields often struggle with limited data availability, as acquiring well-annotated data is both expensive and time-consuming (Pan & Zheng, 2021; Nita et al., 2022; Szczakowska et al., 2023). Moreover, there is also a concern regarding an individual's privacy. Deep learning models typically require large datasets to prevent overfitting, which is a common issue when training models on small or homogeneous datasets (Chlap et al., 2021).

In the domain of handwriting analysis, the recognition of emotions through handwriting analysis has received relatively less attention, despite its potential as a non-intrusive and cost-effective approach (Alai & Afreen, 2023). These methods aim to assess an individual's personality characteristics by studying their handwriting patterns, with the underlying principle that a person's handwriting can reveal insights into their subconscious mind and behavioral tendencies. The current landscape of emotional state recognition underscores the need for effective and accessible methods to identify and address mental health conditions like depression, anxiety, and stress. These conditions significantly impact individuals' quality of life and contribute to substantial healthcare costs globally (World Health Organization, 2017). In the Philippines, nearly 1 in 10 young adults (8.9%) suffer from moderate to severe depressive symptoms, with those experiencing these symptoms being at greater risk of contemplating suicide (Puyat et al., 2021). Early detection and intervention can mitigate the adverse effects of these conditions, and identifying potential indicators through non-invasive means could facilitate timely support and treatment (Esposito et al., 2020). However, there is insufficient public data to develop accurate models for emotion recognition using handwriting and drawing samples.

Collecting and labeling large amounts of data in this domain is time-consuming and expensive, leading to a lack of data and making it difficult to build reliable models for emotion identification from handwriting and drawing samples (Szczakowska et al., 2023; Khan et al., 2024). To address this challenge, data augmentation techniques have shown potential by generating synthetic datasets that cover unexplored input spaces while maintaining correct labels, thereby increasing the size and diversity of training datasets (Wen et al., 2021b). Data augmentation involves generating synthetic data to increase the size and diversity of a dataset, which helps improve the generalization and performance of machine learning models (Hou et al., 2022).

In collecting online handwriting data, several characteristics are recorded, including multiple variables such as the pen tip's X and Y-axis positions, pen status, pressure, azimuth angle, altitude angle, and timestamp. This data is typically gathered using devices like tablets and styluses (Khan et al., 2024). In psychological applications, such as Parkinson's disease detection and emotional state recognition systems, the data is derived from drawings of shapes like clocks, pentagons, circles, and houses, as well as from writing words and sentences (Likforman-Sulem, 2017). These activities are used to gather information on cognitive, emotional, and developmental status.

Most studies on handwriting have focused primarily on movements made while the pen is on the writing surface. However, some studies emphasize the significance of in-air movements, which can only be tracked when the pen tip is within about 1 cm of the surface. Beyond this range, the data is lost (Faundez-Zanuy et al., 2020). Missing data can reduce efficiency, complicate analysis, and introduce biases between complete and incomplete datasets (Barnard & Meng, 1999; Farhangfar et al., 2007). Two common methods to address missing data are deletion and imputation (Cheema, 2014). Deletion involves removing partially observed samples or features, which can create dataset gaps and lead to incorrect parameter estimations (McKnight et al., 2007; Graham, 2009). Imputation, on the other hand, estimates and fills in missing values, preserving the dataset's integrity and allowing for a more accurate and complete analysis (Rubin, 1976). While deletion is straightforward, it can result in significant information loss when missing data is abundant (Guo, 2019). Imputation leverages all available information to fill in missing data, making it a more practical and effective approach.

In previous studies focusing on handwriting data for emotional state recognition, only a few data augmentation techniques have been utilized, primarily relying on geometric transformations. However, these methods have drawbacks, as they require expert knowledge to maintain correct labels (Abayomi-Alli, 2021). For instance, Flores et al. (2021) introduced Gaussian white noise as an augmentation technique to generate additional reliable observations, addressing the imbalance issue in the EMOTHAW dataset. Similarly, Nolazco-Flores et al. (2021; 2022) employed Gaussian random noise to ensure that the data had an equal number of observations, effectively mitigating the imbalance problem. Additionally, existing data augmentation techniques for time series data have primarily focused on univariate datasets, overlooking the complexities and nuances of multivariate time series data, such as handwriting (Yang & Desell, 2022). Various studies have explored data augmentation in handwriting, but imputation remains underexplored despite the presence of missing data (Flores et al., 2021; Najda & Saedd, 2022; Otero et al., 2022).

Additionally, earlier statistical imputation methods like ARIMA, ARFIMA, and SARIMA, along with machine learning techniques such as regression, K-nearest neighbor (KNN), matrix factorization, and MICE, have been used for imputing missing values in multivariate time series.

However, these traditional imputation techniques often fail to perform adequately for multivariate data, as they do not effectively utilize the inherent correlations across features (Pourshahrokhi et al., 2021). Deep generative models like Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) have shown superior performance for data imputation by effectively modeling complex dynamics and learning the underlying data distribution, leading to more accurate and reliable imputation of multivariate time series data (Wang et al., 2024).

VAEs excel in capturing and generating the underlying distribution of real data, making them particularly suitable for generating synthetic tabular datasets (Bang et al., 2024). GANs have also been employed to tackle the imputation challenge, with models like MTS-GAN (Guo et al., 2019) and E2GAN (Luo et al., 2019) utilizing the adversarial training framework to generate realistic imputations for missing values (Wang et al., 2024). However, GANs are notoriously difficult to train, susceptible to mode collapse, and lack evaluation metrics (Iglesias et al., 2023). They also face limitations in modeling the complex distributions of multivariate time series data (Guo et al., 2019; Li et al., 2022). VAEs may produce low-quality data as they optimize for reconstruction loss (Ham et al., 2020).

To overcome these limitations, combining the strengths of VAEs and GANs into a hybrid model called VAE-GAN has been proposed over the years (Ruan et al., 2023). VAE-GAN utilizes the latent variable model of VAE to generate data and uses the discriminator of GAN to evaluate the authenticity of the generated samples (Iglesias et al., 2023). This combination allows the model to enhance its generative capabilities by producing more high-quality and diverse samples of time series datasets (Hu et al., 2023). However, scenarios where missing values are drawn from a different distribution than the training data pose a challenge for VAE-GAN. To address this, researchers have introduced shift-correction (SC) variants, such as SC-β-VAE (Li et al., 2021) and SC-VAE (Qiu et al., 2020), which modify the assumption of the training data distribution to follow a shifted Gaussian. Additionally, the β-VAE variant, a generalization of VAE that balances reconstruction loss and regularization loss through a hyperparameter β, has been explored to improve imputation performance (Qiu et al., 2020).

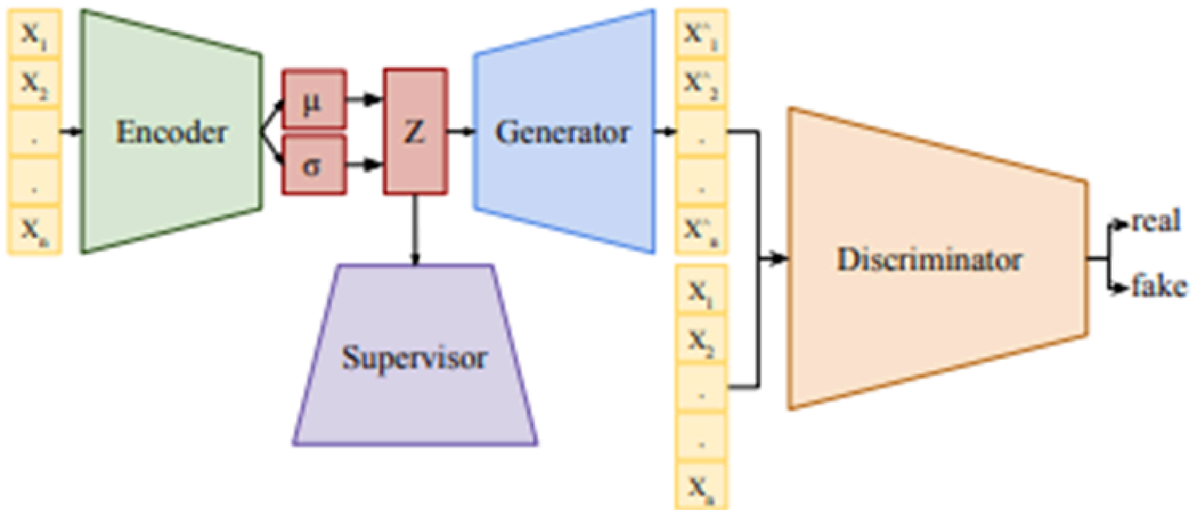
The developed solution of combining VAE and GAN with shift correction and beta regularization (SC-β-VAE-GAN) aims to leverage the strengths of these individual models while mitigating their drawbacks. By integrating the generative capabilities of VAEs (Bang et al., 2024), the adversarial training framework of GANs (Wang et al., 2024), and the shift correction and regularization techniques, the SC-β-VAE-GAN model has the potential to provide a robust and effective solution for data augmentation and imputation. To the best of the authors' knowledge, this is the first approach to utilize a VAE-GAN hybrid for the purpose of data imputation. This approach is particularly valuable in domains where data collection is challenging or time-consuming, such as healthcare, physics, and psychology, as it maximizes the utility of available data while ensuring high-quality synthetic samples and accurate imputation of missing values, even in scenarios where the missing data exhibits specific patterns or is drawn from a different distribution than the training data. Furthermore, the SC-β-VAE-GAN model can potentially be applied to a wide range of applications beyond handwriting recognition, extending its usefulness to other domains or scenarios involving multivariate time series data.

Theoretical Framework

VAE-GAN (Variational Autoencoder Generative Adversarial Network). This is a model that combines the strengths of Variational Autoencoders (VAEs), which learn an efficient latent representation, and Generative Adversarial Networks (GANs), known for generating realistic samples (Mishra, 2024).

Razghandi et al.'s (2023) VAE-GAN model architecture, depicted in Figure 1, includes an encoder that encodes the input sequence (x1, x2, ...) into a Gaussian distribution over the latent space, represented by mean and variance vectors. A supervisor trains the encoder to closely approximate the next time step in the latent space based on the input sequence. Additionally, a generator reconstructs the sequence from the latent space, attempting to produce a sequence that fools the discriminator into considering it real. The discriminator distinguishes between the real input sequences and the generated (fake) sequences from the generator. This model is particularly useful for tasks like sequence generation, where the goal is to generate realistic sequences that match the training data distribution. This makes the VAE-GAN model important for this study.

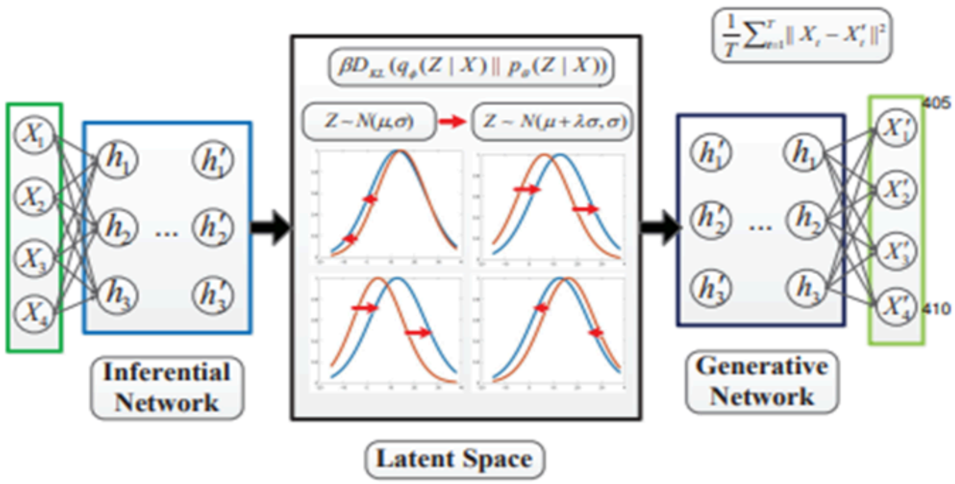
Figure 1. VAE-GAN Model Architecture (Razghandi et al., 2023)



VAE based on Shift Correction. According to Li et al. (2021), this model is a modified standard VAE model designed to fill in missing values in data samples. This correction aims to counteract the deviation caused by missing values. It is applied in the Gaussian latent distribution, where a shift hyperparameter λ is manually set to center the latent distribution, thus correcting the possible bias produced by missing values. In their model, as shown in Figure 2, Li illustrates how the VAE architecture is used to impute missing values.

The architecture includes an inferential network acting as the encoder part of the VAE, which takes input data (A1, A2, ..., An) and encodes it into a latent space representation, parameterized by mean (μ) and variance ($\sigma 2$) vectors. The latent space contains the learned compressed representation of the input data, where each data point is encoded as a distribution in this space, defined by the μ and $\sigma 2$ vectors from the inferential network. A generator acts as the decoder part of the VAE, taking samples from the latent space and generating/reconstructing the output data (B1, B2, ..., Bn). The model incorporates a shift correction mechanism, represented by the term $(1 + \Sigma i\ |xi - xi|)$ in the diagram. This term penalizes temporal shifts between the generated data and the input data, ensuring better alignment in the time dimension. This shift correction-based model is designed to learn a disentangled latent representation of the input data while ensuring good reconstruction quality. The shift correction mechanism specifically addresses the issue of temporal misalignment in the generated data, making it suitable for applications involving time series data or sequential data tasks, as in this study.

Figure 2. VAE based on Shift Correction Model Architecture (Li et al., 2021)



Conceptual Framework

Figure 3. Conceptual Framework of the Study

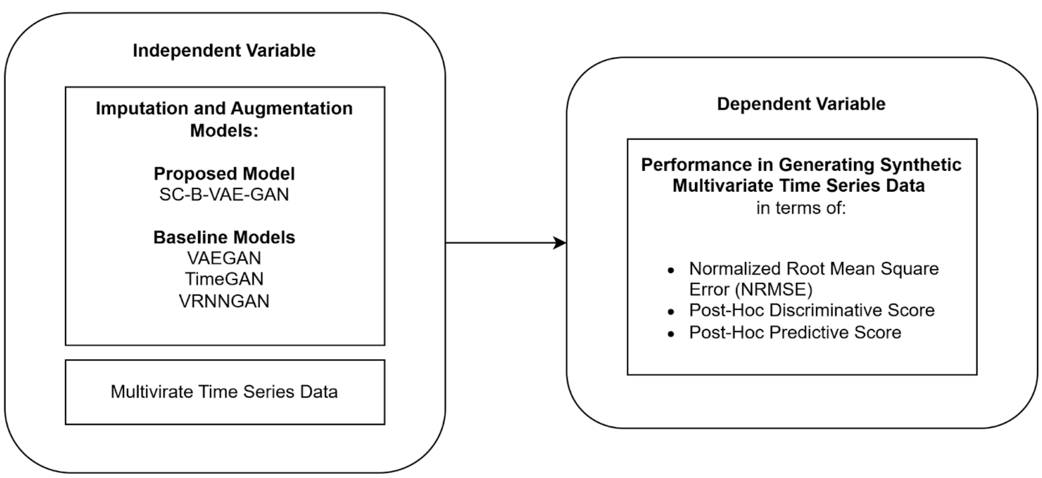


Figure 3 illustrates the conceptual framework of the study, outlining the core components and objective of the study. The primary focus is to evaluate the developed model and compare its performance on baseline augmentation models for generating synthetic multivariate time series data.

The independent variable comprises the developed SC-β-VAE-GAN (Shift Correction Beta Variational Autoencoder Generative Adversarial Network) model, which is a modified approach developed by the researchers. Also, baseline models, namely VAE-GAN, TimeGAN and VRNNGAN, were included for comparative analysis of how the developed and modified model performed based on the performance of similar models. These techniques were trained and applied to the multivariate time series data which serves as the input.

To check how the models performed, dependent variables are presented. Dependent variables include the performance indicators used to assess how accurately these models perform time series synthetic data generation. Specifically, three metrics are employed: Normalized Root Mean Square Error (NRMSE), Post-Hoc Discriminative Score, and Post-Hoc Predictive Score. These metrics quantify the similarity between the generated synthetic data and the real data, evaluating both the statistical properties and the ability to preserve temporal dependencies and patterns.

By systematically evaluating the developed SC-β-VAE-GAN model and the baseline models using these performance metrics, the researchers aim to determine if the developed model will be an effective approach for imputing and augmenting multivariate time series data.

Statement of the Problem

Analyzing online handwriting data has proven useful in various fields. However, challenges such as limited data and missing data have hindered the quality of previous studies. To address these issues, the study developed SC-β-VAE-GAN, a shift correction VAE-GAN model, to generate synthetic data for imputation and augmentation. The study aims to address the following questions:

1. What is the performance of the SC-β-VAE-GAN in generating synthetic data, based on the following metrics:
    1.1. Normalized Root Mean Square Error (NRMSE)
    1.2. Post-Hoc Discriminative Score
    1.3. Post-Hoc Predictive Score
2. What is the performance of VAE-GAN, TimeGAN, and VRNNGAN in generating synthetic data, based on the following metrics:
    2.1. Normalized Root Mean Square Error (NRMSE)
    2.2. Post-Hoc Discriminative Score
    2.3. Post-Hoc Predictive Score
3. Is there a significant difference between SC-β-VAE-GAN and other models, specifically: A. VAE-GAN B. TimeGAN C. VRNNGAN; in terms of generating synthetic data?

Hypotheses

H_0 - There is no significant difference between the performance (NRMSE, Post-Hoc Discriminative Score, Post-Hoc Predictive Score) of SC-β-VAE-GAN and VAE-GAN in terms of generating synthetic data.

H_0 - There is no significant difference between the performance (NRMSE, Post-Hoc Discriminative Score, Post-Hoc Predictive Score) of SC-β-VAE-GAN and TimeGAN in terms of generating synthetic data.

H_0 - There is no significant difference between the performance (NRMSE, Post-Hoc Discriminative Score, Post-Hoc Predictive Score) of SC-β-VAE-GAN and VRNNGAN in terms of generating synthetic data.

Scope and Limitations of the Study

In this study, a system that is capable of generating synthetic data for imputation and augmentation was developed. A Generative Adversarial Network (GAN) framework was incorporated where a Variational Autoencoder (VAE) with shift correction and hyperparameter β was the generator, and the Long Short-Term Memory (LSTM) network was the discriminator. The experimentation included the use of multivariate time series dataset, specifically EMOTHAW – an online handwriting data to generate synthetic data. Additionally, a GPS time series dataset was used to validate the model's applicability to other time series dataset. Furthermore, a comparison between the developed system, SC-β-VAE-GAN, and other systems such as VAE-GAN, TimeGAN and VRNNGAN was conducted.

The study did not include offline handwriting datasets or images containing handwriting data, as it involves different types of data representation. Unlike offline handwriting, which captures static images of written text, online handwriting mainly records dynamic information such as pen position, pressure, and timing as the writing process is happening. Aside from that, classification of these data was not part of the study as our main goal was to generate synthetic data.

Significance of the Study

This study that developed a Shift Correction β-VAE-GAN Model for Imputation and Augmentation of Handwriting Multivariate Time Series Data addressed the challenges associated with limited data and missing data. It can offer valuable insights and applications across different domains.

AI and Machine Learning Professionals - The introduction of shift correction and β adjustments into the VAE-GAN framework has the potential to lead to significant advancements in data augmentation and imputation. These improvements could address critical issues such as enhancing data variability and handling missing values, both essential for effective model training. This advancement could greatly benefit AI and machine learning professionals by providing them with potentially more effective tools for data augmentation and imputation of multivariate time series data.

Graphologists - By improving methods for augmenting and imputing handwriting time series data, these models could provide higher quality data, potentially leading to more accurate and reliable analysis in graphology. This could deepen insights into personality traits, cognitive states, and emotional conditions derived from handwriting, ultimately enhancing the validity and reliability of their findings.

Future Researchers - The advancements developed in this study, particularly the integration of shift correction and Beta into the VAE-GAN model, could be used as reference for future research. Researchers might be able to build upon these improvements to further enhance data handling techniques, explore new applications, and continue advancing the fields of machine learning and data science.

Definition of Terms

1. **Data Augmentation** - A process that involves modifying cases in the training set to increase data diversity without collecting new data, thereby enhancing model generalization.

2. **Data Imputation** - A process that fills in missing values in a dataset using various techniques trained on observed data, restoring complete datasets for analysis or modeling.
3. **Missing Data** - Data that are absent or unavailable in a dataset.
4. **Synthetic Data** - Artificially generated data that imitates the characteristics of real-world data, used for research, testing, and training models when real data is unavailable or sensitive.
5. **Graphology** - The scientific method of assessing personality characteristics by studying handwriting patterns.
6. **Online Handwriting** - Captures the real-time movement of a pen or stylus on a digital surface, recording data such as position, azimuth, altitude, pen status, and pressure for analysis and recognition.
7. **Time Series Data** - A sequence of data points collected at consistent time intervals, used to analyze trends and patterns over time.
8. **Multivariate Time Series Data** - A type of time series data that consists of multiple time-ordered and time-dependent variables.
9. **Variational Autoencoder (VAE)** - A neural network model for unsupervised learning that encodes and decodes data into a latent space, enabling the generation of new data. It is used as the generator component in the VAE-GAN model.
10. **Shift Correction** - An adjustment in VAEs that changes the mean of the latent vector to correct distribution errors caused by missing data.
11. **Generative Adversarial Network (GAN)** - A neural network model with a generator and discriminator trained together to produce realistic data samples. It enhances VAE by improving data realism through adversarial training in the VAE-GAN model.

12. **Long Short-Term Memory (LSTM)** - A recurrent neural network for modeling long sequences of data. It acts as the discriminator in the VAE-GAN model.
13. **Normalized Root Mean Square Error (NRMSE) -** A metric that evaluates the feature error or difference between the original data and the synthetic data, providing a normalized measure of the overall discrepancy between the two datasets.
14. **Post-Hoc Discriminative Score -** A metric that assesses how realistic and similar the generated data is compared to real-world data, indicating the ability of the synthetic data to mimic the statistical properties of the original data.
15. **Post-Hoc Predictive Score -** A metric that evaluates the predictability and the preservation of patterns in the synthetic data, ensuring that it retains the underlying structures necessary for predictive modeling and analysis.

Chapter 2

REVIEW OF RELATED LITERATURE AND STUDIES

The section reviews the literature and studies that support the use of the proposed SC-β-VAE-GAN model for imputation and augmentation of handwriting multivariate time series data. The literature review covers eight recurring themes: handwriting data, time series data, data augmentation, data imputation, the combination of data augmentation and imputation, generative adversarial networks (GANs), variational autoencoders (VAEs), and VAE-GANs. These themes offer a structured approach to understanding the various methods, gaps, challenges, and limitations within the research topic.

I. Handwriting Data

Handwriting is a multi-sensory activity and skill that plays a crucial role in life (Alaei & Alaei, 2023). Each individual possesses a unique handwriting style that could reflect one's personality and psychology. The characteristics drawn from handwritten data motivated researchers to explore this area for over a decade. And it has been used in various applications including user authentication, assessment of neurodegenerative disorders, and classification of handedness, gender, and age groups (Hasan et al., 2024). This can be beneficial for professionals and experts in assessing or addressing specific issues in these fields.

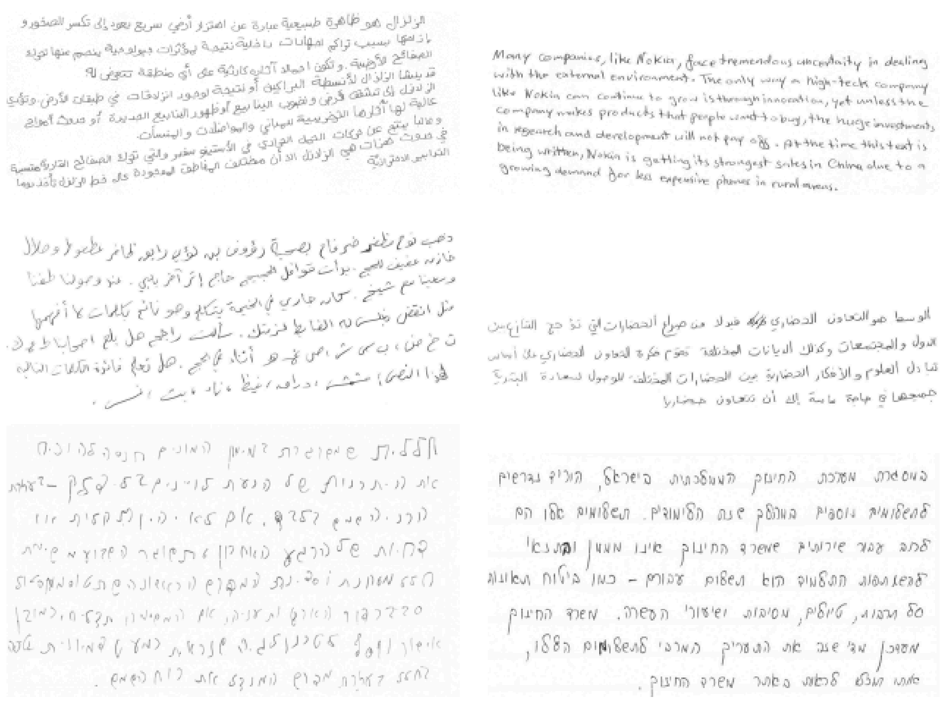
A handwriting system can be categorized into two, online and offline. The latter are typically represented by digitized images extracted from documents, typically from a pen and paper sample containing handwritten data. Meanwhile, online handwriting data requires specialized hardware, including a digitalized tablet or pen to capture signatures directly (Taleb et al., 2020). Extracting features from the two handwriting samples can be categorized into two, micro and macro. The latter describes the overall pictorial characteristics (size, slant, shape, and space between the words and characters), whereas micro features describe the attributes of individual characters/components, for example, the geometry or shape of the individual characters (Alaei & Alaei, 2023). Moreover, offline handwriting samples use image-processing techniques to extract features, while online handwriting samples use signal-processing methods (Velazquez-Flores, 2021). Extracting these characteristics is important as it could help intelligent systems to accurately predict or classify an individual.

1. **Offline Handwriting Data**

Recognition of handwriting and drawing images have been widely researched over the past decade. Offline handwriting can be described as the process of converting handwritten or drawn images to its digital form. It has a variety of applications ranging from digital character conversion to signboard translation and scene image analysis (Singh et al., 2023). Automatic classification such as of gender and age has been a subject for researchers, and it has been proven that several handwriting characteristics relate to a gender. Male usually are more angular, disorderly, and slanted than females. And some studies have also shown that age can affect the writing performance of an individual.

In a study by Rabaev et al. (2022), they developed a system that can automatically detect gender and age that can be beneficial in some fields such as forensics and historical document analysis. They proposed B-ResNet where they used the B-CNN Architecture with ResNet instead of VGG blocks. Furthermore, they used a combination of three offline handwriting datasets namely KHATT which composes of 1000 handwritten Arabic images made by high school and university students, QUWI dataset that includes handwriting samples written in English and Arabic, as well as HHD dataset which contains Hebrew handwritten images for gender detection.

Figure 4. Handwritten images from the KHATT, QUWI, and HHD databases (Rabaev et al., 2022)



These datasets were represented by two parallel ResNet models and concatenated to the bilinear vectors that are fed to the classification layer. The results of their experimentation showed that the proposed model has a higher classification accuracy for gender detection compared to other models except for HHD dataset, and the combination of English and Arabic handwriting images from QUWI dataset. Additionally, the proposed model also yielded higher accuracy for age classification for three and four age classes.

The study of Ozyurt et al. (2024) proposed a novel approach for handwriting signature verification. First, they collected a dataset containing 12,600 handwriting signature images from 420 individuals each containing 30 distinct signatures. The features from these images were extracted using MobileNetV2, and used three different feature selection techniques namely neighborhood component analysis (NCA), Chi2, and mutual_info (MI). Using these three techniques, they were able to select 200, 300, 400, and 500 relevant features from the images. These were fed to different machine learning models such as SVM, KNN, DT, Linear Discriminant Analysis, and Naive Bayes. The results of their experimentation showed that classification without using any feature selection technique yielded a 91.3% accuracy while NCA with 300 features had a high accuracy of 97.7%.

While there is already a lot of advancement in this area, it is still a field that researchers are focusing on because of its demand applicability in different domains including handwritten manuscript recognition, bank form recognition and historical document processing (Wang et al., 2021). Moreover, offline handwriting recognition itself is complicated as there are a lot of variations in writing styles, character types, and complicated structure of texts (Wang et al., 2021). Unlike online handwriting, tasks for offline handwriting recognition can be divided to smaller ones like paragraph detection, text-line segmentation, word/character segmentation, image normalization (Wang et al., 2021).

2. Online Handwriting Data

Online handwriting data captures multiple variables such as the pen tip's X and Y-axis positions, pen status, pressure, azimuth angle, altitude angle, and timestamp (Khan et al., 2024). Researchers use online handwriting analysis to determine and understand various traits of individuals, including personality, neurodegenerative diseases, emotional states, gender, age, and nationality. This data is gathered using devices like tablets and styluses.

The rich metrics provided by online handwriting analysis make it an effective tool for the early detection of health conditions and diseases. It offers a non-invasive diagnostic method that can be conveniently administered by non-technical personnel in the patient's environment without disrupting daily routines. Additionally, it provides a cost-effective solution, requiring minimal infrastructure or medical equipment, and delivers information swiftly and affordably (Otero et al., 2022).

Rahman and Halim (2022) explored the use of tablet devices to non-invasively determine emotional states through handwriting and drawing acquired from the publicly available EMOTHAW (EMOTional State Recognition from HAndwriting and DraWing) database. They specifically used temporal, spectral, and Mel Frequency Cepstral Coefficients (MFCC) for feature extraction and employed BiLSTM networks for classifying these features. Additionally, spatial features such as velocities in the x- and y-directions were examined. Utilizing multiple public benchmark datasets, the research identifies specific activities and features correlating with emotional states like depression, anxiety, and stress, achieving a significant improvement in classification accuracy by 5.32% to 8.9% overing the number of data for training deep learning models like BiLSTM. The researchers recommended deploying data augmentation techniques such as the Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) and spline-based interpolation techniques. These methods generate synthetic data samples to address issues with small datasets that suffer from class imbalance and susceptibility to overfitting.

Khan et al. (2024) also explored identifying negative emotions such as depression, anxiety, and stress through online handwriting and drawing data. A deep learning network, specifically an attention-based transformer, was used to analyze handwriting samples to determine emotional sr baseline methods. The proposed method exhibited promising results of 92.64% on the EMOTHAW dataset. However, they encountered challenges in developing the model due to the limited diversity in handwriting and drawing styles within the training data, impacting the model's reliability and generalizability for emotion identification.

Most online handwriting research focuses on pen-on-paper movements, but in-air movements are also significant. Greco et al. (2023) found that depressed patients spent more time with the pen in-air (Up) and on paper (Down) and took longer to complete tasks compared to healthy participants. Specific tasks like writing words, drawing clocks, and writing sentences revealed longer in-air and overall times for depressed individuals, who also had more unrecognized and fewer total pen strokes. Kunhoth et al. (2023) demonstrated that including in-air features significantly improved machine learning classifiers for diagnosing dysgraphia. Classifiers combining on-surface and in-air features outperformed those using only on-surface features, with the AdaBoost classifier achieving 80.8% accuracy compared to 72.5%. Gargot et al. (2020) found that in-air movements correlate with handwriting quality and speed in typically developing children, with less in-air movement linked to better handwriting. Children with severe dysgraphia showed abnormalities in kinematics, pressure, and tilt. These studies highlight the importance of in-air features in enhancing the accuracy and reliability of handwriting analysis.

However, a limitation of this approach is that tracking pen movements when the tip is in the air is restricted to a range of about 1 cm from the surface, resulting in data loss beyond this range (Faundez-Zanuy et al., 2020). Such data loss can significantly reduce the dataset's size and eliminate valuable information crucial for analysis, potentially biasing results by inadequately representing the original data's scope and nature (Zhu et al., 2024).

II. Time Series Data

Time series data are usually collected from various real-world systems (Chawla et al., 2019) and have been a key area of academic research in different applications (Lim & Zohren, 2021). This area of research is growing over time as changes in data for various processes are recorded and needed for numerous global processes, especially to solve problems related to big data (Torres et al., 2021) and to capture the dynamic changes and patterns that occur over time (Lim & Zohren, 2021). These data are collected from various sources, including sensors, financial markets, and social media platforms, and are used to analyze and forecast future trends and behaviors (Buaton et al., 2019). The importance and quality of time series data lies in their ability to provide valuable insights into complex systems, enabling researchers and practitioners to identify patterns, make predictions, and inform decision-making processes (Lim & Zohren, 2021).

Time series data usually have values for a given entity I at time T. This entity represents the temporal information needed for data analysis, capturing how a specific attribute of interest changes over time. The variable T denotes the specific times at which observations of the entity are recorded, providing a sequential snapshot that allows for trend analysis, forecasting, and other time-dependent evaluations. Understanding the temporal dynamics of an entity through time series data is crucial for making informed decisions and predictions in various fields such as finance, healthcare, and environmental studies (Lim & Zohren, 2021). The analysis and processing of time series data have become increasingly crucial in today's data-driven world, particularly in the context of big data and the Internet of Things (IoT). As the volume and complexity of time series data continue to grow, new methods and techniques are being developed to effectively extract insights and make accurate predictions (Choi et al., 2021).

Within time series data, univariate and multivariate datasets offer distinct perspectives. Univariate data involve a single feature variable, with measurements focusing on one aspect at a time. This type of data is useful for understanding the behavior and trends of a single attribute over time. In contrast, multivariate data involve multiple feature variables measured simultaneously or at different frequencies. This type of data provides a comprehensive view by capturing the interplay between various attributes over time, allowing for a more holistic analysis and richer insights (Weerakody et al., 2021).

1. Multivariate Time Series Data

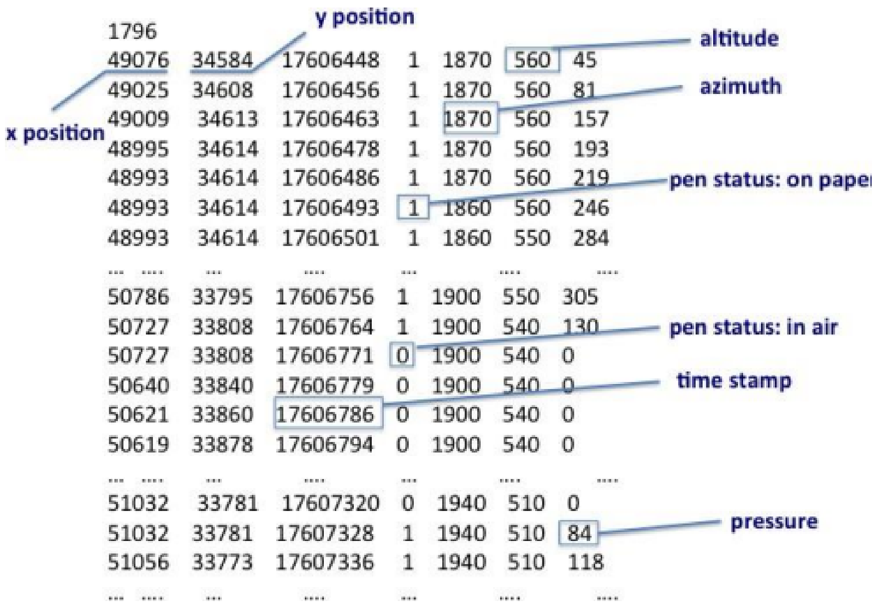
Multivariate time series data is a type of data that contains different types of behaviors in different working periods of the system. These multivariate periodic recorded data contain multiple variables compared to univariate which only has one variable (Sürmeli & Tümer, 2019). Its data has gained importance in various domains such as medicine, finance, multimedia, and nearly every field that operates with temporal datasets. This type of data is characterized not only by individual attributes but also by their interactions. The consideration of multiple attributes can make processes like prediction, pattern finding, and augmentation more complex and tedious (Baydogan & Runger, 2014).

The complexity of multivariate time series (MTS) data requires a more rigorous and complex approach for implementing unsupervised learning and clustering analysis, given that multiple characteristics are present in each log. In addition, as MTS is relatively new compared to univariate time series (UTS), some time-series-related studies have focused only on single-variable time series due to the absence of high-dimensional features (Li & Du, 2021). MTS, on the other hand, is not only multi-dimensional and high-dimensional but also typically has an extended time dimension, numerous attribute variables, and large data volumes. This poses a challenge for data analysis, machine learning, data recognition, and other related domains (Baldán & Benítez, 2021).

2. Handwriting as Multivariate Time Series Data

Handwriting analysis and recognition are fields that extensively utilize multivariate time series data. This type of data involves multiple variables recorded over time, capturing the dynamic and complex nature of handwriting (Morrill et al., 2020), as shown in Figure 5. Each variable, such as pen pressure, stroke direction, and velocity, provides crucial insights into the mechanics and characteristics of handwriting, facilitating a deeper understanding of how handwriting is produced and its distinct attributes (Lee et al., 2022).

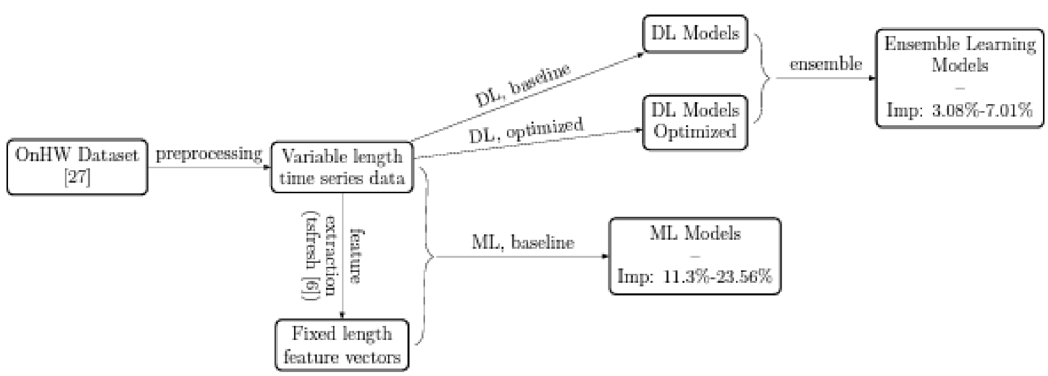
Figure 5. Extracted online handwriting variables (Likforman-Sulem, 2017)



Handwriting recognition is a prominent application of multivariate time series data, where the temporal sequences of different handwriting features are analyzed to accurately interpret written text. The intricacies of these multivariate characteristics pose challenges for traditional learning algorithms, necessitating advanced techniques for effective analysis and recognition (Morrill et al., 2020) Study by Azimi et al, uses online handwriting data in experimenting the different machine learning and deep learning models for handwriting recognition. Researchers found that since the sensor captures a lot of information pieces such as the time it takes to write, cadence and stylization making it more beneficial to understand more about the certain handwriting. (Azimi et al., 2023).

Their contributions include significant accuracy improvements in both Machine Learning (ML) and Deep Learning (DL) classifiers on the OnHW-chars dataset, with ML models showing a 11.3%-23.56% improvement and DL models achieving a 2.17%-4.34% increase over previous benchmarks. Additionally, the use of ensemble learning methods yielded further improvements of 3.08%-7.01%. The researchers also focused on providing explainability for their models, extending the LIME architecture to add interpretability to their multivariate time series data. They ensured the reproducibility and verifiability of their results by making their preprocessing code and models publicly available. Their study not only improved accuracy but also provided insights into why certain models are suitable for specific data types, contributing to the transparency and advancement of handwriting recognition research. Future work suggested includes optimizing ensemble methods and conducting deeper analyses on model explainability and feature importance (Azimi et al., 2023).

Figure 6. Workflow for training ML and DL models with ensemble method (Azimi et al., 2023)



A research by Lintonen & Raty tries to solve the problem of complexity in self-learning of the multivariate time series data by proposing a novel stopping criterion called Peak evaluation using perceptually important points. This criterion algorithmically adds the stopping criteria aiming to make the learning not undergo manually added hyperparameters on the system. It has been applied to different datasets such as heartbeat, wave gesture, basic motions, finger movements, handwriting, and many more. The result shows that it performs well for univariate and multivariate datasets, enhancing the accuracy of datasets, class balancing with a positive class dominance. Researchers, on the other hand, recommend that further study might be conducted following their research, specifically enhancing the data labeling process and examining the effects of multimodality of time series data (Lintonen & Raty, 2019).

III. Data Augmentation

The basic idea of data augmentation is to generate synthetic datasets that cover unexplored input spaces while maintaining correct labels. It is used to artificially increase the size and diversity of a training dataset by creating modified versions of the existing data (Wen et al., 2021). This process involves applying a series of transformations, such as flipping, rotating, scaling, or adding noise, to the original data samples (Keskin, 2023). The resulting augmented dataset contains both the original samples and their transformed counterparts.

The primary purpose of data augmentation is to improve the performance and generalization capabilities of machine learning models, particularly in scenarios where the available training data is limited. In deep learning models, large amounts of data are generally required to prevent overfitting, which is a common concern when a model is trained on a small or homogeneous dataset (Chlap et al., 2021). The use of deep learning models has achieved remarkable success in many fields, including healthcare, climate science, industrial automation, and affective computing. However, these fields often suffer from limited data availability, as it is expensive and time-consuming to acquire well-annotated data (Pan & Zheng, 2021; Nita et al., 2022; Szczakowska et al., 2023), making data augmentation a practical solution. By exposing the model to a larger and more varied dataset, data augmentation aims to enhance the model's ability to learn robust representations and avoid overfitting.

1. Handwriting Data Augmentation

Several studies have been conducted in handwriting analysis by deploying different data augmentation techniques. In a study by Kamran et al. (2020), they proposed a novel approach by using deep transfer learning-based algorithms and different data augmentation techniques to create a robust model that can detect early symptoms of Parkinson’s Disease (PD). Different PD datasets such as HandPD, NewHandPD, and Parkinson's Drawing datasets were collected and used for the experimentation. During the latter, different combinations of datasets and data augmentation techniques such as flipping, thresholding, rotation, illumination, and contrast were fed to the models. The results showed that pre-trained CNN with fine-tuned architecture models yielded high performance, most specifically when the data augmentation technique illumination was employed as it obtained a 99.22% classification accuracy. Furthermore, their experimentation results showed that their proposed model surpasses the state-of-the-art approach when diagnosing early Parkinson’s Disease symptoms.

The study by Hamdi et al. (2021) proposed a novel approach by employing four different data augmentation strategies for online handwriting recognition of multi-language including Arabic and Latin script. They used four databases where 3 datasets namely ADAB, ALTEC-OnDB, and online_KHATT are for Arabic script, and UNIPEN is for Latin script. Their experimentation included the deployment of geometric, frequency, beta-elliptic, and hybrid data augmentation strategies and end-to-end CNN architectures were utilized to test the performance of their proposed method. The results of their experimentation yielded an improvement by reducing the character error rate (CER) and word error rate (WER) significantly.

Another research by Najda and Saedd (2022) implemented augmentation methods to expand the input data of their system for online signature verification. Drawing from current advancements in the field, they selected five augmentation techniques to enhance the dataset. Each method was applied with varying repetition rates for every signature, ranging from ×0 to ×40 times. These augmentation methods include interpolation with modifications by the authors, noise addition to time series also with modifications, signal scaling, signal rotation, and warping time series. Najda and Saedd (2022) consider the common properties in handwritten data. The study used the database SVC 2004, which contains properties such as X coordinate, Y coordinate, Pressure, Interval, Pen state, Azimuth angle, and Elevation Angle. These properties are where patterns are extracted from and enable the creation of new feature metrics such as signature duration, pen lead velocity and acceleration, coordinates of discrete points drawn from the signature line, or means and standard deviations of individual signal components.

In a study by Otero et al. (2022), a multivariate Empirical Mode Decomposition (mEMD) method was proposed to generate artificial handwritten data for diagnosing Essential Tremor (ET) using a Long Short-Term Memory (LSTM) model. Frequency decomposition was performed, and Intrinsic Mode Functions (IMFs) were generated from it. These IMFs of the subjects were then utilized to generate artificial samples for training the LSTM model (see Figure 7.a and figure 7.b below). The BIODARW dataset was employed, comprising a total of 51 samples, with 24 from the ET group and 27 from the control group. The study focused solely on X and Y coordinates, as these handwriting characteristics can be obtained from any digital tablet and provide sufficient information for diagnosing essential tremor. Experimental results indicated a 10% improvement compared to test cases without the generated data.

Figure 7.a. The process of generating IMFs from the original data using mEMD (Otero et al., 2022)

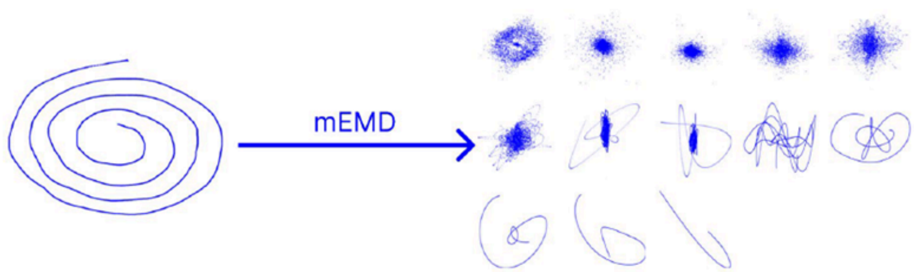
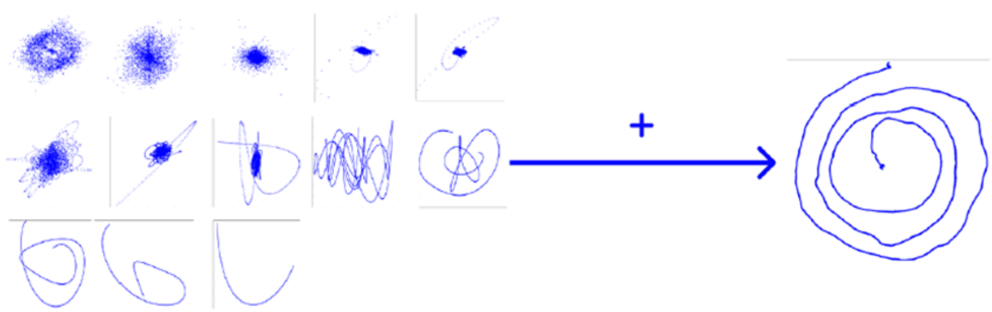


Figure 7.b. The process of generating IMFs from the original data using mEMD (Otero et al., 2022)

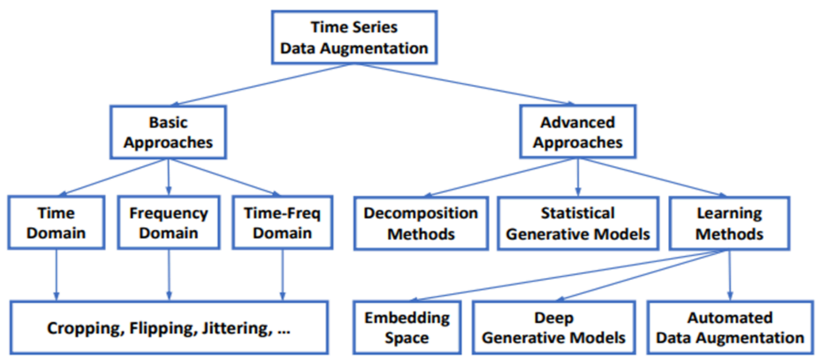


In previous studies focusing on handwriting data for emotional state recognition, only a few data augmentation techniques have been utilized, primarily relying on geometric transformations. However, these methods have drawbacks, as they require expert knowledge to maintain correct labels (Abayomi-Alli, 2021). For instance, Flores et al. (2021) introduced Gaussian white noise data as an augmentation technique to generate additional reliable observations, addressing the imbalance issue in the EMOTHAW dataset. Similarly, Nolazco-Flores et al. (2021; 2022) employed Gaussian random noise to ensure that the data had an equal number of observations, effectively mitigating the imbalance problem.

2. Time Series Data Augmentation

One area where deep learning has shown effectiveness but also faces limitations due to scarce data is time series analysis. Time series data poses unique challenges for data augmentation, as highlighted in Wen's (2021b) survey. This data, characterized by its inherent sequential nature and temporal dependencies, presents significant challenges for data augmentation. Efforts to preserve underlying patterns while augmenting data are complicated, particularly as time series can be transformed into frequency and time-frequency domains for potentially more effective augmentation techniques. However, these methods become more complex with multivariate time series, where interactions across multiple variables over time must be carefully managed. Traditional augmentation methods from image and speech processing often fail to generate valid synthetic time series data. Furthermore, the effectiveness of augmentation strategies is highly task-specific; techniques suitable for classification might not work well for tasks like anomaly detection or forecasting. The issue of class imbalance in time series classification further complicates the creation of a large, balanced set of synthetic data. When certain classes are underrepresented, the synthetic data may not adequately capture the variability within those classes, leading to a model that struggles to distinguish between them. This imbalance can cause the model to misclassify instances, as it may become biased toward the more prevalent classes, resulting in decreased accuracy and increased misclassification, particularly for the minority classes. Additionally, there is a notable deficiency in research on new methods for augmenting multivariate time series data, with many studies focusing solely on univariate data, which involves just a single input channel (Yang & Desell, 2022). The existing time series data augmentation methods across common tasks, including time series forecasting, anomaly detection, and classification, are outlined in Figure 8 (Wen et al., 2021b)

Figure 8. Taxonomy of time series data augmentation techniques (Wen et al., 2021b)



Basic data augmentation methods for time series can be categorized into two domains: time domain and frequency domain. In the time domain, methods such as window cropping, window warping (Gao et al., 2023), flipping (Wen & Angryk, 2024), and noise injection (Kim et al., 2023) are commonly used. These techniques involve manipulating the original time series by extracting slices, compressing or extending time ranges, flipping the sign of the series, and injecting noise patterns like Gaussian noise or spike-like trends. Additionally, label expansion in the time domain is utilized for anomaly detection tasks. In the frequency domain, perturbations in amplitude and phase spectra are applied to enhance time series data augmentation, offering improvements in anomaly detection tasks (Gao et al., 2020; Lee et al., 2019).

Advanced data augmentation methods for time series encompass decomposition-based, statistical generative, learning-based, and automated approaches. Decomposition-based methods, such as STL decomposition followed by recombination (Gao et al., 2020), utilize the components of time series like trend, seasonality, and residual signals to generate new synthetic data. Statistical generative models, like mixture autoregressive models (Kang et al., 2020), simulate time series data by describing conditional distributions based on historical data points. Learning-based methods leverage techniques like embedding space interpolation to generate diverse samples. Automated data augmentation employs reinforcement learning or evolutionary search strategies to automatically search for optimal augmentation policies, yielding significant improvements in classification accuracy for time series datasets (Cheung & Yeung, 2021).

Among advanced data augmentation methods, deep generative models have recently shown the ability to generate realistic high-dimensional data objects, such as images and sequences. DGMs developed for sequential data, like audio and text, can often be extended to model time series data. Generative adversarial networks (GANs) are particularly popular for generating synthetic samples and effectively increasing the training set, though generating effective time series data with GANs remains challenging (Wen et al., 2021b).

Combining basic time-domain methods, such as merging patterns and infusing noise, has been shown by Gao et al. (2024) to outperform using a single method and achieve the best performance in time series classification. While GANs are the primary deep generative model used for time series data augmentation, other models like Deep Autoregressive Networks (DARNs), Normalizing Flows (NFs), and Variational Autoencoders (VAEs) also hold great potential. Wen et al. (2021b) suggested exploring these less investigated models for time series data augmentation, presenting an exciting future opportunity that could lead to new and improved methodologies.

IV. Data Imputation

Missing data can lead to a loss of efficiency, complicate the analysis process, and introduce significant biases due to differences between complete and incomplete datasets (Barnard & Meng, 1999; Farhangfar et al., 2007). When dealing with the inevitable issue of missing data, which can arise from disruptions or malfunctions in data sampling and transmission, two common approaches are deletion and imputation (Cheema, 2014). Deletion involves removing samples or features that are only partially observed, which can leave gaps in the dataset and potentially lead to incorrect parameter estimations (McKnight et al., 2007; Graham, 2009). In contrast, data imputation estimates and fills in missing values, allowing for more accurate and complete analysis. Imputation replaces missing values with estimated ones, preserving the dataset's integrity (Rubin, 1976). According to Guo (2019), although deletion methods are straightforward, they can result in the loss of valuable information when the proportion of missing data is high. In contrast, imputation methods aim to utilize all available information from the observations and fill in the missing data to create a complete dataset, making them a more practical and effective approach than simply discarding missing data.

1. Handwriting Data Imputation

Missing data, or missing values, occur when there is no data stored for certain variables or participants. Data can go missing due to incomplete data entry, equipment malfunctions, lost files, and many other reasons (Bhandari, 2021). In any dataset like handwriting, there is usually some missing data. This can be due to device errors or limitations. Any errors or limitations in the pen tablet device during data recording could result in missing or corrupted values for certain handwriting samples (Faundez-Zanuy et al., 2020). Another reason is incomplete writing samples. It's possible that for some keywords or iterations, a person may have failed to fully complete the writing, leading to missing values in the recorded data (Park et al., 2021).

Akash et al. (2020) proposes a user authentication system based on an individual's handwriting data collected from a pen tablet device. The dataset comprises handwriting samples from 24 individuals, each writing 10 defined keywords 5 times, resulting in 50 samples per person. Six key features are extracted from the handwriting data: writing time, pen pressure, x-axis angle, y-axis angle, horizontal angle, and vertical angle, which capture vital attributes of a person's handwriting style. The authors employed two methods for their system. Method 1 involves basic feature extraction and classification using the Support Vector Machine (SVM) algorithm. Method 2 includes data pre-processing steps to balance the dataset format. Specifically, the Flatten Function is used to convert the 2D data to 1D form, and the Imputation Function fills in missing data values by averaging the same column. After pre-processing, four classification algorithms are used: SVM, Logistic Regression (LR), Linear Discriminant Analysis (LDA), and Random Forest (RF). The experiments show that Method 2 with pre-processing achieves better accuracy, with testing accuracy rates of 87%

for SVM, 85% for LR, 76% for LDA, and 77% for RF. The imputation step plays a crucial role in balancing the dataset by filling in missing values with the column average, allowing the classifiers to work with a consistent and complete dataset format. However, even though this imputation is simple, it can introduce bias and ignore relationships with other variables (Huang, 2023).

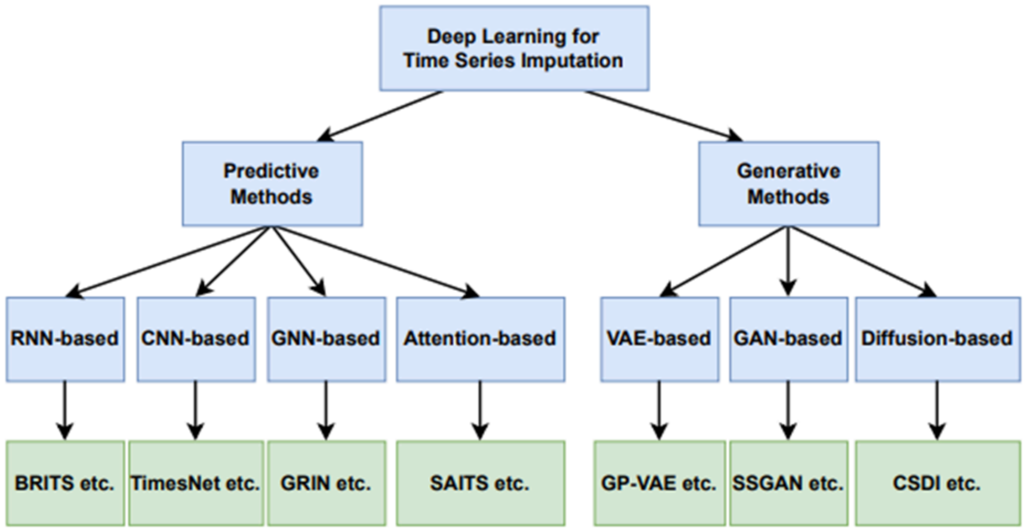
Chang et al. (2020) present a method called "data incubation" to enhance handwriting recognition by strategically synthesizing missing data using a controllable generative model. Handwriting data is characterized by its content (the text written) and style (printed, cursive, neat, etc.), and it is challenging to collect sufficient real data that encompasses all content and style variations. To address this, they propose a controllable generative model trained on the available real data to synthesize new handwriting samples with underrepresented content (e.g., URLs, emails) and rare styles (e.g., cursive, slanted). The synthetic data is then combined with real data to train a more effective handwriting recognizer that can better generalize to unseen content and styles. Their framework analyzes and optimizes the synthetic data generation process by training recognizers on real data, synthetic data, and their combination, helping to identify issues like artifacts in synthetic data or missing style/content variations. By carefully controlling the synthesis process, they achieve a 66% reduction in character error rate compared to training solely on real data, effectively imputing or filling in the missing data variations.

2. Time Series Data Imputation

Multivariate time series data often contains many missing values due to challenges in data collection. In areas like transportation (Zhang et al., 2021), healthcare (Kazijevs & Samad, 2023), and energy (Bülte et al., 2023), problems like sensor failures, unstable environments, and privacy issues can disrupt data collection. To address this, various imputation methods have been developed to fill in the gaps in multivariate time series data.

Earlier statistical imputation methods like ARIMA, ARFIMA, and SARIMA, along with machine learning techniques such as regression, K-nearest neighbor (KNN), matrix factorization, TIDER, and MICE, have been used for imputing missing values in multivariate time series. However, these methods often fail to capture complex temporal relationships and patterns and can suffer from bias due to case dropping or inappropriate data replacement (Wang et al., 2024). These traditional approaches are not ideal for high-dimensional data with fewer samples and perform inadequately for multivariate, highly dynamic data (Pourshahrokhi, 2021). Some studies developed systems that significantly improved time series imputation that can handle the complexity of the data such as missForest (Zhang et al., 2021) but, it still shows limitation as some results showed that the traditional method displayed superior performance than the developed model in different instances. Recently, deep learning methods, including Transformers, Variational AutoEncoders (VAEs), Generative Adversarial Networks (GANs), and diffusion models, have shown superior performance by effectively modeling complex dynamics and learning the true underlying data distribution, leading to more accurate and reliable imputation of multivariate time series data. In Figure 9, the overview of current deep multivariate time series imputation methods is presented (Wang et al., 2024).

Figure 9.The taxonomy of deep learning methods for multivariate time series imputation (Wang et al., 2024)



Deep learning methods have revolutionized the imputation of missing values in multivariate time series data, offering superior performance by effectively modeling complex dynamics and learning the underlying data distribution. Among these, generative models using diffusion techniques have shown promising results. For instance, CSDI (Conditional Score-based Diffusion Models for Imputation) (Tashiro et al., 2021), SSSD (Score-based Stochastic Differential Equations) (Alcaraz and Strodthoff, 2023, TMLR), CSBI (Conditional Score-based Imputation) (Chen et al., 2023), MIDM (Missing Imputation using Diffusion Models) (Wang et al., 2023), PriSTI (Prior-based Score-based Time-series Imputation) (Liu et al., 2023), DA-TASWDM (Diffusion Augmented Time-series with Attention-based Score Weighing and Diffusion Modeling) (Xu et al., 2023), and SPD (Score-based Probabilistic Diffusion) (Bilos et al., 2023) utilize diffusion models combined with attention mechanisms to address missing data under the MCAR (Missing Completely At Random) assumption. These models excel in capturing intricate temporal dependencies and spatial correlations, leading to highly accurate imputations.

In addition to diffusion-based approaches, generative models using Generative Adversarial Networks (GANs) have also been employed to tackle the imputation challenge. Unfortunately, GANs are not designed for sequential data, which is why research has been directed towards the development of hybrid models that use GANs as the underlying global concept (Richter et al., 2023). For example, models like MTS-GAN (Multivariate Time Series Generative Adversarial Network) (Guo et al., 2019), E2GAN (End-to-End Generative Adversarial Network) (Luo et al., 2019), NAOMI (Non-Autoregressive Multiresolution Imputation) (Liu et al., 2019), and SSGAN (Semi-Supervised Generative Adversarial Network) (Miao et al., 2021) utilize the adversarial training framework, typically combining GANs with RNNs (Recurrent Neural Networks), to generate realistic imputations for missing values under the MCAR (Missing Completely At Random) assumption. These GAN-based methods excel in generating high-quality synthetic data that closely resembles the original data distribution, making them highly effective for imputing missing values in multivariate time series.

Variational AutoEncoders (VAEs) have also been extensively explored for imputation tasks. Models such as GP-VAE (Gaussian Process Variational AutoEncoder) (Fortuin et al., 2019), V-RIN (Variational Recurrent Imputation Network) (Mulyadi et al., 2021), and supnotMIWAE (Supervised Non-Missingness Induced Variational AutoEncoder) (Kim et al., 2023) leverage the power of VAEs, often in combination with Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), to handle missing data under various assumptions, including MCAR, MAR (Missing At Random), and MNAR (Missing Not At Random). These VAE-based models are adept at learning latent representations that can effectively capture the underlying structure of the data, resulting in robust and reliable imputation.

According to Rubin's theory (1976), missing data are typically grouped into three categories: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). The assumptions of MCAR, MAR, and MNAR are crucial in imputation models because they define the nature of the missing data and influence the choice of imputation method. MCAR assumes that the missingness is completely random and independent of both observed and unobserved data, simplifying the imputation process. MAR assumes that the missingness is related to the observed data but not the unobserved data, allowing for more informed imputation based on available data. MNAR is the most challenging assumption, where the missingness is related to the unobserved data itself, requiring sophisticated models like VAEs that can learn from incomplete data and infer the missing values accurately. Understanding these assumptions helps in selecting appropriate models and techniques that can handle the specific nature of the missing data, ensuring more reliable and accurate imputations.

V. Data Augmentation and Data Imputation

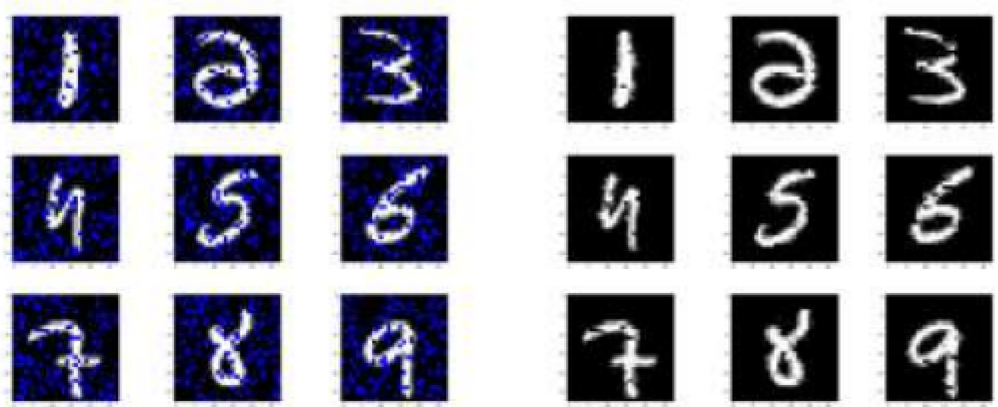
In the realm of machine learning and data analysis, ensuring the quality and completeness of datasets is crucial for building robust models (Zhang et al., 2021). Two fundamental techniques that address these challenges are data augmentation and data imputation. Data augmentation involves generating synthetic data to increase the size and diversity of a dataset, which helps improve the generalization and performance of machine learning models (Hou et al., 2022). Data imputation, on the other hand, focuses on estimating and filling in missing values within datasets to enable more accurate and comprehensive analysis (Huisman, 2000).

Combining data augmentation and data imputation is an innovative approach that takes advantage of the strengths of both techniques to enhance dataset quality. The rationale behind this combination lies in the complementary nature of these methods. Data augmentation enriches the dataset by adding variety, while data imputation fills the gaps caused by missing values. By integrating these techniques, one can simultaneously address data scarcity and missing values, leading to more reliable and effective machine learning models. This combined approach ensures that the dataset is not only more complete but also more diverse, which is essential for training models that generalize well to new, unseen data (Pourshahrokhi et al., 2022). Moreover, in fields where data collection is expensive or time-consuming, such as medicine, physics, and psychology, this strategy is particularly beneficial as it maximizes the utility of available data.

One study that effectively combined data augmentation and data imputation is by Vilardell et al. (2021), who investigated the efficacy of different methods for imputing missing data and generating synthetic data in breast cancer survival analysis. The study evaluated a range of models, including the ModGraProDep model, which integrates log-linear models with Bayesian networks, along with other techniques such as generalized linear models and neural networks. Using two different datasets and examining scenarios of missing data (MCAR and MNAR), the study found that the ModGraProDep models, particularly GM.SAT, consistently outperformed other methods. The inclusion of variables related to molecular subtypes further enhanced predictive performance, confirming the effectiveness of ModGraProDep in both imputing missing data and enhancing data reliability in the context of breast cancer research.

Rath et al. (2023) proposed a novel model for handling missing data and enhancing training data in multivariate time series, particularly applicable to fields like plasma diagnostics. Their approach utilizes dependent state space Student-t processes, leveraging the correlation between input signals with a multivariate Matérn covariance kernel. This model accommodates heavy-tailed noise distributions, enhancing robustness to outliers. Through Bayesian filtering and smoothing, they reduce computational complexity, making it suitable for high-resolution time series. Evaluation on synthetic test cases inspired by plasma diagnostics data demonstrates improved accuracy in imputing missing data and generating augmented data compared to independent models that don't consider signal correlations.

Figure 10. Image reconstruction of the MNIST dataset using F-HMC for imputing missing values (blue pixels) (Pourshahrokhi et al., 2021)



Another significant contribution comes from Pourshahrokhi et al. (2021), who introduced the Folded Hamiltonian Monte Carlo (F-HMC) model, combined with Bayesian inference, to effectively address missing data and preserve privacy by augmenting the original data for research analysis. The F-HMC model works by adapting to the patterns in the existing data to create new samples, which can fill in missing information and enhance the overall dataset. This technique was tested on a dataset from the University of California, San Francisco (UCSF), which includes cancer symptom assessments with missing and sensitive data, as well as the Modified National Institute of Standards and Technology (MNIST) dataset. To illustrate that the proposed method is generalizable to other datasets as shown in figure 10. The study compared the F-HMC model's performance to standard methods like KNN and MICE using measures such as error rates and accuracy. The results indicated that the F-HMC model not only filled in missing values more effectively but also created higher quality synthetic data compared to other methods such as GANs and synthpop, especially in complex, high-dimensional datasets like those often found in healthcare, where other models often fail to perform well.

VI. Generative Adversarial Network (GAN)

Generative adversarial networks is one of the most extensively studied in the field of Artificial Intelligence, specifically deep learning, in the past few years. (Wang et al., 2021). Developed and proposed by Godfellow and his co-researchers, are a type of deep generative model that have been very popular for data generation, as it is capable of producing outputs that are very similar to its training data (Lee, J.,2022). Its ability to efficiently generate decision samples, elimination of possible bias and its compatibility with other neural networks makes it advantageous (Wang et al., 2021). Also, because of its ability to generate data without modeling the probability density function (Yi et al., 2019).

GANs consist of a unique class of neural network models where two networks undergo parallel training processes. One network concentrates on generation, while the other specializes in distinguishing generated data from real ones. This training scheme gained attention due to its usefulness in generating new samples. However, this concept is relatively new which is why it requires further research and different fields (Yi et al., 2019). The two networks, specifically, are the generator (G) and discriminator (D) wherein the discriminator is a binary classifier trained to differentiate between real data from the training set and fake data generated by the generator. While the discriminator improves its classification accuracy, the generator learns to produce data increasingly indistinguishable from real data, thus maximizing the discriminator's errors. (Lee, J, 2022).

The generator and discriminator of GAN are trained simultaneously in a min max manner where its goal is to find the optimal solution for both the generator and the discriminator, where the generator tries to generate realistic samples that can fool the discriminator, while the discriminator tries to distinguish between real and fake samples (Goodfellow et al., 2014).

Figure 11. Loss Function of GAN (Lee, 2022)

$$min\, max\, V(G, D) \; = \; \Xi[log(D(x)] \; + \; \Xi[log(1 \; - \; D(z)]$$

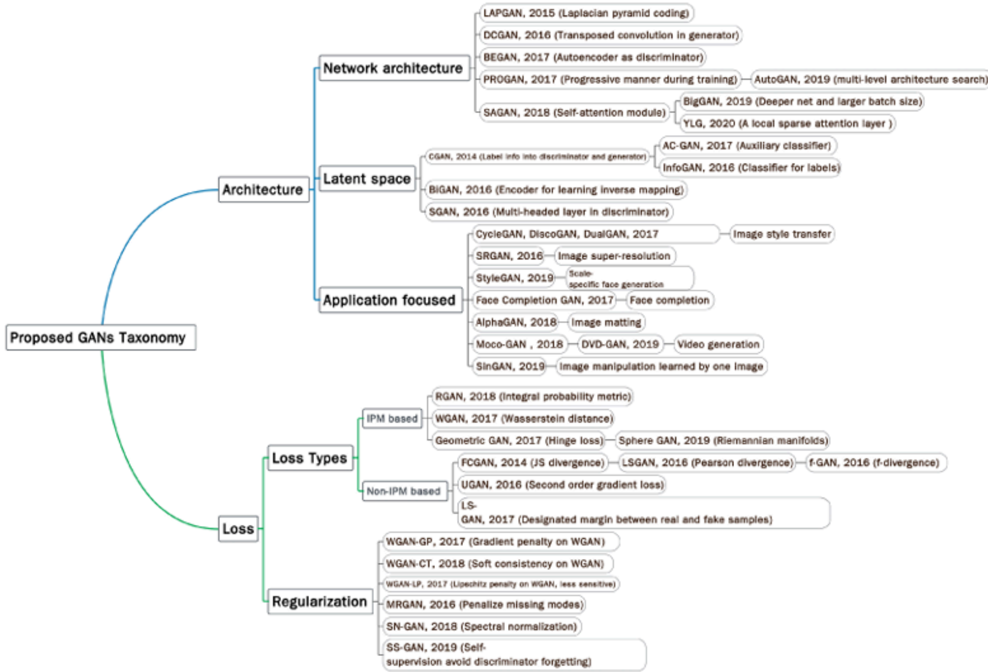
Figure above shows the minmax function where it operates such that the discriminator's goal is to maximize the probability of correctly classifying real samples as real (log(D(x))) and fake samples as fake (log(1 - D(G(z)))). While the generator's goal is to generate samples that can fool the discriminator into classifying them as real. The generator tries to minimize the term log (1 - D(G(z))), which means it wants the discriminator to output a high probability for the generated samples being real. In short , its goal is to make discriminator maximize its capability to correctly classify samples, while generator minimize the ability of discriminator to classify generated data as fake (Lee, 2022).

Despite its advantages, GAN still proposes significant challenges in data generation such as generation of quality data, diversity of generated data and stabilizing training. (Wang et al., 2021). In addition, GANs are very difficult to train and evaluate due to the necessity for their discriminator and generator to achieve Nash equilibrium, which is very difficult during training. Additionally, there are instances where the generator fails to learn the full distribution of datasets, which usually leads to a mode collapse issue (Dai et al., 2017). Following are the reason why there is still a need for further research in this method especially on some forms of data that haven't fully explored yet (Lee, J, 2022).

1. Generative Adversarial Network Variants

A study proposes a taxonomy that divides different GANs into two main variants: the architecture variant and the loss variant. The architecture variants focus on modifications to improve the overall GAN architecture. On the other hand, the loss variant, which is further divided into two categories – loss types and regularization, involves optimizations or additions that penalize the loss function. After application of GAN into image generation, the result of the study shows that addition of self-attention to both generator and discriminator, as in SAGAN, enhances image diversity. For training stability, spectral normalization emerges as an effective, easy to implement, low-cost loss function solution. Current state-of-the-art models like BigGAN and PROGAN produce high-quality, diverse images in computer vision, but applying GANs to video and other domains like time-series and natural language processing lags and presents opportunities for further research. (Wang et al., 2021).

Figure 12.The taxonomy of the recent GANs (Wang et al., 2021)

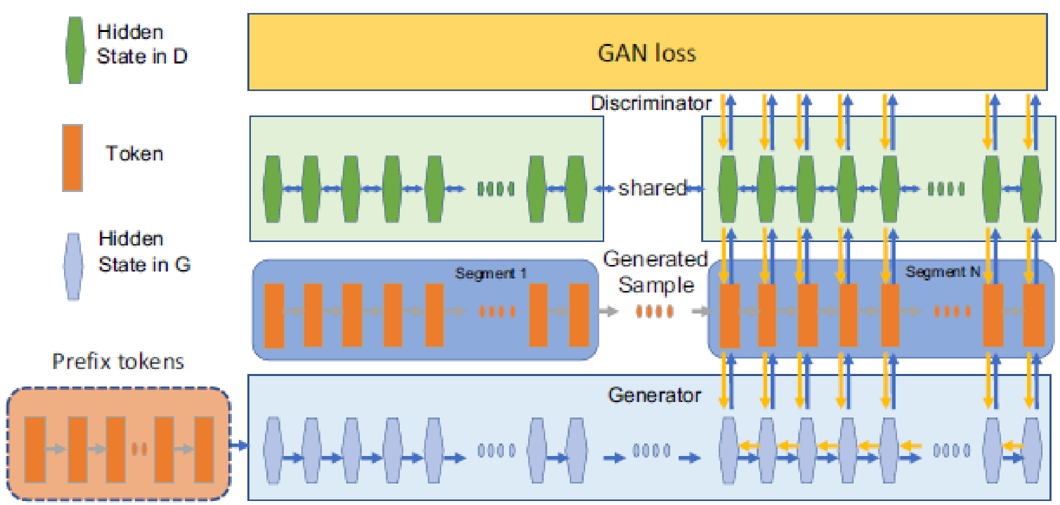


GAN has been studied for applications in various types of data, including natural language generation, music generation, and images (Lee, J., 2022). These applications often focus on areas within computer vision, such as image generation, attribute manipulation, image translation, semantic segmentation, and many more (Wang et al., 2021).

A certain study proposed a GAN architecture, named MELGAN, which is used for conditional audio synthesis. MelGAN is non-autoregressive, fully convolutional, and shows significantly faster than real-time on GPUs and CPUs without optimization tricks. In the study, researchers successfully trained GANs for raw audio generation without adding any distillation or perceptual loss function which results in high-quality text-to-speech synthesis (Kumar et al., 2019). Another study proposes a Transformer-based GAN approach for generating high-quality long symbolic music sequences. The model mainly uses a

Transformer-XL as generator for modeling the long-term dependencies in music and a pre-trained BERT model as its discriminator network which provides feedback to the generator using training. Several techniques like the Gumbel-Softmax trick, truncated backpropagation through time (TBPTT), and gradient penalties are incorporated to stabilize the GAN on training of the discrete sequences. The result of the study shows that the method, with BERT Discriminator, performs better than the baseline models such as Transformer GAN with WGAN, Transformer GAN with WGAN GPen, with PPO-GAN and Transformer-XL (Muhamed et al., 2021).

Figure 13. Transformer-based generative adversarial network (GAN) architecture (Muhamed et al., 2021)



Another GAN research, which proposes GAN-BERT extending the fine tuning process of BERT into two components such as generator G and a discriminator D, taking the form of a semi-supervised GAN (SS-GAN) framework. Although it resembles the behavior of GAN entirely but since it uses unlabeled samples meaning that it has to deal with unsupervised loss components. Results reveal that text classification tasks like topic categorization, question classification and sentiment analysis showed that GAN-BERT can drastically reduce the needs for labeled samples. Reason is mainly because GAN-BERT achieved better performance with only 50-100 labeled examples than a fully supervised BERT model trained on much more data (Croce et al., 2020).

2. Generative Adversarial Network for Time Series Data

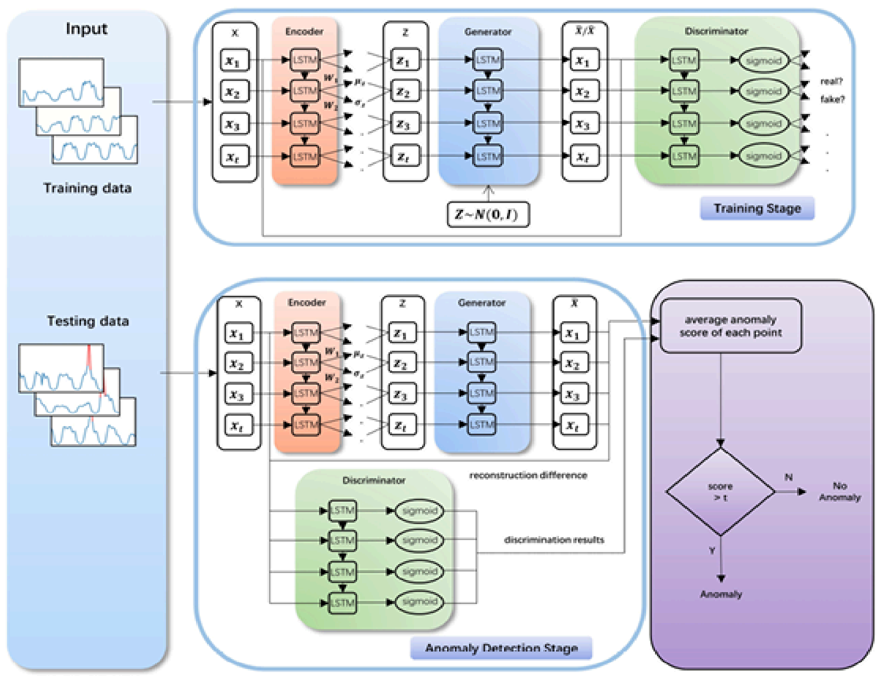
Although GANs have established their advantages in different forms of data, their application in generating synthetic temporal data, such as sequential and time-series data, is still relatively new. The G function and D function of it are almost similar, however the exact architecture of the models differs depending on the data feeded (Lee, J., 2022).

TimeGAN, or Time-series Generative Adversarial Networks, is another framework designed to generate realistic time-series data by preserving the temporal dynamics of variables over time. Introduced by Yoon et al. in 2019, this approach aimed to address the limitation of existing GAN methods for time-series generation, which do not adequately capture the temporal dynamics and correlations found in time-series data. TimeGAN integrates the flexibility of unsupervised GANs with the control of supervised training. It achieves this through a novel combination of an adversarial loss and a stepwise supervised loss, which ensures the model captures the conditional distributions of temporal transitions. Additionally, an embedding network reduces the high dimensionality of the adversarial learning space, enhancing the learning of temporal relationships and improving parameter efficiency. Empirical evaluation of the study shows TimeGAN's superior performance over state-of-the-art benchmarks across various datasets. These results highlight the effectiveness of TimeGAN's supervised loss and embedding network in generating data that closely mimics real-world temporal dynamics. Future work may explore incorporating differential privacy to enhance its utility in sensitive data applications (Yoon et al., 2019)

A study by Niu and his colleagues in Beijing proposed an LSTM-based Variational Autoencoder Generative Adversarial Network (VAE-GAN) for time series anomaly detection. The method monitors equipment states using time series data, dividing the process into two stages: model training, where the model learns the normal data distribution, and anomaly detection, where anomalies are identified by calculating anomaly scores. The LSTM-based VAE-GAN jointly trains the encoder, generator, and discriminator to leverage their combined abilities for more accurate and faster anomaly detection. Experiments using the selected time series dataset showed that this method has a higher F1 value, indicating a better balance between precision and recall in model performance, which enhances the model's ability to correctly identify positive instances while minimizing false positives. Higher precision means the model is more accurate when it predicts positive instances, reducing the number of negative instances incorrectly flagged as positive. Higher recall means the model is better at identifying all actual positive instances, reducing the number of true positives missed by the model. Additionally, it spends less time compared to traditional GAN-based methods, due to the avoidance of the optimization process at the anomaly detection stage. Higher recall means the model is better at identifying all actual anomalies, reducing the number of true anomalies missed by the model. Additionally, it spends less time compared to traditional GAN-based methods, due to the avoidance of the optimization process at the anomaly detection stage.

While some points' anomaly scores are calculated multiple times due to the moving window mechanism, the accuracy remains unaffected. Despite its accuracy and speed, the method has limitations, particularly in detecting successive anomaly subsequences, which requires a redesigned anomaly score module. Future improvements include developing an adaptive threshold adjustment method for quicker use (Niu et al., 2020).

Figure 14. LSTM-based Variational Autoencoder Generative Adversarial Network Architecture (Niu et al., 2020)



Despite its accuracy and speed, the method has limitations, particularly in detecting successive anomaly subsequences, which requires a redesigned anomaly score module. Future improvements include developing an adaptive threshold adjustment method for quicker use (Niu et al., 2020).

Another study that focuses on Time series Data, proposed a Variational Recurrent Neural Network GAN (VRNNGAN), which is a novel GAN framework for synthetic time-series generation. It mainly uses Variational Autoencoder (VAE) as a generator and incorporates the bidirectional RNN as a discriminator. With the recurrent VAE, temporal dynamics are captured and learned in the time varying latent space. The goal of the model is similar to different GAN frameworks, which is to generate realistic time-series data. Researchers uses three datasets and perform mainly four evaluations, such as t-SNE Plots, Post-Hoc Discriminative Score, Post-Hoc Predictive Score and Percentile Score to test the performance of the model and the generated synthetic data. The Post-Hoc Discriminative Score measures how easily a discriminator can tell apart real and synthetic data, with lower scores indicating better performance. The Post-Hoc Predictive Score assesses the accuracy of predictions made using synthetic data, with lower scores showing better preservation of real data patterns. While the t-SNE plot helps visualize how the synthetic data clusters with the real data in 2D/3D (Lee et al., 2022).

The result shows that the VRNNGAN generates realistic synthetic data, particularly excelling in tasks measuring the usefulness of synthetic data in time-series modeling. VRNNGAN performed well in t-SNE plots and Predictive Score evaluations, often outperforming other baseline methods like TimeGAN. While its Discriminator and Percentile Scores were generally strong, VRNNGAN showed variable performance across different datasets, with some limitations on complex, multivariable data such as stock datasets. Comparatively, VRNNGAN often surpassed TimeGAN, VRNN, and CRNNGAN in most evaluations.

However, VRNNGAN's training process is sensitive to hyperparameters, requiring careful tuning to optimize performance (Lee, J., 2022).

VII. Variational Autoencoder (VAE)

Variational Autoencoders (VAEs) are a type of generative model that share architectural similarities with traditional autoencoders. Both consist of an encoder (also called a recognition or inference model) and a decoder (or generative model), and they aim to reconstruct input data while learning from latent representations. The key distinction lies in how VAEs handle the latent space. Unlike regular autoencoders, VAEs create a continuous latent space. They achieve this by having the encoder output two vectors instead of one: a vector of means and a vector of standard deviations. The mean vector determines the central location of the encoded input, while the standard deviation vector defines the range of possible variations around that center. This design allows VAEs to sample encodings randomly within this range, exposing the decoder to different variations of the same input. This approach enables VAEs to capture more nuanced representations of the data, making them powerful tools for tasks like data generation and imputation (Saldanha et al., 2022).

What sets VAEs apart is their ability in capturing and generating the underlying distribution of real data. This capability makes VAEs particularly suitable for generating synthetic tabular datasets. Their strength lies in using variational inference to learn the patterns and relationships within the real data. Once trained, a VAE can produce high-quality synthetic samples by drawing random noise from a Gaussian distribution and passing it through the decoder (Bang et al., 2024).

1. **Variational Autoencoder for Data Augmentation and Imputation**

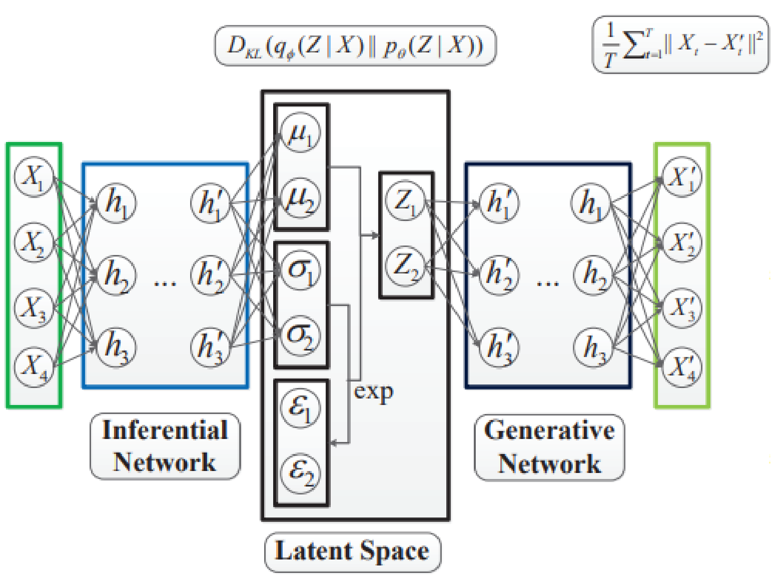
There are studies that used a VAE for augmentation and imputation of data. The study of Paepae et al. (2023) aimed to evaluate the effectiveness of using a variational autoencoder (VAE) for data augmentation to enhance the prediction accuracy of nitrogen (N) and phosphorus (P) concentrations in water quality monitoring. The researchers used a VAE to generate synthetic water quality data and assessed the similarity between real and generated samples using distribution plots and Jensen-Shannon divergence. They then trained various machine learning models, including Deep Neural Networks (DNN), K-Nearest Neighbors (KNN), Extremely Randomized Trees (ERT), Support Vector Regression (SVR), and XGBoost (XGB), on both the original and augmented datasets. The results demonstrated that the VAE-generated data closely mirrored the distribution of the original data. Moreover, the predictive performance, measured by Root Mean Squared Error (RMSE), improved by 10-35% when the datasets were doubled using VAE-generated data. RMSE quantifies the differences between predicted and observed values, which is a measure of model accuracy. Wherein, lower RMSE indicates better predictive performance, as it reflects smaller discrepancies between predictions and actual outcomes. Among the models, KNN and ERT showed the best performance in urban and rural catchments, respectively. The researchers concluded that VAE-based data augmentation significantly enhanced the predictive accuracy of virtual sensors, particularly in urban catchments. This approach allows for the use of fewer surrogate sensors while maintaining accuracy, which is advantageous for cost-effective water resource management.

On the other hand, the study of Boquet et al. (2019) addressed the problem of missing data in road traffic forecasting, which could negatively affect estimation accuracy. The authors proposed using a Variational Autoencoder (VAE) as an unsupervised data imputation method to learn the underlying traffic data distribution. They implemented a VAE with an encoder-decoder architecture to learn a continuous latent space that captured traffic data characteristics. The VAE was trained to minimize reconstruction error while regularizing the latent space. For imputation, missing values were initialized randomly, and the VAE iteratively refined the estimates until convergence. Using a real-world traffic dataset with induced missing data, they compared their VAE imputation against Principal Component Analysis (PCA) and a non-linear autoencoder (AE). The VAE significantly outperformed other methods, especially with Not Missing At Random (NMAR) patterns, showing up to 69.6% Root Mean Squared Error (RMSE) improvement in traffic speed forecasting.

2. Variational Autoencoder Variants

The study of Li et al. (2021) proposes a shift correction β-VAE (SC-β-VAE) model, a variant of the Variational Auto-Encoder (VAE), to address the imputation of specific missing values in multivariate time series data. The standard VAE assumes that training and test data follow the same distribution, which is often violated in real-world scenarios like meteorological or air quality data, where missing values are concentrated in specific periods. This concentration causes a shift in the original probability distribution, leading to decreased imputation performance.

Figure 15. The network architecture of the standard VAE model (Li et al. 2021)



To address this, the authors introduce a shift correction hyperparameter λ to modify the latent space's Gaussian distribution from $N(\mu, \sigma)$ to $N(\mu + \lambda\sigma, \sigma)$, correcting the deviated distribution. Additionally, they extend this to the β-VAE model, introducing a hyperparameter β to control the trade-off between the model's generative ability and disentanglement. The SC-β-VAE outperforms baseline methods, including statistical, machine learning, and generative models, in imputing specific missing values (Missing Not At Random) and random missing values (Missing Completely At Random) in real-world datasets. The study concludes that the SC-β-VAE effectively improves imputation accuracy by correcting the shifted probability distribution and balancing the model's generative and disentanglement abilities.

The study of Qiu et al. (2020) proposes using a variational autoencoder (VAE) framework for imputing missing values in genomic data, such as transcriptome and methylome data. They introduce a shift-correction (SC) variant of VAE to address scenarios where the missing values are drawn from a different distribution than the training data, a common occurrence in certain missing data patterns.

In the SC-VAE, they modify the assumption of the training data distribution to follow a shifted Gaussian. This adjustment allows the model to better impute cases where missing values are concentrated at lower values. Their results demonstrate that for scenarios with missing values skewed towards lower expression levels, the SC-VAE achieves better imputation accuracy compared to traditional methods like K-nearest neighbors and singular value decomposition.

The authors also investigate the effect of varying the latent space regularization strength in VAE, showing that stronger regularization decreases imputation performance. They attribute VAE's imputation ability to the noise injection in the latent space, which is absent in regular deterministic autoencoders. This demonstrates that VAE, particularly the SC variant, can be an effective and computationally efficient alternative to traditional methods for imputing missing values in genomic data, especially in scenarios where missing data exhibits specific patterns.

Additionally, the authors explore the use of β-variational autoencoder (β-VAE), a generalization of VAE that introduces a hyperparameter β to balance the reconstruction loss with the regularization loss. When $\beta > 1$, the latent space is smoother and more disentangled, improving encoding efficiency. Conversely, when $\beta = 0$, the regularization term is removed, resembling a simple autoencoder with noise injected into the latent space. Experiments with varying β (0, 1, 4, 10) under different missing data scenarios (5%, 10%, 30%) show that imputation performance is similar for $\beta = 0$ and $\beta = 1$, while higher β values worsen performance. This suggests that strong regularization does not benefit imputation tasks and that noise injection into the latent space is crucial for VAE's imputation ability.

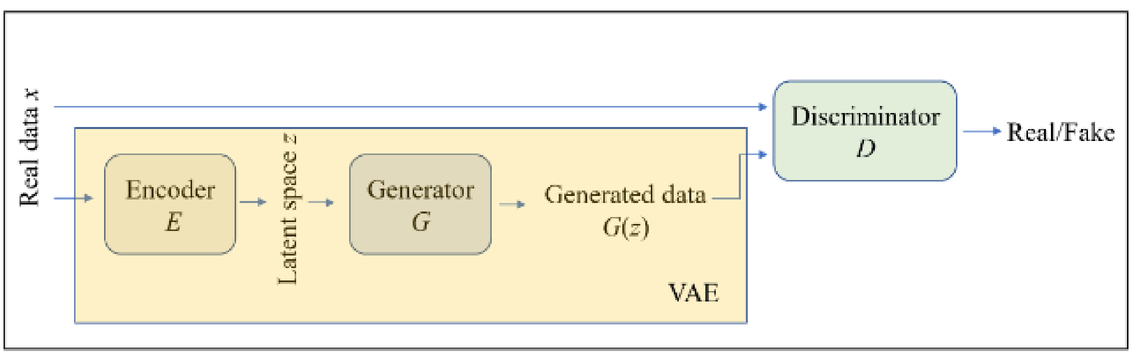
VIII. Variational Autoencoder- Generative Adversarial Network (VAE-GAN)

Generative models are one way to synthesize time series data. Unlike random transformations that use random noise, slicing, cropping, or scaling, generative models take a less direct route and use the distributions of features in the datasets to generate new patterns (Iwana & Uchida, 2021). Some of the popular generative models include Variational Autoencoder (VAE) and Generative Adversarial Network (GAN).

These two generative models have their own strengths and weaknesses. For example, VAE uses the idea of probabilistic inference and reparameterization tricks to get various latent code z (Li, 2023). Meanwhile, GAN is a generative neural model based on a competition between two neural networks (Iglesias et al., 2023). The use of both generator and discriminator has made it possible for a GAN model to produce realistic data. However, these two have disadvantages such that VAE may be stable during the training process, but as it is being optimized to match the reconstruction loss of given inputs, it might produce blurry images (Ham et al., 2020). On the other hand, GAN is typically hard to train and they suffer from mode collapse, instability, and evaluation metrics (Iglesias et al., 2023). Due to this, balancing the generator and discriminator is difficult because in some instances, the discriminator converges faster than the discriminator (Ham et al., 2020).

In image processing, VAEs would typically produce blurred and low quality images since the input from the encoder only uses a simple element-wise error, however, humans see images in its high feature form.

Figure 16. VAE-GAN Architecture (Ruan et al., 2023)



On the other hand, GANs would be able to generate sharp images while also capturing important features from it. However, training of a GAN model would be difficult because the generator learns from a random distribution of z, and its loss function depends on the discriminators. The quality of the images will start to degrade if the generator will be able to fool the discriminator and its ability to classify the real ones from the fake one is nearly 50%, the discriminator might accept strange generation results. To solve this problem, combining the strengths of these two generative models have been proposed over the years. As seen in Figure 16, VAE-GAN utilizes the latent variable model of VAE to generate the data and uses the discriminator of GAN to evaluate the authenticity of the generated samples (Ruan et al., 2023). It uses an encoder to input the data into the latent space, and the encoded latent vector will be used as an input for the decoder or generator. The data reconstructed by the decoder will be used as an input for the discriminator where it will determine whether the generated data is real or fake. A loss function will be used for cross-learning between the two networks, which will improve the generative power of the model (Hu et al., 2023). This combination would allow the model to enhance its generative capabilities by generating more high quality and diverse samples of time series dataset.

1. **Variants of VAE-GAN**

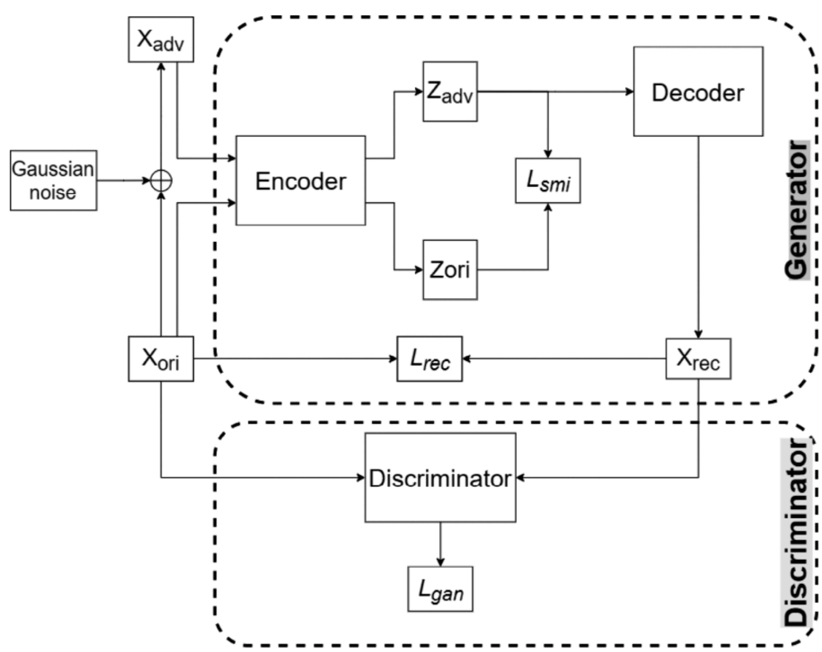
VAEGAN was first proposed by Larsen et al. in 2015 where it uses an end-to-end architecture that is configured in the order of encoder, decoder, and discriminator (Ham et al., 2020). And for the past years, there have been several studies conducted where novel approaches and hybrid models were developed.

For instance, in the study of Ye and Bors (2020), they developed a model capable of learning new tasks without forgetting the previously known knowledge while also retaining meaningful and disentangled representation of data. The Lifelong VAEGAN (L-VAEGAN) is a hybrid model following the lifelong learning framework and VAEGAN architecture. It addresses the gap of Generative Replay Mechanism (GRM) where it is capable of learning multiple tasks while keeping previously learned knowledge using generative models like GAN or VAE, but lacks the ability to learn latent data representation. A two-step optimization algorithm called “wake” and “dreaming” were used to train the hybrid model. This was applied to supervised, semi-supervised, and unsupervised learning. Under supervised learning, L-VAEGAN was trained to learn MNIST to SVHN and MNIST to Fashion lifelong learning tasks. The results of their experimentation showed that L-VAEGAN achieved higher accuracy compared to state-of-the-art models such as LGAN. Meanwhile, training L-VAEGAN under semi-supervised learning where only a small number of labeled data from different databases were considered. There were 1,000 labelled images considered for MNIST dataset, while 10,000 images were considered for Fashion database. The results of their experimentation showed that in a semi-supervised setting, the proposed model outperformed LGAN and has a competitive results compared to other models. Lastly, the training of L-VAEGAN in an unsupervised setting showed that the proposed model could learn multiple tasks while keeping previously known knowledge. Their experimentation showed that L-VAEGAN was able to smoothly interpolate images coming from different sources. Such that, the proposed model was able to transition from a chair to a human face. This could mean that L-VAEGAN could learn from multiple domains while keeping a meaningful latent representation of data without having a catastrophic forgetting.

In another study by Ham et al. (2020), they proposed a model called Unbalanced GAN where it uses VAE as a pre-trained generator to balance the convergence between the GAN's generator and discriminator. First, an input of the dataset was fed to the VAE model and the value of the weights of its decoder will be used to initialize the generator. And the latter was trained using a GAN loss function. A pre-trained generator using VAE was employed to address the issue of mode collapse and to ensure that the GAN model would not converge to a strange distribution. In their experimentation, the proposed model was compared to other GAN models including DCGAN, LSGAN, and WGAN. They used three different datasets namely MNIST, CIFAR-10, and LSUN Bedroom. The results of their experimentation showed that the proposed model has identical support and outperforms other models in terms of inception score. Support values give insight into the distribution of instances for each class in the dataset, with identical support indicating that each class is equally represented. The inception score, in this case, serves as a performance metric to measure the quality and diversity of the generated images.

Chen et al. (2023) proposed D-VAEGAN model to classify DeepFake images by using denoising and VAEGAN framework. Adversarial attacks can bypass current DeepFake detection systems by adding perturbation to the images that are typically not noticeable by the human eyes, hence training an adversarial model such as VAEGAN was proposed to combat this. The generator, VAE was used to remove the perturbations from the images, while the discriminator was used to classify whether the reconstructed image is real or fake.

Figure 17. D-VAEGAN Architecture (Chen et al., 2023)



In the figure above, first, an input of clean image z and adversarial example z′ were used as an input to the encoder where z′ used a denoising technique to remove the perturbation effects from the images. Then using this, features will be mapped out in the latent space and be used by the decoder to reconstruct the images. Loss functions were used to measure the similarity between the original and reconstructed image. Then the output from the decoder will be used as an input for the discriminator to classify whether the reconstructed image is real or fake. An adversarial loss is measured to ensure that the generative capability of the generator is derived from the discriminator loss. Their experimentation included testing this to adversarial attacks such as FGSM, BIM, PGD, DeepFool, and C&W. The results showed that the proposed model yielded robustness in terms of classifying DeepFake images.

2. VAE-GAN for Synthetic Data Generation

In a study by Lee (2022), he proposed a novel VAEGAN model called VRNNGAN for generating time series synthetic data. It uses the GAN framework where the generator uses a recurrent VAE, and a bidirectional RNN was used as the discriminator. The VRNN is the generation G function, while bidirectional RNN is the discriminator D function. In the G function, the VRNN model has the capability to reconstruct and generate time series data. The real sample, labeled as “real”, along with the reconstructed and generated sample that derived from the inference and generation model respectively, labeled as “fake” were fed to the D function where it has a strong capability of classifying which data is real or fake. A loss function was measured to ensure the balance between the generator and discriminator. In the experimentation, the proposed model was evaluated using three different time series datasets namely ARMA, ECG200, and Stocks.

The results of the experimentation showed that the model performed well in different performance metrics such as t-SNE plots where the plots between the synthetic data and original data overlapped in the same area. Meanwhile, only the ECG dataset had the lowest Post-Hoc Discriminative Scores while TimeGan and VRNN performed much better in ARMA and Stocks datasets respectively. On the other hand, the proposed model yielded good results for both Post-Hoc Predictive Score and Percentile Scores. This shows that VRNNGAN is capable of generating synthetic data, however, further improvements might be needed.

The study of Zhang et al. (2023) used VAEGAN to generate synthetic data for sea-land clutter classification of sky-wave-over-the-horizon radar (OHTR). Since deep learning models usually need a large amount of data, class imbalance and scarce data are common problems for sea-land clutter classification of OHTR. To address this problem, an auxiliary classifier was introduced to the VAEGAN architecture, and an AC-VAEGAN model was developed. Unlike typical VAEGAN, the proposed model can specify the class of the synthetic data. The AC-VAEGAN is composed of three parts, namely En, De/G, and D/C. The En is responsible for encoding the data to the latent space, and the output (z_deco) from this is used as an input for the De/G. The latter not only takes z_deco, but it also takes its attribute c as an input, along with random noise vector z_gen and its class attribute c. The loss function for both generator and discriminator is also computed. In their experimentation, they used a benchmark dataset for sea-land clutter of OHTR, and validated it using the MSTAR dataset. The results of their experimentation showed that AC-VAEGAN outperforms AC-GAN in both traditional GAN metrics such as GAN-train and GAN-test, and statistical evaluation including absolute distance (AD), cosine similarity (CS), and Pearson correlation coefficient (PCC). However, it had some limitations such as the loss function lacks an interpretable indicator that can guide the training process of the model.

Synthesis of the Reviewed Literature and Studies

The studies reviewed underscore the importance of handwriting data, particularly in online analysis, for its ability to capture unique features inaccessible through offline methods. Online handwriting data has shown notable performance, especially with deep learning networks, but to achieve effectiveness, large datasets are necessary. Additionally, due to the limitations in recording in-air features with current stylus and tablet technologies, there is often missing data, thereby impacting the integrity of the analysis.

Literatures have supported that to increase dataset size, data augmentation is a more practical method than creating new data from scratch. Efforts to address the issue of low dataset sizes through augmentation encounter challenges, primarily relying on traditional methods that necessitate expert label maintenance and oversight. Moreover, while time series data augmentation techniques are prevalent, they have largely overlooked the nuances of multivariate datasets like handwriting data. Nonetheless, studies have suggested that by combining various methods and promising advancements in deep generative models, there is potential to generate realistic synthetic handwriting samples. These advancements could help extend these models to better handle time series data modeling for high-dimensional, multivariate datasets.

However, imputation techniques specifically for handling missing in-air features in online handwriting data remain underexplored. Existing imputation methods prove ineffective, especially when dealing with data missing completely at random which is often more reflective of real-life scenarios. Consequently, there exists a notable gap in the research around developing imputation techniques tailored to effectively handle missing data in such high-dimensional, multivariate time series datasets while preserving the integrity of the recorded samples.

Despite the advancements of generative adversarial networks (GANs), they still face significant limitations in effectively modeling complex multivariate time series distributions. While recent studies have shown promising results in using GANs to generate synthetic time-series data, challenges remain in optimizing hyperparameters and ensuring robust performance across diverse datasets and applications.

Current VAE-based imputation methods show promise for handling missing data in various domains. However, their effectiveness in imputing missing values in time series data remains a challenge. One notable limitation is the assumption that training and test data follow the same distribution, which is often violated in real-world scenarios. This discrepancy leads to decreased imputation performance, especially when missing values are concentrated in specific periods, causing a shift in the original probability distribution. In addressing this issue, recent studies propose innovative approaches such as the shift correction β-VAE (SC-β-VAE) model. This variant of VAE introduces a shift correction hyperparameter to modify the latent space's Gaussian distribution, effectively correcting the deviated distribution caused by missing values.

Recent studies have started combining Variational Autoencoder (VAE) and Generative Adversarial Network (GAN) to leverage the latent variable model of VAE and the discriminative power of GAN for improving the quality and diversity of generated data. Despite their successful application in synthetic data generation across various domains, there is currently no exploration, based on the author's reviewed studies, of their potential for imputing missing values in multivariate time series data.

Based on these findings, the current study explored the use of the shift correction mechanism to effectively address the distributional shifts caused by missing data, which is a common issue in online handwriting datasets. This mechanism was incorporated into a VAE-GAN model, combining the strengths of both VAEs and GANs. VAEs excel at learning robust representations of complex data distributions, while GANs are adept at generating high-quality synthetic samples. Unlike many existing imputation and augmentation methods designed for univariate data, the developed model is specifically tailored to handle the complexities of multivariate time series data, such as handwriting. By generating synthetic data that closely mimics the characteristics of the original handwriting data, the developed model offered an inexpensive, time-efficient method to address the challenges posed by limited and missing data in handwriting datasets, potentially leading to improved performance in various handwriting analysis tasks. Additionally, to better demonstrate the effectiveness of the developed model, three types of baseline models—VAE-GAN, TimeGAN, and VRNNGAN— were selected for comparative experiments.

Chapter 3

METHODOLOGY

This chapter presents the utilized methodology in the study, as well as an outline of a comprehensive summary of the research design, sources of data, system architecture, research instrument, data generation/gathering procedures, and statistical analysis, which were employed to evaluate the performance of the system using the evaluation metrics that are addressed in the statement of the problem, and to achieve the study's objectives.

Research Design

The study implemented a quantitative quasi-experimental design to evaluate the effectiveness of models based on the augmentation and imputation techniques used. In this design, it is expected that changes in the independent variables will lead to changes in the dependent variables. The independent variables in this study were the imputation and augmentation models, including VAE-GAN, TimeGAN, VRNNGAN, and the developed SC-β-VAE-GAN. These models were assessed in generating synthetic data for two distinct datasets: EMOTHAW and GPS time series data from Greenland. The evaluation was based on several dependent performance metrics, including Normalized Root Mean Square Error (NRMSE), Post-Hoc Discriminative Score, and Post-Hoc Predictive Score.

The quasi-experimental design does not require the use of control and experimental groups, nor does it rely on randomization. This design was well-suited for this study because it does not require the establishment of control and experimental groups, and there is no need to randomly assign participants to different treatments or conditions. Instead, the study utilized pre-existing datasets that were not collected through a randomized process, rendering randomization unnecessary. It was focused on assessing the impact of different imputation and augmentation techniques on the effectiveness of the models.

Sources of Data

The conducted study used a handwriting multivariate time series dataset, which was an essential requirement. The researchers augmented and imputed the EMOTHAW dataset, a novel and publicly available database developed by Likforman-Sulem et al. (2017) for recognizing emotional states through handwriting analysis. The EMOTHAW dataset consists of handwriting samples from 129 participants, all of whom are Master’s and BS students from the Department of Psychology at the Seconda Università di Napoli in Italy. The participants' ages range from 21 to 32 years, with a mean age of 24.8 years.

The database includes seven tasks: drawing interlinking pentagons, drawing a house, writing four Italian words in capital letters, drawing loops with both hands, performing the Clock Drawing Test, and writing a phonetically complete Italian sentence in cursive. These tasks were carefully chosen based on well-established medical and psychological tests used for cognitive impairment detection and personality assessment.

To collect the data, Likforman-Sulem used an INTUOS WACOM series 4 digitizing tablet and an Intuos Inkpen. This setup allowed them to capture not only the on-paper handwriting but also in-air movements, which have proven to be as important as on-surface information in previous studies. The tablet recorded various parameters for each data point, including x-y positions, timestamps, pen status (up or down), pressure, and pen azimuth and altitude angles.

In addition to the handwriting tasks, each participant completed the Italian version of the Depression Anxiety Stress Scales (I-DASS-42) questionnaire. This self-report tool assesses the levels of depression, anxiety, and stress experienced by the participants over the past week, providing ground truth labels for the handwriting data. The researchers have already contacted the authors and successfully obtained the EMOTHAW dataset. The handwriting data is in SVC file format and the I-DASS-42 data is in xls file format.

The researchers also used the Greenland GPS dataset for the validation of the developed SC-β-VAE-GAN model, which was also used in the study of Zhang et al. (2021) to impute data. It is a dataset from Nevada Geodetic Laboratory, the University of Nevada at Reno. The dataset consists of daily GPS coordinate time series from 20 stations in Greenland wherein stations provide data for geodetic and geophysical studies, but due to logistics challenges and hardware malfunctions, especially in harsh polar environments like Greenland, these time series often contain gaps or missing values. The dataset was useful for validation purposes as it represented real-world time series geospatial data with natural occurrences of missing values, making it ideal for evaluating the ability of the developed model to handle and accurately impute missing data in time series.

Research Instrument

The researchers developed a tool to impute and augment data based on the VAE-GAN framework. To address the research question, the tools were evaluted. Additionally, the researchers followed a step-by-step procedure detailed in the experiment paper, which covered data preparation, model evaluation, and comparison with existing models.

The primary programming language utilized for developing the experimental framework is Python, chosen for its extensive libraries and frameworks tailored to machine learning and data processing needs. Complementing Python, Visual Studio

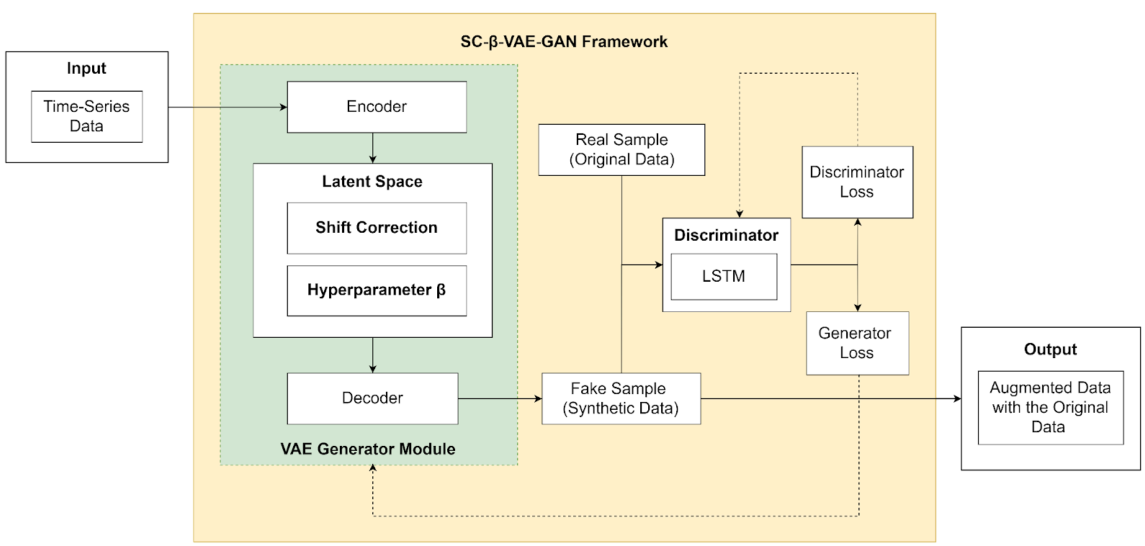
Code (VS Code) was employed for code development and execution as it provides a robust and versatile development environment that supports data exploration, visualization, and comprehensive documentation of the experimental process.

On the other hand, for data processing, especially for time series data, the researchers utilized key Python libraries such as Pandas for data manipulation and analysis, NumPy for numerical computations, Matplotlib and Seaborn for data visualization, and TSFresh for extracting relevant features from time series data.

The experiment paper included detailed procedures, tables, and tools necessary for recording and evaluating the results of the generated synthetic data. This document ensures a systematic approach to the experimentation process and provides a clear framework for assessing the performance and quality of the results.

System Architecture

Figure 18. System Architecture of SC-β-VAE-GAN for Generating Synthetic Data for Imputation and Augmentation



The SC-β-VAE-GAN framework was designed to augment and impute multivariate time-series data. The process began with the input of time-series data, which was fed into the VAE Generator Module. This module consists of an encoder, a latent space with shift correction and hyperparameter β, and a decoder. Initially, the encoder processed the time-series input data, compressing it into a lower-dimensional latent space representation. Within this latent space, a shift correction mechanism adjusted the latent variables, while the hyperparameter β controled the degree of regularization in the VAE. The decoder then took this processed latent space representation and generated synthetic data intended to closely resemble the original time-series input data.

The synthetic data generated by the VAE Generator Module was then passed to the Discriminator Module, which included an LSTM-based discriminator. This discriminator distinguished between real and synthetic data. The real time-series input data was labeled as a real sample, while the synthetic data generated by the decoder was labeled as a fake sample. The discriminator evaluated both real and fake samples, outputting a discriminator loss that reflected how accurately it can differentiate between the two. Concurrently, the framework calculated two types of losses: the discriminator loss and the generator loss. The discriminator loss was incurred when the discriminator incorrectly classified real and fake samples, while the generator loss was based on the generator's ability to produce synthetic data that can fool the discriminator.

The framework operated within an adversarial training loop, where the synthetic data generated by the VAE generator was continuously improved by minimizing the generator loss. Simultaneously, the discriminator was trained to maximize the discriminator loss, thereby enhancing its ability to distinguish real data from synthetic data. This iterative process continued until the synthetic data becomes indistinguishable from the real data. The final output of the SC-β-VAE-GAN framework was augmented data, which comprises both the original time-series data and the high-quality synthetic data produced by the VAE generator. This augmented data can subsequently be utilized for various downstream applications, such as training machine learning models, thus demonstrating the effectiveness and utility of the SC-β-VAE-GAN framework in enhancing original datasets for improved analysis and model performance.

Data Gathering/Generation Procedure

The following steps were followed by the researchers in generating the data:

1. The researchers collected the handwriting samples from the public database, EMOTHAW. This undergone pre-processing where the data were sorted and organized by handling missing values and normalizing the data.
2. After pre-processing, the dataset was fed to the SC-β-VAE-GAN model, following the GAN framework, where VAE acted as a generator, and a Long Short-Term Memory (LSTM) network was the discriminator.
    a. The inferential network or encoder inputted the dataset to the latent space. A Gaussian Distribution was used to map the probability distribution of the dataset in the latent space. Then, using a Gaussian Distribution with shift correction where the same mean and variance was used to sample the latent vector Z. The mean was modified by adding a shift parameter multiplied to the variance. This was used to ensure that the model could make an assumption close to the original probability distribution as much as possible when learning input data with missing values.
    b. A hyperparameter β was used to control the trade-off between the generation and disentanglement of the data. The value was

$0<\beta<1$, which allowed the the β-VAE model to balance the regularization and reconstruction term for better generation and disentanglement of data. This was to make sure that the model generated better data imputation.

c. Adding these two hyperparameters allowed the model to have a better learning of the handwriting dataset that has been inputted. A latent vector Z was sampled using the adjustments in the latent space. Next, using this as an input, the generative network or decoder reconstructed the handwriting time series dataset where synthetic data was generated.

d. Using the synthetic data and the real sample data, these two were fed to the LSTM discriminator where the generated data was classified as real or fake.

e. A loss function for both generator and discriminator was computed. The former was used to measure how well the generated samples are close to the real ones by making the discriminator believe that they are real. Meanwhile a discriminator loss measured how well the LSTM distinguishes between real and fake samples.

3. The performance of the model in generating synthetic data was evaluated using Normalized Root Mean Square Error (NRMSE), Post-Hoc Discriminative Score, and Post-Hoc Predictive Score.

Ethical Consideration

To ensure ethical conduct, the researchers obtained ethical approval first from the ethics committee at the Polytechnic University of the Philippines before conducting the study. The study used the datasets EMOTHAW and Greenland GPS, and ensured appropriate use by strictly adhering to the terms and conditions specified by the dataset providers. The researchers gave proper credits to the authors of the original datasets and complied with data protection regulations. Maintaining anonymity and confidentiality is a priority. All personally identifiable information were anonymized. Transparency was ensured throughout the study by thoroughly documenting the methods, data preprocessing steps, model training procedures, and evaluation metrics. Research findings, including negative results, was reported transparently, providing an honest account of the research process.

Analysis and Statistical Treatment of Data

The following data treatment was used to solve the statement of the problems of this study, as well as to analyze the result of the model performance and its comparison to other imputation and augmentation methods.

1. **Normalized Root Mean Square Error (NRMSE)**

   This metric was used to measure the average discrepancy between the synthetic data generated by SC-β-VAE-GAN and the real data, normalized by the range of the observed data. This metric provided evaluation of the model's accuracy, with lower NRMSE values indicating high performance.

   Each data consist of sequence of data points over time, definite Y_i be the observed value at the i-th point, which can be a vector including the features

$Y_i = (x_i, y_i, p_i, \theta_i)$ and $\widehat{Y}_i$ be the predicted value (synthetic data) at the i-th time point which is $\widehat{Y}_i = (\widehat{x}_i, \widehat{y}_i, \widehat{p}_i, \widehat{\theta}_i)$. The computation of error for each were computed by $Error_{f,i} = y_{f,i} - \widehat{y}_{f,i}$. Then RMSE was calculated for each feature across all time points.

The Root Mean Square Error (RMSE) was calculated using the formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Y_i - \widehat{Y}_i)^2}{N}}$$

To normalize the RMSE and make it comparable across different data, the NRMSE was computed by dividing the RMSE by the mean of the observed values.

$$NRMSE = \frac{RSME}{mean(y)}$$

A lower NRMSE indicates that the generated synthetic data closely resembles the real data, reflecting the model's strong performance in accurately capturing the characteristics of the handwriting samples. This suggests that the model is effective in learning and reproducing the underlying patterns and features of the original data, making the synthetic data highly realistic. Conversely, a higher NRMSE suggests that the synthetic data deviates significantly from the real data, highlighting areas where the model may need further refinement and improvement. This indicates that the model is less effective in replicating the true characteristics of the handwriting samples, resulting in synthetic data that is less realistic and potentially less useful for subsequent analysis or applications. The results for the models was summarized by calculating the mean and standard deviation to facilitate comparison with other models.

2. Post-Hoc Discriminative Score

In the Post-Hoc Discriminative Score task, an LSTM model was trained to classify data as either real or synthetic. The goal is to achieve a classification accuracy of 50%, which indicates that the classifier cannot distinguish between the real and synthetic data. This outcome would mean that the synthetic data is effectively indistinguishable from real-world data, demonstrating its validity as real-world data.

Researchers began by combining the real and synthetic data into a single dataset. Then a 10-fold cross-validation approach was performed to ensure the reliability of the results, with nine folds used for training and the remaining fold for testing. To evaluate the results, the accuracy must be around 50%. The classifier can achieve this 50% accuracy in the following ways:

- Correctly identifying all real data and incorrectly identifying all synthetic data.
- Incorrectly identifying all real data and correctly identifying all synthetic data.
- Correctly identifying some real and some synthetic data, while incorrectly identifying others, such that the total correct guesses amount to 50%.

The main objective was to produce synthetic data such that the performance of the binary classifier is comparable to that of a random classifier. If the classifier achieves 100% accuracy, it means the synthetic datasets have failed to generate realistic data. This leads to the interpretation that the lower the discriminative score, the more realistic the synthetic data, while a higher score indicates a lack of realism. The results for the models was summarized by computing the mean and standard deviation for easier comparison with other models.

3. **Post-Hoc Predictive Score**

The Post-Hoc Predictive Score measured the prediction accuracy of synthetic data generated by a model. To evaluate this, all synthetic data was used, with the last time step of each sample set aside as the prediction target, and all previous time steps used as input for the model. The model was then tested on real data, serving as the ground truth holdout set. Min-max scaling was applied to normalize the data, with each experiment repeated 10 times to ensure consistency and validation.

The Mean Absolute Percentage Error (MAPE) was used to evaluate the model's performance in predicting the last time step. Additionally, MAPE was calculated as the average of the absolute percentage errors between the actual and predicted values. The Mean Absolute Percentage Error (MAPE) was used as the predictive score:

$$MAPE = \frac{\sum_{i=1}^{N} \left(\frac{Y_i - \widehat{Y}_i}{Y_i}\right) x\ 100}{N}$$

To interpret the result, lower MAPE indicates that the synthetic data accurately captured the patterns of the original data, while higher MAPE suggests poor predictive performance. For each model, the mean and standard deviation of MAPE were computed to aid in visualization and comparison, providing a clear assessment of how well the synthetic data generated by the proposed technique preserves the predictive characteristics of the original dataset from different models.

4. Hypothesis Testing

The **Kruskal-Wallis H-test** was employed to examine whether there exists a significant difference in the performance of synthetic data generation among various models, specifically SC-β-VAE-GAN, VAE-GAN, TimeGAN, and VRNNGAN. This non-parametric test was chosen as it does not assume normality of the data and is suitable for comparing more than two independent groups. The computed performance metrics used in this analysis include **Normalized Root Mean Square Error (NRMSE)**, **Discriminative Score**, and **Predictive MAPE Score**.

The formula for the Kruskal-Wallis H-test is:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} R_i^2/n_i - 3(n + 1)$$

Where:

- H= Kruskal-Wallis statistic
- n= Total number of observations
- n_i = Number of observations in group i
- R_i = Sum of ranks in group i

If the Kruskal-Wallis H-test yields a statistically significant result (p-value ≤ 0.05), pairwise comparisons between models are conducted using the **Dwass-Steel-Critchlow-Fligner (DSCF) method**. The DSCF test is a rank-based non-parametric procedure that adjusts for multiple comparisons and provides a robust analysis of pairwise differences.

A p-value ≤ 0.05 in the Kruskal-Wallis H-test or DSCF pairwise comparisons was interpreted as significant, indicating rejection of the null hypothesis (H_0: "There is no significant difference in the performance among the models"). Conversely, a p-value > 0.05 suggests that the null hypothesis cannot be rejected.

Chapter 4

RESULT AND DISCUSSION

The chapter analyzes and presents the study's collected data results, focusing on addressing the problems outlined in the research. The main focus of the research is on generating synthetic data for the imputation and augmentation of handwriting multivariate time series data, particularly to address challenges posed by limited and missing data. The dataset used in this study includes 903 online handwriting time series samples.

The research aims to address three statements of the problem. The first problem statement is how the SC-β-VAE-GAN performs in generating synthetic data based on Normalized Root Mean Square Error (NRMSE), Post-Hoc Discriminative Score, and Post-Hoc Predictive Score. The second problem statement is how VAE-GAN, TimeGAN, and VRNNGAN perform in comparison, using the same evaluation metrics. The final problem statement is whether there are significant differences between SC-β-VAE-GAN and the other models in terms of synthetic data generation performance.

To address the research problem, the researchers applied a statistical approach to derive insights. In particular, the Kruskal-Wallis H-test was utilized to assess whether there is a significant difference in the performance of synthetic data generation across different models, with a focus on SC-β-VAE-GAN compared to other models such as VAEGAN, TimeGAN, and VRNNGAN. The analysis used performance metrics including NRMSE, Discriminative Score, and Predictive MAPE Score.

Dataset

1. EMOTHAW Dataset

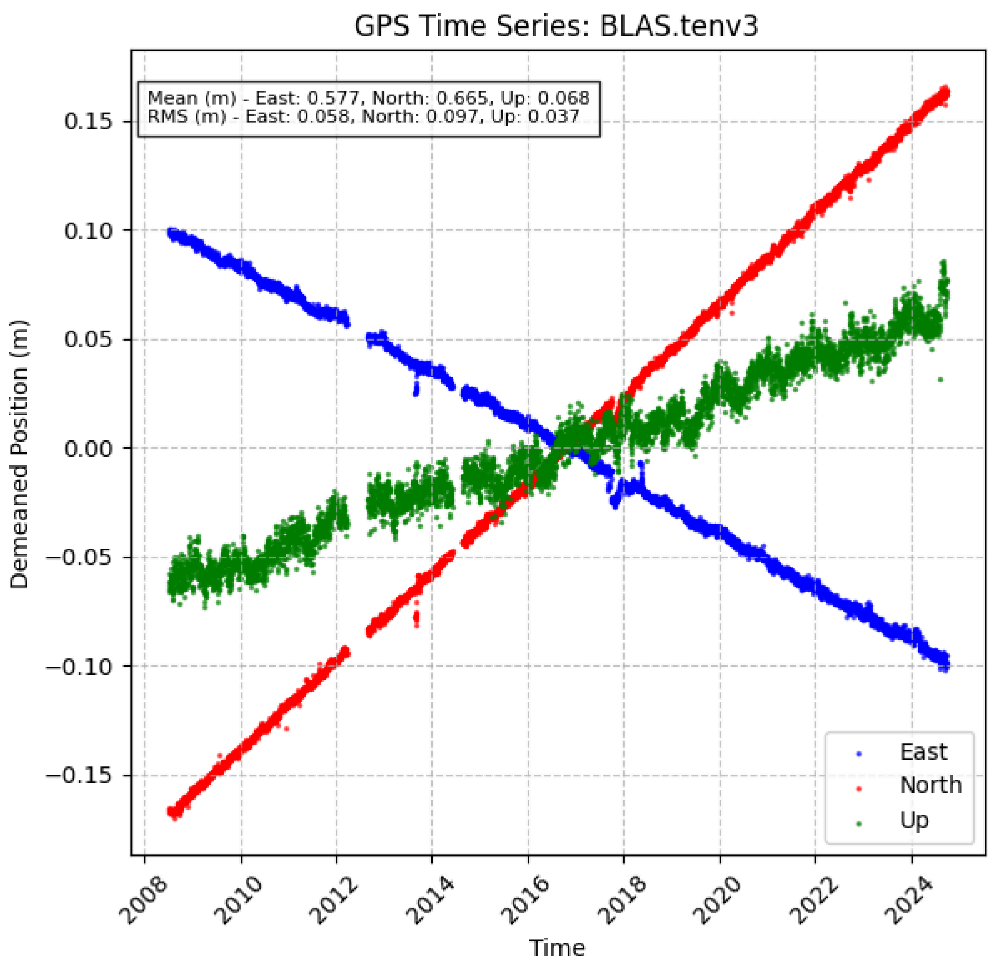
Figure 19. Preprocessed EMOTHAW Time Series Data with columns x, y, timestamp, pen status, azimuth, altitude, and pen pressure

```
47944 33492 0 1 1800 490 67
47949 33506 7 1 1800 500 148
47949 33512 15 1 1800 500 193
47950 33515 22 1 1800 500 228
47950 33519 30 1 1800 500 270
47950 33524 37 1 1810 500 306
47950 33530 45 1 1810 500 341
47950 33536 52 1 1810 500 365
```

The researchers used the EMOTHAW dataset to answer the research questions, developed by Likforman-Sulem et al. (2017), which is publicly available. Permission to use the dataset was requested via a Google Form, and it was subsequently shared with the researchers through email. It includes 903 online handwriting time series samples from 129 participants in Italy. The dataset consists of seven handwriting tasks, such as drawing pentagons, writing Italian words, and performing the Clock Drawing Test, all designed for cognitive impairment detection and personality assessment. The data were stored in .svc files, which were used to save information about handwriting samples. To preprocess the data for analysis, pandas was used to open these files and read the information. Each file contained details such as the position of the pen (x and y), the time when each data point was recorded (timestamp), whether the pen was touching the surface (pen status), the pressure applied on the pen (pressure), and the angle of the pen (azimuth and altitude). To make the data easier to work with, the timestamp was adjusted to start from zero, as shown in Figure 19. Then, MinMaxScaler was applied to change the values of the position, time, and other measurements so they fit within a range from 0 to 1. This ensured that all the features of the data were on the same scale, making it easier for the model to learn from the data.

2. Greenland GPS dataset

Figure 20. Plotted GPS Time Series Data from BLAS station using coordinates such as East, North, and Up from 2008 to 2024



To validate the developed model's applicability to other time series dataset, the researchers used the Greenland GPS dataset which was also used by Zhang et al. (2021) in their study. The researchers selected the same 20 stations used by Zhang et al. (2021), namely BLAS, DGJG, DKSG, HJOR,

HMBG, HRDG, JGBL, JWLF, KMJP, KMOR, KUAQ, KULL, LBIB, LEFN, MARG, MSVG, NRSK, QAAR, UTMG, and YMER. Since it is a public dataset, these data were retrieved and downloaded from the website of Nevada Geodetic Laboratory (NGL), at the University of Nevada at Reno. Each station has a time series dataset encoded in ascii text and in two types of reference frame. The IGS14 is part of the International GNSS Service global reference frame, while the NA are localized frames used for North American data. Moreover, the NGL has also offered different processed time series data and the researchers have chosen the 24-hour final solution as it offered daily position of the time series data, as well as the IGS14 reference frame, as it was the reference frame used by Zhang et. al (2021) in their study. The files were in .tenv3 format and it consisted of different properties including the station name, different time and date formats when the GPS was recorded such as the standard date (in YYMMMDD format), decimal year (yyyy.yyyy), modified Julian day (__MJD), , GPS week (week), and the day of GPS week (d). Aside from that, the GPS time series also consisted of a longitude degrees of reference meridian (reflon), and positional components, mainly East, North, and Up. These three composed of an integer (_e0(m), ____n0(m), u0(m)) and fractional (__east(m), _north(m), ____up(m)) parts of the coordinates. As shown in Figure 20, the three coordinates were plotted using their fractional value from the BLAS station. Each component's mean were calculated along with its demeaned position. The latter was integrated to better visualize the plot around the coordinate. The researchers used these three features as the input for the model. To ensure that the coordinates will be closely similar to the original data, an additional pre-processing step was integrated by the researchers. Other features such as uncertainties of the coordinates (sig_e(m), sig_n(m), sig_u(m)), and the correlation between these three (__corr_en, __corr_eu, __corr_nu) were taken into consideration and used to further normalize the data using MinMax scaler, which was helpful as all the values of the features (east, north, and up coordinates) were on the same scale that made it easier for the model to learn.

Statement of the Problem 1: Performance Evaluation of SC-β-VAE-GAN

To evaluate the SC-β-VAE-GAN model's performance in generating synthetic data, the researchers first analyzed the output visually. Figure 21 provides an example of real handwriting time series data from the EMOTHAW dataset, while Figure 22 displays synthetic handwriting time series data generated by the SC-β-VAE-GAN model based on the real data, both visualized using the developed tool.

Figure 21. Example of Original EMOTHAW Time Series Data Visualized as Plots

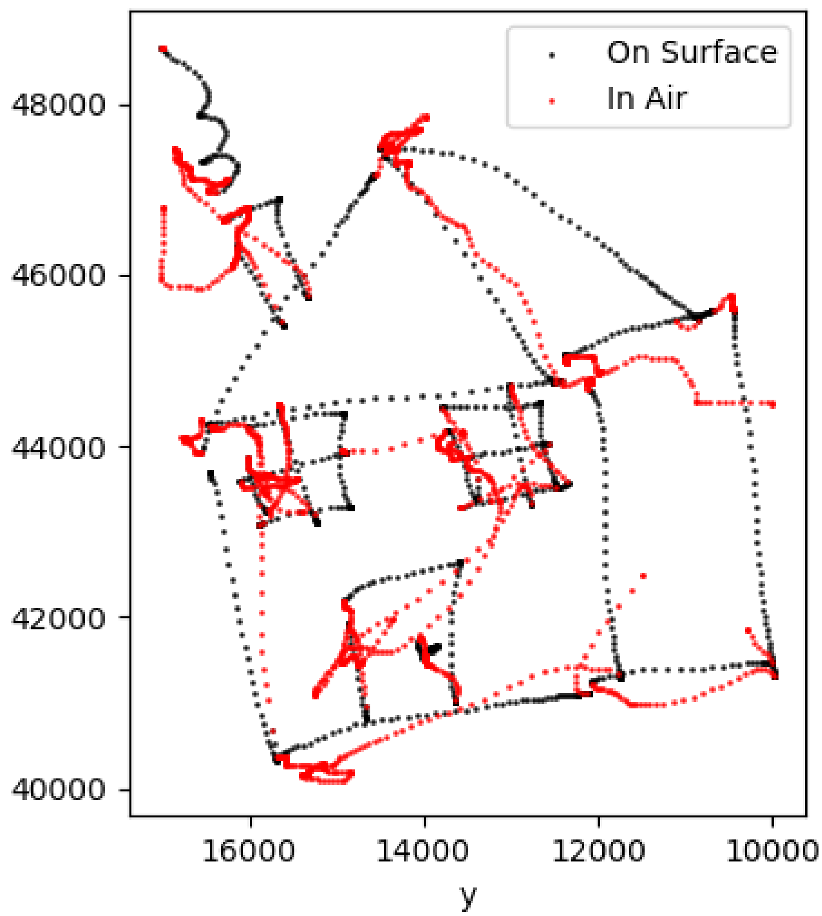


Figure 21 shows a visualization of the original EMOTHAW time series data from Likforman-Sulem et al. (2017) in the form of a plot. The data is separated into two categories: "On Surface" and "In Air," visualized using black and red, respectively. This approach allows for a clear distinction between data captured when the pen or tool is in contact with a surface and when it is lifted into the air.

Figure 22. Example of Synthetic EMOTHAW Time Series Data Generated by SC-β-VAE-GAN Visualized as Handwriting

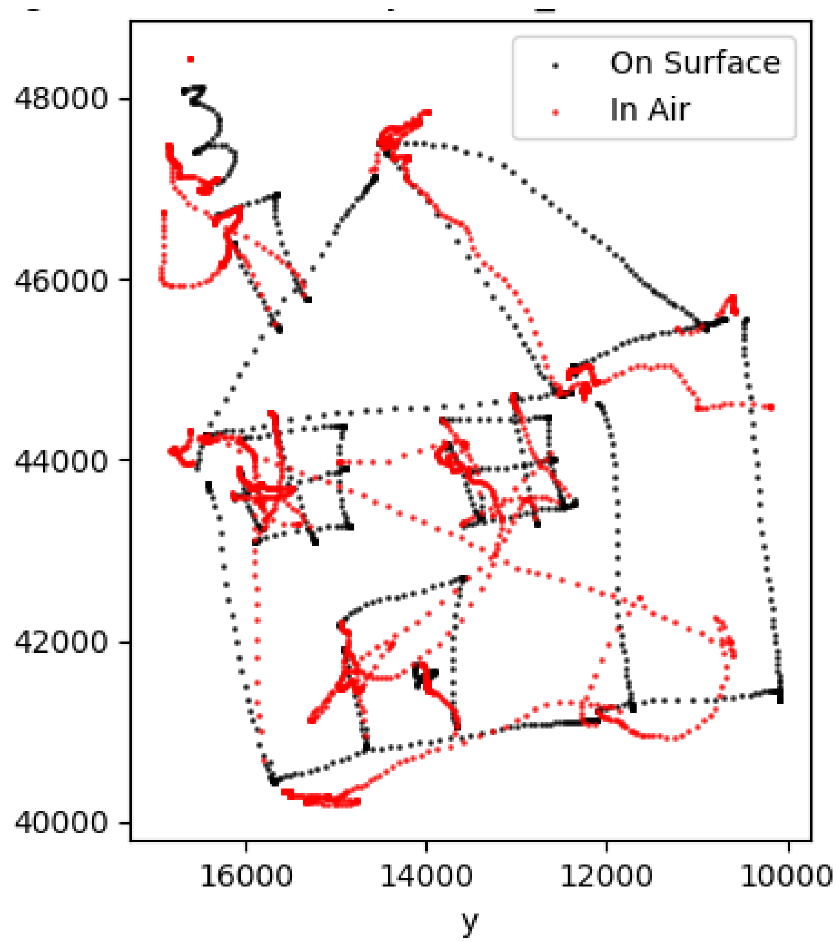


Figure 22 presents a plot illustrating synthetic EMOTHAW time series data produced by SC-β-VAE-GAN. The data is categorized into two groups: "On Surface" and "In Air," represented by black and red, respectively. This visualization emphasizes the difference between moments when the pen or tool interacts with a surface and when it moves through the air.

Table 1

Comparison of Original and Synthetic EMOTHAW Time Series Data (Black for Original Values, Green for Synthetic Values, Red for Missing Values)

	1	2	3	4	5
X1	NaN	NaN	44508	44508	44508
X2	44603	44608	44612	44614	44615
Y1	NaN	NaN	10564	10634	10714
Y2	10211	10265	10383	10500	10627
Time stamp1	NaN	NaN	4260	4267	4275
Time stamp2	4177	4185	4192	4200	4207
Pen status1	NaN	NaN	0	0	0
Pen status2	0	0	0	0	0
Azimuth1	NaN	NaN	2060	2040	2030
Azimuth2	2080	2070	2060	2040	2030
Altitude1	NaN	NaN	460	450	450
Altitude2	460	460	460	450	450
Pressure1	NaN	NaN	0	0	0
Pressure2	0	0	0	0	0

To illustrate the changes made by the model in augmentation and imputation, Table 1 presents a snippet of a side-by-side comparison of the real (Likforman-Sulem et al., 2017), and synthetic EMOTHAW time series data, generated using the developed tool. Each column represents data at a specific time point.

- Rows labeled with 1 correspond to the original data values.
- Rows labeled with 2 represent the synthetic data generated by the tool.
- Missing values are highlighted in red (NaN), indicating gaps in the original data.

- Original data values are displayed in black, while the synthetic data values are shown in green to differentiate them clearly.

As shown in table 1, synthetic data generated by the model was able to augment and impute missing values like in the first two rows, the original data (X1) had missing values (NaN) on its first two columns, and by using the developed model, these values were imputed and augmented (X2), solving the missing gaps in the time series data.

Table 2

Performance of SC-β-VAE-GAN

Model	Normalized Root Mean Square Error	Post-Hoc Discriminative Score	Post-Hoc Predictive Score
SC-β-VAE-GAN	.92%	50.79%	5.35%

Table 2 addresses the first problem statement by showcasing the performance of the SC-β-VAE-GAN model in generating synthetic data based on its Normalized Root Mean Square Error (NRMSE), Post-Hoc Discriminative Score, and Post-Hoc Predictive Score. The model achieved an NRMSE of .92%, demonstrating that the synthetic data generated by the model is very similar to the real data, with only a small discrepancy when normalized across the data range. This indicates that the model effectively captures the underlying characteristics of the handwriting samples, as evident in the synthetic data shown in Figure 21 (Lee et al., 2022). For the Post-Hoc Discriminative Score, the model achieved a value of 50.79%, suggesting that the classifier could not reliably distinguish between real and synthetic data, with accuracy close to random chance (50%) . This implies that the synthetic data is nearly indistinguishable from the real data in terms of its features, demonstrating the model's ability to produce highly realistic samples, as reflected in the similarity between the real and synthetic data, illustrated in Figures 20 and 21 . Lastly, the model achieved a Post-Hoc Predictive Score of 5.35%, indicating that the synthetic data exhibits a low prediction error when used to forecast future time steps, suggesting the model captures temporal patterns in the data accurately. This is demonstrated in the side-by-side comparison of real and synthetic EMOTHAW time series data in Table 1, where the synthetic data effectively imputes missing values ("NaN") and preserves temporal consistency, highlighting its strong predictive performance (Paepae et al., 2023; Boquet et al., 2019).

Statement of the Problem 2: Performance Evaluation of VAEGAN, TimeGAN, and VRNNGAN

Evaluation of the performance of the baseline models – VAEGAN, TimeGAN, and VRNNGAN in generating synthetic data based on its Normalized Root Mean Square Error (NRMSE), Post-Hoc Discriminative Score, and Post-Hoc Predictive Score.

Table 3

Result of the VAEGAN , TimeGAN and VRNNGAN

Performance of VAEGAN, TimeGAN and VRNNGAN			
Model	**Normalized Root Mean Square Error**	**Post-Hoc Discriminative Score**	**Post-Hoc Predictive Score**
VAEGAN	4.11%	54.01%	6.11%
TimeGAN	90.76%	99.70%	23.67%
VRNNGAN	3.76%	53.35%	7.12%

Table 3 presents the performance of baseline models—VAEGAN, TimeGAN, and VRNNGAN—in generating synthetic data, evaluated using Normalized Root Mean Square Error (NRMSE), Post-Hoc Discriminative Score, and Post-Hoc Predictive Score.

Both the VAEGAN and VRNNGAN models demonstrated low NRMSE values of 4.11% and 3.76%, respectively, indicating that their generated synthetic data closely resembles the real data. This suggests that these models effectively capture the underlying patterns of the original dataset with minimal discrepancies. In contrast, TimeGAN exhibited a significantly higher NRMSE of 90.76%, indicating that the synthetic data generated by this model differs substantially from the real data. This highlights its poor ability to replicate the characteristics of the original dataset (Paepae et al., 2023; Boquet et al., 2019).

In terms of Post-Hoc Discriminative Scores, VAEGAN and VRNNGAN achieved scores close to 50%, specifically 54.01% and 53.35%, respectively. These scores indicate that the synthetic data generated by these models are nearly indistinguishable from the real data, as the classifier’s accuracy is close to random guessing. This demonstrates the strong realism of the data produced by these models. However, TimeGAN recorded a much higher discriminative score of 99.70%, meaning that the classifier can easily distinguish between real and synthetic data. This implies that the synthetic data generated by this model lacks realism and does not convincingly mimic the features of real-world data (Lee et al., 2022).

For the Post-Hoc Predictive Score, VAEGAN achieved a score of 6.11%, while VRNNGAN closely followed with 7.12%. These low values indicate that the synthetic data from these models effectively captures temporal patterns and maintains predictive consistency, meaning it performs well in forecasting tasks that can be used for imputation. In contrast, TimeGAN recorded a significantly higher predictive score of 23.67%, reflecting poor predictive performance. The synthetic data from this model fails to preserve the temporal dependencies and patterns of the original dataset (Lee et al., 2022).

Statement of the Problem 3: Performance Comparison of SC-β-VAE-GAN with VAEGAN, TimeGAN, and VRNNGAN

The section presents the evaluation of the developed model, SC-β-VAE-GAN model against three existing augmentation methods: VAEGAN, TimeGAN, and VRNNGAN. The analysis lean towards assessing their relative performance in generating synthetic handwriting multivariate time series data. The evaluation metrics included Normalized Root Mean Square Error (NRMSE), Post-Hoc Discriminative Score, and Post-Hoc Predictive Score, which collectively provided insights into the quality, realism, and predictive utility of the generated data. By comparing these metrics , the study aimed to identify whether SC-β-VAE-GAN offered a significant improvement over the baseline models and to substantiate its effectiveness in imputation and augmentation tasks.

Table 4

Evaluation of Normalized Root Mean Square Error

Normalized Root Mean Square Error		
Model	EMOTHAW Dataset	
	Mean	**Std.**
VAEGAN	4.11%	3.02%
TimeGAN	90.76%	86.33%
VRNNGAN	3.76%	2.91%
SC-β-VAE-GAN	**.92%**	**1.23%**

Table 4 below compares the normalized root mean square error for the SC-β-VAE-GAN model against other models (VAEGAN, TimeGAN, and VRNNGAN) on the EMOTHAW dataset. NRMSE is used to evaluate the reconstruction error of the generated synthetic handwriting time series data. Lower values indicate better performance. The SC-β-VAE-GAN model demonstrates the lowest mean NRMSE (0.92%) and a relatively low standard deviation (1.23%), suggesting it outperforms the competing models in terms of reconstruction accuracy and stability (Paepae et al., 2023; Boquet et al., 2019).

Table 5

Evaluation of Post Hoc Discriminative Score

Discriminative Score		
Model	EMOTHAW Dataset	
	Mean	**Std.**
VAEGAN	54.01%	.76%
TimeGAN	99.70%	3.99%
VRNNGAN	53.35%	.59%
SC-β-VAE-GAN	**50.79%**	**.47%**

Table 5 displays the post-hoc discriminative score of the developed model and baseline models, the metric measures the realism of the generated synthetic handwriting data by evaluating how well it can be distinguished from real data using a classifier. Score nearest to 50% signifies higher realism wherein the SC-β-VAE-GAN model achieves the mean discriminative score (50.79%) nearest to 50% which also has a standard deviation (0.47%), indicating superior performance in generating realistic and indistinguishable synthetic data compared to the baseline models (Lee et al., 2022).

Table 6

Evaluation of Post Hoc Predictive Score

Predictive MAPE Score		
Model	EMOTHAW Dataset	
	Mean	**Std.**
VAEGAN	6.11%	3.52%
TimeGAN	23.67%	4.11%
VRNNGAN	7.12%,	3.52%
SC-β-VAE-GAN	**5.35%**	**3.37%**

Table 6 provides a comparison of the predictive Mean Absolute Percentage Error (MAPE) score for the SC-β-VAE-GAN model and the baseline models. The Predictive MAPE evaluates the utility of the generated synthetic handwriting time series data in downstream predictive timestamps, with lower scores (below 5%) indicating better predictive performance. The SC-β-VAE-GAN model achieves a mean score of 5.35% with a standard deviation of 3.37%, coming closest to the optimal performance range, and outperforms most baseline models except for VAEGAN (6.11%). This result underscores its efficacy in generating data that supports accurate predictions (Lee et al., 2022).

To address SOP 3 and test the associated hypotheses, Kruskal-Wallis analysis followed by Dwass-Steel-Critchlow-Fligner pairwise comparisons were conducted. These analyses were performed to determine if there were significant differences between SC-β-VAE-GAN and three other generative models (VAE-GAN, TimeGAN, and VRNNGAN) in terms of synthetic data generation. The performance was evaluated using NRMSE, Post-Hoc Predictive Scores, and Post-Hoc Discriminative Scores. The results are summarized in Table 7 and the accompanying pairwise comparisons.

Table 7

Kruskal Wallis Hypothesis Testing

	x^2	df	p	ε^2
NRMSE	**2562**	**3**	**<.001**	**0.709**
PREDICTIVE	**2074**	**3**	**<.001**	**0.574**
DISCRIMINATIVE	**3611**	**3**	**<.001**	**1.000**

Table 7 shows the results of the **Kruskal-Wallis Hypothesis Testing**, which was employed to evaluate whether significant differences exist among models across three performance metrics: NRMSE, Predictive Scores, and Discriminative Scores. For NRMSE, the test statistic (x^2 = 2562) with 3 degrees of freedom and a computed p < 0.001 demonstrates significant differences among the models. Since the computed p-value is less than the significance level of 0.05 indicating that at least one model exhibits significantly different NRMSE performance. Similarly, for Predictive Scores, the test statistic (x^2 =2074) and a computed p < 0.001 lead to rejection of (H0), confirming significant differences among the models in predictive performance. Lastly, for Discriminative Scores, the test statistic (x^2 = 3611) and p<0.001, highlighting significant differences in discriminative ability across the models. This table confirms significant differences in the quality of synthetic data generated by various baseline models and the developed model. It validates that the superior performance of the developed model across the metrics—post-hoc discriminative score (assessing class separability), post-hoc predictive score (measuring predictive accuracy), and NRMSE (evaluating reconstruction error)—is both statistically significant and robust (Lee et al., 2022; Paepae et al., 2023; Boquet et al., 2019).

Succeeding table presents the result of the Dwass-Steel-Critchlow-Fligner pairwise comparisons to identify specific differences between SC-β-VAE-GAN and the other models.

Table 8

SC-β-VAE-GAN and VAEGAN pairwise comparisons

SC-β-VAE-GAN and VAEGAN			
	W	**p**	**Conclusion**
NRMSE	**40.148**	**<.001**	**Reject H_0**
PREDICTIVE	**6.86**	**<.001**	
DISCRIMINATIVE	**60.01**	**<.001**	

Table 8 reveals the comparison between SC-β-VAE-GAN and VAE-GAN which shows significant differences in all metrics. For NRMSE, the W-value of 40.148 and $p<0.001$ indicate significant difference, since $p < 0.05$ which demonstrates SC-β-VAE-GAN's superior accuracy. Similarly, for Predictive Scores, the W-value of 6.86 and $p<0.001$ confirm significant differences and validate SC-β-VAE-GAN's better predictive performance. For Discriminative Scores, the W-value of 60.01 and $p < 0.001$, confirming SC-β-VAE-GAN's superior discriminative capability. These findings reject the null hypothesis and confirm that SC-β-VAE-GAN outperforms VAE-GAN in generating synthetic data with superior quality across all key metrics: NRMSE (reconstruction accuracy), post-hoc predictive score (predictive reliability), and post-hoc discriminative score (class separability) (Lee et al., 2022; Paepae et al., 2023; Boquet et al., 2019).

Table 9

SC-β-VAE-GAN and TIMEGAN pairwise comparisons

SC-β-VAE-GAN and TIMEGAN			
	W	**p**	**Conclusion**
NRMSE	**51.572**	**<.001**	**Reject H_0**
PREDICTIVE	**51.91**	**<.001**	
DISCRIMINATIVE	**60.1**	**<.001**	

The table 9 compares SC-β-VAE-GAN with TimeGAN, where significant differences are observed also in all metrics. For NRMSE, a W-value of 51.572 and p<0.001 indicating SC-β-VAE-GAN's a significant difference with accuracy metric. For Predictive Scores, the W-value of 51.91 and p < 0.001 confirming SC-β-VAE-GAN's significant advantage in predictive performance and for Discriminative Scores with a W-value of 60.1 and p<0.001 highlighting SC-β-VAE-GAN's stronger discriminative ability compared to TimeGAN. These findings reject the null hypothesis and proves that SC-β-VAE-GAN significantly outperforms TimeGAN across all metrics, including reconstruction accuracy (NRMSE), predictive reliability, and discriminative capability, in generating synthetic data (Lee et al., 2022; Paepae et al., 2023; Boquet et al., 2019).

Table 10

SC-β-VAE-GAN and VRNNGAN pairwise comparisons

SC-β-VAE-GAN and VRNNGAN			
	W	**p**	**Conclusion**
NRMSE	**42.787**	**<.001**	**Reject H_0**
PREDICTIVE	**13.51**	**<.001**	
DISCRIMINATIVE	**60.1**	**<.001**	

Table 10 shows the third pairwise comparison of SC-β-VAE-GAN against VRNNGAN and identifies significant differences across all metrics. For NRMSE, a W-value of 42.787 with $p < 0.001$ indicates a notable variation in results between the two models. And for Predictive Scores, a W-value of 13.51 with $p<0.001$ also reflects a significant difference in the predictive outcomes. Then lastly, the Discriminative Scores comparison, with a W-value of 60.1 and $p < 0.001$, highlights a substantial difference in discriminative results between SC-β-VAE-GAN and VRNNGAN. These results also imply that the null hypothesis must be rejected, which signifies that SC-β-VAE-GAN performs significantly better than VRNNGAN across all evaluated metrics in generating synthetic data specifically has better reconstruction accuracy, predictive reliability, and discriminative ability (Lee et al., 2022, p. 59; Paepae et al., 2023; Boquet et al., 2019).

The tables above show that upon performing the Kruskal-Wallis test followed by pairwise comparisons, the null hypothesis that there are no significant differences between the performance of SC-β-VAE-GAN and the other models (VAE-GAN, TimeGAN, and VRNNGAN) in terms of NRMSE, Post-Hoc Discriminative Score, and Post-Hoc Predictive Score are all rejected. The results revealed significant differences in all metrics, with p-values less than 0.001, confirming that SC-β-VAE-GAN differs significantly from VAE-GAN (Ham et al., 2020), TimeGAN (Yoon et al., 2019), and VRNNGAN (Lee et al., 2022) in generating synthetic data.

Cross-Dataset Validation of SC-β-VAE-GAN

Table 11

Performance of SC-β-VAEGAN using GPS Time Series Dataset

Model	Normalized Root Mean Square Error	Post-Hoc Discriminative Score	Post-Hoc Predictive Score
SC-β-VAE-GAN	1.15%	51.71%	0.41%

To test the generalizability of the SC-β-VAE-GAN model, its performance was evaluated on a different time-series dataset, specifically GPS time series data. This cross-dataset validation aimed to determine whether the model's effectiveness extends beyond handwriting datasets, highlighting its potential applicability in other domains.

Table 11 presents the performance metrics of SC-β-VAE-GAN when applied to the GPS time series dataset. The results include the Normalized Root Mean Square Error (NRMSE), Post-Hoc Discriminative Score, and Post-Hoc Predictive Score, which measure the model's ability to generate realistic and temporally consistent synthetic data. An NRMSE of 1.15% indicates high reconstruction accuracy (Paepae et al., 2023; Boquet et al., 2019), while a Post-Hoc Discriminative Score of 51.71% suggests that the synthetic data generated closely resembles real data, as it remains nearly indistinguishable by classifiers (Lee et al., 2022). Additionally, the Post-Hoc Predictive Score of 0.41% demonstrates the model's good temporal prediction capabilities in this new context (Lee et al., 2022). This dataset is known to have missing values, specifically its spatial properties such as the coordinates (east, north, and up) which could greatly impact geodetic research (Zhang et al., 2021). These findings reinforce the SC-β-VAE-GAN's adaptability to different time-series domains.

Comparison of Models Trained on SC-β-VAE-GAN Synthetic and Original Data

The section presents a comparison of classification performance between a model trained on a combination of SC-β-VAE-GAN synthetic data and original data, and another model trained exclusively on original data. The analysis evaluates the impact of incorporating synthetic data generated by SC-β-VAE-GAN on model performance, based on metrics such as precision, recall, F1-score, and support, using online handwriting multivariate time series data.

The EMOTHAW dataset from Likforman-Sulem et al. (2017) was used for this experiment, as it is the most recent and publicly available online handwriting time series dataset. The classifier employed is an emotion recognition system based on an attention-based transformer model from Khan et al. (2024), which, to the researchers' knowledge, as of this writing, is the latest and best-performing model for classifying emotions using handwriting time series data, achieving an accuracy of 92.64%. The emotions included in the dataset are normal, depression, anxiety, and stress. For the experiment, two models were trained, first with only the original EMOTHAW dataset and the second with a combination of the original EMOTHAW data and synthetic data generated by SC-β-VAE-GAN. The dataset was labeled into four categories—normal, depression, anxiety, and stress—based on the DASS (Depression, Anxiety, and Stress Scale) scores provided in the accompanying Excel file, which contained the scores for each user. For the dataset division, the researchers followed the 80/20 rule, using 80% of the data for training and 20% for testing. In the first model, 721 samples from the original EMOTHAW dataset were used for training, and 182 samples were used for testing. In the second model, a total of 1442 samples were used for training, consisting of 721 original samples augmented with 721 synthetic samples. The testing set remained the same as in the first model, with 182 samples.

Figure 23. Precision Comparison of Models Trained on Original and SC-β-VAE-GAN Synthetic Data

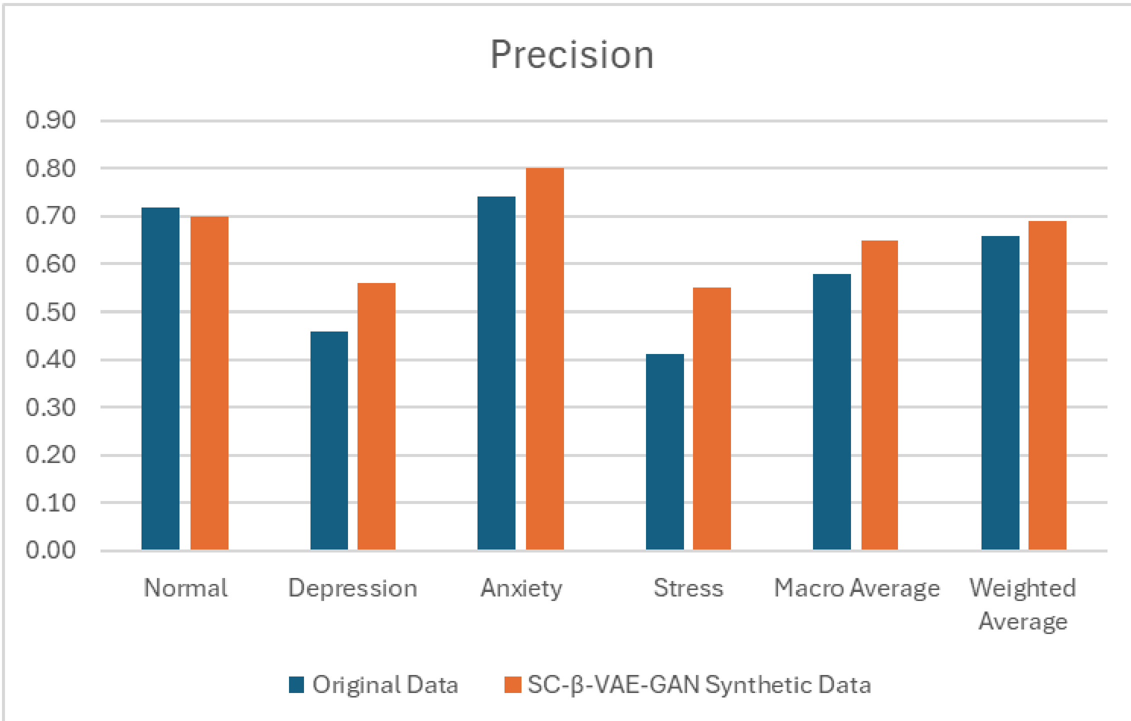


Figure 23 presents the comparison of precision scores between models trained on the original EMOTHAW dataset and those trained on a combination of the original data and synthetic data generated by SC-β-VAE-GAN. Precision measures the accuracy of positive predictions, indicating how many of the predicted positive instances are actually correct. The blue bars represent the scores for the models trained on the original data, while the orange bars indicate the scores for the models trained on the synthetic data. The precision for the Normal class is slightly higher for the original data (0.72) compared to synthetic data (0.70). For the Depression class, the synthetic data model shows improved precision (0.56 vs. 0.46). Similarly, in the Anxiety class, the synthetic data model performs better (0.80 vs. 0.74), as well as in the Stress class (0.55 vs. 0.41). The macro average precision is higher for the synthetic data model (0.65 vs. 0.58), and the weighted average precision also improves (0.69 vs. 0.66), suggesting that synthetic data enhances overall model performance across all classes (Niu et al., 2020).

Figure 24. Recall Comparison of Models Trained on Original and SC-β-VAE-GAN Synthetic Data

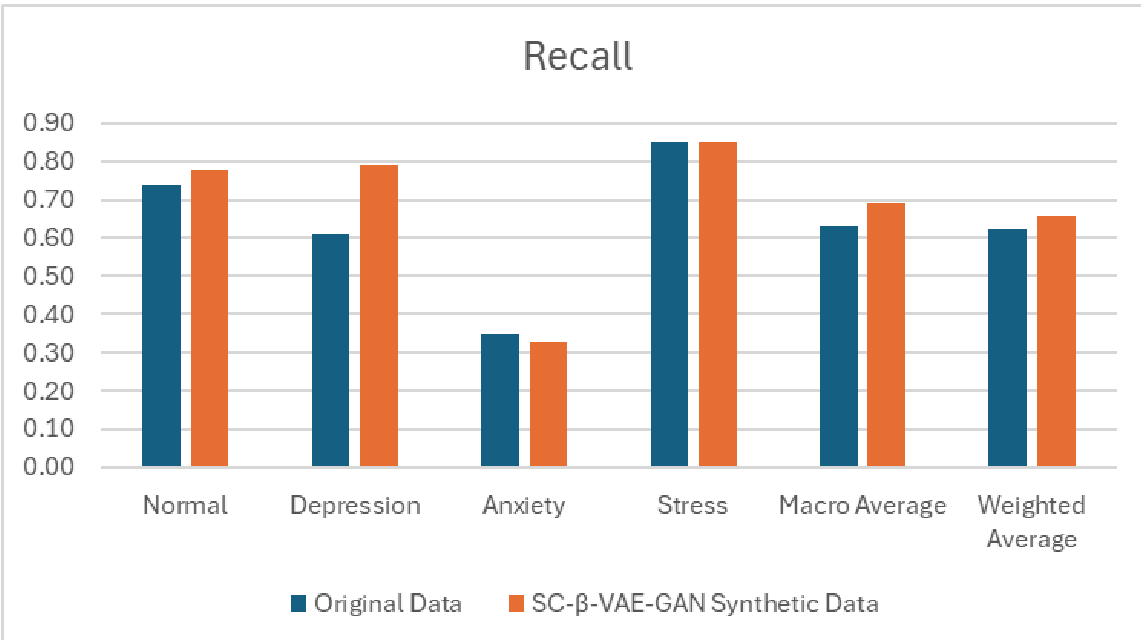


Figure 24 presents the recall results comparing the models trained on the original data and the SC-β-VAE-GAN synthetic data. Recall measures how many actual positive instances were correctly identified by the model. For the "Normal" class, the synthetic data model achieves a higher recall (0.78 vs. 0.74). In the "Depression" class, the synthetic data model significantly improves recall (0.79 vs. 0.61). However, for the "Anxiety" class, recall slightly decreases with synthetic data (0.33 vs. 0.35), and recall remains the same for the "Stress" class (0.85). The macro average recall is higher for the synthetic data model (0.69 vs. 0.63), and the weighted average recall also improves (0.66 vs. 0.62).. This suggests that the synthetic data model performs slightly better when considering the distribution of different classes (Niu et al., 2020).

Figure 25. F1-Score Comparison of Models Trained on Original and SC-β-VAE-GAN Synthetic Data

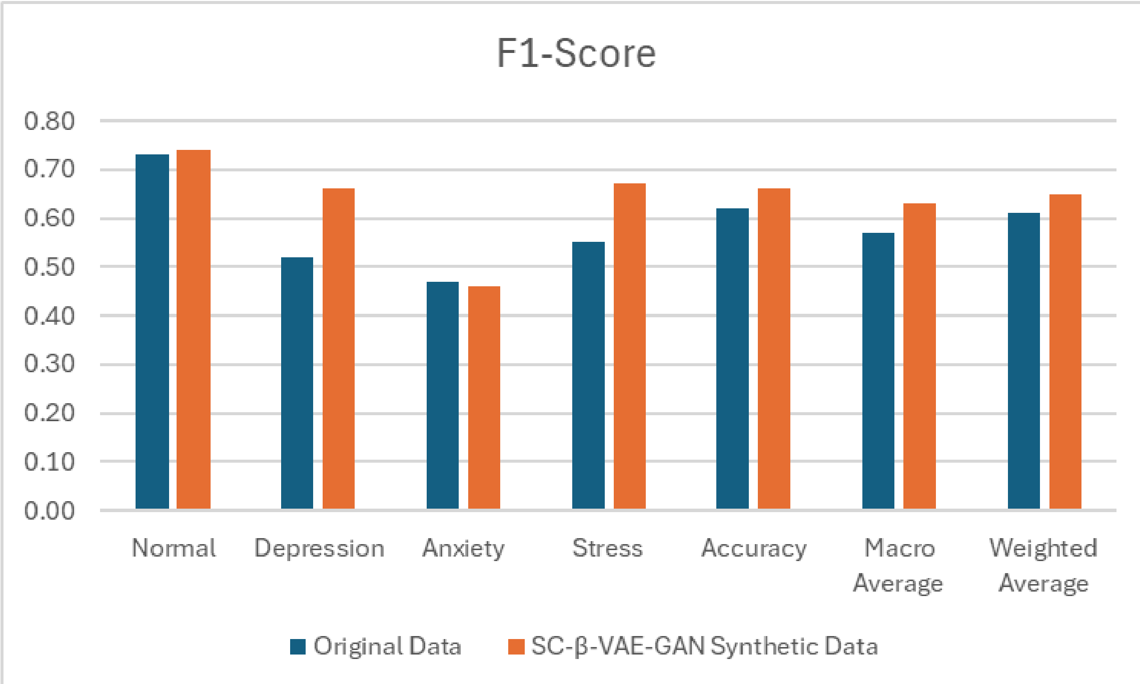


Figure 25 presents the F1 scores comparing the models trained on the original EMOTHAW dataset and the SC-β-VAE-GAN synthetic data. The F1 score balances precision and recall, with a higher value indicating better performance in correctly identifying both positive and negative instances (Niu et al., 2020). For the "Normal" class, the synthetic data model slightly outperforms the original (0.74 vs. 0.73). The "Depression" class shows a significant improvement with synthetic data (0.66 vs. 0.52). The "Anxiety" class shows little change (0.46 vs. 0.47), while the "Stress" class improves with synthetic data (0.67 vs. 0.55). The macro average F1 score increases from 0.57 to 0.63, and the weighted average F1 score improves from 0.61 to 0.65, demonstrating the positive impact of synthetic data on overall model performance.

Figure 26. Support Comparison of Models Trained on Original and SC-β-VAE-GAN Synthetic Data

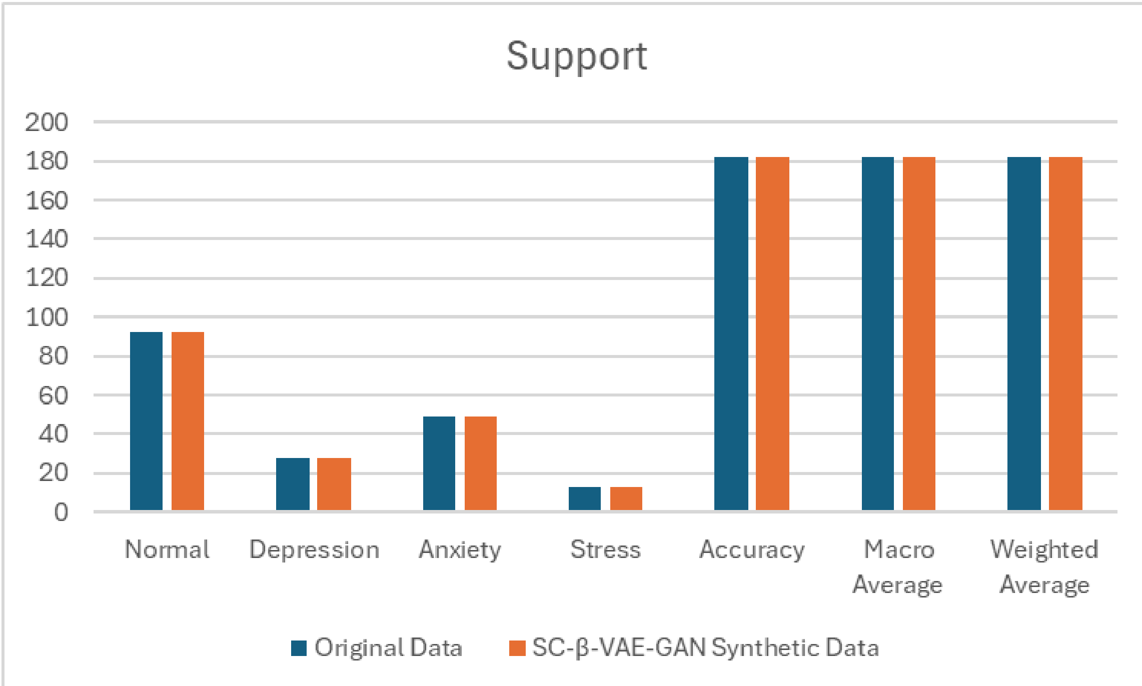


Figure 26 presents the support comparing the models trained on the original EMOTHAW dataset and the SC-β-VAE-GAN synthetic data. Support values give insight into the distribution of instances for each class in the dataset. For Normal class Both the original data and the SC-β-VAE-GAN synthetic data have 92 instances of the "Normal" class. For Depression class There are 28 instances of the "Depression" class in both the original and synthetic datasets. For The "Anxiety" class has 49 instances in both datasets. For Stress The "Stress" class has 13 instances in both datasets. Accuracy, Macro Average, and Weighted Average: These values represent the total number of instances across all classes (182 in this case), and since there is no change in the class distribution between the original and synthetic data, the support values are identical for both datasets.

Figure 27. Confusion Matrix of Model Trained on Original Data

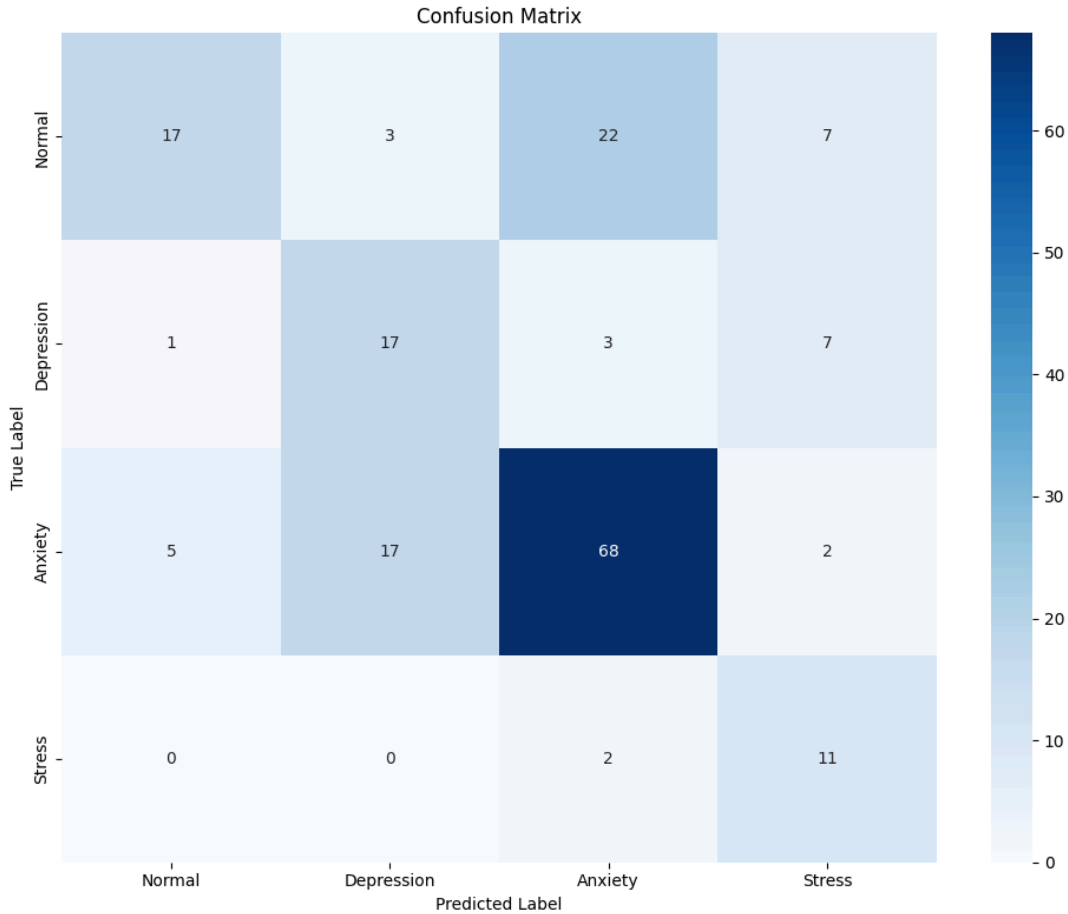


Figure 27 shows the confusion matrix for the model trained on the original data, revealing that the model performs well in predicting the "Anxiety" class, with 68 correct predictions, but struggles with distinguishing between "Normal" and "Anxiety," as 22 instances of "Normal" were misclassified as "Anxiety." It also confuses "Depression" with "Anxiety," as 17 "Anxiety" instances were predicted as "Depression." The model is better at predicting "Stress," with 11 correct predictions, though it still misclassifies instances of other classes as "Stress" (7 instances from both "Normal" and "Depression"). Additionally, there are misclassifications between "Normal" and "Depression" (3 instances), and "Normal" and "Anxiety" (5 instances), highlighting the model's difficulty in distinguishing between these classes. Overall, the model shows some degree of misclassification, particularly between "Normal," "Anxiety," and "Depression" (Wen, 2021b).

Figure 28. Confusion Matrix of Model Trained on SC-β-VAE-GAN Synthetic Data

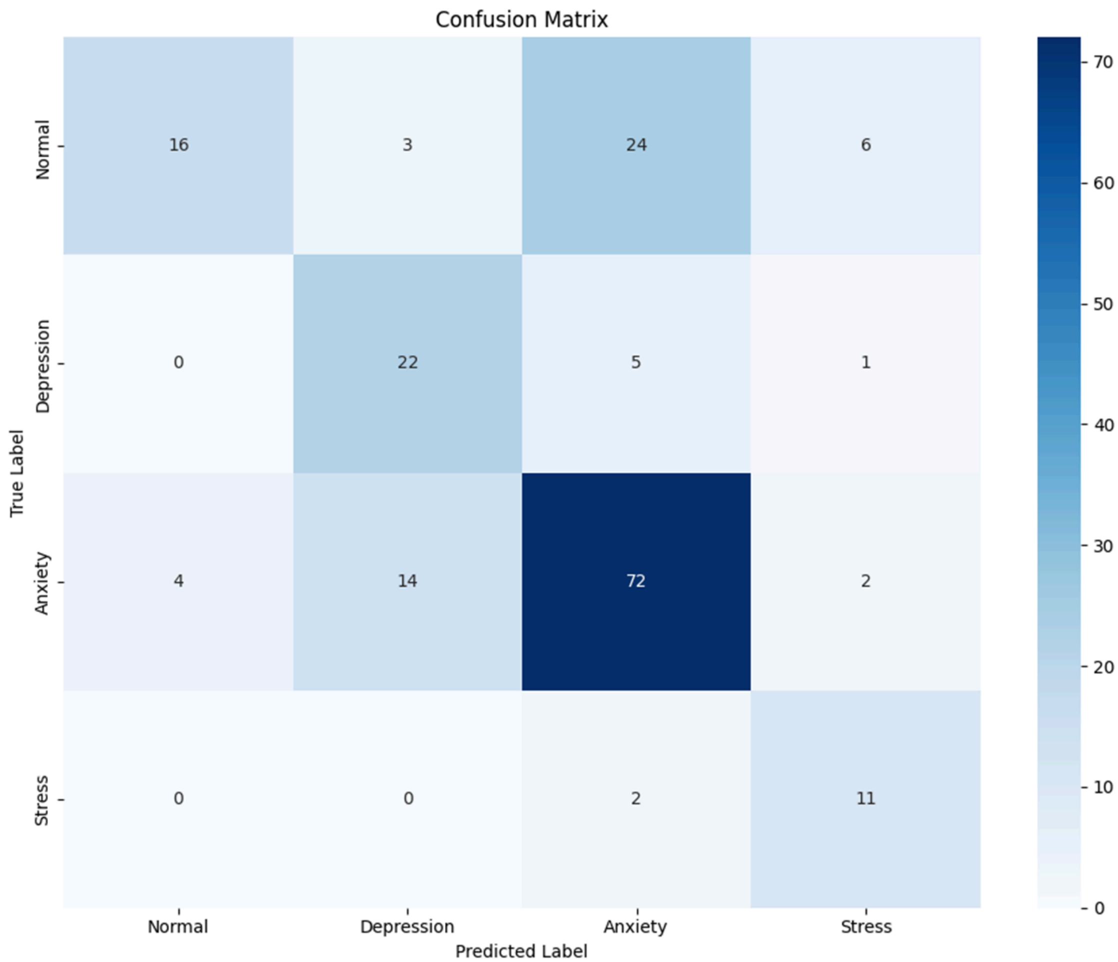


Figure 28 shows the confusion matrix for the model trained on SC-β-VAE-GAN Synthetic Data, demonstrating strong performance in predicting "Depression" and "Anxiety," with 22 correct predictions for "Depression" and 72 for "Anxiety." However, the model struggles to distinguish "Normal" from "Anxiety," misclassifying 24 "Normal" instances as "Anxiety," and has difficulty distinguishing "Normal" from "Stress," with 6 "Normal" instances misclassified as "Stress." While the model improves in the

"Depression" and "Anxiety" classes, challenges remain in accurately classifying certain instances, particularly between "Normal" and "Anxiety" (Wen, 2021b).

Chapter 5

SUMMARY OF FINDINGS, CONCLUSIONS, AND RECOMMENDATIONS

The chapter discusses the findings and includes a summary of the results, recommendations for future research, and conclusions. The study addresses the challenges of limited and missing data in handwriting time series, which can compromise the quality of research outcomes. To address these issues, the SC-β-VAE-GAN model was developed to generate synthetic handwriting data through imputation and augmentation. The research evaluates the model's performance using metrics such as Normalized Root Mean Square Error (NRMSE), Post-Hoc Discriminative Score, and Post-Hoc Predictive Score, comparing its effectiveness to baseline models including VAEGAN, TimeGAN, and VRNNGAN, while also analyzing significant differences among them. The EMOTHAW dataset, consisting of 903 samples from 129 participants completing various handwriting tasks, served as the foundation for this analysis.

Summary of Findings

The main objective of this study is to develop SC-β-VAE-GAN, a shift correction VAE-GAN model, for generating synthetic data to support imputation and augmentation. The following are the key research findings:

1. The performance of the SC-β-VAE-GAN model demonstrated its ability to generate synthetic data that closely resembled real handwriting time series data, effectively augmenting and imputing missing values while preserving temporal patterns. The model achieved a Normalized Root Mean Square Error (NRMSE) of 0.92%, indicating minimal discrepancy

between the synthetic and real data. With a Post-Hoc Discriminative Score of 50.79%, the classifier was unable to distinguish between the real and synthetic data, highlighting the model's ability to generate realistic data, as a classifier with random chance accuracy would also score around 50%. Furthermore, the Post-Hoc Predictive Score of 5.35% indicated low prediction error, highlighting the model's capability to capture temporal patterns and accurately forecast future time steps. These results confirmed the SC-β-VAE-GAN's effectiveness in generating high-quality synthetic data for augmentation and imputation.

2. The performance of VAEGAN and VRNNGAN demonstrated low NRMSE values of 4.11% and 3.76%, respectively, indicating that their synthetic data closely mirrored the real data. These models also achieved Post-Hoc Discriminative Scores of around 50%, showing that the generated data was nearly indistinguishable from the real data. Their Post-Hoc Predictive Scores (6.11% for VAEGAN and 7.12% for VRNNGAN) suggested strong temporal consistency and predictive accuracy. In contrast, TimeGAN performed poorly across all metrics, with a high NRMSE of 90.76%, a Post-Hoc Discriminative Score of 99.70%, and a Post-Hoc Predictive Score of 23.67%, indicating that its synthetic data failed to replicate the real data's characteristics or temporal patterns effectively. These findings highlight that while VAEGAN and VRNNGAN produced realistic and reliable synthetic data, TimeGAN struggled to generate realistic data and maintain temporal consistency.
3. The performance comparison between the SC-β-VAE-GAN model and three other generative models—VAEGAN, TimeGAN, and VRNNGAN—demonstrated that SC-β-VAE-GAN significantly

outperformed the baseline models across key metrics. It achieved the lowest Normalized Root Mean Square Error (NRMSE) of 0.92%, indicating superior reconstruction accuracy. In terms of Post-Hoc Discriminative Score, SC-β-VAE-GAN scored 50.79%, closest to the ideal value of 50%, suggesting that its synthetic data was most indistinguishable from real data. Additionally, SC-β-VAE-GAN performed well on the Post-Hoc Predictive Score, with a mean of 5.35%, outshining most models except VAEGAN. Statistical tests, including Kruskal-Wallis analysis and pairwise comparisons, confirmed significant differences in performance, with p-values less than 0.001, leading to the rejection of the null hypothesis. These results collectively indicate that SC-β-VAE-GAN offers substantial improvements over VAEGAN, TimeGAN, and VRNNGAN in terms of data reconstruction, realism, and prediction, making it particularly effective for augmentation and imputation tasks.

4. To further evaluate the model's performance, SC-β-VAE-GAN was tested on GPS data, demonstrating its effectiveness beyond handwriting data. The model achieved an NRMSE of 1.15%, indicating highly accurate reconstruction of the data. A Post-Hoc Discriminative Score of 51.71% showed that the synthetic data was nearly identical to real data, while a Post-Hoc Predictive Score of 0.41% highlighted its strong ability to predict future patterns. These results confirm that SC-β-VAE-GAN can adapt to different types of time series data, showcasing its flexibility and reliability in diverse applications.
5. The performance of models trained on the original EMOTHAW dataset and those trained with a combination of original data and synthetic data generated by SC-β-VAE-GAN showed that the models trained with

synthetic data outperformed those trained on the original data across most metrics. Precision, recall, F1-score, and support values all indicated improvements when synthetic data was incorporated, particularly in identifying depression and stress. The synthetic data model achieved better performance in terms of overall precision (0.65 vs. 0.58) and recall (0.66 vs. 0.62), with notable improvements in depression (precision: 0.56 vs. 0.46) and stress (precision: 0.55 vs. 0.41). However, there was little impact on anxiety recall, and the models showed some misclassification challenges, particularly between normal and anxiety classes. The synthetic data model also improved F1 scores for depression (0.66 vs. 0.52) and stress (0.67 vs. 0.55), while maintaining similar accuracy to the original model. Overall, adding synthetic data improved the classification system, resulting in more balanced performance and better generalization across all emotional categories.

Conclusion

Upon a thorough analysis of the results, this study demonstrated that the SC-β-VAE-GAN model effectively addresses critical challenges in handwriting time-series analysis by providing a solution for data augmentation and imputation. The experimentation involved testing under standardized conditions with 100 epochs for all models, including the baseline VAEGAN, VRNNGAN, and TIMEGAN, where SC-β-VAE-GAN consistently outperformed its counterparts while maintaining an optimal balance between quality and computational efficiency (Mumuni & Mumuni, 2022).

1. SC-β-VAE-GAN performs well in generating high-quality synthetic data. The model minimizes reconstruction errors, produces realistic outputs, and supports predictive tasks effectively, all of which are necessary for

augmentation and imputation tasks. This is similar to the findings of Li et al. (2021), who discovered that the inclusion of shift correction in VAE ensures smooth data continuity, reduces inconsistencies, and maintains representativeness. These benefits are also evident when applied to a VAE-GAN model. Additionally, the use of disentangled representation learning enables the model to focus on key features of handwriting time-series data, resulting in accurate and efficient data generation (Wang et al., 2022).

2. VAEGAN and VRNNGAN perform moderately in generating synthetic data, while TIMEGAN underperforms in this task. VAEGAN and VRNNGAN generate realistic and temporally consistent data due to their combination of Variational Autoencoders (VAE) and Generative Adversarial Networks (GAN) (Lee, 2022). In contrast, TIMEGAN, which lacks a VAE, struggles to produce authentic and temporally coherent data, possibly due to its limited 100 training epochs, which hinder the optimization of its temporal embedding mechanisms. Prior research by Yoon et al. (2019) has shown that TIMEGAN's performance improves substantially when trained for around 1000 epochs, enabling it to fully exploit its generative potential. However, longer training periods are computationally intensive, which is why only a 100-epoch benchmark was adopted to ensure fair and consistent comparisons across models, as highlighted by Mumuni & Mumuni (2022).
3. The SC-β-VAE-GAN model performs significantly better than baseline models such as TimeGAN, VRNNGAN, and VAEGAN in generating synthetic data. Its design, which includes shift correction and imputation, helps it effectively handle variability in handwriting data (Li et al., 2021).

While baseline models face challenges such as sensitivity to hyperparameters, slow processing, and difficulties in managing data distribution shifts, SC-β-VAE-GAN overcomes these issues to produce more accurate and realistic synthetic outputs (Lee, J., 2022; Yoon et al., 2019; Li et al., 2021). Despite its success, SC-β-VAE-GAN is limited in computational efficiency, which could pose a bottleneck for large datasets. The fixed 100-epoch training constraint, though ensuring fairness, may not fully optimize performance in all cases, impacting the scalability and outcomes of both SC-β-VAE-GAN and other baseline models (Mumuni & Mumuni, 2022).

4. SC-β-VAE-GAN enhances performance on GPS time series data. Due to the strength of incorporating shift correction and β-VAE enhancements, the model can adapt to non-random missing data and maintain data continuity, addressing challenges of missing data and complex temporal patterns in this dataset. This sets it apart from benchmark models like MissForest, which struggle with distribution shifts caused by concentrated missing data (Zhang et al., 2021). The model is effective for real-world applications, making it a promising solution for handling intricate and incomplete datasets (Lee et al., 2022; Li et al., 2021).
5. Incorporating synthetic data generated by SC-β-VAE-GAN into emotion classification models can significantly improve their performance. This is because adding the synthetic data generated by SC-β-VAE-GAN in the training model addresses issues of low dataset size and missing values (Pourshahrokhi et al., 2022). Specifically, the classifier performs well in detecting depression and stress. However, the impact is limited for anxiety due to class imbalance in the used dataset (EMOTHAW), which

led to misclassification challenges between dominant classes like normal and anxiety (Nolazco-Flores et al., 2021; 2022).

This research sets a foundation for future studies to explore the application of SC-β-VAE-GAN in other time-series domains, such as GPS and healthcare data, and to extend its capabilities to multimodal integration, including speech and physiological signals. SC-β-VAE-GAN proved transformative in addressing data limitations in handwriting analysis, offering solutions for data augmentation and imputation while paving the way for advancements in generative modeling and handwriting-based emotion detection systems.

Recommendations

The SC-β-VAE-GAN model has demonstrated significant potential for enhancing data augmentation and imputation in time series applications. To further improve its performance and expand its applications in this field, the following recommendations are proposed:

1. The researchers suggest comparing the SC-β-VAE-GAN model with newer models and techniques, especially since one of the baseline models, TimeGAN, didn't perform well on the standard tests. This comparison should include models developed around the same time as this research, like the Data Augmentation with Analytic Wavelets (DAAW) by Kulevome (2024), as well as other models in development. Doing so will help ensure the performance benchmarks are up-to-date and include new ideas from the latest advancements.
2. The researchers propose exploring hybrid approaches for SC-β-VAE-GAN, including integration with lightweight neural network

architectures on generator or discriminator, as well as investigating parallel processing techniques and model compression strategies. Integrating lightweight architectures is found crucial, based on findings, to reduce the computational burden and improve the scalability, enabling the model to perform efficiently in resource-constrained environments. These approaches are envisioned as potential solutions to address the model's inherent complexity while preserving its good performance. This recommendation is due to the result that, despite the SC-β-VAE-GAN achievement of favorable results with its 100-epoch benchmark, the significant computational demands of extended iterations highlight a potential bottleneck, specifically when applied to large datasets or deployed in real-world applications requiring faster processing times.

3. The researchers recommend using SC-β-VAE-GAN to generate additional synthetic data specifically for underrepresented emotional classes. Because it was found that using augmented data from SC-β-VAE-GAN showed improved performance particularly in detecting depression and stress, yet it has experienced challenges with anxiety classification due to class imbalance. The confusion matrices revealed misclassification issues between normal and anxiety classes, which suggest that future studies should focus on using SC-β-VAE-GAN to augment minority classes. By targeting underrepresented emotions with SC-β-VAE-GAN, the model can achieve a more balanced dataset, reducing bias toward majority classes. This approach could potentially address the classification challenges observed in the current results and could enhance the overall model performance.

References:

Abayomi‑Alli, O., Damaševičius, R., Maskeliūnas, R., & Abayomi‑Alli, A. (2020). BiLSTM with data augmentation using interpolation methods to improve early detection of Parkinson disease. *Annals of Computer Science and Information Systems*. https://doi.org/10.15439/2020f188

Akash, M. A. H., Begum, N., Rahman, S., Shin, J., Amiruzzaman, M., & Islam, M. R. (2020). User Authentication Through Pen Tablet Data Using Imputation and Flatten Function. *Conference: 2020 3rd IEEE International Conference on Knowledge Innovation and Invention (ICKII)*. https://doi.org/10.1109/ickii50300.2020.9318975

Alaei, F., & Alaei, A. (2023). Review of age and gender detection methods based on *handwriting analysis. Neural Computing & Applications (Print), 35*(33), 23909–23925. https://doi.org/10.1007/s00521-023-08996-x

Alai, S., & Afreen, M. (2023). HANDWRITING ANNALYSIS FOR DETECTION OF PERSONALITY TRAITS USING MACHINE LEARNING APPROACH. International Research Journal of Modernization in Engineering Technology and Science. https://doi.org/10.56726/irjmets41243

Alcaraz, J. L., & Strodthoff, N. (2023). Diffusion-based time series imputation and forecasting with structured state space models. *Transactions on Machine Learning Research*.

Annaki, I., Rahmoune, M., & Bourhaleb, M. (2024). Overview of data augmentation techniques in time series analysis. *International Journal of Advanced Computer Science and Applications*, *15*(1). https://doi.org/10.14569/ijacsa.2024.01501118

Azimi, H., Chang, S., Gold, J., & Karabina, K. (2023). Improving accuracy and explainability of online handwritten character recognition. *International Journal on Document Analysis and Recognition*. https://doi.org/10.1007/s10032-023-00456-5

Bang, S. J., Kang, M. J., Lee, M., & Lee, S. M. (2024). STO-CVAE: state transition-oriented conditional variational autoencoder for data augmentation in disability classification. *Complex & Intelligent Systems*, 10(3), 4201–4222. https://doi.org/10.1007/s40747-024-01370-x

Baldán, F. J., & Benítez, J. M. (2021). Multivariate times series classification through an *interpretable representation. Information Sciences*, 569, 596–614. https://doi.org/10.1016/j.ins.2021.05.024

Barnard, J., & Meng, X.-L. (1999). Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research*.

Baydogan, M. G., & Runger, G. (2014). Learning a symbolic representation for multivariate time series classification. *Data Mining and Knowledge Discovery*, 29(2), 400–422. https://doi.org/10.1007/s10618-014-0349-y

Bhandari, P. (2021, December 8). Missing data: Types, explanation, & imputation. *Scribbr*. Retrieved from https://www.scribbr.com/statistics/missing-data/

Biloš, M., Rasul, K., Schneider, A., Nevmyvaka, Y., & Günnemann, S. (2022). Modeling Temporal Data as Continuous Functions with Stochastic Process Diffusion. *arXiv*. https://doi.org/10.48550/arxiv.2211.02590

Boquet, G., Lopez Vicario, J., Morell, A., & Serrano, J. (2019, April 17). Missing data in traffic estimation: A variational autoencoder imputation method. 2019 IEEE International Conference on Acoustics, *Speech and Signal Processing (ICASSP),* Brighton, UK. https://doi.org/10.1109/ICASSP.2019.8683011

Buaton, R., Mawengkang, H., Zarlis, M., Effendi, S., Pardede, A. M. H., Maulita, Y., Fauzi, A., & Novriyenni, N. (2019). Time series optimization on data mining. *Journal of Physics*. Conference Series, 1235(1), 012014. https://doi.org/10.1088/1742-6596/1235/1/012014

Bülte, C., Kleinebrahm, M., Yilmaz, H. Ü., & Gómez-Romero, J. (2023). Multivariate time series imputation for energy data using neural networks. *Energy and AI*, 13, 100239. https://doi.org/10.1016/j.egyai.2023.100239

Chang, J. R., Bresler, M., Chherawala, Y., Delaye, A., Deselaers, T., Dixon, R., & Tuzel, O. (2020). Data incubation -- synthesizing missing data for handwriting recognition. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2110.07040

Chawla, N., Cheng, W., V.Zhang, C., Song, D., Chen, Y., Feng, X., Cheng, W., , Lumezanu, C., (2019). A Deep Neural Network for Unsupervised Anomaly Detection and Diagnosis in Multivariate Time Series Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 1409–1416. doi:10.1609/aaai.v33i01.33011409

Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research*, 84(4), 487–508. Chen, Y., Deng, W., Fang, S., Li, F., Yang, N. T., Zhang, Y., et al. (2023). Provably convergent Schrödinger bridge with applications to probabilistic time series imputation. In *Proceedings of* *the International Conference on Machine Learning (ICML)*.https://doi.org/10.3102/0034654314532697

Chen, P., Xu, M., & Qi, J. (2023). DeepFake detection against adversarial examples based on D‐VAEGAN. *IET Image Processing*. https://doi.org/10.1049/ipr2.12973

Cheung, T.-H., & Yeung, D.-Y. (2021). MODALS: Modality-agnostic automated data augmentation in the latent space. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., & Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5), 545–563. https://doi.org/10.1111/1754-9485.13261

Choi, K., Yi, J., Park, C., & Yoon, S. (2021). Deep Learning for Anomaly Detection in Time-Series Data: Review, analysis, and guidelines. *IEEE Access*, 9, 120043–120065. https://doi.org/10.1109/access.2021.3107975

Croce, D., Castellucci, G., & Basili, R. (2020). GAN-BERT: Generative Adversarial Learning for Robust Text Classification with a Bunch of Labeled Examples. https://doi.org/10.18653/v1/2020.acl-main.191

Dai, Q., Li, Q., Tang, J., & Wang, D. (2017, November 21). Adversarial network embedding. arXiv.org. https://arxiv.org/abs/1711.07838

Donders, A. R. T., Van der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59, 1087–1091.

Esposito, A., Raimo, G., Maldonato, M., Vogel, C., Conson, M., & Cordasco, G. (2020). Behavioral sentiment analysis of depressive states. In Proceedings of the 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom) (pp. 209–214). IEEE. https://doi.org/10.1109/CogInfoCom50765.2020.9237856

Faundez-Zanuy, M., Fierrez, J., Ferrer, M. A., Diaz, M., Tolosana, R., & Plamondon, R. (2020). Handwriting biometrics: Applications and future trends in e-Security and e-Health. *Cognitive Computation*, 12(5), 940–953. https://doi.org/10.1007/s12559-020-09755-z

Farhangfar, A., Kurgan, L. A., & Pedrycz, W. (2007). A novel framework for imputation of missing values in databases. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 37(5), 692-709.

Flores, O. A. V. (2021). Front-end modeling for emotional state recognition. *RePEc*. https://hdl.handle.net/11285/648403

Fortuin, V., Baranchuk, D., Rätsch, G., & Mandt, S. (2019). GP-VAE: Deep Probabilistic Time Series imputation. *arXiv*. https://doi.org/10.48550/arxiv.1907.04155

Gao, J., Song, X., Wen, Q., Wang, P., Sun, L., & Xu, H. (2020). RobustTAD: Robust time series anomaly detection via decomposition and convolutional neural networks. *MileTS'20: 6th KDD Workshop on Mining and Learning from Time Series*, 1–6.

Gao, Z., Li, L., & Xu, T. (2023). Data augmentation for Time-Series Classification: An extensive empirical study and comprehensive survey. *arXiv*. https://doi.org/10.48550/arxiv.2310.10060

Gao, Y., Wang, Y., & Wang, Q. (2023). Improving the Transferability of Time Series Forecasting with Decomposition Adaptation. *ArXiv, abs/2307.00066.*

Gargot, T., Asselborn, T., Pellerin, H., Zammouri, I., Anzalone, S. M., Casteran, L., Johal, W., Dillenbourg, P., Cohen, D., & Jolly, C. (2020). Acquisition of handwriting in children with and without dysgraphia: A computational approach. *PloS One*, 15(9), e0237575. https://doi.org/10.1371/journal.pone.0237575

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. *arXiv.org*. https://arxiv.org/abs/1406.2661

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60(1), 549–576.

Greco, C., Raimo, G., Amorese, T., Cuciniello, M., Mcconvey, G., Cordasco, G., Faundez-Zanuy, M., Vinciarelli, A., Callejas-Carrion, Z., & Esposito, A. (2023). Discriminative power of handwriting and drawing features in depression. *International Journal of Neural Systems*, 34(02). https://doi.org/10.1142/s0129065723500697

Guo, Z., Wan, Y., & Ye, H. (2019). A data imputation method for multivariate time series based on generative adversarial network. *Neurocomputing*, 360, 185–197. https://doi.org/10.1016/j.neucom.2019.06.007

Hamdi, Y., Boubaker, H., & Alimi, A. M. (2021). Data augmentation using geometric, frequency, and beta modeling approaches for improving multi-lingual online handwriting recognition. *International Journal on Document Analysis and*

Recognition (IJDAR), 24(3), 283–298. https://doi.org/10.1007/s10032-021-00376-2

Ham, H., Jun, T. J., & Kim, D. (2020). Unbalanced GANs: Pre-training the Generator of Generative Adversarial Network using Variational Autoencoder. *arXiv*. https://doi.org/10.48550/arxiv.2002.02112

Hasan, T., Rahim, M. A., Shin, J., Nishimura, S., & Hossain, M. N. (2024). Dynamics of Digital Pen-Tablet: Handwriting analysis for person identification using machine and deep learning techniques. *IEEE Access*, 12, 8154–8177. https://doi.org/10.1109/access.2024.3352070

Hou, J., Jiang, H., Wan, C., Yi, L., Gao, S., Ding, Y., & Xue, S. (2022). Deep learning and data augmentation based data imputation for structural health monitoring system in multi-sensor damaged state. *Measurement*, 196, 111206. https://doi.org/10.1016/j.measurement.2022.111206

Hu, M., Jiang, S., Jia, F., Yang, X., & Li, Z. (2023). Improved prediction of aquatic beetle diversity in a stagnant pool by a One-Dimensional convolutional neural network using variational autoencoder generative adversarial Network-Generated data. *Applied Sciences*, *13*(15), 8841. https://doi.org/10.3390/app13158841

Huang, L., Song, M., Shen, H., Hong, H., Gong, P., Deng, H., & Zhang, C. (2023). Deep learning methods for Omics data imputation. *Biology*, 12(10), 1313. https://doi.org/10.3390/biology12101313

Huisman, M. (2000). Imputation of missing item responses: Some simple techniques. *Quality & Quantity*, 34, 331–351.

Iglesias, G., Talavera, E., González-Prieto, Á., Mozo, A., & Gómez-Canaval, S. (2023). Data Augmentation techniques in time series domain: a survey and taxonomy. *Neural Computing & Applications,* 35(14), 10123–10145. https://doi.org/10.1007/s00521-023-08459-3

Iwana, B. K., & Uchida, S. (2021). An empirical survey of data augmentation for time series classification with neural networks. *PloS One*, 16(7), e0254841. https://doi.org/10.1371/journal.pone.0254841

Kamran, I., Naz, S., Razzak, I., & Imran, M. (2020). Handwriting dynamics assessment using deep neural network for early identification of Parkinson's disease. *Future Generation Computer Systems*. https://doi.org/10.1016/j.future.2020.11.020

Kang, Y., Hyndman, R. J., & Li, F. (2020). GRATIS: Generating time series with diverse and controllable characteristics. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 13(4), 354-376.

Kapp, A., Hansmeyer, J., & Mihaljević, H. (2023). Generative Models for Synthetic Urban Mobility Data: A Systematic Literature review. *ACM Computing Surveys*, *56*(4), 1–37. https://doi.org/10.1145/3610224

Kazijevs, M., & Samad, M. D. (2023). Deep imputation of missing values in time series health data: A review with benchmarking. *Journal of Biomedical Informatics*, 144, 104440. https://doi.org/10.1016/j.jbi.2023.104440

Keskin D. (2023,). Synthetic data and data augmentation. *Medium*. Retrieved from https://medium.com/@dogankeskin01/synthetic-data-and-data-augmentation-c022029dd660

Khan, Z. M., Xia, Y., Aurangzeb, K., Khaliq, F., Alam, M., Khan, J. A., & Anwar, M. S. (2024). Emotion detection from handwriting and drawing samples using an attention-based transformer model. PeerJ. *Computer Science*, 10, e1887. https://doi.org/10.7717/peerj-cs.1887

Kim, G., Yoo, H., Cho, H., & Chung, K. (2023). Defect detection model using time series data augmentation and transformation. *Computers, Materials & Continua/Computers, Materials & Continua*, 0(0), 1–10. https://doi.org/10.32604/cmc.2023.046324

Kulevome, D. K. B., Wang, H., Cobbinah, B. M., Mawuli, E. S., & Kumar, R. (2024). Effective time-series Data Augmentation with Analytic Wavelets for bearing fault diagnosis. *Expert Systems With Applications*, *249*, 123536. https://doi.org/10.1016/j.eswa.2024.123536

Kunhoth, J., Maadeed, S. A., Saleh, M., & Akbari, Y. (2023). Exploration and analysis of On-Surface and In-Air handwriting attributes to improve dysgraphia disorder diagnosis in children based on machine learning methods. *Biomedical Signal Processing and Control*, 83, 104715. https://doi.org/10.1016/j.bspc.2023.104715

Kumar, K., Kumar, R., Thibault, D. B., Gestin, L., Teoh, W. Z., Sotelo, J., Alexandre, D. B., Bengio, Y., & Courville, A. (2019, October 8). *MELGAN: Generative Adversarial Networks for Conditional Waveform Synthesis*. arXiv.org.

Lee, A. L. A., Wah, L. L., Min, L. H., & Chen, O. S. (2022). Revisiting handwriting fundamentals through an interdisciplinary framework. *The Malaysian Journal of Medical Sciences the Malaysian Journal of Medical Science*, 29(1), 18–33. https://doi.org/10.21315/mjms2022.29.1.3https://arxiv.org/abs/1910.06711

Lee, C. H. J. (2022). *VRNNGAN: A Recurrent VAE-GAN Framework for Synthetic Time-Series (Doctoral dissertation)*.

Lee, T. K.-M., Kuah, Y. L., Leo, K.-H., Sanei, S., Chew, E., & Zhao, L. (2019). Surrogate rehabilitative time series data for image-based deep learning. In *EUSIPCO 2019* (pp. 1–5).

Li, J. (2023). Data Generation and Latent Space Based Feature Transfer Using ED-VAEGAN, an Improved Encoder and Decoder Loss VAEGAN Network. In *Proceedings of the 2023 2nd International Conference on Artificial Intelligence, Internet and Digital Economy (ICAID 2023) Volume 9* (pp. 123–135). https://doi.org/10.2991/978-94-6463-222-4_12

Li, H., & Du, T. (2021). Multivariate time-series clustering based on component relationship networks. *Expert Systems With Applications*, 173, 114649. https://doi.org/10.1016/j.eswa.2021.114649

Li, J., Ren, W., & Han, M. (2021). Variational auto-encoders based on the shift correction for imputation of specific missing in multivariate time series. *Measurement*, 186, 110055. https://doi.org/10.1016/j.measurement.2021.110055

Likforman-Sulem, L., Esposito, A., Faundez-Zanuy, M., Clemencon, S., & Cordasco, G. (2017). EMOTHAW: a novel database for emotional state recognition from handwriting and drawing. IEEE Transactions on Human-machine Systems, 47(2), 273–284. https://doi.org/10.1109/thms.2016.2635441

Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: a survey. Philosophical Transactions - Royal Society. Mathematical, Physical and Engineering Sciences/Philosophical Transactions - Royal Society. *Mathematical,* *Physical and Engineering Sciences*, 379(2194), 20200209. https://doi.org/10.1098/rsta.2020.0209

Lintonen, T., & Raty, T. (2019). Self-learning of multivariate time series using perceptually important points. *IEEE/CAA Journal of Automatica Sinica*, 6(6), 1318–1331. https://doi.org/10.1109/jas.2019.1911777

Liu, M., Huang, H., Feng, H., Sun, L., Du, B., & Fu, Y. (2023). PRISTI: A Conditional Diffusion Framework for Spatiotemporal Imputation. *arXiv*. https://doi.org/10.48550/arxiv.2302.09746

Liu, R., Liu, W., Zheng, Z., Wang, L., Mao, L., Qiu, Q., & Ling, G. (2023). Anomaly-GAN: A data augmentation method for train surface anomaly detection. *Expert Systems With Applications*, *228*, 120284. https://doi.org/10.1016/j.eswa.2023.120284

Luo, Y., Zhang, Y., Cai, X., & Yuan, X. (2019). E2GAN: End-to-end generative adversarial network for multivariate time series imputation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19)* (pp. 3094–3100). AAAI Press.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. Guilford Press.

Miao, X., Wu, Y., Wang, J., Gao, Y., Mao, X., & Yin, J. (2021). Generative semi-supervised learning for multivariate time series imputation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10), 8983–8991. https://doi.org/10.1609/aaai.v35i10.17086

Mishra, S. (2024). An introduction to VAE-GANs. *Weights & Biases*. Retrieved from https://wandb.ai/shambhavicodes/vae-gan/reports/An-Introduction-to-VAE-GANs--VmlldzoxMTcxMjM5

Morrill, J., Fermanian, A., Kidger, P., & Lyons, T. (2020, June 1). A generalised signature method for multivariate time series feature extraction. *arXiv.org*. https://arxiv.org/abs/2006.0087

Muhamed, A., Li, L., Shi, X., Yaddanapudi, S., Chi, W., Jackson, D., Suresh, R., Lipton, Z. C., & Smola, A. J. (2021). Symbolic Music Generation with Transformer-GANs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1), 408–417. https://doi.org/10.1609/aaai.v35i1.16117

Mumuni, A., & Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. Array, 16, 100258. https://doi.org/10.1016/j.array.2022.100258

Nolazco-Flores, J. A., Faundez-Zanuy, M., Velázquez-Flores, O. A., Cordasco, G., & Esposito, A. (2021). Emotional state recognition performance improvement on a handwriting and drawing task. *IEEE Access*, 9, 28496–28504. https://doi.org/10.1109/access.2021.3058443

Nolazco-Flores, J. A., Faundez-Zanuy, M., Velázquez-Flores, O. A., Del-Valle-Soto, C., Cordasco, G., & Esposito, A. (2022). Mood state detection in handwritten tasks using PCA–MFCBF and automated machine learning. *Sensors*, 22(4), 1686. https://doi.org/10.3390/s22041686

Nita, S., Bitam, S., Heidet, M., & Mellouk, A. (2022). A new data augmentation convolutional neural network for human emotion recognition based on ECG signals. *Biomedical Signal Processing and Control*, 74, 103580. https://doi.org/10.1016/j.bspc.2022.103580

Niu, Z., Yu, K., & Wu, X. (2020). LSTM-Based VAE-GAN for Time-Series Anomaly Detection. *Sensors*, 20(13), 3738. https://doi.org/10.3390/s20133738

Otero, J. F. A., López-de-Ipina, K., Caballer, O. S., & others. (2022). EMD-based data augmentation method applied to handwriting data for the diagnosis of essential tremor using LSTM networks. *Scientific Reports,* 12(1), 12819. https://doi.org/10.1038/s41598-022-16741-y

Ozyurt, F., Majidpour, J., Rashid, T. A., & Koc, C. (2024). Offline Handwriting Signature Verification: A transfer learning and feature selection approach. *arXiv*. https://doi.org/10.48550/arxiv.2401.09467

Paepae, T., Bokoro, P., & Kyamakya, K. (2023). Data augmentation for a Virtual-Sensor-Based nitrogen and phosphorus monitoring. *Sensors*, 23(3), 1061. https://doi.org/10.3390/s23031061

Pan, B., & Zheng, W. (2021). Emotion recognition based on EEG using generative adversarial nets and convolutional neural network. *Hindawi*. https://doi.org/10.1155/2021/2520394

Park, W., Babushkin, V., Tahir, S., & Eid, M. (2021). Haptic guidance to support handwriting for children with cognitive and fine motor delays. *IEEE Transactions on Haptics*, 14(3), 626–634. https://doi.org/10.1109/toh.2021.3068786

Pourshahrokhi, N., Kouchaki, S., Kober, K. M., Miaskowski, C. A., & Barnaghi, P. M. (2021). A Hamiltonian Monte Carlo model for imputation and augmentation of healthcare data. *arXiv*. Retrieved from https://api.semanticscholar.org/CorpusID:232105260

Puyat, J. H., Gastardo-Conaco, M. C., Natividad, J., & Banal, M. A. (2021). Depressive symptoms among young adults in the Philippines: Results from a nationwide cross-sectional survey. Journal of Affective Disorders Reports, 3, 100073. https://doi.org/10.1016/j.jadr.2020.100073

Qiu, Y. L., Zheng, H., & Gevaert, O. (2020). Genomic data imputation with variational auto-encoders. *Gigascience*, 9(8). https://doi.org/10.1093/gigascience/giaa082

Rabaev, I., Alkoran, I., Wattad, O., & Litvak, M. (2022). Automatic Gender and Age Classification from Offline Handwriting with Bilinear ResNet. *Sensors*, *22*(24), 9650. https://doi.org/10.3390/s22249650

Rath, K., Rügamer, D., Bischl, B., Von Toussaint, U., & Albert, C. G. (2023). Dependent state space Student‐t processes for imputation and data augmentation in plasma diagnostics. *Contributions to Plasma Physics*, 63(5–6). https://doi.org/10.1002/ctpp.202200175

Richter, A., Ijaradar, J., Wetzker, U., Jain, V., & Frotzscher, A. (2022). Multivariate Time Series Imputation: A Survey on available Methods with a Focus on hybrid GANs. *TechRxiv*. https://doi.org/10.36227/techrxiv.21572070.v1

Ruan, D., Chen, X., Gühmann, C., & Yan, J. (2023). Improvement of Generative adversarial Network and its application in bearing fault diagnosis: a review. *Lubricants*, *11*(2), 74. https://doi.org/10.3390/lubricants11020074

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.

Savarana. M.K, Roopa. M.S, Arunalatha, J.S, Venugopal. K.R. (2024). Navigating Data Scarcity in Multivariate Time Series Forecasting: A Hybrid Model Perspective. 2017 IEEE Region 10 Symposium (TENSYMP), 1–7. https://doi.org/10.1109/tensymp61132.2024.10752270

Saldanha, J., Chakraborty, S., Patil, S., Kotecha, K., Kumar, S., & Nayyar, A. (2022). Data augmentation using Variational Autoencoders for improvement of respiratory disease classification. *PloS One*, 17(8), e0266467. https://doi.org/10.1371/journal.pone.0266467

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology*, 57(1), 1.

Singh, S., Sharma, A., & Chauhan, V. K. (2023). Indic script family and its offline handwriting recognition for characters/digits and words: a comprehensive survey. *Artificial Intelligence Review*, *56*(S3), 3003–3055. https://doi.org/10.1007/s10462-023-10597-y

Steve, M. N., Olusegun, J., & Paul, H. (2024, October 16). Addressing class imbalance with synthetic data generation. https://www.researchgate.net/publication/385592781_Addressing_class_imbalance_with_synthetic_data_generation

Sürmeli, B. G., & Tümer, M. B. (2019). Multivariate Time Series Clustering and its Application in Industrial Systems. Cybernetics and Systems, 51(3), 315–334. https://doi.org/10.1080/01969722.2019.1691851

Szczakowska, P., Wosiak, A., & Żykwińska, K. (2023). Improving automatic recognition of emotional states using EEG data augmentation techniques. *Procedia Computer Science*. https://doi.org/10.1016/j.procs.2023.10.419

Taleb, C., & Likforman-Sulem, L. (2020). Improving deep learning Parkinson's disease detection through data augmentation training. In Lecture Notes in Computer Science (Vol. 11915). https://doi.org/10.1007/978-3-030-37548-5_7

Tashiro, Y., Song, J., Song, Y., & Ermon, S. (2021). CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In *Proceedings of the Neural Information Processing Systems*.

Torres, J. F., Hadjout, D., Sebaa, A., Martínez-Álvarez, F., & Troncoso, A. (2021). Deep learning for Time Series Forecasting: A survey. *Big Data*, 9(1), 3–21. https://doi.org/10.1089/big.2020.0159

Velázquez Flores, O. A. (2021). Front-end modeling for emotional state recognition. *RePEc*. https://hdl.handle.net/11285/648403

Vilardell, M., Buxó, M., Clèries, R., Martínez, J. M., Garcia, G., Ameijide, A., Font, R., Civit, S., Marcos-Gragera, R., Vilardell, M. L., Carulla, M., Espinàs, J. A., Galceran, J., Izquierdo, A., & Borràs, J. M. (2020). Missing data imputation and synthetic data simulation through modeling graphical probabilistic dependencies between variables (ModGraProDep): An application to breast cancer survival. *Artificial Intelligence in Medicine*, 107, 101875. https://doi.org/10.1016/j.artmed.2020.101875

Wang, J., Du, W., Cao, W., Zhang, K., Wang, W., Liang, Y., & Wen, Q. (2024). Deep learning for multivariate Time Series Imputation: a survey. *arXiv*. https://doi.org/10.48550/arxiv.2402.04059

Wang, Y., Xiao, W., & Li, S. (2021). Offline Handwritten text recognition using Deep Learning: A review. *Journal of Physics. Conference Series, 1848*(1), 012015. https://doi.org/10.1088/1742-6596/1848/1/012015

Wang, X., Chen, H., Tang, S., Wu, Z., & Zhu, W. (2022, November 21). Disentangled representation learning. arXiv.org. https://arxiv.org/abs/2211.11695

Wang, X., Zhang, H., Wang, P., Zhang, Y., Wang, B., Zhou, Z., & Wang, Y. (2023). An observed value consistent diffusion model for imputing missing values in multivariate time series. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (SIGKDD)*.

Wang, Z., She, Q., & Ward, T. E. (2021). Generative adversarial networks in computer vision. *ACM Computing Surveys*, 54(2), 1–38. https://doi.org/10.1145/3439723

Weerakody, P. B., Wong, K. W., Wang, G., & Ela, W. (2021). A review of irregular time series data handling with gated recurrent neural networks. *Neurocomputing*, *441*, 161–178. https://doi.org/10.1016/j.neucom.2021.02.046

Wen, J., & Angryk, R. A. (2024). Class-Based Time series data augmentation to mitigate extreme class imbalance for solar flare prediction. *arXiv.* https://doi.org/10.48550/arxiv.2405.20590

Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., & Xu, H. (2021b). Time series data augmentation for deep learning: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence* (IJCAI-21). https://doi.org/10.24963/ijcai.2021/631

Weerakody, P. B., Wong, K. W., Wang, G., & Ela, W. (2021). A review of irregular time series data handling with gated recurrent neural networks. *Neurocomputing*, 441, 161–178. https://doi.org/10.1016/j.neucom.2021.02.046

World Health Organization. (2017). Depression and other common mental disorders: Global health estimates. Retrieved from https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf.

Xu, J., Lyu, F., & Yuen, P. C. (2023). Density-aware temporal attentive step-wise diffusion model for medical time series imputation. In *Proceedings of the 2023 ACM International Conference on Information and Knowledge Management (CIKM)*.

Ye, F., & Bors, A. G. (2020). Learning latent representations across multiple data domains using lifelong VAEGAN. In *Lecture notes in computer science* (pp. 777–795). https://doi.org/10.1007/978-3-030-58565-5_46

Yang, H., & Desell, T. (2022). Robust augmentation for multivariate time series classification. *arXiv* . https://doi.org/10.48550/arxiv.2201.11739

Yi, X., Walia, E., & Babyn, P. S. (2019). Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58, 101552. https://doi.org/10.1016/j.media.2019.101552

Yoon, J., Jarrett, D., & Van Der Schaar, M. (2019). Time-series generative adversarial networks. *Neural Information Processing Systems*, 32, 5508–5518. https://papers.nips.cc/paper/8789-time-series-generative-adversarial-networks.pdf

Zhang, S., Gong, L., Zeng, Q., Li, W., Xiao, F., & Lei, J. (2021). Imputation of GPS coordinate time series using MissForest. *Remote Sensing*, 13(12), 2312. https://doi.org/10.3390/rs13122312

Zhang, X., Wang, Z., Lu, K., & Pan, Q. (2023). Data augmentation and classification of Sea-Land clutter for Over-the-Horizon Radar using AC-VAEGAN. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2301.00947

Zhou, Y., Aryal, S., & Bouadjenek, M. R. (2024). Review for Handling Missing Data with special missing mechanism. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2404.04905

Zhu, S., Ziwei, X., & Li, Y. (2024). Electricity theft detection in smart grids based on omni-scale CNN and AutoXGB. IEEE Access, 12, 15477–15492. https://doi.org/10.1109/access.2024.3358683

APPENDICES

Appendix 1: Experiment Paper

Objective:

The objective of the experiment is to answer the Statement of the Problem.

Materials and Equipment:

- The developed SC-β-VAE-GAN model for generating time-series synthetic data.
- The two datasets:
  - EMOTHAW dataset
  - GPS Time Series Data
- Experiment Paper
- Pen and Paper

Procedure:

1) Researchers prepared an experiment paper to serve as a guide for evaluating the developed system.
2) Then researchers will generate synthetic data for the given data set, specifically with EMOTHAW dataset.
3) After generating and iterating on the data, the performance of the proposed model, SC-β-VAE-GAN, in generating synthetic data will be evaluated using

metrics such as Normalized Root Mean Square Error (NRMSE), Post-Hoc Discriminative Score, and Post-Hoc Predictive Score.

Result of the SC-β-VAE-GAN Model

Performance of SC-β-VAE-GAN			
Model	**Normalized Root Mean Square Error**	**Post-Hoc Discriminative Score**	**Post-Hoc Predictive Score**

4) The evaluation will also be repeated with other baseline models, including VAE-GAN, TimeGAN, and VRNNGAN, with the results documented in the following table:

Result of the VAEGAN , TimeGAN and VRNNGAN

Performance of VAEGAN, TimeGAN and VRNNGAN			
Model	**Normalized Root Mean Square Error**	**Post-Hoc Discriminative Score**	**Post-Hoc Predictive Score**
VAEGAN			
TimeGAN			
VRNNGAN			

5. To address the hypothesis, a Kruskal-Wallis Test with Pairwise Comparison will be used between the metric results of the proposed model and other baseline models, specifically VAE-GAN, Time-GAN, and VRNNGAN.

Evaluation of Normalized Root Mean Square Error

Normalized Root Mean Square Error		
Model	EMOTHAW Dataset	
	Mean	**Std.**
VAE-GAN		
TimeGAN		
VRNNGAN		
SC-β-VAE-GAN		

Evaluation of Post Hoc Discriminative Score

Discriminative Score		
Model	EMOTHAW Dataset	
	Mean	**Std.**
VAE-GAN		
TimeGAN		
VRNNGAN		
SC-β-VAE-GAN		

Evaluation of Post Hoc Predictive Score

Predictive MAPE Score		
Model	EMOTHAW Dataset	
	Mean	**Std.**
VAE-GAN		
TimeGAN		
VRNNGAN		
SC-β-VAE-GAN		

Kruskal Wallis Hypothesis Testing

	x^2	df	p	ε^2
NRMSE				
PREDICTIVE				
DISCRIMINATIVE				

SC-β-VAE-GAN and VAEGAN pairwise comparisons

SC-β-VAE-GAN and VAEGAN			
	W	p	Conclusion
NRMSE			
PREDICTIVE			
DISCRIMINATIVE			

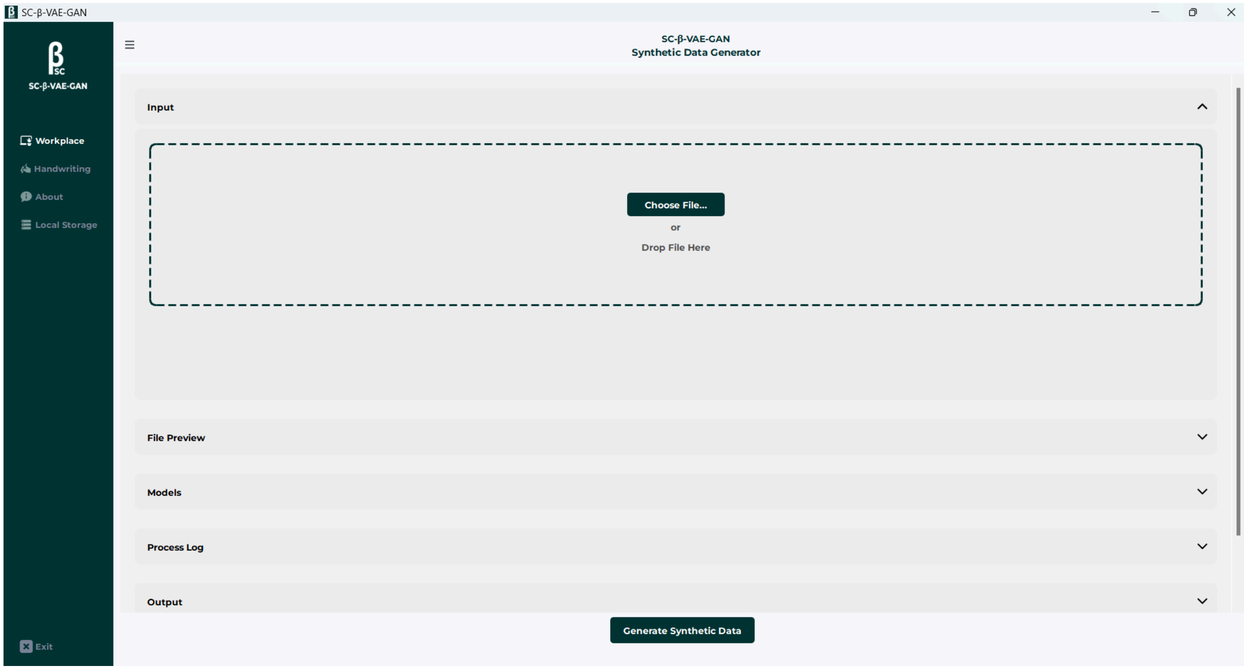
SC-β-VAE-GAN and TIMEGAN pairwise comparisons

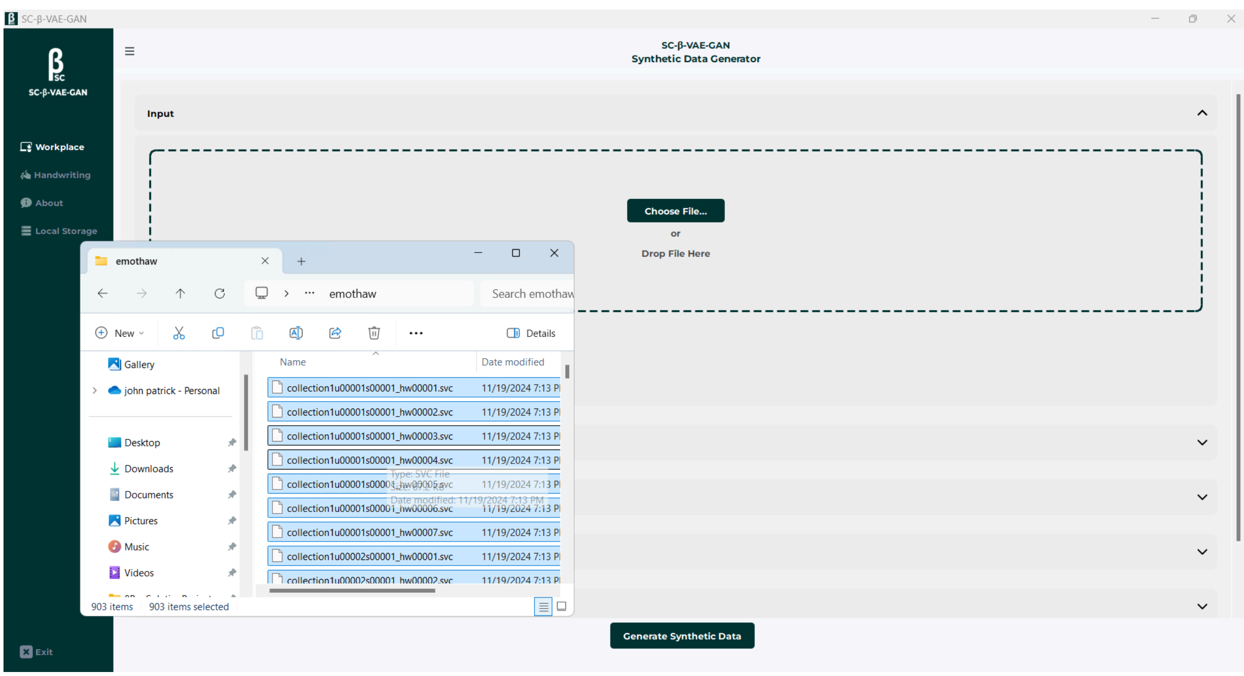
SC-β-VAE-GAN and TIMEGAN			
	W	p	Conclusion
NRMSE			
PREDICTIVE			
DISCRIMINATIVE			

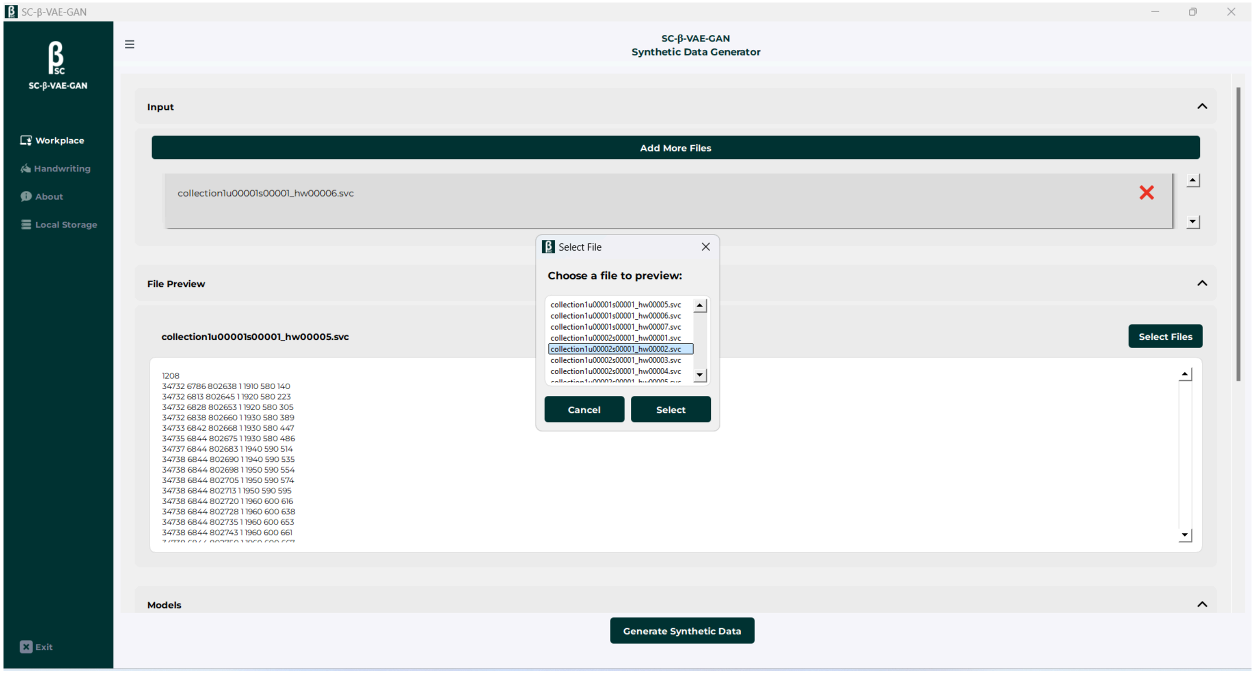
SC-β-VAE-GAN and VRNNGAN pairwise comparisons

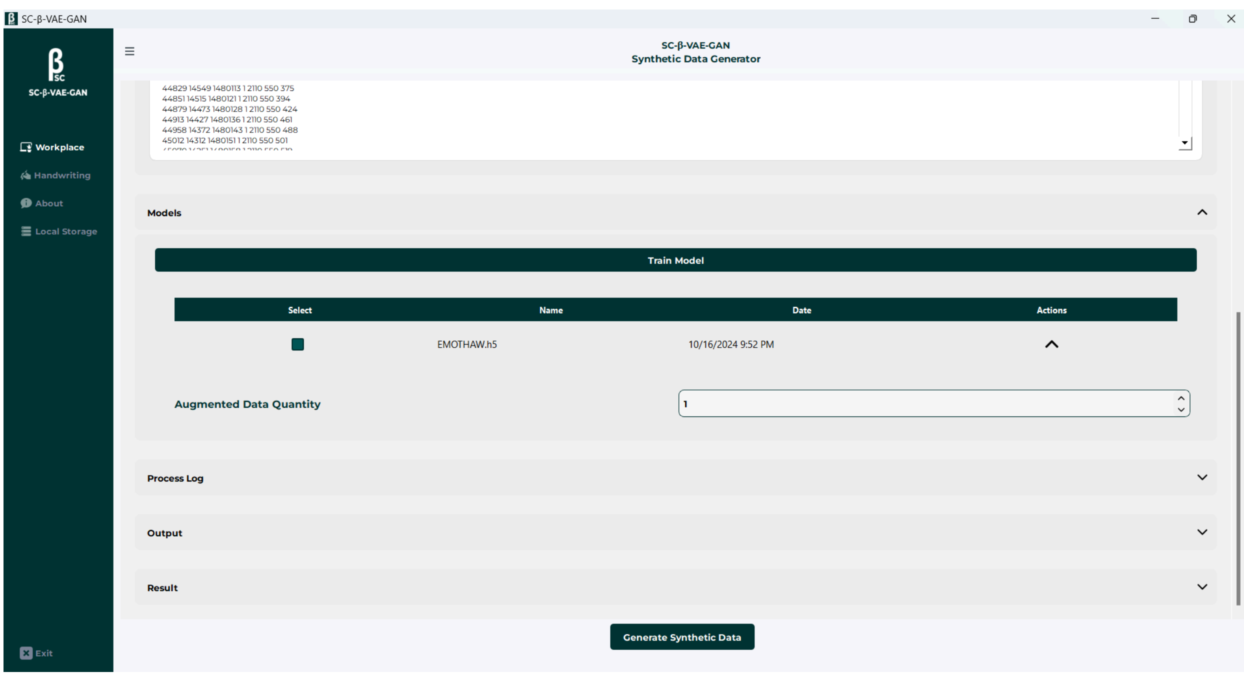
SC-β-VAE-GAN and VRNNGAN			
	W	p	Conclusion
NRMSE			
PREDICTIVE			
DISCRIMINATIVE			

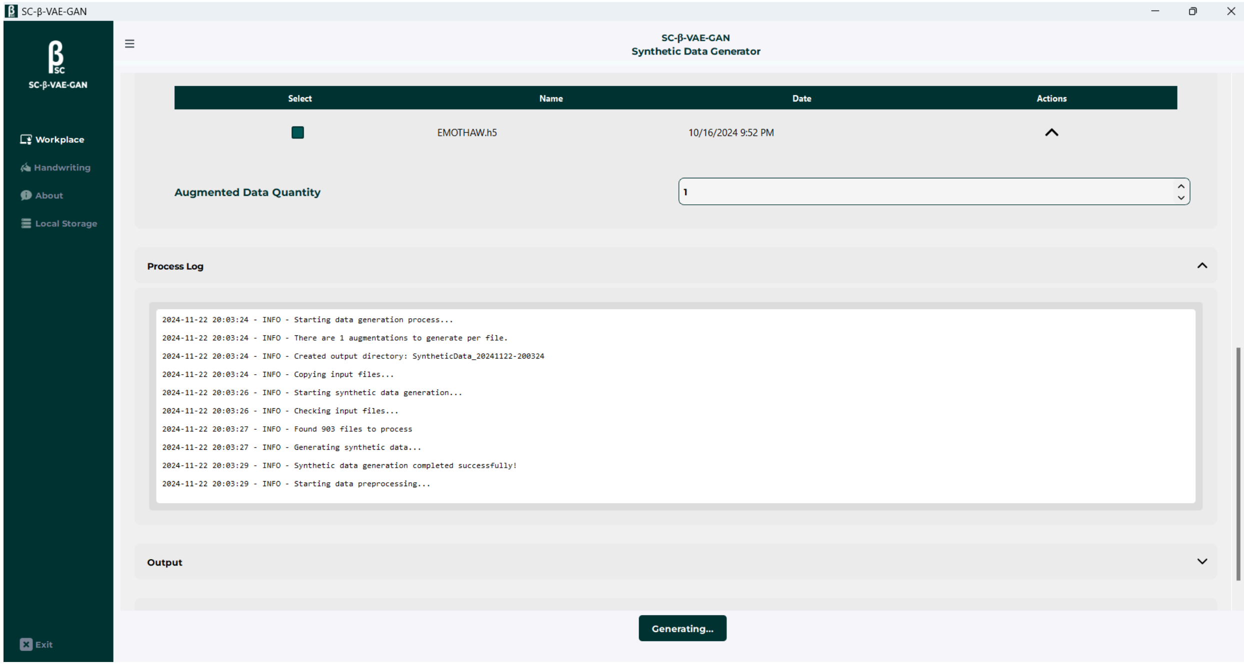
Appendix 2: Screen Layout of the Tool

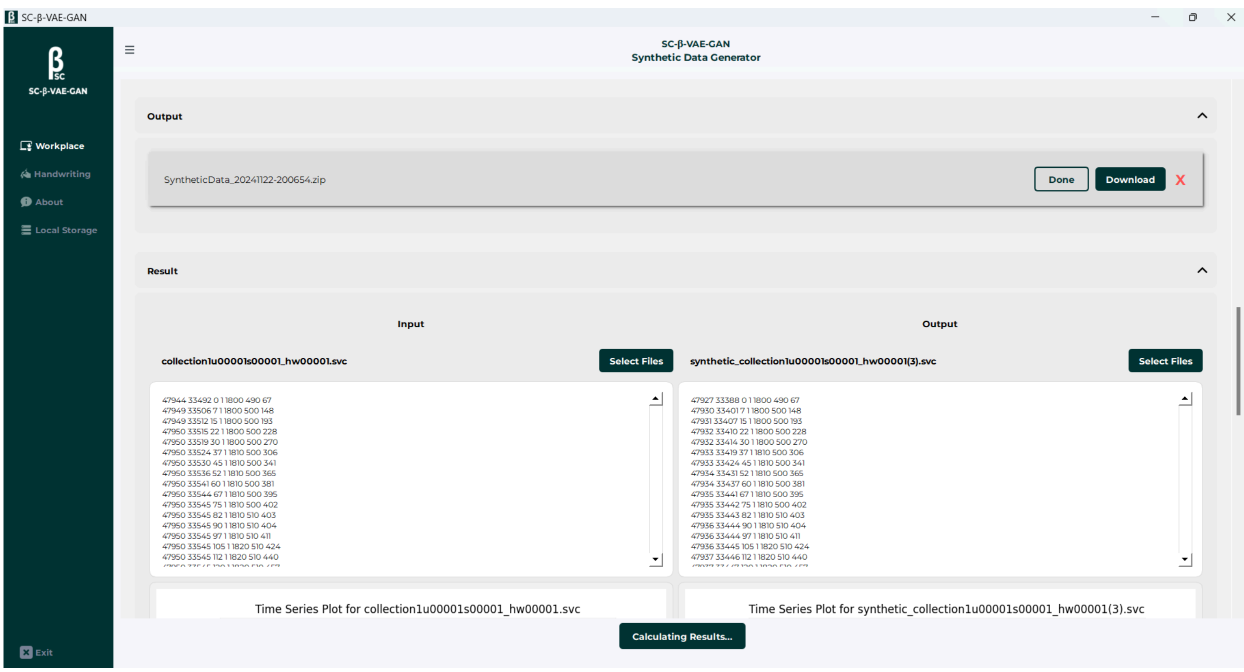


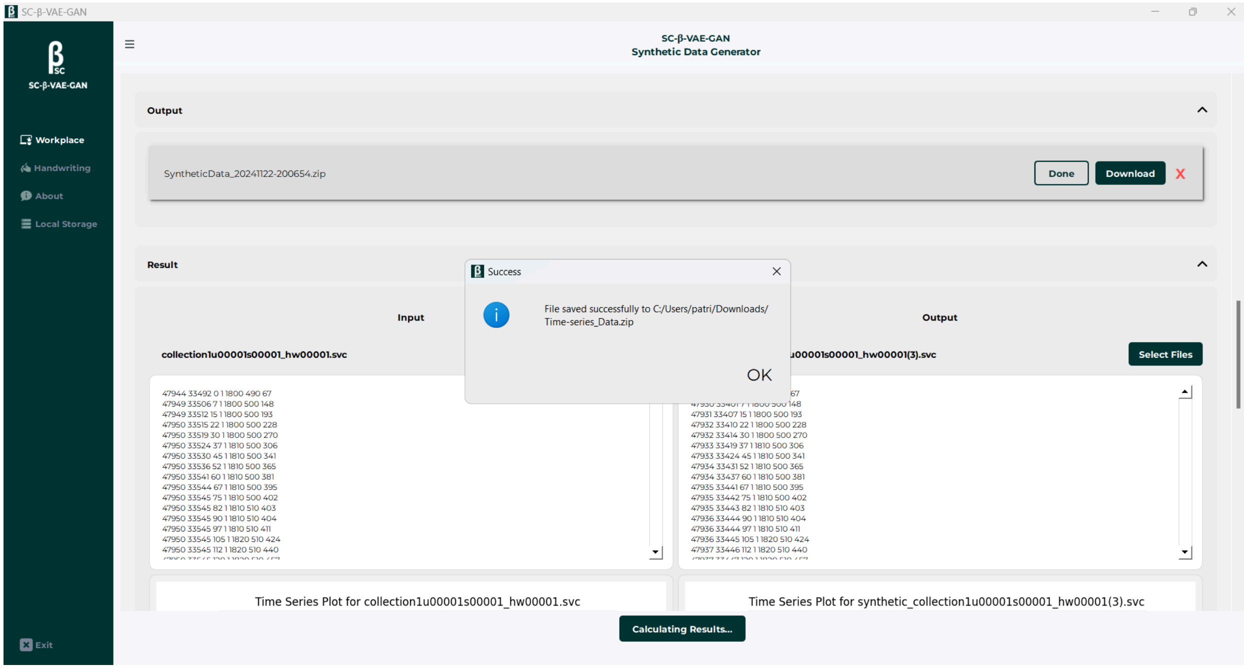


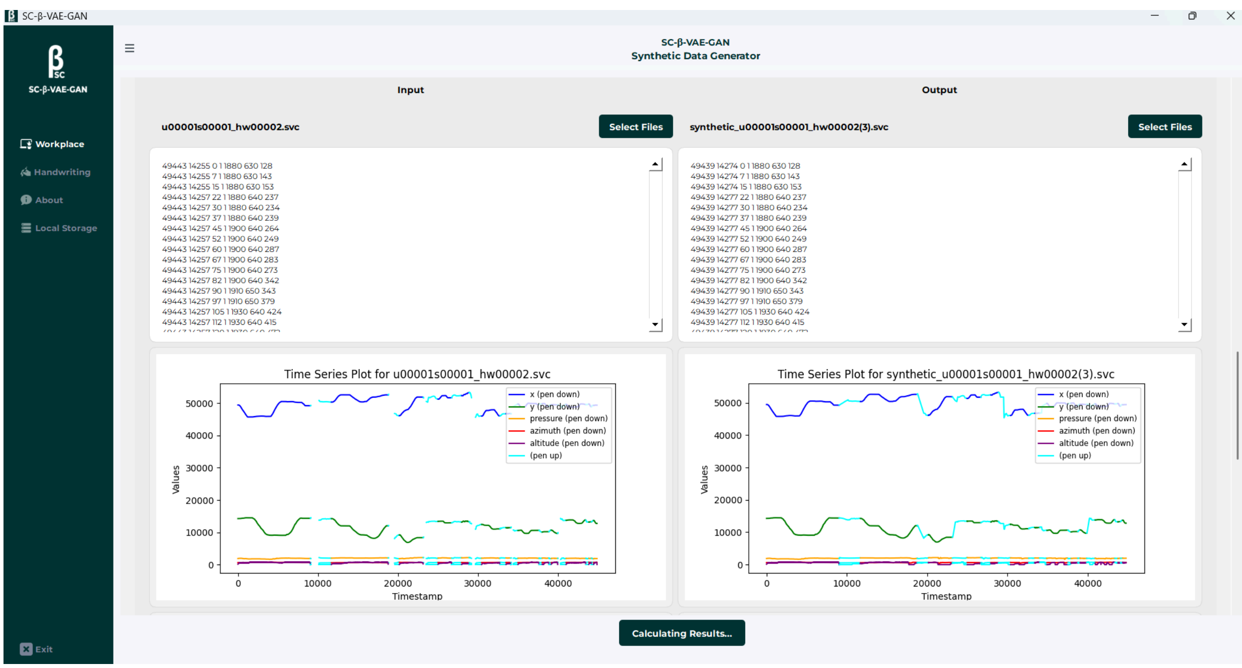

SC-β-VAE-GAN
Synthetic Data Generator
Workplace
Handwriting
About
Local Storage
Exit
Input
Add More Files
collection1u00001s00001_hw00006.svc
Select File
Choose a file to preview:
Cancel
Select
File Preview
collection1u00001s00001_hw00005.svc
Select Files
Models
Generate Synthetic Data

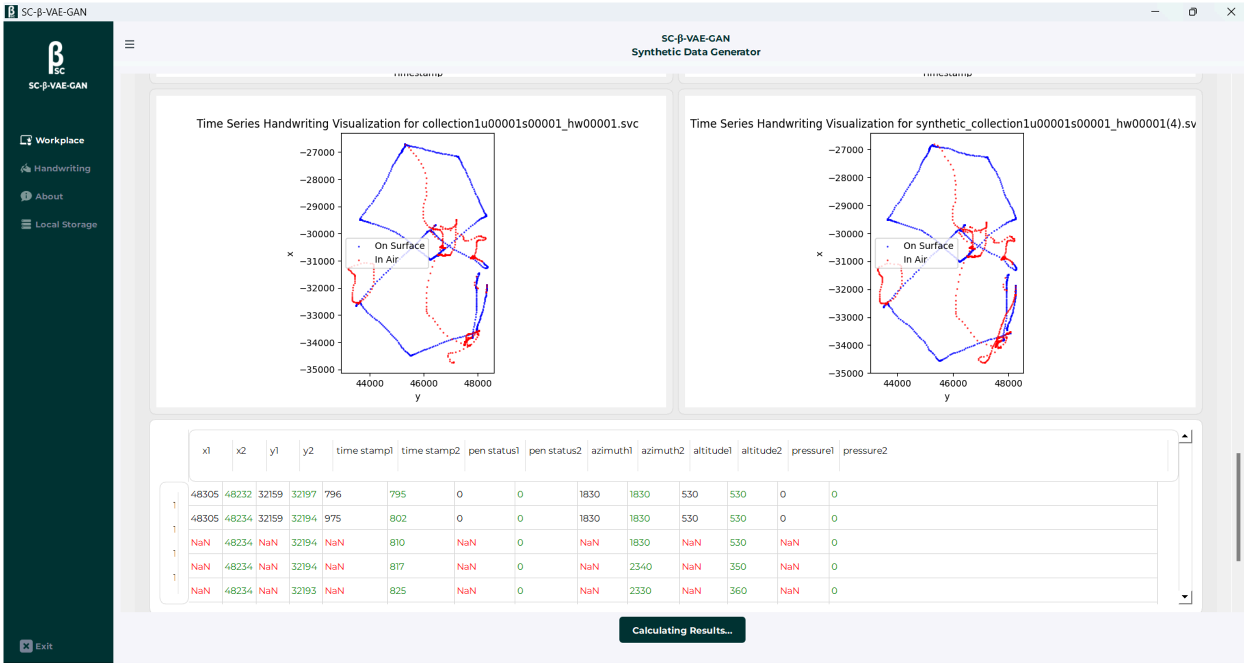

SC-β-VAE-GAN
Synthetic Data Generator
Workplace
Handwriting
About
Local Storage
Exit
Models
Train Model
Select
Name
Date
Actions
EMOTHAW.h5
10/16/2024 9:52 PM
Augmented Data Quantity
Process Log
Output
Result
Generate Synthetic Data

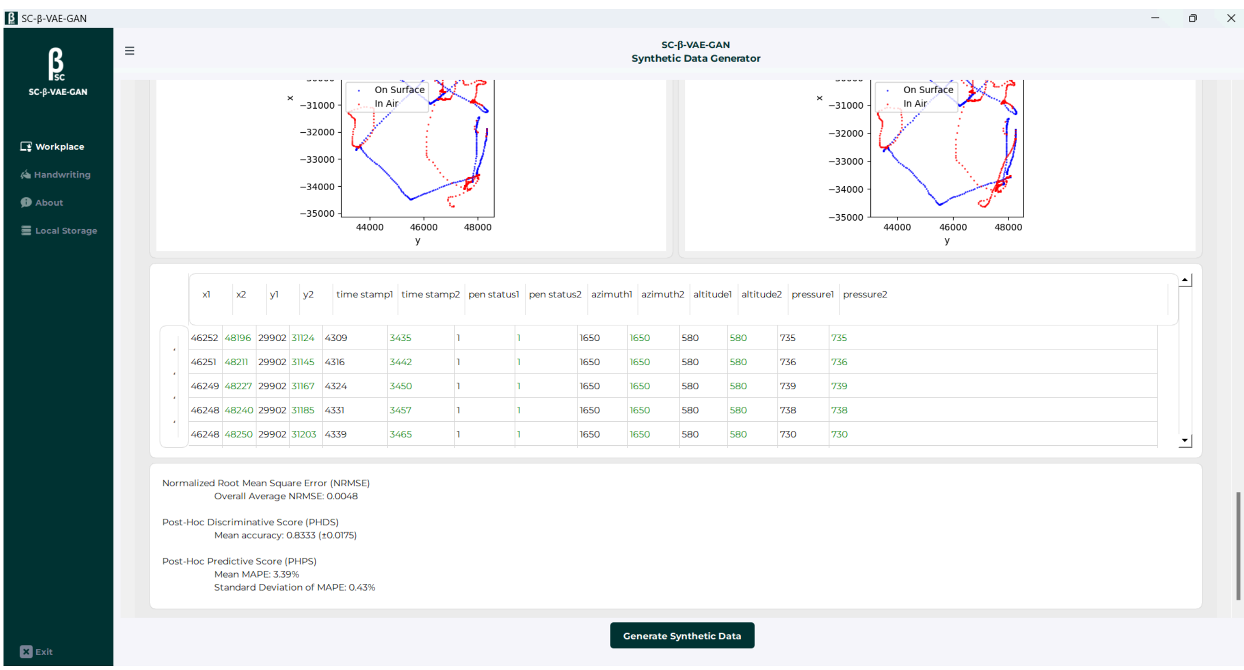

SC-β-VAE-GAN
SC-β-VAE-GAN
Synthetic Data Generator
Workplace
Handwriting
About
Local Storage
Exit
Select
Name
Date
Actions
EMOTHAW.h5
10/16/2024 9:52 PM
Augmented Data Quantity
Process Log
2024-11-22 20:03:24 - INFO - Starting data generation process...
2024-11-22 20:03:24 - INFO - There are 1 augmentations to generate per file.
2024-11-22 20:03:24 - INFO - Created output directory: SyntheticData_20241122-200324
2024-11-22 20:03:24 - INFO - Copying input files...
2024-11-22 20:03:26 - INFO - Starting synthetic data generation...
2024-11-22 20:03:26 - INFO - Checking input files...
2024-11-22 20:03:27 - INFO - Found 903 files to process
2024-11-22 20:03:27 - INFO - Generating synthetic data...
2024-11-22 20:03:29 - INFO - Synthetic data generation completed successfully!
2024-11-22 20:03:29 - INFO - Starting data preprocessing...
Output
Generating...

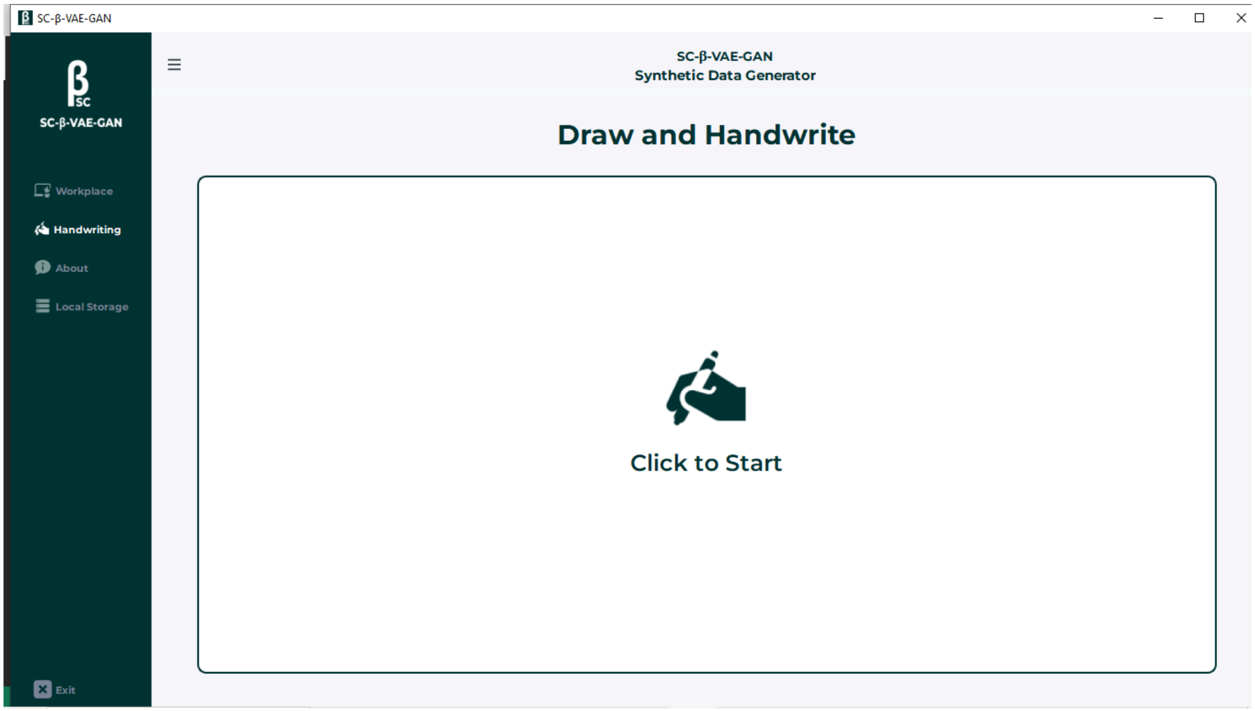

SC-β-VAE-GAN
SC-β-VAE-GAN
Synthetic Data Generator
Workplace
Handwriting
About
Local Storage
Exit
Output
SyntheticData_20241122-200654.zip
Done
Download
Result
Input
Output
collection1u00001s00001_hw00001.svc
Select Files
synthetic_collection1u00001s00001_hw00001(3).svc
Select Files
Time Series Plot for collection1u00001s00001_hw00001.svc
Time Series Plot for synthetic_collection1u00001s00001_hw00001(3).svc
Calculating Results...

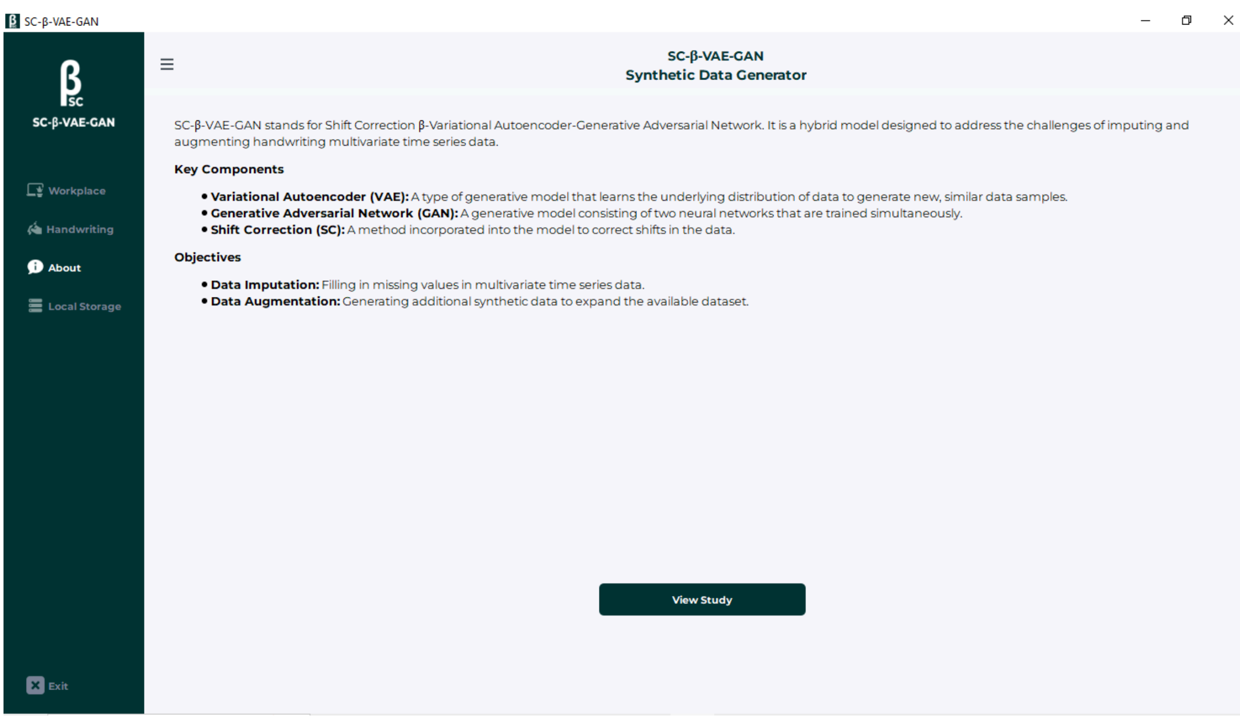

SC-β-VAE-GAN
SC-β-VAE-GAN
Synthetic Data Generator
Workplace
Handwriting
About
Local Storage
Exit
Output
SyntheticData_20241122-200654.zip
Done
Download
X
Result
Success
File saved successfully to C:/Users/patri/Downloads/
Time-series_Data.zip
OK
Input
Output
collection1u00001s00001_hw00001.svc
u00001s00001_hw00001(3).svc
Select Files
47944 33492 0 1 1800 490 67
47949 33506 7 1 1800 500 148
47949 33512 15 1 1800 500 193
47950 33515 22 1 1800 500 228
47950 33519 30 1 1800 500 270
47950 33524 37 1 1810 500 306
47950 33530 45 1 1810 500 341
47950 33536 52 1 1810 500 365
47950 33541 60 1 1810 500 381
47950 33544 67 1 1810 500 395
47950 33545 75 1 1810 500 402
47950 33545 82 1 1810 510 403
47950 33545 90 1 1810 510 404
47950 33545 97 1 1810 510 411
47950 33545 105 1 1820 510 424
47950 33545 112 1 1820 510 440
47931 33407 15 1 1800 500 193
47932 33410 22 1 1800 500 228
47932 33414 30 1 1800 500 270
47933 33419 37 1 1810 500 306
47933 33424 45 1 1810 500 341
47934 33431 52 1 1810 500 365
47934 33437 60 1 1810 500 381
47935 33441 67 1 1810 500 395
47935 33442 75 1 1810 500 402
47935 33443 82 1 1810 510 403
47936 33444 90 1 1810 510 404
47936 33444 97 1 1810 510 411
47936 33445 105 1 1820 510 424
47937 33446 112 1 1820 510 440
Time Series Plot for collection1u00001s00001_hw00001.svc
Time Series Plot for synthetic_collection1u00001s00001_hw00001(3).svc
Calculating Results...

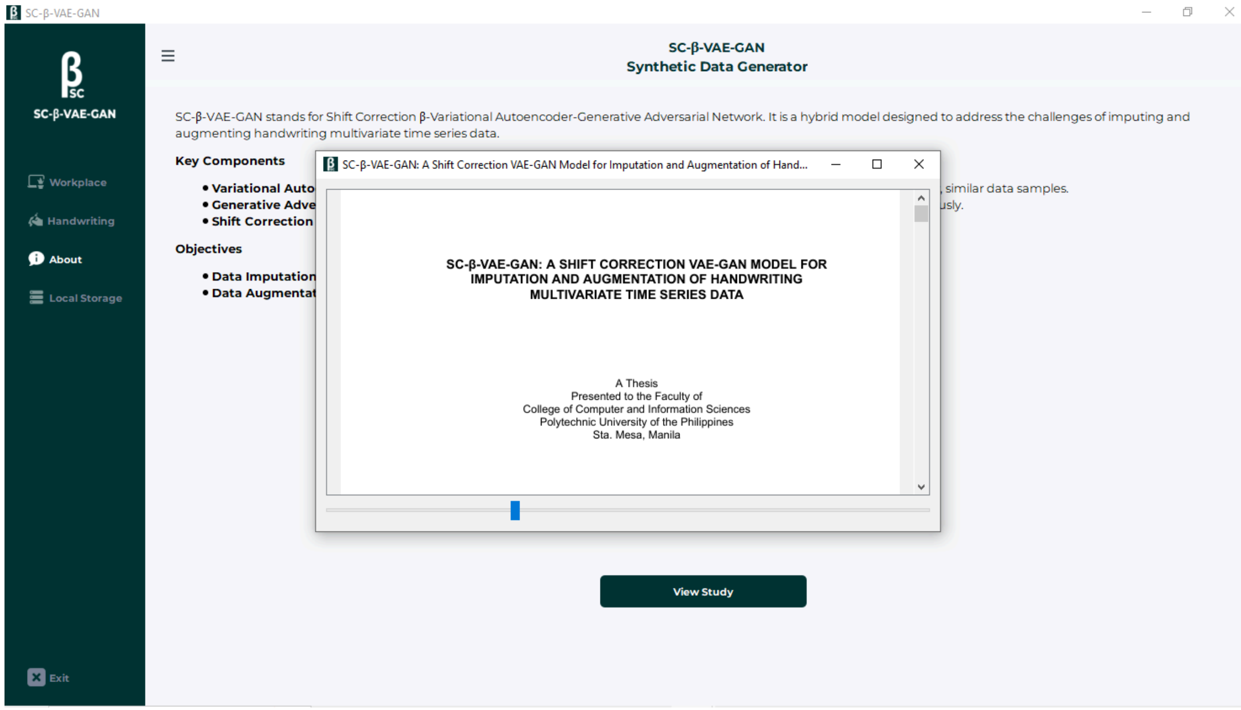

SC-β-VAE-GAN
SC-β-VAE-GAN
Synthetic Data Generator
Workplace
Handwriting
About
Local Storage
Exit
Input
Output
u00001s00001_hw00002.svc
Select Files
synthetic_u00001s00001_hw00002(3).svc
Select Files
49443 14255 0 1 1880 630 128
49443 14255 7 1 1880 630 143
49443 14255 15 1 1880 630 153
49443 14257 22 1 1880 640 237
49443 14257 30 1 1880 640 234
49443 14257 37 1 1880 640 239
49443 14257 45 1 1900 640 264
49443 14257 52 1 1900 640 249
49443 14257 60 1 1900 640 287
49443 14257 67 1 1900 640 283
49443 14257 75 1 1900 640 273
49443 14257 82 1 1900 640 342
49443 14257 90 1 1910 650 343
49443 14257 97 1 1910 650 379
49443 14257 105 1 1930 640 424
49443 14257 112 1 1930 640 415
49439 14274 0 1 1880 630 128
49439 14274 7 1 1880 630 143
49439 14274 15 1 1880 630 153
49439 14277 22 1 1880 640 237
49439 14277 30 1 1880 640 234
49439 14277 37 1 1880 640 239
49439 14277 45 1 1900 640 264
49439 14277 52 1 1900 640 249
49439 14277 60 1 1900 640 287
49439 14277 67 1 1900 640 283
49439 14277 75 1 1900 640 273
49439 14277 82 1 1900 640 342
49439 14277 90 1 1910 650 343
49439 14277 97 1 1910 650 379
49439 14277 105 1 1930 640 424
49439 14277 112 1 1930 640 415
Time Series Plot for u00001s00001_hw00002.svc
Time Series Plot for synthetic_u00001s00001_hw00002(3).svc
x (pen down)
y (pen down)
pressure (pen down)
azimuth (pen down)
altitude (pen down)
(pen up)
Values
Timestamp
0
10000
20000
30000
40000
50000
Calculating Results...

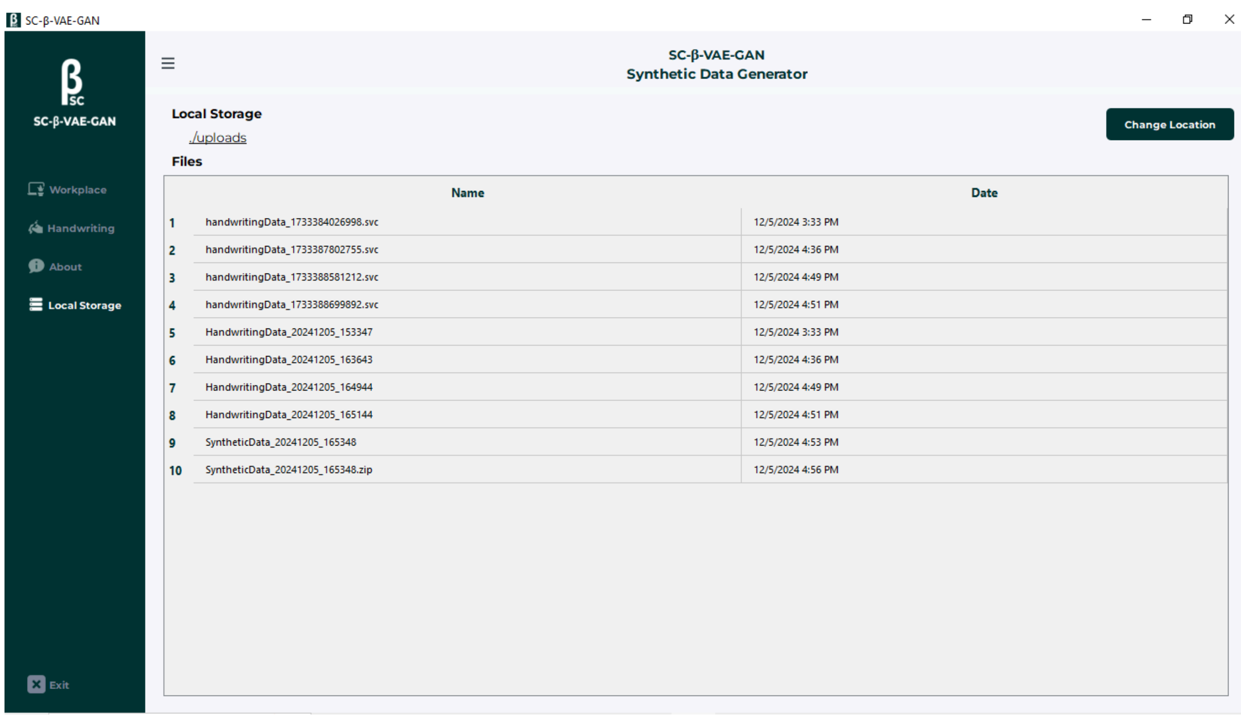

SC-β-VAE-GAN
SC-β-VAE-GAN
Synthetic Data Generator
SC-β-VAE-GAN
Workplace
Handwriting
About
Local Storage
Exit
Time Series Handwriting Visualization for collection1u00001s00001_hw00001.svc
Time Series Handwriting Visualization for synthetic_collection1u00001s00001_hw00001(4).sv
On Surface
In Air
x
y
x1 x2 y1 y2 time stamp1 time stamp2 pen status1 pen status2 azimuth1 azimuth2 altitude1 altitude2 pressure1 pressure2
48305 48232 32159 32197 796 795 0 0 1830 1830 530 530 0 0
48305 48234 32159 32194 975 802 0 0 1830 1830 530 530 0 0
NaN 48234 NaN 32194 NaN 810 NaN 0 NaN 1830 NaN 530 NaN 0
NaN 48234 NaN 32194 NaN 817 NaN 0 NaN 2340 NaN 350 NaN 0
NaN 48234 NaN 32193 NaN 825 NaN 0 NaN 2330 NaN 360 NaN 0
Calculating Results...


SC-β-VAE-GAN
SC-β-VAE-GAN
Synthetic Data Generator
SC-β-VAE-GAN
Workplace
Handwriting
About
Local Storage
Exit
On Surface
In Air
x
y
x1 x2 y1 y2 time stamp1 time stamp2 pen status1 pen status2 azimuth1 azimuth2 altitude1 altitude2 pressure1 pressure2
46252 48196 29902 31124 4309 3435 1 1 1650 1650 580 580 735 735
46251 48211 29902 31145 4316 3442 1 1 1650 1650 580 580 736 736
46249 48227 29902 31167 4324 3450 1 1 1650 1650 580 580 739 739
46248 48240 29902 31185 4331 3457 1 1 1650 1650 580 580 738 738
46248 48250 29902 31203 4339 3465 1 1 1650 1650 580 580 730 730
Normalized Root Mean Square Error (NRMSE)
Overall Average NRMSE: 0.0048
Post-Hoc Discriminative Score (PHDS)
Mean accuracy: 0.8333 (±0.0175)
Post-Hoc Predictive Score (PHPS)
Mean MAPE: 3.39%
Standard Deviation of MAPE: 0.43%
Generate Synthetic Data


SC-β-VAE-GAN
SC-β-VAE-GAN
Workplace
Handwriting
About
Local Storage
Exit
SC-β-VAE-GAN
Synthetic Data Generator
Draw and Handwrite
Click to Start


SC-β-VAE-GAN
SC-β-VAE-GAN
Workplace
Handwriting
About
Local Storage
Exit
SC-β-VAE-GAN
Synthetic Data Generator
SC-β-VAE-GAN stands for Shift Correction β-Variational Autoencoder-Generative Adversarial Network. It is a hybrid model designed to address the challenges of imputing and augmenting handwriting multivariate time series data.
Key Components
• Variational Autoencoder (VAE): A type of generative model that learns the underlying distribution of data to generate new, similar data samples.
• Generative Adversarial Network (GAN): A generative model consisting of two neural networks that are trained simultaneously.
• Shift Correction (SC): A method incorporated into the model to correct shifts in the data.
Objectives
• Data Imputation: Filling in missing values in multivariate time series data.
• Data Augmentation: Generating additional synthetic data to expand the available dataset.
View Study


SC-β-VAE-GAN
SC-β-VAE-GAN
Synthetic Data Generator
SC-β-VAE-GAN
Workplace
Handwriting
About
Local Storage
Exit
SC-β-VAE-GAN stands for Shift Correction β-Variational Autoencoder-Generative Adversarial Network. It is a hybrid model designed to address the challenges of imputing and augmenting handwriting multivariate time series data.
Key Components
• Variational Auto
• Generative Adve
• Shift Correction
similar data samples.
usly.
Objectives
• Data Imputation
• Data Augmentat
SC-β-VAE-GAN: A Shift Correction VAE-GAN Model for Imputation and Augmentation of Hand...
SC-β-VAE-GAN: A SHIFT CORRECTION VAE-GAN MODEL FOR IMPUTATION AND AUGMENTATION OF HANDWRITING MULTIVARIATE TIME SERIES DATA
A Thesis
Presented to the Faculty of
College of Computer and Information Sciences
Polytechnic University of the Philippines
Sta. Mesa, Manila
View Study


SC-β-VAE-GAN
SC-β-VAE-GAN
Synthetic Data Generator
SC-β-VAE-GAN
Workplace
Handwriting
About
Local Storage
Exit
Local Storage
./uploads
Change Location
Files
Name
Date
1 handwritingData_1733384026998.svc 12/5/2024 3:33 PM
2 handwritingData_1733387802755.svc 12/5/2024 4:36 PM
3 handwritingData_1733388581212.svc 12/5/2024 4:49 PM
4 handwritingData_1733388699892.svc 12/5/2024 4:51 PM
5 HandwritingData_20241205_153347 12/5/2024 3:33 PM
6 HandwritingData_20241205_163643 12/5/2024 4:36 PM
7 HandwritingData_20241205_164944 12/5/2024 4:49 PM
8 HandwritingData_20241205_165144 12/5/2024 4:51 PM
9 SyntheticData_20241205_165348 12/5/2024 4:53 PM
10 SyntheticData_20241205_165348.zip 12/5/2024 4:56 PM

Appendix 3: Implementation Report

Introduction

This study developed a data augmentation model using the VAEGAN framework with Beta and Shift Correction for imputation which has a name of SC-Beta-VAEGAN. The model is primarily designed to augment multivariate time-series data, specifically the Emothaw Dataset, which contains 903 online handwriting samples from 129 participants. Also, the main goal is to address challenges posed by limited and missing data in datasets used for handwriting recognition. The developed system then compares the SC-Beta-VAEGAN model with existing baseline models for time-series augmentation, such as VAEGAN, TimeGAN, and VRNNGAN, using various metrics including Post-Hoc Discriminative Score, Post-Hoc Predictive Score, and NRMSE to determine the quality of synthetic data generated.

Problem Statement

This study aimed to contribute to the study of data augmentation and the development of new approaches by creating a system that integrates augmentation and imputation into a single framework. To achieve this, the study developed a system that incorporates Shift Correction and Imputation within the VAEGAN architecture.

The study seeks to address the following challenges:

1. Limited resources for handwriting data due to the time-consuming process of data collection for handwriting.
2. Missing handwriting data caused by device limitations, which can negatively affect the quality of the dataset.

3. Lack of an augmentation model that integrates imputation into the augmentation process while considering the multivariate nature of handwriting data.

The study aimed to accomplish the following objectives:

1. Develop a system capable of augmenting and imputing handwriting multivariate time-series data by integrating Shift Correction and Beta into the VAEGAN architecture.
2. Incorporate Imputation and Shift Correction into data augmentation, which, to the researchers' knowledge, represents the first study to combine both approaches.
3. Compare the performance of SC-β-VAE-GAN with other existing augmentation models, such as VAEGAN, TimeGAN, and VRNNGAN, to evaluate the quality of the synthetic data generated and if there is a significant difference on the performance.

Datasets

This study utilizes the EMOTHAW dataset as its main dataset which are developed by Likforman-Sulem et al. (2017). It is a publicly available multivariate handwriting time-series dataset designed for emotional state recognition. It includes handwriting samples from 129 participants, aged 21 to 32 years (mean age 24.8), who completed seven handwriting tasks chosen based on psychological and medical tests. Data was collected using an INTUOS WACOM digitizing tablet, capturing parameters like x-y positions, timestamps, pen pressure, and pen movements (on and off paper). Ground truth labels for the handwriting data were provided using the Italian version of the Depression Anxiety Stress Scales (I-DASS-42) questionnaire. The dataset is available in SVC and XLS file formats. Also for validation, Greenland GPS Dataset is also test on the system to check if the model works with another time-series data, as it provides real-world geospatial time-series data with missing values, making it ideal for evaluating the model's data imputation capabilities.

Collection 1 User 3 Emothaw handwriting task 2 Sample data

u00003s00001_hw00002 - Notepad
File Edit Format View Help

```
3262
49040 15748 7376735 1 1970 610 25
49040 15748 7376742 1 1970 610 128
49040 15748 7376750 1 1970 610 178
49040 15748 7376757 1 1970 610 228
49040 15748 7376765 1 1950 610 262
49019 15748 7376772 1 1950 610 295
48981 15746 7376780 1 1950 610 325
48926 15741 7376787 1 1950 610 347
48853 15737 7376795 1 1950 620 382
48764 15733 7376802 1 1950 620 380
48661 15729 7376810 1 1950 620 394
48546 15726 7376817 1 1950 620 393
48420 15722 7376825 1 1950 620 391
48282 15718 7376832 1 1950 620 392
48133 15714 7376840 1 1940 620 384
```

GPS TENV3 Sample data

BLAS - Notepad
File Edit Format View Help

```
site YYMMMDD yyyy.yyyy __MJD week d reflon _e0(m) __east(m) ____n0(m) _north(m) u0(m) ____up(m) _ant(m) sig_e(m) sig_n(m)
sig_u(m) __corr_en __corr_eu __corr_nu _latitude(deg) _longitude(deg) __height(m)
BLAS 08JUL08 2008.5175 54655 1487 2  -23.0    512  0.674555   8833621  0.499398   484  0.005451  0.0000 0.000715 0.000867
0.004620  0.049702 -0.029873  0.430087  79.5386071327  -22.9747181354   484.00545
BLAS 08JUL09 2008.5202 54656 1487 3  -23.0    512  0.675501   8833621  0.497378   484  0.003480  0.0000 0.000714 0.000865
0.004647  0.039800 -0.019273  0.423218  79.5386071146  -22.9747180888   484.00348
BLAS 08JUL10 2008.5229 54657 1487 4  -23.0    512  0.676397   8833621  0.498661   484  0.002200  0.0000 0.000738 0.000891
0.004833  0.061547 -0.037363  0.407489  79.5386071261  -22.9747180445   484.00220
BLAS 08JUL11 2008.5257 54658 1487 5  -23.0    512  0.675272   8833621  0.498006   484  0.002578  0.0000 0.000728 0.000885
0.004793  0.046815 -0.031710  0.396252  79.5386071203  -22.9747181000   484.00258
BLAS 08JUL12 2008.5284 54659 1487 6  -23.0    512  0.676043   8833621  0.498672   484  0.006384  0.0000 0.000722 0.000899
0.004865  0.019606 -0.046881  0.481937  79.5386071262  -22.9747180620   484.00638
BLAS 08JUL13 2008.5311 54660 1488 0  -23.0    512  0.676751   8833621  0.498068   484  0.010199  0.0000 0.000761 0.000935
0.005132  0.074008 -0.013105  0.448013  79.5386071208  -22.9747180271   484.01020
BLAS 08JUL14 2008.5339 54661 1488 1  -23.0    512  0.676900   8833621  0.499213   484  0.006875  0.0000 0.000780 0.000980
0.005358  0.061688 -0.033201  0.471513  79.5386071311  -22.9747180197   484.00687
BLAS 08JUL15 2008.5366 54662 1488 2  -23.0    512  0.676978   8833621  0.498513   484  0.008276  0.0000 0.000796 0.000906
0.004955  0.072611 -0.005174  0.390575  79.5386071248  -22.9747180159   484.00828
BLAS 08JUL16 2008.5394 54663 1488 3  -23.0    512  0.674206   8833621  0.497145   484  0.004298  0.0000 0.000776 0.000929
0.005131  0.082076 -0.042347  0.411443  79.5386071125  -22.9747181526   484.00430
BLAS 08JUL17 2008.5421 54664 1488 4  -23.0    512  0.676286   8833621  0.498368   484  0.009644  0.0000 0.000762 0.000885
0.004791  0.093073 -0.072012  0.366798  79.5386071235  -22.9747180500   484.00964
BLAS 08JUL18 2008.5448 54665 1488 5  -23.0    512  0.675580   8833621  0.497995   484  0.009169  0.0000 0.000754 0.000911
0.004841  0.048877  0.017040  0.414951  79.5386071202  -22.9747180848   484.00917
BLAS 08JUL19 2008.5476 54666 1488 6  -23.0    512  0.673946   8833621  0.498332   484  0.009371  0.0000 0.000733 0.000883
0.004689  0.027906  0.015767  0.409376  79.5386071232  -22.9747181654   484.00937
BLAS 08JUL20 2008.5503 54667 1489 0  -23.0    512  0.676050   8833621  0.498865   484  0.006523  0.0000 0.000718 0.000872
0.004677  0.019815 -0.033769  0.424343  79.5386071279  -22.9747180616   484.00652
BLAS 08JUL21 2008.5530 54668 1489 1  -23.0    512  0.676886   8833621  0.498151   484  0.004547  0.0000 0.000757 0.000890
0.004929  0.017493 -0.055487  0.433170  79.5386071216  -22.9747180204   484.00455
BLAS 08JUL22 2008.5558 54669 1489 2  -23.0    512  0.674921   8833621  0.498670   484  0.004084  0.0000 0.000741 0.000901
0.004861  0.015175 -0.025342  0.427671  79.5386071262  -22.9747181173   484.00408
BLAS 08JUL23 2008.5585 54670 1489 3  -23.0    512  0.675862   8833621  0.497767   484  0.008385  0.0000 0.000755 0.000933
0.005011  0.050954 -0.016277  0.427005  79.5386071181  -22.9747180709   484.00839
```

Time Frame

Activity	Status	Date
Chapter 1-3 Documentation	Done	May 24, 2024
Development of the System	Done	July 20, 2024
Testing of the System	Done	October 10, 2024
Chapter 4-5 Documentation	Done	December 10, 2024

Implementation Procedures

1. Prepared the EMOTHAW dataset and performed data augmentation using the developed tool. The program repository is included below:

   https://github.com/Java-rice/Thesis-Project

2. Evaluated the synthetic data generated by the SC-β-VAE-GAN model using Normalized Root Mean Square Error, Post-Hoc Discriminative Score, and Post-Hoc Predictive Score.
3. Cleaned the baseline models, VAEGAN, VRNNGAN, and TimeGAN, so they could process the EMOTHAW Data.
4. Evaluated the synthetic data generated by baseline models, VRNNGAN, TimeGAN, and VAEGAN using Normalized Root Mean Square Error, Post Hoc Discriminative Score, Post Hoc Predictive Score.
5. Employed the Kruskal-Wallis H-test with Dwass-Steel-Critchlow-Fligner pairwise comparison to evaluate the stated hypothesis

H_0 - There is no significant difference between the performance (NRMSE, Post-Hoc Discriminative Score, Post-Hoc Predictive Score) of SC-*β*-VAE-GAN and VAE-GAN in terms of generating synthetic data.

H_0 - There is no significant difference between the performance (NRMSE, Post-Hoc Discriminative Score, Post-Hoc Predictive Score) of SC-*β*-VAE-GAN and TimeGAN in terms of generating synthetic data.

H_0 - There is no significant difference between the performance (NRMSE, Post-Hoc Discriminative Score, Post-Hoc Predictive Score) of SC-*β*-VAE-GAN and VRNNGAN in terms of generating synthetic data.

6. Performed cross-dataset validation experiments using the GPS dataset on SC-β-VAE-GAN to generate time series synthetic data and obtained the metrics NRMSE, Post-Hoc Discriminative Score, and Post-Hoc Predictive Score to measure how effective the system is on different data.
7. Employed a classification system to compare the performance of models trained on a combination of SC-β-VAE-GAN synthetic data and original data with models trained solely on original handwriting multivariate time-series data. Performance was measured using Precision, Recall, F1-Score, and Support to compare the classification model with synthetic data to the model without synthetic data.

Issues and Concerns

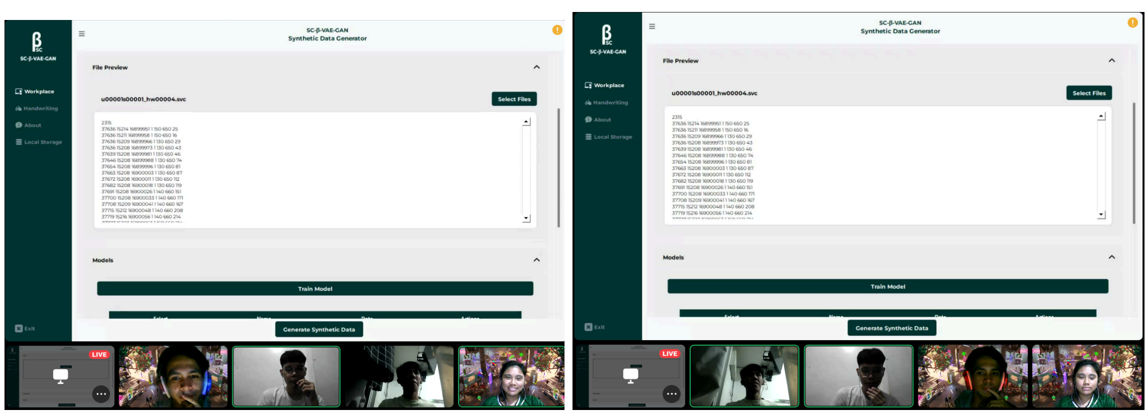
During the experimentation in our study, there are several challenges were encountered that impacted the experimentation process:

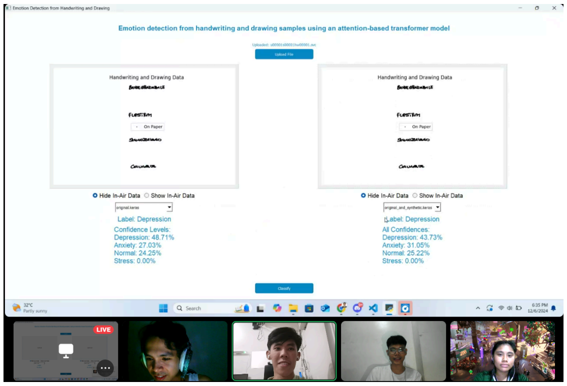
- Cleaning the baseline models to accept Emothaw data is time-consuming.

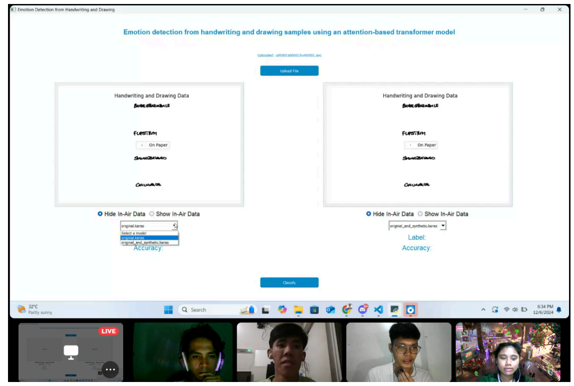
- Some baseline models take significant time to generate synthetic data; therefore, to balance efficiency with quality in testing, we set the parameters of all models at the same rate.
- The initial statistical tool, specifically the Chi-square test, was found to be inapplicable to our study, as confirmed by a statistician. This led to shifting into the suggested hypothesis testing tool, the Kruskal-Wallis test with pairwise comparison for each hypothesis.
- The classifier, based on the research "Emotion Detection from Handwriting and Drawing Samples Using an Attention-Based Transformer Model" by Khan et al., was initially unclear and not working due to an incomplete repository, so the system was studied and fixed based on their framework before testing it with SC-β-VAE-GAN-generated synthetic data.

Implementation Documentation

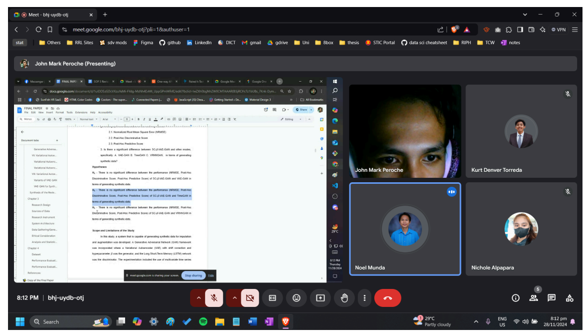
Experimentation







Statement of the Problem 3 (Statistical Tool) Consultation



Appendix 4: Revision Matrices

Thesis Proposal Revision Matrix

Polytechnic University of the Philippines
COLLEGE OF COMPUTER AND INFORMATION SCIENCES
Sta. Mesa, Manila

THESIS PROPOSAL DEFENSE REVISION MATRIX

PROPOSAL TITLE	SECTION	GROUP NO.
SC-β-VAE-GAN: A SHIFT CORRECTION VAE-GAN MODEL FOR IMPUTATION AND AUGMENTATION OF HANDWRITING MULTIVARIATE TIME SERIES DATA	BSCS 3-3	1

Document Part	COMMENTS/SUGGESTIONS/RECOMMENDATIONS by the Panel	ACTION TAKEN by the Proponent/s	Page No.	REMARKS (Panelist Approval by indicating the date) Panel 1	Panel 2	Panel 3	Panel 4
Title							
Preliminaries							
Chapter 1							
Chapter 2							
Chapter 3							
References							
Appendices							
General Observations	1. How will the synthetic data be validated? 2. Recommend implementing real-time handwriting input.	1. Validation of the tool and its output synthetic data will be based on its metric measurement (NRMSE, Post Hoc Predictive Score, Post Hoc Discriminative Score) and comparison with other baseline models. 2. The "Handwriting in Real Time" feature has been included in the screen mockup of the tool within the manuscript.	1. 69 2. 96				

Legend:
Name of Panelist 1: Lydinar D. Dastas
Name of Panelist 2: Iluminada Vivien R. Domingo
Name of Panelist 3: Angelica P. Payne
Name of Panelist 4:

CCIS Office N206 2nd Floor North Wing, PUP A. Mabini Campus, Anonas Street, Sta. Mesa, Manila 1016
Trunk Line: 335-1787 or 335-1777 local 272
Website: www.pup.edu.ph | Email: ccis@pup.edu.ph

THE COUNTRY'S 1st POLYTECHNICU

Tool Defense Revision Matrix

Polytechnic University of the Philippines
COLLEGE OF COMPUTER AND INFORMATION SCIENCES
Sta. Mesa, Manila

BSCS THESIS TOOL REVISION MATRIX

PROPOSAL TITLE	SECTION	GROUP NO.
SC-β-VAE-GAN: A SHIFT CORRECTION VAE-GAN MODEL FOR IMPUTATION AND AUGMENTATION OF HANDWRITING MULTIVARIATE TIME SERIES DATA	4-3	1

Proposal/Tool Part	COMMENTS/SUGGESTIONS/RECOMMENDATIONS by the Panel	ACTION TAKEN by the Proponent/s	Page No.	REMARKS (Panelist Approval by indicating the date) Panel 1	Panel 2	Panel 3	Panel 4
Title*							
Preliminaries							
Chapter 1*							
Chapter 2*							
Chapter 3*							
Tool (Software/Hardware)	1. **Increase the size of the canvas to make it larger** (Improvement). 2. **Merge the file preview features into one table in the result section for easier comparison.** (Improvement). 3. **Create a UI for validating generated synthetic data to enhance result interpretation** (Improvement)	**1 and 2.** Increased the canvas size for better visibility and merged file preview features into one table for easier comparison. **3.** A user interface was developed to validate generated synthetic data through classification testing.			10/21/24	ANGIE PAYNE	
References							
General Observations							

Name of Panelist 1: Lydinar D. Dastas
Name of Panelist 2: Iluminada Vivien R. Domingo
Name of Panelist 3: Angelica P. Payne
Name of Panelist 4:

***Revisions should be minimal in a way that it will not change the initially approved proposal and that proposal changes should NOT be accommodated in order to satisfy the inadequacies of the tool created.**

CCIS Office N206 2nd Floor North Wing, PUP A. Mabini Campus, Anonas Street, Sta. Mesa, Manila 1016
Trunk Line: 335-1787 or 335-1777 local 272
Website: www.pup.edu.ph | Email: ccis@pup.edu.ph

THE COUNTRY'S 1st POLYTECHNICU

Appendix 5: Ethical Clearance

Republic of the Philippines
POLYTECHNIC UNIVERSITY OF THE PHILIPPINES
OFFICE *of the* VICE PRESIDENT *for* RESEARCH, EXTENSION, *and* DEVELOPMENT
RESEARCH MANAGEMENT OFFICE
UNIVERSITY RESEARCH ETHICS CENTER

Date: November 08, 2024

To/For: **JOHN PATRICK D. LAGATUZ**
JOHN MARK P. PEROCHE
KURT DENVER P. TORREDA
NICHOLE N. ALPAPARA

Subject: Ethical Clearance

From: **Julie Charmain O. Bonifacio**
Chief, Research Ethics Section

This is to inform you that the documentary requirements you submitted for your research project titled **"SC-ß-VAE-GAN: A SHIFT CORRECTION VAE-GAN MODEL FOR IMPUTATION AND AUGMENTATION OF HANDWRITING MULTIVARIATE TIME SERIES DATA"** passed the evaluation of the PUP Research Ethics Committee (REC) in accordance with the requirements set by the Philippine Health Research Ethics Board (PHREB).

UREC Code	UREC-2024 - 2146
Type of Review	EXEMPTED
Approval Date	November 08, 2024
Expiry Date	November 07, 2025
PUP-UREC Decision	**Approved**

The standard conditions of this approval are as follows:

1. Conduct the project following the submitted and approved research protocol and other documentary requirements.
2. If changes are to be made in the research project/study that will affect the research participants, an amendment of the research protocol must be submitted to urec@pup.edu.ph before implementing such changes.
3. When ethical clearance is about to expire, the researcher must apply to resubmit the research protocol.
4. A final report/terminal report must be submitted when the research project/study is complete.
5. Researchers must advise the PUP-UREC (email: urec@pup.edu.ph) in writing if the research project/study has been discontinued.

You may now commence your research project/study. Good luck.

S423, 4th Floor South Wing, PUP A. Mabini Campus, Anonas Street, Sta. Mesa, Manila 1016
Trunk Line: 335-1787 or 335-1777 local 235/357
Website: www.pup.edu.ph | Email: vpredl@pup.edu.ph

THE COUNTRY'S 1st POLYTECHNICU

SOCOTEC ISO 9001

PAB ACCREDITED QMS CERTIFICATION BODY

ISO 9001:2015 CERTIFIED
CERTIFICATE NUMBER: SCP000413Q

Appendix 6: Biographical Statement

Nichole Alpapara is an undergraduate student pursuing the degree of Bachelor of Science in Computer Science at the Polytechnic University of the Philippines. She has a strong foundation in HTML, CSS, JavaScript, and ReactJS for web development. She is experienced in both frontend and backend development which she acquired from creating web applications and integrating APIs through her part-time job as a web developer. She also has certification in cybersecurity fundamentals and networking essentials, with hands-on experience gained through internships and collaborative team projects.

John Patrick Lagatuz is an undergraduate student pursuing the degree of Bachelor of Science in Computer Science at the Polytechnic University of the

Philippines. He has a passion for software development. Through internships and part-time work as a Software Developer, he has gained experience building and maintaining applications, handling both front-end and back-end tasks. He is proficient in languages like PHP, Python, Java, C, and Kotlin, which made him capable of designing software solutions, managing databases, and integrating user interfaces with backend systems.

John Mark Peroche is an undergraduate student pursuing the degree of Bachelor of Science in Computer Science at the Polytechnic University of the Philippines. He has a strong background in web and software development. He is skilled in multiple programming languages and frameworks, including Tailwind, JavaScript, and React. He has gained valuable experience through various projects, an internship, and a part-time job, where he has taken on roles such as Front-end Developer and Back-end Developer. In these positions, he has demonstrated proficiency in building user interfaces, developing applications, and integrating databases.

Kurt Denver Torreda is an undergraduate student pursuing the degree of Bachelor of Science in Computer Science at the Polytechnic University of the Philippines. He has a passion for Data Science. He has a good foundation in Python, machine learning and data analytics, which he has been able to apply to his internship and in various projects. He has experience in creating scripts for automated trading and customizing trading indicators using Pine Script. He also has a background in web and software development. He applied his skills in React, Tailwind, and JavaScript in many school projects and in his part-time job as a Front-end Developer.

Appendix 7

Consultation Document

This document certifies the review and approval of statistical methodologies used in SOP-3 for hypothesis testing. The objective of this analysis is to assess and compare the performance metrics of Sc-BetaVAEGAN, VAEGAN, TimeGAN, and VRNNGAN using valid statistical tools, specifically confirming that the **Kruskal-Wallis H-test with post hoc pair-wise comparison** is suitable and valid for evaluating the differences in performance metrics among the models. The previously considered Chi-square test was deemed inappropriate for this purpose.

Method Revisions:

1. **Previous Method Considered**: Chi-square test.
2. **Reason for Exclusion**: The Chi-square test is designed for categorical data and was found unsuitable for analyzing continuous performance metrics in this context.

Consultation Outcome: Upon consultation, it was concluded that the **Kruskal-Wallis H-test with post hoc pair-wise comparison** provides the most accurate and relevant results for hypothesis testing based on the nature of the data and study objectives.

I, the undersigned, have reviewed the statistical methodology used in SOP-3 and confirm the following:

1. The selection of the **Kruskal-Wallis H-test with post hoc pair-wise comparison** is appropriate for the hypothesis testing performed in SOP-3.
2. The Chi-square test, initially considered, was correctly identified as not applicable due to the data structure and research goals.
3. The applied methodology adheres to statistical standards and best practices.

By signing below, I approve the use of the **Kruskal-Wallis H-test with post hoc pair-wise comparison** as the statistical tool for SOP-3. This approval ensures the statistical rigor of the study and the validity of its conclusions.

Statistician's Information:

- **Name: NOEL P. MUNDA**
- **Title/Position: Master Teacher I and Research Consultant**
- **Affiliation: DepEd Mamatid National High School**

Signature over Printed Name:

NOEL P. MUNDA, LPT, PhD (c)
November 29, 2024