SYNTHETIC DATA-DRIVEN PREDICTIVE MAINTENANCE FOR COMPUTED TOMOGRAPHY SCAN MACHINE USING NEURAL NETWORK

By

MOHAMMAD HABIB SHAH ERSHAD BIN MOHD AZRUL SHAZRIL

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Master of Science

January 2025

FK 2025 1

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Master of Science

SYNTHETIC DATA-DRIVEN PREDICTIVE MAINTENANCE FOR COMPUTED TOMOGRAPHY SCAN MACHINE USING NEURAL NETWORK

By

MOHAMMAD HABIB SHAH ERSHAD BIN MOHD AZRUL SHAZRIL

January 2025

Chairman : Associate Professor Syamsiah binti Mashohor, PhD
Faculty : Engineering

The absence of abnormal data in predictive maintenance (PdM) for CT scan machines poses significant challenges, particularly in identifying crucial condition indicators, estimating health indices, and modeling degradation patterns necessary for predicting machine failures and remaining useful life. These limitations hinder the development of robust PdM models capable of effectively detecting anomalies and predicting failures in critical medical equipment. To address these challenges, this study aims to apply the Mahalanobis Distance (MD) method to analyze the real IoT sensor data collected from CT scan machines for detecting anomalies. A critical MD threshold value of 3.5485, corresponding to a 95% confidence level, was used to identify abnormal data points. Following the anomaly detection, synthetic abnormal data is generated using two methods which are Noise Addition (NA) and Tabular Generative Adversarial Networks (TGAN). The study aims to assess the reliability of the generated synthetic data and evaluate the performance of machine learning (ML) models in classifying anomalies for PdM in CT scan machines. An Artificial Neural Network (ANN) was employed to classify the normal and abnormal including real and

synthetic data. The results showed that the NA method achieved a classification accuracy of 98.64%, while the TGAN method demonstrated a higher accuracy of 99.53%. Additionally, TGAN-generated data closely resembled real-world anomalies, resulting in superior performance metrics such as 100% for precision, recall, and F1-scores, with fewer misclassifications. In conclusion, TGAN proved to be a more reliable approach for generating synthetic data and improving PdM model performance, particularly in situations where real abnormal data is limited. These findings highlight the importance of advanced data generation techniques like TGAN in enhancing the accuracy, reliability, and robustness of predictive maintenance systems for CT scan machines.

Keywords: Anomaly Detection, Mahalanobis Distance, Synthetic data, Noise, TGAN, ANN, Predictive Maintenance, IoT, CT scan machine, Machine Learning

SDG: GOAL 3: Good Health and Well-being, GOAL 9: Industry, Innovation and Infrastructure, GOAL 12: Responsible Consumption and Production

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Master Sains

RAMALAN PENYELENGGARAAN BERASASKAN DATA SINTETIK UNTUK MESIN IMBAS TOMOGRAFI BERKOMPUTER YANG MENGGUNAKAN RANGKAIAN NEURAL

Oleh

MOHAMMAD HABIB SHAH ERSHAD BIN MOHD AZRUL SHAZRIL

Januari 2025

Perngerusi : Profesor Madya Syamsiah binti Mashohor, PhD
Fakulti : Kejuruteraan

Ketiadaan data abnormal dalam ramalan penyelenggaraan (*PdM*) bagi mesin imbasan *CT* menimbulkan cabaran yang besar, terutamanya dalam mengenal pasti penunjuk keadaan yang kritikal, menganggarkan indeks kesihatan, dan memodelkan corak kemerosotan yang diperlukan untuk meramal kegagalan mesin dan jangka hayat yang kegunaan yang masih ada. Kekurangan ini menghalang pembangunan model *PdM* yang kukuh yang mampu mengesan kelainan dan meramal kegagalan dalam peralatan perubatan yang penting secara efektif. Untuk mengatasi cabaran tersebut, kajian ini bertujuan untuk mengaplikasikan kaedah Jarak Mahalanobis (*MD*) untuk menganalisis data sensor *IoT* sebenar yang dikumpul dari mesin imbasan *CT* untuk mengesan kelainan. Nilai ambang kritikal MD ditetapkan pada 3.5485, yang menggunakan tahap keyakinan 95%, untuk mengenal pasti titik data tidak normal. Selepas pengesanan kelainan, data sintetik tidak normal dijana menggunakan dua kaedah, iaitu Penambahan Hingar (*NA*) dan Rangkaian Adversarial Generatif Tabular (*TGAN*). Kajian ini bertujuan untuk menilai kebolehpercayaan data sintetik yang dihasilkan dan

menilai prestasi model pembelajaran mesin dalam mengklasifikasikan kelainan untuk *PdM* dalam mesin imbasan *CT*. Sebuah Rangkaian Neural Buatan (*ANN*) digunakan untuk mengklasifikasikan data normal dan tidak normal termasuk data sebenar dan sintetik. Keputusan menunjukkan bahawa kaedah *NA* mencapai ketepatan klasifikasi sebanyak 98.64%, sementara kaedah *TGAN* menunjukkan ketepatan yang lebih tinggi, iaitu 99.53%. Selain itu, data yang dihasilkan oleh *TGAN* lebih menyerupai kelainan dunia sebenar, yang menghasilkan metrik prestasi yang lebih unggul seperti ketepatan, kepekaan, dan skor F1 sebanyak 100%, dengan lebih sedikit kesilapan pengklasifikasian. Kesimpulannya, *TGAN* terbukti sebagai pendekatan yang lebih boleh dipercayai untuk menghasilkan data sintetik dan meningkatkan prestasi model *PdM*, terutama dalam situasi di mana data tidak normal yang sebenar terhad. Penemuan ini menekankan pentingnya teknik penjanaan data canggih seperti *TGAN* dalam meningkatkan ketepatan, kebolehpercayaan, dan kekuatan sistem penyelenggaraan ramalan bagi mesin imbasan *CT*.

Kata Kunci: Pengesahan Anomali, Jarak Mahalanobis, Data sintetik, Hingar, TGAN, ANN, Ramalan penyelenggaraan, IoT, Mesin imbasan CT, Pembelajaran mesin

SDG: MATLAMAT 3: Kesihatan Yang Baik dan Kesejahteraan, MATLAMAT 9: Industri, Inovasi, dan Infrastruktur, MATLAMAT 12: Penggunaan dan Pengeluaran Bertanggungjawab

ACKNOWLEDGEMENTS

Alhamdulillah, praise to Allah S.W.T for giving me the opportunity and strength to conclude this meaningful journey of my final year thesis even though many obstacles and challenges need to be faced.

I want to express a huge appreciation to my project supervisor, Associate Professor Dr. Syamsiah binti Mashohor and co-supervisor Professor Mohd Fadlee A. Rasid for allowing me to do the project and providing invaluable guidance along the journey. Their vision, sincerity and motivation have deeply inspired me. Working and studying under their supervision was a privilege and an honor. I am grateful for their acceptance and patience with me throughout the project and thesis preparation.

I would like to extend my sincere gratitude to the Ministry of Health Malaysia (Kementerian Kesihatan Malaysia - KKM) for their generous financial support in funding this project. Their contribution has been instrumental in enabling the successful completion of this research. The funding provided by KKM not only facilitated the acquisition of necessary resources but also allowed for the exploration of innovative approaches in predictive maintenance for CT scan machines. I am deeply thankful for their commitment to advancing healthcare technology and research, which has made this work possible.

I would like to express my sincere gratitude to En. Azizi Ali for his exceptional contribution as a team member on this project. His technical expertise, unwavering dedication, and valuable insights have played a significant role in the successful

completion of this research. En. Azizi's commitment to overcoming challenges and his collaborative spirit have been a source of motivation and inspiration throughout the journey. I am truly thankful for his hard work and support, and it has been a privilege to work alongside him on this project. I would like to express my heartfelt appreciation to Nur Fatinah Hafiz, a dear friend and dedicated group member, for her invaluable support and contributions throughout this project.

I would like to express my deepest appreciation to my team members for their hard work and dedication throughout the predictive maintenance project. Their technical expertise, insightful ideas, and collaborative spirit were instrumental in overcoming the challenges we faced. I am truly grateful for the time and effort they devoted to ensuring the project's success, and I look forward to future opportunities to work together.

Additionally, I would like to extend my deepest gratitude to the staff of the Department of Computer and Communication System Engineering who manage the postgraduate student affairs. Their continuous support, timely assistance, and diligent efforts in handling the administrative aspects of the program have greatly facilitated my journey. I sincerely appreciate their hard work in ensuring smooth processes, which allowed me to focus on my research and studies.

Lastly, I would like to express my deepest gratitude to my beloved family for their unwavering support, patience, and encouragement throughout this journey. I am forever grateful for your unconditional support, which has given me the strength to

keep moving forward. This achievement would not have been possible without each and every one of you. Thank you for all your encouragement!

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Master of Arts. The members of the Supervisory Committee were as follows:

Syamsiah binti Mashohor, PhD
Associate Professor
Faculty of Engineering
Universiti Putra Malaysia
(Chairman)

Mohd Fadlee bin A. Rasid, PhD
Professor
Faculty of Engineering
Universiti Putra Malaysia
(Member)

ZALILAH MOHD SHARIFF, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 8th May 2025

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
CGAN	Conditional Generative Adversarial Network
CM	Confusion Matrix
CNN	Convolutional Neural Network
CrM	Corrective Maintenance
CSV	Comma Separated Value
CT	Computed Tomography
CTGAN	Conditional Tabular Generative Adversarial Network
DICOM	Digital Imaging and Communications In Medicine
DL	Deep Learning
DPM	Digital Predictive Maintenance
DSS	Decision Support System
EKF	Extended Kalman Filter
ENBANN	Enhanced Naive Bayes Artificial Neural Network
FFE	First Failure Event
FMECA	Failure Modes and Effects Analysis
FN	False Negatives
FP	False Positives
FPR	False Positive Rate
FR	Failure Rate
GAAL	Generative Adversarial Active Learning

GAN	Generative Adversarial Network
ICU	Intensive Care Unit
IoT	Internet of Things
IR4.0	Industrial Revolution 4.0
KDE	Kernel Density Estimation
KL	Kullback-Leibler
KNN	K-Nearest Neighbour
LINAC	Linear Accelerator
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MD	Mahalanobis Distance
MLP	Multilayer Perceptron
MTFB	Mean Time Between Failures
MTTD	Mean Time to Detection
MTTR	Mean Time to Repair
NA	Noise Addition
NF	Noise Factor
NILM	Non-Intrusive Load Monitoring
NPV	Negative Predictive Value
PCA	Principal Component Analysis
PdM	Predictive Maintenance
PM	Preventive Maintenance
PPGAN	Privacy Preserving Generative Adversarial Network
PPV	Positive Predictive Value

PV-PP	Photovoltaic Power Plant
RMSE	Root-Mean Square Error
RNN	Recurrent Neural Networks
ROC	Receiver Operating Characteristic
RUL	Remaining Useful Life
SAX-HCBOP	Symbolic Aggregate Approximation – Histogram Coefficient Based Bag of Patterns
SCADA	Supervisory Control and Data Acquisition
SD	Synthetic Data
SME	Small and Medium Enterprise
SVM	Support Vector Machine
SWCVAE	Sliding-Window Convolutional Variational Autoencoder
TGAN	Tabular Generative Adversarial Network
TN	True Negatives
TP	True Positives
TPR	True Positive Rate
TSA	Time Series Analysis
TVAE	Transfer Variational Autoencoders
VMC	Vertical Machining Centre

CHAPTER 1

INTRODUCTION

1.1 Research Background

Computed Tomography (CT) scan machines are crucial medical imaging devices that use X-rays and computer technology to produce detailed cross-sectional images. They play a key role in modern healthcare by enabling accurate diagnoses, guiding treatment decisions, and supporting preventive medicine through early detection, especially in cancer screening (P. R. Patel & De Jesus, 2023). The CT scans generate images in the Digital Imaging and Communications in Medicine (DICOM) format, which allows for easy storage and sharing of medical data (see Figure 1.1), and improving diagnostic efficiency (Abdillah et al., 2017). For example, the CT imaging is widely used in hospitals which it hold great potential for early disease detection particularly in thorax body part (Shah et al, 2021).

The integration of Internet of Things (IoT) sensors in CT scan machines is crucial for improving maintenance strategies by providing real-time monitoring and data collection on critical machine parameters. These sensors continuously monitor variable factors such as machine performance and environmental conditions, offering valuable insights into the operational state of the equipment. This constant stream of data enables early detection of anomalies, allowing maintenance teams to address potential issues before they result in costly breakdowns or downtime (Ayvaz & Alpay, 2021; Ong et al., 2022). By leveraging the data gathered from IoT sensors, healthcare facilities can enhance the reliability of their CT scanners, ensuring uninterrupted

operation and optimal performance, which is essential in environments where equipment precision and uptime are critical.

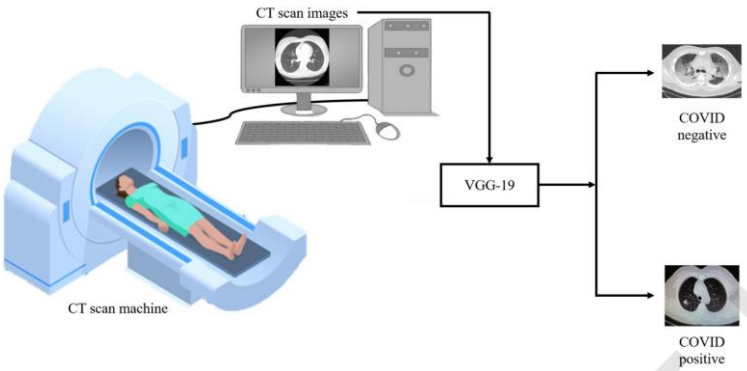


Figure 1.1: The flow diagram of the CT scan system screening a patient (Shah et al., 2021)**.**

The use of synthetic data (SD) is equally important in predictive maintenance (PdM), particularly when real-world data is scarce or imbalanced (Voronov et al., 2023). The SD can be generated to simulate machine behavior, especially in cases where abnormal or failure data is limited. This allows for the development and training of predictive models capable of recognizing rare or critical failure patterns. By incorporating SD, maintenance systems can be made more comprehensive and effective, even when actual failure occurrences are infrequent. The inclusion of SD helps ensure that PdM models can function accurately, leading to better predictions and improved system reliability (Patwardhan et al., 2016). This is especially valuable in industries like healthcare, where equipment failures are not only costly but can also impact patient care.

PdM is vital for ensuring the reliability and efficiency of CT scan machines by detecting issues before they lead to system failures. Unlike preventative maintenance,

which follows a strict schedule, PdM leverages real-time data from IoT sensors to monitor critical parameters allowing early detection of anomalies that signal potential problems (Einabadi et al., 2023). Machine learning models can analyze this data, classifying abnormalities and enabling timely interventions. By continuously monitoring both the machine and environmental conditions, such as room temperature and humidity, PdM reduces unscheduled downtime, optimizes machine performance, and extends the equipment's lifespan. This proactive approach is particularly important for medical industries that rely on CT scanners daily, as it minimizes the risk of failure while maintaining optimal operating conditions.

The combination of IoT sensors, SD and PdM techniques forms a powerful strategy for maintaining CT scan machines. IoT sensors provide continuous real-time monitoring of machine performance and environmental factors, while SD compensates for the lack of sufficient real-world failure data. The PdM leverages these data sources to predict and prevent potential failures, improving the overall reliability and efficiency of CT scan machines. By adopting these technologies, medical industries can ensure uninterrupted operation, minimize system downtime, and extend the lifespan of their equipment, ultimately enhancing patient care and operational efficiency.

1.2 Problem Statements

1. The absence of abnormal data in PdM for CT scan machines presents a significant challenge, as it restricts the ability to identify crucial condition indicators, estimate the machine's health index, and model degradation patterns essential for predicting failures and the remaining useful lifetime of the equipment (Fathi et al., 2021). Without abnormal data, accurately modeling failure patterns becomes difficult, resulting in ineffective maintenance strategies and an increased risk of machine breakdowns, which could disrupt critical medical services (Abdallah et al., 2023). Moreover, predictive models specifically for CT scan machines rely heavily on abnormal data to detect early signs of potential failure, making it challenging to train these models effectively when such data is unavailable (Filios et al., 2023). This problem occurs because routine maintenance practices often fail to capture data during malfunctioning states, focusing instead on data from normal operations, which limits the effectiveness of PdM strategies in capturing early warning signs of critical breakdown types, such as mechanical wear, electronic failures, or software glitches.

2. Even when abnormal data is available for PdM in CT scan machines, finding an effective method to balance the dataset and mimic real-world conditions remains a significant challenge. Due to the typically imbalanced nature of CT scan machine maintenance data, with abnormal instances being rare, conventional methods struggle to create a synthetic yet representative distribution of abnormal data. Ensuring that generated SD accurately reflects

the complex patterns of real operational anomalies specific to CT scan machines is critical for training robust machine learning models. However, many existing techniques fail to replicate the nuanced variations found in actual abnormal data from CT scan machines, leading to less reliable predictions and a higher risk of undetected failures. Additionally, selecting suitable SD generation models for CT scan machines is challenging due to their varying strengths and weaknesses, which may lead to discrepancies between synthetic and real data in specific applications (He, 2024). This makes the process of effectively balancing and modeling abnormal data an ongoing issue in PdM for CT scan machines. Creating synthetic data becomes essential to simulate these rare but critical abnormal conditions, aiding in the enhancement of maintenance strategies by allowing more comprehensive training of predictive models.

3. Selecting an appropriate PdM model for CT scan machines is a complex task due to the diverse characteristics of the dataset, such as the variability in sensor data and the presence of anomalies. Different machine learning algorithms excel in different scenarios, and the choice of model depends on the specific learning objectives, such as whether the focus is on fault detection, remaining useful life prediction, or anomaly classification for CT scan machines. Moreover, there is a constant trade-off between achieving high accuracy and maintaining interpretability, particularly in high-stakes environments like healthcare, where understanding model outputs is crucial (Arena et al., 2024). This makes it imperative to carefully assess and select models that not only perform well but also provide actionable insights for PdM of CT scan

machines. Regular maintenance practices, typically involving routine checks and component replacements, often overlook the integration of advanced predictive analyses, which could preemptively address these failures by anticipating them before they result in operational downtime.

1.3 Aim and Objectives

The aim of this project is to assess the reliability of synthetic abnormal data generated for machine learning models. The objectives includes:

1. To analyze the real data and generate synthetic abnormal data using the Mahalanobis Distance (MD) method for effective anomaly detection.
2. To generate synthetic abnormal data from real IoT sensor data collected from CT scan machines using Noise Addition (NA) technique and Tabular Generative Adversarial Networks (TGANs).
3. To design and evaluate the performance of the Artificial Neural Network classification model of anomalies data for PdM of CT scan machines.

1.4 Scope of Work

Throughout 2023, sensor data from a CT scan machine environment was collected monthly for PdM purposes. The dataset includes six key features which are electrical current, radiation levels, temperature, humidity, and vibration data from the X and Y axes of the CT scan donut. Various IoT sensors were used for data collection, including the Schneider PM5110 for current, the RAD-S101 for radiation, the AM103 for temperature and humidity, and the MK926 MEMS for vibration. The data was

captured on a monthly basis to monitor machine performance, with time interpolation techniques employed to address any missing values. The project will be implemented using Python 3.9.18 on an ASUS TUF F17 laptop with an Nvidia GTX 1650 Ti GPU and an AMD Ryzen 7 processor. The following versions of essential libraries will be utilized: pandas 1.4.2 for data manipulation and preprocessing, matplotlib 3.5.2 and seaborn 0.11.1 for data visualization, numpy 1.26.4 and scipy 1.13.1 for numerical computations and scientific functions, and tensorflow 2.9.0 for building deep learning models. The tabgan 2.0.5 library, specifically tabgan.sampler, will be used to generate synthetic data, and scikit-learn (sklearn) 1.1.1 will be employed for machine learning algorithms and model evaluation. Additionally, Jupyter 1.0.0 will be used for interactive development and documentation.

The scope of work for this project encompasses using MD for anomaly detection, NA and TGAN for generating synthetic data, and ANN for model prediction. The project will initiate with data preparation, involving collection and preprocessing of data from industrial sensors to ensure quality and consistency. Following this, MD will be implemented to identify anomalies in the dataset, with established thresholds to enhance detection accuracy. To address data limitations, particularly in underrepresented conditions, NA will be applied to existing data points to simulate variability, and TGAN will be employed to generate synthetic data that mirrors the statistical properties of the original dataset. The synthetic data will be used to train an ANN, designed with specific layers and activation functions suitable for the data characteristics. The ANN will undergo training and validation, using techniques like cross-validation to ensure the model generalizes well on unseen data. Model performance will be evaluated using metrics such as accuracy, precision, and recall.

The deliverables will include a fully functional anomaly detection system, a synthetic data generator, a trained predictive model, comprehensive documentation, and a final report summarizing methodologies and findings. This approach ensures a good predictive maintenance system that enhances operational reliability and efficiency.

1.5 Contribution of Thesis

The contributions of this thesis are multi-faceted and aimed at advancing PdM for CT scan machines using machine learning and SD generation. First, this research introduces a robust approach for detecting anomalies in IoT sensor data collected from CT scan machines, leveraging MD to accurately classify normal and abnormal conditions. Additionally, the thesis explores two methods for generating synthetic real abnormal data which are NA and TGAN. By employing these methods, the research addresses the issue of data imbalance. Finally, the study demonstrates the effectiveness of both data generation techniques by evaluating them through an ANN model, measuring performance metrics such as accuracy, precision, recall, F1-score and others. This research ultimately contributes to improved PdM models for CT scan machines, ensuring early detection of potential failures and enhancing machine reliability.

1.6 Thesis Outline

This thesis explores the generation of reliable SD for PdM of CT scan machines. Chapter 1 introduces the study, outlining its objectives and significance. Chapter 2 reviews relevant literature on PdM, SD generation techniques, anomaly detection

methods using MD, and the application of ANNs. Chapter 3 details the research methodology, including data collection, anomaly detection, SD generation and ANN development. Chapter 4 presents the results and discussion, comparing the effectiveness of the two SD generation methods and their impact on ANN performance for PdM. Finally, Chapter 5 concludes the study, summarizing findings, contributions, and recommendations for future research.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This literature review has been conducted to gather information regarding related works on predictive maintenance (PdM) using IoT sensors and synthetic data (SD) creation. The review of previous works includes all aspects of the methodology related to numerical data processing for PdM. This chapter also reviews topics such as the datasets, pre-processing, SD generation, and classification.

2.2 Predictive Maintenance on Industrial Machine

In this era Industrial Revolution 4.0 (IR4.0), most industries utilized machines to enhance the quality and productivity of their products. The use of machines in industry can save time and reduce the costs of hiring workers. However, when these machines break down, it can hinder the growth of an industry. According to Abdelillah et al. (2023), IR4.0 has revolutionized manufacturing processes. This transformation is achieved by utilizing advanced sensing and data analytics technologies. These technologies enhance efficiency in manufacturing, with PdM playing a vital role. PdM is crucial in ensuring the reliability and availability of production systems. Therefore, some industries have adopted PdM as a measure to reduce the impacts caused by machine breakdowns. PdM is a proactive approach that aims to predict potential equipment failures before they occur, enabling timely interventions and minimizing unplanned downtime (Alves et al., 2020). For example, Mourabit et al. (2020)

developed a PdM model for semiconductor failures using a Multilayer Perceptron (MLP) algorithm, achieving a good accuracy of 91.95%. Other than that, industries like oil and gas also use PdM to ensure the machines and components are always in the best condition. According to the research (Barrera et al., 2022), the clustering algorithms and autoencoder was used for fault detection of gas turbines, achieving 96.45% accuracy for the test dataset and 94.88% for the synthetic dataset. PdM has revolutionized various sectors by enabling proactive management of equipment and systems, ensuring optimal performance and minimizing unexpected failures. This technological advancement is particularly transformative in the medical industry, where the reliability of diagnostic and therapeutic devices is critical to patient care. The integration of PdM with the industrial Internet of Things (IoT) and advanced analytics has further amplified its impact in the medical industry (Bukhsh & Stipanovic, 2020). Table 2.1 shows the example of PdM in the medical industry. Through these examples, it is evident that PdM not only enhances operational efficiency but also supports broader global objectives of health, innovation, and sustainability.

Table 2.1: Table 2.1: PdM in the Medical Industry

Authors/Years	Problem statement	Purpose	Performance metrics	Limitation
(Zamzam et al., 2021)	There is a lack of comprehensive strategic maintenance models that integrate preventive, corrective, and replacement maintenance strategies with predictive models using a large medical equipment dataset.	To develop a prioritisation assessment and predictive system for strategic maintenance management of medical equipment in public healthcare facilities.	SVM got highest accuracy 99.42% for preventive maintenance (PM) and 99.80% for replacement programme. K-Nearest neighbour (KNN) got highest accuracy for corrective maintenance (CrM) which is 98.93%.	A limitation of the study is the need for manual intervention in determining criteria weightages for reliability assessments, which introduces subjectivity and potential inconsistencies in the prioritization process.
(Ahmed et al., 2021)	The subjectivity and uncertainty in expert-driven decisions for healthcare facility maintenance prioritization pose challenges, necessitating a more objective, systematic, and automated prioritization framework.	To develop a systematic and automated approach for maintenance prioritization in healthcare facilities using machine learning and Neutrosophic logic.	AUC-ROC for Decision Tree: 0.917, Accuracy for Decision Tree: 0.90, AUC-ROC for KNN: 0.947, Accuracy for KNN: 0.86, AUC-ROC for Naïve Bayes: 0.855.	The prioritization process still involves some level of subjectivity, as expert input is needed for criteria weight assignment, even though the model attempts to minimize it with neutrosophic logic.

Table 2.1: Continued

Authors/Years	Problem statement	Purpose	Performance metrics	Limitation
(Kim et al., 2022)	Predicting downtime for medical linear accelerators (LINACs) is challenging due to complex failure causes, and sudden machine failures can disrupt patient treatments.	To predict downtime and maintain the optimal condition of a medical LINAC using a 20-year maintenance dataset and time series analysis.	The autoregressive integrated moving average (ARIMA) model predicted downtime with a mean absolute percentage error (MAPE) of 3%, indicating a stable and reliable model.	The ARIMA model used for time series prediction does not handle missing values. well and assumes linear relationships between data points, which may not always represent real-world scenarios accurately.
(Li et al., 2022)	The current maintenance and quality control processes for medical equipment lack integration and fail to fully utilize the potential of data, leading to frequent equipment malfunctions and reduced operational efficiency in hospirtal.	To enhance the maintenance and quality control of medical equipment using information fusion technology to ensure equipment reliability, maintainability, and safety in hospitals.	The performance was measured through the pass rates of equipment quality control tests before and after maintenance, with results showing a significant improvement in pass rates, demonstrating the effectiveness of the proposed system.	The system is still immature and has not been fully applied to major medical systems, requiring further improvements for widespread implementation.

Table 2.1: Continued

Authors/Years	Problem statement	Purpose	Performance metrics	Limitation
(Abd Rahman et al., 2023)	The current healthcare maintenance practices focus on corrective and preventive maintenance, which leads to high operational costs and frequent device failures, highlighting the need for a predictive maintenance strategy to improve efficiency.	To develop a predictive model using ensemble classifier for medical device failure to improve maintenance strategies and reduce costs in healthcare facilities.	The ensemble classifier achieved the best results with 79.50% accuracy, 88.36% specificity, and 77.43% precision after optimization, outperforming other machine learning and deep learning models.	The model relies on structured data from existing medical devices, and the inclusion of unstructured data like maintenance notes is suggested as future work to further improve prediction accuracy
(Zamzam et al., 2023)	Medical equipment failures can disrupt healthcare services and jeopardize patient safety, while traditional reactive and preventive maintenance approaches areinefficient and costly, highlighting the need for predictive maintenance systems.	To develop failure analysis predictive models for medical equipment, focusing on predicting the first failure event (FFE), failure-to-year ratio (FYR), and failure rectification group (FRG), to support smart maintenance practices in healthcare.	The SVM model for FFE achieved the highest accuracy of 96.9%, the DT model for FYR achieved 83.9% accuracy, and the ANN model for FRG achieved 76.7% accuracy.	The study's predictive models may be affected by imbalanced datasets, and further work is needed to ensure better dataset integrity and to integrate the models with existing hospital systems for real-time analysis.

Table 2.1: Continued

Authors/Years	Problem statement	Purpose	Performance metrics	Limitation
(Prasetya et al., 2023)	The unpredictability of machine breakdowns, particularly in medical mask machines, leads to frequent failures of different components, causing delays and inefficiencies in the production process.	To implement the Weibull analysis method in designing PdM for a medical mask machine to reduce downtime and random failures.	The mean time between failures (MTBF) and the failure rate (FR) are used as performance metrics, with an error of 1.8% between automated and manual calculations.	The Weibull analysis method requires highly specific and ideal initial conditions for accurate predictions. In real-world scenarios, it can be difficult to obtain such ideal data, which may limit the effectiveness of the PdM.
(Rahman et al., 2023)	Medical device failures and high maintenance costs, coupled with a lack of strategic maintenance frameworks, necessitate the development of predictive systems to mitigate unplanned downtime and maintenance costs.	To propose a predictive model for medical device failure using Artificial Intelligence (AI) driven analysis of multimodal maintenance records, integrating structured and unstructured data.	The optimized Ensemble Classifier achieved the highest accuracy of 88.80%, with 94.41% specificity, 88.82% recall, 88.46% precision, and an F1 score of 88.84%.	The predictive model uses historical data for predictions, but the study does not address the integration of real-time data from devices, which could enhance predictive accuracy and timeliness.

Table 2.1: Continued

Authors/Years	Problem statement	Purpose	Performance metrics	Limitation
(Gallab et al., 2024)	complexity, high costs, and the risk of disruptions in patient care. Current maintenance practices often lead to unplanned downtime and high costs, highlighting the need for a more proactive, data-driven approach.	To propose a new Digital Predictive Maintenance (DPM) framework to facilitate the transition from traditional maintenance to predictive maintenance, with a particular focus on healthcare facilities, aiming to improve equipment reliability, reduce downtime, and optimize resource allocation.	The study evaluates the proposed DPM framework based on its ability to reduce Mean Time to Detection (MTTD), minimize Mean Time to Repair (MTTR), and lower system downtime through real-time monitoring and predictive capabilities.	The reliance on interconnected digital systems and cloud storage increases the risk of data breaches and cyberattacks, especially given the sensitive nature of healthcare data.

2.3 Computed Tomography Scan Machine

A CT scan machine is a vital medical equipment used for non-invasive diagnostic imaging, providing detailed cross-sectional images of anatomical structures with high clarity and contrast resolution (Saad & Moore, 2022; X. Zhou & Liu, 2024). CT scanner has evolved significantly over the years, with key technological advancements shaping its history. The development of CT scanners has seen quantum leaps in the dimensionality of clinical CT data, with innovations like helical spiral data acquisition, multi-slice CT, wide-cone CT, dual-source CT, and spectral CT contributing to improved performance in terms of volume coverage, temporal resolution, and material differentiation capabilities (Hsieh & Flohr, 2021; Taguchi et al., 2022). CT imaging, a non-invasive diagnostic technique, utilizes various signals to provide high-sensitivity cross-sectional scans, offering fast scanning times and clear images for the examination of various diseases in both medical and industrial settings (X. Zhou & Liu, 2024). These machines utilize rotating X-ray beams and detectors to capture images, exposing patients to higher levels of ionizing radiation compared to standard radiography (Saad & Moore, 2022). The first CT scanner was developed in the early 1970s, leading to enthusiastic adoption until the late 1970s. This was followed by regulatory challenges in the 1980s due to perceived overuse, and subsequent innovations reigniting growth in the 1990s and 2000s (Bhidé et al., 2020). Each component inside a CT scan machine is crucial, a malfunction in any part can result in significant downtime and substantial costs. Figure 2.1 shows the use of CT scan machines increased every year especially during pandemic Covid 19.

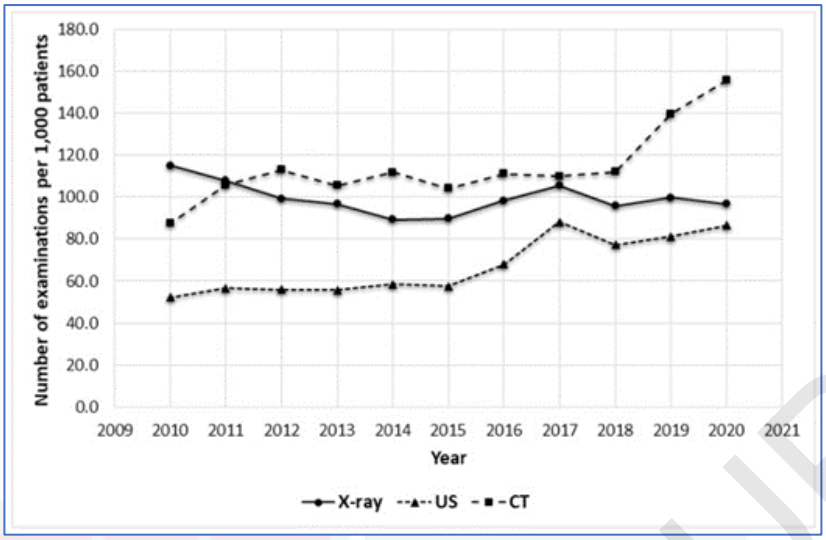


Figure 2.1: Annual change of CT scan machine usage (Winder et al., 2021)**.**

Based on research H. Zhou et al. (2024), a model has been developed for predicting anomalies in CT equipment using real-time status data, aimed at assisting hospitals in making accurate maintenance decisions. The Symbolic Aggregate Approximation Histogram Coefficient based Bag of Patterns (SAX-HCBOP) model, applied on the IoMT dataset, achieved an accuracy of 0.904, recall of 0.747, precision of 0.417, and an F1-score of 0.535. Common maintenance tasks for CT scan machines include electrical safety checks, physical checks, performance checks, and functional checks such as validating transient current detectors and thermal cutouts (Kwan et al., 2020).

According to Adachi (2008), common maintenance malfunctions for CT scan machines include issues with the gantry unit, reconstitution unit, and control unit. Problems related to the X-ray tube, detector, and abnormality detection mechanisms can also occur, impacting the image quality and diagnostic capabilities of the machine (Yuichi, 2007). However, the quality of the sensor data was low, which poses a limitation as the current model might not be as effective if data quality improves.

Therefore, PdM is essential to preemptively address potential malfunctions, ensuring consistent operation and minimizing costly disruptions.

2.4 Internet of Thing Sensors and Parameters

The evolution from traditional maintenance strategies like reactive and preventive maintenance to PdM has been driven by the need for enhanced operational efficiency and reduced downtime in industrial systems (Hector & Panjanathan, 2024; Mallioris et al., 2024; Postiglione & Monteleone, 2024). The integration of IoT sensors and data analytics has paved the way for a proactive, data-driven approach to maintenance, shifting away from the traditional reactive and preventive strategies. IoT-enabled sensors have become the backbone of PdM, enabling continuous monitoring of critical equipment and machinery. These sensors collect a wealth of real-time data, including vibration, temperature, pressure, and other operational parameters, which are then transmitted to cloud-based platforms for analysis (Alves et al., 2020; Wahid et al., 2022). Table 2.2 shows the sensors used and the selected parameters for PdM while Figure 2.2 provides a summary of the most selected parameters for conducting PdM in the year 2023. The integration of IoT sensors for PdM poses various challenges that need to be addressed. According to Kazmi et al. (2023) one of the challenges in using IoT sensors for PdM is the efficient acquisition, processing, and transmission of large amounts of sensor data.

This challenge is addressed by utilizing a hybrid approach that involves edge processing to reduce the burden on centralized cloud platforms, as seen in the study on IoT sensor nodes for industrial motors. Additionally, the need for sophisticated data

analytics techniques, such as machine learning algorithms, to handle the massive amounts of sensor data generated by IoT devices is highlighted as a key challenge in PdM applications (Omol et al., 2024). Furthermore, the integration of deep learning methods for PdM in Industrial IoT settings poses challenges related to data quality, interpretability, scalability, and ethical concerns, emphasizing the need for advanced techniques and strategies to overcome these obstacles (Jeba Emilyn. et al., 2024). The integration of IoT sensors for PdM enhances operational efficiency and reduces downtime but it also presents challenges in data management and analytics that necessitate advanced techniques and strategies to overcome.

Table 2.2: IoT sensor parameters used for PdM.

Authors	Sensor Parameter	Purpose
(Taşcı et al., 2023)	Weight, speed, temperature, electric current, vacuum, air pressure	To develop a machine learning-based predictive maintenance approach that can predict the Remaining Useful Life (RUL) of machines on production lines in a manufacturing environment.
(Pagano, 2023)	Temperature, vibration	To propose a new approach for predictive maintenance in an industrial plant. The method combines LSTM neural networks with Bayesian inference to monitor and assess the health of the plant.
(R. Patil et al., 2023)	Temperature, vibration, rotational speed, torque, tool wear, process temperature	To propose a system for predictive maintenance using machine learning techniques, applied on Vertical Machining Centre (VMC). The system focuses on monitoring the health of industrial machines by collecting and analyzing sensor data to predict potential failures before they occur.

Table 2.2: Continued

(Achouch et al., 2023)	Voltage, pressure, flow, humidity, temperature	To apply machine learning algorithms to predict failures and estimate the RUL of a TA-48 multistage centrifugal compressor used in a plant.
(Shrivastava et al., 2023)	Current, spindle speed, vibration	Use Enhanced Naive Bayes Artificial Neural Network (ENBANN) to present a PdM model that leverages sensor data and machine learning techniques to improve maintenance in industrial settings. This approach was applied is a cutting machine used in the firewood industry.
(Gupta et al., 2023)	Vibration, acceleration	Develop a scalable and economical maintenance 4.0 solution for airport baggage handling systems using IoT sensors and machine learning.
(X. Wang et al., 2023)	Voltage, current, torque, temperature, pressure, resistance	Develop a PdM system combining data-driven and knowledge-based methods to enhance the operational efficiency of industrial robots. This research was applied to welding robots in a new energy automotive welding workshop, where these robots are used extensively in production lines for automobile manufacturing.

Table 2.2: Continued

(Fordal et al., 2023)	Temperature, vibration, power usage	To develop a predictive maintenance platform that integrates sensor data and ANN to improve maintenance and decision-making in industrial environments, it was applied on splitting saw machine.
(Fordal et al., 2023)	Temperature, vibration, power usage	To develop a predictive maintenance platform that integrates sensor data and ANN to improve maintenance and decision-making in industrial environments, it was applied on splitting saw machine.
(Burmeister et al., 2023)	Multiple sensors, historical data	Demonstrate the potential of utilizing production data alone to predict future maintenance and validate the use of interpretable machine learning models for root cause analysis, not specific on what machine.
(Hadi et al., 2023)	Vibration	Develop a fault classification model for PdM using AutoML techniques to improve accuracy and reduce manual intervention on ball bearings.

Table 2.2: Continued

(Raparthi et al., 2023)	Temperature, Pressure, Vibration, error log, Maintenance records	Develop a PdM framework for IoT devices using Time Series Analysis (TSA) and Deep Learning (DL) to enhance reliability and operational efficiency, not specify on any machine.

Figure 2.2 illustrates the key parameters selected for PdM based on the literature reviewed in Table 2.2. The radar chart highlights two critical factors: Temperature and Vibration, which are emphasized due to their frequent selection and significant impact on machine health. Temperature is essential for detecting overheating or abnormal thermal conditions, which can signal potential mechanical or electrical failures. Vibration is another crucial factor, as abnormal vibrations often indicate mechanical issues such as misalignment, imbalance, or wear in rotating equipment. These two parameters are commonly chosen in PdM strategies due to their reliability in providing early signs of equipment degradation, thus enhancing the ability to perform timely maintenance and avoid unexpected breakdowns.

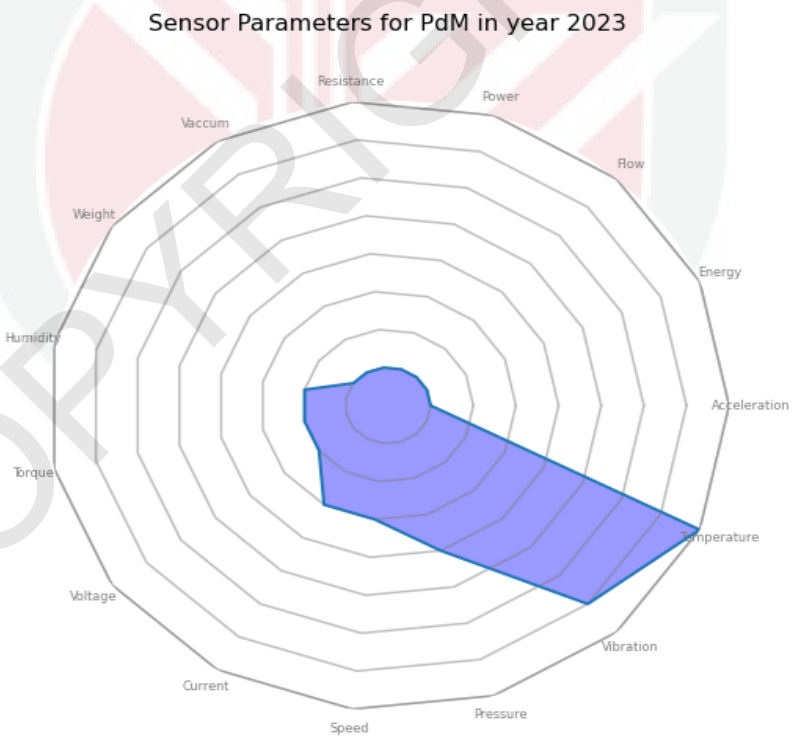


Figure 2.2: Summary of selected sensor parameters in PdM for industrial machine for year 2023.

2.5 Data Pre-processing

Data preprocessing is the process of transforming raw data into a clean and usable format by handling missing values, normalizing, transforming, and encoding data. Data preprocessing is a crucial step in various research fields, including oceanography and medical analytics, involving the preparation of data for further analysis (Ashwini Tuppad & Shantala Devi Patil, 2023; Panalaran et al., 2024). It encompasses tasks like data cleaning, transformation, and reduction to ensure that the data is accurate, consistent, and suitable for computational algorithms (Varma et al., 2023). The importance of data preprocessing techniques also to address imbalanced data, showing a significant improvement in forecasting accuracy when rare events are considered (Pisa et al., 2019). Benhar et al. (2020) conducted a systematic literature review on data preprocessing for heart disease classification, analyzing the impact of preprocessing techniques on classification performance and comparing different classifier-preprocessing combinations in terms of accuracy rate. Other than that, Fan et al. (2021) provided an in-depth review of existing preprocessing methods for efficient knowledge discovery from building operational data, discussing potential applications in smart building energy management.

In PdM also data pre-processing plays an important role especially when dealing with sensor data that often contains missing values and noise (Latyshev, 2018). One example Jiang et al. (2022) employed feature fusion to combine multiple categories of electrical attributes into a single feature using fused feature F as formulated as (2.1), in this case data preprocessing used to predict remaining useful life of equipment. The

f represents the feature of each category (e.g., current, voltage, power, power factor, frequency), and Ψ denotes the concatenation operation.

$$F = \Psi(f1, f2, \dots, fC) \tag{2.1}$$

The research by Elkateb et al. (2024) highlights the use of data normalization and feature selection in data pre-processing processes to identify outliers and enhance model performance through significant feature identification. This z-score normalization was applied in a PdM approach for industrial applications using (2.2).

$$D_j' = \frac{D_j - \underline{D_j}}{\sigma_{D_j}} \tag{2.2}$$

The formula provided is used for z-score normalization, a common technique in data preprocessing. In this context, Dj' represents the normalized value of the data point within the data category j. Dj is the original data point i in the same category j, $\underline{D_j}$ denotes the mean (average) of all data points in category j and σ_{D_j} stands for the standard deviation of the data points in category j, which measures the dispersion of the data points around the mean. This z-score normalization process adjusts the values of data points so that they have a mean of 0 and a standard deviation of 1 within their category. According to (Henderi, 2021), MinMax scaling normalization is another common technique used in data preprocessing. It transforms data values into a specific range, typically between 0 and 1, using the (2.3).

$$v_{norm} = \frac{v_i - v_{min}}{v_{max} - v_{min}} \tag{2.3}$$

Where v_{norm} is the normalized value, v_i is the actual data point, and v_{min} and v_{max} are the minimum and maximum values of the feature, respectively. MinMax normalization ensures that all features contribute equally to the analysis, preventing features with larger ranges from dominating the model's learning process (Chepino et al., 2023; C. Patel et al., 2022). This process is useful in predictive maintenance applications where sensor data with varying ranges, such as temperature, vibration, and humidity, needs to be consistently scaled for accurate pattern recognition.

Lastly, data preprocessing has been implemented in machine fault diagnosis by extracting relevant information from vibration signals to represent the multidimensional features which allows an accurate analysis (Hassan et al., 2024). In data preprocessing for PdM, ensuring clean, accurate, and well-structured data is critical for reliable model training and accurate predictions.

2.6 Anomaly Data Generation

Anomaly data generation involves creating abnormal instances within datasets to enhance the training of anomaly detection algorithms. According to Dang et al. (2022), this process is important as the accuracy of anomaly detection algorithms heavily relies on the quantity and quality of training data. By generating various forms of anomalies such as stuck-at, offset, drift, noise, outlier, and spike, researchers can improve the robustness of these algorithms. There are many approaches to generate anomaly data, Clifton et al. (2011) stated that anomaly data can be effectively generated by applying a multivariate extension of the generalized Pareto distribution, which models extremal data points and provides insights into tail behavior beyond conventional normal data

models. The advanced approach by using Generative Adversarial Network (GAN), for example Chen et al. (2022) generates the anomaly data using a GAN-based predictive model, where the generator produces potential anomalies that help the discriminator enhance detection accuracy. According to Liu et al. (2020), abnormal data can be generated using the Generative Adversarial Active Learning (GAAL) approach, where a generator creates potential outliers that help in training the discriminator to accurately distinguish between normal and outlier data. Figures 2.3 and 2.4 display two types of GAAL architecture on how abnormal data was generated. In summary, the generation of anomaly data plays a critical role in enhancing the robustness and accuracy of detection models to ensure the effectiveness of the result generated.

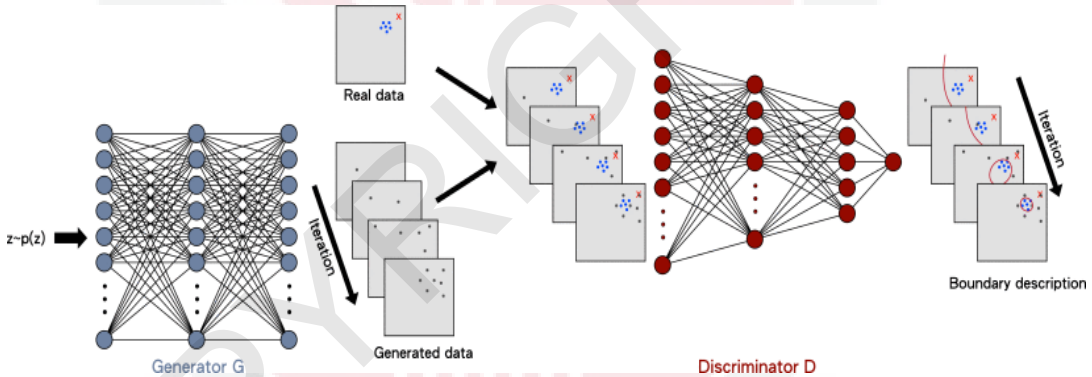


Figure 2.3: Single Objective Generative Adversarial Active Learning Architecture (Liu et al., 2020)**.**

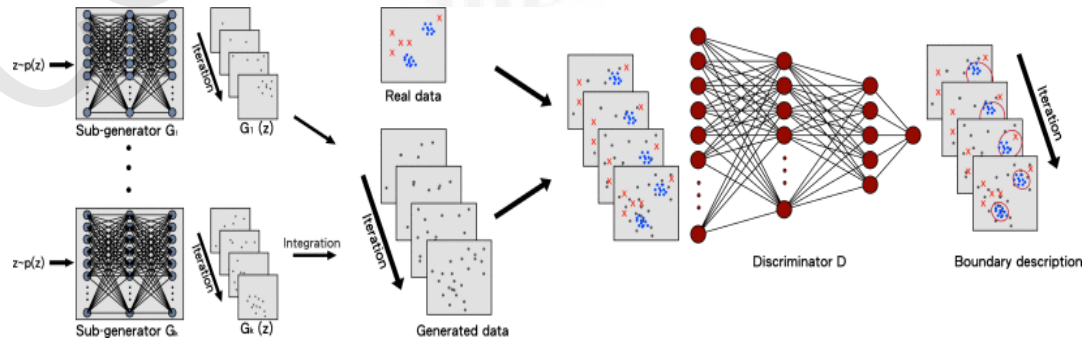


Figure 2.4: Multiple Objective Generative Adversarial Active Learning Architecture (Liu et al., 2020)**.**

Synthetic data generation has emerged as an important technique in fields where access to real-world data is limited or constrained by privacy concerns. As the demand for robust AI models increases, so does the need for high-quality SD particularly in industrial applications. SD finds diverse applications across various fields, including computer vision, online safety, data science, social sciences, healthcare analytics, and more. It is extensively used for training computer vision models to overcome challenges like data annotation efforts, imbalance, and privacy concerns (Delussu et al., 2024). High-fidelity SD aids in data augmentation, supporting machine learning algorithm training, addressing missing data, and generating virtual patient cohorts for personalized medicine (Z. Wang et al., 2024). In social sciences, SD is gaining traction, challenging traditional representational models and emphasizing a relational approach to understand its implications (Offenhuber, 2024). In healthcare analytics, SD shows promise in informing policy, enhancing privacy, and improving predictive analytics, albeit facing challenges related to data quality, bias, and privacy concerns (Giuffrè & Shung, 2023).

The most advanced and popular generated SD model GAN is used in various domains of application. Studies highlight the potential of GAN variants like Conditional Tabular GAN (CTGAN) and CopulaGAN, as well as Transfer Variational Autoencoders (TVAE), in creating SD that mirrors real data distributions, aiding in data augmentation and model performance enhancement (Miletic & Sariyar, 2024; Vaz & Figueira, 2024). According to Afonso & Giufrida (2023) SD generated by GANs improve the sweet pepper detection models, enhancing overall performance. This showcases recent applications of SD in agricultural computer vision. In medical, GANs are utilized for generating synthetic time-series medical records of dementia

patients, ensuring privacy preservation and maintaining data quality for distribution in healthcare applications (Ashrafi et al., 2023). Figure 2.5 shows the average pattern of real and synthetic data with different type of GANs where y-axis is data value and x-axis is the time, according to results Privacy Preserving GAN (PPGAN) is the best model for this application.

The use of synthetic data in PdM is gaining the attention as it addresses the challenges of data scarcity and enhances model training. By generating SD that mimic real-world conditions, industries can improve fault detection and predictive capabilities (Hakami, 2024). According to Eklund (2006) SD enhances predictive maintenance by simulating realistic fault scenarios, enabling the training of classification systems despite limited real labeled data, thus improving fault detection performance. Various method was used to generate synthetic data for PdM, Klein and Bergmann (2019) use physical Fischertechnik model factory to reproduce real failures, enabling the generation of synthetic predictive maintenance data through various methods. Other than that, Generative adversarial networks (GANs) are employed to generate synthetic condition monitoring data for aircraft engines, enhancing the dataset for predictive maintenance analysis (Fu et al., 2019). In PdM, GANs offer a more sophisticated approach to generating synthetic abnormal data by capturing complex patterns in the sensor data, leading to more realistic simulations and better model training compared to traditional methods like Synthetic Minority Over Sampling Technique (SMOTE) or simple oversampling.

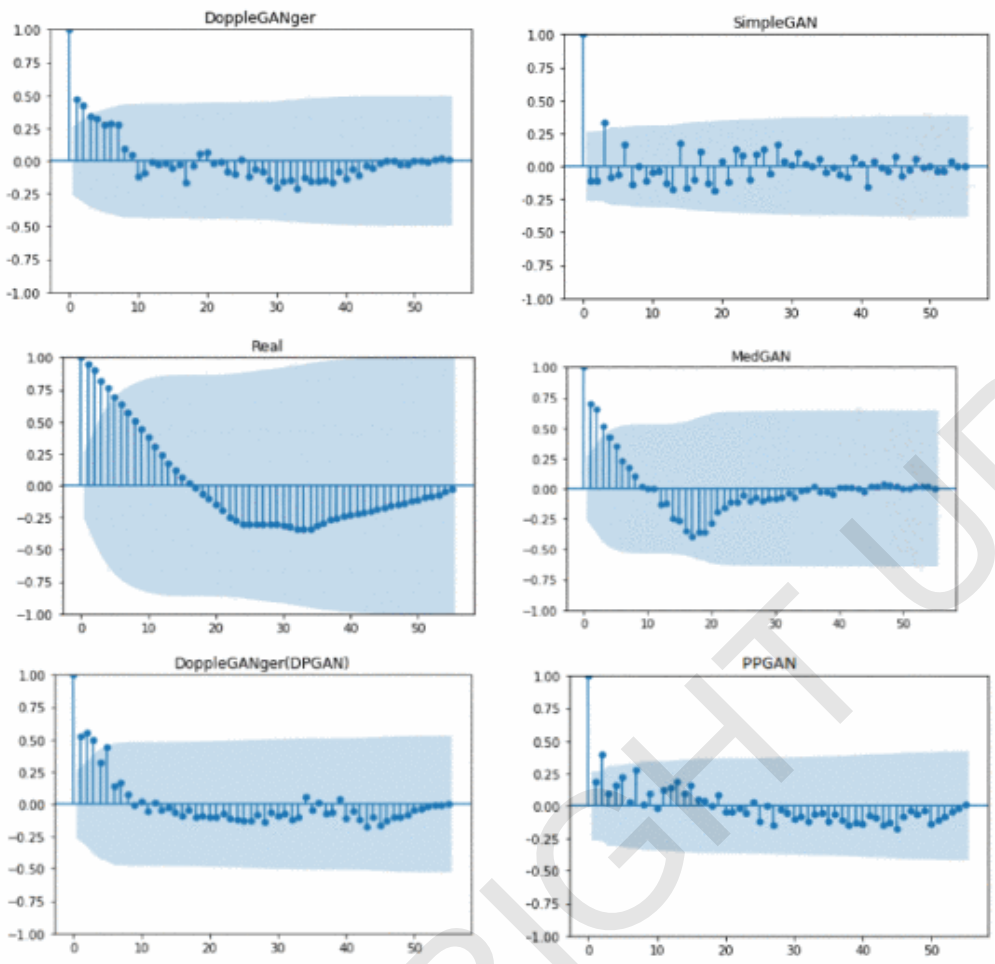


Figure 2.5: Comparison of time series data (Ashrafi et al., 2023)**.**

TGAN and its variants have been employed in various works to enhance the generation of tabular data, addressing challenges such as data privacy, correlation preservation, and causal relationships. These studies demonstrate the versatility and effectiveness of GANs in synthesizing high-quality tabular datasets. TGAN is designed to generate synthetic tabular data, effectively handling both discrete and continuous variables. It outperforms traditional statistical models in capturing inter-column correlations and is scalable for large datasets (Xu & Veeramachaneni, 2018). Other than that, TGAN are utilized to create a large set of synthetic data based on a smaller original dataset of experimental observations on ultra-high-performance concrete (UHPC) (Marani et al., 2020). According to (Amrith et al., 2023; Ma et al., 2024), SD produced by TGAN has

been found to enhance the accuracy of predictive models, surpassing the performance of models trained only on original datasets. TGANs are adept at producing data that retains the statistical properties of the original datasets, which ensures the realism and utility of the generated samples for training purposes, as noted by (Nigam & Srivastava, 2023; Tun & Pisica, 2023).

Various methods have been proposed to evaluate the quality and accuracy of SD. These methods include using clustering indices like the rate of change in linear regression correlation coefficients, Dunn index, and silhouette coefficient for time-series datasets (Seol et al., 2024). According to Meiser et al. (2024) the quality and accuracy of SD were evaluated using Non-intrusive load monitoring (NILM) models for transferability and a reliable evaluation methodology based on Cochran's sample size. The study by Iantovics and Enăchescu (2022) employed a mathematically grounded data-quality assessment method, incorporating predictive power analysis of independent variables and validating the results using metrics such as sensitivity, specificity, and accuracy.

There are still challenges to using SD in real world applications. Bias in SD remains a significant concern, as it can lead to unfair outcomes when models trained on such data are deployed in real-world scenarios. According to Whitney and Norman (2024), SD can lead to false confidence due to increased dataset diversity and representation, potentially compromising model performance. Moreover, SD raises concerns about circumventing consent for data usage, complicating governance and ethical practices by disconnecting data from those impacted by algorithmic harm. In summary, SD generation is a rapidly evolving field with significant implications for both research and industry.

2.7 Anomaly Data Detection

Not all data have abnormalities, sometimes anomalies are subtle and hidden within complex patterns, making their detection both challenging and essential for ensuring the reliability and integrity of a system. Detecting anomalies in data analysis is essential because it significantly influences the quality and dependability of machine learning models (Azimi & Pahl, 2024). According to Patil et al. (2024) detecting anomalies ensures the integrity of scientific research by upholding honesty and truthfulness, especially in critical fields like clinical research. Furthermore, in the technological realm, identifying anomalies promptly is essential to prevent potential problems caused by unexpected events, highlighting the significance of anomaly detection in maintaining system security and performance (A. Sharma & Khanna, 2024). According to Nassif et al. (2021) there are three types of anomaly detection which are Point anomalies, Contextual anomalies and Collective anomalies. Point anomalies where a single data instance can be considered anomalous for the remainder of the data, the instance is called a point anomaly and is regarded as the simplest anomaly form. If in a particular context a data instance is anomalous, but not in another context, it is called a contextual anomaly. Lastly, if a set of associated data instances is anomalous for the entire dataset, it is called a collective anomaly. Anomaly detection has been widely researched and applied across various fields. For example, by applying Principal Component Analysis (PCA) to time series data, businesses can enhance their decision-making processes and mitigate risks through early identification of anomalies (Dani et al., 2024). Other than that, Shukla & Sengupta (2020) do anomaly detection by employing a combination of Hierarchical Clustering

and LSTM Neural Networks to accurately identify outliers in time-series data, the architecture as shown in Figure 2.6.

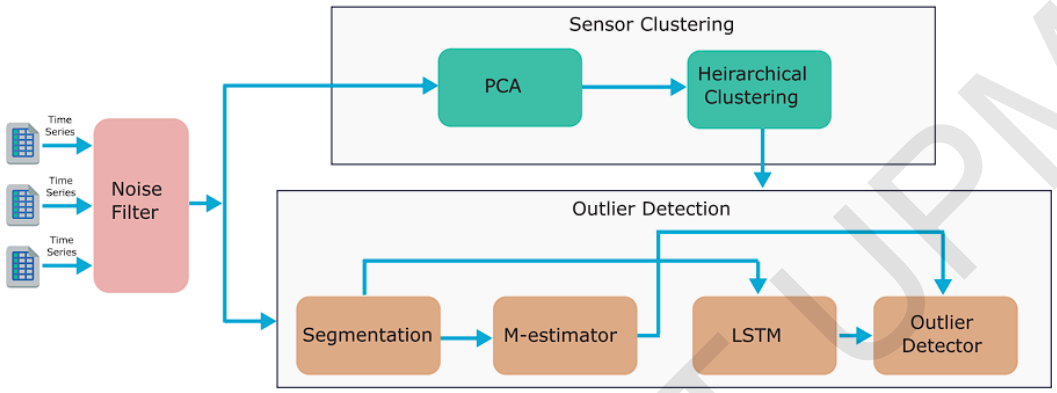


Figure 2.6: Architecture for process detection of outliers (Shukla & Sengupta, 2020)**.**

Next, an unsupervised anomaly detection method based on a sliding-window convolutional variational autoencoder (SWCVAE) was proposed (T. Chen et al., 2020), which effectively identifies anomalies in industrial robots by capturing both spatial and temporal patterns in multivariate time series data. Additionally, Asakura et al. (2020) utilized the Mahalanobis–Taguchi method to conduct anomaly detection in a logistic operating system, calculating the MD based on multidimensional feature vectors derived from vibration data to identify deviations from normal operational states. According to Wu & Zhang (2006) measures the distance between a point and a distribution, accounting for correlations among variables. It is defined as (2.4).

$$MD = \sqrt{(x-\mu)^T S^{-1}(x-\mu)} \qquad (2.4)$$

Where here x is the observation, μ is the mean, and S is the covariance matrix. In industrial settings, MD helps in detecting faults by evaluating the state of equipment

based on multidimensional sensor outputs (Shunji & Hisae, 2012). To conclude, these diverse approaches to anomaly detection demonstrate the growing sophistication in monitoring systems, ensuring more accurate predictions and minimizing potential disruptions across various industrial and operational environments.

2.8 Predictive Maintenance Model

PdM models have gained significant attention in various industries due to their ability to anticipate and prevent equipment failures, ultimately leading to cost savings and increased efficiency. PdM models can be categorized into statistical models, machine learning models, deep learning models, and hybrid models. According to case study (Oprea et al., 2019), a statistical PdM model was employed to assess the reliability indicators of the photovoltaic power plant (PV-PP) components. This model involved calculating failure rates based on historical operational data from the PV-PP. By utilizing probabilistic models, particularly Markov chains, the study was able to predict future failures and maintenance requirements, demonstrating the effectiveness of statistical approaches in predictive maintenance. In another research from Wang et al. (2019), the statistical PdM model was applied by integrating a model-based prognostics method, specifically utilizing an Extended Kalman Filter (EKF) combined with a linearization approach. This method was designed to predict the fatigue crack growth in aircraft fuselage panels. Next according to the research J. Wang et al. (2020), a statistical PdM model was applied through an integrated fault diagnosis and prognosis approach specifically designed for wind turbine bearings with limited degradation data. This approach combined wavelet transform-based fault diagnosis with a Bayesian framework and particle filter for the prognosis of the bearing's RUL.

In machine learning model types, SVM, decision trees and random forests are utilized for classification tasks. A medium Gaussian SVM was applied in Lin (2021) study, which involved extracting features from vibration signals and using these features to classify the health status of motor bearings. The study analyzed the performance of different Gaussian kernel functions (fine, medium, and coarse) and found that the medium Gaussian SVM achieved the highest accuracy in fault diagnosis, with a total accuracy of 96%. In addition, Random Forest was employed for fault diagnosis and the classification of supervisory control and data acquisition (SCADA) real-time data, predicting the state of line production. The Random Forest model demonstrated superior performance with a 95% accuracy, highlighting its effectiveness in ensuring predictive maintenance and improving production performance while avoiding energy wastage (Zermane and Drardja, 2022). Figure 2.7 shows the trend of machine learning model for PdM.

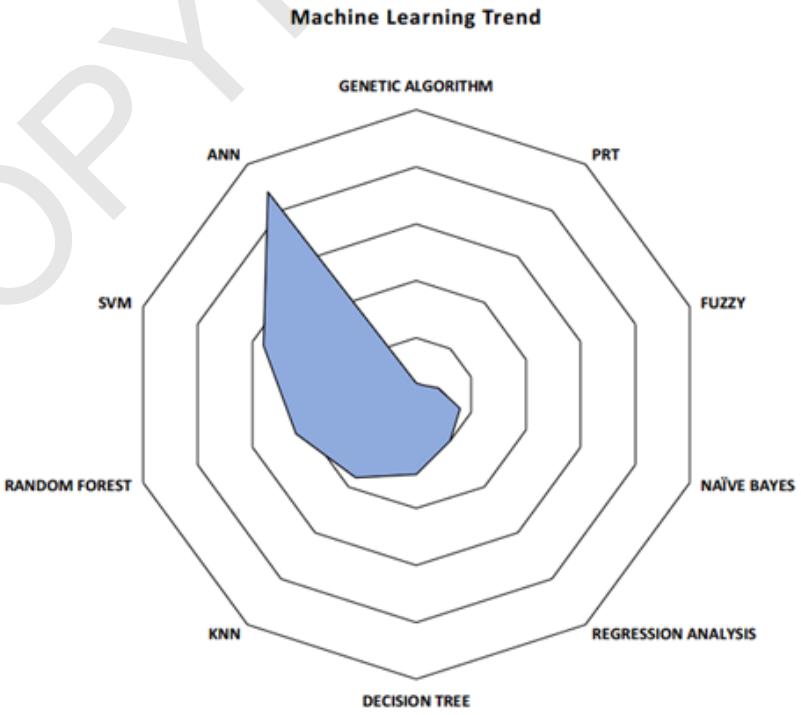


Figure 2.7: Trend machine learning model for PdM (Samatas et al., 2021)**.**

PdM also utilizes deep learning due to their ability to enhance equipment reliability and reduce downtime. Research indicates that deep learning techniques, such as CNNs and recurrent neural networks (RNNs), are effective in analyzing time-series data from machinery to predict failures before they occur (Dube, 2024; Mishra et al., 2024). According to Tanyıldız et al. (2024) these models leverage large datasets to identify patterns and anomalies, which traditional methods may overlook. Bampoula et al. (2021) utilized LSTM autoencoders to develop a deep learning model for PdM in cyber-physical production systems, demonstrating its effectiveness in estimating equipment health status and remaining useful life.

The most recent version of PdM model types is a hybrid version that combines multiple models together. The hybrid model described by Wang et al. (2020) integrates data-driven and model-based approaches for PdM of high-speed railway power equipment. It combines an LSTM-RNN-powered maintenance predictor, which forecasts maintenance timing based on historical data, with a sample generator based on physical degradation and failure models to proactively produce simulated data, addressing data deficiency issues. The architecture of LSTM-RNN as shown in Figure 2.8 below. In another paper by Abood et al. (2023), deep learning is utilized through a hybrid model combining a CNN and a Conditional Generative Adversarial Network (CGAN) to enhance PdM for electromechanical systems. CNN is employed for feature extraction from the input data, while the CGAN is used to generate and classify data, resulting in improved fault diagnosis accuracy. This hybrid approach significantly enhances prediction accuracy compared to standalone models, achieving an average F-Score of 100% on the testing dataset.

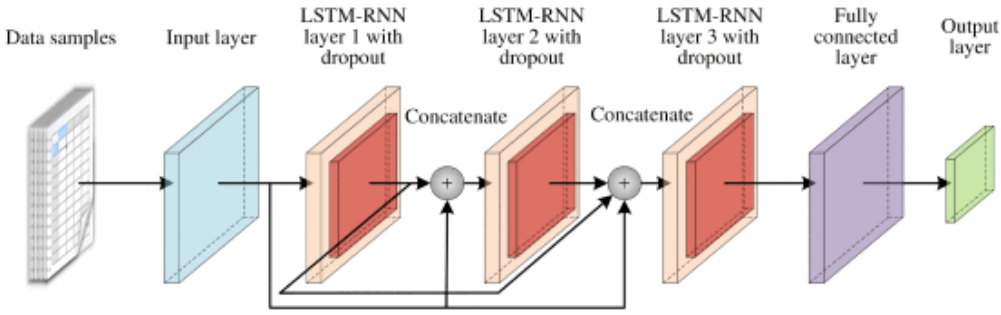


Figure 2.8: LSTM-RNN architecture (Wang et al., 2020)**.**

Combination of IoT and PdM technologies is emerging as a transformative strategy in healthcare maintenance. According to Niyonambaza et al. (2020), reactive maintenance in hospitals leads to significant operational disruptions, which an IoT-based PdM framework using temperature monitoring and machine learning aims to mitigate, achieving up to 96% accuracy in predicting equipment failures. However this approach does not integrate a complete maintenance scheduling system. Similarly, Shamayleh et al. (2020) highlight the inefficacies of corrective maintenance for medical equipment, advocating for an IoT-based PdM that employs vibration signals to predict failures with a 96% accuracy using SVM classifiers. This method requires broader validation across various equipment types. Furthermore, Bani et al. (2022) discuss the limitations of traditional maintenance strategies in handling the Haemodialysis Reverse Osmosis Water Purification System, introducing an IoT-based system that effectively predicts system breakdowns through LSTM models. However, the scalability of this prototype remains a challenge. On the other hand, Priya Talawar et al. (2023) explore machine learning applications for PdM in healthcare IoT, although the models used demonstrated poor sensitivity, necessitating continuous model retraining and better integration with hospital systems. Guissi et al. (2024) address the need for predictive maintenance in MRI systems by developing an IoT-based solution

that uses CNN models to detect anomalies in MRI cooling systems with over 98% accuracy, although the scope of data for model training was limited. Research from Manchadi et al. (2024) propose an IoT-based architecture for Intensive Care Unit (ICU) ventilators to enhance reliability and reduce risks of malfunctions, focusing on continuous monitoring of critical components, yet facing challenges with data security. Collectively, these studies underscore the potential of IoT and PdM to revolutionize maintenance practices in healthcare, though each approach has specific limitations that must be addressed to fully realize their effectiveness. This result of IoT and PdM on medical industry will be benchmarking what accuracy should achieve for this research.

The trend in PdM models is increasingly characterized by the integration of advanced data analytics and machine learning techniques. In summary, PdM models have evolved from traditional statistical approaches to more sophisticated machine learning, deep learning, and hybrid models, each offering unique advantages in various industrial applications. As these models continue to advance, they are poised to significantly enhance equipment reliability, reduce downtime, and optimize operational efficiency across multiple sectors. The integration of advanced analytics and AI-driven techniques will likely play a critical role in shaping the future of PdM, making it an indispensable tool in modern industry.

2.9 Performance Measures for Predictive Maintenance

The effectiveness of a classification algorithm in PdM should be evaluated by its ability to accurately forecast failures, reduce the occurrence of false positives and negatives, and reliably estimate the remaining useful life of assets. To gauge this

effectiveness, performance measures are critical, encompassing several key parameters such as accuracy, precision, sensitivity (or recall), and the area under the Receiver Operating Characteristic (ROC) curve (AUC-ROC). These metrics are fundamentally derived from the confusion matrix, which provides a detailed breakdown of the algorithm's predictions, categorizing them into true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). By analyzing these metrics, one can gain insights into the algorithm's predictive capabilities, ensuring that it not only identifies potential failures but does so with a high degree of reliability and minimal error.

In a binary or binomial classification task, the confusion matrix (CM) serves as a vital tool for evaluating model performance by summarizing the TP, TN, FP, and FN predictions. To further extend this evaluation framework to more complex classification problems, the concept of a hierarchical CM has been proposed. This approach addresses the peculiarities of hierarchical classification problems, proving its applicability across various scenarios and providing a generalized framework for evaluation (Riehl et al., 2023). As shown in Figure 2.9, the confusion matrix visually represents these elements, offering a clear and concise method for assessing the performance of classification models.

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 2.9: Example of CM (Draelos, 2019)**.**

Sensitivity (2.5) also known as the true positive rate or recall, and specificity (2.6) also known as the true negative rate, are key metrics for evaluating the performance of classification models, particularly in binary classification. Sensitivity measures the model's ability to correctly identify positive instances within the dataset, calculated as the ratio of TP to the sum of true TP and FN. On the other hand, specificity measures the model's ability to correctly identify negative instances, calculated as the ratio of TN to the sum of TN and FP. Together, these metrics provide a comprehensive view of a model's performance, highlighting its effectiveness in distinguishing between positive and negative cases within the data.

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.5}$$

$$Specificity = \frac{TN}{FP + TN} \tag{2.6}$$

Accuracy (2.7) is another crucial metric used to evaluate the performance of a classification model. It measures the overall correctness of the model by calculating the proportion of all correctly predicted instances (both TP and TN) out of the total

number of instances in the dataset. In other words, accuracy tells us how often the model is right across all predictions.

$$Accuracy = \frac{TP + TN}{P + N} \tag{2.7}$$

Another parameter that can be obtained from the confusion matrix is Positive Predictive Value (PPV), also known as precision, which measures the proportion of positive predictions that are actually correct. Similarly, Negative Predictive Value (NPV) is another important metric derived from the CM, assessing the proportion of negative predictions that are true negatives. Both PPV (2.8) and NPV (2.9) offer valuable insights into the reliability of a model's predictions, particularly in contexts where the costs of FP or FN are significant.

$$PPV = \frac{TP}{TP + FP} \tag{2.8}$$

$$NPV = \frac{TN}{TN + FN} \tag{2.9}$$

Additionally, the F-measure or F1-score can be obtained using formula (2.10), which can be obtained from recall (sensitivity) and precision (PPV). It provides a balanced metric that considers both FP and FN, making it particularly useful when dealing with imbalanced datasets. All of these formulas are derived from the work of (Chicco & Jurman, 2023).

$$F - measure = \frac{2 \times precision \times recall}{precision + recall} \tag{2.10}$$

The classification report of PdM models involves assessing various machine learning algorithms for predicting and analyzing machine performance (D. B. Sharma et al., 2023; Zanke et al., 2024), providing a detailed evaluation through metrics such as precision, recall, F1-score, and support for each class, typically representing different machine states like "Normal," "Needs Maintenance," and "Failure." These metrics help gauge the model's accuracy in predicting each class, with high precision reducing false positives, high recall ensuring critical issues are detected, and the F1-score balancing precision and recall.

The ROC curve is a graphical representation used to evaluate the performance of binary classification systems (Fawcett, 2006). It is created by plotting the True Positive Rate (TPR), also known as sensitivity, against the False Positive Rate (FPR) at various threshold settings.

$$FPR = \frac{FP}{FP + TN} \tag{2.11}$$

An ideal ROC curve will have its points concentrated at the top left corner of the graph (position (0, 1)), where the TPR is 100% and the FPR is 0%. This point represents a perfect classifier, indicating maximum sensitivity and specificity. In contrast, a point on the diagonal line connecting (0, 0) to (1, 1) indicates a classification that is no better than random guessing, with 50% sensitivity and 50% specificity. A point above this diagonal signifies good classification, whereas a point below it indicates weak prediction. Figure 2.10 shows graphical representation of the ROC curve, which is used to evaluate the performance of binary classification models where blue color is the best classifier. The ROC curve is a powerful tool for visualizing the trade-off

between sensitivity and specificity in a classifier's performance. By analyzing the shape and position of the curve, one can assess the effectiveness of a model in distinguishing between the positive and negative classes.

The AUC-ROC (Area Under the Receiver Operating Characteristic Curve) is a vital metric for evaluating the performance of binary classification models, providing a single scalar value that reflects the model's ability to distinguish between positive and negative classes across all possible thresholds. By measuring the trade-off between the TPR and the FPR, AUC offers a comprehensive, threshold-independent assessment of classification effectiveness (X. Liu et al., 2023; Yang et al., 2022, 2023). This metric is widely used in various fields, including biomedical signal analysis, security frameworks, and anomaly detection, due to its ability to encapsulate the entire ROC curve into a single, insightful measure of a classifier's overall performance. Table 2.3 shows the classification accuracy value based on the area of AUC. This Figure 2.11 illustrates how the AUC varies across different ROC curves, with each curve representing a classifier's performance in distinguishing between positive and negative classes. The AUC values range from 0.5, representing random guessing, to 0.9, indicating a highly effective classifier. Higher AUC values, such as 0.9, demonstrate strong discriminative ability with high sensitivity and low false positives, while lower AUC values, closer to 0.5, indicate a weaker performance.

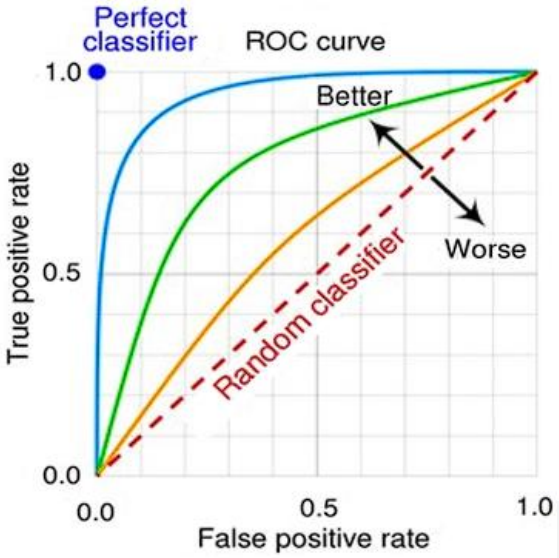


Figure 2.10: Understanding ROC Curve (Campion & Campion, 2023)**.**

Table 2.3: Example of AUC accuracy result (Ustyannie et al., 2021)

AUC value	Classification
0.90 – 1.00	Excellent
0.80 – 0.90	Good
0.70 – 0.80	Fair
0.60 – 0.70	Poor
<0.60	Failure

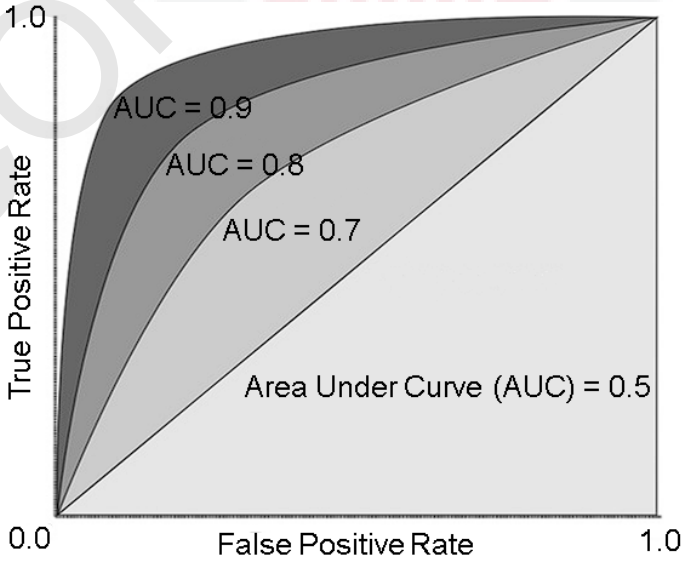


Figure 2.11: AUC ROC graph (Laher et al., 2017)**.**

K-fold cross-validation is a robust statistical method used to assess the performance of machine learning models by partitioning the dataset into 'k' subsets or folds. Each fold serves as a testing set while the remaining k-1 folds are used for training, allowing for a comprehensive evaluation of model accuracy and generalization capabilities. This technique helps mitigate overfitting by ensuring that every data point is used for both training and validation across different iterations, thus providing a more reliable estimate of model performance compared to a single train-test split (Herrera-Chavez et al., 2024). Research indicates that the choice of 'k' can significantly impact the results, with smaller values leading to higher variance in performance estimates, while larger values may increase computational costs (Admojo & Nurul Rismayanti, 2024). The optimal choice of K in K-fold cross-validation can vary depending on the dataset and the specific goals of the analysis. Research from Imran et al. (2023) suggests that smaller values of K, such as K=5 or K=10, are often preferred because they strike an optimal balance between bias and variance. These values facilitate effective training and validation while keeping computational costs manageable. However, some studies suggest that increasing K can lead to more reliable estimates of model performance, particularly in smaller datasets, as it maximizes the training data used in each fold (Marcot & Hanea, 2021; Ridwan et al., 2023). Figure 2.12 shows how data was split into training and testing using K-fold analysis. In summary, K-fold cross-validation effectively balances bias and variance in model evaluation, with the choice of 'K' being crucial for optimizing performance estimates and computational efficiency.

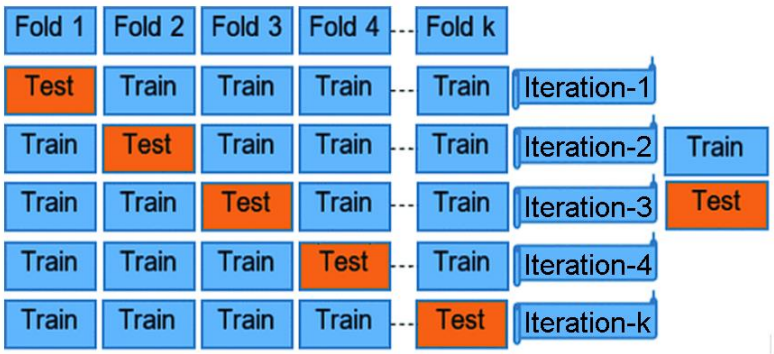


Figure 2.12: Data split between training and testing for every fold.

2.10 Summary

This chapter provides a comprehensive overview of PdM and its critical role in various industry applications, including the maintenance of high-cost industrial machines and sensitive medical equipment like CT scan machines. PdM is essential in these contexts due to the high costs associated with maintenance and the disruptive nature of service interruptions, particularly in medical settings where machine downtime can directly impact patient care.

One of the significant challenges faced in PdM is the lack of sufficient anomaly data, as real-world failure events are rare. This has led researchers to propose the use of synthetic data generation to augment datasets and ensure that machine learning models are effectively trained to detect potential failures. Among the methods proposed for synthetic data generation, GANs have emerged as one of the most effective techniques. GANs are used to generate realistic synthetic data that simulate fault conditions, helping to balance datasets and improve model training.

Anomaly detection using the MD involves identifying unusual data points by measuring their distance from the mean of a dataset, taking into account the covariance

among the variables. This method is effective for detecting outliers in multivariate data, where anomalies are not obvious in univariate analysis. Following anomaly detection, predictive modeling using ANN can be employed. ANNs are capable of learning complex patterns from the data, making them suitable for predicting future outcomes based on historical data. This approach is particularly useful in scenarios where the relationships between variables are nonlinear and intricate, as ANNs can adjust their internal parameters to best fit the observed data, leading to robust and accurate predictions.

The synthetic data generated through GANs and other methods are typically tested using machine learning classifiers to evaluate their usefulness in PdM. These classifiers assess the ability of synthetic data to improve the detection of anomalies and potential failures in machines. Performance metrics commonly used in PdM classification include accuracy, precision, recall, and F1-score, which measure the model's effectiveness in predicting machine failures. By using these metrics, researchers can evaluate how well the predictive maintenance models perform when synthetic data is incorporated, ultimately improving machine uptime and reducing maintenance costs.

This chapter lays the foundation for understanding the challenges and solutions in PdM, particularly the role of synthetic data and machine learning in overcoming data imbalance and improving predictive accuracy in both industrial and critical medical equipment contexts.

CHAPTER 3

METHODOLOGY

3.1 Introduction

The focus of this study is to determine the effect of synthetic data (SD) on a Predictive Maintenance (PdM) model. In this chapter, the process for developing a system capable of classifying abnormalities in Computed Tomography (CT) scan machines based on both real and SD is outlined. The framework involves data preprocessing, SD generation, anomaly detection, machine learning development, and model evaluation to determine the condition of the CT scan machine. The methodology presented in this chapter provides a comprehensive approach to leveraging both real and SD for enhancing the PdM model's ability to accurately classify abnormalities in CT scan machines.

3.2 Overview of the Framework

The overall methodology, as shown in Figure 3.1, begins with the acquisition of data, which is downloaded from the relevant dashboard. This data undergoes a preprocessing stage to ensure that all features are balanced, it means all the features have the same amount of data or each row of the data of CSV must have value (not a blank space). Following this, the data is validated against the scanning log to identify any variations that occur during scanning and non-scanning periods for each feature.

Once validated, the data is normalized, and the Mahalanobis Distance (MD) is calculated. An ellipse is then created using the MD to establish a critical value that serves as a threshold to differentiate between normal and abnormal conditions.

In PdM for CT scan machine, normal data refers to sensor readings and operational parameters such as current, temperature, vibration, and radiation when the machine is functioning correctly without any signs of degradation or failure. These values remain within expected thresholds based on historical data and manufacturer specifications. In contrast, abnormal data consists of sensor readings that deviate significantly from normal conditions, indicating potential faults or impending failures. This may include excessive current fluctuations, unusual temperature spikes, irregular vibration patterns, or radiation anomalies, which could require maintenance intervention.

The next step involves the generation of SD using two distinct methods. The first method, noise addition (NA), involves adding noise to the existing data to create new SD, thereby expanding the head and tail of the real data distribution. The second method involves generating abnormal conditions from the real data based on the MD and employing a Tabular Generative Adversarial Network (TGAN) to balance the normal and abnormal classes. The SD represents the real failure of the machine by checking whether the outliers in the left and right tails exceed the normal condition values for certain features, ensuring that the generated anomalies align with real failure patterns.

In Figure 3.1, the main reason for implementing two distinct paths for data handling and balancing (NA and TGAN) is to leverage both statistical simplicity and AI-driven

complexity effectively. NA adds random noise to data, preserving its underlying distribution and enhancing dataset robustness through controlled variability, making it ideal for quick and efficient data augmentation without extensive computational resources. Conversely, TGAN excels in generating synthetic data for underrepresented classes, addressing class imbalances by learning and replicating complex data distributions. This method adapts over time, improving with new data, crucial for dynamic systems. Integrating both methods capitalizes on NA's efficiency and TGAN's sophisticated data generation capabilities, ensuring a robust, versatile approach to creating a balanced dataset that enhances the generalization capability of predictive models.

Both SD generation methods are then subjected to class verification using the MD, where the generated SD is compared against the previously established threshold. If the number of data for both classes remain imbalanced, the SD creation process is repeated using either NA or TGAN respectively. Once the data is balanced, it is inserted into the Artificial Neural Network (ANN) model for training. Finally, the performance of both SD generation methods is compared to determine which approach is more effective for enhancing PdM in CT scan machines. The stages involved to achieve each objective are also labeled in the figure.

Balancing sensor data sampling is essential to ensure that machine learning models accurately distinguish between normal and abnormal conditions. When data is imbalanced, with normal sensor readings significantly outnumbering abnormal readings, the model tends to become biased toward the dominant class, leading to poor anomaly detection and higher false negatives. This can result in missed failures in

predictive maintenance, increasing the risk of unexpected breakdowns. By maintaining a balanced dataset, the model receives equal exposure to both normal and abnormal conditions, improving its ability to generalize, detect anomalies effectively, and reduce prediction errors.

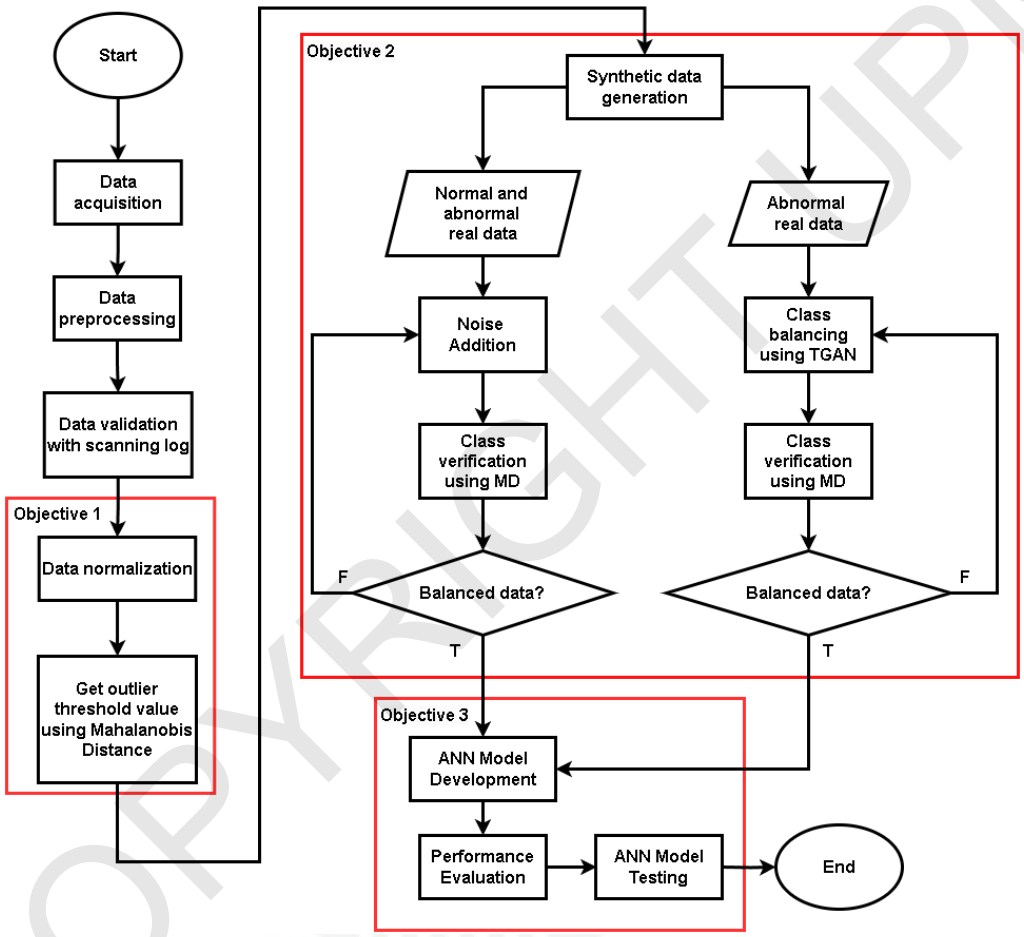


Figure 3.1: Overall flowchart for this project.

3.3 Datasets of Computed Tomography Scan Machine

Datasets for this research from Internet of Things (IoT) connected sensors and they were collected by downloading them from the provided dashboard. All of the sensors for this research were set up by the appointed technical partner of the project at a

hospital located in Klang Valley Selangor. Overall flow for sensor setup to end user as shown in Figure 3.2 where all the sensor data was sent through the IoT gateway and it was saved into MongoDB before the user can visualize or download it.

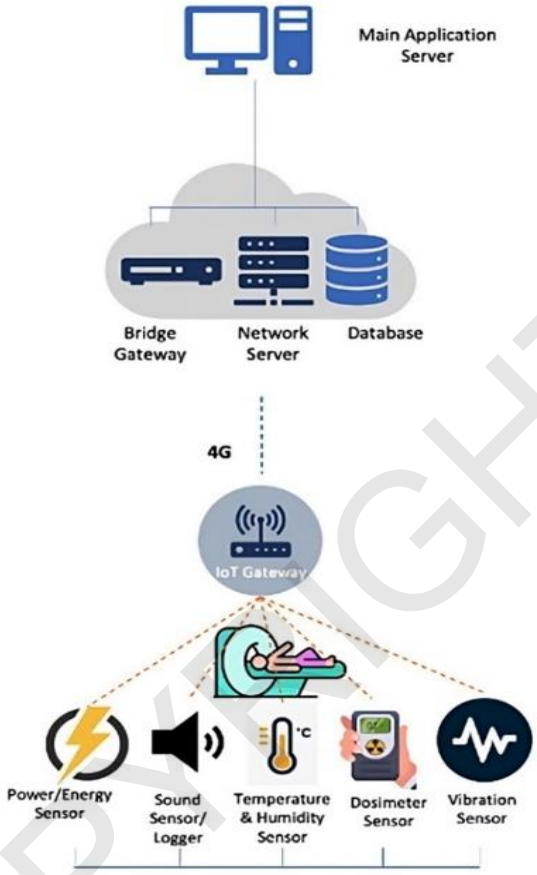


Figure 3.2: Network topology of IoT CT scan monitoring system.

The CT scan machine use for this project is Aquilion Lightning TSX-035A, this machine is a 16-row, 32-slice helical CT scanner developed by Canon Medical Systems, designed for high-performance imaging with advanced features. One of its key capabilities is dose reduction, utilizing technologies such as AIDR3D Enhanced and Exposure 3D to minimize radiation exposure while maintaining image quality. The scanner also includes a subtraction feature, which effectively removes bone and

calcium from datasets, enhancing diagnostic accuracy. Additionally, its navigation mode assists operators through each step of the examination with interactive computer graphics and animations, improving workflow efficiency. The intelligent filming function allows for automatic compilation of images into a predefined layout, streamlining the imaging process. Moreover, the Single Energy Metal Artifact Reduction (SEMAR) technology significantly reduces metal artifacts, ensuring clear visualization of soft tissue structures even in the presence of metal implants. The details specification for the CT scan machine was put into Appendix E.

There are four types of sensors set up inside the CT scan room, which are the AM103, MK926 MEMS, Schneider PM5110, and RAD-S101. The first sensor, the AM103, is used to capture temperature and humidity around the CT scan machine. According to Szabo et al., (2022), these parameters are crucial for detecting potential overheating issues in the equipment. The next sensor, the MK926 MEMS, detects accelerometer data, or more precisely, vibrations. This sensor can capture data along three axes: X, Y, and Z but only the X and Y axes were selected. The Z-axis motion is often minimal or negligible since the machine is designed to rotate in a single plane, making the Z-axis motion less relevant for detecting vibrations or anomalies in this setup.

The Schneider PM5110 power meter sensor captures multiple parameters, such as power, energy, voltage, and current. However, only the current was selected as a parameter for the PdM model, as it is a direct indicator of potential overload conditions that could lead to overheating. Finally, since the CT scan machine produces radiation, the RAD-S101 sensor was employed to detect radiation levels before, during, and after the scanning process. Analysis of the maintenance reports revealed that the X-ray tube

in this machine was frequently replaced. This raised concerns about potential radiation leakage from the X-ray tube, which could pose significant health risks to both patients and radiologists. All these sensors are non-invasive because the CT scan machine is under radiology care and is still under leasing services by the manufacturer.

The data received from the sensors have varying timestamps due to differences in sensor specifications and the data load capacity that the server gateway can manage. Specifically, current, acceleration on the X and Y axes are recorded every 1 minute, radiation is logged every 3 minutes, while temperature and humidity readings are captured every 10 minutes. The amount of data in Table 3.1 also varies due to occasional connectivity losses. Additionally, during emergencies, the person in charge may temporarily remove the sensor from the CT scan machine to facilitate a smoother and faster process for aiding patients, further contributing to the inconsistencies in the amount of data.

Table 3.1: The amount of IoT data points collected for all features in 2023.

Features/ Month	Current	Radiation	Temperature and Humidity	X and Y
January	25841	8604	3089	25839
February	24445	8143	2923	24444
March	36943	9994	4407	10000
April	35518	11791	4292	35508
May	36607	12194	4419	36595
June	35920	11972	3275	35913
July	22197	8067	2869	22206
August	35100	11693	4139	35090
September	32604	10856	3788	32601
October	34537	12001	4297	36152
November	34514	11953	4247	35895
December	34523	11867	4427	35663

In Table 3.1, the amount of data collected throughout 2023 is shown. From this table, it can be concluded that the collected data was not consistent every month, particularly in January, February, June, and July. The amount of current data ranged from 22,197 to 36,943, and the amount of acceleration data ranged from 22,206 to 36,663, reflecting a large volume of data as these sensors captured readings every minute. In contrast, the amount of radiation sensor data ranged from 8,067 in July to 12,194 in May, which is lower compared to current and acceleration data due to the sensor's inability to capture data at intervals shorter than three minutes. The temperature and humidity sensor, which captured data every 10 minutes, recorded an amount ranging from 2,869 to 4,427, as the environment in the CT scan room does not change rapidly. These

variations in the amount of data collected are also due to unavoidable factors such as emergencies or connectivity issues that occasionally interrupt data transmission.

3.4 Data Pre-processing

The data will undergo preprocessing to ensure that the inputs fed into the machine learning model are clean, consistent, and reliable. As illustrated in Figure 3.3, which depicts the data preprocessing flow, the data will be merged into a single Comma Separated Value (CSV) file and sorted according to the timestamp. Any redundant timestamps will be combined into a single entry to ensure that each row has a unique timestamp. Blank areas in the dataset will be filled using time interpolation (3.1) method since this data is in time series (Kumar Srivastav & Vaidya, 2020). This method allows for the accurate estimation of missing data points in a time series, ensuring a continuous and complete dataset for more reliable analysis and decision-making. The values y_1 and y_2 are the known data values at times t_1 and t_2, respectively. The time t represents the point at which the value is missing. The value y is the estimated value at time t, based on interpolation between y_1 and y_2. This dataset will then be matched with scanning durations from the scanning logbook to analyze the differences between scanning and non-scanning periods. The scanning logbook covers the period from January 2023 to July 2023.

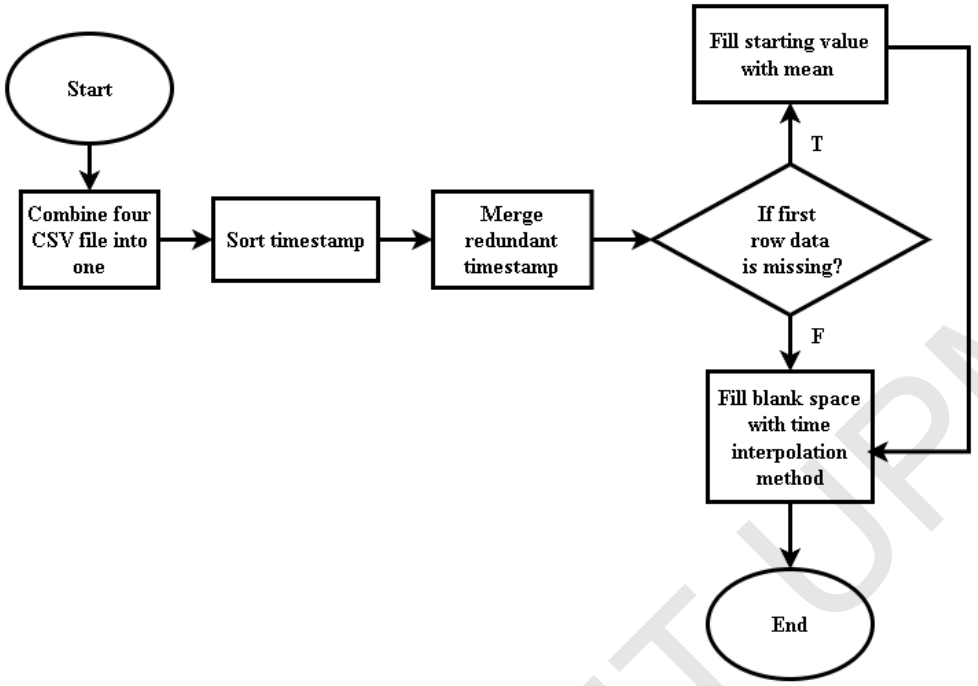


Figure 3.3: Flowchart of data preprocessing.

$$y = y_1 + \frac{(y_2 - y_1)}{(t_2 - t_1)} \times (t - t_1) \quad (3.1)$$

The scanning log contains numerous parameters, but only three were selected for analysis: the CT scan machine name, and the start and end times of the scanning sessions. Since the log includes data from two CT scan machines (CT1 and CT2), only the data from CT1 was used in this research since the IoT sensors are only deployed on CT1 machine. The start and end times were converted into timestamps to align with the sensor data timestamps. This alignment allows to visualize the changes during scanning and non-scanning periods.

Next, the data was normalized to ensure that each sensor's readings contribute equally to the analysis, preventing features with larger numerical ranges from dominating the model's learning process. This process involves adjusting the data to a common scale without distorting the differences in the ranges of values. In this research,

normalization helps bring all sensor data, such as current, acceleration on the X and Y axes, radiation, temperature, and humidity, to a comparable range. This is particularly important because the different sensors have varying units of measurement and scales, such as current measured in amperes and temperature in degrees Celsius. If left unnormalized, these variations could skew the analysis and lead to inaccurate results. Min-max scaling (3.2) was chosen for normalization in this research because, according to Laska and Yolanda (2024), although z-score normalization is resistant to outliers, it can obscure the original data distribution, which may not be ideal for all datasets.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3.2}$$

Where X' is the normalized value, X is the original value, X_{min} is the minimum value in the dataset, and $Xmax$ is the maximum value in the dataset. This formula scales the original value X to a range between 0 and 1, ensuring that all data points are proportionally adjusted to fall within this normalized range. This process helps in standardizing the data, making it easier for the machine learning model to learn patterns without being influenced by differences in the scales of various features.

3.5 Synthetic Data Generation

In this study, the MD, as formulated in (3.3), is employed for outlier detection in a multivariate dataset. The MD was used because it effectively captures the multivariate nature of data, allowing for a more nuanced understanding of outliers in complex datasets (Dhamale & Kashikar, 2025). Other than that, this method performance is superior in scenarios with heavy-tailed distributions, where ML methods may struggle

due to their reliance on specific distributional assumptions (Cabana et al., 2021). Moreover, the choice of MD over more conventional ML methods is further supported by its targeted efficacy in defining the spatial relationships in data. Asakura et al. (2020) use MD to derive abnormal data based on normal data, illustrating its practical application in anomaly detection. While both MD and ML methods are instrumental in data analysis, MD is a statistical measure specifically designed to calculate the distance between a point and a multivariate distribution, considering the correlations between variables. Conversely, ML encompasses a broader range of algorithms used for tasks such as classification, regression, and outlier detection. The MD serves as a critical tool within ML for identifying multivariate outliers, highlighting its role in the broader scope of machine learning applications, particularly in outlier detection scenarios involving multivariate data (Cansiz, 2023). The MD measures the distance between a data point x and the centroid μ of the distribution, considering the correlations among the features. It is mathematically expressed as:

$$MD = (x - \mu)^T \Sigma^{-1} (x - \mu) \qquad (3.3)$$

Where Σ is the covariance matrix, and Σ^{-1} is its inverse. Outliers are identified by comparing the squared MD, MD^2 against a critical value derived from the Chi-squared $(\chi 2)$ distribution with degrees of freedom equal to the number of features. For n features, the critical value is determined as $\chi^2_{0.95,n}$, corresponding to the 95% confidence level. A 95% confidence level is commonly used in this context as it provides a practical balance between certainty and precision. With a 5% chance of incorrectly identifying a data point as an outlier, the 95% confidence level maintains relatively narrow confidence intervals, allowing for more precise analysis while controlling the likelihood of false identifications.

As shown in Figure 3.4, different confidence levels influence the range of the MD. A 99% confidence level would provide more certainty by extending the interval, reducing the likelihood of falsely identifying outliers but at the cost of reduced precision due to wider intervals. Conversely, a 90% confidence level would narrow the interval, making detection more precise but increasing the risk of identifying normal data points as outliers. Therefore, the 95% confidence level provides a reasonable trade-off for minimizing errors while maintaining usable precision in multivariate outlier detection (Hazra, 2017; Simundic, 2008).

Although the boundary in two-dimensional space is an ellipse, the critical value remains a single scalar, as it uniformly applies across all directions in the feature space, ensuring that the ellipse represents a contour of constant MD. Data points with squared MD exceeding this critical value are considered outliers.

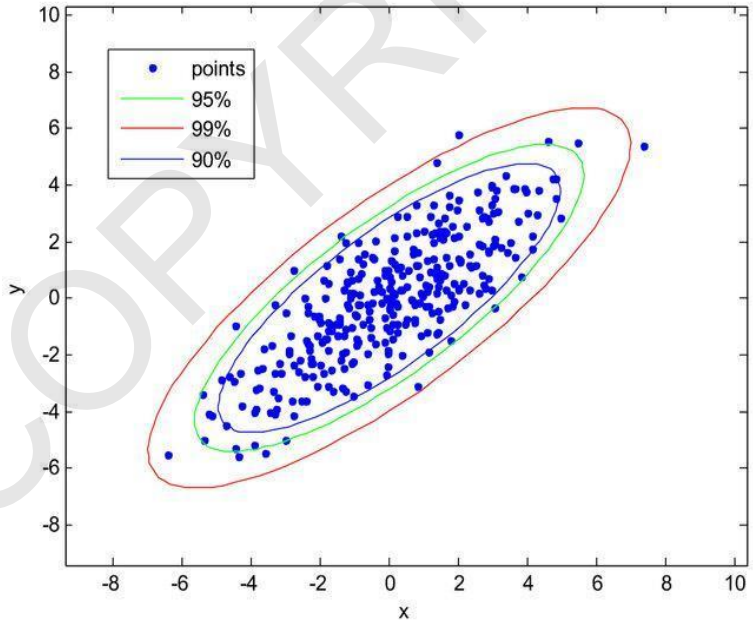


Figure 3.4: Ellipse with different confidence levels (Zhu et al., 2020)**.**

For visualization in two-dimensional feature space, ellipses are constructed to represent the boundary within which 95% of the data points are expected to lie. The

ellipse is defined by the covariance matrix of the two selected features. The eigenvalues, λ_i and eigenvectors, v_i of this covariance matrix are obtained by solving the characteristic equation $\det(\Sigma - \lambda I) = 0$, where I is the identity matrix (Andrew et al., 1993). The lengths of the ellipse's axes are given by (3.4).

$$L_{ellipse} = 2 \times \sqrt{\chi^2_{0.95,2}} \times \sqrt{\lambda_i} \tag{3.4}$$

Where $\chi^2_{0.95,2}$ is the critical value from the Chi-squared distribution with 2 degrees of freedom, and λ_i are the eigenvalues corresponding to the principal directions determined by the eigenvectors, v_i. This method provides a statistical approach to identifying outliers, distinguishing between normal and anomalous observations based on their MD. The combination of eigenvalues and eigenvectors ensures that MD accounts for the actual shape and orientation of the data distribution, Figure 3.5 shows ellipse with different eigenvalues and Figure 3.6 shows ellipse with different eigenvectors. Different eigenvalues will make the shape of the ellipse different, meanwhile different eigenvectors will decide the direction of the ellipse.

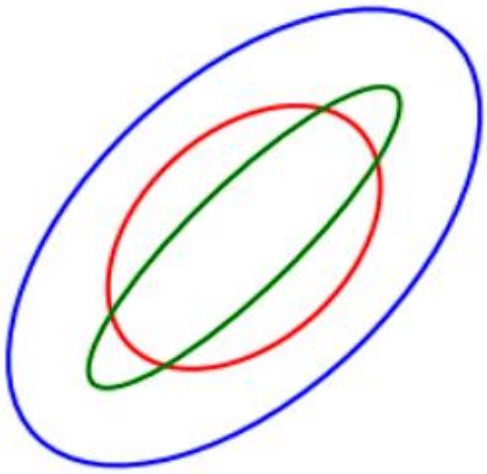


Figure 3.5: Ellipse with different eigenvalues.

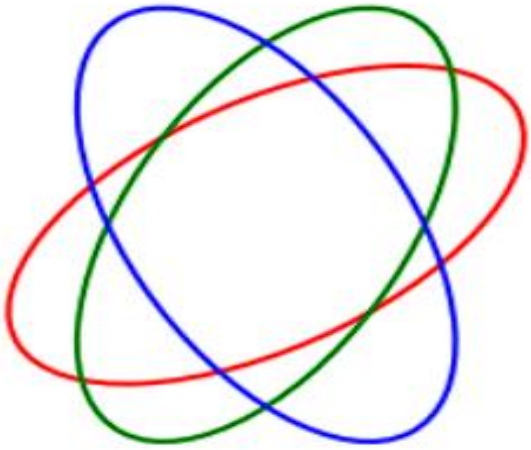


Figure 3.6: Ellipse with different eigenvectors.

Noise was introduced to the normalized features of the dataset. The noise was generated from a normal distribution with a mean of 0 and a standard deviation proportional to the original feature's standard deviation, scaled by a noise factor (NF) of 2. The NF is essentially a scaling factor applied to the standard deviation of the original dataset features when generating noise. By setting the NF to 2, noise is generated at a level that is double the standard deviation of each feature, which effectively amplifies the variability within the dataset. This approach ensures that the synthetic data points are spread further across the range of possible values, thereby simulating potential abnormal conditions that could occur in real scenarios.

When comparing NF of 2 to other NFs, it was observed that noise levels lower than 2 (e.g., NF less than 2) introduced only minimal variability in the dataset. While these lower levels maintained the data within safe bounds, they did not effectively simulate potential outlier scenarios or significant abnormal conditions, making them less suitable for tasks such as anomaly detection and predictive maintenance. Conversely, noise levels higher than 2 (e.g., NF > 2) resulted in unrealistic values for several features, particularly humidity, where values exceeded 100%, violating the physical

and sensor-imposed limits. This led to a deterioration in the quality of the SD, as the extreme values no longer reflected plausible operating conditions.

The value of 2 was chosen because exceeding this threshold would surpass the sensor's specification range. To ensure that the synthetic abnormal data are both realistic and useful, it is crucial to generate them within the sensor's detection limits. This involves understanding the limits and characteristics of the sensors used in data collection to mimic real-world examples (Romanelli & Martinelli, 2023). This step effectively expanded the head and tail of the distribution, simulating conditions that could be considered abnormal. The choice of the NF value is particularly significant because it determines the extent to which the dataset's original distribution is altered without exceeding the physical and operational constraints of the sensors involved.

Table 3.2: Minimum and maximum value with different noise factor.

Features		Current (A)	Radiation (uSv/h)	X (g)	Y (g)	Temperature (°C)	Humidity (%)
NF = 0	Min	0	0	-1.01	-0.005	17.2	29.5
	Max	67.58	2684352.98	0	0.05	31.2	67
NF = 1	Min	-7.56	-376616.75	-1.2	-0.01	15.47	18.36
	Max	70.11	2647380.15	0.19	0.05	33.75	73.35
NF = 2	Min	-17.17	-701647.91	-1.44	-0.03	12.85	3.65
	Max	72.49	2773961.52	0.26	0.06	36.19	100.21
NF = 3	Min	-27.39	-1286105.54	-1.66	-0.03	9.46	-11.60

	Max	78.98	2917183.91	0.47	0.08	36.74	116.25
NF = 4	Min	-37.39	-1542467.85	-1.92	-0.05	5.46	-45.31
	Max	73.66	2937549.43	0.54	0.10	43.13	143.87
Sensor Range	Min	1	0.15	-16	-16	-20	0
	Max	500	36000	+16	+16	70	100

In this context, maintaining the NF at or below 2 prevents the synthetic data from reaching values that are not only implausible but also beyond what the sensors could realistically detect, thereby preserving the integrity and applicability of the synthetic dataset for predictive modelling and anomaly detection tasks. Table 3.2 shows minimum and maximum value with different noise factors, and the humidity range will be outside sensor range if the NF is more than 2. Negative value of current and radiation will not be considered because sensor reading should not below zero value.

After the noise was added, the data was reverted to its original scale using the minimum and maximum values saved from the normalization step. This ensured that the noisy data remained within the same range as the original data, preserving the integrity of the dataset. The MD was recalculated for each point in the synthetic dataset. Data points were then labeled as "0" as normal or "1" as abnormal based on whether their squared MD exceeded the previously determined critical value. This labeling process ensured a clear and consistent classification of the SD.

A visual comparison between the real and SD distributions was made using histograms to illustrate the effects of the NA and the creation of abnormal conditions. This

comparison is shown in Figure 3.7, providing a clear example of how the SD compares to the original dataset. It is shown that the range of values for the SD parameters are expanded at moderate frequencies compared to the real data that focused on a smaller range of values with a higher peak.

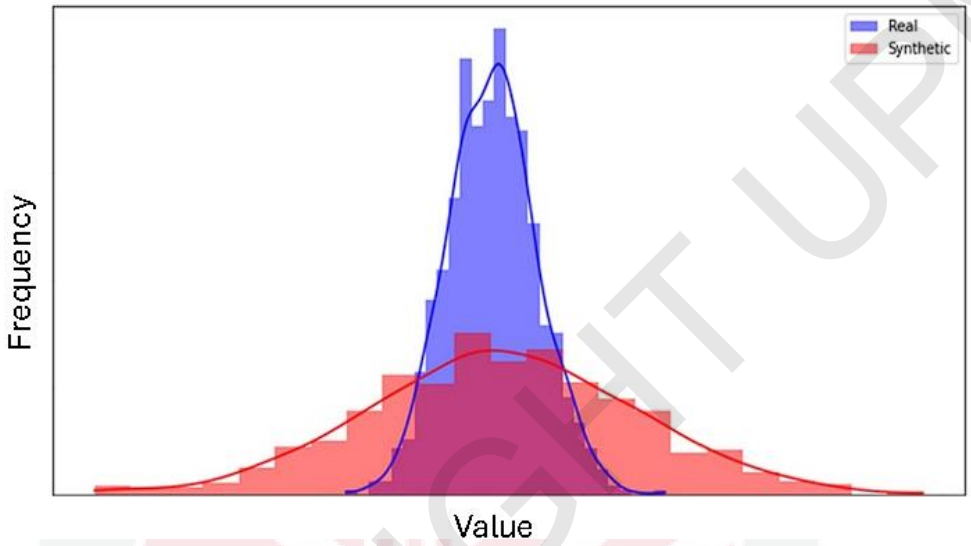


Figure 3.7: Distribution of real and synthetic data.

After identifying the outliers, it was observed that the dataset became imbalanced, with the number of outliers differing from the number of normal data points. To address this imbalance, second method was used which is TGAN was employed to generate SD and balance the dataset. The TGAN is subset for the GAN where GANs are composed of two neural networks, termed the generator and the discriminator, which contest with each other in a game-theoretic scenario. The SD generated by TGAN has been shown to improve the accuracy of predictive models, outperforming those trained solely on original datasets (Amrith et al., 2023; Ma et al., 2024). TGANs generate data that maintains the statistical characteristics of the original datasets, ensuring that the generated samples are realistic and useful for training (Nigam & Srivastava, 2023; Tun & Pisica, 2023). TGAN is specifically designed for generating high-quality SD data in

tabular form, which is particularly suitable for the structured nature of the dataset. The TGAN model consists of a generator that learns to create synthetic samples that resemble the normal data points, and a discriminator that attempts to distinguish between real data and SD. The training process continues iteratively until the generator produces SD so that the discriminator cannot reliably differentiate from real data. The architecture of the TGAN model used in this study is illustrated in Figure 3.8.

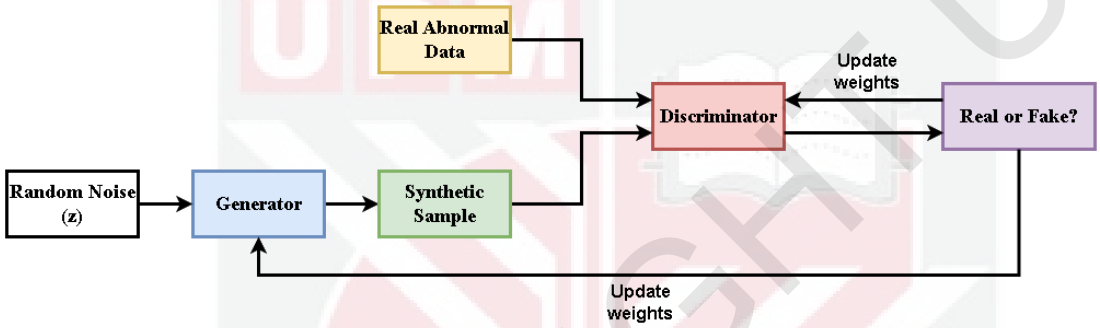


Figure 3.8: The GAN architecture use for tabular data (Sauber-Cole & Khoshgoftaar, 2022)**.**

The SD generated by TGAN was then added to the original dataset, particularly to the minority class, in this case abnormal data ensuring that the dataset for both classes (normal and abnormal) will become balanced. This balance was crucial for subsequent analysis, as it ensured a more equitable representation of both normal and anomalous data points. Balancing the data is critical for training ANN models, as imbalanced data can lead to model bias where the model overly favors the majority class, potentially overlooking significant patterns in the minority class. This can adversely affect key performance metrics such as accuracy, precision, and recall, which are essential for evaluating the effectiveness of the model in predicting both normal and abnormal conditions accurately. TGAN is designed specifically to generate synthetic tabular

data, it adapts the traditional GAN architecture to work with mixed-type data (numerical and categorical features) typically found in tabular datasets.

By integrating the MD for robust outlier detection with TGAN for data augmentation, this method effectively addresses the challenges of working with an imbalanced, unlabeled dataset, enabling more reliable and comprehensive analysis. To assess TGAN's performance in creating SD, Kullback-Leibler (KL) Divergence is used. KL Divergence measures how one probability distribution diverges from a second, reference probability distribution (Bouhlel & Dziri, 2019). Additionally, the correlation matrix is employed to identify similar structures in the data, helping visualize whether the SD resembles the real data (Pham et al., 2024). Another visualization method used to assess TGAN is the Kernel Density Estimation (KDE) plot, which provides a smoothed, continuous approximation of the data distribution, making it particularly useful for comparing the distributions of real and SD (Plesovskaya & Ivanov, 2021). These combined techniques ensure a thorough and reliable evaluation of the TGAN-generated SD, enhancing the overall quality of the analysis.

3.6 Machine Learning Model Development

To evaluate the effectiveness of SD generated by both NA and TGAN methods, the data was fed into an ANN model designed for PdM, as illustrated in Figure 3.9. These models leverage vast datasets from IoT sensors and historical maintenance records, enabling real-time insights into equipment health. The integration of ANNs in predictive maintenance has shown remarkable results, including a 92% prediction

accuracy compared to traditional methods at 78%, and a 35% reduction in system downtime (Abdulrazzq et al., 2024).

ANN excel in handling non-linear models, which makes them highly effective for predictive maintenance (Kaynak & Ervural, 2024). Their ability to analyze complex data patterns is integral to developing sustainable maintenance strategies for road bridge infrastructure (Gunderia et al., 2024). For this research, the ANN model was structured to handle six input features: Current, Radiation, Temperature CT, Humidity CT, and X and Y accelerations. These features were carefully selected due to their significance in monitoring operational conditions relevant to PdM.

The architecture of the ANN consisted of an input layer, two hidden layers, and an output layer. Each of the hidden layers was fully connected to the preceding and succeeding layers, allowing the network to capture complex, non-linear relationships in the data. The neurons in the hidden layers applied activation functions, introducing the necessary non-linearity to enhance the model's learning capacity. This structure was crucial for ensuring that the ANN could effectively learn from the SD generated by both methods. To properly assess the model's performance, the dataset was divided into 70% training, 20% validation, and 10% testing. This data splitting is crucial for training the model effectively while also allowing for proper validation and independent testing to ensure the model's generalization on unseen data.

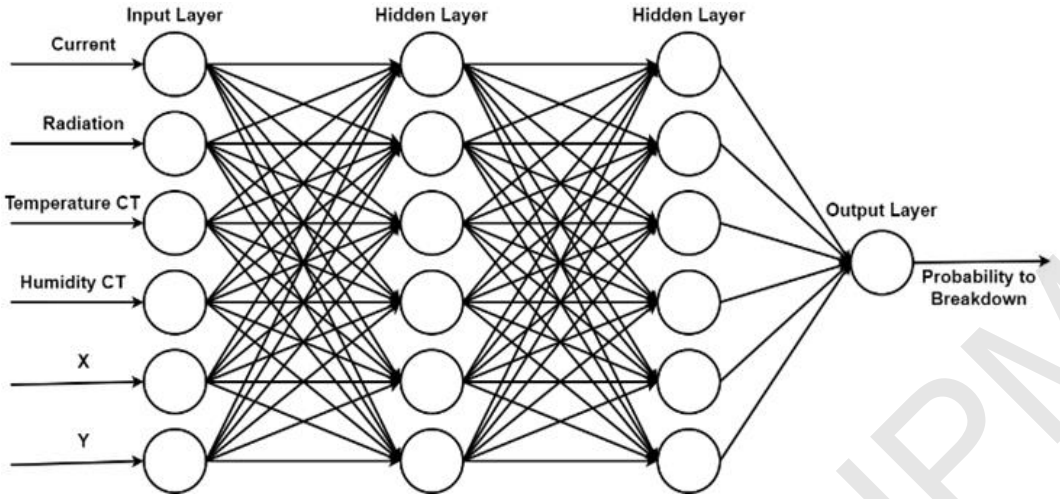


Figure 3.9: The proposed ANN architecture.

The output layer of the ANN consisted of a single neuron, which generated a probability score indicating the likelihood of a breakdown under the given conditions. This output was used for binary classification, where 0 represented normal operation and 1 indicated abnormal conditions. This output was essential for PdM, as it provided a direct measure of the risk associated with each data point parameter for ANN model as shown in Table 3.3.

The threshold of 0.5 was chosen to differentiate between normal and abnormal conditions based on established anomaly detection principles. Burmeister et al. (2023)suggested that values approaching 1 are associated with anomalies, while scores below 0.5 correspond to normal observations. This threshold aligns with common classification practices in predictive maintenance, ensuring a clear decision boundary between operational and failure states. Additionally, Brownlee (2021) outlined four key considerations in threshold selection which are (1) predicted probabilities are not always calibrated, (2) training and evaluation metrics may differ, (3) imbalanced class distribution can affect decision boundaries, and (4) the cost of misclassification must be considered. Given these factors, a 0.5 threshold provides a balance between

sensitivity and specificity, reducing false positives while ensuring early fault detection in CT scan machines.

To assess the quality of the SD, the ANN model was trained separately on datasets generated by NA and TGAN. For each dataset, the model was trained using a loss function such as binary cross-entropy to minimize the difference between the predicted probabilities and the actual outcomes. The model's performance was then evaluated based on its accuracy in predicting breakdowns in both normal and abnormal conditions. By comparing the predictive accuracy of the ANN when trained on data from the NA method versus the TGAN method, insights were gained into which technique was more effective at creating SD that accurately represented abnormal conditions. This comparison was crucial for determining the most appropriate approach for future SD generation in PdM applications, guiding the selection of methods that best support reliable and accurate predictions.

The development and implementation of algorithms in this research are implemented using Python version 3.9.18, running on an AMD Ryzen 7 4000 series processor with an NVIDIA GeForce GTX 1650 Ti GPU. This programming language was chosen due to its extensive libraries and tools that support machine learning and SD generation.

Table 3.3: Setting for ANN model.

ML Parameters	Setting
Activation	Sigmoid
Optimizer	Adam
Loss	Binary cross entropy
Epoch	100
Batch size	32
Learning rate	0.001

3.7 Performance Evaluation

To evaluate the performance of SD generation, all the real and synthetic data will be used in the ANN development. Then, the real anomaly data collected from the sensors will be tested as ANN model validation. The anomaly data is obtained in November 2023 after the confirmation from the CT scan machine repair report provided by the technical partner. The performance of the ANN model was evaluated using a comprehensive set of metrics that are essential for assessing PdM models. These metrics included accuracy, precision, sensitivity (recall), specificity, F1-score, and the AUC-ROC. Each of these metrics was derived from the confusion matrix, which provides a detailed breakdown of the model's predictions.

The performance of the ANN model was evaluated using a confusion matrix, which categorized the model's predictions into four groups: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). From these categories, according to (Fawcett, 2006) several key performance metrics were derived. Accuracy (3.5) was calculated as the proportion of correctly predicted instances (both TP and

TN) out of the total number of instances, providing a general measure of the model's correctness. The formula for accuracy is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.5)$$

Precision or Positive Predictive Value (PPV) in (3.6) assessed the proportion of positive predictions that were actually correct, reflecting the model's reliability in identifying positive cases. It was calculated as:

$$Precision = \frac{TP}{TP + FP} \quad (3.6)$$

Sensitivity (also known as recall) in (3.7) measured the model's ability to correctly identify positive instances, highlighting its effectiveness in detecting true positives. The formula for sensitivity is:

$$Sensitivity = \frac{TP}{TP + FN} \quad (3.7)$$

Specificity (3.8) evaluated the model's capability to correctly identify negative instances, ensuring that false positives were minimized. It was calculated as:

$$Specificity = \frac{TN}{FP + TN} \quad (3.8)$$

Finally, the F1-Score (3.9) which is the harmonic mean of precision and recall, provided a balanced metric that was particularly valuable for datasets with imbalanced classes, where both precision and recall needed to be considered equally. The formula for the F1-score is:

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3.9)$$

Together, these metrics provided a comprehensive view of the ANN model's performance, allowing for a detailed assessment of its predictive accuracy and reliability in the context of PdM.

The Receiver Operating Characteristic (ROC) curve was used to graphically evaluate the model's performance by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings. The AUC-ROC was then calculated to provide a single scalar value that reflected the model's ability to distinguish between positive and negative classes across all possible thresholds. A higher AUC value indicated better model performance, with values closer to 1.0 representing excellent discriminative ability.

To ensure robust evaluation and mitigate overfitting, 5-fold cross-validation was employed. In this method, the dataset was partitioned into five subsets or folds. Each fold served as a testing set while the remaining four folds were used for training the model. This process was repeated five times, with each fold being used exactly once as the testing set. The performance metrics, including accuracy, precision, sensitivity, specificity, and F1-score, were calculated for each fold, and the average performance across all folds was reported. Using fold, K=5 strikes a balance between bias and variance, providing a reliable estimate of the model's generalization capabilities while keeping computational costs manageable. This approach ensured that the model's performance was consistently evaluated across different subsets of the data, providing a thorough assessment of its predictive power.

3.8 Summary

In this study, a comprehensive approach was employed to enhance and evaluate PdM models using both real and SD. Outliers were first detected using MD, allowing for the identification of abnormal conditions without labeled data, which led to an imbalance in the dataset. This use of MD to detect outliers is important, as it identifies deviations based purely on the inherent statistical characteristics of the data. By detecting these abnormalities without pre-labeled data, MD contributes to creating an imbalanced dataset, as the number of identified anomalies is typically much smaller than the number of normal data points. To address this, SD was generated through NA and TGAN. This SD was then incorporated into an ANN model designed for PdM. The model's performance was rigorously evaluated using key metrics derived from the confusion matrix, including accuracy, precision, sensitivity, specificity, and F1-score, as well as the AUC-ROC. To ensure a reliable assessment, 5-fold cross-validation was utilized, striking a balance between bias and variance. The prediction validation method will be compares using data collected from September 2023 to December 2023. The models are evaluated by predicting machine breakdowns, with the results assessed against real data and the maintenance report.

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1 Introduction

In this experiment, the dataset consisted of 88272 total numbers of real and SD combining normal and abnormal. The MD was used to label the data normal or abnormal using the threshold value of 3.5485. The SD was generated by using two methods which are TGAN and NA. These dataset are used in training and tests for classification using ANN model. The performance of each algorithm is analyzed and discussed. Later, the developed model is validated using real anomaly data collected from the study site in November 2023.

4.2 Data Preprocessing of IoT Sensors

The predictive maintenance IoT system for the CT scan machine utilizes four specialized sensors, each tailored to monitor key operational parameters crucial for the machine's performance and reliability, the setup has been reported in Mohd Azrul Shazril et al. (2023). The AM103 Temperature, Humidity, and CO2 Sensor, depicted in Figure 4.1, is a MEMS-based device designed to measure the ambient temperature, humidity, and CO2 levels in the CT scan room. It employs the LoRaWAN communication protocol to ensure efficient data transmission. For radiation safety, the RAD-S101 Radiation Dosimeter, as shown in Figure 4.2, is utilized. This sensor is capable of detecting hard beta rays, gamma rays, and X-rays, which are essential for

adhering to radiation safety standards. It communicates data through RS485 and MODBUS-RTU protocols. The importance of vibration analysis in the maintenance of the CT scan machine is addressed by the MK926 MEMS Accelerometer Vibration Sensor, featured in Figure 4.3. This sensor has a three-axis system and is primarily used to monitor the rotational movements of the gantry. Finally, the Schneider PM5110 Power Meter, shown in Figure 4.4 to monitor current use by CT scan machine. Together, these sensors form a robust monitoring framework that aids in the early detection of potential faults, thereby improving maintenance strategies and prolonging the machine's operational life.

Figure 4.1: AM103 sensor put inside CT scan room to detect the environemnt.

Figure 4.2: RAD-S101 radiation sensor attach opposite with where radiation beam come out.

Figure 4.3: MK926 MEMS accelerometer attach on gantry to detect the vibration of the machine.

Figure 4.4: Reading shown from power meter sensor, Schneider PM5110.

In the preprocessing stage, the downloaded data was combined into a single CSV file and sorted based on the timestamp. Figure 4.5 illustrates the data after sorting, where the highlighted values indicate entries with the same timestamp. These duplicate timestamps were merged, as depicted in Figure 4.6, to eliminate redundancy and ensure clean data for subsequent analysis. In Figure 4.7, a mean value was inserted into the first row of any feature with missing values. The use of the mean value to replace missing values in the dataset is a common practice aimed at maintaining statistical integrity. The mean provides a central tendency measure, which is particularly useful in datasets where the parameter values do not vary widely. By using the mean value, the impact on the dataset's overall statistical distribution is minimized, preventing the introduction of bias that could occur with more extreme substitution values, such as the minimum or maximum. This step is crucial because the time interpolation method requires an initial value to begin the process. Finally, as shown in Figure 4.8, the missing values were filled using the time interpolation method, which calculates missing data points based on the surrounding time intervals and existing

data. This approach ensures continuity and accuracy in the dataset, preparing it for further analysis.

timestamp	Current(Amp)	Radiation(uSv/h)	X(g)	Y(g)	Temperature(°C)	Humidity(%)
1675180808	3.899995089					
1675180808		0				
1675180835			-0.989	0.021		
1675180880	3.98473525					
1675180907			-0.975	0.023		
1675180934					20.8	41
1675180952	4.004533291					
1675180979			-0.989	0.023		
1675181024	3.659770489					
1675181024		0				
1675181051			-0.988	0.026		
1675181096	3.892496586					
1675181123			-0.993	0.022		

Figure 4.5: Highlighted redundant data.

timestamp	Current(Amp)	Radiation(uSv/h)	X(g)	Y(g)	Temperature(°C)	Humidity(%)
1675180808	3.899995089	0				
1675180835			-0.989	0.021		
1675180880	3.98473525					
1675180907			-0.975	0.023		
1675180934					20.8	41
1675180952	4.004533291					
1675180979			-0.989	0.023		
1675181024	3.659770489	0				
1675181051			-0.988	0.026		
1675181096	3.892496586					
1675181123			-0.993	0.022		

Figure 4.6: Highlighted merge timestamp.

timestamp	Current(Amp)	Radiation(uSv/h)	X(g)	Y(g)	Temperature(°C)	Humidity(%)
1675180808	3.899995089	0	-0.98128	0.023398	20.97837838	47.8511803
1675180835			-0.989	0.021		
1675180880	3.98473525					
1675180907			-0.975	0.023		
1675180934					20.8	41
1675180952	4.004533291					
1675180979			-0.989	0.023		
1675181024	3.659770489	0				
1675181051			-0.988	0.026		
1675181096	3.892496586					
1675181123			-0.993	0.022		

Figure 4.7: Highlighted missing data filled with mean value.

timestamp	Current(Amp)	Radiation(uSv/h)	X(g)	Y(g)	Temperature(°C)	Humidity(%)
1675180808	3.899995089	0	-0.98128	0.023398	20.97837838	47.8511803
1675180835	3.931772649	0	-0.989	0.021	20.94015444	46.3830702
1675180880	3.98473525	0	-0.98025	0.02225	20.87644788	43.9362201
1675180907	3.992159515	0	-0.975	0.023	20.83822394	42.4681101
1675180934	3.999583781	0	-0.98025	0.023	20.8	41
1675180952	4.004533291	0	-0.98375	0.023	20.8	41.3305509
1675180979	3.87524724	0	-0.989	0.023	20.8	41.8263773
1675181024	3.659770489	0	-0.98837	0.024875	20.8	42.6527546
1675181051	3.747042775	0	-0.988	0.026	20.8	43.148581
1675181096	3.892496586	0	-0.99112	0.0235	20.8	43.9749583
1675181123	3.927694142	0	-0.993	0.022	20.8	44.4707846

Figure 4.8: Highlighted missing data fill using time interpolation method.

4.3 Data Validation with Scan Log

Figures 4.9 to Figure 4.14 illustrate the sensor readings during scanning events in April 2023, providing a clear view of how each sensor responded in real-time when the CT scan machine was in operation. The data presented in this section focuses on 1st April 2023 until 31st April 2023 as the best example since this month had no missing data. Data from other months, which contain gaps, have been placed in Appendix C. Visualization for August 2023 to December 2023 not included in Appendix C due to no scanning log from the hospital. Total data captured for year 2023 is 841333 but 46095 of data was selected for further analysis due to the constraint that MD had limitation for large dataset. The 46095 data was selected based on complete data for 6 features each timestamp and it was normalized before going into SD generation.

The mapping between recorded sensor reading and the patient's scanning time helps to understand the relationship between the scanning duration (in red) and the corresponding sensor reading (in blue), revealing critical insights into the machine's behavior and performance during actual usage. To determine when the CT machine should undergo maintenance, the system analyzes changes in these sensor readings.

Data collected from January to December, which sometimes experiences connection losses (thus necessitating the use of time interpolation to fill missing data), provides a robust basis for this analysis. For instance, an unusually high radiation value may indicate an ongoing scan, which is expected, but persistent high values or unexpected readings in other parameters like current or vibration could signal operational issues, prompting a maintenance check. Importantly, the input data used in our models represent real scenarios in the CT scan room, as evidenced by the changes recorded in the data during scanning and non-scanning states. Figures 4.9 to 4.14 show the data used for machine learning changes during scanning and non-scanning periods, illustrating the dynamic nature of the operational environment. Most of the sensor data is at baseline values during the non-scanning state, but there are changes during the scanning state. For baseline values, current is 3 Amp, radiation is 0 uSv/hr, X-axis is -1.0g, Y-axis is 0.02g, temperature is 20.5℃ and humidity is 50%. From this, it can be concluded that the data transmitted more frequently is a better match compared to the others. The overlapping values (red line overlaps with blue line) is the exact matching time of scanning and sensor reading. There are also non-overlapping values due the mismatch and sensor's activation time that require a longer period such as radiation sensor.

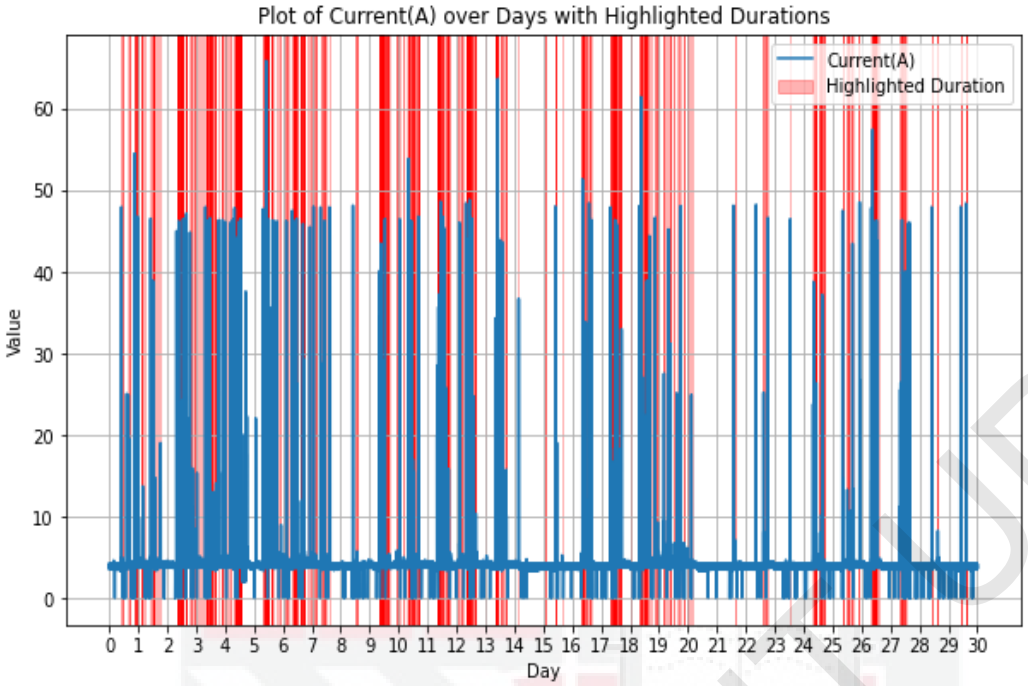


Figure 4.9: Current data for one month with highlighted scanning duration.

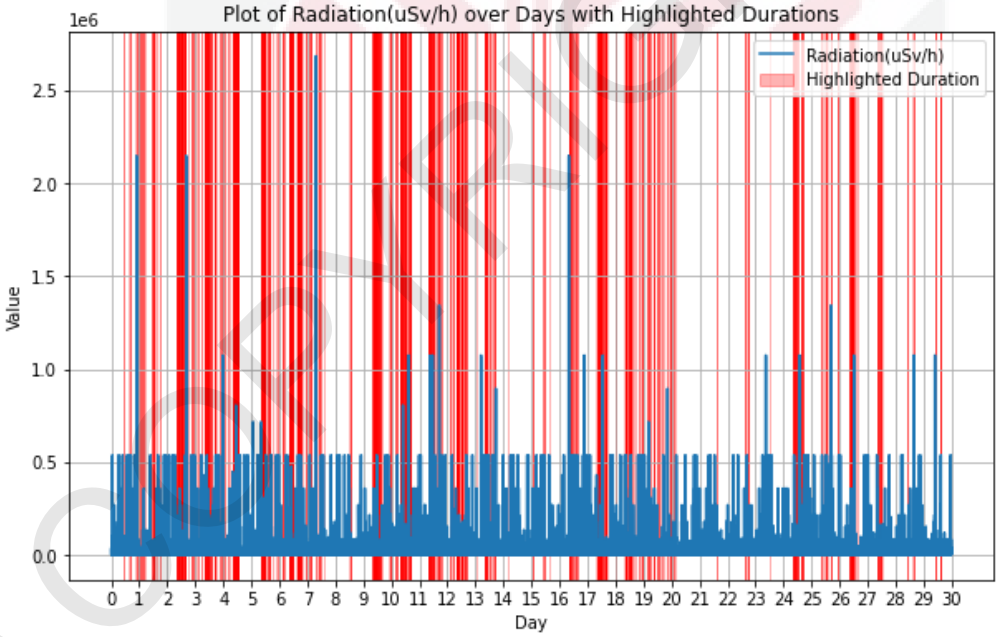


Figure 4.10: Radiation data for one month with highlighted scanning duration.

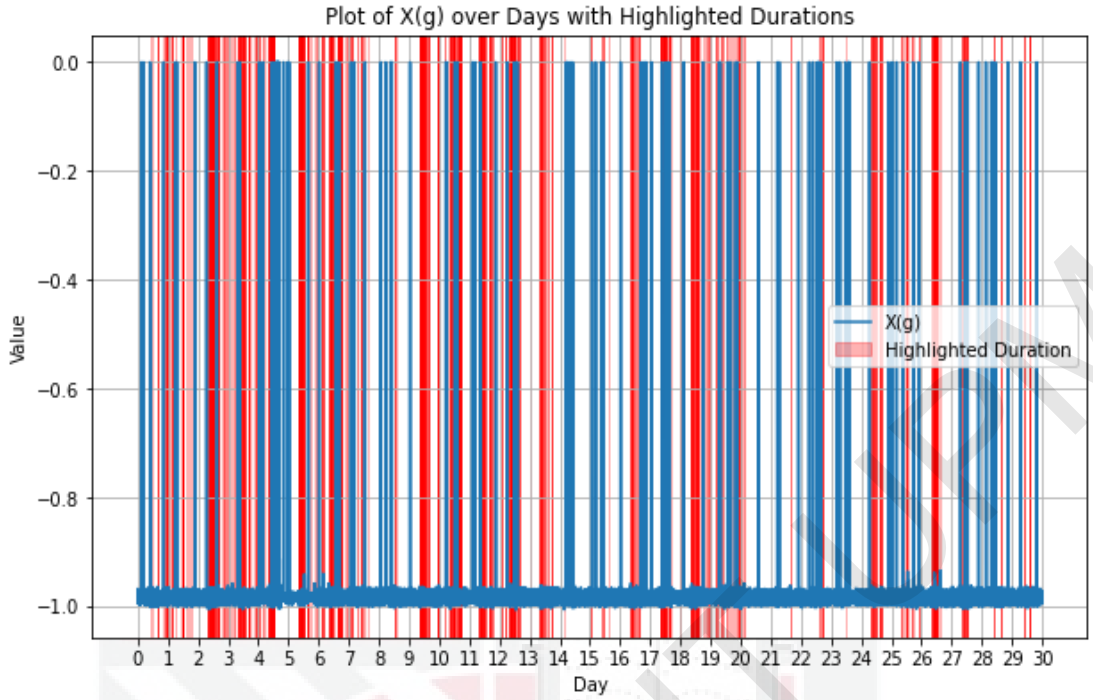


Figure 4.11: X axis data for one month with highlighted scanning duration.

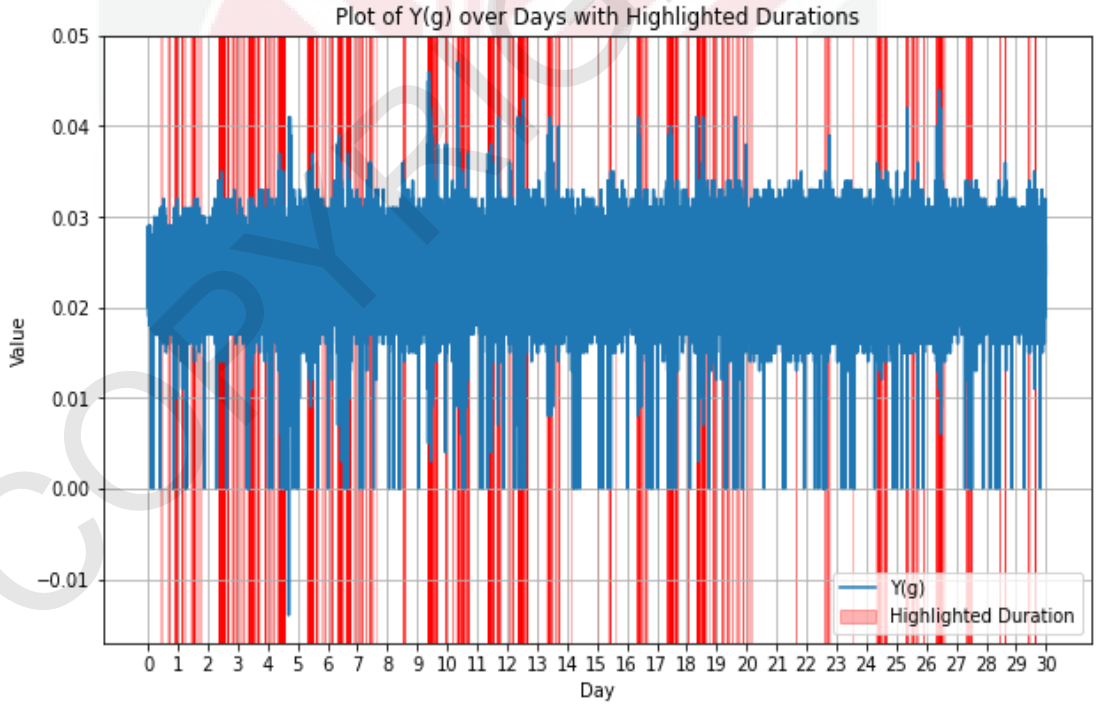


Figure 4.12: Y axis data for one month with highlighted scanning duration.

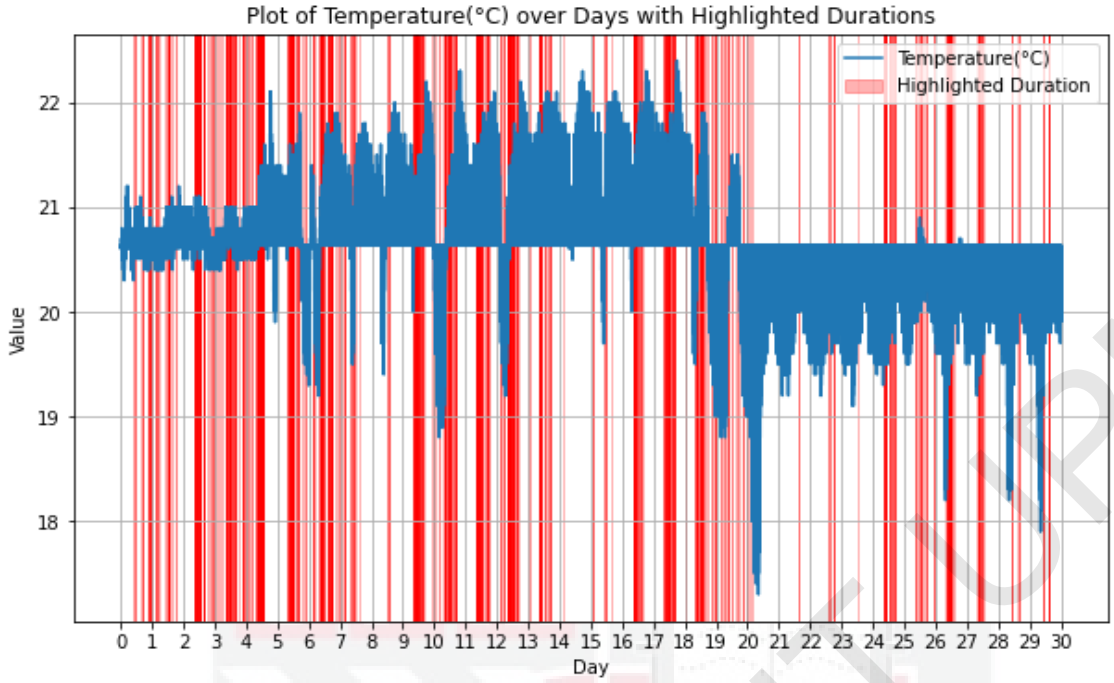


Figure 4.13: Temperature data for one month with highlighted scanning duration.

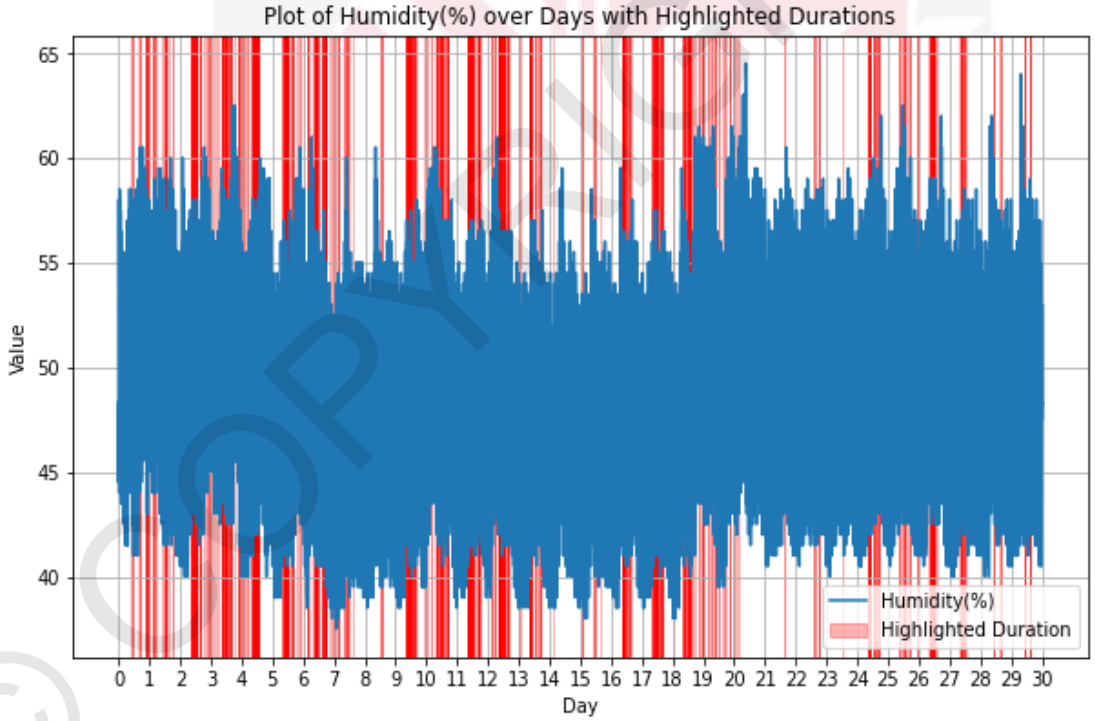


Figure 4.14: Humidity data for one month with highlighted scanning duration.

4.4 Synthetic Data Generation

Before generating SD using NA and TGAN, MD was applied for anomaly detection on the original dataset. This method helped identify outliers, particularly abnormal data points essential for predictive maintenance. However, due to the limited number of abnormal instances, the dataset was imbalanced, making it challenging to develop a robust detection model. To address this, NA was used to expand the tails of the data distribution by introducing controlled noise, creating additional abnormal data points. Simultaneously, TGAN was employed to generate synthetic data that accurately captured the underlying distribution of the real dataset, ensuring the synthetic data maintained key statistical properties. These combined methods aimed to create a more balanced dataset, with MD being applied again to evaluate the effectiveness of anomaly detection on the newly generated data.

4.4.1 Anomaly Detection using Mahalanobis Distance

The outlier detection process was carried out using MD on six operational features: current, radiation, temperature, humidity, and acceleration along the X and Y axes. A pairwise comparison of these features is depicted in the grid of scatter plots shown in Figures 4.15 and 4.16. Each scatter plot represents the relationship between two features, with an ellipse (labeled with a red line) indicating the 95% confidence level based on the critical MD value of 3.5485. The ellipses in each plot define the boundary where the data points are considered normal. Any data points falling outside of these ellipses are identified as outliers, signaling potential abnormal operating conditions in

the CT scan machine. These anomalies could indicate irregularities that may lead to operational issues or machine failure if not addressed.

To interpret the scatter plots, the first row illustrates the current value on the x-axis against other features on the y-axis. Subsequent rows follow a similar pattern which are the second row displays radiation on the x-axis and the other features on the y-axis, the third row features temperature, the fourth row shows humidity, and the fifth and sixth rows display acceleration along the X and Y axes, respectively. This arrangement helps in visually comparing each feature against the others to spot anomalies and trends.

The 95% confidence interval ensures that most of the normal data points are captured within the ellipse, providing a robust method for identifying outliers. The assumption here is that data points within the ellipse are considered normal, while those outsides are presumed to be abnormal. This assumption is critical for the effective implementation of predictive maintenance as it directly influences the decision-making process regarding the health of the CT scan machine. Additionally, based on the scatter plots in Figures 4.15 and 4.16, the color of the scatter plots changes based on the density of data points, where yellow indicates a high density of data, and dark blue indicates a low density. This color coding helps in visually assessing the concentration of data points and further aids in identifying areas of interest for anomaly detection. The visualization in Figures 4.15 and 4.16 effectively demonstrates how MD is used for multivariate anomaly detection, helping to identify abnormal conditions that require attention for predictive maintenance purposes. The plot shows the ellipse for every feature that is used in the ANN model development and the ellipse indicate

which data point is inside the range of values (inside ellipse) and outliers (outside ellipse).

Figures 4.17 and 4.18 depict scatter plots that incorporate synthetic data generated using the NA method, where the data is spread widely due to extending the head and tail of the distribution. This approach aims to enrich the dataset by enhancing the representation of abnormal data points, which are critical for the predictive model's accuracy. On the other hand, Figures 4.19 and 4.20 show scatter plots utilizing synthetic data generated by the TGAN. Unlike NA, TGAN focuses solely on using the abnormal data to generate new synthetic instances, maintaining a closer fidelity to the original data distribution but enhancing the dataset's diversity.

Based on the MD results, 1,959 data points were labeled as abnormal out of a total of 46,095, highlighting the minority presence of abnormal conditions. To address this imbalance, synthetic data generation can be used to create a balanced dataset, which can then be fed into the ANN model for improved training and prediction accuracy.

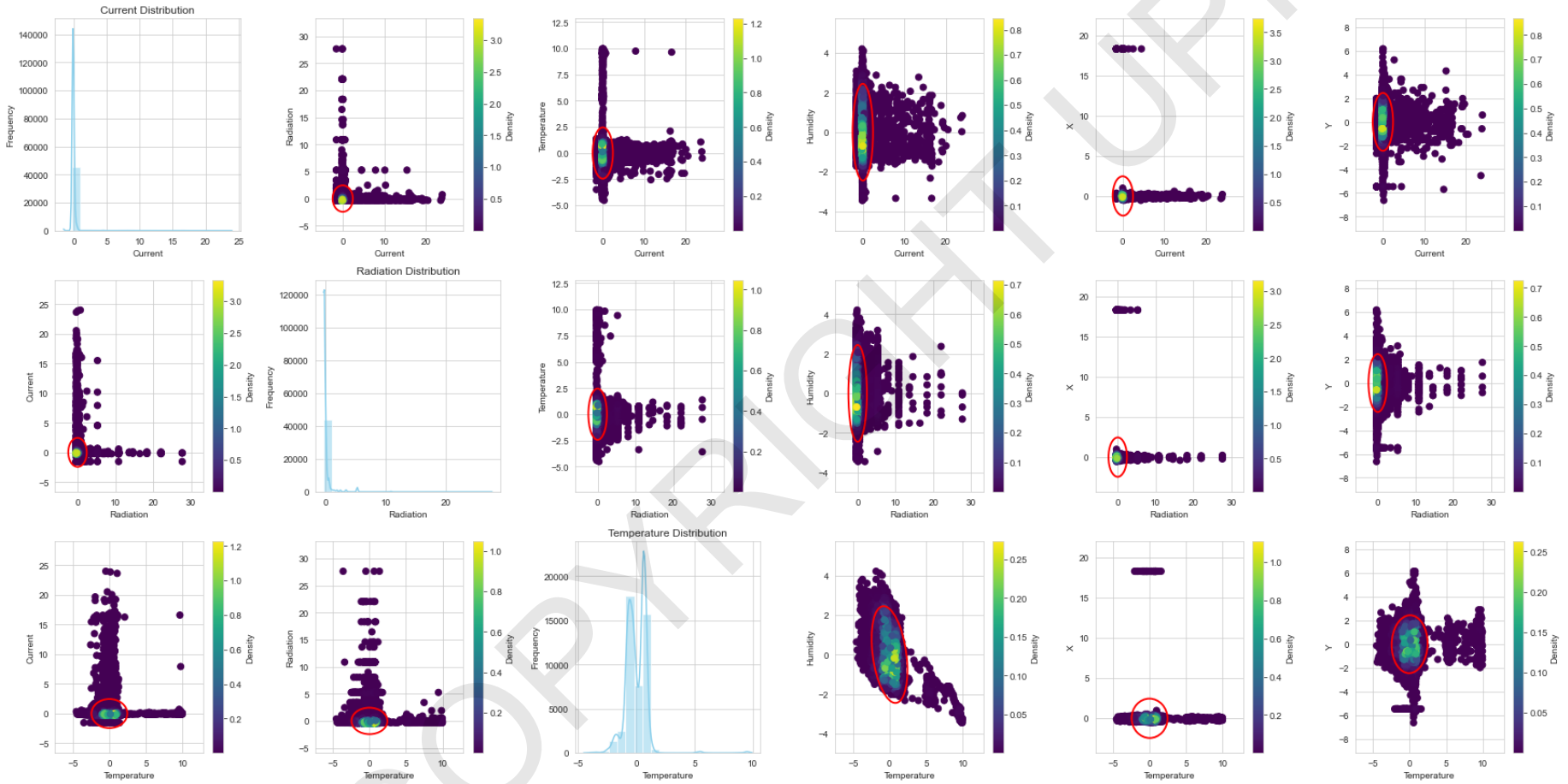


Figure 4.15: Ellipse generated from MD on multivariate original data (upper view).

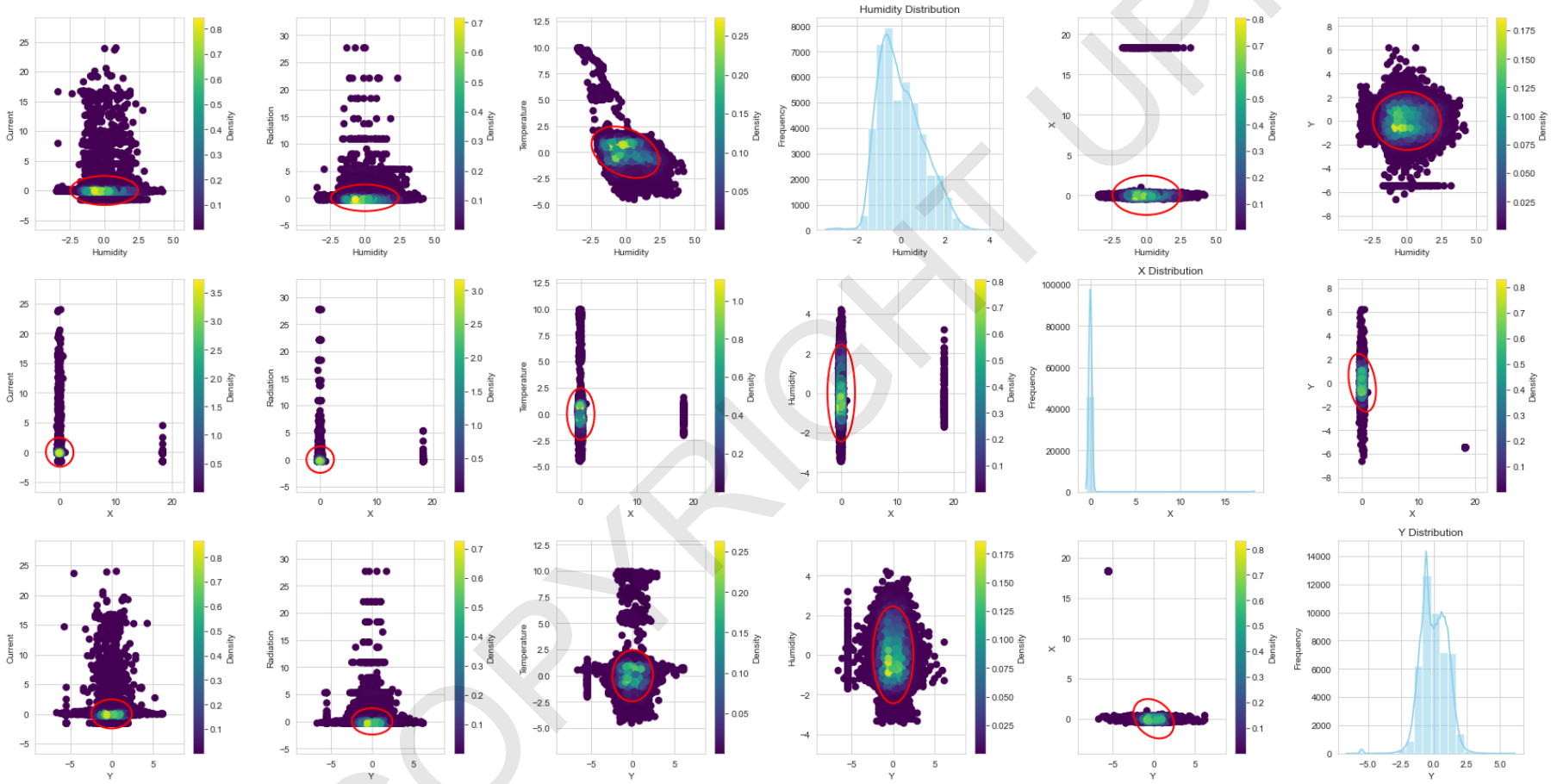


Figure 4.16: Ellipse generated from MD on multivariate original data (lower view).

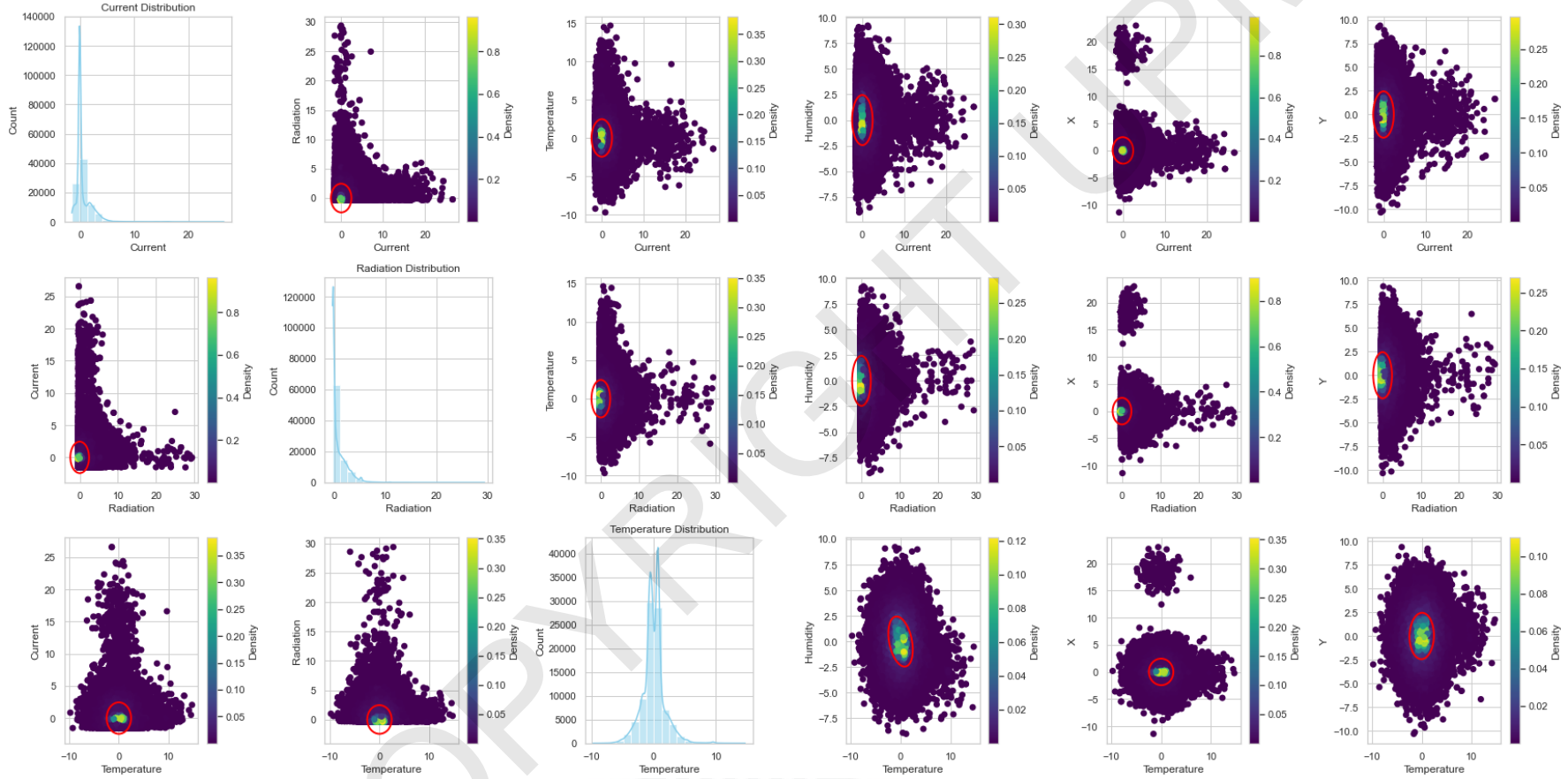


Figure 4.17: Data distribution after apply NA (upper view).

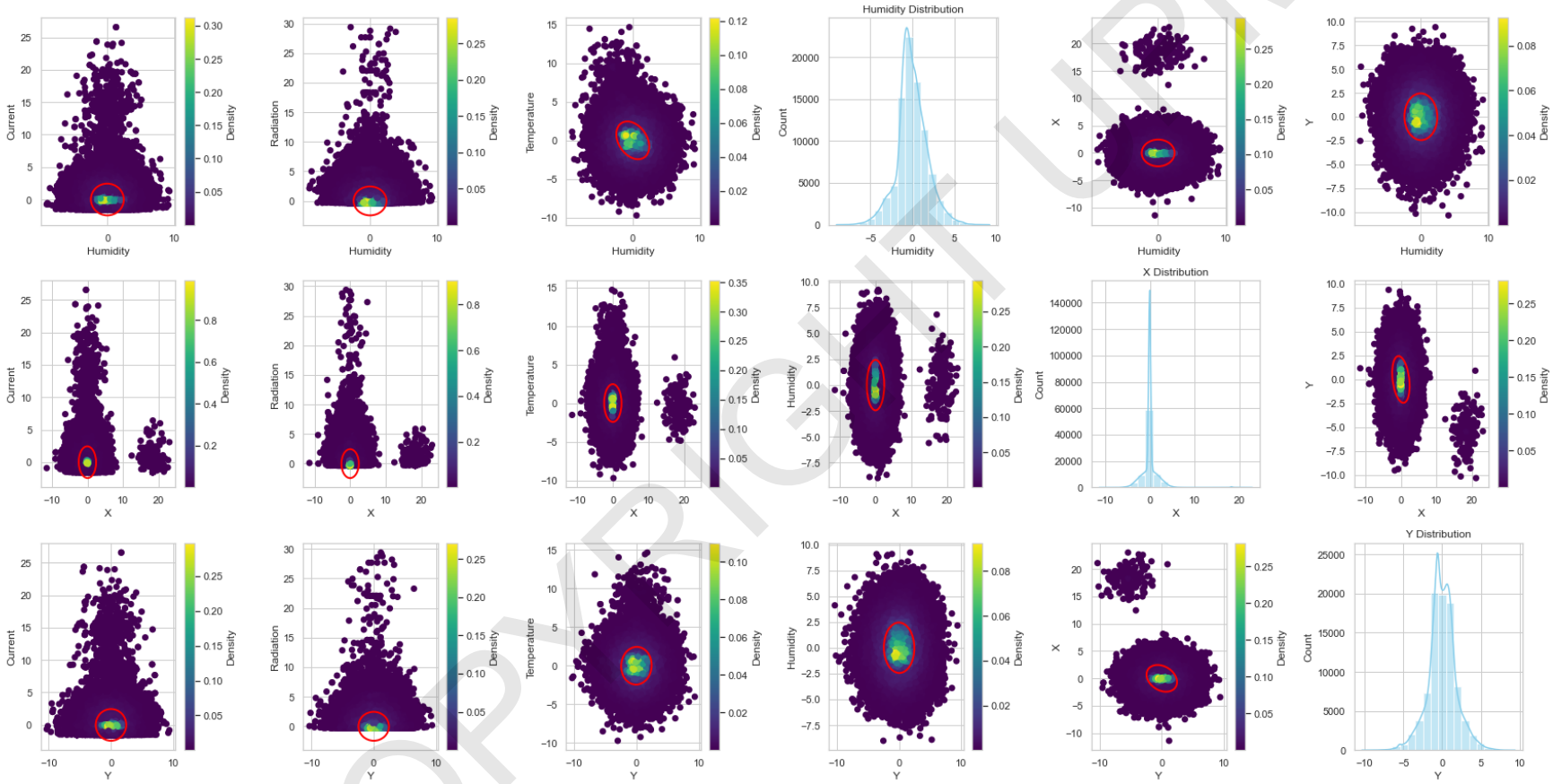


Figure 4.18: Data distribution after apply NA (lower view).

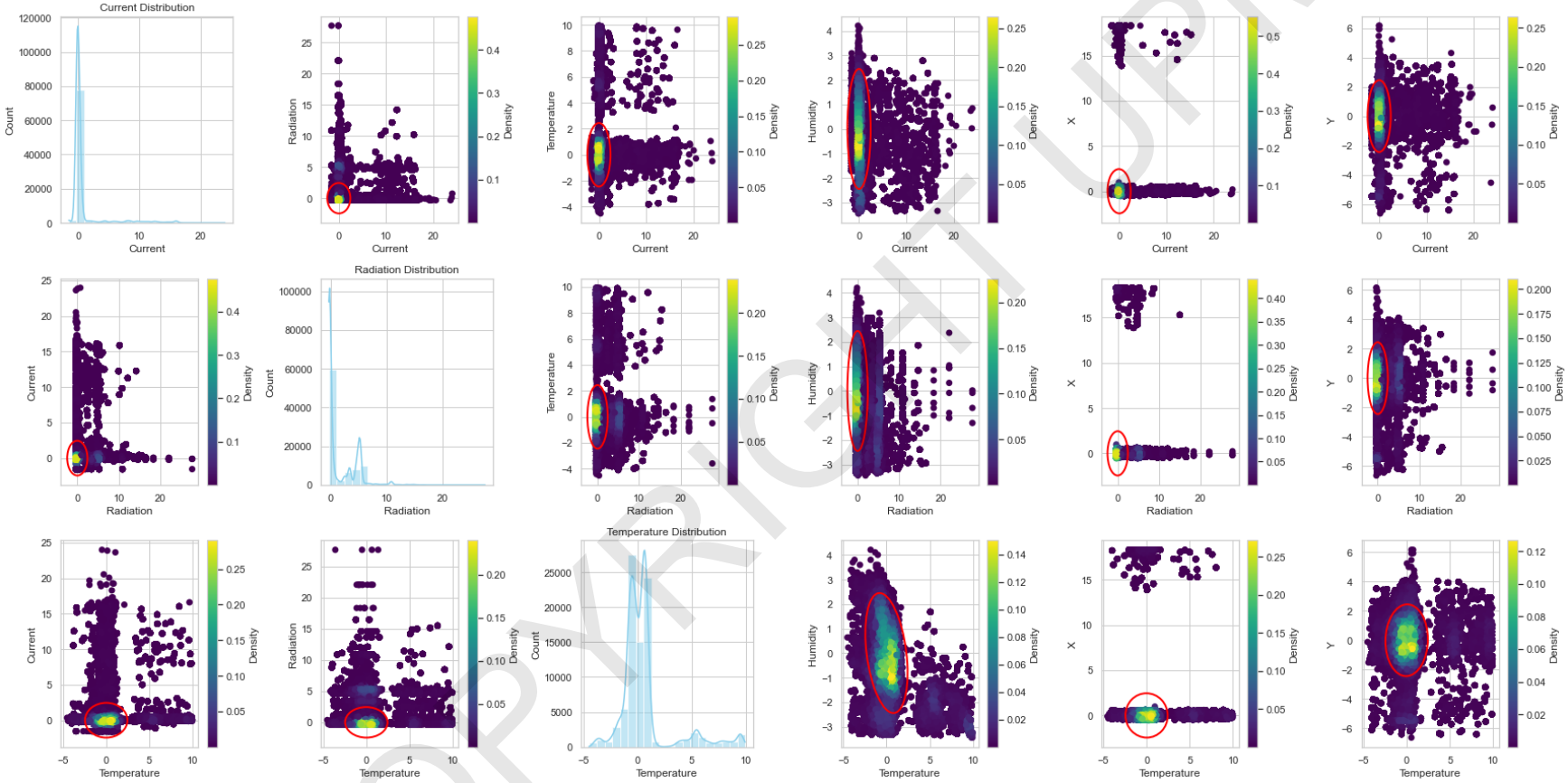


Figure 4.19: Data distribution after apply TGAN (upper view).

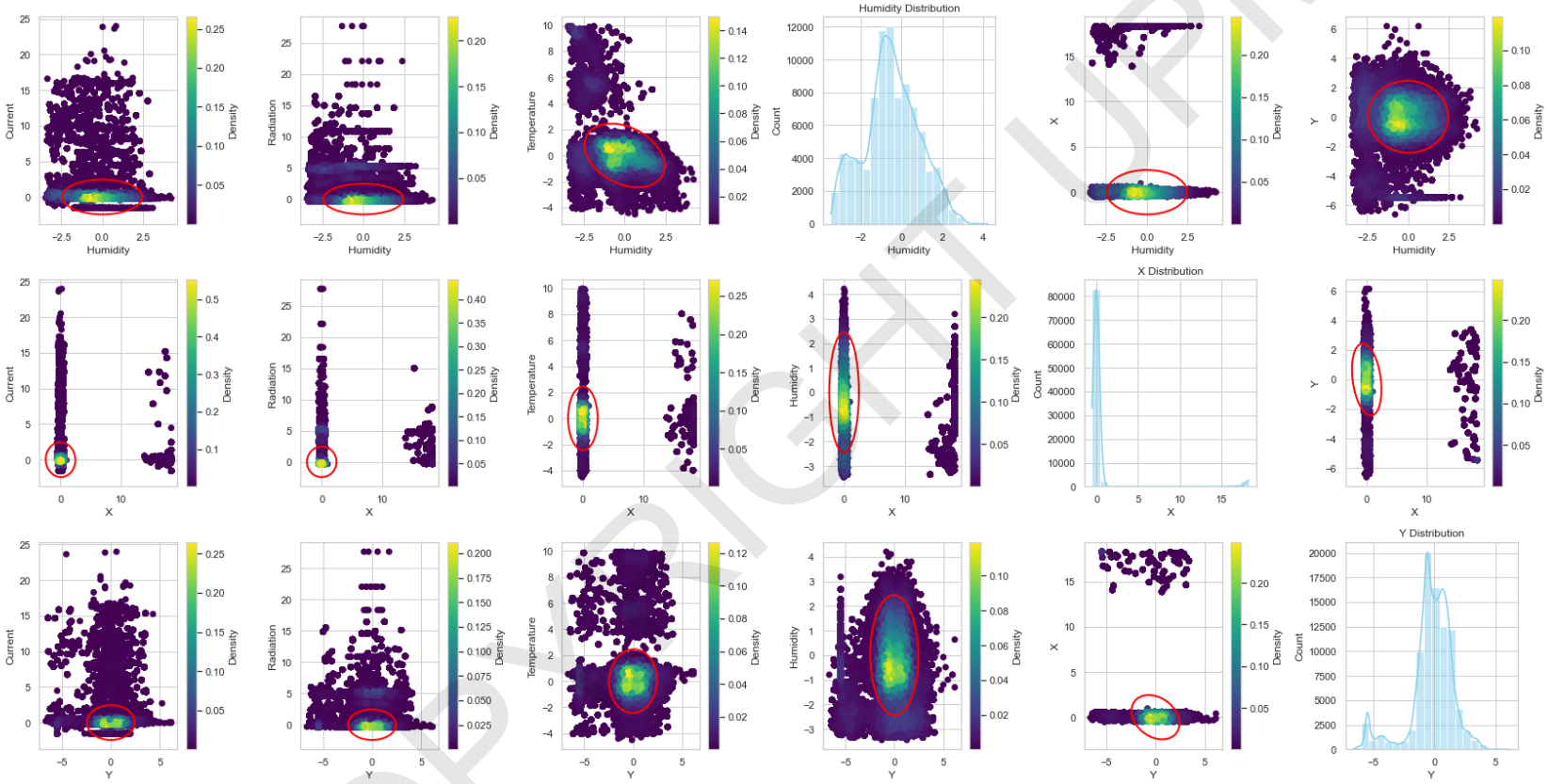


Figure 4.20: Data distribution after apply TGAN (lower view).

4.4.2 Synthetic Data Generation using Noise Addition

In the process of generating SD using NA, a NF of 2 was selected after careful consideration of how various noise levels impacted the dataset. This decision was made to strike a balance between introducing realistic variability and maintaining adherence to the sensor specifications, particularly for humidity. During the evaluation of different NF, it was found that using an NF greater than 2 caused the humidity values to exceed the sensor's specified range of 0% to 100%. Since humidity is physically constrained within this range, exceeding it would result in unrealistic data, which would not be useful for predictive maintenance models aimed at detecting real-world anomalies. With an NF of 2, the generated data for humidity remained within acceptable limits, ranging from 3.65% to 100.21%, ensuring that the noise introduced appropriate but realistic variations.

For radiation, the SD ranged from 0 µSv/h to 2,773,961.52 µSv/h, and any negative values were discarded as they are physically impossible for radiation readings. While the SD exceeded the sensor's maximum measurement capability of 36,000 µSv/h, this was not considered problematic because the radiation readings in this dataset represent cumulative values over a 3-minute period. The higher values reflect the cumulative nature of the measurements rather than real-time sensor limitations, making the exceedance acceptable for this use case.

For current, the SD ranged from 0A to 72.49A. As with radiation, negative values were discarded, as they are not realistic in normal operating conditions. The upper limit remained within the sensor's maximum of 500A, ensuring that the generated data

remained realistic while still expanding the head and tail of the distribution. For the X-axis and Y-axis accelerations, the generated data ranged from -1.44g to 0.26g and -0.03g to 0.06g, respectively, which remained well within the sensor's range of -16g to +16g for both axes. These values introduced slight variations in acceleration, providing realistic but expanded distributions without generating extreme outliers. For temperature, the SD ranged from 12.85°C to 36.19°C, which was comfortably within the sensor's operational range of -20°C to 70°C. The NA produced moderate variability in this feature, maintaining a realistic temperature range.

This synthetic dataset will undergo MD analysis again, using the centroid of the real data to calculate the distance and verify which points are abnormal. This process will be repeated until the synthetic abnormal data has the same amount as the normal real data. Figure 4.21 displays the difference in distribution between the real data and SD using NA, showcasing the expansion of both the head and tail of the data distribution without introducing unrealistic negative values for any of the features. The blue bars represent the distribution of the original data, while the red bars show the distribution after noise has been introduced. The overlap between the original and noisy data signifies that the noisy data remains within the normal operating range of the sensor. The figure demonstrates that, despite the introduction of noise, the synthetic data maintains a reasonable distribution relative to the original, staying within acceptable sensor limits. Additional results using different noise factors are included in Appendix D for further comparison.

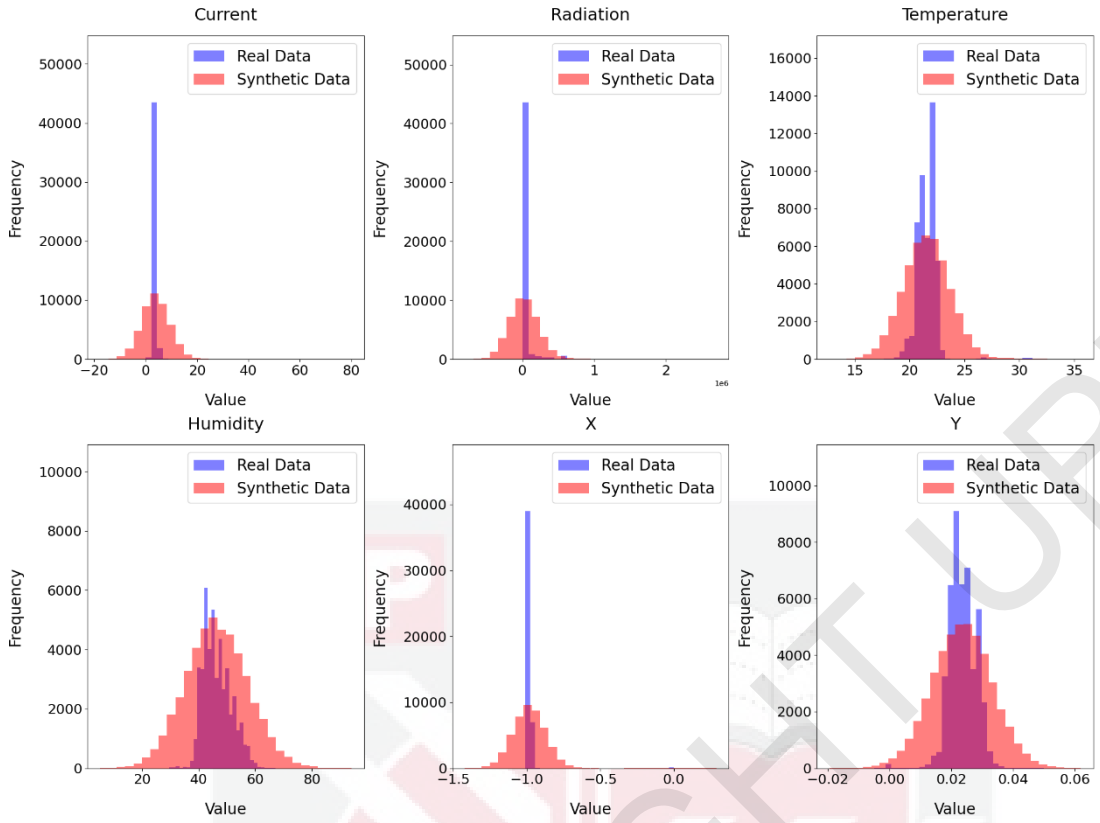


Figure 4.21: Distribution of real and synthetic data using NA.

The SD generation process was conducted in three cycles, producing different quantities of synthetic abnormal data points: 15747 in the first cycle, 21073 in the second, and 15236 originally generated in the third. However, to align the synthetic data with the actual distribution of real data, only 5357 data points from the third cycle were randomly selected for use. This approach was necessary to balance the number of synthetic abnormal data points with the real abnormal data, which totaled 1959. Thus, the combined total of synthetic and real abnormal data reached 44136, precisely matching the count of real normal data. The final dataset for the ANN model training comprised 44136 abnormal data points (both real and synthetic) balanced with 44,136 real normal data points, bringing the total dataset to 88272 data points. All of the detail about the data show in Table 4.1 below.

Table 4.1: Distribution of real and synthetic data using NA.

Real normal			44136
Real Abnormal			1959
Synthetic Abnormal	Cycle 1	15747	42177
	Cycle 2	21073	
	Cycle 3 (Randomly Select)	5357	
Total			88272

4.4.3 Synthetic Data Generation using Tabular Generative Adversarial Network

The results of using TGAN to generate synthetic abnormal data demonstrate a successful balancing of the dataset, where the abnormal real data was initially a minority. The TGAN model was applied solely to the abnormal real data, allowing it to focus on generating realistic synthetic samples that enhance the representation of these critical points. As depicted in Figures 4.22, the distribution of synthetic abnormal data closely aligns with the real abnormal data. Prior to the generation of SD, the dataset was heavily skewed towards normal instances, limiting the model's ability to accurately detect and predict anomalies. After TGAN augmentation, the dataset became more balanced, which is essential for improving the model's training and performance, particularly in detecting rare abnormal conditions. In comparison, the NA method generated SD using the entire dataset, including both normal and abnormal instances, resulting in a broader augmentation. However, TGAN's targeted approach on abnormal data provided a more effective solution for balancing the minority class, enhancing the model's ability to focus on anomaly detection.

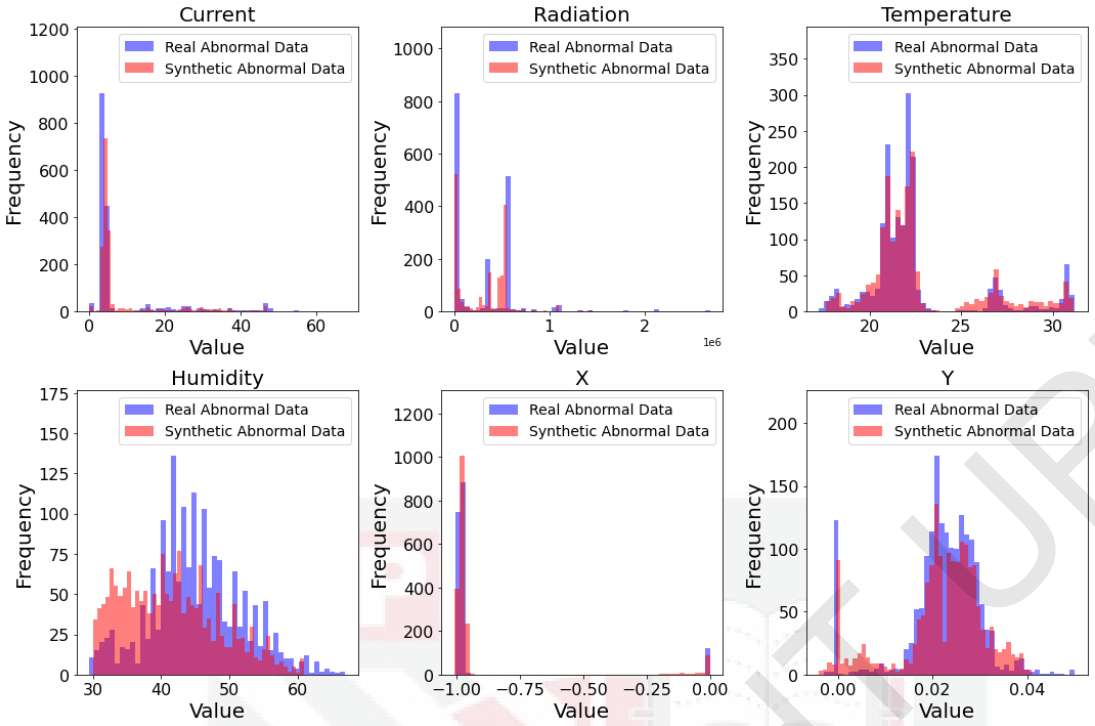


Figure 4.22: Distribution of real and synthetic abnormal data using TGAN.

The KL Divergence results provide a quantitative measure of how closely the SD generated by TGAN matches the real data for each feature. The KL Divergence values are as follows, with Current being 2.789, Radiation at 1.568, Temperature at 0.257, Humidity at 0.228, X at 0.215, and Y at 0.938. These values are close to 0 and indicate that SD closely aligns with the real data, as evidenced by the low KL Divergence values. KL Divergence, also known as Kullback-Leibler Divergence, is a statistical measure used to determine how one probability distribution diverges from a second, expected probability distribution. It quantifies the amount of information lost when using one distribution to approximate another, thus providing a metric for the similarity between the generated synthetic data and the original real data distributions.

The correlation matrices shown in Figure 4.23 provide a detailed comparison of the relationships between features in the real and synthetic datasets. The matrix on the left represents the real data, while the one on the right corresponds to the SD generated

using TGAN. In both matrices, correlation values range from -1 to 1, where values closer to 1 indicate a strong positive correlation, and those closer to -1 indicate a strong negative correlation between features. A value close to 0 suggests little or no correlation.

Upon examining the two matrices, it was observed that TGAN successfully replicated the general structure of correlations from the real data. For example, key relationships between features are preserved across both datasets. However, there are some minor discrepancies in the strength of correlations, such as the correlation between specific feature pairs (e.g., -0.7 in the real data versus -0.41 in the SD for temperature, humidity, x-axis and y-axis features). These differences may arise due to TGAN's focus on generating SD for the minority abnormal class, which can slightly alter some feature relationships. Despite these variations, the overall similarity between the two matrices indicates that TGAN maintains the essential correlations, ensuring the SD remains representative of the real data for the predictive maintenance task.

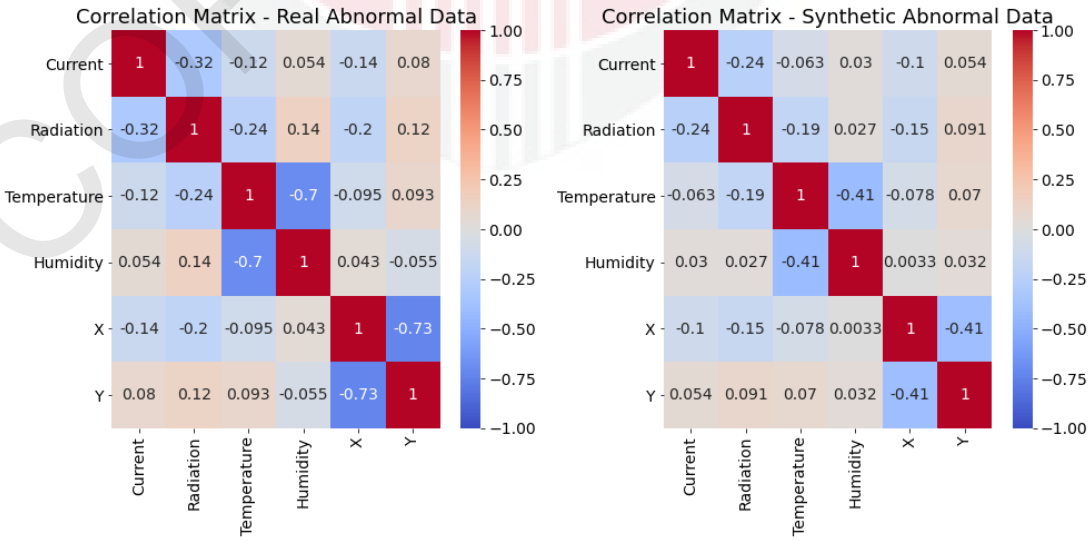


Figure 4.23: Correlation matrix of real and synthetic data using TGAN.

The Kernel Density Estimation (KDE) plots presented in Figures 4.24 compare the distribution of real data (orange) and SD (blue) generated using TGAN. KDE plots provide a smooth representation of the data's probability density, making it easier to observe the overall distribution patterns. In most plots, the SD closely aligns with the real data, with similar peaks and downturns indicating that TGAN has effectively captured the underlying distribution.

However, there are slight differences in some regions where the SD shows higher or lower peaks compared to the real data. These variations suggest that while TGAN has successfully modeled the overall structure, some deviations are present, particularly in the tail regions of the distributions, where the density of the SD either slightly overestimates or underestimates the real data. Despite these minor discrepancies, the KDE plots demonstrate that the SD retains the key characteristics of the real data, ensuring its usefulness for tasks like PdM.

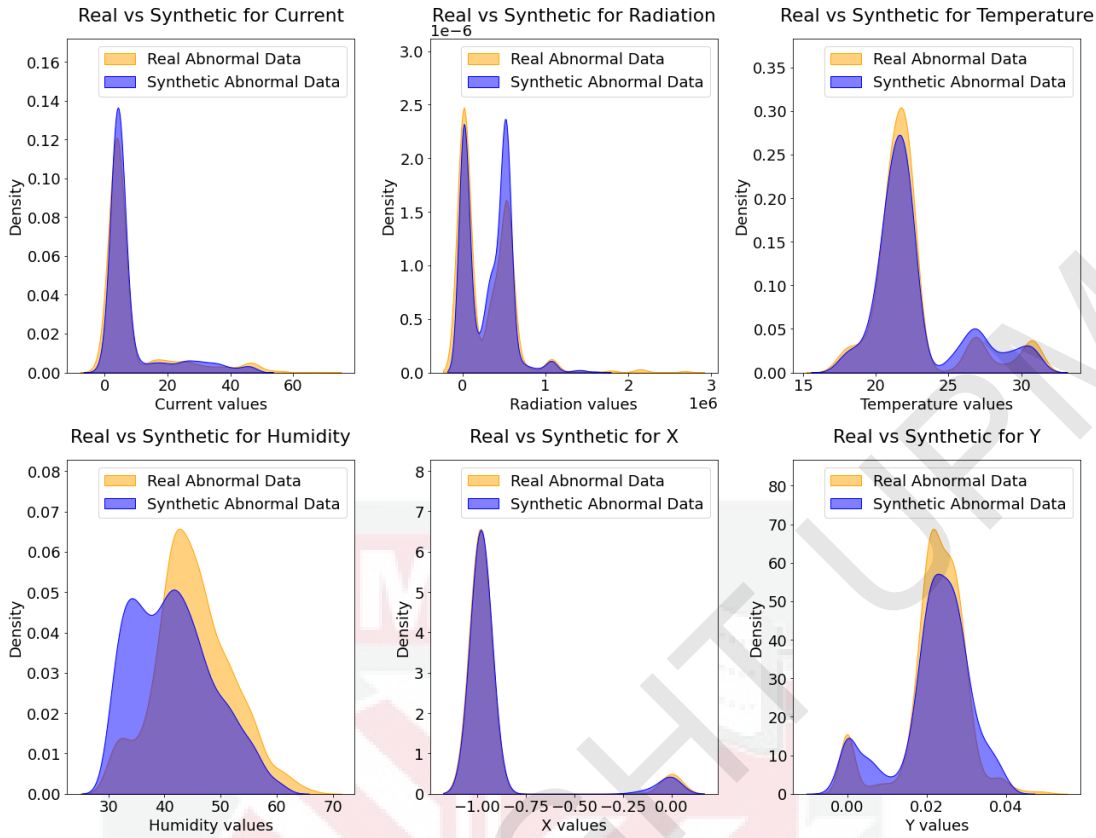


Figure 4.24: KDE plot of real and synthetic abnormal data.

The TGAN-based SD generation process effectively balances the dataset by generating abnormal data until the total of synthetic abnormal and real abnormal data equals the amount of real normal data. Notably, this TGAN generates 42177 rows of SD directly without needing to run the process multiple times, as TGAN efficiently learns and replicates the data distribution in a single run. The KDE plots, correlation matrices, and KL Divergence results confirm that TGAN preserves the key characteristics and relationships of the real data, though some discrepancies are evident, particularly for features like current and radiation. Despite these differences, the overall alignment between the real and SD ensures that the SD is representative and valuable for enhancing model performance in predictive maintenance tasks, especially in improving the detection and analysis of abnormal conditions.

4.5 Machine Learning Model Performance

In this analysis, an ANN model was utilized to predict anomalies in a CT scan machine. The dataset includes both real-world abnormal data and synthetic abnormal data generated using NA and TGAN. By combining these approaches, the class imbalance problem was mitigated, enabling the model to generalize more effectively. The performance of the ANN model was evaluated using various metrics, providing insights into its ability to distinguish between normal and abnormal machine conditions. Figures 4.25 and Figure 4.26 show the training and validation graph for both methods. The overall comparison of the model performance, as shown in Table 4.2, highlights that the TGAN method consistently achieves higher accuracy and lower loss compared to the NA method. Specifically, TGAN results in an average accuracy of 99.53%, which is 0.89% higher than NA's average accuracy of 98.64%. Additionally, the TGAN method has a significantly lower average loss of 1.43%, compared to the NA method's 4.05%. These results confirmed the effectiveness of TGAN in generating SD and improving the overall performance of the PdM model. The benchmark performance from the literature review is discussed in section 2.8.

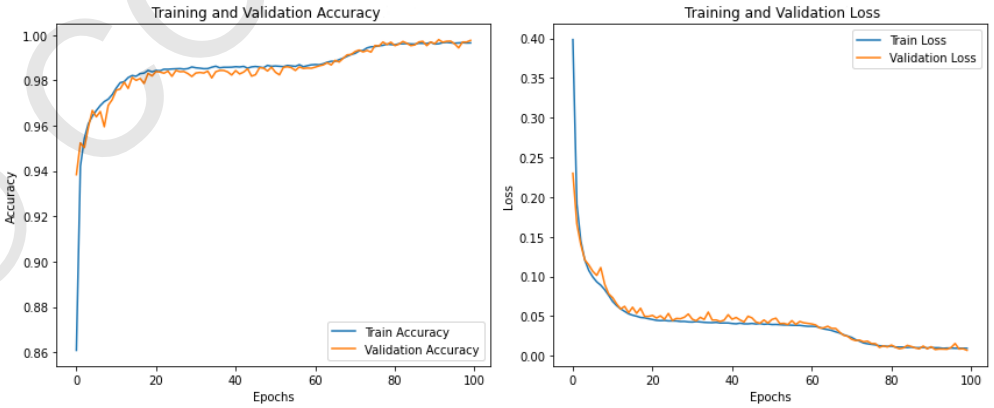


Figure 4.25: Training and validation graph using NA.

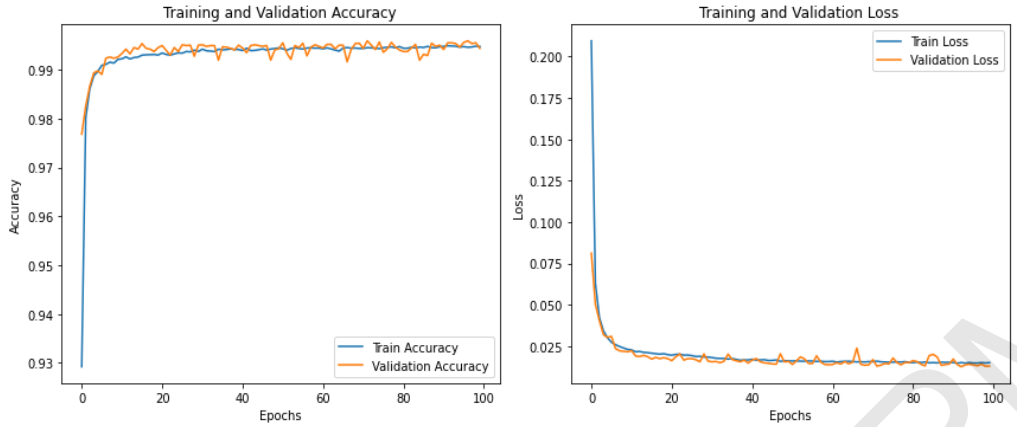


Figure 4.26: Training and validation graph using TGAN.

Tables 4.3 to 4.4 and Figures 4.27 to 4.28 compare the performance of the ANN model using NA and TGAN data generation methods. In Table 4.2, the NA method shows strong results with precision of 0.97 for class 0 and 0.99 for class 1, while recall and F1-scores are both around 0.98. Table 4.3, showing the TGAN method, achieves perfect precision, recall, and F1-scores of 1.00 for both classes, clearly outperforming the NA method. The ROC curves in Figures 4.19 (NA) and 4.20 (TGAN) show perfect discrimination, each with an AUC of 1.00. Despite similar ROC results, the TGAN method consistently provides better overall performance with flawless accuracy, precision, recall, and F1-scores. While the NA method performs well, it does not match the robustness of TGAN in these metrics. This comparison demonstrates that although both methods are highly effective, TGAN offers superior reliability. From these evaluations, it is observed that TGAN is successfully generated SD for abnormal class better due to the generator and discriminator operations which eliminate the unrealistic abnormal data. The NA generated SD based on the distribution of normal data and may limit the creation of the SD.

Table 4.2: Accuracy and loss of k-fold validation for both methods.

Fold	NA (%)		TGAN (%)	
	Accuracy	Loss	Accuracy	Loss
1	98.83	3.55	99.54	1.45
2	98.80	3.60	99.55	1.36
3	97.76	6.49	99.54	1.43
4	99.65	0.94	99.55	1.39
5	98.15	5.67	99.48	1.51
Average	98.64	4.05	99.53	1.43

Table 4.3: Classification report for NA.

	Precision	Recall	F1-score	Support
Normal (0)	0.97	0.99	0.98	4330
Abnormal (1)	0.99	0.97	0.98	4410
Accuracy			0.98	8740
Macro Average	0.98	0.98	0.98	8740
Weighted Average	0.98	0.98	0.98	8740

Table 4.4: Classification report for TGAN.

	Precision	Recall	F1-score	Support
Normal (0)	1.00	1.00	1.00	4393
Abnormal (1)	1.00	1.00	1.00	4347
Accuracy			1.00	8740
Macro Average	1.00	1.00	1.00	8740
Weighted Average	1.00	1.00	1.00	8740

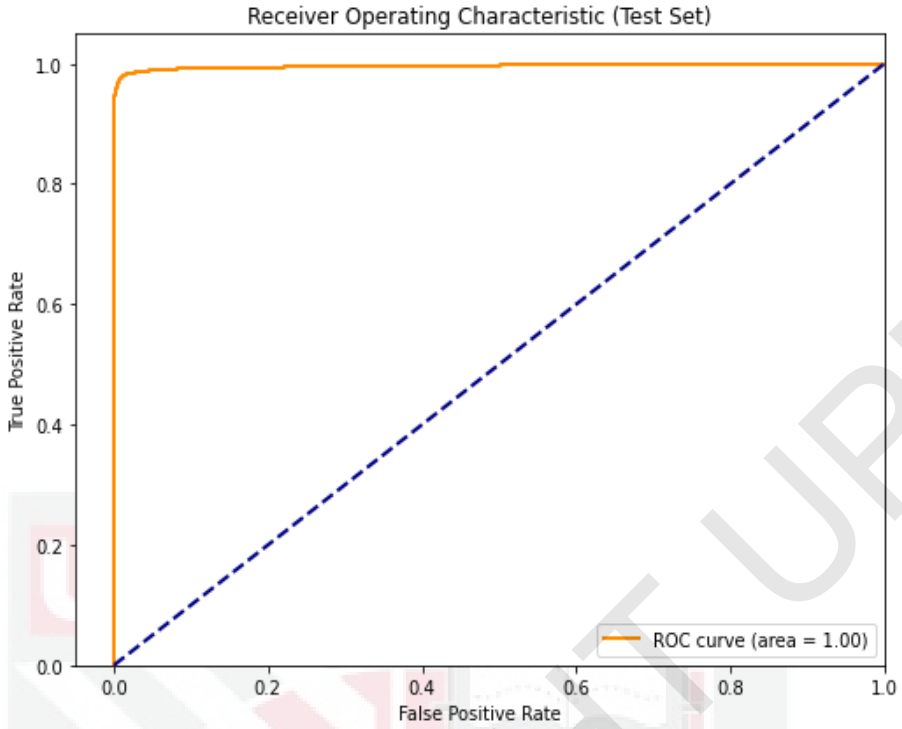


Figure 4.27: ROC AUC for NA.

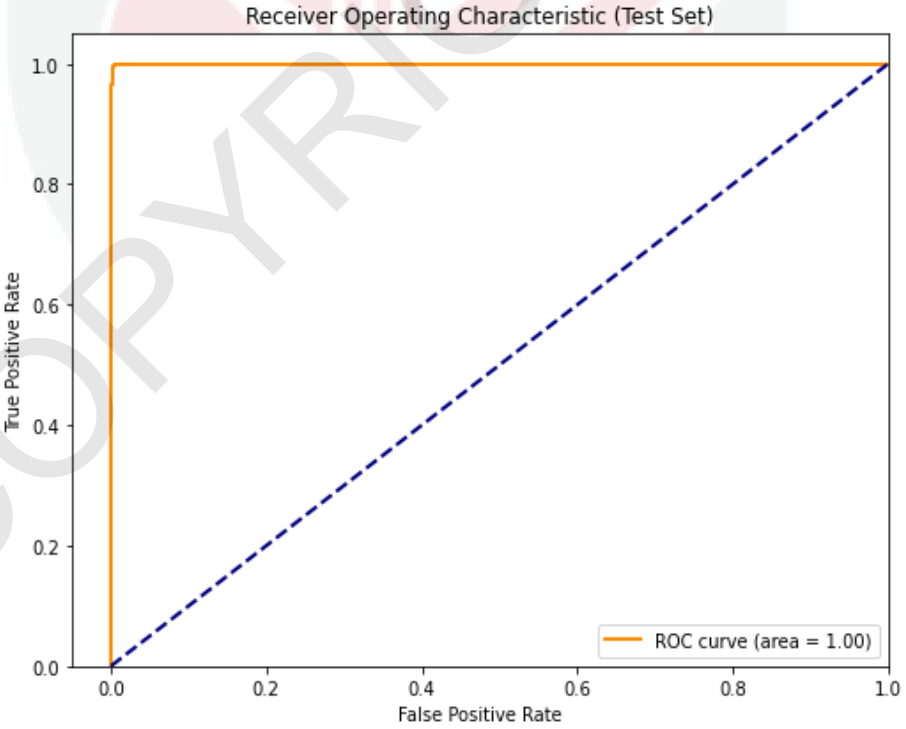


Figure 4.28: ROC AUC for TGAN.

Comparing the confusion matrices in Table 4.5 (NA method) and Table 4.6 (TGAN method) reveals a clear advantage for the TGAN method in terms of classification

performance. In the NA method, 24 instances of class 0 were misclassified as class 1, and 137 instances of class 1 were misclassified as class 0. In contrast, the TGAN method shows much fewer misclassifications, with only 19 misclassified instances for class 0 and 21 misclassified instances for class 1. This difference is particularly notable for class 1, where the TGAN method significantly reduces misclassification errors from 137 to 21. Overall, the TGAN method demonstrates better performance, with fewer misclassified data points, especially for abnormal data, leading to more reliable and accurate model predictions.

The inclusion of synthetic data using the TGAN method significantly enhances the model's predictive capabilities compared to not having synthetic data. Without the addition of synthetic data, the model may suffer from a lack of diversity in the training examples, particularly for rare abnormal conditions that are critical for training robust predictive models. This limitation can lead to higher error rates and less confidence in the model's ability to generalize to new, unseen scenarios. With synthetic data, however, the model gains a more comprehensive understanding of potential variations and complexities in the data, enhancing its ability to accurately identify and classify both normal and abnormal conditions.

Table 4.5: Confusion matrix for NA.

Confusion Matrix NA (Test Set)			
Actual	Normal (0)	4306	24
	Abnormal (1)	137	4273
		Predicted	

Table 4.6: Confusion matrix for TGAN.

Confusion Matrix TGAN (Test Set)			
Actual	Normal (0)	4374	19
	Abnormal (1)	21	4326
		Predicted	

The models produced using both methods (NA and TGAN) were compared separately from the ANN model based on SD generated by expanding the head and tail from Mohd et al. (2023) due to differences in data used during model development. Figure 4.29 presents the prediction result comparison between the NA (orange) and TGAN (green) models, while Figure 4.30 exclusively shows the results from the Mohd et al.

(2023) model (blue). The model validation stage is performed by comparing the prediction results on the collected data from the study site from September 2023 until December 2023. In both figures, the Y-axis represents the average prediction values, and the X-axis represents days, where Day 1 corresponds to 1 September 2023, and Day 122 corresponds to 31 December 2023. The maintenance event occurred on Day 99 (8 December 2023), marked by the dashed line.

In Figure 4.30, the results from the Mohd et al. (2023) model (blue line) indicate that the machine is predicted to break down, as all readings were high before maintenance. In contrast, Figure 4.29 illustrates the comparison between the NA (orange) and TGAN (green) models. The TGAN model provides a clearer distinction between real breakdowns and normal machine conditions compared to NA. The NA model consistently shows high predictions, making it more sensitive, while the TGAN model suggests a lower risk after maintenance compared to Mohd et al. (2023).

The superior performance of TGAN is due to its ability to generate synthetic data using all six input features (current, radiation, temperature, humidity, X, Y) with Mahalanobis Distance (MD) as the output feature. This approach enables TGAN to preserve meaningful feature relationships, ensuring that the generated data closely resembles real-world conditions. For instance, under normal circumstances, low radiation levels correspond to a stable machine state with lower current usage. By maintaining these feature correlations, TGAN produces more accurate breakdown predictions.

In contrast, the NA method, which expands the tail of the data distribution, randomly assigns values to the features. This can result in physically unrealistic combinations, such as high radiation with low current usage, which is unlikely to occur in a real CT scan machine. As a result, the NA model becomes overly sensitive, frequently predicting breakdowns even when the machine is in a normal operating state. Although this sensitivity makes NA useful for early warning detection, it is less reliable than TGAN in distinguishing between normal and breakdown conditions.

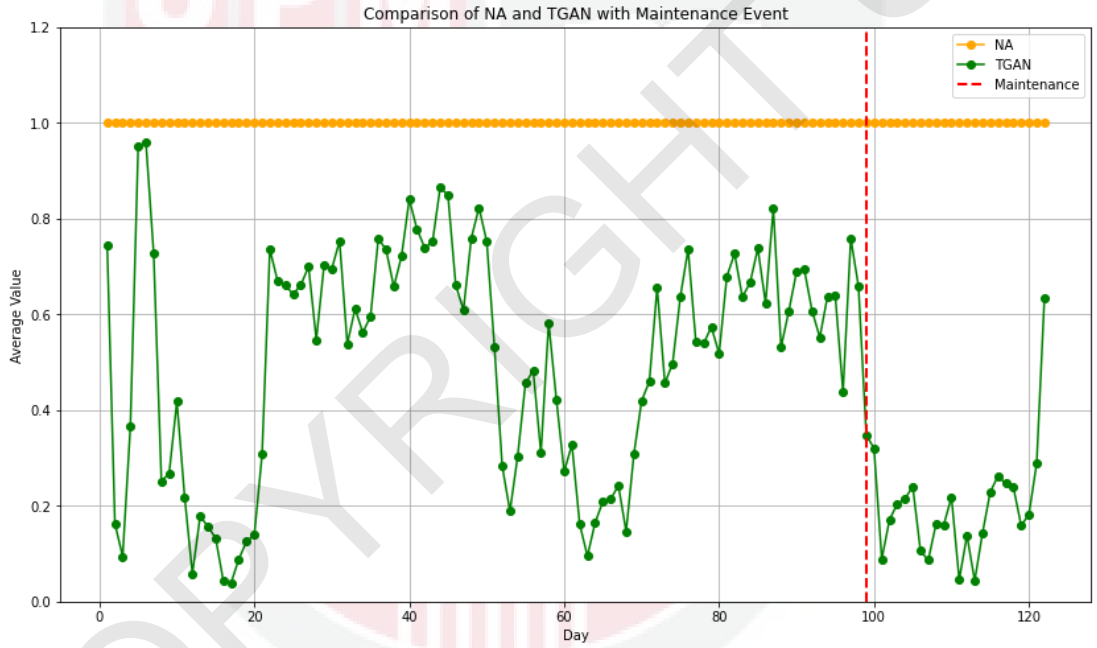


Figure 4.29: Prediction result from NA and TGAN models using data September 2023 to December 2023.

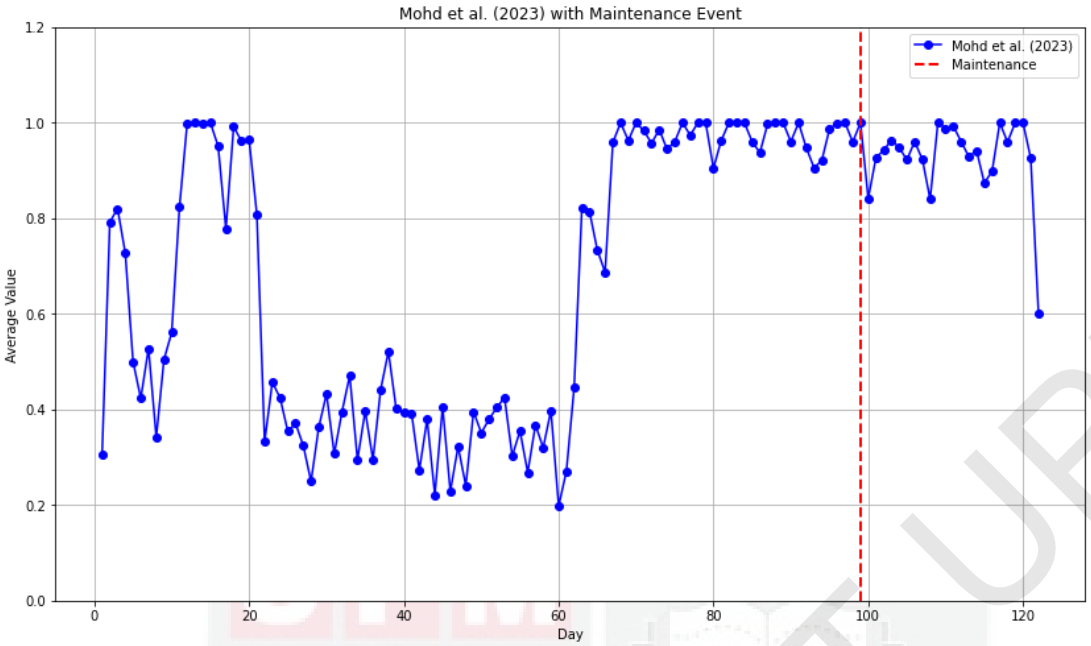


Figure 4.30: Prediction result from Mohd et al. (2023) model using data September 2023 to December 2023.

After investigating, the main reason that led to the need for maintenance, it was determined that temperature was a critical factor. Figures 4.31 to 4.36 show the readings for November for all features, where only the temperature readings were significantly lower than expected, with an average temperature of 21°C. According to Alahmari (2022) operating below the recommended temperature range, typically around 22.2 to 23.8 degrees Celsius (72 to 75 degrees Fahrenheit), can adversely affect the machine's operational efficiency and may lead to increased mechanical wear or failure. Based on these results, it can be concluded that both the TGAN and Mohd et al. (2023) models successfully predict the breakdown. However, it cannot be determined which one outperforms the other since there was only one maintenance report available.

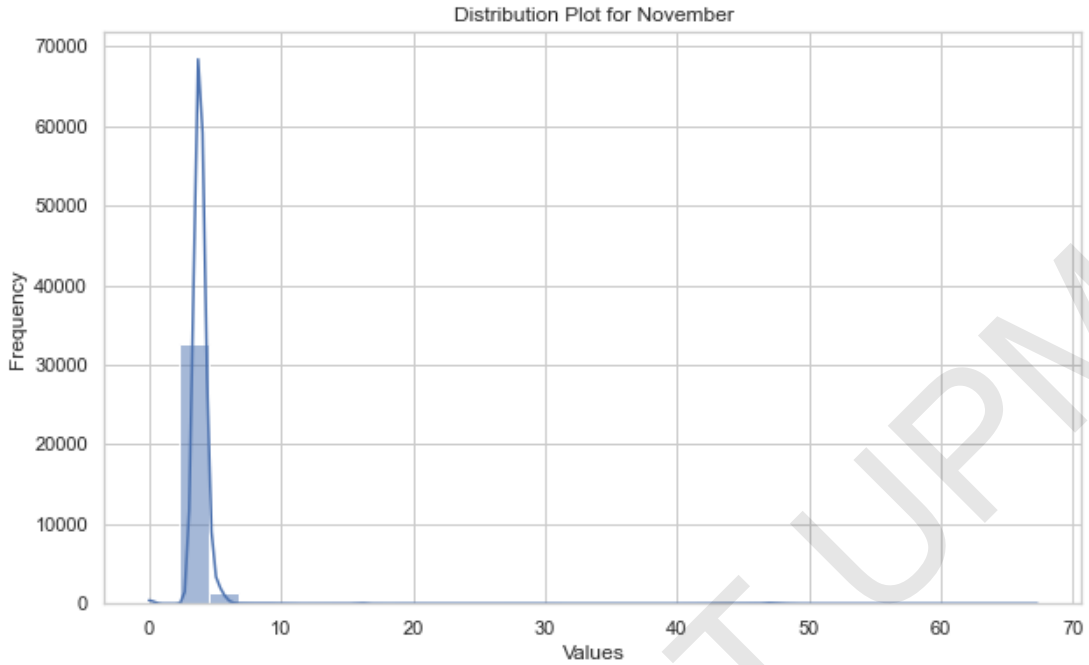


Figure 4.31: Current reading on November 2023.

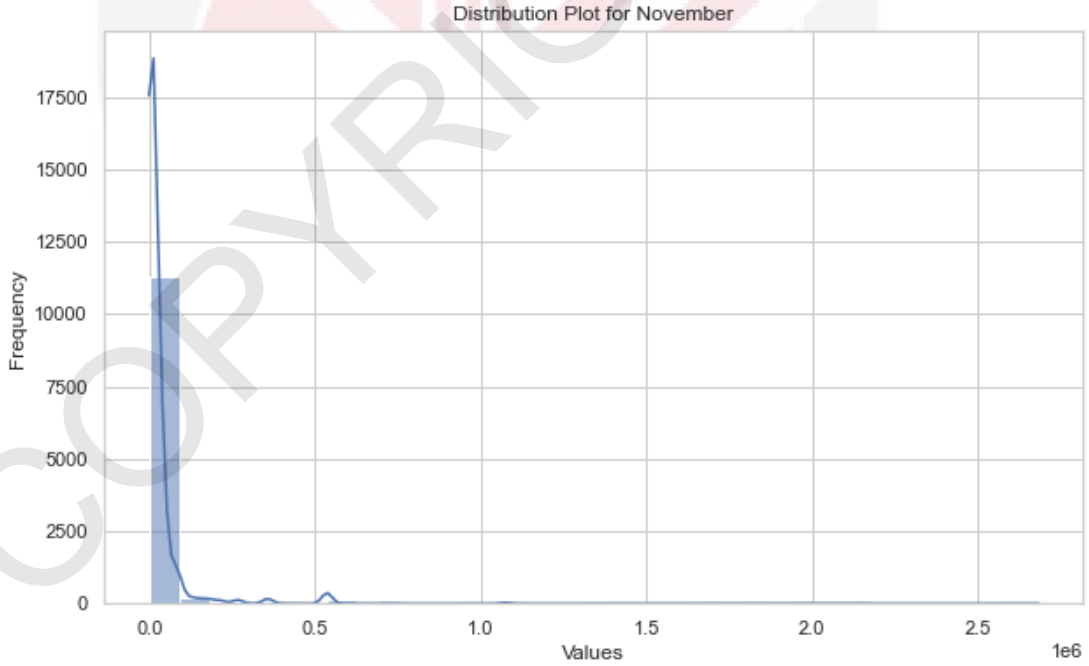


Figure 4.32: Radiation reading on November 2023.

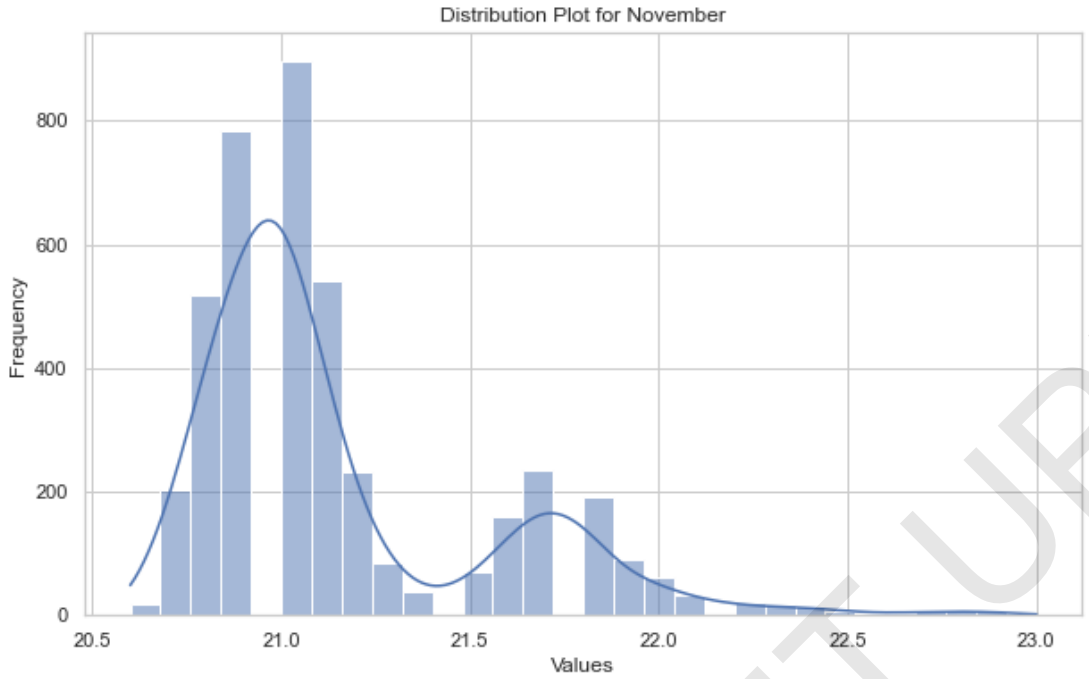


Figure 4.33: Temperature reading on November 2023.

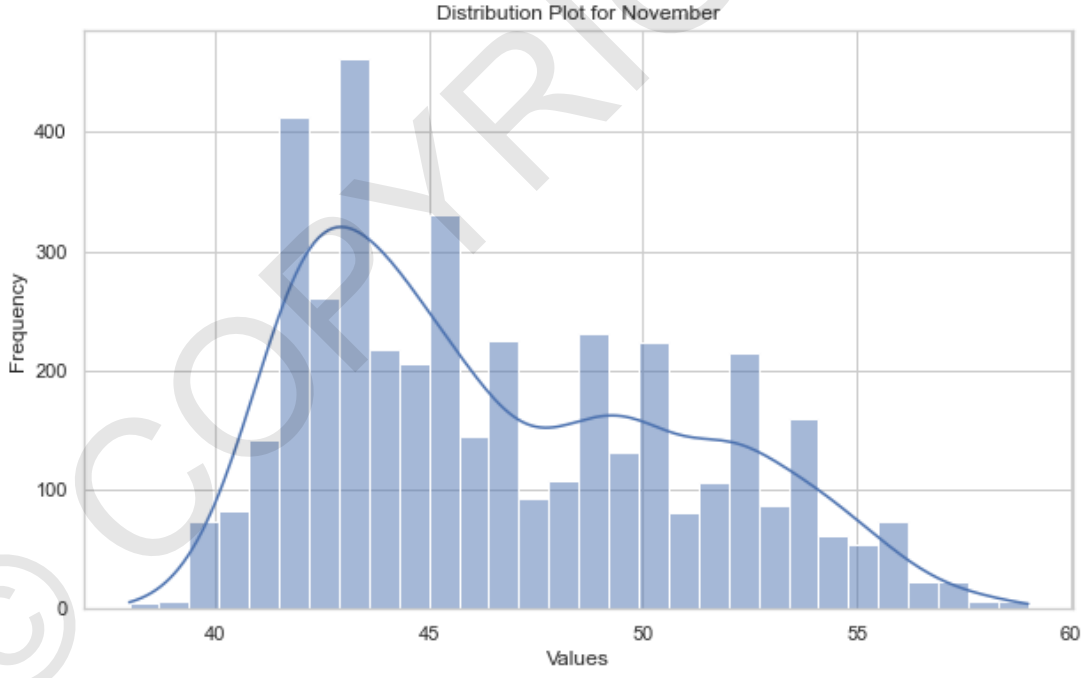


Figure 4.34: Humidity reading on November 2023.

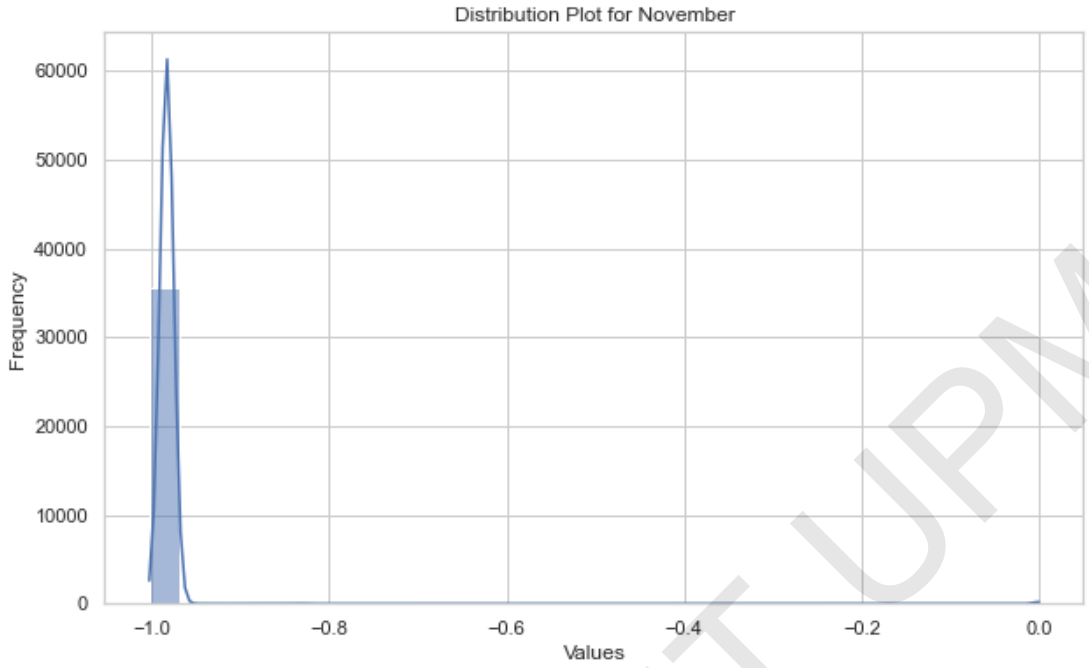


Figure 4.35: Vibration of x-axis reading on November 2023.

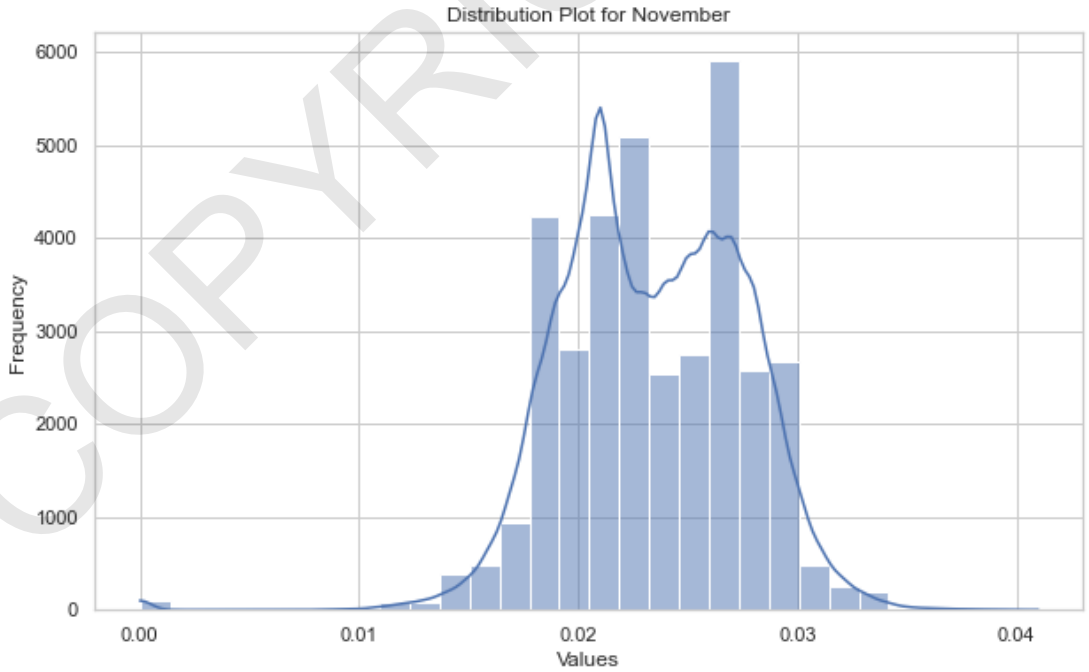


Figure 4.36: Vibration of x-axis reading on November 2023.

4.6 Limitations

The research presents several limitations that impact the accuracy and generalizability of the findings. One major challenge is the synchronization of sensor data, as different sensors operate on varying intervals, such as current and acceleration data collected every minute, radiation every three minutes, and temperature and humidity every ten minutes. This inconsistency complicates real-time analysis and may require interpolation techniques, which can introduce errors. Another limitation is that while real-time data is crucial for predictive maintenance, some sensors, like those measuring radiation, provide cumulative values rather than real-time readings, which can delay anomaly detection and affect timely decision-making.

Moreover, the application of MD for abnormal detection faces constraints when dealing with too much data, as large datasets can make it difficult to accurately define elliptical boundaries for outlier detection, which can affect the efficiency and accuracy of the analysis. The NA method used to generate synthetic abnormal data also presents a challenge, as it is specifically designed for the current sensor setup and environment. Its applicability to other sensor types or environments would require further research and adaptation, limiting its broader usability. Additionally, the lack of a comprehensive CT scan machine maintenance report complicates the validation of prediction results, as currently, only one maintenance report is available. This is insufficient for drawing conclusive insights. More maintenance reports would allow for a better evaluation of the model's predictive capabilities.

The prediction results, based on real breakdown data validation, are produced by averaging hourly breakdown values into daily predictions. While this approach highlights changes in prediction patterns, it does not pinpoint the exact time of potential breakdowns. This limitation requires technicians to regularly monitor these pattern changes, with alarms triggered by high prediction results. However, the limited amount of real abnormal data remains a significant constraint. Despite the use of synthetic data generation methods like NA, the scarcity of real-world abnormal instances could reduce the model's generalization capability and its ability to detect rare events effectively. More real abnormal data is needed to improve the robustness and accuracy of the predictive maintenance model.

4.7 Summary

The dataset used in the experiment consists of 88,272 data points, combining both real and SD, classified as normal or abnormal using MD. The SD was generated using two methods which are NA and TGAN. After preprocessing, missing values were filled using time interpolation. An ANN model was employed to predict anomalies, with TGAN outperforming NA in terms of accuracy and loss. TGAN achieved 99.53% accuracy, compared to NA's 98.64%. The final validation of the ANN model on real breakdown that happened in November 2023 shows that the model is able to predict the breakdown of the CT scan machine and this matched to the maintenance report provided by the vendor. Despite limitations, such as the scarcity of real abnormal data and data synchronization challenges, the study demonstrates the effectiveness of SD in enhancing predictive maintenance models. All algorithms are done using Python 3.7.9 version and the source code used as attached in Appendix A.

CHAPTER 5

CONCLUSION

5.1 Conclusions

In conclusion, this study addresses the challenge of improving PdM for CT scan machines using data collected from IoT sensors. The research begins by identifying three key problems which are the absence of an effective anomaly detection method for high-dimensional sensor data, the lack of real abnormal data and the difficulty in building a robust model capable of predicting both real and SD. These issues create significant obstacles in accurately detecting anomalies and failures in CT scan machines, which are crucial for timely maintenance and minimizing machine downtime.

The objectives of the study were to implement a reliable method to detect abnormality, generate SD that mimics abnormal machine conditions, and evaluate the effectiveness of different approaches in enhancing the PdM model's predictive capabilities. Specifically, the research aimed to detect abnormality in the dataset using the MD and to generate SD using two techniques, NA and TGAN. The final goal was to assess the performance of an ANN model trained on both real and synthetic data.

To achieve these objectives, two methods for generating SD were employed. The NA method involved adding noise to the real data to simulate abnormal conditions, while TGAN focused on generating synthetic abnormal data to balance the dataset. The MD method, using a critical value of 3.5485 with a 95% confidence level, successfully

separated 46,095 real data points into 1,959 abnormal and 44,136 normal data points. Both NA and TGAN methods successfully generated 44,136 synthetic abnormal data points, effectively balancing both classes in the dataset.

The results showed that while the NA method achieved a model accuracy of 98.64%, the TGAN method produced SD that closely aligned with the real data distribution, resulting in a higher model accuracy of 99.53% and superior performance metrics such as lower loss, higher precision, recall, and F1-scores. The TGAN method also produced fewer misclassifications, making it a more reliable approach for detecting anomalies in CT scan machines. In the model validation stage, three methods were compared on real data collected from September 2023 until December 2023. The maintenance work was reported in November 2023.

Further analysis revealed that the high prediction rate of machine breakdowns was due to the machine operating below the recommended temperature range. This factor is significant because prolonged operation under suboptimal temperature conditions can lead to mechanical wear, reduced lubrication efficiency, and increased component stress, ultimately accelerating mechanical degradation. The TGAN model effectively captured these patterns, demonstrating better visualization of pattern changes in predicting machine breakdowns compared to NA and the head-and-tail expansion methods.

This study demonstrates that TGAN is a more effective method for generating synthetic abnormal data and enhancing the performance of PdM models, particularly when real abnormal data is limited. The findings highlight the importance of using

advanced data generation techniques like TGAN to improve the reliability and accuracy of predictive maintenance systems in industrial applications, ensuring timely detection of potential failures and reducing downtime.

5.2 Future Work

Several areas for future work can enhance the accuracy and scope of this research. One significant direction is developing frameworks that synchronize data collection across sensors, ensuring that all data streams are captured in real-time and at consistent intervals. This would resolve the current issue of asynchrony between different sensors, such as radiation, current, and temperature, enabling more accurate analysis.

Improving real-time radiation monitoring is another key focus, moving away from cumulative readings and implementing continuous, real-time data collection techniques. This would help in more immediate abnormal detection and quicker response times. Additionally, optimizing the MD method or exploring alternative techniques that handle large datasets more effectively could enhance the accuracy of outlier detection in extensive data collections. Exploring dimensionality reduction methods or developing new computational algorithms could address this issue.

The NA method, while effective in the current setup, should be tested and adapted for other types of sensors and environments. Future research should aim to generalize the NA approach, making it applicable across various industrial settings. Another crucial step is the collection of more CT scan machine maintenance reports to further validate

the predictive maintenance results. Gathering additional maintenance data will provide better support for assessing the accuracy of predictions.

The current prediction results are presented as daily averages, which has limitations in identifying the exact time a breakdown may occur. Therefore, future work should focus on improving the indication of the precise breakdown duration to enhance maintenance scheduling and response times. Finally, the collection of more real abnormal data remains a priority. Collaborating with industry partners from different manufacturer to gather real-world abnormal data or creating simulated failure scenarios could help expand the dataset, which would improve the model's performance in detecting rare events and anomalies.

REFERENCES

Abd Rahman, N. H., Mohamad Zaki, M. H., Hasikin, K., bd Razak, N. A. A., Ibrahim, A. K., & Lai, K. W. (2023). Predicting medical device failure: a promise to reduce healthcare facilities cost through smart healthcare management. *PeerJ Computer Science*, *9*, e1279–e1279. https://doi.org/10.7717/peerj-cs.1279

Abdallah, A., Rognvaldsson, T., Fan, Y., Pashami, S., & Ohlsson, M. (2023). Discovering Premature Replacements in Predictive Maintenance Time-to-Event Data. *PHM Society Asia-Pacific Conference*, *4*(1). https://doi.org/10.36001/phmap.2023.v4i1.3609

Abdelillah, F. M., Nora, H., Samir, O., & Benslimane, S. M. (2023). Predictive Maintenance Approaches in Industry 4.0: A Systematic Literature Review. *2023 IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*. https://doi.org/10.1109/wetice57085.2023.10477802

Abdillah, B., Bustamam, A., & Sarwinda, D. (2017). Image processing based detection of lung cancer on CT scan images. *Journal of Physics: Conference Series*, *893*, 012063. https://doi.org/10.1088/1742-6596/893/1/012063

Abdulrazzq, R. A., Sajid, N. M., & Hasan, M. S. (2024). Artificial intelligence-driven predictive maintenance in IoT systems. *South Florida Journal of Development*, *5*(12), e4781. https://doi.org/10.46932/sfjdv5n12-030

Abood, A. M., Nasser, A. R., & Al-Khazraji, H. (2023). Predictive maintenance of electromechanical systems based on enhanced generative adversarial neural network with convolutional neural network. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *12*(4), 1704. https://doi.org/10.11591/ijai.v12.i4.pp1704-1712

Achouch, M., Dimitrova, M., Dhouib, R., Ibrahim, H., Adda, M., Sattarpanah Karganroudi, S., Ziane, K., & Aminzadeh, A. (2023). Predictive Maintenance and Fault Monitoring Enabled by Machine Learning: Experimental Analysis of a TA-48 Multistage Centrifugal Plant Compressor. *Applied Sciences (Switzerland)*, *13*(3). https://doi.org/10.3390/app13031790

Adachi, K. (2008). *Maintenance support information management apparatus, X-ray CT system, and maintenance service center apparatus of medical system* (Issue JP2008279281A).

Admojo, F. T., & Nurul Rismayanti. (2024). Estimating Obesity Levels Using Decision Trees and K-Fold Cross-Validation: A Study on Eating Habits and Physical Conditions. *Indonesian Journal of Data and Science*, *5*(1), 37–44. https://doi.org/10.56705/ijodas.v5i1.126

Afonso, M., & Giufrida, V. (2023). Editorial: Synthetic data for computer vision in agriculture. *Frontiers in Plant Science*, *14*. https://doi.org/10.3389/fpls.2023.1277073

Ahmed, R., Nasiri, F., & Zayed, T. (2021). A novel Neutrosophic-based machine learning approach for maintenance prioritization in healthcare facilities. *Journal of Building Engineering*, *42*, 102480. https://doi.org/10.1016/j.jobe.2021.102480

Alahmari, A. (2022). *Cold CT scanner rooms: A simple solution for the patient comfort and for hypothermia cases*. *5*, 7–9. https://doi.org/10.36811/ojrmi.2022.110043

Alves, F., Badikyan, H., Antonio Moreira, H. J., Azevedo, J., Moreira, P. M., Romero, L., & Leitao, P. (2020). Deployment of a Smart and Predictive Maintenance System in an Industrial Case Study. *2020 IEEE 29th International Symposium on Industrial Electronics (ISIE)*, 493–498. https://doi.org/10.1109/ISIE45063.2020.9152441

Amrith, V., S, D., S, S. K., Vajipayajula, S., Srinivasan, K., Thangavel, S. K., & Kumar, T. G. (2023). An early malware threat detection model using Conditional Tabular Generative Adversarial Network. *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1–8. https://doi.org/10.1109/ICCCNT56998.2023.10307903

Andrew, A. L., Chu, K.-W. E., & Lancaster, P. (1993). Derivatives of Eigenvalues and Eigenvectors of Matrix Functions. *SIAM Journal on Matrix Analysis and Applications*, *14*(4), 903–926. https://doi.org/10.1137/0614061

Arena, S., Florian, E., Sgarbossa, F., Sølvsberg, E., & Zennaro, I. (2024). A conceptual framework for machine learning algorithm selection for predictive maintenance. *Engineering Applications of Artificial Intelligence*, *133*, 108340. https://doi.org/10.1016/j.engappai.2024.108340

Asakura, T., Yashima, W., Suzuki, K., & Shimotou, M. (2020). Anomaly detection in a logistic operating system using the mahalanobis-taguchi method. *Applied Sciences (Switzerland)*, *10*(12). https://doi.org/10.3390/app10124376

Ashrafi, N., Schmitt, V., Spang, R. P., Möller, S., & Voigt-Antons, J.-N. (2023). Protect and Extend - Using GANs for Synthetic Data Generation of Time-Series Medical Records. *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*, 171–176. https://doi.org/10.1109/QoMEX58391.2023.10178496

Ashwini Tuppad, & Shantala Devi Patil. (2023). Data Pre-processing Issues in Medical Data Classification. *Journal of Advanced Zoology*, *44*(S6), 1079–1084. https://doi.org/10.17762/jaz.v44iS6.2361

Ayvaz, S., & Alpay, K. (2021). Predictive maintenance system for production lines in manufacturing: A machine learning approach using IoT data in real-time. *Expert Systems with Applications*, *173*, 114598. https://doi.org/10.1016/j.eswa.2021.114598

Azimi, S., & Pahl, C. (2024). Anomaly analytics in data-driven machine learning applications. *International Journal of Data Science and Analytics*. https://doi.org/10.1007/s41060-024-00593-y

Bampoula, X., Siaterlis, G., Nikolakis, N., & Alexopoulos, K. (2021). A Deep Learning Model for Predictive Maintenance in Cyber-Physical Production Systems Using LSTM Autoencoders. *Sensors*, *21*(3), 972. https://doi.org/10.3390/s21030972

Bani, N. A., Noordin, M. K., Hidayat, A. A., Kamil, A. S. A., Amran, M. E., Kasri, N. F., Muhtazaruddin, M. N., & Muhammad-Sukki, F. (2022). Development of Predictive Maintenance System for Haemodialysis Reverse Osmosis Water Purification System. *2022 4th International Conference on Smart Sensors and Application (ICSSA)*, 23–28. https://doi.org/10.1109/icssa54161.2022.9870965

Barrera, J. M., Reina, A. R., Maté, A., & Trujillo, J. (2022). Fault detection and diagnosis for industrial processes based on clustering and autoencoders: a case of gas turbines. *International Journal of Machine Learning and Cybernetics*, *13*, 3113–3129. https://doi.org/10.1007/s13042-022-01583-x

Benhar, H., Idri, A., & Fernández-Alemán, J. L. (2020). Data preprocessing for heart disease classification: A systematic literature review. *Computer Methods and Programs in Biomedicine*, *195*, 105635. https://doi.org/10.1016/j.cmpb.2020.105635

Bhidé, A., Datar, S., & Stebbins, K. (2020). *Computed Tomography (CT)-Beyond Traditional X-Rays*. https://www.hbs.edu/ris/Publication%20Files/20-004_49d8866f-1dac-4418-9d74-b65195a5098a.pdf

Bouhlel, N., & Dziri, A. (2019). Kullback–Leibler Divergence Between Multivariate Generalized Gaussian Distributions. *IEEE Signal Processing Letters*, *26*(7), 1021–1025. https://doi.org/10.1109/LSP.2019.2915000

Brownlee, J. (2021). *A Gentle Introduction to Threshold-Moving for Imbalanced Classification*. https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/

Bukhsh, Z. A., & Stipanovic, I. (2020). Predictive Maintenance for Infrastructure Asset Management. *IT Professional*, *22*, 40–45. https://doi.org/10.1109/mitp.2020.2975736

Burmeister, N., Frederiksen, R. D., Hog, E., & Nielsen, P. (2023). Exploration of Production Data for Predictive Maintenance of Industrial Equipment: A Case Study. *IEEE Access*, *11*, 102025–102037. https://doi.org/10.1109/ACCESS.2023.3315842

Cabana, E., Lillo, R. E., & Laniado, H. (2021). Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators. *Statistical Papers*, *62*(4), 1583–1609. https://doi.org/10.1007/s00362-019-01148-1

Campion, M. A., & Campion, E. D. (2023). Machine learning applications to personnel selection: Current illustrations, lessons learned, and future research. *Personnel Psychology*, *76*(4), 993–1009. https://doi.org/10.1111/peps.12621

Cansiz, S. (2023). *Mahalanobis Distance and Multivariate Outlier Detection in R*.
Chen, P., Liu, H., Xin, R., Carval, T., Zhao, J., Xia, Y., & Zhao, Z. (2022). Effectively Detecting Operational Anomalies In Large-Scale IoT Data Infrastructures By Using A GAN-Based Predictive Model. *Computer Journal*, *65*(11), 2909–2925. https://doi.org/10.1093/comjnl/bxac085

Chen, T., Liu, X., Xia, B., Wang, W., & Lai, Y. (2020). Unsupervised anomaly detection of industrial robots using sliding-window convolutional variational autoencoder. *IEEE Access*, *8*, 47072–47081. https://doi.org/10.1109/ACCESS.2020.2977892

Chepino, B. G., Yacoub, R. R., Aula, A., Saleh, M., & Sanjaya, B. W. (2023). EFFECT OF MINMAX NORMALIZATION ON ORB DATA FOR IMPROVED ANN ACCURACY. *Journal of Electrical Engineering, Energy, and Information Technology (J3EIT)*, *11*(2), 29. https://doi.org/10.26418/j3eit.v11i2.68689

Chicco, D., & Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, *16*(1), 4. https://doi.org/10.1186/s13040-023-00322-4

Clifton, D. A., Hugueny, S., & Tarassenko, L. (2011). Pinning the tail on the distribution: A multivariate extension to the generalised Pareto distribution. *2011 IEEE International Workshop on Machine Learning for Signal Processing*, 1–6. https://doi.org/10.1109/MLSP.2011.6064572

Dang, T.-B., Le, D.-T., Kim, M., & Choo, H. (2022). A General Model for Long-short Term Anomaly Generation in Sensory Data. *2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, 1–5. https://doi.org/10.1109/IMCOM53663.2022.9721783

Dani, S. K., Thakur, C., Nagvanshi, N., & Singh, G. (2024). Anomaly Detection using PCA in Time Series Data. *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, 1–6. https://doi.org/10.1109/IATMSI60426.2024.10502929

Delussu, R., Putzu, L., & Fumera, G. (2024). Synthetic Data for Video Surveillance Applications of Computer Vision: A Review. *International Journal of Computer Vision*, *132*, 4473–4509. https://doi.org/10.1007/s11263-024-02102-x

Dhamale, P. S., & Kashikar, A. S. (2025). Outlier detection in cylindrical data based on Mahalanobis distance. *Communications in Statistics - Simulation and Computation*, *54*(2), 331–341. https://doi.org/10.1080/03610918.2023.2252630

Draelos, R. (2019). *Measuring Performance: The Confusion Matrix*. https://glassboxmedicine.com/2019/02/17/measuring-performance-the-confusion-matrix/

Dube, A. (2024). Application of Deep Learning in Predictive Maintenance of Aircraft Engines. *Darpan International Research Analysis*, *12*(3), 83–100. https://doi.org/10.36676/dira.v12.i3.58

Einabadi, B., Mahmoodjanloo, M., Baboli, A., & Rother, E. (2023). Dynamic predictive and preventive maintenance planning with failure risk and opportunistic grouping considerations: A case study in the automotive industry. *Journal of Manufacturing Systems*, *69*, 292–310. https://doi.org/10.1016/j.jmsy.2023.06.012

Eklund, N. H. W. (2006). Using Synthetic Data to Train an Accurate Real-World Fault Detection System. *The Proceedings of the Multiconference on "Computational Engineering in Systems Applications,"* 483–488. https://doi.org/10.1109/CESA.2006.4281700

Elkateb, S., Métwalli, A., Shendy, A., & Abu-Elanien, A. E. B. (2024). Machine learning and IoT – Based predictive maintenance approach for industrial applications. *Alexandria Engineering Journal*, *88*, 298–309. https://doi.org/10.1016/j.aej.2023.12.065

Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. *Frontiers in Energy Research*, *9*, 652801. https://doi.org/10.3389/fenrg.2021.652801

Fathi, K., van de Venn, H. W., & Honegger, M. (2021). Predictive Maintenance: An Autoencoder Anomaly-Based Approach for a 3 DoF Delta Robot. *Sensors*, *21*(21), 6979. https://doi.org/10.3390/s21216979

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Filios, G., Karatzas, S., Krousarlis, M., Nikoletseas, S., Panagiotou, S. H., & Spirakis, P. (2023). Realizing Predictive Maintenance in Production Machinery Through Low-Cost IIoT Framework and Anomaly Detection: A Case Study in a Real-World Manufacturing Environment. *2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*, 338–346. https://doi.org/10.1109/DCOSS-IoT58021.2023.00062

Fordal, J. M., Schjølberg, P., Helgetun, H., Skjermo, T. Ø., Wang, Y., & Wang, C. (2023). Application of sensor data based predictive maintenance and artificial neural networks to enable Industry 4.0. *Advances in Manufacturing*, *11*(2), 248–263. https://doi.org/10.1007/s40436-022-00433-x

Fu, Q., Wang, H., Zhao, J., & Yan, X. (2019). A Maintenance-prediction Method for Aircraft Engines using Generative Adversarial Networks. *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*, 225–229. https://doi.org/10.1109/ICCC47050.2019.9064184

Gallab, M., Ahidar, I., Zrira, N., & Ngote, N. (2024). Towards a Digital Predictive Maintenance (DPM): Healthcare Case Study. *Procedia Computer Science*, *232*, 3183–3194. https://doi.org/10.1016/j.procs.2024.02.134

Giuffrè, M., & Shung, D. L. (2023). Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *Npj Digital Medicine*, *6*(1), 186. https://doi.org/10.1038/s41746-023-00927-3

Guissi, M., El Yousfi Alaoui, M. H., Belarbi, L., & Chaik, A. (2024). IoT for predictive maintenance of critical medical equipment in a hospital structure. *Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska*, *14*, 71–76. https://doi.org/10.35784/iapgos.6057

Gunderia, A., Agarwal, P., Nikam, V. B., & Bambole, A. N. (2024). AI Approaches for Predictive Maintenance in Road Bridge Infrastructure. *2024 International Conference on Smart Applications, Communications and Networking (SmartNets)*, 1–6. https://doi.org/10.1109/SmartNets61466.2024.10577641

Gupta, V., Mitra, R., Koenig, F., Kumar, M., & Tiwari, M. K. (2023). Predictive maintenance of baggage handling conveyors using IoT. *Computers and Industrial Engineering*, *177*, 109033. https://doi.org/10.1016/j.cie.2023.109033

Hadi, R. H., Hady, H. N., Hasan, A. M., Al-Jodah, A., & Humaidi, A. J. (2023). Improved Fault Classification for Predictive Maintenance in Industrial IoT Based on AutoML: A Case Study of Ball-Bearing Faults. *Processes*, *11*(5). https://doi.org/10.3390/pr11051507

Hakami, A. (2024). Strategies for overcoming data scarcity, imbalance, and feature selection challenges in machine learning models for predictive maintenance. *Scientific Reports*, *14*(1), 9645. https://doi.org/10.1038/s41598-024-59958-9

Hassan, I. U., Panduru, K., & Walsh, J. (2024). Review of Data Processing Methods Used in Predictive Maintenance for Next Generation Heavy Machinery. *Data*, *9*(5), 69. https://doi.org/10.3390/data9050069

Hazra, A. (2017). Using the confidence interval confidently. *Journal of Thoracic Disease*, *9*(10), 4124–4129. https://doi.org/10.21037/jtd.2017.09.14

He, X. (2024). Technical Perspective: Synthetic Data Needs a Reproducibility Benchmark. *ACM SIGMOD Record*, *53*(1), 64–64. https://doi.org/10.1145/3665252.3665266

Hector, I., & Panjanathan, R. (2024). Predictive maintenance in Industry 4.0: a survey of planning models and machine learning techniques. *PeerJ Computer Science*, *10*, e2016. https://doi.org/10.7717/peerj-cs.2016

Henderi, H. (2021). Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer. *IJIIS: International Journal of Informatics and Information Systems*, *4*(1), 13–20. https://doi.org/10.47738/ijiis.v4i1.73

Herrera-Chavez, A. I., Rodríguez-Martínez, E. A., Flores-Fuentes, W., Rodgíruez-Quiñonez, J. C., García-Gallegos, J. C., Montiel-Ross, O. H., Gonzàalez-Navarro, F. F., & Sergiyenko, O. (2024). Multi-label Image Classification for Ocular Disease Diagnosis Using K-fold Cross-Validation on the ODIR-5K Dataset. *2024 IEEE 33rd International Symposium on Industrial Electronics (ISIE)*, 1–6. https://doi.org/10.1109/ISIE54533.2024.10595740

Hsieh, J., & Flohr, T. (2021). Computed tomography recent history and future perspectives. *Journal of Medical Imaging*, *8*(05), 052109. https://doi.org/10.1117/1.JMI.8.5.052109

Iantovics, L. B., & Enăchescu, C. (2022). Method for Data Quality Assessment of Synthetic Industrial Data. *Sensors*, *22*(4), 1608. https://doi.org/10.3390/s22041608

Jeba Emilyn., J., Vinod Kumar., D., David Samuel Azariya., S., Prakash., M., & Sam Thamburaj., A. (2024). Deep Learning-based Predictive Maintenance for Industrial IoT Applications. *2024 International Conference on Inventive Computation Technologies (ICICT)*, 1197–1202. https://doi.org/10.1109/ICICT60155.2024.10544377

Jiang, Y., Dai, P., Fang, P., Zhong, R. Y., Zhao, X., & Cao, X. (2022). A2 -LSTM for predictive maintenance of industrial equipment based on machine learning. *Computers & Industrial Engineering*, *172*, 108560. https://doi.org/10.1016/j.cie.2022.108560

Kaynak, G., & Ervural, B. (2024). Predictive Maintenance Planning Using a Hybrid ARIMA-ANN Model. *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, *13*(3), 618–632. https://doi.org/10.17798/bitlisfen.1466339

Kazmi, M., Shoaib, M., Aziz, A., Khan, H., & Qazi, S. (2023). An Efficient IIoT-Based Smart Sensor Node for Predictive Maintenance of Induction Motors. *Computer Systems Science and Engineering*, *47*, 255–272. https://doi.org/10.32604/csse.2023.038464

Kim, K. H., Sohn, M.-J., Lee, S., Koo, H.-W., Yoon, S.-W., & Madadi, A. K. (2022). Descriptive Time Series Analysis for Downtime Prediction Using the Maintenance Data of a Medical Linear Accelerator. *Applied Sciences*, *12*, 5431. https://doi.org/10.3390/app12115431

Klein, P., & Bergmann, R. (2019). Generation of Complex Data for AI-based Predictive Maintenance Research with a Physical Factory Model. *Proceedings of the 16th International Conference on Informatics in Control, Automation and Robotics*, 40–50. https://doi.org/10.5220/0007830700400050

Kumar Srivastav, A., & Vaidya, D. (2020). *Interpolation*. https://www.wallstreetmojo.com/interpolation/

Kwan, I. L. W., Soysa, W. H., Cheung, D. C. L., & Liang, W. R. (2020). *Maintenance systems and methods for medical devices* (Issue SG10202002519UA).

Laher, A. E., Watermeyer, M. J., Buchanan, S. K., Dippenaar, N., Simo, N. C. T., Motara, F., & Moolla, M. (2017). A review of hemodynamic monitoring techniques, methods and devices for the emergency physician. *The American Journal of Emergency Medicine*, *35*(9), 1335–1347. https://doi.org/10.1016/j.ajem.2017.03.036

Laska, R. R., & Yolanda, A. M. (2024). A Comparative Study of Z-Score and Min-Max Normalization for Rainfall Classification in Pekanbaru. *Journal of Data Science*, *2024*(1). https://doi.org/10.61453/jods.v2024no04

Latyshev, E. (2018). Sensor Data Preprocessing, Feature Engineering and Equipment Remaining Lifetime Forecasting for Predictive Maintenance. *International Conference on Data Analytics and Management in Data Intensive Domains*, 226–231. https://api.semanticscholar.org/CorpusID:58005342

Li, J., Mao, Y., & Zhang, J. (2022). Maintenance and Quality Control of Medical Equipment Based on Information Fusion Technology. *Computational Intelligence and Neuroscience*, *2022*, 1–11. https://doi.org/10.1155/2022/9333328

Lin, S.-L. (2021). Application of Machine Learning to a Medium Gaussian Support Vector Machine in the Diagnosis of Motor Bearing Faults. *Electronics*, *10*(18), 2266. https://doi.org/10.3390/electronics10182266

Liu, X., Pan, Y., Ding, Y., & Pan, Z. (2023). Distributed One-Pass Online AUC Maximization. *Discrete Applied Mathematics*, *341*, 322–336. https://doi.org/10.1016/j.dam.2023.08.026

Liu, Y., Li, Z., Zhou, C., Jiang, Y., Sun, J., Wang, M., & He, X. (2020). Generative Adversarial Active Learning for Unsupervised Outlier Detection. *IEEE Transactions on Knowledge and Data Engineering*, *32*(8), 1517–1528. https://doi.org/10.1109/TKDE.2019.2905606

Ma, H., Geng, M., Wang, F., Zheng, W., Ai, Y., & Zhang, W. (2024). Data Augmentation of a Corrosion Dataset for Defect Growth Prediction of Pipelines Using Conditional Tabular Generative Adversarial Networks. *Materials*, *17*(5), 1142. https://doi.org/10.3390/ma17051142

Mallioris, P., Aivazidou, E., & Bechtsis, D. (2024). Predictive maintenance in Industry 4.0: A systematic multi-sector mapping. *CIRP Journal of Manufacturing Science and Technology*, *50*, 80–103. https://doi.org/10.1016/j.cirpj.2024.02.003

Manchadi, O., BEN-BOUAZZA, F.-E., Dehbi, Z. E. O., Edder, A., Tafala, I., Et-Taoussi, M., & Jioudi, B. (2024). An Internet of Things-based Predictive Maintenance Architecture for Intensive Care Unit Ventilators. *International Journal of Advanced Computer Science and Applications*, *15*(2), 932–943. https://doi.org/10.14569/ijacsa.2024.0150294

Marani, A., Jamali, A., & Nehdi, M. L. (2020). Predicting Ultra-High-Performance Concrete Compressive Strength Using Tabular Generative Adversarial Networks. *Materials*, *13*(21), 4757. https://doi.org/10.3390/ma13214757

Marcot, B. G., & Hanea, A. M. (2021). What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? *Computational Statistics*, *36*(3), 2009–2031. https://doi.org/10.1007/s00180-020-00999-9

Meiser, M., Duppe, B., & Zinnikus, I. (2024). Generation of meaningful synthetic sensor data — Evaluated with a reliable transferability methodology. *Energy and AI*, *15*, 100308. https://doi.org/10.1016/j.egyai.2023.100308

Miletic, M., & Sariyar, M. (2024). Challenges of Using Synthetic Data Generation Methods for Tabular Microdata. *Applied Sciences*, *14*(14), 5975. https://doi.org/10.3390/app14145975

Mishra, Dr. N., Haval, Dr. A. M., Mishra, A., & Dash, S. S. (2024). Automobile Maintenance Prediction Using Integrated Deep Learning and Geographical Information System. *Indian Journal of Information Sources and Services*, *14*(2), 109–114. https://doi.org/10.51983/ijiss-2024.14.2.16

Mohd Azrul Shazril, M. H. S. E., Mashohor, S., Amran, M. E., Fatinah Hafiz, N., Rahman, A. A., Ali, A., Rasid, M. F. A., Safwan A. Kamil, A., & Azilah, N. F. (2023). Predictive Maintenance Method using Machine Learning for IoT Connected Computed Tomography Scan Machine. *2023 IEEE 2nd National Biomedical Engineering Conference (NBEC)*, 42–47. https://doi.org/10.1109/NBEC58134.2023.10352591

Mourabit, Y. El, El, Y., Zougagh, H., & Wadiai, Y. (2020). Predictive System of Semiconductor Failures based on Machine Learning Approach. *International Journal of Advanced Computer Science and Applications*, *11*(12), 199–203. https://doi.org/10.14569/ijacsa.2020.0111225

Nassif, A. B., Talib, M. A., Nasir, Q., & Dakalbab, F. M. (2021). Machine Learning for Anomaly Detection: A Systematic Review. In *IEEE Access* (Vol. 9, pp. 78658–78700). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2021.3083060

Nigam, A., & Srivastava, S. (2023). Generating Realistic Synthetic Traffic Data using Conditional Tabular Generative Adversarial Networks for Intelligent Transportation Systems. *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, 2881–2886. https://doi.org/10.1109/ITSC57777.2023.10422234

Niyonambaza, I., Zennaro, M., & Uwitonze, A. (2020). Predictive Maintenance (PdM) Structure Using Internet of Things (IoT) for Mechanical Equipment Used into Hospitals in Rwanda. *Future Internet*, *12*, 224. https://doi.org/10.3390/fi12120224

Offenhuber, D. (2024). Shapes and frictions of synthetic data. *Big Data & Society*, *11*(2). https://doi.org/10.1177/20539517241249390

Omol, E., Mburu, L., & Onyango, D. (2024). Anomaly Detection In IoT Sensor Data Using Machine Learning Techniques For Predictive Maintenance In Smart Grids. *International Journal Of Science, Technology & Management*, *5*, 201–210. https://doi.org/10.46729/ijstm.v5i1.1028

Ong, K. S. H., Wang, W., Niyato, D., & Friedrichs, T. (2022). Deep-Reinforcement-Learning-Based Predictive Maintenance Model for Effective Resource Management in Industrial IoT. *IEEE Internet of Things Journal*, *9*(7), 5173–5188. https://doi.org/10.1109/JIOT.2021.3109955

Oprea, S. V., Bara, A., Preotescu, D., & Elefterescu, L. (2019). Photovoltaic Power Plants (PV-PP) Reliability Indicators for Improving Operation and Maintenance Activities. A Case Study of PV-PP Agigea Located in Romania. *IEEE Access*, *7*, 39142–39157. https://doi.org/10.1109/ACCESS.2019.2907098

Pagano, D. (2023). A predictive maintenance model using Long Short-Term Memory Neural Networks and Bayesian inference. *Decision Analytics Journal*, *6*, 100174. https://doi.org/10.1016/j.dajour.2023.100174

Panalaran, S., Suntoyo, & Sulisetyono, A. (2024). Pre-processing data and window function testing on wave spectrum analysis. *IOP Conference Series: Earth and Environmental Science*, *1298*(1), 012037. https://doi.org/10.1088/1755-1315/1298/1/012037

Patel, C., Pandey, A., Wadhvani, R., & Patil, D. (2022). Forecasting Nonstationary Wind Data Using Adaptive Min-Max Normalization. *2022 1st International Conference on Sustainable Technology for Power and Energy Systems (STPES)*, 1–6. https://doi.org/10.1109/STPES54845.2022.10006473

Patel, P. R., & De Jesus, O. (2023). *CT Scan* (Updated 2023 Jan 2). StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK567796/

Patil, H., Swati, Sudheendramouli, H. C., Pratyush, K., Parikh, S., & Vishwakarma, R. K. (2024). Algorithms for Detection of Anomalies in Data From Clinical Studies and Medical Registries. *2024 International Conference on Intelligent Systems for Cybersecurity (ISCS)*, 1–7. https://doi.org/10.1109/ISCS61804.2024.10581281

Patwardhan, A., Verma, A. K., & Kumar, U. (2016). *A Survey on Predictive Maintenance Through Big Data* (pp. 437–445). https://doi.org/10.1007/978-3-319-23597-4_31

Pejić Bach, M., Topalović, A., Krstić, Ž., & Ivec, A. (2023). Predictive Maintenance in Industry 4.0 for the SMEs: A Decision Support System Case Study Using Open-Source Software. *Designs*, *7*(4), 98. https://doi.org/10.3390/designs7040098

Pham, N.-H., Vo, K.-L., Vu, M. A., Nguyen, T., Riegler, M. A., Halvorsen, P., & Nguyen, B. T. (2024). *Correlation Visualization Under Missing Values: A Comparison Between Imputation and Direct Parameter Estimation Methods* (pp. 103–116). https://doi.org/10.1007/978-3-031-53302-0_8

Pisa, I., Santin, I., Vicario, J. L., Morell, A., & Vilanova, R. (2019). Data Preprocessing for ANN-based Industrial Time-Series Forecasting with Imbalanced Data. *2019 27th European Signal Processing Conference (EUSIPCO)*, 1–5. https://doi.org/10.23919/EUSIPCO.2019.8902682

Plesovskaya, E., & Ivanov, S. (2021). An Empirical Analysis of KDE-based Generative Models on Small Datasets. *Procedia Computer Science*, *193*, 442–452. https://doi.org/10.1016/j.procs.2021.10.046

Postiglione, A., & Monteleone, M. (2024). Predictive Maintenance with Linguistic Text Mining. *Mathematics*, *12*, 1089. https://doi.org/10.3390/math12071089

Prasetya, M. A., Adhim, F. I., & Al Kindhi, B. (2023). Implementation of Weibull Analysis Method in Designing Predictive Maintenance for Medical Mask Machine. *2023 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA)*, 519–527. https://doi.org/10.1109/ICAMIMIA60881.2023.10427705

Priya Talawar, M., Pant, H., Sayyed, M., Summaiya Tamboli, M., Shaik, M., & Ganesan, V. (2023). Predictive Maintenance in Healthcare IoT: A Machine Learning-based Approach Predictive Maintenance in Healthcare IoT: A Machine Learning-based Approach. *Eur. Chem. Bull*, *2023*, 15909–15922. https://doi.org/10.48047/ecb/2023.12.si4.1420

R. Patil, C., K. Jadhav, S., L. Bardiya, A., P. Davande, A., & P. Raverkar, M. (2023). Machine Learning-Based Predictive Maintenance of Industrial Machines. *International Journal of Computer Trends and Technology*, *71*(3), 50–56. https://doi.org/10.14445/22312803/ijctt-v71i3p108

Rahman, N. H. A., Hasikin, K., Razak, N. A. A., Al-Ani, A. K., Anni, D. J. S., & Mohandas, P. (2023). Medical Device Failure Predictions Through AI-Driven Analysis of Multimodal Maintenance Records. *IEEE Access*, *11*, 93160–93179. https://doi.org/10.1109/access.2023.3309671

Raparthi, M., Dodda, S. B., & Maruthi, S. (2023). Predictive Maintenance in IoT Devices using Time Series Analysis and Deep Learning. *Dandao Xuebao/Journal of Ballistics*, *35*(3), 1–10. https://doi.org/https://doi.org/10.52783/dxjb.v35.113

Ridwan, M., Wahyono, T., Iriani, A., Darmastuti, R., Sukisno, & Nugroho, A. H. (2023). K-Fold Cross-Validation Analysis for Student Status Prediction Using Classification Models. *2023 IEEE 7th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 247–252. https://doi.org/10.1109/ICITISEE58992.2023.10405343

Riehl, K., Neunteufel, M., & Hemberg, M. (2023). Hierarchical confusion matrix for classification performance evaluation. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *72*(5), 1394–1412. https://doi.org/10.1093/jrsssc/qlad057

Romanelli, F., & Martinelli, F. (2023). Synthetic Sensor Measurement Generation With Noise Learning and Multi-Modal Information. *IEEE Access*, *11*, 111765–111788. https://doi.org/10.1109/ACCESS.2023.3323038

Saad, M. S., & Moore, S. (2022). Computed Tomography Scan. In *A Medication Guide to Internal Medicine Tests and Procedures* (pp. 78–84). Elsevier. https://doi.org/10.1016/B978-0-323-79007-9.00017-9

Samatas, G. G., Moumgiakmas, S. S., & Papakostas, G. A. (2021). Predictive Maintenance - Bridging Artificial Intelligence and IoT. *2021 IEEE World AI IoT Congress (AIIoT)*, 0413–0419. https://doi.org/10.1109/AIIoT52608.2021.9454173

Sauber-Cole, R., & Khoshgoftaar, T. M. (2022). The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey. *Journal of Big Data*, *9*(1), 98. https://doi.org/10.1186/s40537-022-00648-6

Seol, S., Yoon, J., Lee, J., & Kim, B. (2024). Metrics for Evaluating Synthetic Time-Series Data of Battery. *Applied Sciences*, *14*(14), 6088. https://doi.org/10.3390/app14146088

Shah, V., Keniya, R., Shridharani, A., Punjabi, M., Shah, J., & Mehendale, N. (2021). Diagnosis of COVID-19 using CT scan images and deep learning techniques. *Emergency Radiology*, *28*(3), 497–505. https://doi.org/10.1007/s10140-020-01886-y

Shamayleh, A., Awad, M., & Farhat, J. (2020). IoT Based Predictive Maintenance Management of Medical Equipment. *Journal of Medical Systems*, *44*(4), 72. https://doi.org/10.1007/s10916-020-1534-8

Sharma, A., & Khanna, S. (2024). Anomaly detection using modified linearity based grey swarm optimization algorithm and efficient hyper parameter optimization techniques. *Social Informatics Journal*, *3*(1), 65–84. https://doi.org/10.58898/sij.v3i1.65-84

Sharma, D. B., Sripradha, Nikita, Kodipalli, A., Rao, T., & B R, R. (2023). Machine Predictive Maintenance Classification Using Machine Learning. *2023 International Conference on Computational Intelligence for Information, Security and Communication Applications (CIISCA)*, 308–313. https://doi.org/10.1109/CIISCA59740.2023.00066

Shrivastava, M., Singhal, P., & Bhuvana, J. (2023). INTEGRATING SENSOR DATA AND MACHINE LEARNING FOR PREDICTIVE MAINTENANCE IN INDUSTRY 4.0. *Proceedings on Engineering Sciences*, *5*(S1), 55–62. https://doi.org/10.24874/PES.SI.01.007

Shukla, R. M., & Sengupta, S. (2020). Scalable and Robust Outlier Detector using Hierarchical Clustering and Long Short-Term Memory (LSTM) Neural Network for the Internet of Things. *Internet of Things (Netherlands)*, *9*, 100167. https://doi.org/10.1016/j.iot.2020.100167

Shunji, M., & Hisae, S. (2012). *Anomaly detection method and anomaly detection system.* (Patent US20120041575A1). U.S. Patent and Trademark Office.

Simundic, A.-M. (2008). Confidence interval. *Biochemia Medica*, 154–161. https://doi.org/10.11613/BM.2008.015

Szabo, I., Tulbure, A. A., Fleseriu, D., & Tulbure, A. A. (2022). Predictive Maintenance Efficiency Study Using Wireless Sensor Clusters (WSCs). *2022 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR)*, 1–6. https://doi.org/10.1109/AQTR55203.2022.9802022

Taguchi, K., Ballabriga, R., Campbell, M., & Darambara, D. G. (2022). Photon Counting Detector Computed Tomography. *IEEE Transactions on Radiation and Plasma Medical Sciences*, *6*(1), 1–4. https://doi.org/10.1109/TRPMS.2021.3133808

Tanyıldız, H., Batur Şahin, C., & Batur Dinler, Ö. (2024). Disrupting Downtime: Different Deep Learning Journeys into Predictive Maintenance Anomaly Detection. *NATURENGS MTU Journal of Engineering and Natural Sciences Malatya Turgut Ozal University*, *5*(1), 47–53. https://doi.org/10.46572/naturengs.1490748

Taşcı, B., Omar, A., & Ayvaz, S. (2023). Remaining useful lifetime prediction for predictive maintenance in manufacturing. *Computers and Industrial Engineering*, *184*, 109566. https://doi.org/10.1016/j.cie.2023.109566

Tun, T. P., & Pisica, I. (2023). Synthetic Electricity Consumption Data Generation Using Tabular Generative Adversarial Networks. *2023 58th International Universities Power Engineering Conference (UPEC)*, 1–6. https://doi.org/10.1109/UPEC57427.2023.10294666

Ustyannie, W., Setyaningsih, E., & Iswahyudi, C. (2021). Optimization of software defects prediction in imbalanced class using a combination of resampling methods with support vector machine and logistic regression. *JURNAL INFOTEL*, *13*(4), 176–184. https://doi.org/10.20895/infotel.v13i4.726

Varma, D., Nehansh, A., & Swathy, P. (2023). Data Preprocessing Toolkit : An Approach to Automate Data Preprocessing. *International Journal of Scientific Research in Engineering and Management*, *07*(03). https://doi.org/10.55041/ijsrem18270

Vaz, B., & Figueira, Á. (2024). GANs in the Panorama of Synthetic Data Generation Methods. *ACM Transactions on Multimedia Computing, Communications, and Applications*. https://doi.org/10.1145/3657294

Voronov, S., Krysander, M., & Frisk, E. (2023). Predictive Maintenance of Lead-Acid Batteries with Sparse Vehicle Operational Data. *International Journal of Prognostics and Health Management*, *11*(1). https://doi.org/10.36001/ijphm.2020.v11i1.2608

Wahid, A., Breslin, J. G., & Intizar, M. A. (2022). Prediction of Machine Failure in Industry 4.0: A Hybrid CNN-LSTM Framework. *Applied Sciences*, *12*, 4221. https://doi.org/10.3390/app12094221

Wang, J., Liang, Y., Zheng, Y., Gao, R. X., & Zhang, F. (2020). An integrated fault diagnosis and prognosis approach for predictive maintenance of wind turbine bearing with limited samples. *Renewable Energy*, *145*, 642–650. https://doi.org/10.1016/j.renene.2019.06.103

Wang, Q., Bu, S., & He, Z. (2020). Achieving Predictive and Proactive Maintenance for High-Speed Railway Power Equipment With LSTM-RNN. *IEEE Transactions on Industrial Informatics*, *16*(10), 6509–6517. https://doi.org/10.1109/TII.2020.2966033

Wang, X., Liu, M., Liu, C., Ling, L., & Zhang, X. (2023). Data-driven and Knowledge-based predictive maintenance method for industrial robots for the production stability of intelligent manufacturing. *Expert Systems with Applications*, *234*, 121136. https://doi.org/10.1016/j.eswa.2023.121136

WANG, Y., GOGU, C., BINAUD, N., BES, C., & FU, J. (2019). A model-based prognostics method for fatigue crack growth in fuselage panels. *Chinese Journal of Aeronautics*, *32*(2), 396–408. https://doi.org/10.1016/j.cja.2018.11.010

Wang, Z., Draghi, B., Rotalinti, Y., Lunn, D., & Myles, P. (2024). *High-Fidelity Synthetic Data Applications for Data Augmentation*. https://doi.org/10.5772/intechopen.113884

Whitney, C. D., & Norman, J. (2024). Real Risks of Fake Data: Synthetic Data, Diversity-Washing and Consent Circumvention. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1733–1744. https://doi.org/10.1145/3630106.3659002

Winder, M., Owczarek, A. J., Chudek, J., Pilch-Kowalczyk, J., & Baron, J. (2021). Are We Overdoing It? Changes in Diagnostic Imaging Workload during the Years 2010–2020 including the Impact of the SARS-CoV-2 Pandemic. *Healthcare*, *9*(11), 1557. https://doi.org/10.3390/healthcare9111557

Wu, N., & Zhang, J. (2006). Factor-analysis based anomaly detection and clustering. *Decision Support Systems*, *42*(1), 375–389. https://doi.org/10.1016/j.dss.2005.01.005

Xu, L., & Veeramachaneni, K. (2018). *Synthesizing Tabular Data using Generative Adversarial Networks*. https://doi.org/10.48550/arXiv.1811.11264

Yang, Z., Xu, Q., Bao, S., Cao, X., & Huang, Q. (2022). Learning With Multiclass AUC: Theory and Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(11), 7747–7763. https://doi.org/10.1109/TPAMI.2021.3101125

Yang, Z., Xu, Q., Bao, S., He, Y., Cao, X., & Huang, Q. (2023). Optimizing Two-Way Partial AUC With an End-to-End Framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(8), 10228–10246. https://doi.org/10.1109/TPAMI.2022.3185311

Yuichi, K. (2007). *X-RAY CT APPARATUS AND ITS CONTROL METHOD*. http://v3.espacenet.com/textdoc?DB=EPODOC&IDX=JP2009056066

Zamzam, A. H., Al-Ani, A. K. I., Wahab, A. K. A., Lai, K. W., Satapathy, S. C., Khalil, A., Azizan, M. M., & Hasikin, K. (2021). Prioritisation Assessment and Robust Predictive System for Medical Equipment: A Comprehensive Strategic Maintenance Management. *Frontiers in Public Health*, *9*, 782203. https://doi.org/10.3389/fpubh.2021.782203

Zamzam, A. H., Hasikin, K., & Wahab, A. K. A. (2023). Integrated failure analysis using machine learning predictive system for smart management of medical equipment maintenance. *Engineering Applications of Artificial Intelligence*, *125*, 106715. https://doi.org/10.1016/j.engappai.2023.106715

Zanke, T., Suryawanshi, R., Wath, S., Mulgir, S., & Jagtap, S. (2024). *Predictive Maintenance Model for Industrial Equipment* (pp. 227–236). https://doi.org/10.1007/978-981-97-0975-5_20

Zermane, H., & Drardja, A. (2022). Development of an efficient cement production monitoring system based on the improved random forest algorithm. *The International Journal of Advanced Manufacturing Technology*, *120*(3–4), 1853–1866. https://doi.org/10.1007/s00170-022-08884-z

Zhou, H., Liu, Q., Liu, H., Chen, Z., Li, Z., Zhuo, Y., Li, K., Wang, C., & Huang, J. (2024). Healthcare facilities management: A novel data-driven model for predictive maintenance of computed tomography equipment. *Artificial Intelligence in Medicine*, *149*, 102807. https://doi.org/10.1016/j.artmed.2024.102807

Zhou, X., & Liu, Z. (2024). *Computerized Tomography* (pp. 101–134). https://doi.org/10.1007/978-981-97-1455-1_4

Zhu, M., Ge, G., Yang, Y., Du, Z., & Yang, J. (2020). Uncertainty evaluation of straightness in coordinate measuring machines based on error ellipse theory integrated with Monte Carlo method. *Measurement Science and Technology*, *31*(3), 035008. https://doi.org/10.1088/1361-6501/ab5334

APPENDICES

Appendix A

Noise Addition (NA)

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt



# Load the clean data

data = pd.read_csv("latest_data_clean_nomean.csv")



# Function to normalize a column

def normalize(column):

    min_val = column.min()

    max_val = column.max()

    return (column - min_val) / (max_val - min_val), min_val, max_val



# Function to add noise to a column following its distribution

def add_noise(column, noise_factor=0):

    noise = np.random.normal(0, column.std() * noise_factor, len(column))

    return column + noise
```

```
# Function to revert normalization
def revert_normalization(normalized_column, min_val, max_val):
    return normalized_column * (max_val - min_val) + min_val

# Normalize the data
normalized_data = data.copy()
min_max_values = {}
for feature in data.columns:
    if feature != 'target':  # Exclude target column if exists
        normalized_data[feature], min_val, max_val = normalize(data[feature])
        min_max_values[feature] = (min_val, max_val)

# Generate noisy data
noisy_normalized_data = normalized_data.copy()
for feature in normalized_data.columns:
    if feature != 'target':  # Exclude target column if exists
        noisy_normalized_data[feature] = add_noise(normalized_data[feature])

# Revert normalization
noisy_data = noisy_normalized_data.copy()
for feature in noisy_normalized_data.columns:
```

```
    if feature != 'target':  # Exclude target column if exists
        min_val, max_val = min_max_values[feature]
        noisy_data[feature] = revert_normalization(noisy_normalized_data[feature],
min_val, max_val)

# Visualize the distribution before and after adding noise
fig, axes = plt.subplots(nrows=2, ncols=3, figsize=(15, 10))

for i, ax in enumerate(axes.flat):
    if i < len(data.columns):
        ax.hist(data[data.columns[i]], bins=30, color='blue', alpha=0.5, label='Original')
        ax.hist(noisy_data[noisy_data.columns[i]], bins=30, color='red', alpha=0.5,
label='Noisy')
        ax.set_title(data.columns[i])
        ax.legend()

plt.tight_layout()
plt.show()

# Save the noisy data to a new CSV file
noisy_data.to_csv("NF0.csv", index=False)
```

```
# Print the minimum and maximum values from noisy data

min_max_noisy_data = noisy_data.agg(['min', 'max'])


# Display the results

print(min_max_noisy_data)
```

Tabular Generative Adversarial Network (TGAN)

```
# Import necessary libraries

import pandas as pd

import numpy as np

from tabgan.sampler import GANGenerator

from sklearn.model_selection import train_test_split

from sklearn.metrics import mean_squared_error

from sklearn.preprocessing import MinMaxScaler

from scipy.stats import ks_2samp, entropy, wasserstein_distance

import matplotlib.pyplot as plt

import seaborn as sns


# Load your dataset from 'datatoday.csv'

data = pd.read_csv("datatoday.csv")
```

```
# Separate input features (Current, Radiation, Temperature, Humidity, X, Y) and the
target variable (Mahalanobis Distance)
train_data = data[['Current', 'Radiation', 'Temperature', 'Humidity', 'X', 'Y']]
target_data = pd.DataFrame(data['Mahalanobis Distance'])

# Normalize the input features using Min-Max scaling
scaler = MinMaxScaler()
train_data_normalized = pd.DataFrame(scaler.fit_transform(train_data),
columns=train_data.columns)

# Split the data into train and test sets (e.g., 90% train, 10% test)
train, test, target_train, target_test = train_test_split(train_data_normalized,
target_data, test_size=0.1, random_state=42)

# Generate synthetic data using GANGenerator
new_train, new_target = GANGenerator(
    gen_params={"batch_size": 500, "epochs": 10, "patience": 5}
(.generate_data_pipe(train, target_train, test)

# Save the generated synthetic data
new_train.to_csv('synthetic_train.csv', index=False)
```

```
new_target.to_csv('synthetic_target.csv', index=False)


# Resize synthetic data to match real data size
new_train_resized = new_train.sample(n=len(train),
random_state=42).reset_index(drop=True)


# Inverse transform the normalized data back to original scale
new_train_resized_original =
pd.DataFrame(scaler.inverse_transform(new_train_resized),
columns=new_train_resized.columns)
train_original = pd.DataFrame(scaler.inverse_transform(train),
columns=train.columns)
```

Artificial Neural Network (ANN)

```
# Importing necessary libraries
import numpy as np
import pandas as pd
import tensorflow as tf
from sklearn.model_selection import train_test_split, KFold
from sklearn.preprocessing import MinMaxScaler
```

```
from sklearn.metrics import confusion_matrix, classification_report,
precision_score, recall_score, f1_score, roc_curve, auc
import matplotlib.pyplot as plt
import seaborn as sns


# Loading dataset
data = pd.read_csv("Label.csv")


# Splitting features (X) and labels (Y)
X = data.iloc[:,:-1].values
Y = data.iloc[:,-1].values


# Splitting dataset into 70% training, 20% validation, 10% testing
X_train, X_temp, Y_train, Y_temp = train_test_split(X, Y, test_size=0.3,
random_state=0)
X_val, X_test, Y_val, Y_test = train_test_split(X_temp, Y_temp, test_size=0.33,
random_state=0)  # 33% of 30% = 10%


# Feature scaling
sc = MinMaxScaler()
X_train = sc.fit_transform(X_train)
```

```
X_val = sc.transform(X_val)

X_test = sc.transform(X_test)


# Implementing K-fold cross-validation with K=5

kfold = KFold(n_splits=5, shuffle=True, random_state=0)


fold_no = 1


# Loop over each fold

for train_index, val_index in kfold.split(X_train):

    X_train_fold, X_val_fold = X_train[train_index], X_train[val_index]

    Y_train_fold, Y_val_fold = Y_train[train_index], Y_train[val_index]


    # Initializing the ANN

    ann = tf.keras.models.Sequential()


    # Adding layers

    ann.add(tf.keras.layers.Dense(units=6, activation="relu"))

    ann.add(tf.keras.layers.Dense(units=6, activation="relu"))

    ann.add(tf.keras.layers.Dense(units=1, activation="sigmoid"))
```

```
# Compiling the ANN
ann.compile(optimizer="adam", loss="binary_crossentropy", metrics=['accuracy'])


# Fitting the ANN
print(f'Training fold {fold_no}...')
history = ann.fit(X_train_fold, Y_train_fold, batch_size=16, epochs=100,
                  validation_data=(X_val_fold, Y_val_fold))


# Saving the model for the current fold
ann.save(f"test2024_fold_{fold_no}.h5")


# Making predictions on the validation fold
Y_pred_fold_prob = ann.predict(X_val_fold)
Y_pred_fold = (Y_pred_fold_prob > 0.5).astype("int32")


# Confusion matrix
cm = confusion_matrix(Y_val_fold, Y_pred_fold)


# Calculating metrics
accuracy = np.sum(Y_pred_fold == Y_val_fold) / len(Y_val_fold)
precision = precision_score(Y_val_fold, Y_pred_fold)
```

```
recall = recall_score(Y_val_fold, Y_pred_fold)  # Sensitivity

f1 = f1_score(Y_val_fold, Y_pred_fold)


tn, fp, fn, tp = cm.ravel()

specificity = tn / (tn + fp)  # True Negative Rate


# Print out metrics for each fold

print(f"Fold {fold_no} Results:")

print(f"Accuracy: {accuracy:.4f}")

print(f"Precision: {precision:.4f}")

print(f"Sensitivity (Recall): {recall:.4f}")

print(f"Specificity: {specificity:.4f}")

print(f"F1 Score: {f1:.4f}")


# ROC Curve and AUC

fpr, tpr, _ = roc_curve(Y_val_fold, Y_pred_fold_prob)

roc_auc = auc(fpr, tpr)


# Plot ROC curve for each fold

plt.figure()
```

```
    plt.plot(fpr, tpr, color='darkorange', lw=2, label=f'ROC curve (area =
{roc_auc:.2f})')
    plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
    plt.xlim([0.0, 1.0])
    plt.ylim([0.0, 1.05])
    plt.xlabel('False Positive Rate')
    plt.ylabel('True Positive Rate')
    plt.title(f'Receiver Operating Characteristic (Fold {fold_no})')
    plt.legend(loc="lower right")
    plt.show()

    # Plotting training and validation accuracy/loss for each fold
    plt.figure(figsize=(12, 5))

    # Accuracy
    plt.subplot(1, 2, 1)
    plt.plot(history.history['accuracy'], label='Train Accuracy')
    plt.plot(history.history['val_accuracy'], label='Validation Accuracy')
    plt.title(f'Training and Validation Accuracy (Fold {fold_no})')
    plt.xlabel('Epochs')
    plt.ylabel('Accuracy')
```

```
plt.legend()


# Loss
plt.subplot(1, 2, 2)
plt.plot(history.history['loss'], label='Train Loss')
plt.plot(history.history['val_loss'], label='Validation Loss')
plt.title(f'Training and Validation Loss (Fold {fold_no})')
plt.xlabel('Epochs')
plt.ylabel('Loss')
plt.legend()


plt.tight_layout()
plt.show()


# Confusion matrix visualization for each fold
plt.figure(figsize=(10, 7))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title(f'Confusion Matrix (Fold {fold_no})')
plt.show()
```

```
    # Increment fold number
    fold_no += 1
# Making predictions on the test set
Y_pred_test_prob = ann.predict(X_test)  # Get probabilities
Y_pred_test = (Y_pred_test_prob > 0.5).astype("int32")  # Apply threshold of 0.5

# Confusion matrix for the test set
cm_test = confusion_matrix(Y_test, Y_pred_test)

# Calculating metrics for the test set
accuracy_test = np.sum(Y_pred_test == Y_test) / len(Y_test)
precision_test = precision_score(Y_test, Y_pred_test)
recall_test = recall_score(Y_test, Y_pred_test)  # Sensitivity
f1_test = f1_score(Y_test, Y_pred_test)

tn_test, fp_test, fn_test, tp_test = cm_test.ravel()
specificity_test = tn_test / (tn_test + fp_test)

# Print out test metrics
print("\nTest Set Results:")
print(f'Accuracy: {accuracy_test:.4f}")
```

```
print(f"Precision: {precision_test:.4f}")

print(f"Sensitivity (Recall): {recall_test:.4f}")

print(f"Specificity: {specificity_test:.4f}")

print(f"F1 Score: {f1_test:.4f}")


# Printing classification report

print("\nClassification Report:")

print(classification_report(Y_test, Y_pred_test))


# ROC Curve and AUC for test set

fpr_test, tpr_test, _ = roc_curve(Y_test, Y_pred_test_prob)

roc_auc_test = auc(fpr_test, tpr_test)


# ROC Curve and AUC for test set with adjusted x-axis start

plt.figure(figsize=(8, 6))  # Adjusting the figure size

plt.plot(fpr_test, tpr_test, color='darkorange', lw=2, label=f'ROC curve (area =

{roc_auc_test:.2f})')

plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')

plt.xlim([-0.05, 1.0])  # Shifting the x-axis slightly to the left

plt.ylim([0.0, 1.05])

plt.xlabel('False Positive Rate')
```

```
plt.ylabel('True Positive Rate')

plt.title('Receiver Operating Characteristic (Test Set)')

plt.legend(loc="lower right")


# Adjusting the margins for better visualization

plt.subplots_adjust(left=0.1, right=0.9, top=0.9, bottom=0.1)


plt.show()


# Confusion matrix visualization for test set

plt.figure(figsize=(10, 7))

sns.heatmap(cm_test, annot=True, fmt='d', cmap='Blues')

plt.xlabel('Predicted')

plt.ylabel('Actual')

plt.title('Confusion Matrix (Test Set)')

plt.show()
```

Appendix B

Dataset

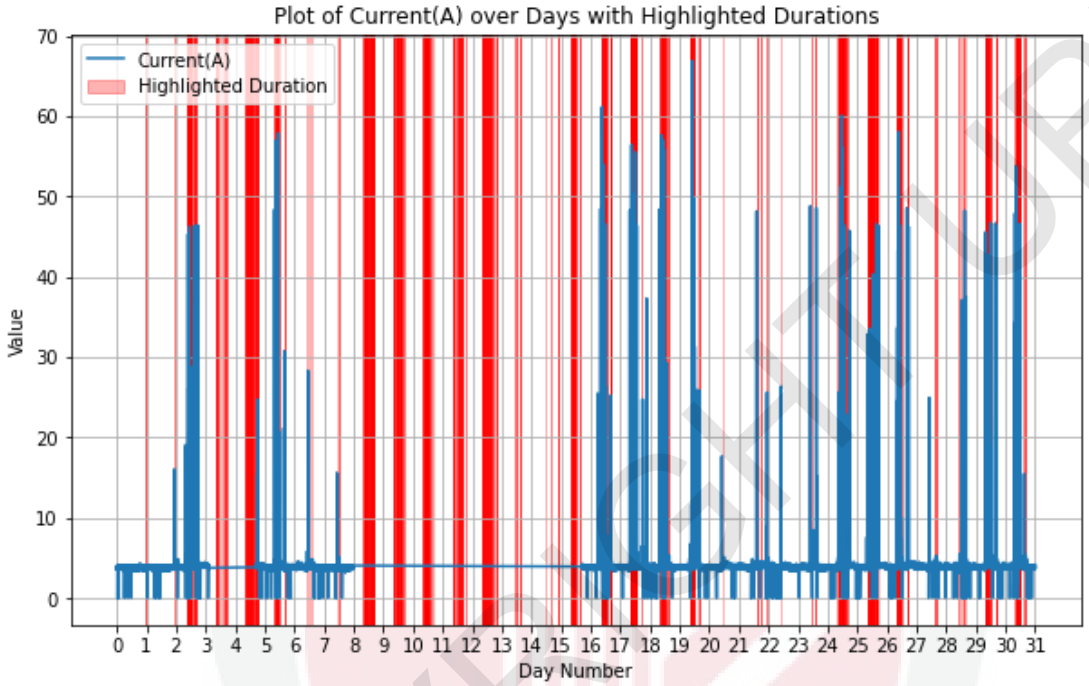
Current	Radiation	Temperature	Humidity	X	Y
4.626756	19004.28	20.8	53.5	-0.991	0.027
4.256624	89478.48	21.5	51.5	-0.985	0.026
4.265372	0	22.2	52	-0.986	0.021
4.34155	46182.44	21.3	42	-0.979	0.019
3.850556	10324.44	22.1	39	-0.998	0.017
3.54746	0	19.5	47	-0.994	0.026
3.496731	21471.96	22.5	47	-0.985	0.021
4.042864	9157.71	22.2	41.5	-0.972	0.006
3.900106	0	22	48	-0.995	0.02
4.113165	2365.07	21.3	49.5	-0.99	0.023
3.521462	0	20.3	43.5	-0.987	0.028
3.616104	0	21	55.5	-0.976	0.026
3.52326	536870.9	21	44.5	-0.988	0.017
5.329096	47197.44	19.9	41.5	-0.981	0.022
4.175244	0	20.9	41.5	-0.978	0.016
3.976616	0	21	56.5	-0.99	0.024
4.083233	0	22.2	45	-0.97	0.036
5.522747	39768.21	22.3	49.5	-0.991	0.028
3.849946	38347.92	20.7	41	-0.991	0.018
4.047408	0	22	53	-0.976	0.027
3.76789	10764.32	21.5	40.5	-0.984	0.021
3.567756	12521.77	22.2	50.5	-0.994	0.026
4.15998	16711.93	22.3	38.5	-0.977	0.024
3.825842	0	21.1	42.5	-0.985	0.025
4.176133	41297.76	20.9	48	-0.988	0.025
3.827674	19611.72	21.4	43.5	-0.983	0.021

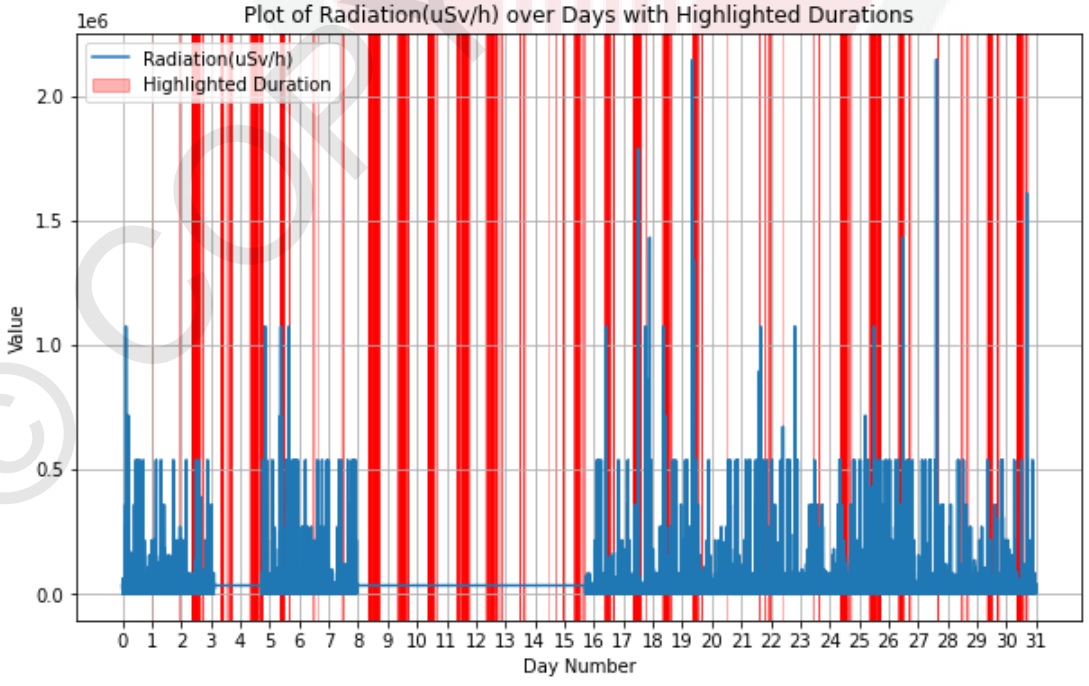
5.273796	20648.88	22.8	38	-0.983	0.02
3.80056	12820.79	22.2	42	-0.977	0.028
3.624221	41297.76	20.8	49.5	-0.98	0.019
3.809787	0	22.3	43	-0.977	0.027
3.896916	0	21	54.5	-0.987	0.02
3.898157	0	21.2	55.5	-0.98	0.028
3.97947	0	21.6	54	-0.988	0.029
3.421705	39768.21	21.4	45.5	-0.981	0.022
3.383224	0	21.1	51	-0.984	0.025

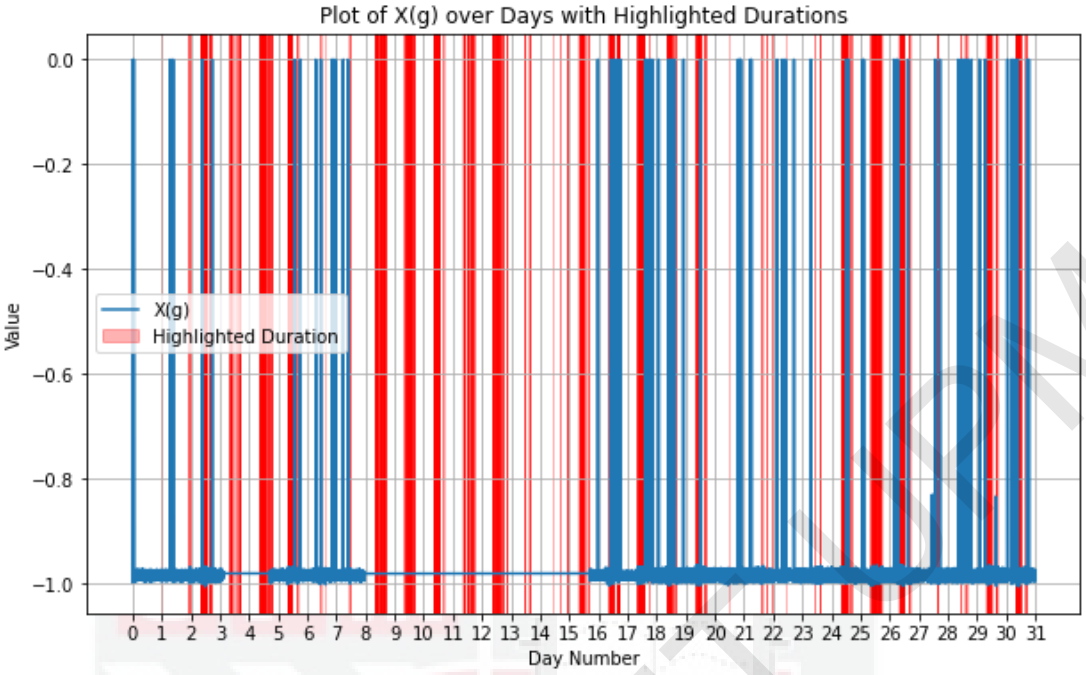
Appendix C

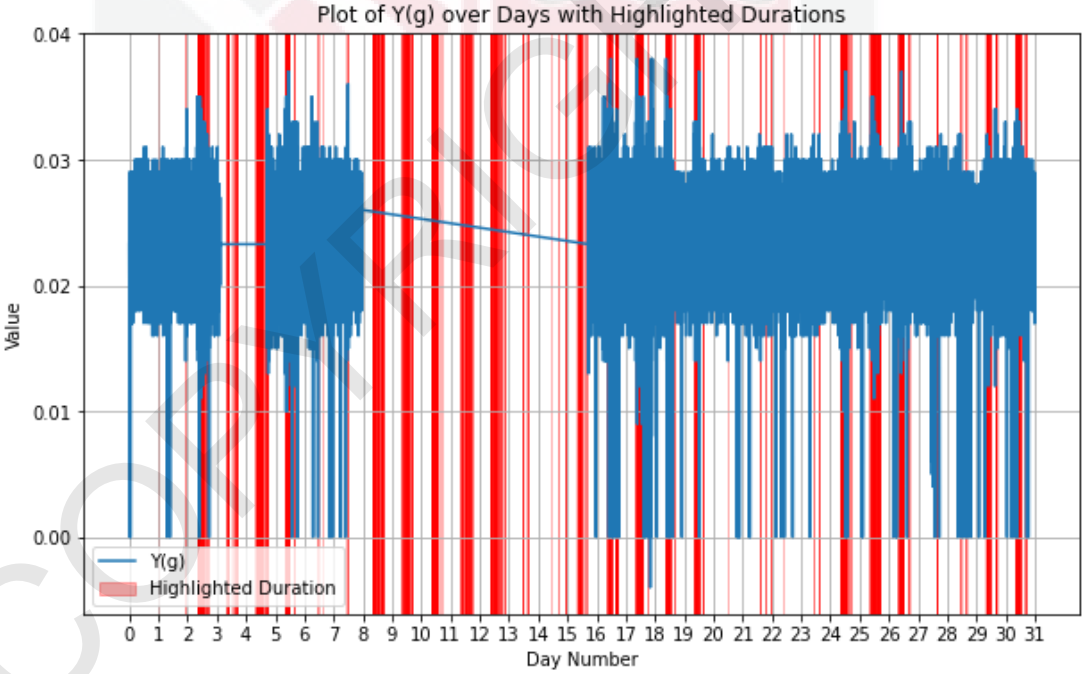
DATA MAPPING

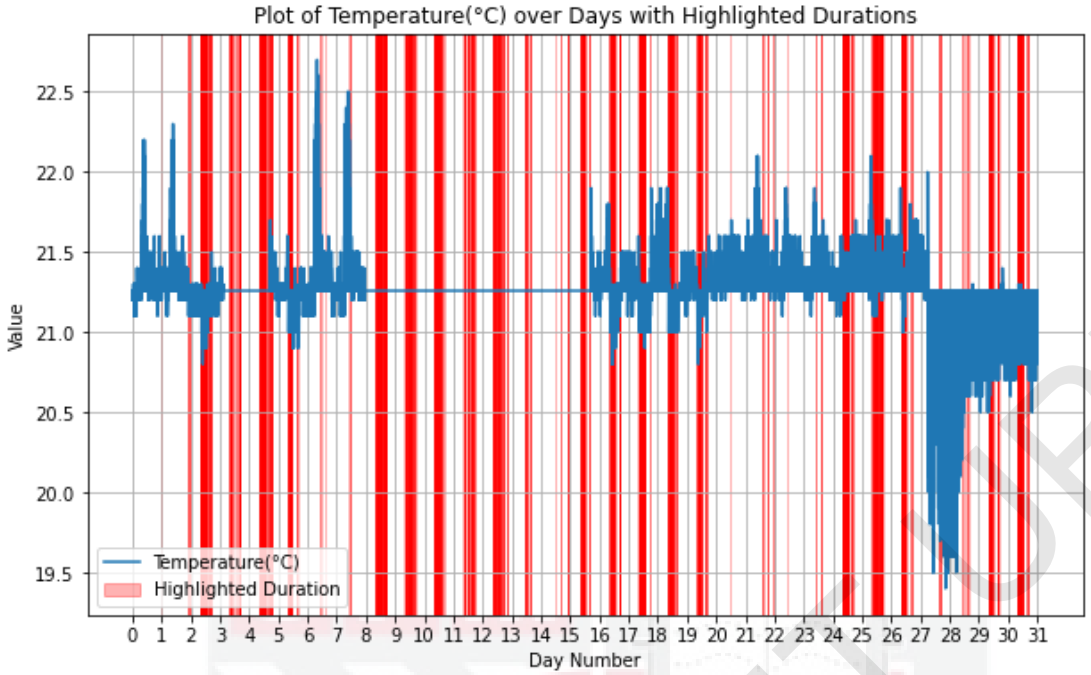
January

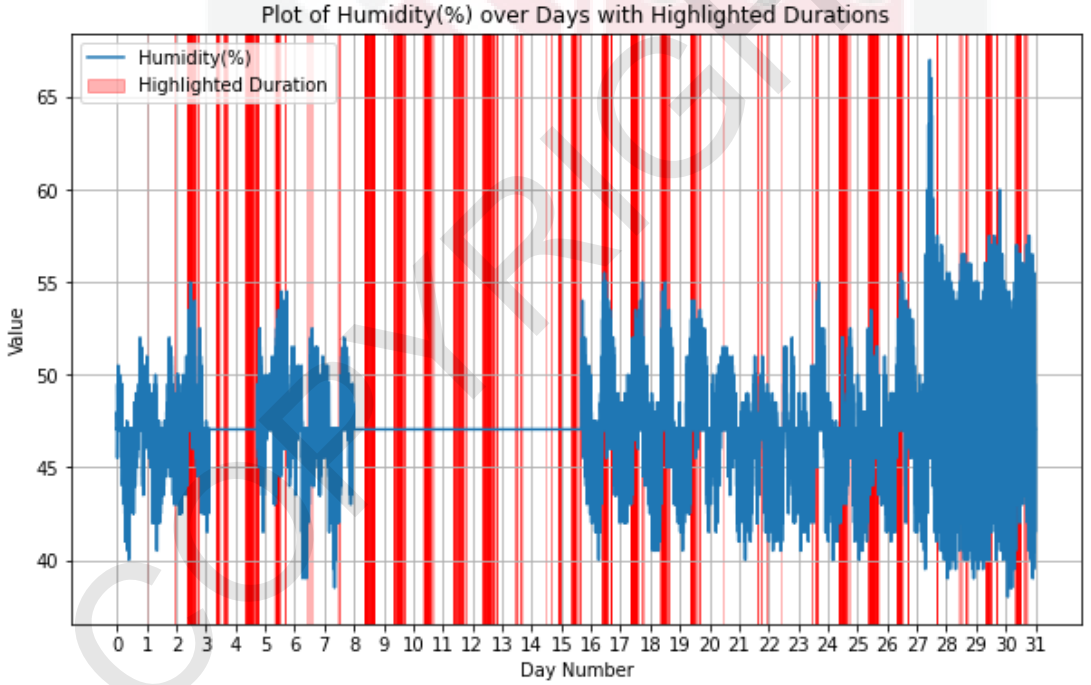




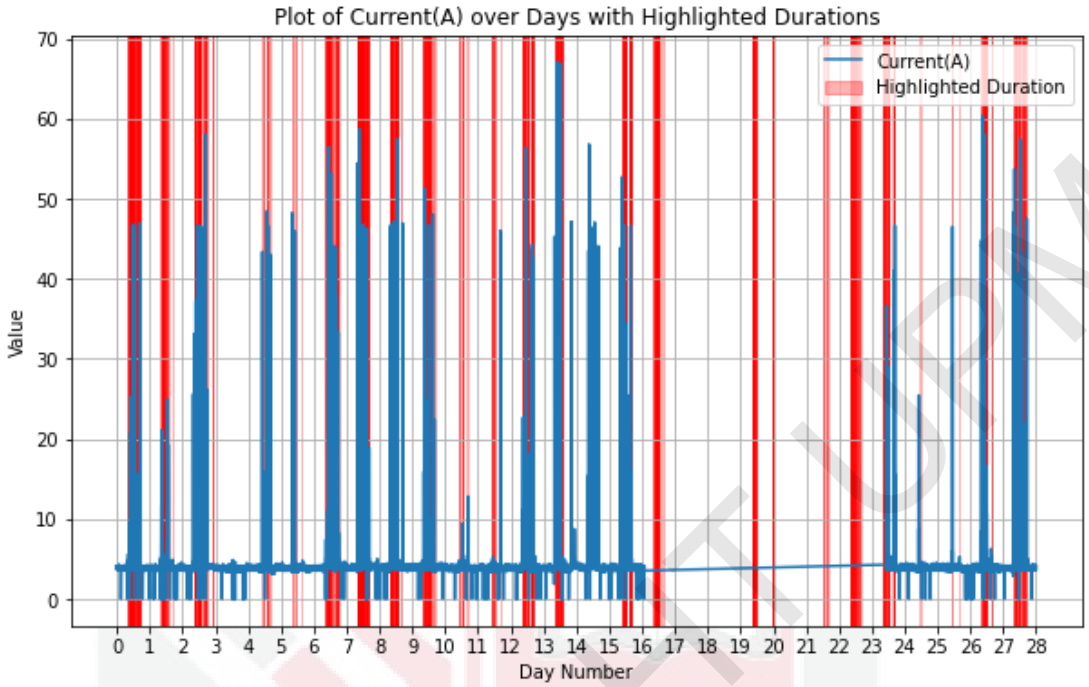

Plot of X(g) over Days with Highlighted Durations
Value
Day Number
X(g)
Highlighted Duration

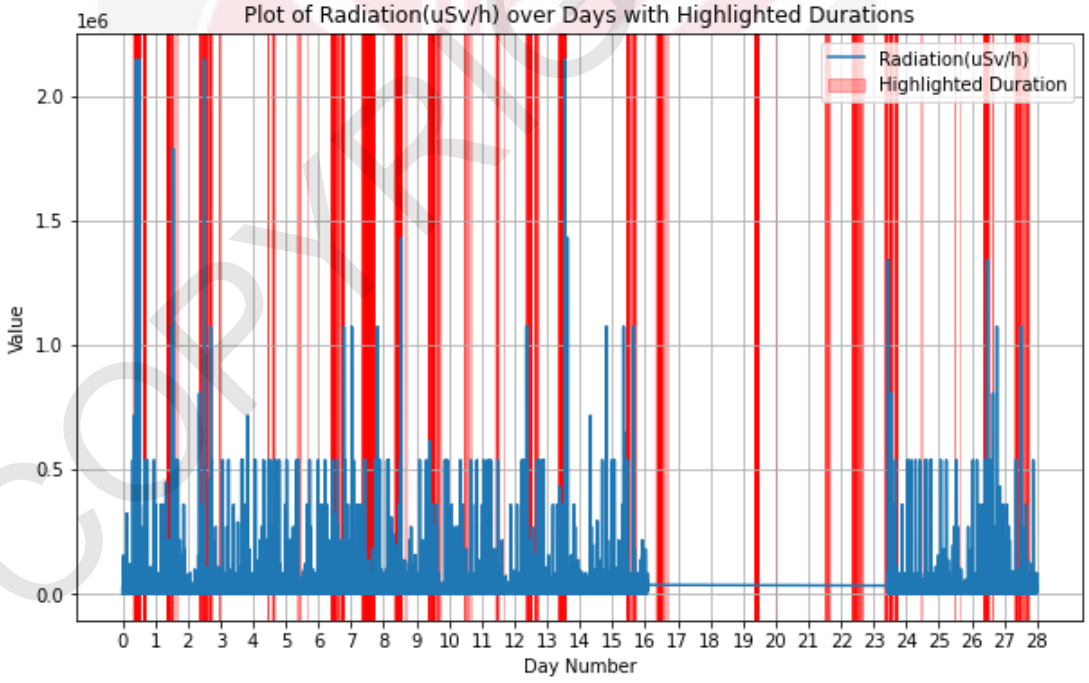

Plot of Y(g) over Days with Highlighted Durations
Value
Day Number
Y(g)
Highlighted Duration

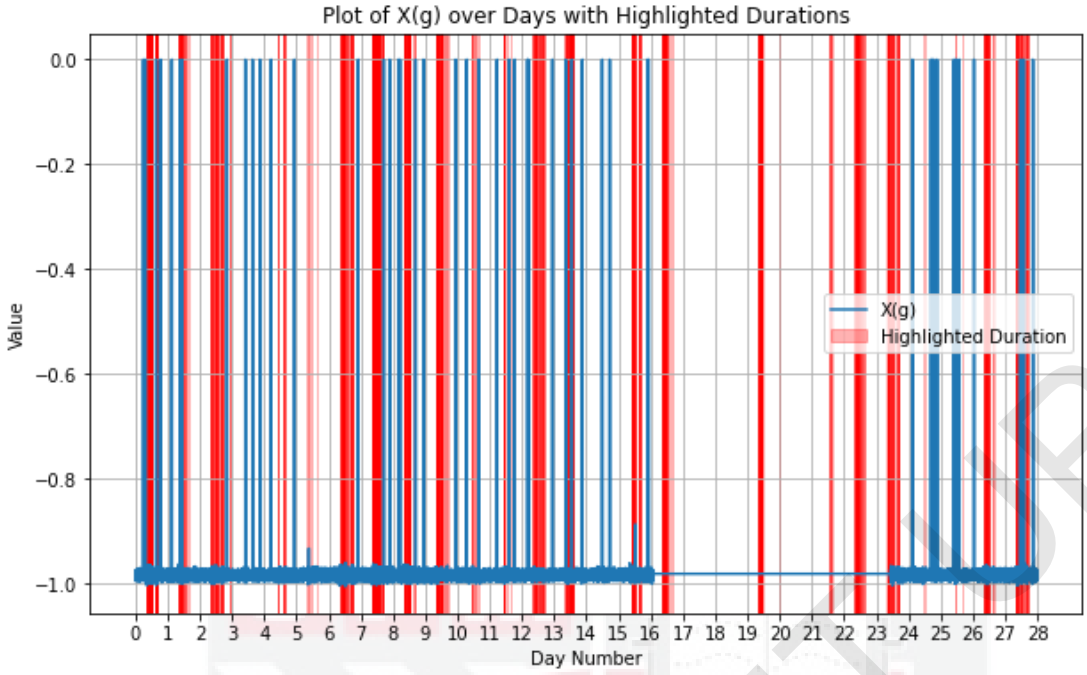


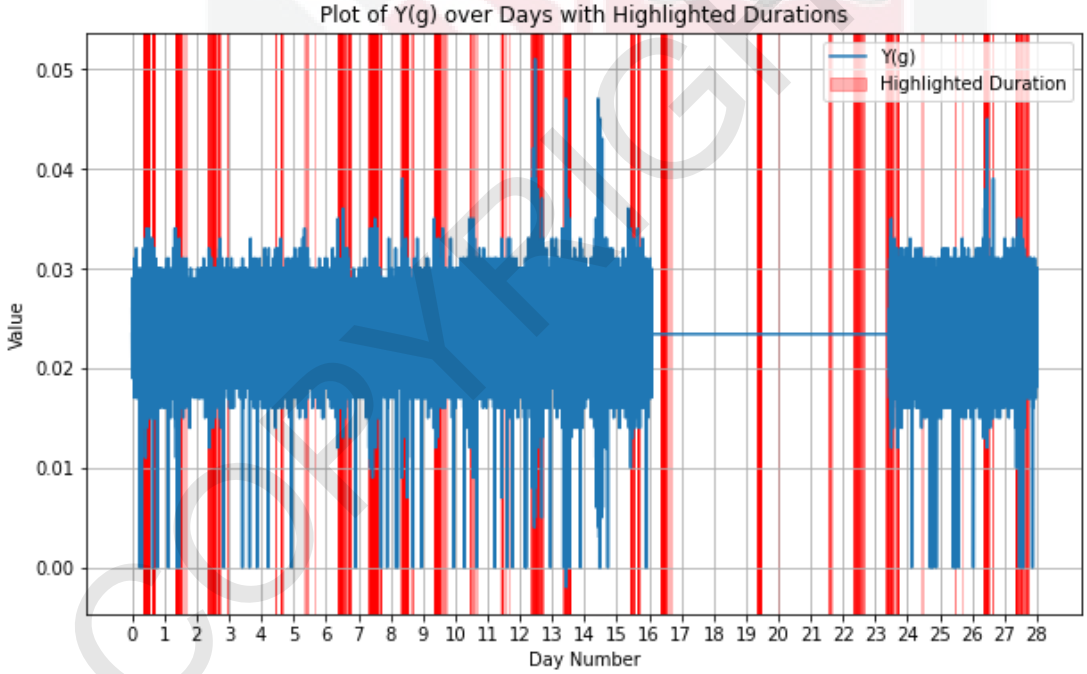


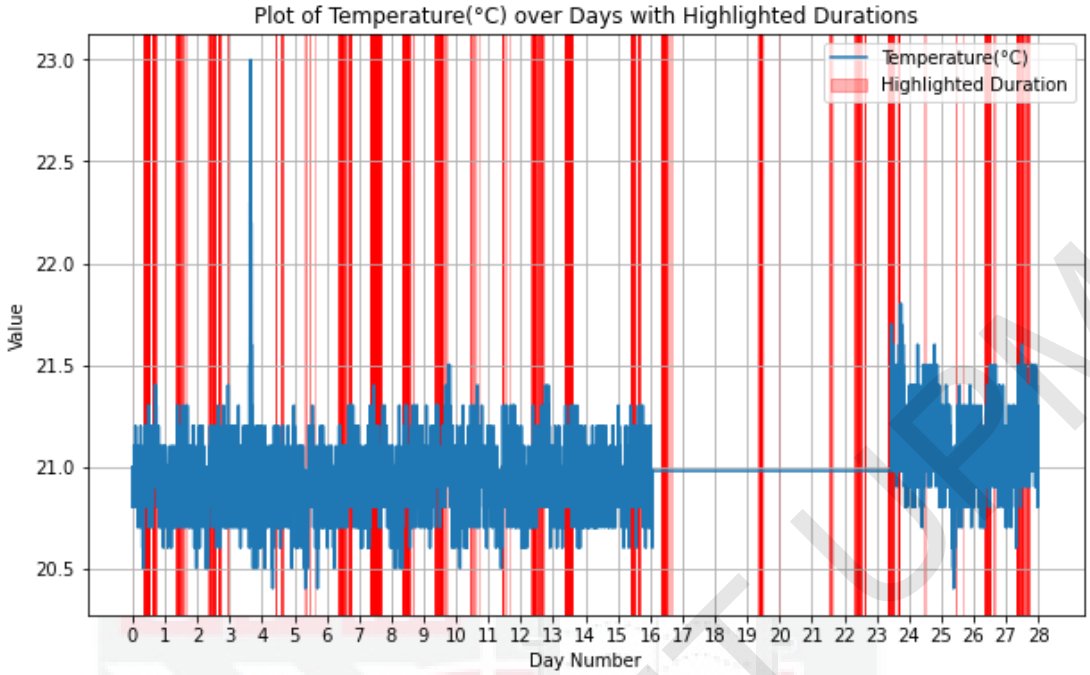
February

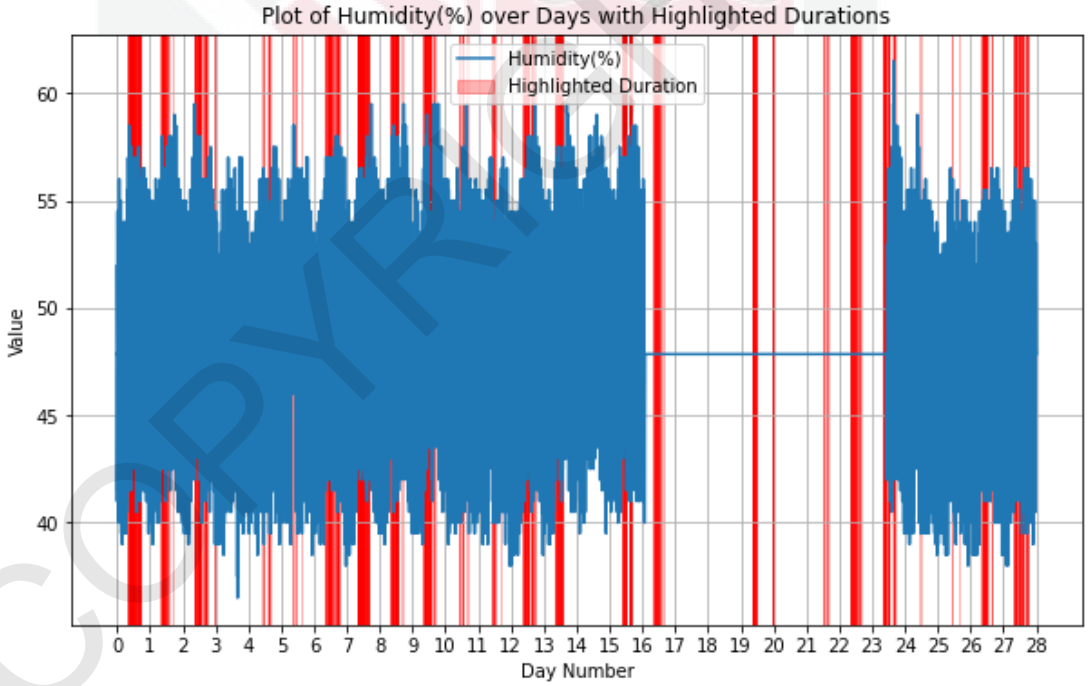




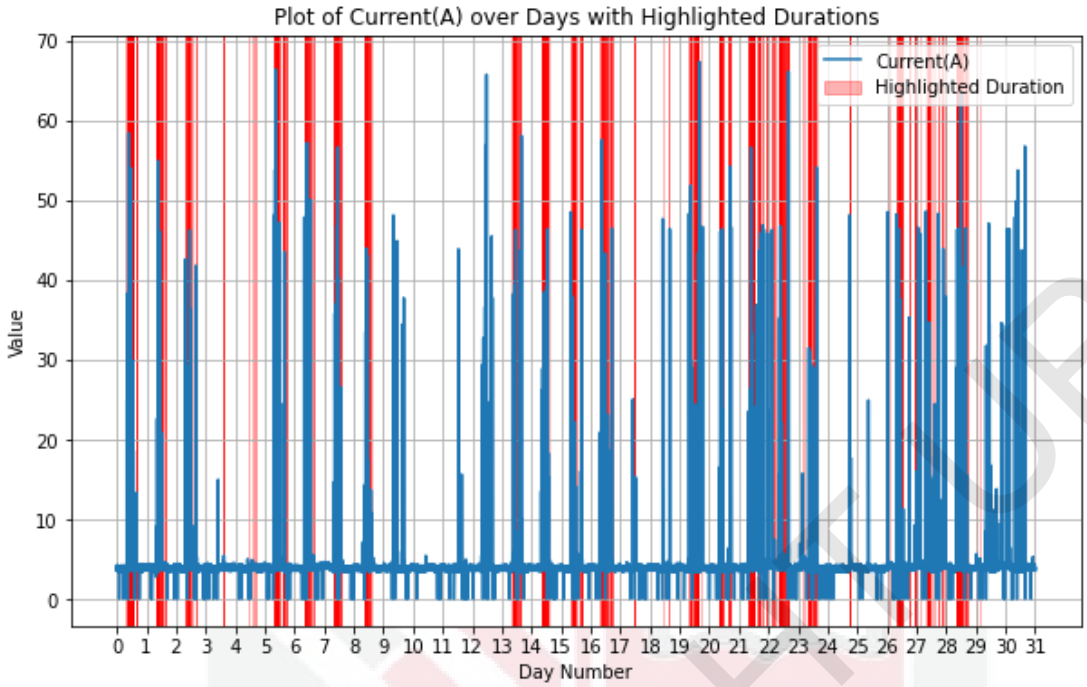


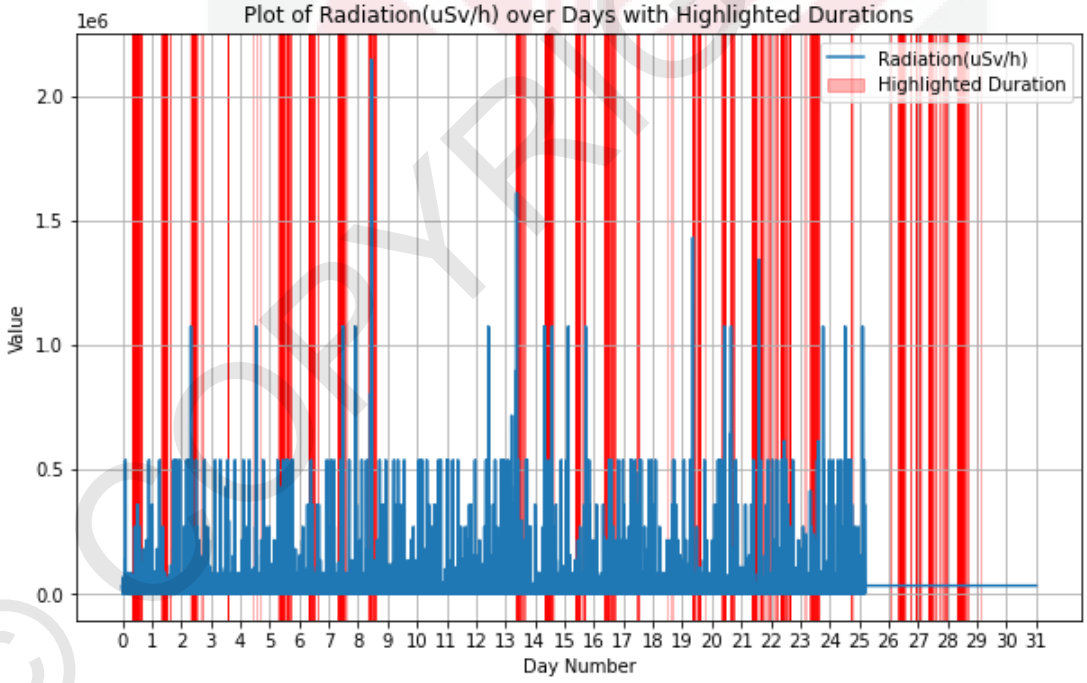


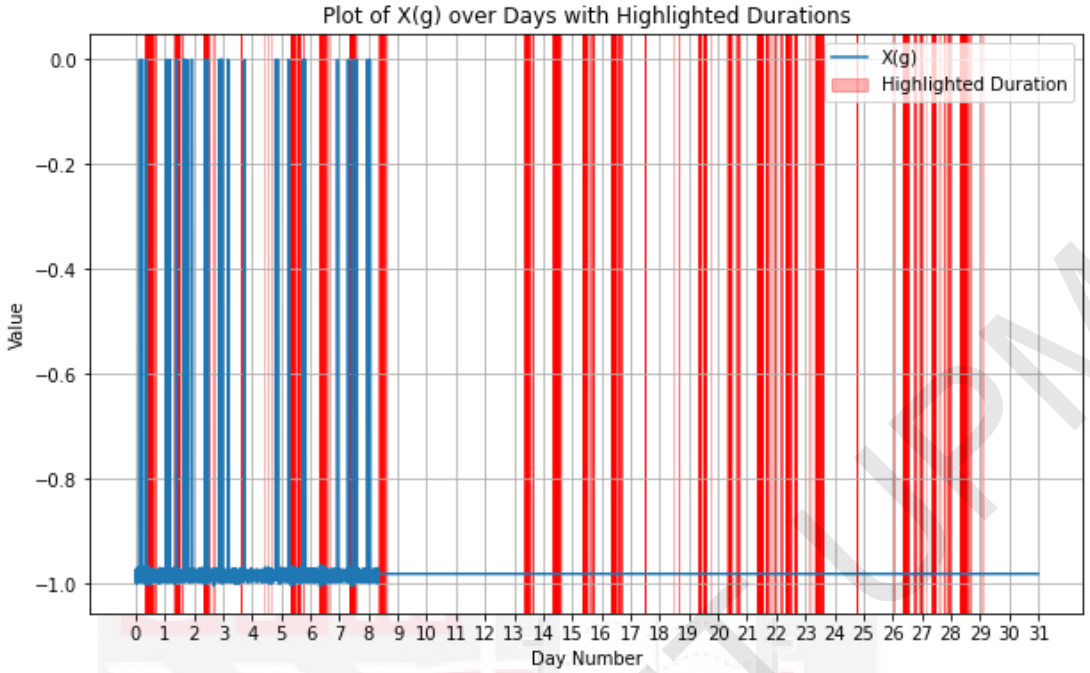


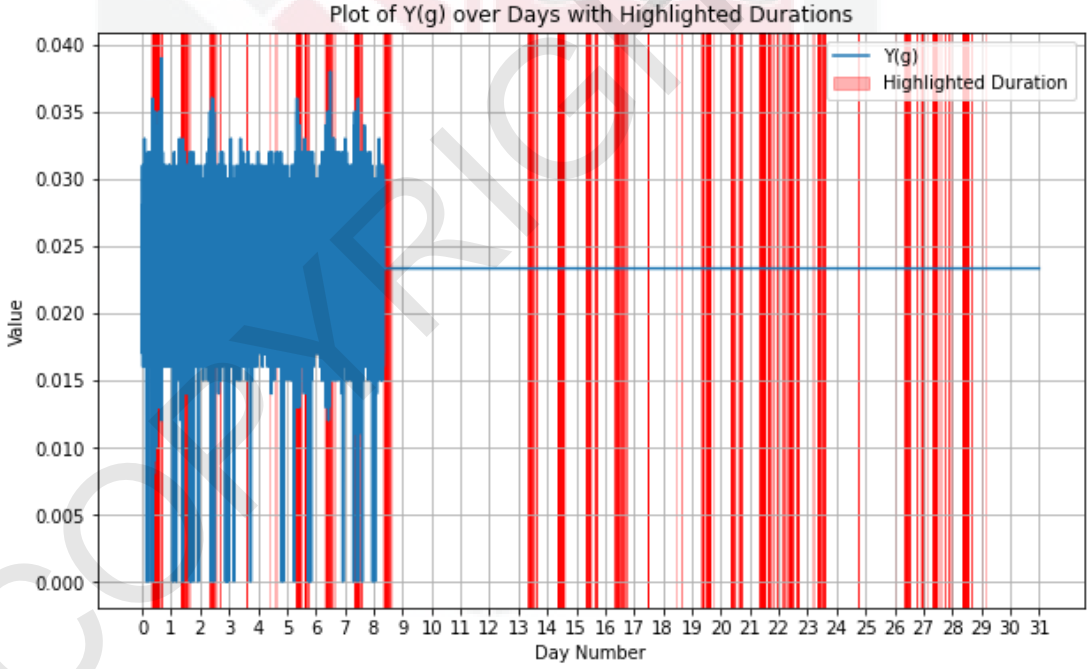


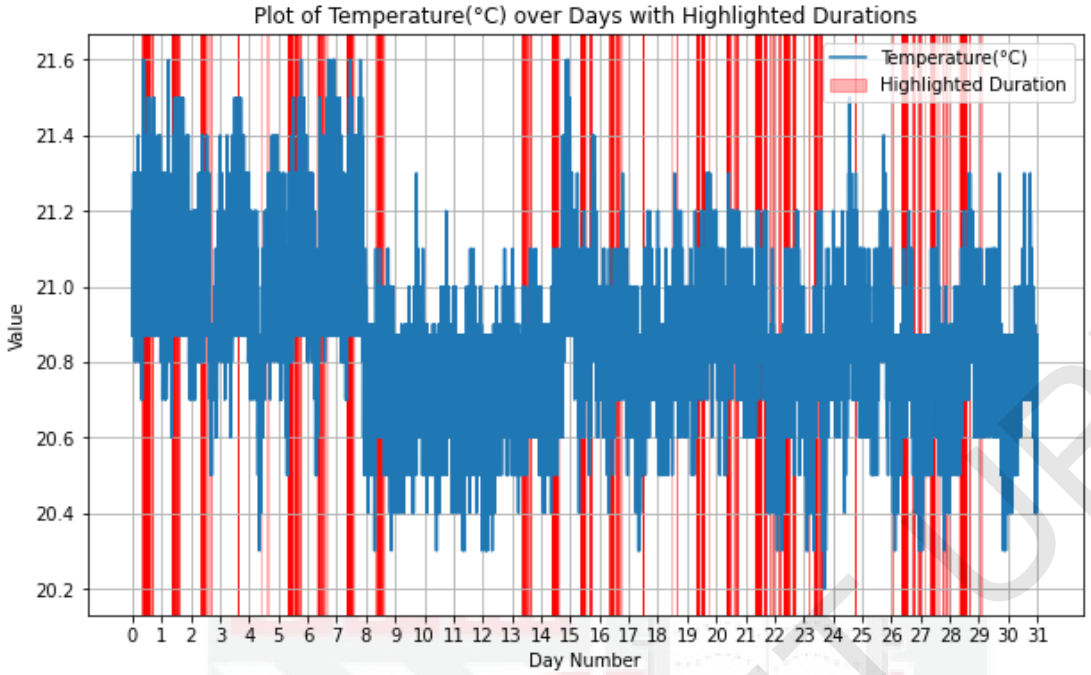
March

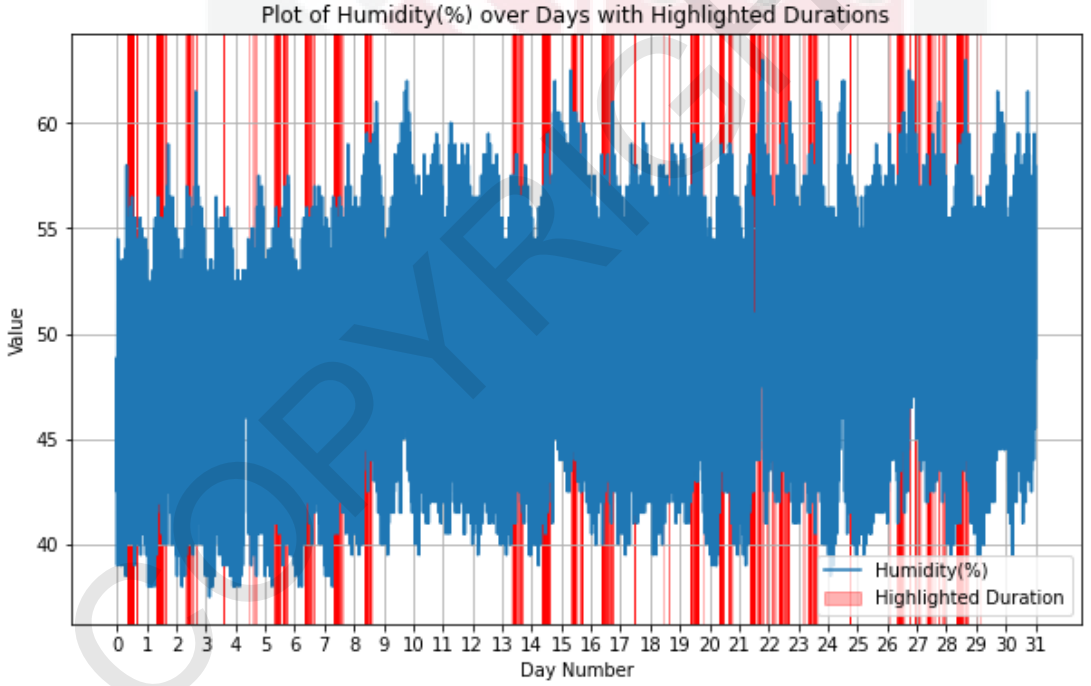




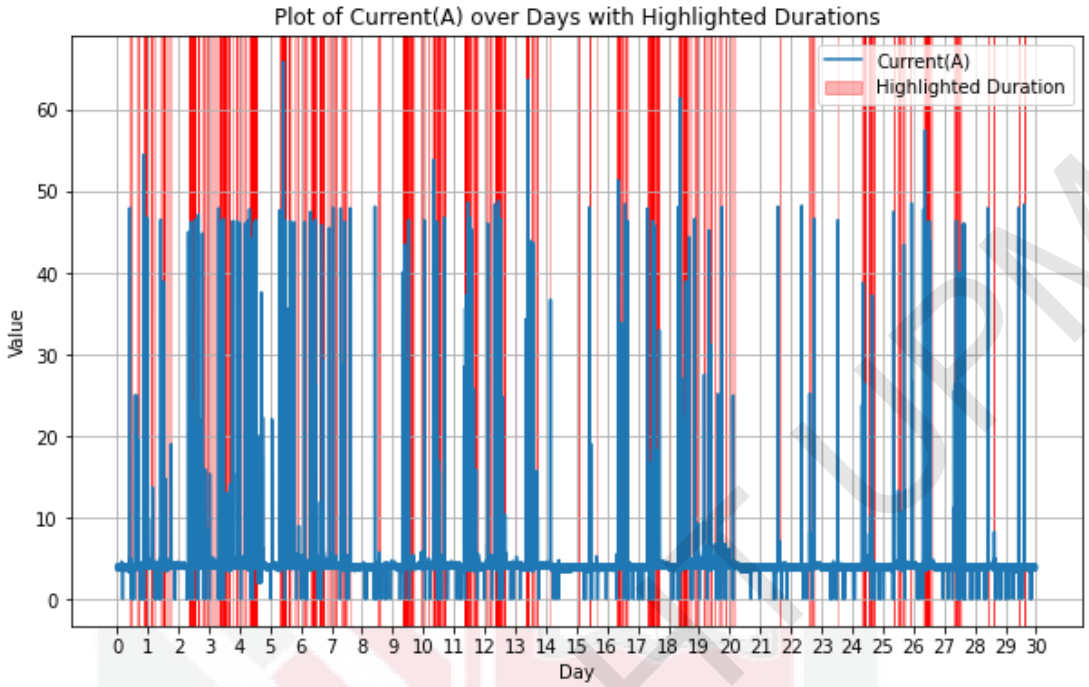


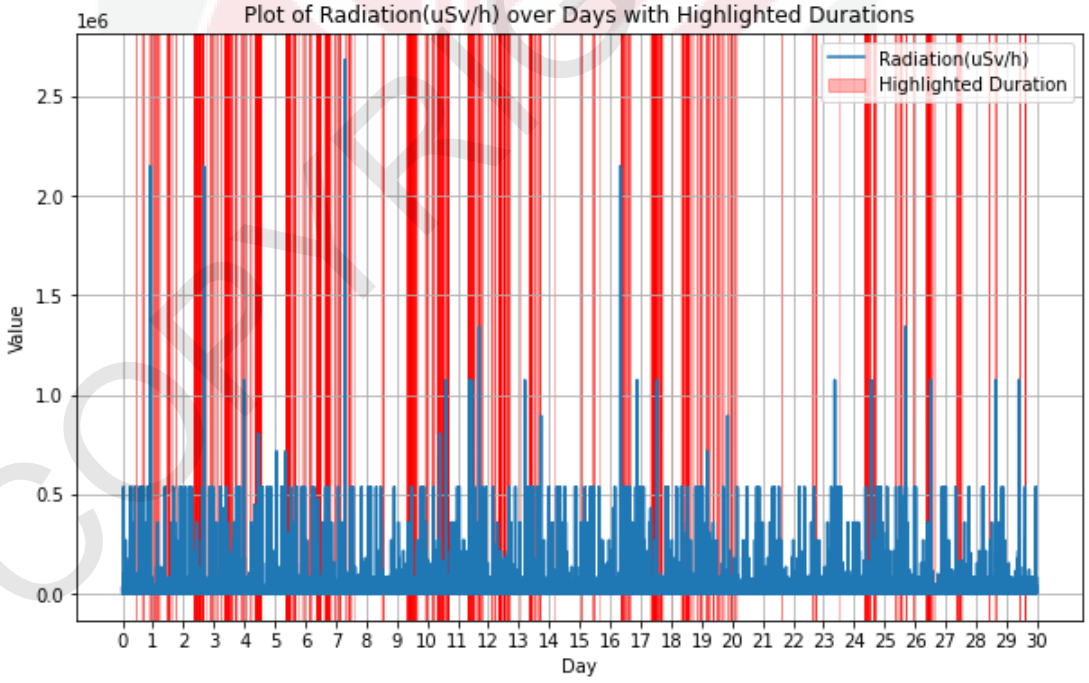


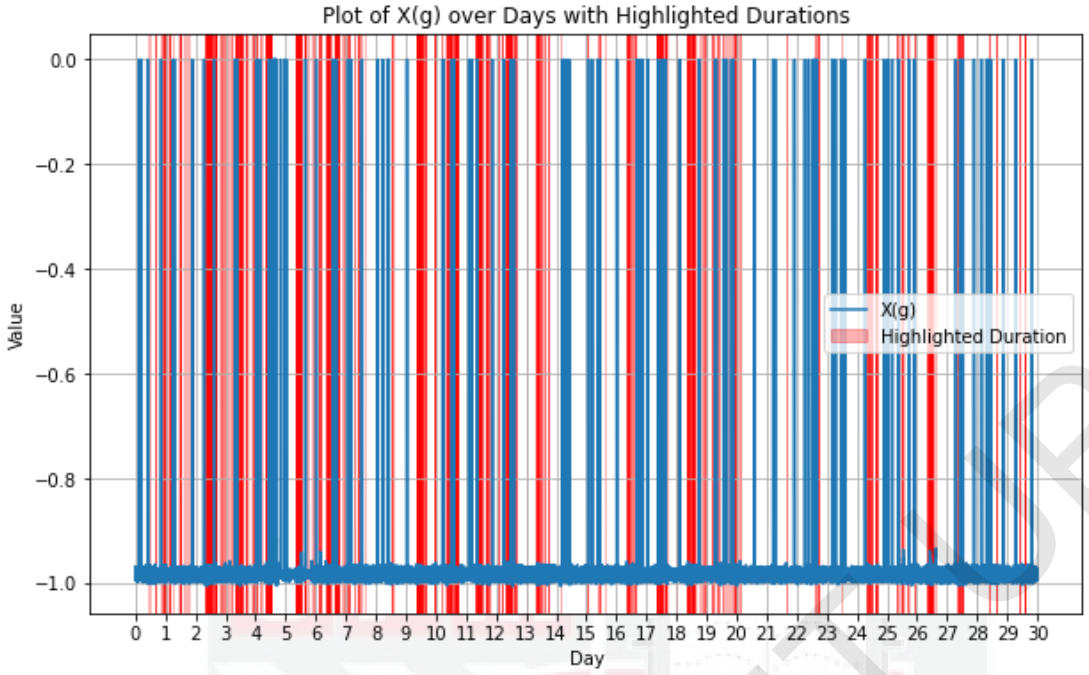


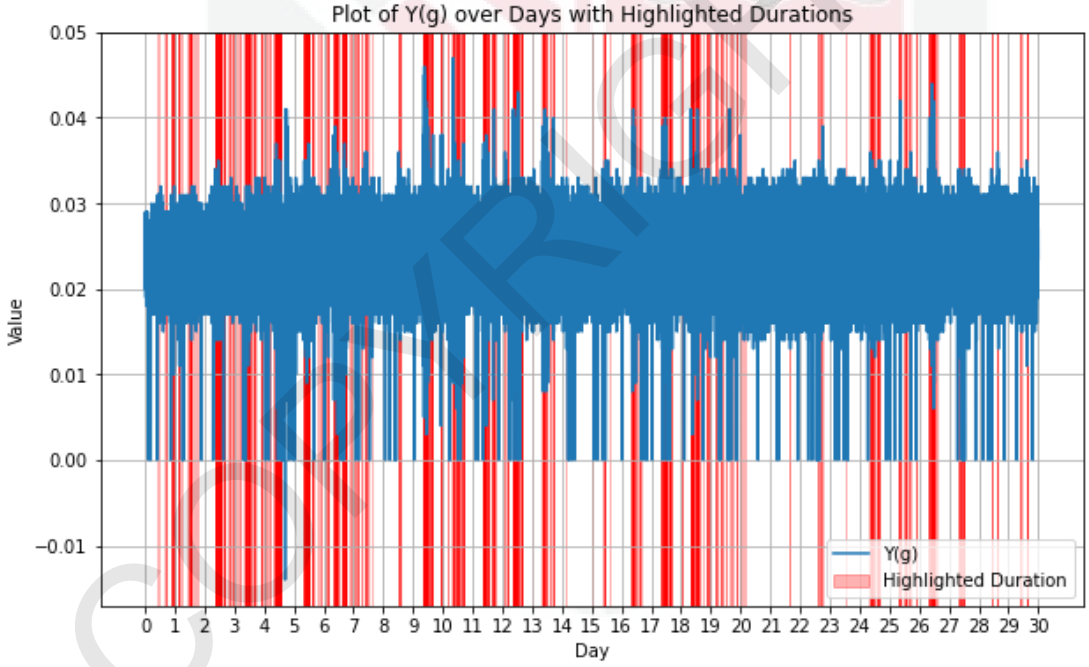


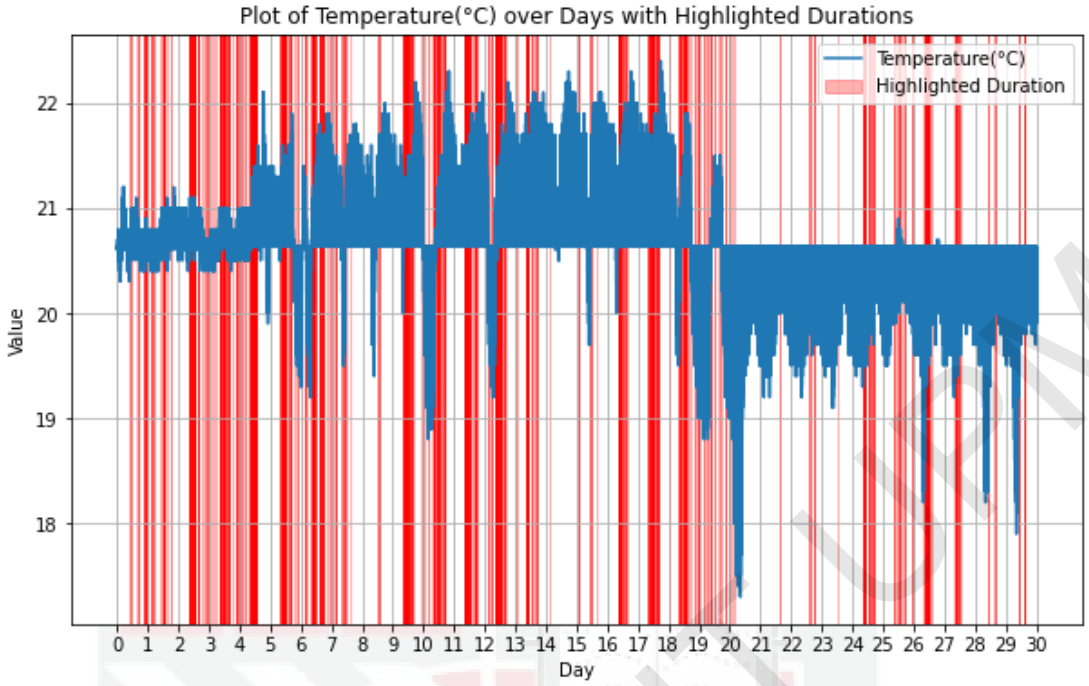
April

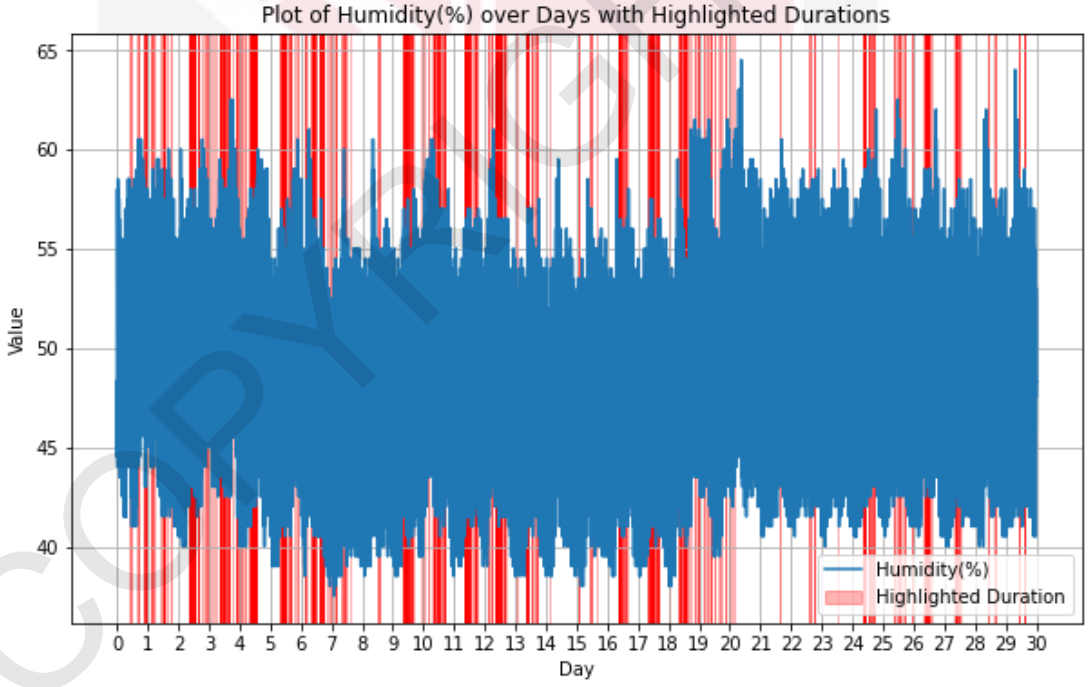




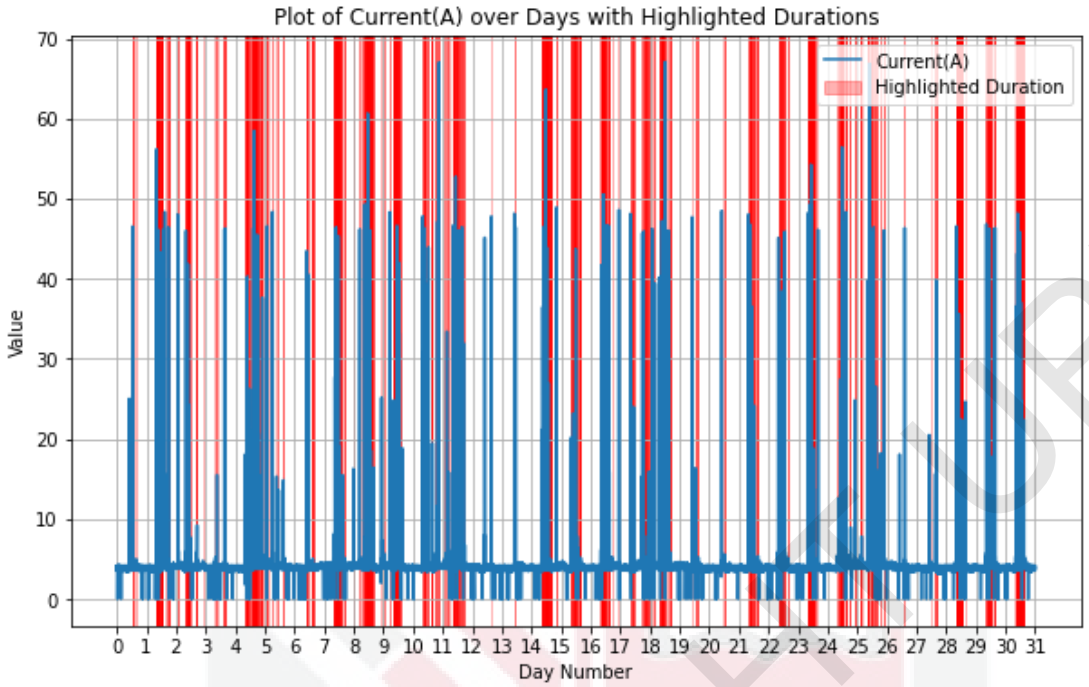


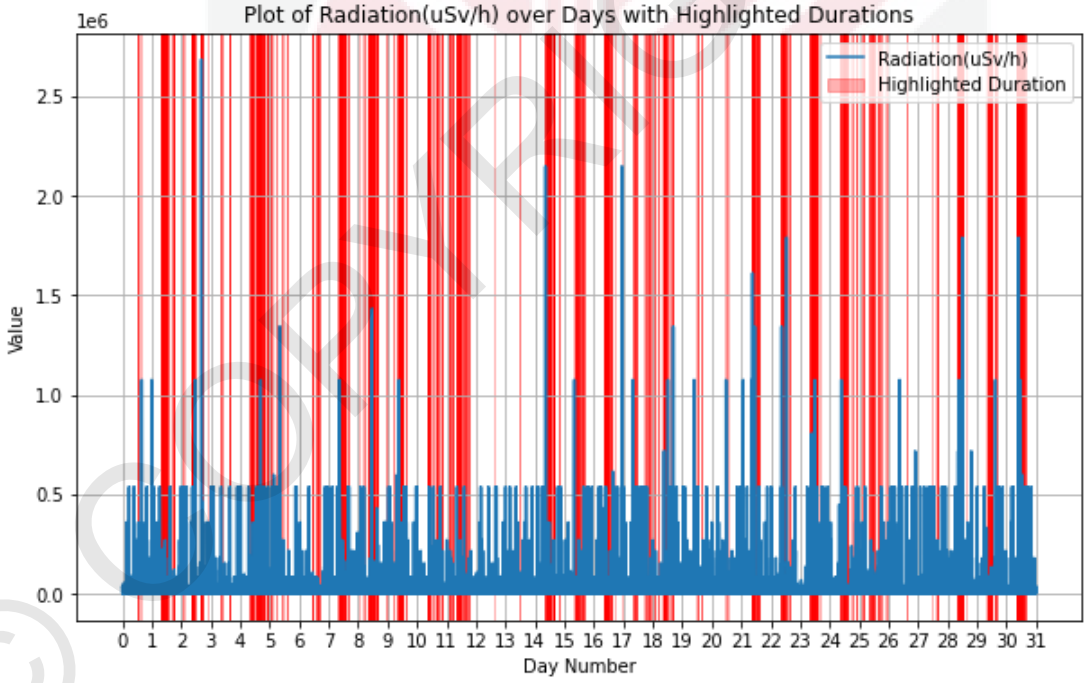


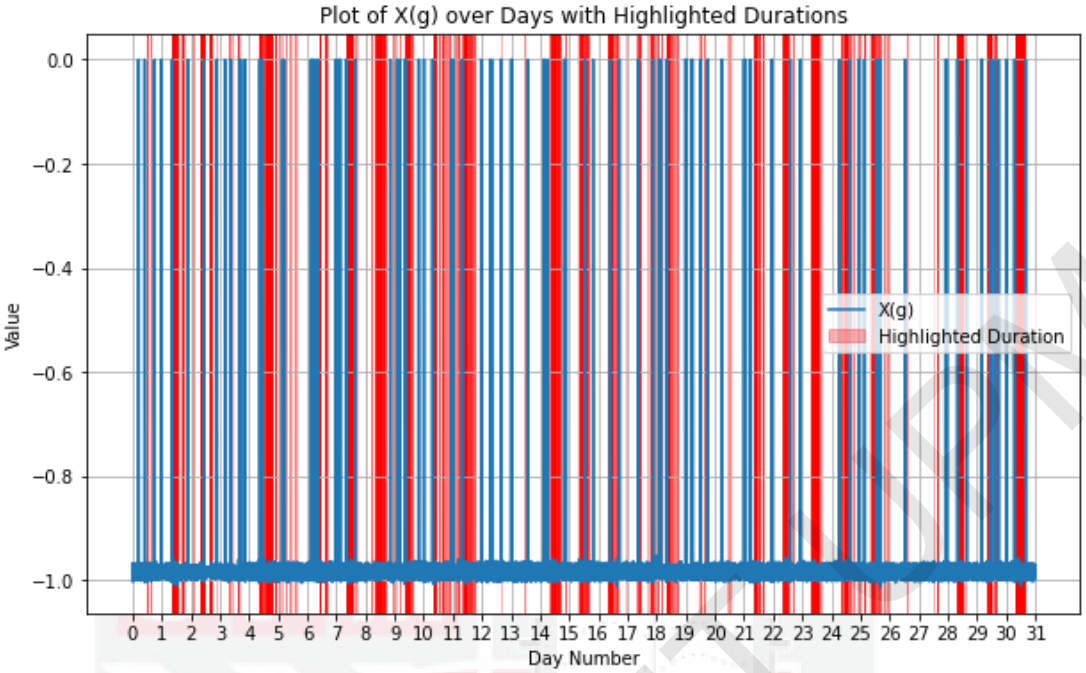


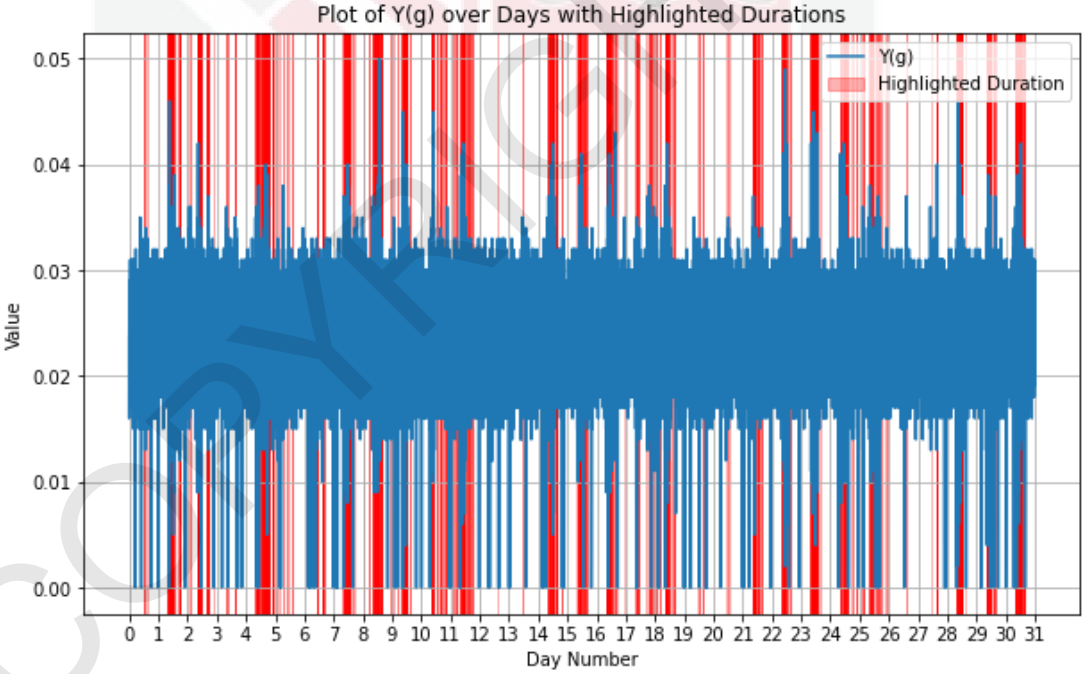


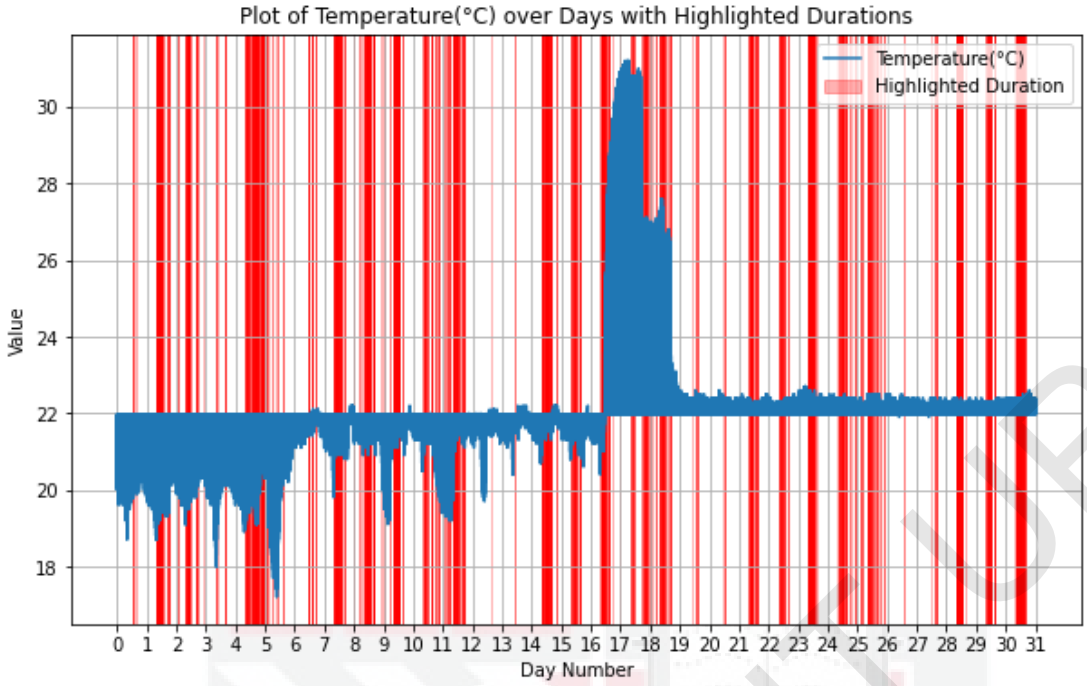
May

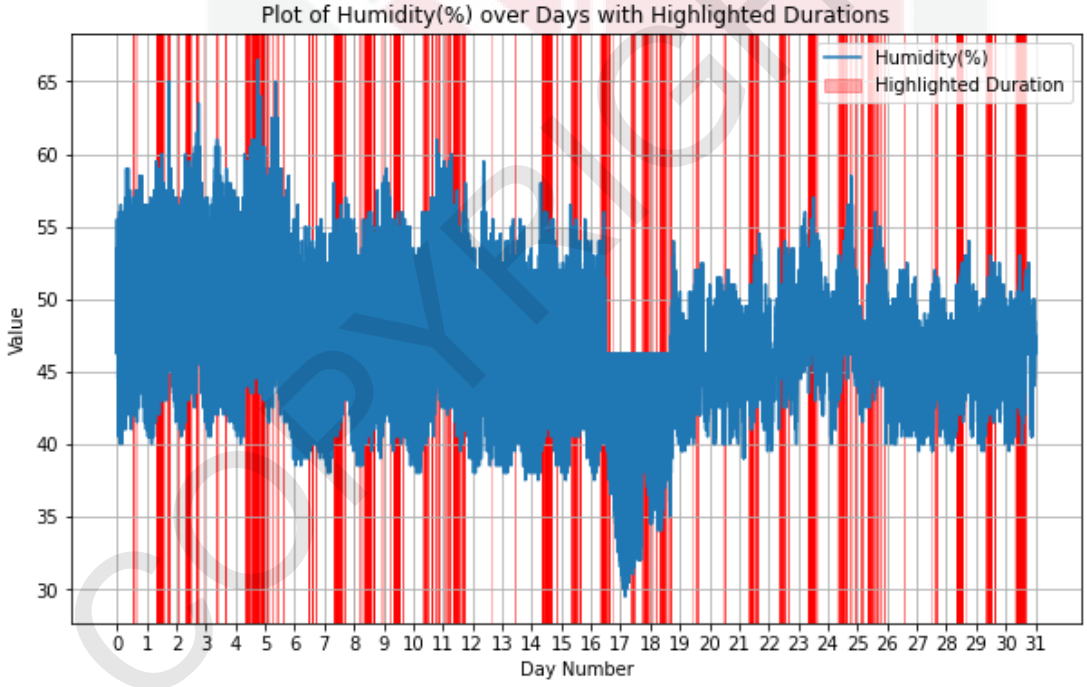




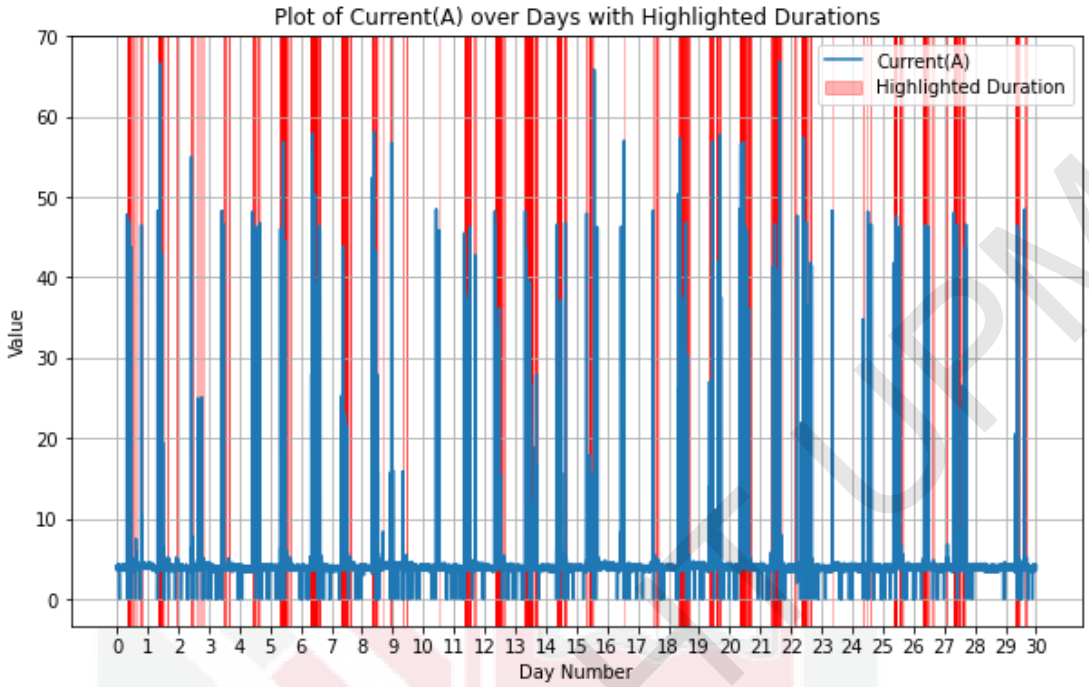

Plot of X(g) over Days with Highlighted Durations
Value
Day Number
X(g)
Highlighted Duration

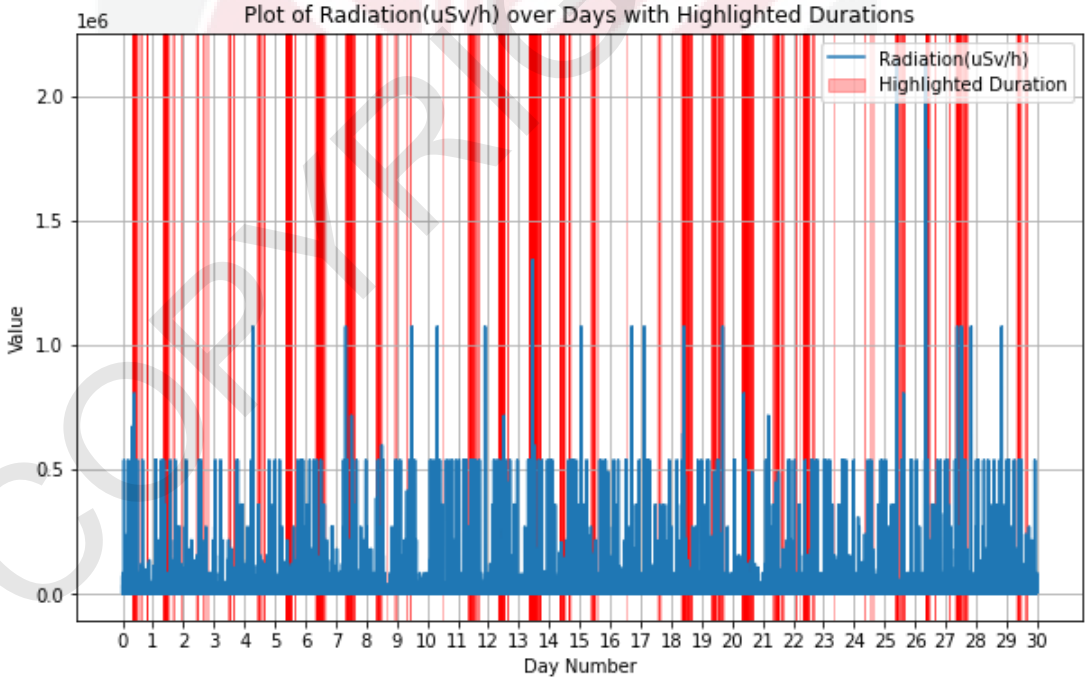

Plot of Y(g) over Days with Highlighted Durations
Value
Day Number
Y(g)
Highlighted Duration

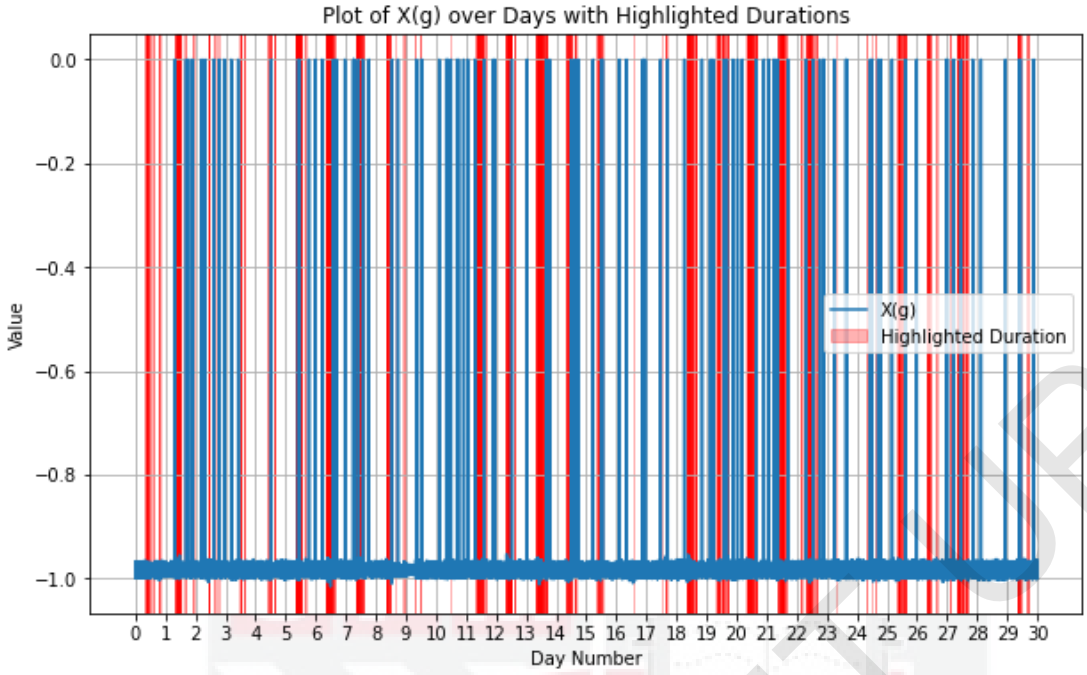


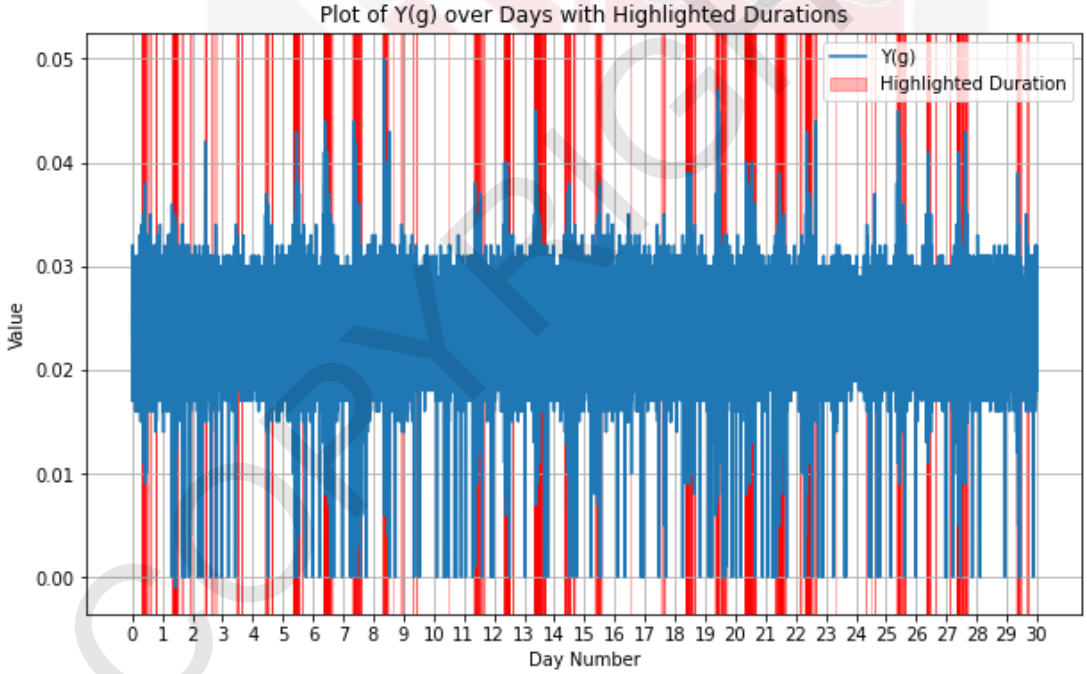


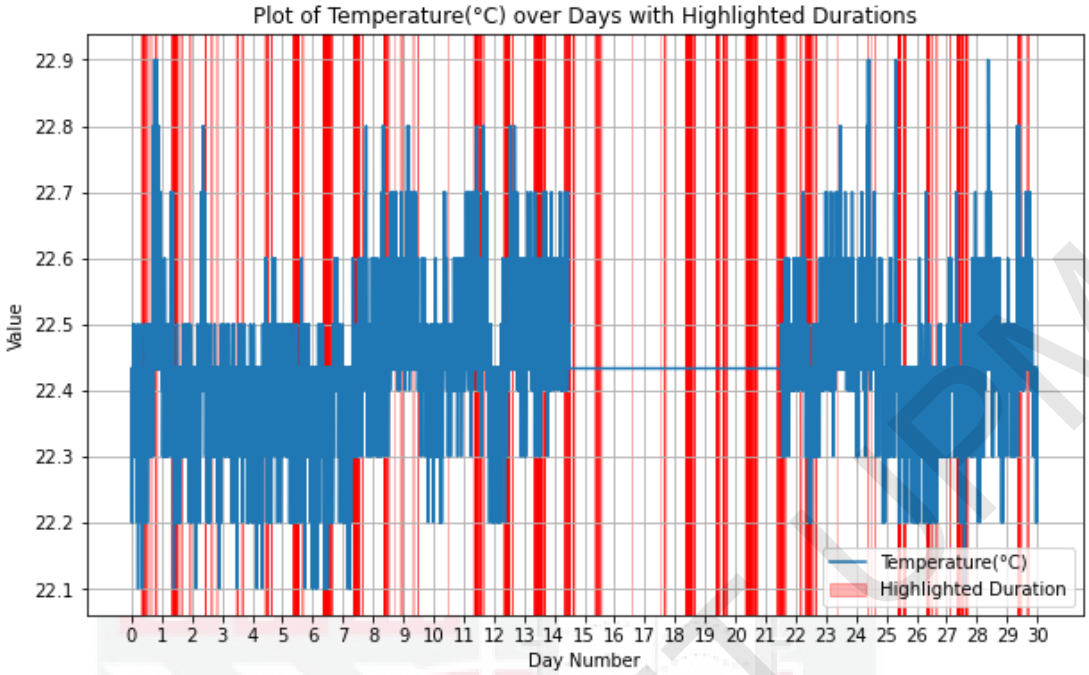
June

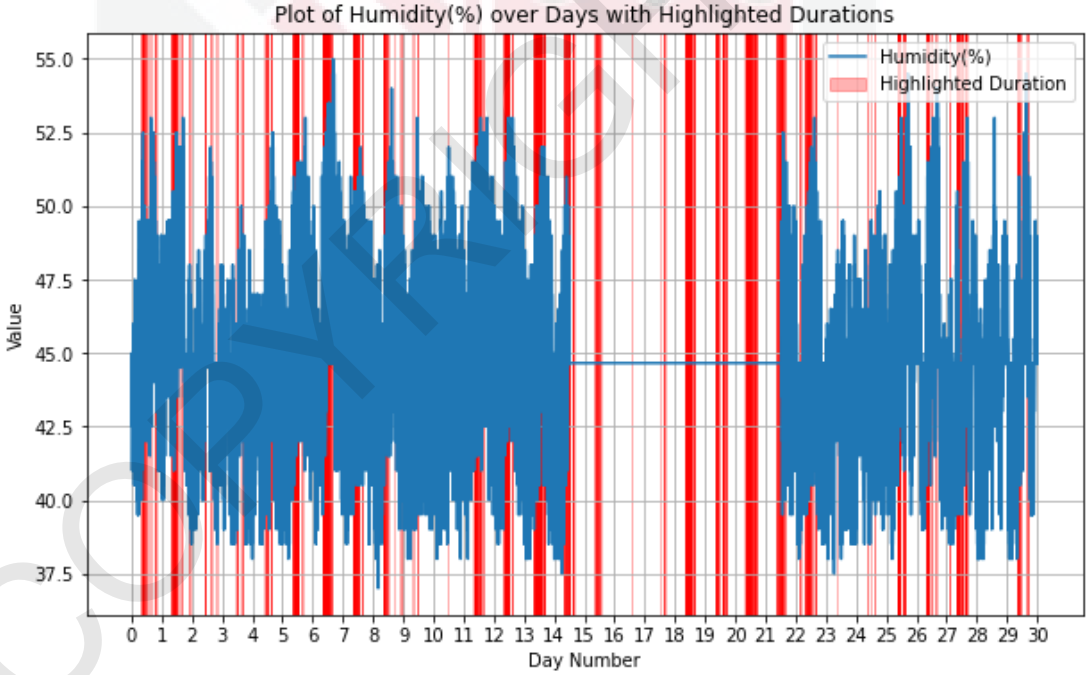




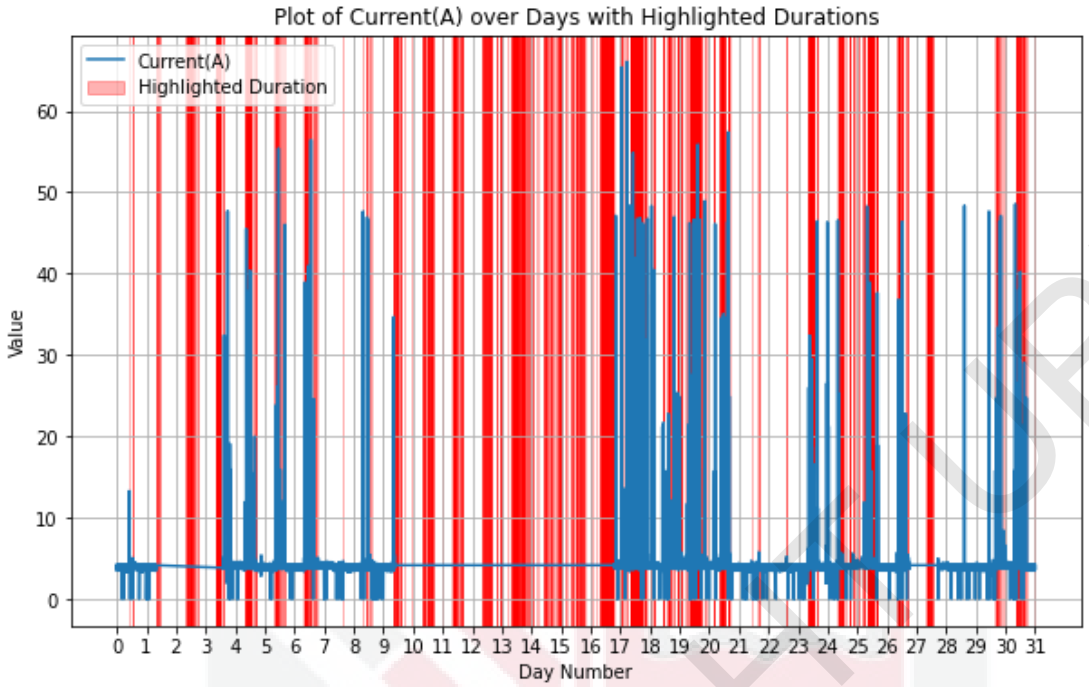


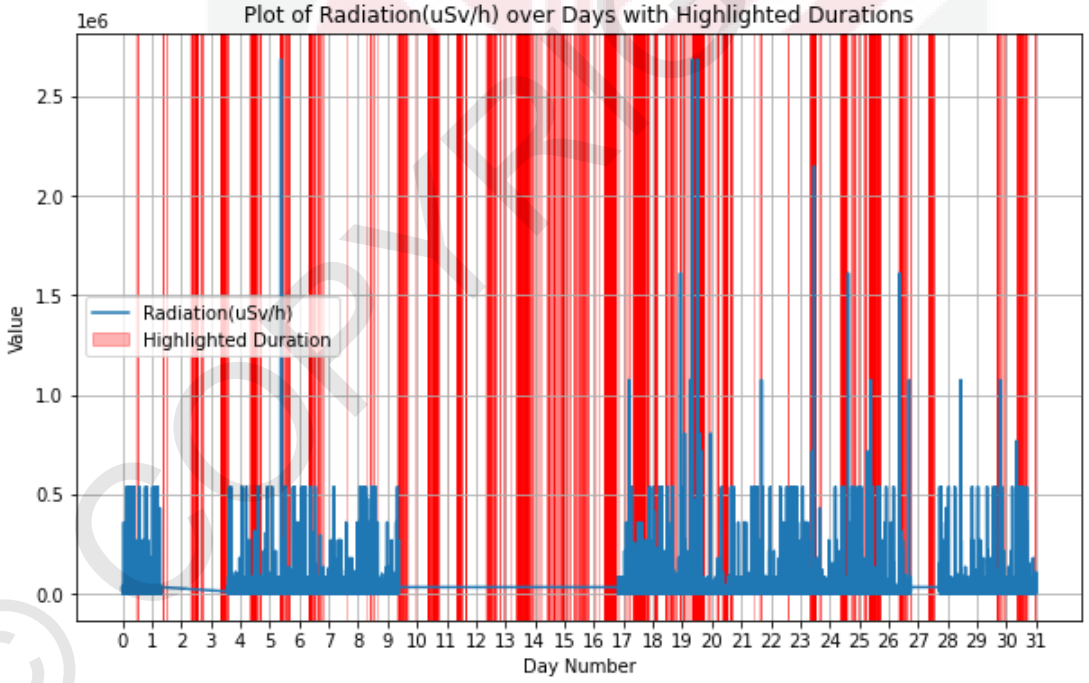


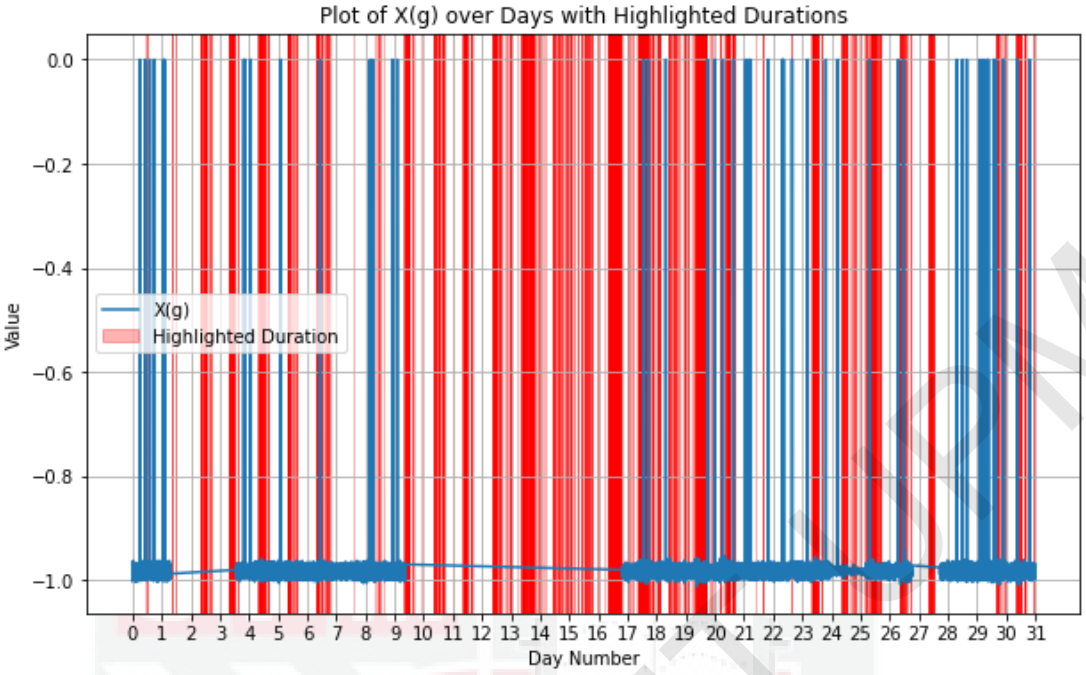


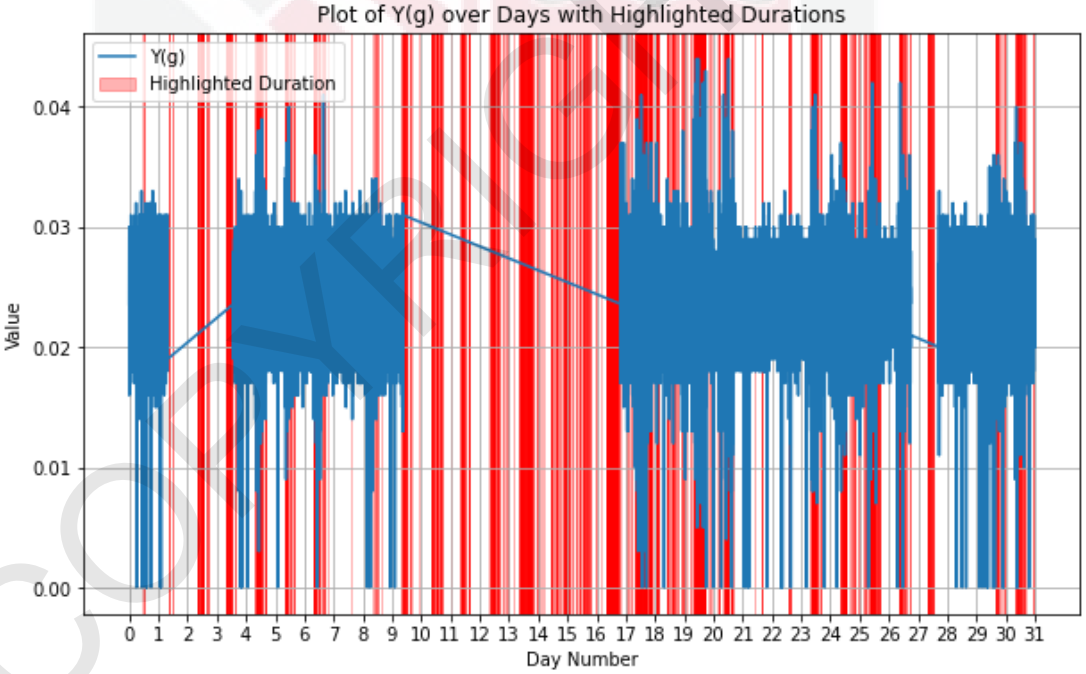


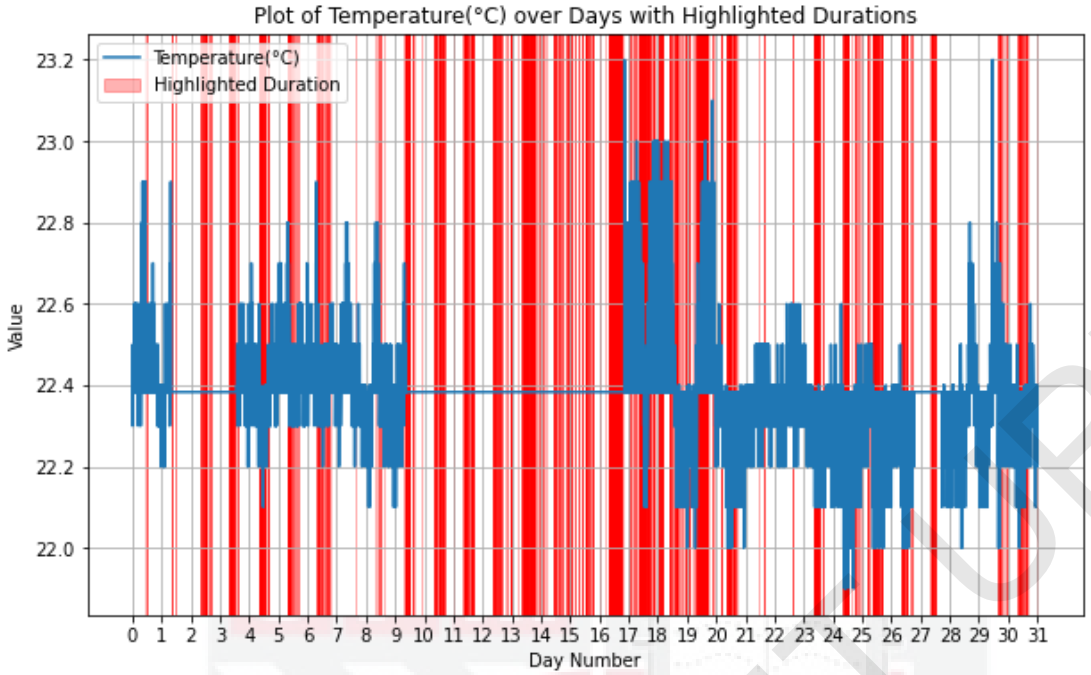
July

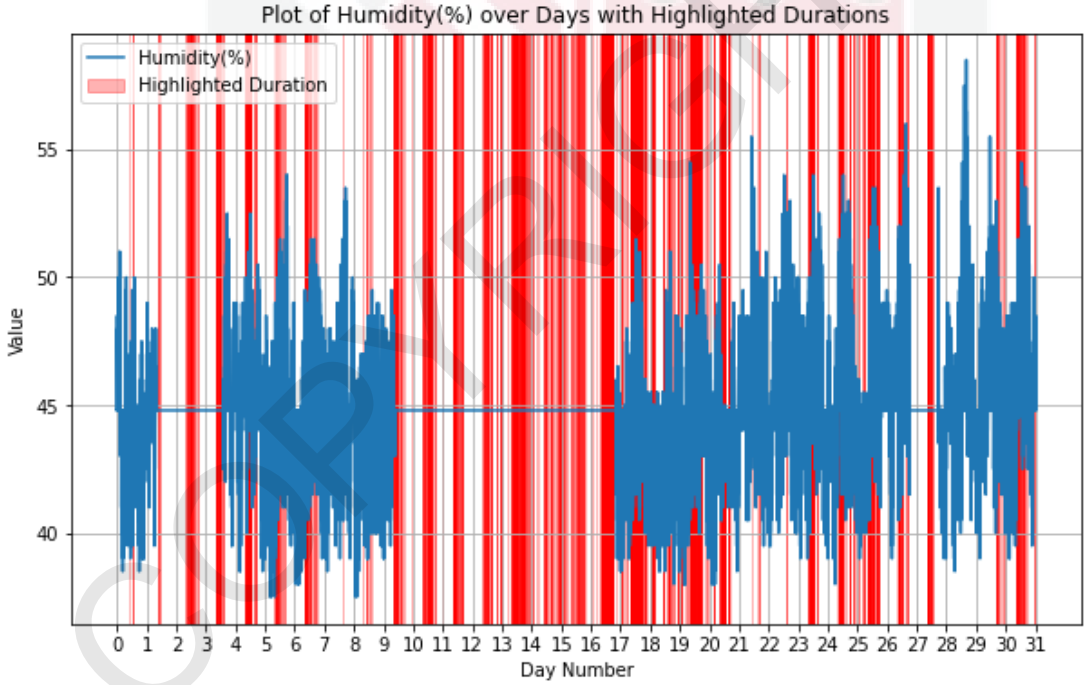





Plot of X(g) over Days with Highlighted Durations
X(g)
Highlighted Duration
Value
Day Number

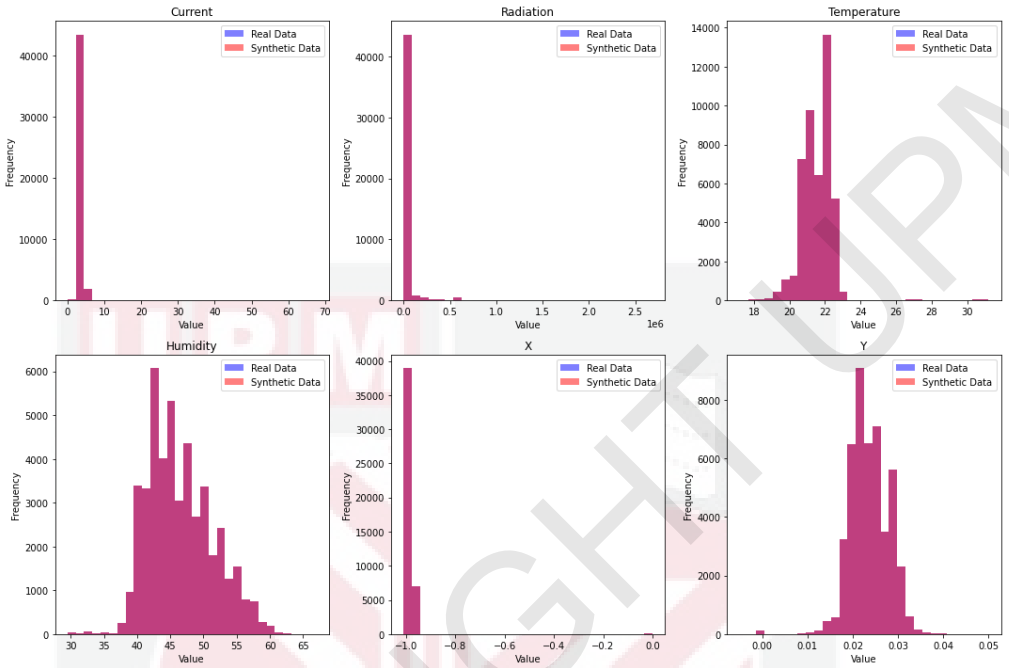

Plot of Y(g) over Days with Highlighted Durations
Y(g)
Highlighted Duration
Value
Day Number



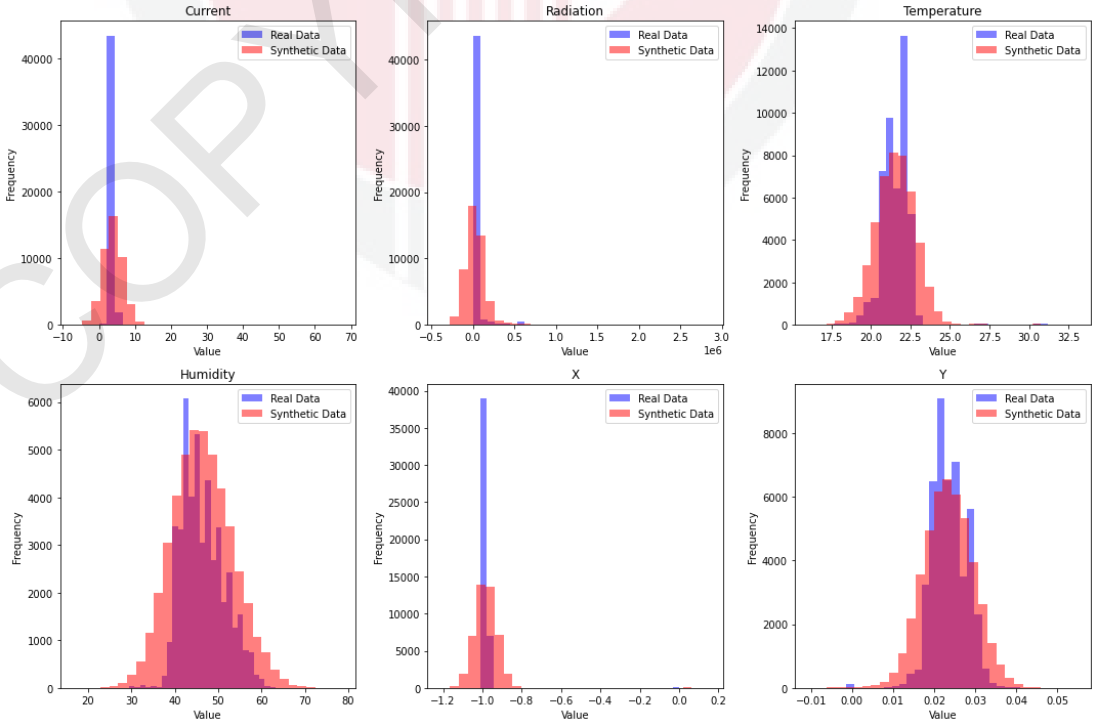


Appendix D

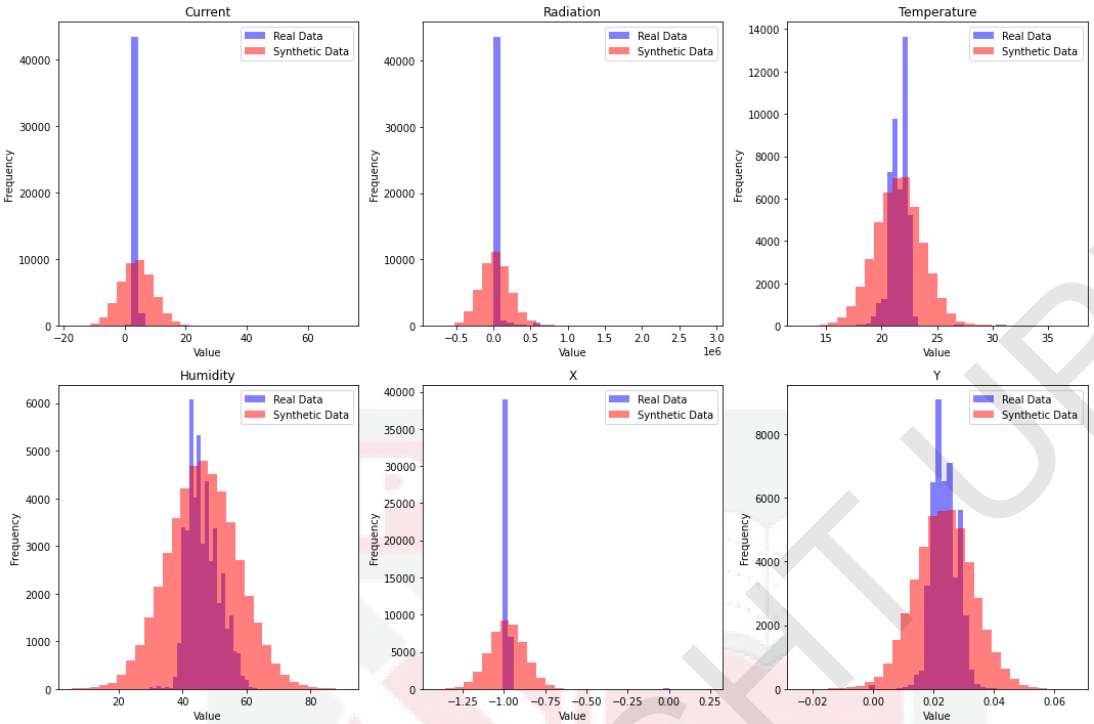
Noise Factor = 0



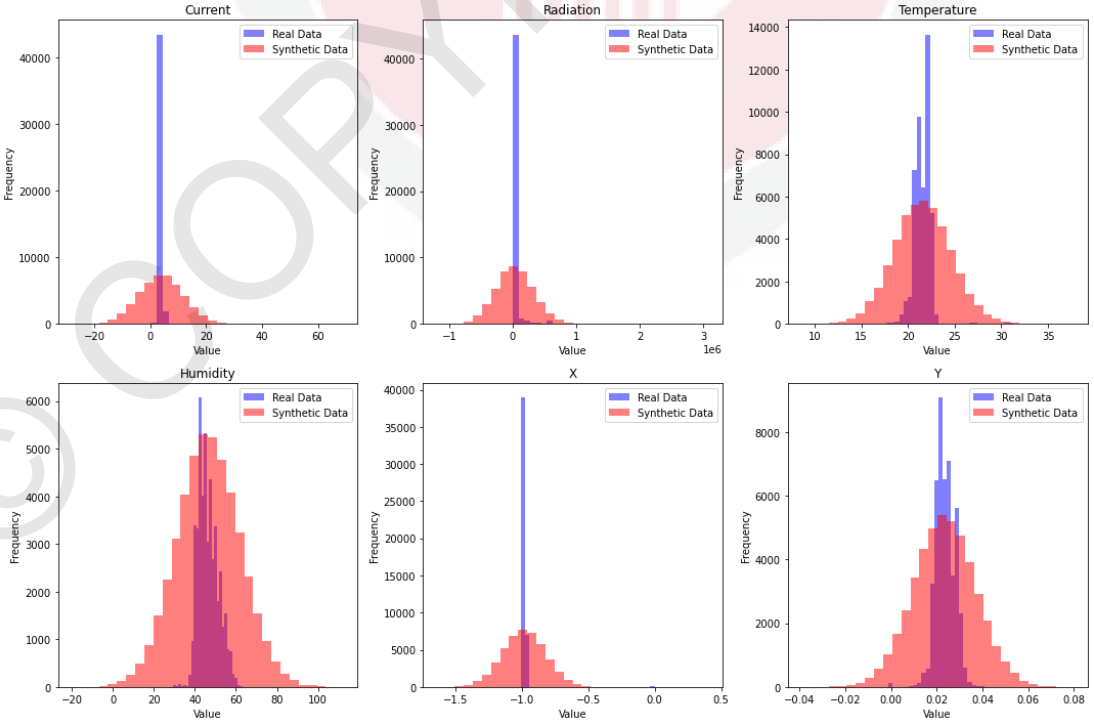
Noise Factor = 1



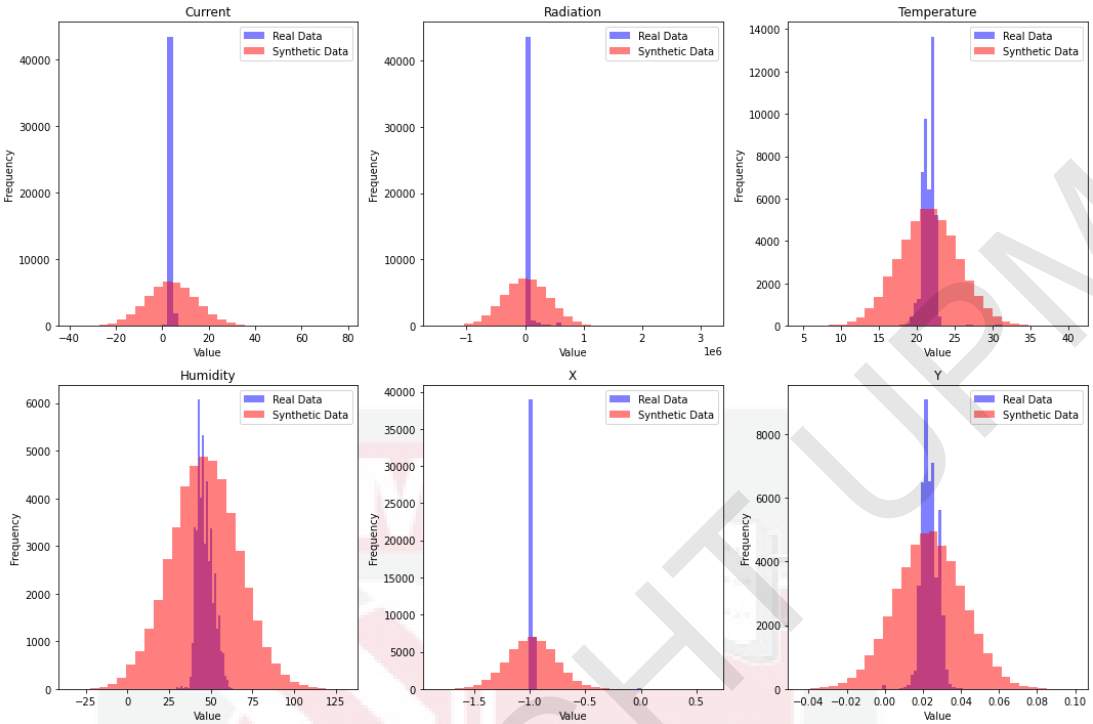
Noise Factor = 2



Noise Factor = 3



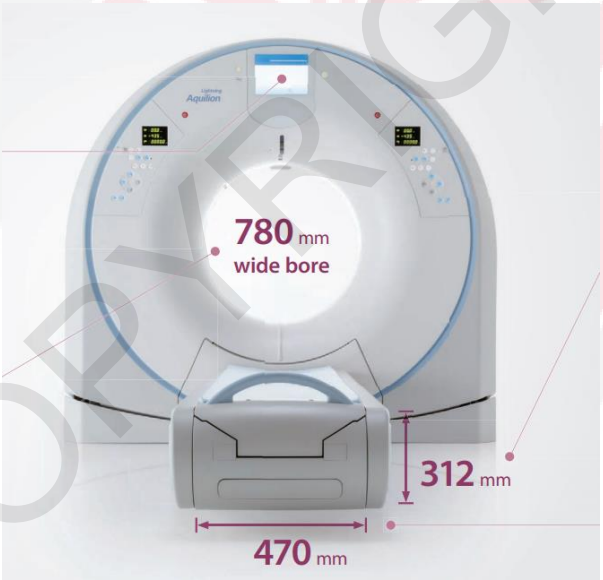
Noise Factor = 4



Appendix E

1. CT Scan Machine (The Aquilion Lightning TSX-035A)

    - A 16-row, 32-slice helical CT scanner from Canon Medical Systems.
    - Features:
        1. Dose reduction: Features like AIDR3D Enhanced and Exposure 3D reduce radiation dose.
        2. Subtraction: Removes bone and calcium from data sets.
        3. Navigation mode: Guides the operator through each step of the examination with computer graphics and animation.
        4. Intelligent filming: Automatically compiles images in a predefined layout.
        5. SEMAR: Reduces metal artifacts so that soft tissue structures can be clearly visualized.



2. Milesight AM103 (CO2 & Humidity & Temperature Sensor)

    - CO2 & Humidity & Temperature, Vivid Emoticon Indication, Traffic Light Status Indicator
    - 2 × 2700 mAh ER14505 Li-SOCl2 Replaceable Batteries (10 min interval).
    - Low Power Consumption & Smart Screen Mode

- AM103 detects the temperature beyond the range from 0°C to 40°C, the screen will turn off automatically.
- Dimensions: 68 × 65 × 20.5 mm (2.68 × 2.56 × 0.81 in)
- Data transmission range: Up to 2 km in urban areas and 15km in rural areas.
-

3. Schneider PM5110 (Energy meter)

- Energy meter with a 96mm-by-96mm backlit LCD display.
- The meter provides Class 0.5S accuracy per IEC 62053-22 standard and 64 samples per cycle.
- Measure Energy, Active and Reactive Power, Voltage, Current, Frequency, Power Factor and up to 15th Harmonic.
- The meter will function in a 50Hz or 60Hz network and accepts supply voltage ranging from 100 to 415 VAC and 125 to 250 VDC.
- Line rated current for this meter is 1A or 5A input and will support Single Phase and Neutral, Three Phase, or Three Phase and Neutral configurations.
- The range of measurement voltage between Phases is 35 to 690 VAC at 47 to 63 Hz.
- Measurement range between Phase and Neutral is 20 to 400VAC at 47 to 63 Hz.
- Communication protocol is Modbus RTU and ASCII 2 wires with RS485 port support. T
- The meter also has one digital pulse output and 33 configurable alarms.
- Product dimensions are as follows: width 3.78 inches (96 millimeters), depth 2.83 inches (72 millimeters), height 3.78 inches (96 millimeters) and product weight 13.4 ounces (380 g).
-

4. RAD-S101 (Nuclear radiation sensor)
    - Size with length 110*width 65*height 33mm.
    - Measuring range is 0.15 to 36000 μSv/h.
    - Measurement type are hard beta rays, gamma rays and X-rays.
    - Sensitivity about 450cpm/mR/h.

- Power supply is around 7-24VDC

5. MK926 MEMS (Accelerometer sensor)
    - Sensor can capture data along three axes: X, Y, and Z.
    - Vibration frequency and amplitude of the accelerometer to monitor the compaction speed, velocity, and excitation force.
    - Calculate the excitation force at that moment through amplitude and frequency.

BIODATA OF STUDENT

Mohammad Habib Shah Ershad bin Mohd Azrul Shazril, born in 1999 and originally from Kelantan, is the fourth of seven siblings. He graduated with a Bachelor's degree with Honors in Computer Communication System Engineering from Universiti Putra Malaysia (UPM) in 2022 and is currently pursuing a Master's degree at the same university. His research focuses on machine learning, synthetic data generation, and computer vision, with plans to further his studies by pursuing a PhD in the future.

PUBLICATIONS

M. H. S. E. B. M. A. Shazril, S. Mashohor, M. E. Amran, N. F. Hafiz, A. Ali, M. S. Naseri, and M. F. A. Rasid. (2024). Assessment of IoT-Driven Predictive Maintenance Strategies for Computed Tomography Equipment: A Machine Learning Approach. *IEEE Access*, vol. 12, pp. 195505-195515. doi: 10.1109/ACCESS.2024.3518516.

Conference

M. H. S. E. B. M. A. Shazril, S. Mashohor, M. E. Amran, N. F. Hafiz, A. A. Rahman, A. Ali, M. F. A. Rasid, A. S. A. Kamil, and N. F. Azilah. (2023). Predictive Maintenance Method using Machine Learning for IoT Connected Computed Tomography Scan Machine. *2023 IEEE 2nd National Biomedical Engineering Conference (NBEC)*, 42–47. https://doi.org/10.1109/NBEC58134.2023.10352591

N. F. Hafiz, S. Mashohor, **M. H. S. E. B. M. A. Shazril**, M. F. A. Rasid, and A. Ali. (2023). Comparison of Mel Frequency Cepstral Coefficient (MFCC) and Mel Spectrogram Techniques to Classify Industrial Machine Sound. *2023 15th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, Kuala Lumpur, Malaysia, pp. 273-278. doi: 10.1109/SKIMA59232.2023.10387339.