ISSN: 2634 - 8853

Journal of Engineering and Applied Sciences Technology

SCIENTIFIC Research and Community

Review Article

Open Access

Water Infrastructure Data Visualization Framework

Tanay Kulkarni[1]* and Devashri Karve[2]

[1]Water Infrastructure Professional, USA

[2]Asset Management Consultant, USA

ABSTRACT

The management of aging water infrastructure is a significant challenge for urban utilities, especially as the costs of repairs and replacements continue to escalate. This paper presents an innovative digital twin framework for Watertown City that integrates heterogeneous datasets-sensor readings, customer consumption data, and detailed pipe metadata-to offer a comprehensive visualization and predictive analytics platform. The application, implemented using Streamlit, provides interactive dashboards for water network exploration, usage analysis, pipe condition assessment, machine learning-based pipe-break prediction, and clustering-based network management. This paper details the data sources, preprocessing techniques, model development, visualization strategies, and experimental results. The paper also discusses insights gleaned from the analysis and outlines future enhancements to support real-time decision-making in water infrastructure management.

***Corresponding author**

Tanay Kulkarni, Water Infrastructure Professional, USA.

Received: January 06, 2024; **Accepted: January** 12, 2024; **Published:** January 24, 2024

Keywords: Data Visualization, Water Infrastructure, Digital Twins, Visualization Framework, IoT

Introduction

Water utilities worldwide are challenged by the continuous degradation of aging infrastructure, which often leads to inefficiencies such as non-revenue water (NRW), increased leakage incidents, and unscheduled maintenance. In response to these challenges, there is a growing trend toward leveraging digital twin technology—a virtual representation of physical assets—to support predictive maintenance and operational optimization.

The project addresses these challenges by creating an interactive web-based application that integrates multiple data streams. The application not only visualizes sensor data from tanks, valves, and pumps but also explores customer water usage and pipe metadata. The digital twin framework allows for both descriptive analytics (e.g., historical analysis of sensor readings) and predictive analytics (e.g., forecasting pipe breaks). This paper describes the architecture, methodology, and experimental results associated with the project, and demonstrates how such a system can aid water utilities in making data-driven decisions.

Related Work

Digital twin technology has rapidly gained prominence in the water sector, particularly as utilities seek to manage complex, distributed networks of aging assets. Prior work by Bentley Systems Inc. and Autodesk Inc. has underscored the importance of integrating hydraulic modeling, SCADA data, and GIS information into a unified framework for real-time monitoring and decision support. Applications such as Bentley OpenFlows WaterSight provide a robust platform for the operational management of water networks by integrating remote sensing data with hydraulic models.

Recent research in asset management has demonstrated that data-driven approaches can improve maintenance scheduling and reduce costs. Machine learning techniques, including decision tree regressors and clustering algorithms, have been applied to predict infrastructure failures and classify assets based on risk profiles. The Watertown City project builds upon these insights by combining state-of-the-art visualization methods with predictive analytics to form a digital twin that offers both operational insights and strategic guidance.

Data Sources and Preprocessing

Data Sources

The application leverages three primary datasets:

Sensor Data: This includes continuous readings from various sensors placed across the water network. Key parameters include:

- **Water Tank Levels:** Measurements of water volume in storage tanks.
- **Flow Readings:** Flow rates in pipes, valves, and pumps.
- **Pressure Readings:** Pressure measurements at critical points in the network.

Customer Consumption Records: Monthly water usage data from city residents are segmented by zone. This data is critical for understanding consumption patterns and detecting anomalies that might indicate leakages or system inefficiencies.

Pipe Metadata: Detailed information about the physical characteristics of the water network's pipes, such as:

- **Diameter and Length:** Key geometric attributes.
- **Material Type:** Typically Ductile Iron or Cast Iron.
- **Installation Year:** Which aids in assessing pipe aging.
- **Number of Reported Breaks:** Historical records of failures.
- **Bed-Soil pH, Discharge, and Pressure:** Environmental and operational factors affecting pipe longevity.

Data Preprocessing

Given the heterogeneity of the data, a comprehensive preprocessing pipeline was developed:

- **Filtering and Aggregation:** Sensor data is filtered by time range and aggregated using statistical metrics such as the mean and standard deviation to smooth out noise.
- **Unit Conversion:** The application supports dynamic switching between US Customary and SI units to enhance interpretability for different user groups.
- **Outlier Treatment:** For the pipe metadata, outlier values are "snapped" to bounds determined by the interquartile range (IQR). This helps mitigate the influence of extreme values when visualizing and modeling.
- **Normalization:** All features used for clustering and machine learning are scaled between 0 and 1. This normalization is essential for distance-based algorithms such as KMeans clustering.
- **Data Partitioning:** For predictive analytics, the pipe metadata is partitioned into training (75%) and testing (25%) sets. The training set is further split into training and validation subsets to optimize model parameters.

Methodology

Application Architecture

The digital twin application is implemented using Streamlit, a Python-based framework that supports rapid prototyping of interactive web applications. The application is structured into five primary tabs:

- **Water Network Exploration:** Visualizes raw and filtered sensor data over time.
- **Water Usage Exploration:** Displays customer water usage trends using line charts, bar graphs, and pie charts.
- **Pipe Data Exploration:** Allows interactive exploration of pipe metadata, including correlations between different features.
- **Pipe-Break Prediction:** Utilizes a Decision Tree Regressor to predict future pipe breaks based on user-selected features.
- **Water Network Management:** Applies KMeans clustering to classify pipes into distinct risk groups for targeted maintenance.

Visualization Strategies

Interactive visualizations are central to the application's design. The following approaches were implemented:

Time Series and Distribution Charts: For sensor data, time series plots are combined with histograms to provide insights into both temporal trends and data distribution.

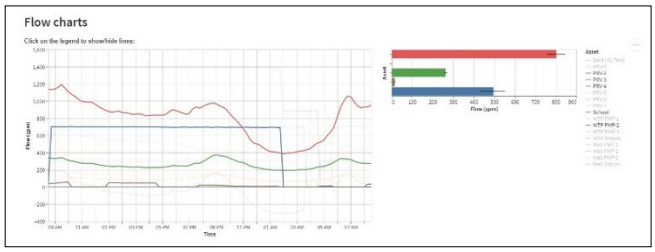

Figure 1: Sensor Data Time Series Visualization

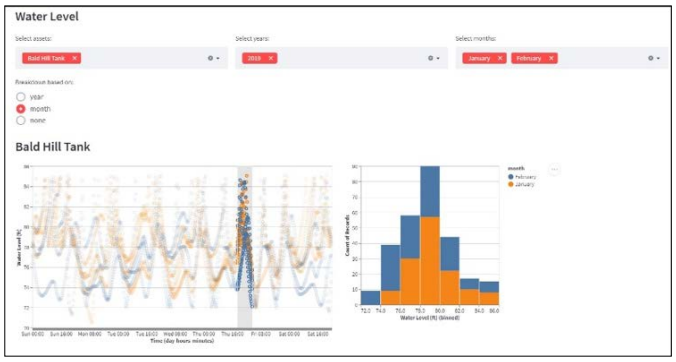

Figure 2: Bar Chart Showing Average Sensor Readings

Zone-Based Analysis: Customer water usage is visualized using color-coded charts that segment data by city zones. This multi-faceted approach enables the identification of anomalies that may indicate leakage or demand spikes.

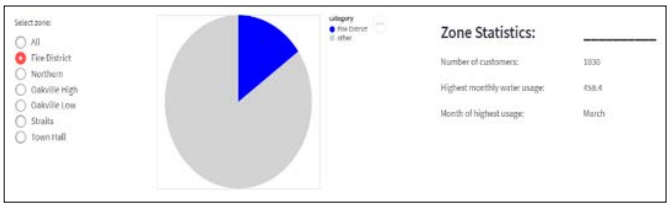

Figure 3: Monthly Water Usage Visualization by Zone

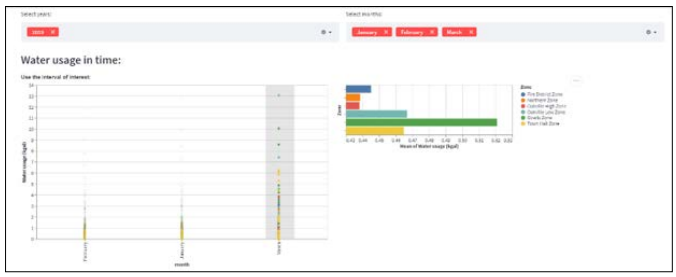

Figure 4: Scatter Plot of Customer Water Usage

Interactive Sidebar and Data Tables: For pipe data exploration, an interactive sidebar allows users to switch between units and select specific pipe attributes for in-depth analysis.

Figure 5: Pipe Data Exploration Settings Sidebar

Machine Learning: Pipe-Break Prediction

The predictive model uses a Decision Tree Regressor to forecast the number of pipe breaks. The key aspects include:

- **Feature Selection:** Users can interactively select from a range of predictor variables including pipe diameter, length, bed-soil pH, number of customers, discharge, pressure, and pipe age.
- **Model Training:** The data is split into training (75%) and

testing (25%) datasets, with further subdivision for validation. This hierarchical split ensures that the model's performance is robustly evaluated.

- **Performance Metrics:** The model achieves a prediction accuracy of approximately 68.5% on the testing dataset when all features are included. Additional metrics such as precision, recall, and f1-score are computed to gauge model performance.
- **Dynamic Prediction:** The application features an input interface for entering custom values. This "what-if" analysis tool allows users to simulate changes in pipe conditions and immediately observe the predicted number of breaks.

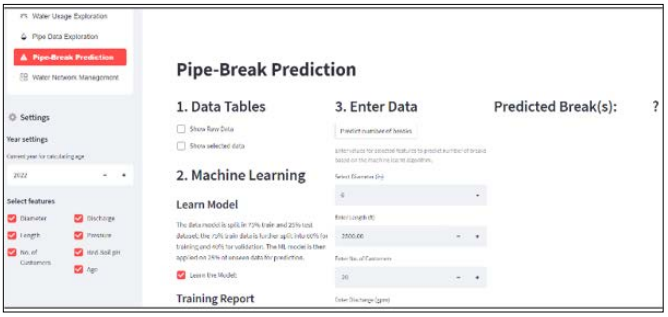

Figure 6: Pipe-Break Prediction Settings Sidebar

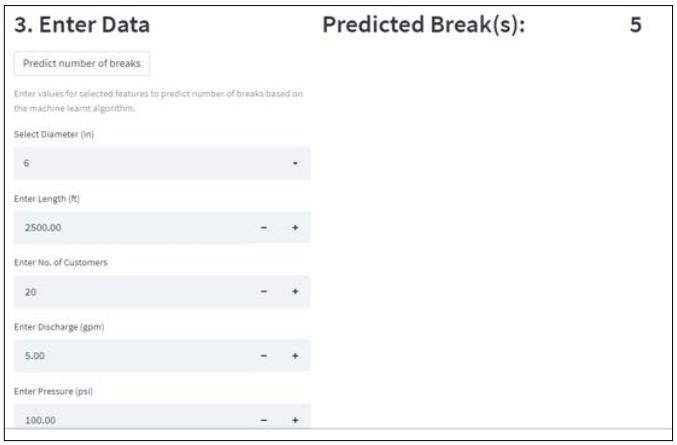

Figure 7: Dynamic Prediction of Pipe Breaks

Clustering

KMeans clustering is applied to classify pipes into distinct groups based on their attributes. The process involves:

- **Preprocessing:** Data scaling and outlier correction are performed prior to clustering. This step ensures that all features contribute equally to the distance calculations used by the KMeans algorithm.
- **Determination of Cluster Count:** The Elbow Method is used to identify the optimal number of clusters, which is determined to be three for the current dataset.
- **Interpretation:** The clustering results are visualized using box plots that compare key pipe attributes across clusters. One particular cluster (denoted in orange) is characterized by longer pipes, higher break counts, and lower bed-soil pH values, which are indicative of higher risk.

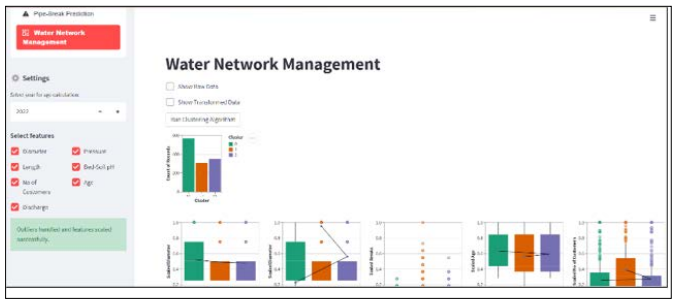

Figure 8: Clustering Settings Sidebar for Water Network Management

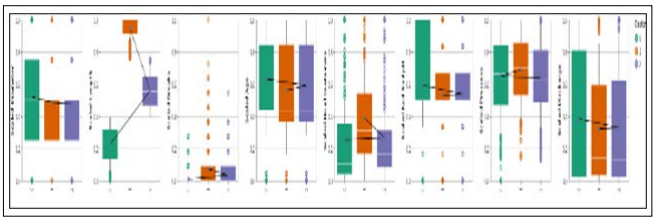

Figure 9: KMeans Clustering Results

Experimental Results and Discussion

Pipe Data Analysis

Exploratory analysis of the pipe metadata revealed several important trends. Visualizations comparing pipe diameter, installation year, and the number of breaks indicated that the majority of the pipes in the network are Ductile Iron and Cast Iron types installed over 70 years ago. This observation is critical as it suggests that a significant portion of the network is nearing or has exceeded its theoretical life expectancy.

For instance, the chart displaying "Diameter vs. Year of Installation" clearly shows that many of the older pipes fall within the age bracket where deterioration is most pronounced.

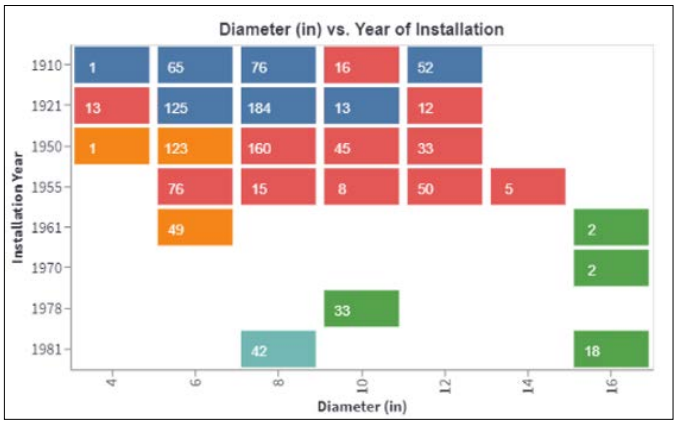

Figure 10: Diameter vs. Year of Installation

Additionally, the "Year of Installation vs. Sum of Pipe Breaks" visualization indicates that pipes installed prior to 1955 account for the majority of reported failures. This finding aligns with engineering expectations regarding material degradation over time.

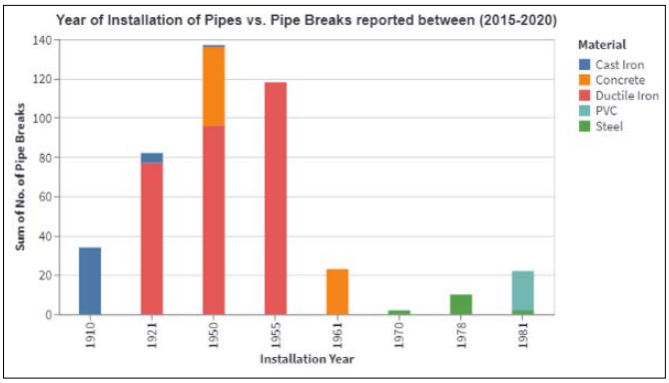

Figure 11: Installation Year vs. Sum of Pipe Breaks

Impact of Operational and Environmental Factors

The analysis further extends to operational parameters such as discharge and pressure. The "Discharge vs. Installation Year" chart reveals that pipes installed before 1955 not only exhibit a higher frequency of breaks but also carry larger volumes of discharge. This implies that a failure in these older pipes could have a widespread impact on network reliability and service continuity.

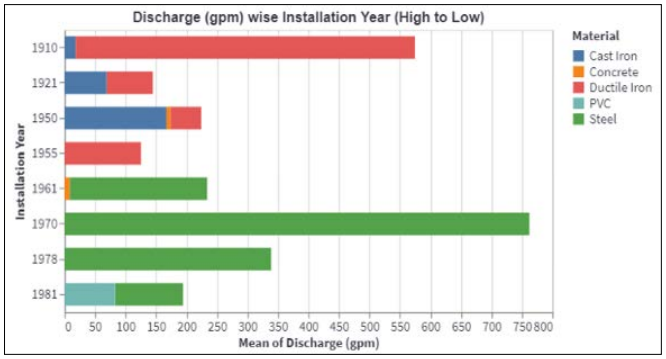

Figure 12: Discharge vs. Installation Year

Environmental factors are also considered in the analysis. The "Bed Soil pH vs. Sum of Pipe Breaks" visualization demonstrates that pipes buried in soils with a pH below 6 (indicative of high acidity) are more susceptible to corrosion and failure. This correlation is critical for maintenance planning, as it highlights the need for targeted inspections in areas with adverse soil conditions.

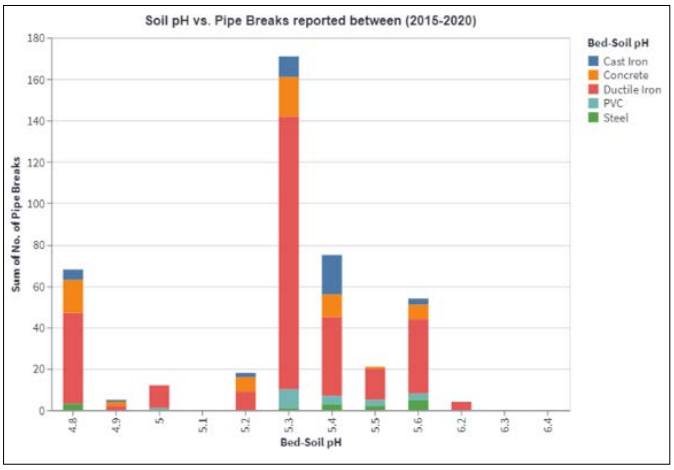

Figure 13: Bed Soil pH vs. Sum of Pipe Breaks

Machine Learning Model Performance

The Decision Tree Regressor used for predicting pipe breaks demonstrated a testing accuracy of approximately 68.5% when all predictor variables are included. This moderate level of performance suggests that while the model captures key trends in the data, there remains scope for improvement—potentially through the incorporation of additional features or the exploration of alternative modeling approaches such as Random Forests.

The interactive nature of the predictive module allows utility managers to experiment with different combinations of features and observe corresponding changes in prediction performance. This "what-if" analysis is invaluable for understanding the sensitivity of the model to various input parameters and for identifying the most influential factors driving pipe breakage.

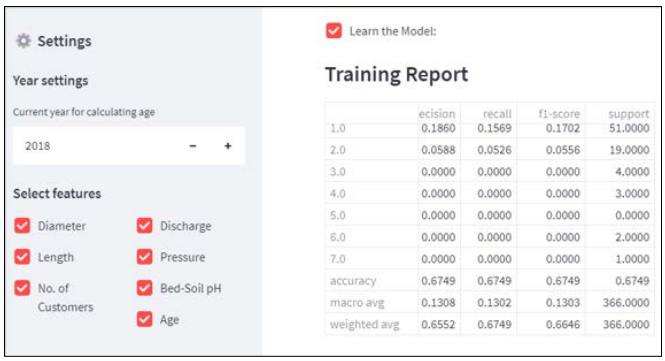

Figure 14: ML Model Building Interface

Clustering and Network Management

KMeans clustering of the pipe metadata yielded three distinct clusters. The optimal number of clusters was determined using the Elbow Method, which indicated diminishing returns beyond three clusters. The clusters were analyzed based on key features such as pipe length, number of breaks, discharge, and environmental factors like soil pH. One cluster, characterized by longer pipes with a high incidence of breaks and low soil pH, was identified as being at particularly high risk. This insight is critical for prioritizing inspections and scheduling proactive maintenance.

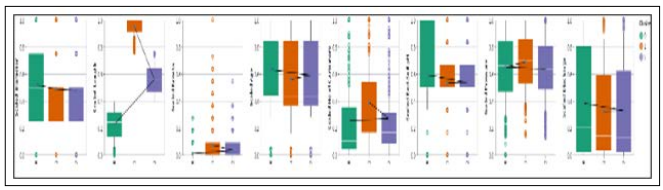

Figure 15: Clustering Visualization

Discussion

Integration of Data and Analytics

The digital twin approach exemplified in this project successfully integrates diverse data streams into a cohesive framework for water infrastructure management. By combining sensor data, customer consumption records, and pipe metadata, the application provides a holistic view of network performance. The interactive dashboards and visualizations empower stakeholders to identify trends, detect anomalies, and make informed decisions regarding maintenance and asset replacement.

Practical Implications for Utilities

The findings of this study have several practical implications:

- **Preventive Maintenance:** The predictive model can serve as an early warning system, flagging pipes that are likely to fail.

This enables utilities to prioritize repairs before catastrophic failures occur.

- **Resource Allocation:** The clustering analysis aids in segmenting the water network into risk categories. High-risk clusters can be targeted for immediate inspection and repair, optimizing the allocation of limited resources.
- **Operational Efficiency:** The integration of real-time sensor data with historical records offers a dynamic view of network performance. This supports timely interventions and adaptive operational strategies.

Limitations and Future Enhancements

While the current framework demonstrates significant promise, several limitations must be acknowledged:

- **Model Accuracy:** The pipe-break prediction model, while effective, has an accuracy of 68.5%. Future work may involve expanding the dataset, incorporating additional features (e.g., real-time environmental factors), and testing alternative algorithms such as ensemble methods.
- **Real-Time Data Integration:** Currently, the system is based on historical data snapshots. Integrating real-time data streams from SCADA systems would greatly enhance the operational utility of the digital twin.
- **Scalability:** The framework, while robust for Watertown City, needs to be evaluated for scalability across larger and more complex water networks.

Future Work

Several avenues exist for enhancing the digital twin framework:

- **Real-Time Monitoring and SCADA Integration:** Integrating live data feeds from SCADA systems will enable continuous monitoring and dynamic response. This real-time integration would further bridge the gap between predictive analytics and operational decision-making.
- **Enhanced Predictive Modeling:** Exploring alternative machine learning algorithms, such as Random Forests or Gradient Boosting Machines, may improve prediction accuracy. Additionally, incorporating temporal models (e.g., time-series forecasting) could capture evolving network conditions more effectively.
- **Expanded Data Sources:** Future iterations could integrate additional data sources such as customer billing records, weather data, and geographic information systems (GIS). This would allow for a more comprehensive analysis of water demand, environmental stressors, and asset performance.
- **User Interface Enhancements:** Although the current Streamlit-based interface is functional, further improvements in usability and visualization—such as incorporating 3D network models and augmented reality (AR) overlays—could enhance stakeholder engagement and decision support.
- **Predictive Maintenance Strategies:** Integrating the insights from the predictive and clustering modules into a broader asset management framework can facilitate proactive maintenance scheduling. Decision-support algorithms could be developed to automatically recommend inspection and repair schedules based on real-time risk assessments.

Conclusion

This paper has presented a detailed technical overview of the digital twin framework designed to support the proactive management of water infrastructure assets. Through the integration of sensor data, customer consumption records, and detailed pipe metadata, the application delivers a multifaceted analytical platform that enables both descriptive and predictive insights.

Key contributions include:

- The development of an interactive web-based application that enables real-time visualization and filtering of sensor and consumption data.
- The implementation of a machine learning model (Decision Tree Regressor) for predicting pipe breaks, along with a dynamic interface for "what-if" analysis.
- The application of KMeans clustering to segment the network into risk-based categories, supporting targeted maintenance and resource allocation.
- A comprehensive data preprocessing pipeline that addresses unit conversion, outlier treatment, and normalization—ensuring robust analysis and visualization.

The experimental results underscore the importance of integrating diverse datasets and leveraging modern analytics to manage aging water infrastructure. While the current model provides a strong foundation, further work in real-time integration, enhanced predictive modeling, and expanded data sources is necessary to fully realize the potential of a digital twin for water networks [1,2].

The insights derived from this project are directly applicable to the strategic planning and maintenance scheduling challenges faced by urban water utilities. By harnessing the power of interactive visualization and machine learning, the digital twin framework not only facilitates operational efficiency but also lays the groundwork for a more resilient and sustainable water infrastructure.

References

1. Bentley Systems Inc. OpenFlows WaterSight. Bentley https://www.bentley.com/en/products/product-line/hydraulics-and-hydrology-software/openflows-watersight.
2. CMU-IDS-2022 Water Infrastructure Data Visualization for Watertown City. GitHub repository https://github.com/CMU-IDS-2022/final-project-digitaltwinfrastructure.