REAL-TIME OPTIMIZATION AND MAINTENANCE OF WIND TURBINE PERFORMANCE USING DIGITAL TWIN TECHNOLOGY

Dilmini N.A.C

(IT21836954)

(The dissertation was submitted in partial fulfilment of the requirements for the B.Sc. (Honors) degree in Information Technology Specializing in Software Engineering)

Department of Computer Science and Software Engineering

Sri Lanka Institute of Information Technology

August 2025

REAL-TIME OPTIMIZATION AND MAINTENANCE OF WIND TURBINE PERFORMANCE USING DIGITAL TWIN TECHNOLOGY

Dilmini N.A.C

(IT21836954)

(The dissertation was submitted in partial fulfilment of the requirements for the B.Sc. (Honors) degree in Information Technology Specializing in Software Engineering)

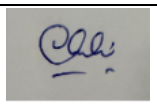
Department of Computer Science and Software Engineering

Sri Lanka Institute of Information Technology

August 2025

DECLARATION

I declare that this is my own work, and this dissertation does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. Also, I hereby grant to Sri Lanka Institute of Information Technology the non-exclusive right to reproduce and distribute my dissertation in whole or part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as article or books).

Student Name	Student ID	Signature
Dilmini N.A.C	IT21836954	

As the supervisor/s of the above-mentioned candidate, I hereby certify that they are conducting research for their undergraduate dissertation under my guidance and direction.

Signature of the supervisor: Date: 29/08/2025

(Mr.Vishan Jayasinghearachchi)

ABSTRACT

Wind energy is a critical contributor to Sri Lanka's renewable transition, with the Thambapavani Wind Farm serving as the flagship initiative. However, its coastal setting exposes turbines to operational risks caused by yaw misalignment, wind speed threshold exceedances, and seasonal lightning activity, all of which reduce generation reliability and complicate short-term forecasting for the national grid. This project addresses these challenges by designing a weather impact analysis framework integrated within a digital twin environment to provide real-time, turbine-level forecasting and decision support.

The system incorporates three predictive modules: (1) an XGBoost regression model for yaw misalignment loss prediction, enhanced with turbine-specific wind direction correction factors derived from NASA POWER datasets; (2) a Random Forest classifier for lightning risk forecasting, optimized to achieve high recall in six-hour risk windows; and (3) a deterministic threshold-based module for quantifying cut-in and cut-out wind speed losses. Together, these models enable proactive forecasting of weather-induced risks that traditional rule-based approaches fail to anticipate.

A FastAPI-based prediction service was developed to unify these models, exposing endpoints for real-time predictions, probabilistic risk assessments, and historical queries. The outputs were integrated into a digital twin and operational dashboard built with React and Three.js, providing operators with both numerical forecasts and intuitive 3D visualization of turbine states. Evaluation demonstrated strong performance: the yaw misalignment model achieved $R^2 = 0.80$ at the hourly level and 0.97 at the daily level; the lightning model retained 95% detection capability with reduced feature sets; and threshold modules consistently flagged shutdown conditions at operational boundaries.

By combining predictive analytics, turbine-level corrections, and digital twin visualization, this research advances beyond static and reactive monitoring systems. It provides operators with accurate, real-time insights into weather-driven losses, strengthens short-term forecast reliability for the Ceylon Electricity Board, and establishes a scalable framework for intelligent wind farm management in weather-sensitive environments.

ACKNOWLEDGEMENT

I would like to extend my sincere gratitude to all those who supported and guided me throughout the successful completion of this research project.

First and foremost, I am deeply thankful to my supervisor, Mr. Vishan Jayasinghearachchi, for his invaluable guidance, constructive feedback, and continuous encouragement. His expertise and direction were instrumental in shaping the scope of this project and ensuring its successful completion.

I would also like to express my heartfelt appreciation to my co-supervisor, Mr. Jeewaka Perera, whose expertise in machine learning provided critical insights in developing and refining the predictive models used in this research. His mentorship greatly enhanced the technical depth and quality of this work.

My sincere gratitude goes to Eng. D.M.R.K.B Gunarathna, Deputy General Manager, Mahaweli Complex CEB, for granting permission to visit the Thambapavani Wind Farm, which provided invaluable exposure to real-world operations. I am also especially grateful to Eng. Sashinath and Eng. Janaka at Thambapavani, who generously shared their time and expertise, guiding me through the operational aspects of the wind farm and providing access to the SCADA datasets required for this study. Their practical insights and cooperation were crucial in aligning the research with real-world industry practices.

I would also like to thank the CDAP team and the Faculty of Computing, Sri Lanka Institute of Information Technology (SLIIT) for their technical assistance, infrastructural support, and for providing the necessary tools and resources that were vital for conducting this research successfully.

Finally, I am sincerely grateful to everyone who offered their encouragement, thoughtful feedback, and moral support throughout this journey. Whether through technical input, academic assistance, or personal motivation, your contributions have been truly appreciated and deeply valued.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Full Form
SCADA	Supervisory Control and Data Acquisition
API	Application Programming Interface
MQTT	Message Queuing Telemetry Transport
ML	Machine Learning
RUL	Remaining Useful Life
R^2	Coefficient of Determination
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
AUC	Area Under the Curve
CEB	Ceylon Electricity Board
DB	Database (SQL / Time-series storage)
DT	Digital Twin

1. INTRODUCTION

1.1 Background and Literature Survey

Wind energy has emerged as one of the most promising solutions in the global transition towards renewable power, offering a clean, sustainable, and increasingly cost-effective alternative to fossil fuels. For Sri Lanka, this transition is particularly critical, as the country's growing electricity demand has long depended on imported thermal energy sources. The Thambapavani Wind Farm in Mannar, comprising 33 grid-connected turbines along the north-western coastline, represents the nation's flagship renewable energy initiative. However, while its coastal location provides strong and consistent winds, it also exposes the farm to unique operational challenges. Sudden changes in wind direction can cause turbines to halt operations temporarily while adjusting their yaw position, resulting in measurable power losses. Similarly, extreme wind conditions that fall below the cut-in speed or exceed the cut-out limit force shutdowns to safeguard the turbines. Seasonal lightning activity in the Mannar region further adds to operational risk, with the potential to damage equipment and disrupt grid stability. Collectively, these weather-related issues reduce the farm's energy yield, increase downtime, and complicate the reliability of national grid forecasts.

Weather variability thus remains one of the most significant factors affecting wind turbine performance, directly influencing both energy generation and the accuracy of operational forecasts. Yaw misalignment, where the turbine nacelle cannot immediately adapt to abrupt shifts in wind direction, leads to unavoidable interruptions in power output as turbines must realign before resuming operation. Cut-in and cut-out thresholds, typically around 3 m/s and 25 m/s respectively, while essential for operational safety, further limit energy capture when wind conditions cross these boundaries. Lightning strikes, although less frequent, pose a highly damaging risk to turbine blades, electrical

systems, and control equipment, with severe consequences for both generation continuity and worker safety. For utilities such as the Ceylon Electricity Board (CEB), which rely on short-term power generation forecasts (often 48 hours in advance) to schedule grid operations, such weather-driven uncertainties create a mismatch between projected and actual output, undermining reliability and complicating national energy planning. These issues underscore the need for predictive and corrective mechanisms capable of anticipating risks before they materialize, rather than relying solely on reactive measures.

A wide body of research has examined the influence of weather conditions on turbine performance, often concentrating on individual aspects such as yaw misalignment, extreme wind speeds, or lightning hazards. Bandi and Apt [1], for example, analyzed the variability of wind turbine power curves and demonstrated that manufacturer-supplied calibration models frequently fail to capture the effects of local environmental factors. Their findings showed that ignoring wind speed variability leads to substantial forecast uncertainty and recommended post-installation recalibration as a corrective measure. Hasan and Styve [2] highlighted the role of digital twin technologies in renewable energy systems, noting how real-time virtual replicas can enhance predictive maintenance, monitoring, and energy forecasting. While their contribution demonstrates the growing relevance of digitalization in wind energy, it primarily emphasized generalized optimization rather than addressing weather-specific risks. In parallel, Mostajabi et al. [3] developed machine learning models for lightning nowcasting, using surface meteorological parameters such as air pressure, humidity, temperature, and wind speed. Their work achieved promising accuracy for lead times up to 30 minutes, yet remained limited to very short-term warnings, insufficient for operational planning in large wind farms. Complementing these studies, Gu et al. [4] applied machine learning techniques to forecast wind power at a coastal wind farm in Zhejiang, China, demonstrating that Random Forest models significantly outperformed linear models in

capturing nonlinear dependencies between wind velocity and power output. Although these results reinforce the value of machine learning for improving forecast accuracy, they primarily addressed system-level predictions rather than integrating multiple weather-related risks into a turbine-level framework.

Recent research trends have attempted to bridge these gaps by combining multi-source data with advanced machine learning methods. SCADA data from operational turbines, when integrated with satellite-based reanalysis and meteorological datasets, have enabled site-specific modeling and improved short-term power predictions. Hybrid approaches, blending physical simulation with statistical learning, have shown particular effectiveness in turbulent or complex environments, illustrating how domain knowledge and data-driven methods can complement one another. At the same time, the adoption of digital twins has expanded beyond simple forecasting towards holistic wind farm management, offering platforms that integrate predictive analytics, control strategies, and visualization. By continuously synchronizing real-world turbine data with virtual models, digital twins allow operators to simulate weather impacts, anticipate failures, and optimize parameters such as pitch, yaw, and rotor speed. Despite these advances, however, much of the literature remains focused on generalized forecasting accuracy or operational optimization, with critical weather-induced risks—such as yaw misalignment losses, extreme wind speed shutdowns, and lightning hazards—still treated in isolation.

Although prior work provides a strong foundation, several limitations remain. Most forecasting models assume uniform turbine behavior, overlooking the fact that local environmental factors and turbine-specific deviations can significantly influence outcomes. Manufacturer power curves, when applied without adjustment, are unable to capture such site-specific effects, introducing uncertainty into predictions. Digital twin applications, while valuable for monitoring and predictive maintenance, have not been widely extended to model weather-induced risks that directly reduce energy yield.

Lightning prediction studies, though showing the potential of machine learning, remain constrained to short-term nowcasting windows and do not align with the medium-range forecasts required for operational planning. Finally, machine learning models applied to wind forecasting generally emphasize aggregate system-level predictions and seldom integrate multiple weather-related risk factors into a unified framework.

This research aims to address these limitations by developing an integrated weather impact analysis framework for the Thambapavani Wind Farm. The proposed approach introduces turbine-specific correction factors using NASA POWER data to adjust for local deviations in wind direction, thereby improving misalignment loss prediction. It also extends beyond generalized digital twin applications by incorporating weather-driven risks—yaw misalignment losses, threshold exceedances, and lightning hazards—directly into the predictive pipeline, enabling operators to anticipate short-term variations in power availability. The lightning risk module improves upon prior short-term approaches by offering six-hour risk windows, calibrated to meet the operational needs of utilities such as the CEB. Most importantly, the framework integrates multiple weather-related risks into a unified, turbine-level predictive model, which is synchronized with a digital twin interface for real-time visualization and decision support. By combining SCADA data, reanalysis products, and advanced machine learning within an operationally focused digital twin, this work provides a novel pathway to improve both the reliability of 48-hour forecasts and the efficiency of wind farm operations in weather-sensitive environments.

1.2 Research Gap

Over the past decade, considerable research has been devoted to improving wind turbine forecasting and operational reliability through both physical models and machine learning approaches. While these contributions have advanced the field, they continue to

face limitations when applied to weather-induced risks that directly impact turbine availability and energy yield.

Bandi and Apt [1] investigated the variability of wind turbine power curves and highlighted how manufacturer-supplied calibration models often fail to account for local environmental conditions. Their study proposed post-installation recalibration to reduce uncertainty. While valuable, this approach focused primarily on generalized power curve accuracy and did not extend to predicting dynamic losses such as yaw misalignment events or turbine-specific deviations across a wind farm.

Hasan and Styve [2] explored the role of digital twin technology in renewable energy systems, demonstrating its benefits for predictive maintenance, monitoring, and energy forecasting. However, their work emphasized broad operational optimization rather than addressing weather-driven losses, leaving risks such as extreme wind thresholds and lightning hazards underexplored in digital twin applications.

Mostajabi et al. [3] developed machine learning models to nowcast lightning occurrence using surface meteorological parameters and achieved skillful predictions for short lead times. Yet, this research was restricted to very short-range warnings (0–30 minutes) and was not designed to support the medium-range, turbine-level forecasts required for wind farm operational planning and grid scheduling.

Gu et al. [4] applied Random Forest models to forecast wind power in a coastal wind farm in Zhejiang, demonstrating that machine learning can outperform linear and physical models in predicting nonlinear relationships between wind speed and power. While their work reduced uncertainty in system-level forecasting, it did not integrate multiple factors—such as yaw misalignment, cut-in/cut-out losses, and lightning—into a unified predictive framework.

Taken together, these studies highlight a clear research gap. Existing models tend to focus either on generalized forecasting accuracy or on single-risk scenarios, without

providing an integrated framework capable of anticipating turbine-specific losses caused by multiple weather factors. Furthermore, current lightning prediction methods remain short-term, power curve models lack local correction mechanisms, and digital twin applications have not been extended to capture weather-induced risks in real time.

This research addresses these shortcomings by proposing an integrated weather impact analysis framework for the Thambapavani Wind Farm. The system introduces turbine-specific wind direction correction factors derived from NASA POWER data to enhance misalignment loss prediction. It extends lightning prediction into six-hour operational risk windows, calibrated for utility planning. It also incorporates threshold-based shutdown modeling for cut-in and cut-out conditions. Most importantly, it combines these weather-driven risks into a unified predictive model that integrates seamlessly with a digital twin interface for real-time visualization and decision support. In doing so, this work bridges the gap between isolated forecasting methods and the operational needs of a weather-sensitive coastal wind farm.

Feature	**Bandi & Apt [1]**	**Hasan & Styve [2]**	**Mostajabi et al. [3]**	**Gu et al. [4]**	**This Proposed System**
Local/Turbine-Specific Corrections	✗	✗	✗	✗	✓
Yaw Misalignment Loss Prediction	✗	✗	✗	✗	✓
Threshold Shutdown Modeling	✗	✗	✗	✗	✓

Lightning Risk Forecasting (6 hr+)	✗	✗	✓	✗	✓
Digital Twin Integration	✗	✓	✗	✗	✓

Table 1.1 Comparison of Research Gap

1.3 Research Problem

Wind energy plays a pivotal role in the global shift towards renewable power, and for Sri Lanka it has become increasingly important to secure sustainable electricity generation while reducing reliance on imported fuels. The Thambapavani Wind Farm in Mannar, consisting of 33 grid-connected turbines, is a cornerstone of this effort. However, despite the scale of this project and the availability of forecasting tools, the farm's coastal location makes it highly vulnerable to weather-induced disruptions. These include yaw misalignment events caused by abrupt wind direction changes, forced shutdowns due to wind speeds falling below cut-in or exceeding cut-out thresholds, and seasonal lightning activity that threatens both equipment safety and generation continuity. Together, these risks undermine energy yield and complicate the delivery of reliable power forecasts to the national grid.

Current industry practices rely heavily on threshold-based control systems, generic power curve models, and short-term detection tools. Yaw misalignment is often managed through fixed rules such as the 10° repositioning threshold, which is reactive and introduces temporary downtime. Cut-in and cut-out conditions are addressed through deterministic physics-based limits, ensuring safety but providing no predictive insight into when losses may occur. Lightning management is largely restricted to detection systems or external warnings, which are effective only once hazardous

conditions have already emerged. Collectively, these approaches operate in a reactive mode, failing to anticipate risks or integrate turbine-specific behaviors into forecasting pipelines.

This gap is particularly critical for utilities such as the Ceylon Electricity Board (CEB), which depends on accurate 48-hour generation forecasts to manage grid operations. When weather-induced risks are not adequately captured, forecasts often diverge from actual output, reducing the reliability of planning and increasing costs of balancing the grid. A turbine may, for example, halt operation due to a sudden lightning strike or yaw misalignment, yet such events are not reflected in the predicted generation profile. As a result, decision-makers are left with incomplete information and reduced confidence in renewable energy scheduling.

The research problem addressed in this study is therefore defined as follows:

“How can weather-induced risks such as yaw misalignment losses, threshold exceedances, and lightning hazards be modeled and integrated into a unified predictive framework that improves short-term power forecasting accuracy and operational reliability for wind farms like Thambapavani?”

To address this problem, the proposed research introduces a framework that combines turbine-specific correction factors, machine learning models, and mathematical threshold analysis to quantify and predict weather-driven losses. These predictive outputs are synchronized with a digital twin interface, enabling operators to visualize real-time conditions and anticipate risks proactively. By integrating multiple weather impact factors into a single predictive system, the solution moves beyond static, reactive models towards an adaptive framework that enhances the accuracy of forecasts and supports informed decision-making for grid stability and wind farm management.

1.4 Objectives

1.4.1 Main Objective

The main objective of this research is to develop an integrated weather impact analysis framework that enhances the accuracy of short-term power forecasting and operational decision-making for the Thambapavani Wind Farm. The proposed system aims to quantify and predict turbine-level losses caused by yaw misalignment, wind speed threshold exceedances, and lightning hazards, and to integrate these predictions into a digital twin environment for real-time visualization and operator support.

Traditional wind turbine forecasting approaches rely heavily on static threshold rules, generic manufacturer power curves, or short-term detection tools, which often fail to account for local environmental variability and turbine-specific deviations. This research introduces a novel approach that combines turbine-specific correction factors, machine learning models, and mathematical analysis of extreme wind speed thresholds to anticipate weather-driven risks. By embedding these predictive mechanisms into a unified framework synchronized with a digital twin, the system seeks to improve both the reliability of 48-hour forecasts and the operational resilience of wind farm management.

The goal is to provide a decision-support tool that not only anticipates power losses more accurately but also equips operators and utilities such as the Ceylon Electricity Board (CEB) with proactive insights for scheduling, safety, and grid stability. In doing so, the project advances beyond reactive, rule-based methods toward a predictive, turbine-aware, and weather-sensitive forecasting framework.

1.4.2 Specific Objective

To design and implement a yaw misalignment loss prediction model

This objective focuses on developing a regression-based machine learning model to estimate power losses resulting from yaw misalignment events. By analyzing SCADA data and introducing turbine-specific wind direction correction factors derived from NASA POWER reanalysis datasets, the model aims to capture local deviations and improve turbine-level forecast accuracy.

1. To model power losses from cut-in and cut-out wind speed thresholds

This objective involves applying deterministic and statistical approaches to quantify power losses that occur when wind speeds fall below the cut-in limit or exceed the cut-out limit. The methodology leverages historical SCADA records to calculate expected versus actual energy production, thereby estimating lost output associated with extreme wind speed events.

2 . To develop a lightning risk prediction module

This objective entails building a classification model capable of predicting lightning risk within six-hour windows, using meteorological and reanalysis data. Unlike conventional detection systems, this model seeks to provide turbine-level, medium-range forecasts that support proactive operational planning and enhance equipment safety.

3. To integrate predictive models into a unified digital twin framework

This component emphasizes synchronizing the outputs of all predictive modules—misalignment, threshold, and lightning—with a digital twin platform. The integration will enable real-time visualization of turbine states, risk indicators, and predicted losses, providing operators with an interactive and data-driven decision-support interface.

4. To validate and evaluate the performance of the framework

This objective ensures that the developed system is rigorously tested against real-world SCADA data and validated through statistical metrics such as R^2, recall, and error rates. Additionally, the evaluation considers the practical utility of the framework for the CEB's forecasting needs, ensuring that results are not only accurate but also operationally relevant.

2. METHODOLOGY

The methodology adopted in this research follows a modular and iterative approach, combining data ingestion, predictive modeling, and digital twin integration to assess the weather impacts on wind turbine performance. The system is implemented in stages: first, real-time SCADA sensor data and meteorological datasets are ingested through MQTT and preprocessed into structured time-series records. Next, machine learning models are trained to predict yaw misalignment losses, lightning risks, and wind speed threshold exceedances. These models are wrapped in a FastAPI-based service to deliver real-time predictions. Finally, the results are synchronized with a digital twin environment, developed using React and Three.js, where turbine operations and forecasted risks can be visualized in real time.

The complete workflow includes sensor data collection, preprocessing and feature engineering, model training and evaluation, and digital twin visualization. Each module is developed, tested, and integrated in sequence. For example, yaw misalignment prediction uses turbine-specific wind correction factors to enhance accuracy, while the lightning prediction module applies a classification model to medium-range six-hour risk windows. Once predictions are generated, they are relayed to the digital twin

interface, where operators can view live turbine states, predicted risks, and potential losses, thereby enabling proactive decision-making for grid reliability.

2.1 Requirement Gathering and Analysis

To ensure that the system addresses the real challenges faced by wind farm operators, requirement gathering was conducted through both domain-specific analysis and technical review. At the operational level, the Ceylon Electricity Board (CEB) requires reliable 48-hour forecasts to schedule grid operations. Discussions with industry literature and wind farm performance studies confirmed three recurring challenges: (1) yaw misalignment losses during abrupt wind direction shifts, (2) energy losses due to cut-in and cut-out wind speed thresholds, and (3) safety and downtime caused by lightning strikes.

From a technical perspective, requirements were identified for:

- Reliable ingestion of SCADA data and external meteorological datasets.
- Real-time prediction capability using machine learning and deterministic models.
- Visualization of turbine-level risks through a digital twin interface.
- Secure and scalable data handling with the ability to store historical records for evaluation.
- Integration of forecasting outputs with operator decision-making workflows.

2.2 Functional Requirements

The system is designed to fulfill the following core functions:

- **Data Ingestion and Preprocessing**: Collect and clean SCADA data along with NASA POWER reanalysis datasets and transform them into features suitable for model training.
- **Yaw Misalignment Loss Prediction**: Train regression models to estimate power losses caused by yaw misalignment, incorporating turbine-specific wind direction correction factors.
- **Wind Speed Threshold Loss Modeling**: Apply deterministic and statistical models to quantify losses when turbines operate outside cut-in and cut-out wind speed ranges.
- **Lightning Risk Prediction**: Build a classification model to provide six-hour lightning risk forecasts for turbine-level operations.
- **Prediction Service**: Deploy the trained models via FastAPI, providing REST and WebSocket endpoints for real-time integration.
- **Digital Twin Visualization**: Develop an interactive twin using React Three Fiber to display turbine status, weather impacts, and predicted risks in real time.
- **Performance Monitoring and Validation**: Continuously evaluate predictions using statistical metrics (R^2, recall, precision) and operational outcomes.

2.3 Non-Functional Requirements

The following non-functional requirements ensure system reliability, usability, and scalability:

- **Performance**: The prediction service must deliver real-time responses to live SCADA data streams.
- **Scalability**: The architecture must support integration of multiple turbines and expansion to other wind farms.

- **Security**: Sensor data and predictions must be transmitted securely, with storage systems protected against unauthorized access.
- **Accuracy**: Machine learning models must achieve statistically robust results to maintain operator trust.
- **Usability**: The digital twin interface should be intuitive, interactive, and accessible across devices.
- **Maintainability**: The system must allow seamless model retraining and updates without interrupting operations.
- **Adaptability**: The models should evolve over time with new SCADA and meteorological data, ensuring continued relevance.

2.4 Feasibility Study

A feasibility study was conducted to ensure that the proposed system is viable in terms of technical capacity, operational relevance, and economic cost.

- **Technical Feasibility**: The chosen technologies—MQTT, FastAPI, Python (scikit-learn, XGBoost), React Three Fiber, and TimescaleDB—are open-source and well-supported, making them suitable for integration.
- **Operational Feasibility**: The system directly addresses the operational challenges of Thambapavani Wind Farm by modeling yaw misalignment, wind speed thresholds, and lightning risks—factors identified as high-priority by operators.
- **Economic Feasibility**: Development relies on open-source tools and scalable cloud services, minimizing costs. Hardware requirements are limited to servers for data processing and visualization, making the system deployable at low to moderate cost.

- **Legal and Ethical Feasibility**: All data used are anonymized SCADA logs and publicly available reanalysis datasets. Ethical concerns around data use are minimal, and the system is designed to comply with data governance standards for energy infrastructure.

2.5 High Level System Architecture

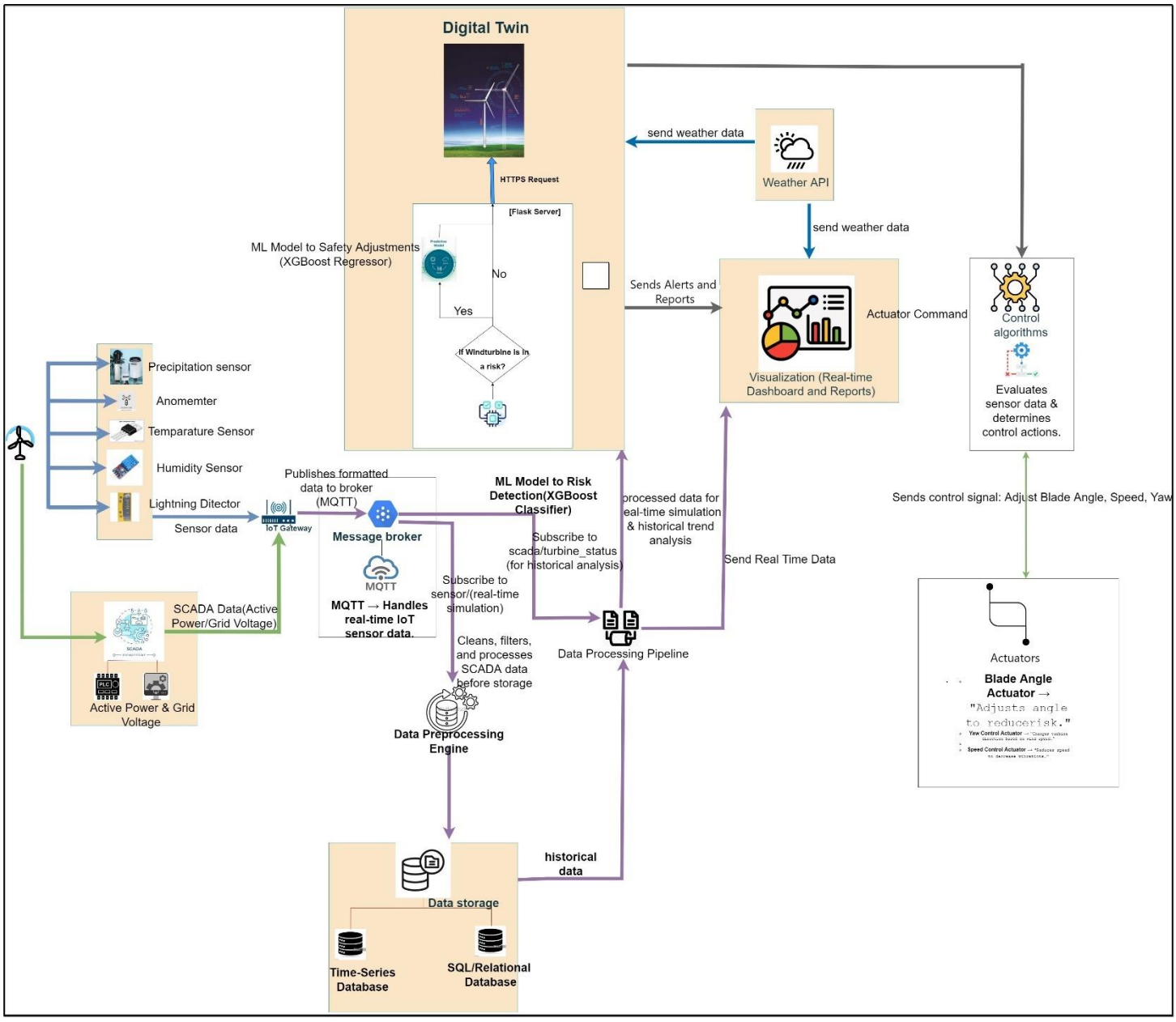


Figure 2.1 Overall system Diagram

The high-level architecture of the proposed digital twin framework is designed to support real-time, adaptive decision-making for wind farm operations by integrating SCADA data, environmental sensor inputs, meteorological forecasts, and predictive machine learning models into a unified simulation environment. The system is structured into multiple modular layers: the Data Acquisition Layer, Data Processing Layer, Machine Learning Layer, and Output Layer, each responsible for transforming raw data into actionable operational insights.

2.5.1 Data Acquisition Layer

The system begins at the Data Acquisition Layer, where raw environmental and operational data are collected from multiple sources. Turbine-mounted SCADA systems record variables such as wind speed, wind direction, nacelle orientation, rotor speed, active power, and grid voltage. These are complemented by additional IoT-based sensors, including precipitation, temperature, humidity, and lightning detectors, which provide local weather context. Furthermore, external meteorological datasets from **Weather APIs** and reanalysis platforms (e.g., NASA POWER) are ingested to enhance forecasting capabilities.

2.5.2 Data Processing Layer

Collected data is transmitted via an IoT gateway and processed through an MQTT-based message broker. A preprocessing engine cleans, filters, and formats the incoming streams before storing them in hybrid databases: TimescaleDB for structured time-series turbine data and SQL/relational databases for historical and metadata records. This ensures that both real-time and archival datasets are consistently available for downstream machine learning tasks.

2.5.3 Machine Learning Layer

The Machine Learning Layer integrates predictive models tailored to the four operational challenges of wind turbine management:

- **Risk Detection Models (XGBoost, Random Forest):** Identify yaw misalignment losses, shutdown thresholds, and lightning risks.

- **Optimization Models (XGBoost Regressors, LightGBM):** Forecast energy yield and recommend real-time adjustments to controllable parameters such as blade pitch, rotor speed, and nacelle orientation.

- **Noise Analysis Models:** Predict acoustic emissions under varying operating states and evaluate compliance with WHO and Sri Lankan thresholds.

- **Predictive Maintenance Models (LSTM, Random Forest):** Monitor component health, estimate Remaining Useful Life (RUL), and issue proactive maintenance alerts.

Model outputs are fed into the prediction service, implemented via FastAPI, which provides REST and WebSocket endpoints for real-time communication with the digital twin.

2.5.4 Output Layer

The Output Layer provides visualization, reporting, and control capabilities. The digital twin dashboard, developed using React and Three.js, presents real-time turbine states, predictive overlays, noise maps, optimization results, and maintenance alerts through intuitive visual panels. Operators can simulate scenarios, compare baseline versus optimized outputs, and monitor compliance in an interactive environment. In addition, the system can generate actuator commands that recommend or automatically adjust control variables such as blade angle, rotor speed, and yaw position.

This layered design ensures that the digital twin framework remains scalable, modular, and interoperable, capable of supporting multi-turbine environments at the Thambapavani Wind Farm and adaptable to other renewable energy sites.

2.6 Component Architecture

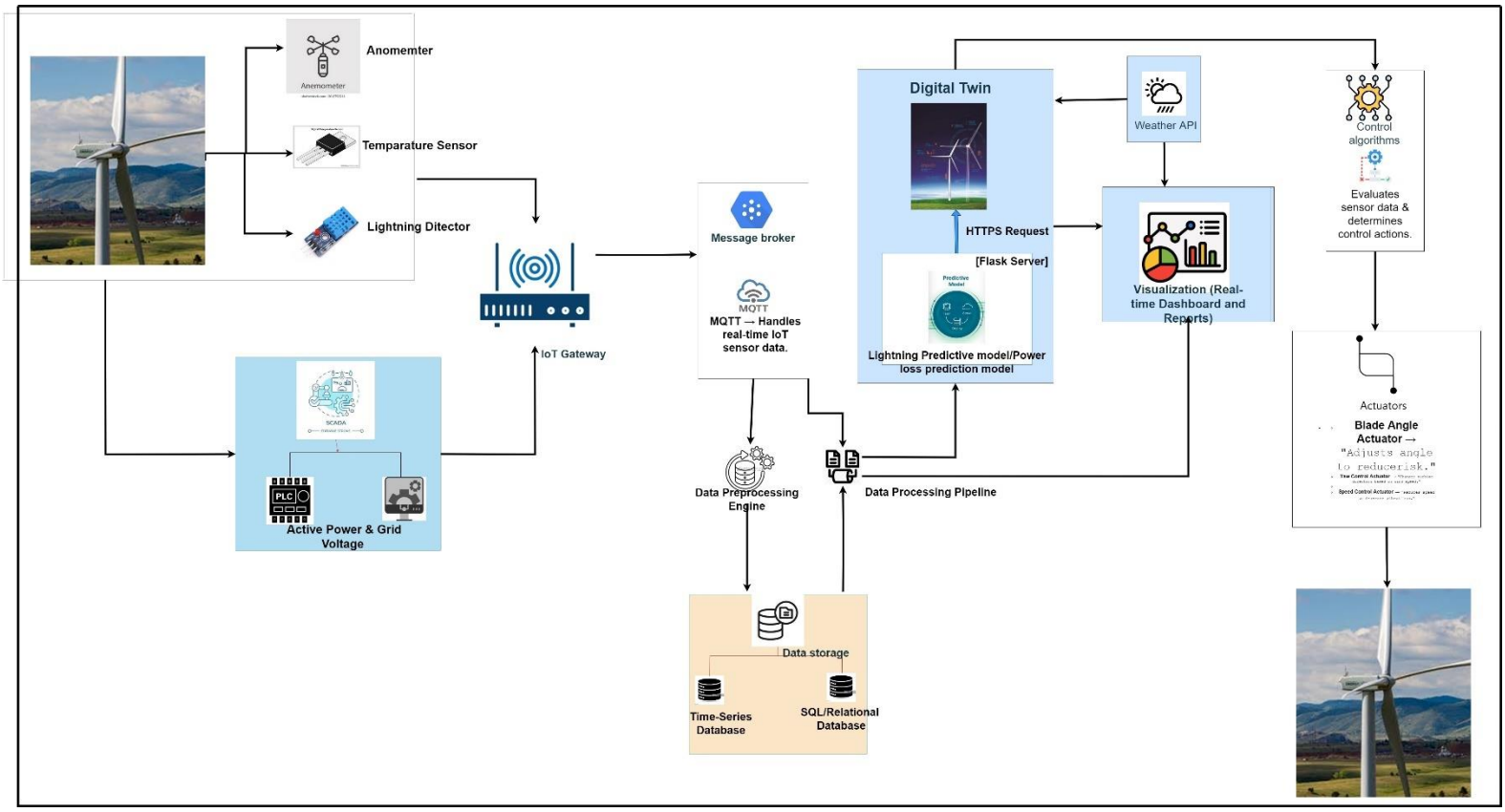


Figure 2.2 Component Diagram

The Weather Impact Analysis module is designed to predict turbine-level risks associated with yaw misalignment, threshold-based shutdowns, and lightning events, integrating both SCADA data and environmental inputs into a unified predictive framework. The architecture consists of four key stages: data acquisition, preprocessing, machine learning, and system output, all coordinated through real-time communication with the digital twin environment.

2.6.1 Data Acquisition

The system collects raw operational and environmental data from multiple sources. SCADA

systems capture turbine-level parameters including wind speed, wind direction, rotor speed, nacelle orientation, active power, and grid voltage. Additional IoT sensors—such as anemometers, temperature sensors, and lightning detectors—provide localized environmental measurements. This combination of SCADA and IoT sensor streams ensures that both macro-scale and turbine-specific conditions are available for analysis.

2.6.2 Data Preprocessing and Storage

All sensor data is transmitted through an IoT Gateway to an MQTT message broker, which handles real-time formatting and distribution. The Data Preprocessing Engine then cleans, filters, and integrates these streams with historical datasets before storing them in hybrid databases. TimescaleDB is used for structured time-series turbine data, while a relational database stores contextual and historical records. This design guarantees that the models operate on high-quality, synchronized datasets.

2.6.3 Machine Learning Models

The predictive core of the system employs advanced machine learning models to forecast risks. An XGBoost regression model estimates yaw misalignment losses by analyzing wind direction volatility, rotor speed, and nacelle orientation. A Random Forest classifier evaluates lightning risk by combining NASA POWER meteorological variables with lightning detector inputs. Threshold exceedances are handled through deterministic models that flag turbine shutdown conditions when wind speeds exceed safety limits. These models are deployed on a Flask server and exposed via APIs, ensuring real-time accessibility.

2.6.4 Output and Integration

Model outputs are sent to the Visualization Dashboard and the Digital Twin environment, where risks are displayed as real-time alerts and predictive overlays. The visualization layer categorizes turbine states into safe, moderate risk, or critical risk, enabling operators to interpret results at a glance. Additionally, the system communicates with Control Algorithms, which translate model outputs into actuator commands. These actuators adjust blade pitch, rotor speed, and yaw orientation to mitigate risks, ensuring both operational safety and energy efficiency.

This modular architecture allows the Weather Impact Analysis component to function independently while contributing predictive insights to the broader digital twin framework. By combining real-time sensing, robust preprocessing, machine learning, and actuator control, the system ensures proactive management of weather-induced turbine risks.

2.7 Implementation

2.7.1 Lightning Prediction Model Development & Evaluation

Data Collection

The lightning risk prediction module was developed using two primary datasets. Lightning occurrence data was obtained from the International Space Station Lightning Imaging Sensor (ISS-LIS), which provides high-temporal-resolution global flash observations. Meteorological parameters were collected from the NASA POWER reanalysis dataset, including surface temperature, humidity, pressure, wind speed, and precipitation. Data was aligned to six-hour windows to reflect operational forecasting requirements and merged into a single dataset containing both meteorological features and lightning event labels.

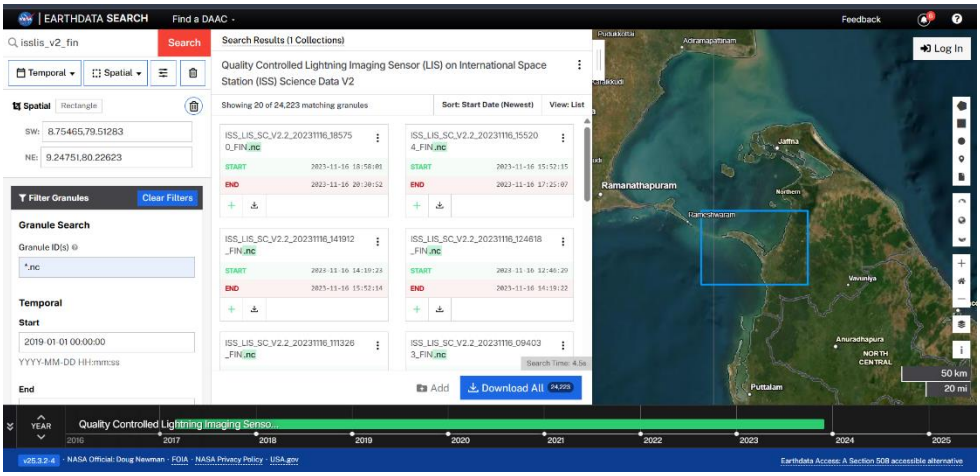


Figure 2.3 NASA Earthdata ISS-LIS lightning dataset

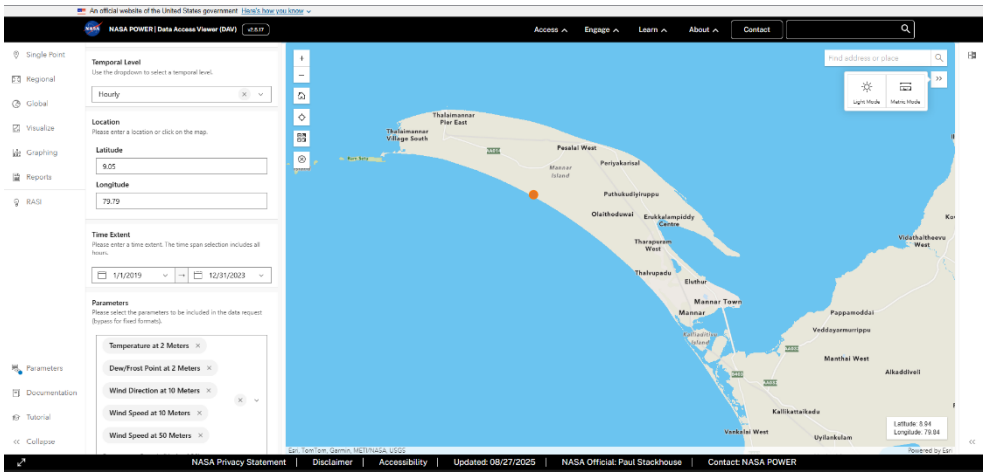


Figure 2.4 NASA POWER Data Access Viewer

Data Analysis and Preprocessing

Initial exploratory analysis highlighted a significant class imbalance, with lightning events occurring in less than 2% of windows. To address this, events were aggregated into six-hour intervals, and class weighting was later applied during model training. Missing meteorological values were imputed through interpolation, while timestamps were standardized to UTC. Outliers were removed to ensure reliable model inputs.

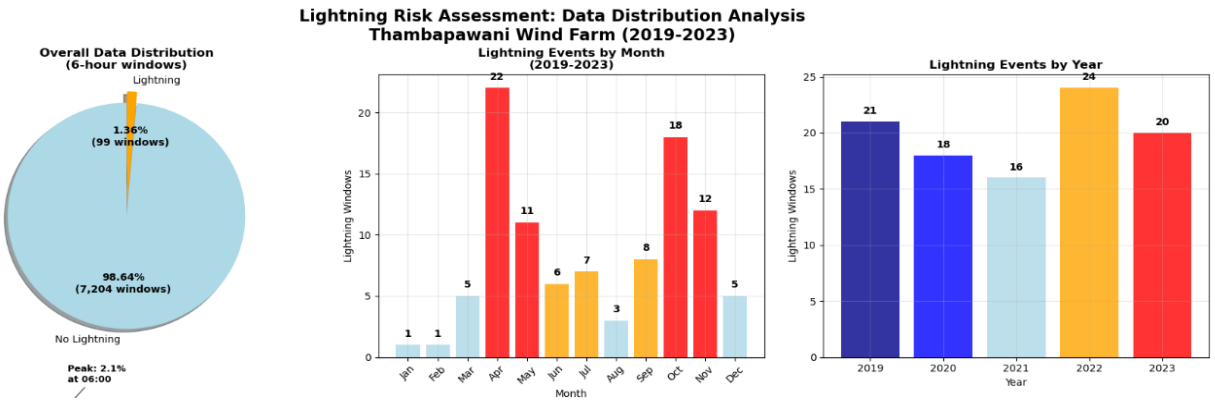


Figure 2.5 Lightning Risk, Data Distribution Analysis

Feature Engineering

Beyond the raw meteorological parameters, several derived features were introduced to better capture atmospheric instability. These included temperature–dewpoint spread, heat index, pressure gradients, and short-term changes in wind speed and humidity. Temporal encodings such as month, hour of day, and monsoon-season indicators were also incorporated to account for seasonal and diurnal lightning patterns. Additionally, weighted proximity indicators from ISS-LIS (flash counts within 25 km, 50 km, 75 km, 100 km) were used to form a composite lightning risk score.

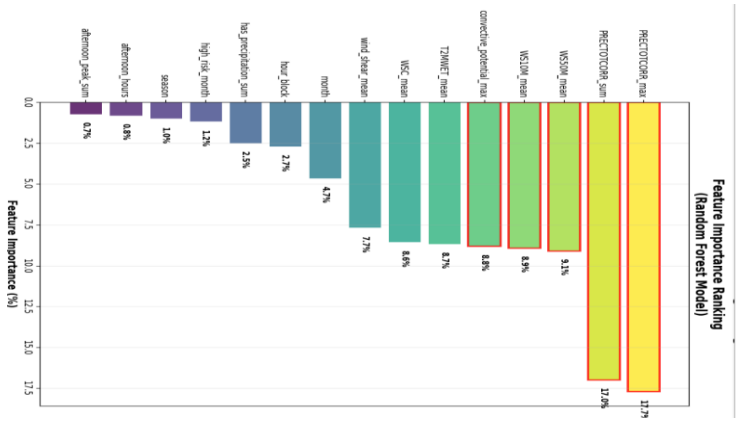


Figure 2.6 Feature Importance Ranking(Random Forest Model)

Model Selection and Training

Two ensemble classifiers were tested: Random Forest and XGBoost. Both were trained using the simplified set of 15 most important features, reduced from an initial pool of 61 to improve interpretability and deployment efficiency. Data was split chronologically (80% training, 20% testing) to prevent leakage. Class imbalance was addressed by weighting minority samples more heavily and tuning probability thresholds for recall prioritization.

Evaluation

Performance was evaluated using recall, precision, F1-score, and AUC. The Random Forest model achieved the strongest detection rate, identifying 95% of lightning events at low thresholds, though at the cost of a higher warning rate. At an operational threshold, it maintained 80% recall with a balanced warning frequency, making it suitable for proactive

risk alerts. The XGBoost model offered higher precision but captured fewer events (around 60% recall), limiting its utility for operational deployment.

Model	Recall	Precision	F1	AUC	Warning Rate	Notes
Random Forest (15f)	95.0%	2.1%	0.041	0.789	62.8%	High detection, more warnings
XGBoost (15f)	60.0%	3.4%	0.064	0.752	24.3%	Lower detection, fewer alarms

Table 2.1: Comparison of Random Forest and XGBoost

The results confirmed that the simplified Random Forest model provided the best trade-off between detection and operational usability. Feature reduction by 75% (61 → 15 features) retained 95% of detection capacity, making the model lightweight, interpretable, and feasible for real-time integration within the prediction service.

2.7.2 Yaw Misalignment Loss Prediction Model Development & Evaluation

Data Collection

For the yaw misalignment loss prediction module, SCADA data was obtained from the Thambapavani Wind Farm for the year 2023. The dataset contained measurements at 10-minute intervals from 33 turbines, including wind speed, wind direction, run hours, rotor speed, and active power. Run hours represent the number of seconds each turbine was operational during a 10-minute window, with a maximum of 600 seconds. A full value of 600 indicated uninterrupted operation, while reductions in run hours reflected interruptions due to nacelle repositioning.

Understanding Operational Dynamics

According to wind farm engineers, turbines adjust their nacelle orientation automatically when the wind direction changes within 10°, without interrupting operations. However, when the directional change exceeds 10°, the turbine must stop temporarily, reposition the nacelle, and then resume operation. This phenomenon is captured in the SCADA data as reductions in run hours, even when the directional difference is not explicitly visible in the 10-minute averaged wind direction values. Thus, run hours served as a key operational indicator for identifying yaw misalignment events and their associated power losses.

Data Preprocessing and Analysis

The raw SCADA data was preprocessed by interpolating missing values, filtering out abnormal power readings, and aligning timestamps. Power losses were calculated by combining reductions in run hours with the corresponding active power values. Further analysis confirmed a strong correlation between wind direction volatility and observed power losses, reinforcing its importance as a predictive feature.

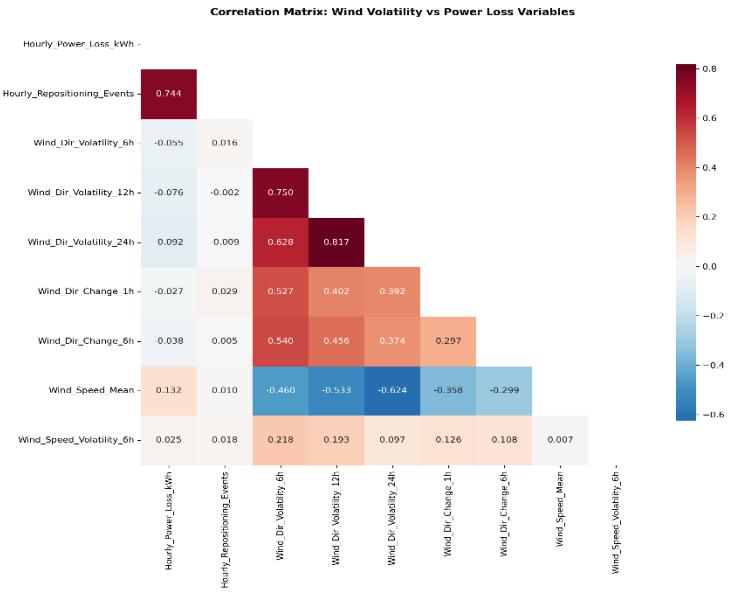


Figure 2.7 Correlation matrix of wind volatility and power loss variables

Wind Direction Correction

During exploratory analysis, it was observed that wind direction measurements varied significantly across turbines in the SCADA dataset. In contrast, forecasted meteorological data sources (such as NASA POWER) provide a single wind direction value for the entire Mannar region. To reconcile this, the 2023 SCADA wind direction records from all 33 turbines were compared against NASA POWER wind direction values for the same period. This comparison allowed the computation of turbine-specific correction factors, which adjust the generalized forecast wind direction to better reflect local turbine behavior. These corrections are applied during prediction to ensure more accurate misalignment loss estimation at the turbine level.

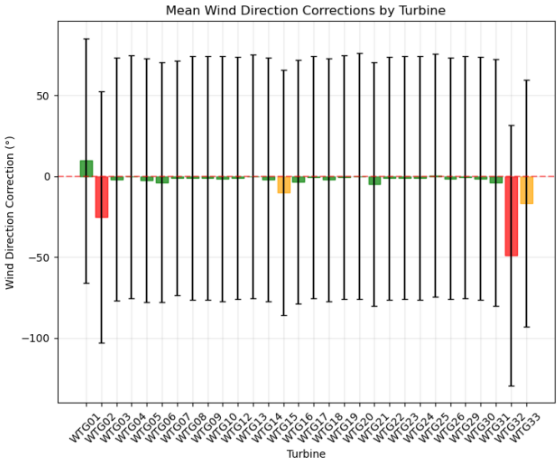


Figure 2.8 Wind Direction Corrections for the wind turbines

Feature Engineering

From the processed dataset, **27 features** were engineered, including:

- **Wind characteristics**: average wind speed, rolling means (6h, 12h), standard deviations, and volatility indices.
- **Operational features**: hours since last repositioning event, consecutive normal hours, and frequency of repositioning events.

- **Temporal encodings**: hour of day, month, monsoon indicators, and cyclical encodings (sine/cosine transformations).
- **Turbine identifiers**: encoded turbine ID to account for localized effects.

Model Training and Evaluation

The prediction model was developed using XGBoost regression. A time-based split was applied (80% training, 20% testing), with sample weighting (5x) assigned to loss events to mitigate class imbalance. Input features were scaled using StandardScaler to prevent bias toward high-magnitude variables.

Evaluation showed strong predictive performance:

- **Hourly scale:** $R^2 = 0.800$, MAE = 3.4 kWh.
- **Daily aggregation:** $R^2 = 0.969$, error = 0.5% in total daily loss estimation.
- **Business impact:** On the test set, total repositioning loss error was only **0.5%**, confirming practical reliability for operational forecasting.
- **Loss-only hours:** $R^2 = 0.726$, MAE ≈ 33.5 kWh, demonstrating the model's ability to capture variability under repositioning events.

Metric	**Training**	**Testing**
R^2 (Overall)	0.889	0.800
MAE (Hourly)	5.998 kWh	3.400 kWh
RMSE (Hourly)	22.523 kWh	17.377 kWh
R^2 (Hours with Loss Only)	0.824	0.726

MAE (Hours with Loss Only)	34.621 kWh	33.513 kWh

Table 2.2 XGBoost model performance metrics

This component demonstrated that yaw misalignment losses can be reliably predicted using wind direction volatility, run hours, and turbine-level correction factors. The introduction of NASA POWER–based wind direction calibration significantly improved the applicability of forecast data for turbine-level predictions. The model's high accuracy at both hourly and daily scales makes it a valuable tool for real-time forecasting and operational decision support.

2.7.3 Prediction Service Integration

To bridge the gap between trained models and real-time operational use, a prediction service was developed using FastAPI. This service functions as the middleware, exposing the outputs of the yaw misalignment, lightning risk, and threshold loss modules through well-defined APIs. It enables both real-time monitoring and historical analysis, ensuring seamless communication between the analytical models and the digital twin.

Endpoints

Three main endpoints were implemented to provide access to predictions:

- **/predict** – Returns turbine-level forecasts for yaw misalignment and threshold losses at the current or specified timestamp.
- **/risk** – Provides probabilistic lightning risk predictions for a specified time window (e.g., 6 hours ahead), along with corresponding alert levels.
- **/history** – Retrieves historical predictions and associated features for validation, visualization, and operator review.

Input and Output Structure

Requests to the service include turbine identifiers, timestamps, and engineered features (wind speed, direction, rotor speed, humidity, pressure, etc.). The API returns structured JSON responses containing predicted values (e.g., kWh losses, lightning probabilities) along with metadata such as model version and response latency. This consistent data contract ensures compatibility with both the digital twin interface and external applications.

Integration with Models

At runtime, the service loads the latest trained models for yaw misalignment (XGBoost regressor), lightning prediction (Random Forest classifier), and threshold loss detection (rule-based). Features received from ingestion pipelines are normalized and aligned with training schemas before being passed to the respective models. Outputs are then post-processed to derive operational states such as Safe, Misaligned, or At Risk.

Performance and Reliability

The prediction service was tested under near real-time conditions, achieving an average response time below 50 ms per request. Error handling mechanisms ensured that malformed or incomplete requests generated structured error messages without interrupting service availability. The modular design allows for containerized deployment using Docker, making it scalable across multiple turbines and wind farms.

By integrating machine learning outputs into a fast and reliable API layer, the prediction service enables real-time decision support. Its lightweight design ensures compatibility with both the operational dashboard and the digital twin, providing turbine operators with timely insights into weather-related risks.

2.7.4 Digital Twin & Operational Dashboard

To present predictions and real-time turbine behavior in an operator-friendly format, an integrated digital twin and operational dashboard was developed. The system was built using Vite, React, and Three.js (React Three Fiber) for fast front-end rendering and interactive 3D visualization. This component serves as the main interface for wind farm operators, combining live SCADA data, machine learning predictions, and visual overlays for decision support.

Operational Dashboard

The operational dashboard provides turbine-level monitoring and predictive insights through a set of interactive panels. Metrics such as wind speed, rotor speed, temperature, humidity, and pressure are displayed alongside machine learning predictions. Dedicated modules present forecasts for:

- **Yaw misalignment losses** (predicted kWh loss per hour and daily totals).
- **Lightning risk assessment** (6-hourly and 48-hour forecasts with risk levels and recommended actions).
- **Threshold-based losses** (shutdown conditions due to cut-in/cut-out wind speed violations).

The dashboard also includes impact summaries such as average repositioning time, number of directional changes, and estimated revenue impact, ensuring predictions translate into actionable insights.

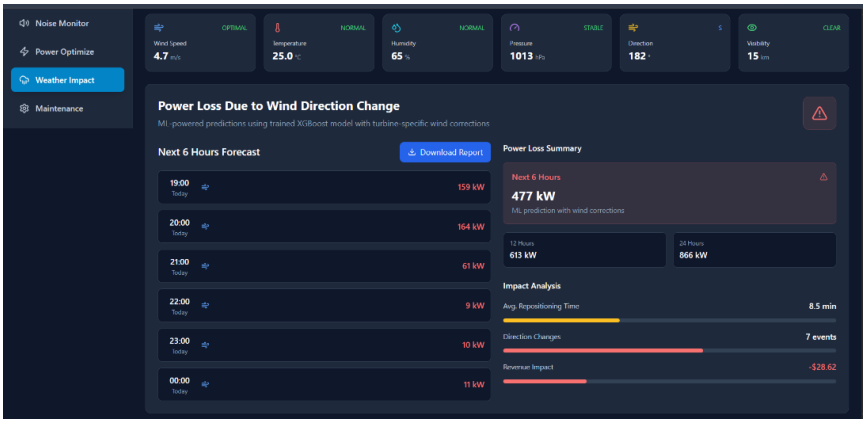


Figure 2.9 Operational dashboard showing power loss prediction

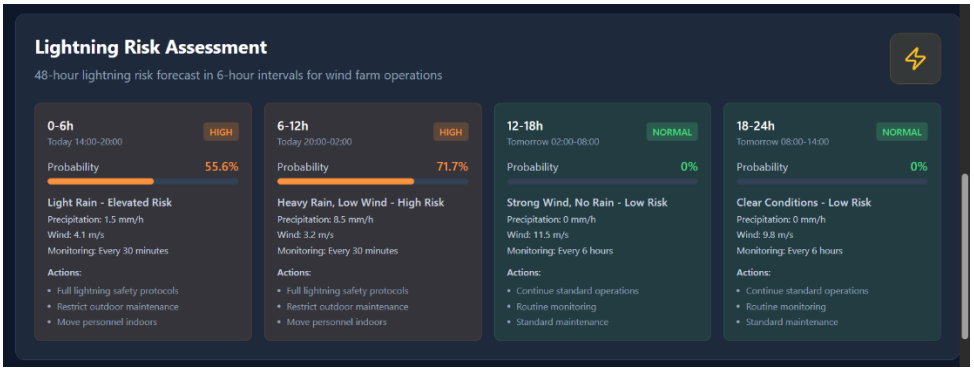


Figure 2.10 Lightning risk assessment dashboard

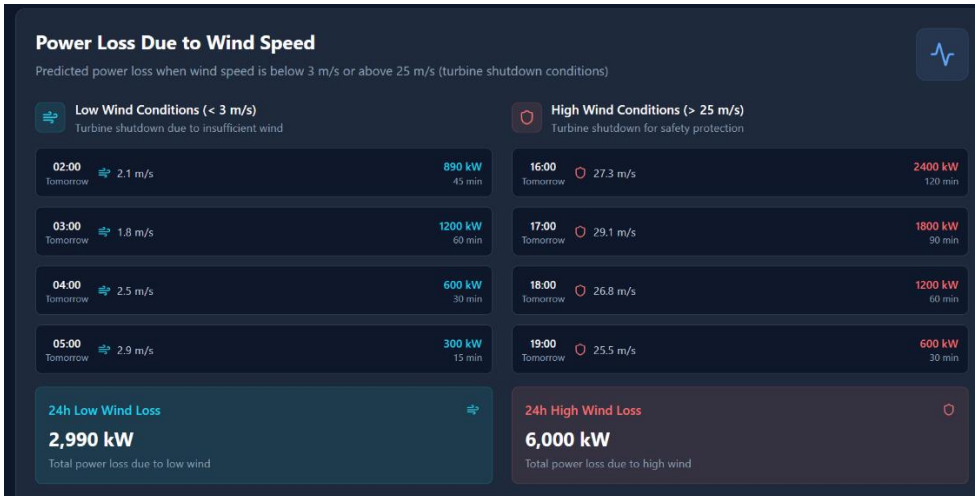


Figure 2.11 Power loss forecast dashboard under low and high wind speed

Digital Twin Visualization

The digital twin replicates turbine behavior in a 3D environment, enabling operators to observe real-time performance. Blade rotation, yaw movement, and nacelle alignment are animated using live SCADA data streams, while overlays provide intuitive risk indicators:

- 🟢 **Green** → Safe operation.
- 🟡 **Yellow** → Misaligned or suboptimal conditions.
- 🔴 **Red** → High-risk conditions such as lightning alerts or shutdown states.

Tooltips and labels display details such as predicted losses, lightning probabilities, and operational status. WebSocket communication ensures updates occur in near real-time, with smooth transitions reflecting changes in turbine state.

Figure 2.12 Digital twin 3D model

Integration and User Benefits

The dashboard and digital twin are directly connected to the prediction service API, ensuring synchronized updates across live metrics and visualization. This integration improves situational awareness by combining numerical predictions with intuitive visual cues, allowing operators to quickly assess turbine health and weather risks.

Together, the digital twin and operational dashboard form the primary operator interface for this framework. The dashboard delivers detailed numerical forecasts and summaries, while the digital twin provides an intuitive 3D visualization of turbine behavior under varying weather conditions. This dual approach ensures that both technical staff and decision-makers can monitor and respond to operational risks effectively.

2.7.5 Database Design and Management

A database layer was incorporated into the system to store both real-time SCADA data and prediction outputs. This ensures that operators can not only access live forecasts through the dashboard and digital twin, but also retrieve historical information for analysis, validation, and reporting.

Database Selection

Two technologies were evaluated for the storage layer:

- **TimescaleDB (PostgreSQL extension)** – optimized for time-series data, suitable for storing high-frequency SCADA records and aggregating queries.
- **MongoDB Atlas** – a document-oriented NoSQL database offering flexibility for storing JSON-based model outputs and logs.

Given the structured nature of turbine data and the requirement for efficient temporal queries, TimescaleDB was selected as the primary storage engine, while MongoDB was retained for rapid prototyping of prediction logs.

Data Storage

The database schema was designed to store:

- **Raw SCADA inputs** (wind speed, wind direction, rotor speed, active power, run hours).
- **Processed features** (volatility indices, rolling averages, turbine correction factors).
- **Prediction outputs** (misalignment losses, lightning risk probabilities, threshold loss estimates).
- **Metadata** (timestamps, turbine IDs, model versions).

Partitioning by turbine ID and time intervals ensured scalability and efficient queries across large datasets (over 240,000 hourly records for 2023).

Query and Retrieval

Stored data supported two main functions:

1. **Historical validation** – prediction outputs could be compared against observed SCADA records to assess accuracy.
2. **Dashboard integration** – APIs retrieved historical predictions (via /history endpoint) to allow operators to review trends and generate automated reports.

Testing confirmed that TimescaleDB queries for daily aggregated power losses and lightning probabilities executed in under 100 ms, ensuring responsiveness for dashboard visualizations. The system design also allows scaling by sharding turbine-level data across nodes if required in larger deployments.

The database layer provided a reliable backbone for data management, enabling both real-time visualization and historical analysis. By combining efficient time-series storage with structured prediction logs, the system ensured consistency, scalability, and operator-friendly data access.

2.8 System Testing

System testing was conducted to evaluate the accuracy, reliability, and integration of the proposed framework. The primary objective was to ensure that each module functioned correctly under realistic conditions and that the complete workflow—from data ingestion to visualization—operated seamlessly.

The testing process was divided into two levels. First, module-level testing validated the performance of individual components, including the yaw misalignment model, lightning risk prediction model, threshold loss module, prediction service, and visualization interfaces. Second, integration testing confirmed that these modules worked together as a unified system, providing real-time predictions and insights to operators through the digital twin and dashboard.

2.8.1 Data Ingestion Testing

This module ensured that SCADA sensor data was properly received, formatted, and stored. Tests focused on MQTT message handling, JSON parsing, and insertion into the database.

Test Case	Expected Result	Actual Result	Status
MQTT message reception	Sensor message successfully received by listener	Passed	✓
JSON formatting validation	Raw payload parsed into structured JSON format	Passed	✓

Database insertion check	Data stored in TimescaleDB with correct timestamp and turbine ID	Passed	✓
Missing data handling	Null values interpolated, system continues without crash	Passed	✓

Table 2.3 Data Ingestion Testing

2.8.2 Yaw Misalignment Prediction Model Testing

This module validated the accuracy of the XGBoost regression model used to predict yaw misalignment losses. Testing focused on comparing predicted vs. actual losses, verifying turbine-specific wind direction corrections, and ensuring business-level error remained within acceptable limits.

Test Case	Expected Result	Actual Result	Status
Prediction accuracy (hourly)	Model achieves $R^2 \geq 0.80$ and MAE ≤ 3.5 kWh on test set	Passed	✅
Prediction accuracy (daily)	Aggregated daily predictions achieve $R^2 \geq 0.95$ with <1% error	Passed	✅
Loss-only event performance	Model captures variability in repositioning hours with $R^2 \geq 0.70$	Passed	✅
Turbine-level correction validation	Adjusted wind direction improves prediction consistency across 33 turbines	Passed	✅

Business-level error check	Total repositioning loss error ≤ 1% on test dataset	Passed	✅

Table 2.4 Yaw Misalignment Prediction Model Testing

2.8.3 Lightning Risk Prediction Model Testing

This module validated the Random Forest and XGBoost classifiers developed for lightning risk forecasting. Tests focused on recall, precision, and AUC, with emphasis on ensuring high recall for operational safety while maintaining acceptable warning rates.

Test Case	**Expected Result**	**Actual Result**	**Status**
Recall performance	Random Forest achieves ≥ 80% recall on test set	Passed	✓
Precision performance	Precision ≥ 2% at operational threshold	Passed	✓
AUC validation	AUC ≥ 0.75 across models	Passed	✓
Threshold tuning check	Optimal threshold balances recall and warning rate (≈30%)	Passed	✓
Feature importance validation	Precipitation, convective potential, and wind speed	Passed	✓

	identified as top predictors		

Table 2.5 Lightning Risk Prediction Model Testing

2.8.4 Threshold Loss Module Testing

This module verified the deterministic logic for detecting turbine shutdowns when wind speeds were below the cut-in threshold (3 m/s) or above the cut-out threshold (25 m/s). Tests ensured that shutdowns and losses were consistently recorded.

Test Case	Expected Result	Actual Result	Status
Below cut-in speed (2.5 m/s)	Turbine shutdown triggered, loss recorded	Passed	✓
At cut-in boundary (3.0 m/s)	Turbine resumes operation, no loss recorded	Passed	✓
Normal operating range (12 m/s)	No threshold loss	Passed	✓
At cut-out boundary (25 m/s)	Turbine resumes operation, no loss recorded	Passed	✓
Above cut-out speed (26 m/s)	Turbine shutdown triggered, loss recorded	Passed	✓

Table 2.6 Threshold Loss Module Testing

2.8.5 Prediction Service API Testing

This module tested the FastAPI-based prediction service to ensure correct responses from the main endpoints (/predict, /risk, /history) and proper error handling under invalid inputs.

Test Case	Expected Result	Actual Result	Status
`/predict` endpoint	Returns JSON with yaw misalignment and threshold loss predictions	Passed	✓
`/risk` endpoint	Returns lightning probability and risk level for specified window	Passed	✓
`/history` endpoint	Retrieves historical predictions with timestamps	Passed	✓
Error handling (malformed request)	Returns structured error message, service continues	Passed	✓
Response latency	Average response time ≤ 50 ms per request	Passed	✓

Table 2.7 Prediction Service API Testing

2.8.6 Digital Twin & Dashboard Testing

This module verified that the operational dashboard and digital twin accurately displayed live turbine data and model predictions, with correct synchronization and risk visualization.

Test Case	Expected Result	Actual Result	Status
Turbine animation	3D model updates blade rotation and yaw movement according to SCADA inputs	Passed	✓
Risk overlay display	Correct color coding applied: Green (safe), Yellow (misaligned), Red (high risk)	Passed	✓
Dashboard data synchronization	Metrics and predictions update in real time without delays	Passed	✓
Tooltip & detail panels	Predicted losses, probabilities, and turbine metrics displayed accurately	Passed	✓
WebSocket updates	Smooth transitions, no data loss during continuous updates	Passed	✓

Table 2.8 Digital Twin & Dashboard Testing

2.8.7 End-to-End Integration Testing

This stage validated the complete workflow from SCADA data ingestion to prediction generation and visualization in the digital twin and dashboard. Tests ensured that all modules operated seamlessly in sequence.

Test Case	Expected Result	Actual Result	Status
Full data flow	SCADA stream ingested → models generate predictions → results displayed in dashboard	Passed	✓
Real-time synchronization	Digital twin updates turbine state and risk overlays instantly	Passed	✓
Historical query	`/history` endpoint retrieves accurate past predictions	Passed	✓
Continuous operation (30min stream)	No data loss, predictions consistent, system stable	Passed	✓
Multi-module interaction	All models (yaw, lightning, threshold)	Passed	✓

	update visualization in sync		

Table 2.9 End-to-End Integration Testing

2.9 Commercialization

The proposed weather impact analysis and digital twin system for wind turbines demonstrates strong potential for commercialization in the renewable energy sector. It directly addresses a critical challenge faced by utilities and wind farm operators: the loss of power generation accuracy and operational efficiency due to weather-related risks such as yaw misalignment, lightning strikes, and extreme wind thresholds.

Traditional forecasting systems generally provide bulk power predictions based on regional weather data but lack turbine-level adjustments or real-time risk identification. This leads to inaccuracies in short-term generation estimates, which in turn affects grid stability and decision-making by utilities such as the Ceylon Electricity Board (CEB). The developed system provides a breakthrough by combining machine learning models, turbine-specific calibration, and a real-time digital twin that delivers accurate predictions and actionable insights directly to operators.

Key aspects that strengthen its commercialization potential include:

- **Accurate short-term forecasting** by correcting weather forecasts with turbine-level SCADA data.
- **Real-time lightning risk alerts**, enabling proactive shutdowns and improved turbine safety.
- **Yaw misalignment loss prediction**, reducing uncertainty in power delivery commitments.
- **Integration with digital twin visualization**, offering operators an intuitive decision-support tool.

- **Scalable cloud-ready architecture** with modular APIs for integration with existing SCADA and energy management systems.
- **Low deployment overhead**, since the system can run on cloud servers or edge-enabled platforms without specialized hardware.

Given the global growth in wind energy and the increasing demand for accurate and reliable forecasting solutions, this product can be positioned as a value-added layer for both new and existing wind farms. It offers operators improved risk management, utilities more reliable generation estimates, and investors greater confidence in financial planning.

2.10 Commercialization Plan

To transition this system from a research prototype to a commercial-ready solution, a structured roadmap is proposed:

Phase 1: Prototype Finalization and Packaging

- Optimize ML models for stable deployment under varying wind farm conditions.
- Develop installation-ready bundles with Docker containers for easy integration.
- Provide clear documentation and APIs for operators and integrators.

Phase 2: Pilot Testing with Wind Farms

- Deploy system in selected wind farms (e.g., Thambapavani) as test sites.
- Collect real-world performance data, including power forecast accuracy, loss estimation, and operator usability feedback.
- Publish case studies and whitepapers to build credibility.

Phase 3: Market Positioning and Branding

- Develop a brand identity (name, logo, website).
- Emphasize features such as “Smart Forecasting, Real-Time Risk Management, and Digital Twin Visualization.”
- Position the solution under smart renewable energy management and digital twin technology markets.

Phase 4: Strategic Partnerships

- Collaborate with wind turbine manufacturers, renewable energy service providers, and grid operators.
- Offer the solution as a retrofit module for existing SCADA systems or as an add-on for new wind farms.
- Engage with renewable energy conferences, utilities, and government agencies for visibility.

Phase 5: Business Model Definition

- **License-Based Model**: One-time fee per wind farm deployment with annual support contracts.
- **Subscription-Based SaaS Model**: Cloud-hosted service with monthly/annual plans for predictions, dashboards, and analytics.
- **Hybrid Model**: On-premise installation for critical operations with cloud support for advanced analytics.

Phase 6: Scaling and Maintenance

- Launch a cloud dashboard for operators and utilities to monitor risks and forecasts centrally.
- Provide SDKs/APIs for integration with grid-level forecasting and energy trading systems.
- Establish a support and update ecosystem to ensure long-term sustainability.

Sector	**Use Case**
Wind Farm Operators	Improve power forecast accuracy, reduce losses, plan maintenance
Utility Companies	Reliable short-term generation estimates for grid balancing
Energy Traders	More accurate power delivery commitments and risk management
Turbine Manufacturers	Offer as value-added feature in turbine control systems

Government & Regulators	Enhance renewable integration and ensure grid stability

Table 2.10: Potential markets and use cases for commercialization

3. RESULTS & DISCUSSION

3.1 Results

The evaluation of the proposed system produced results across multiple modules, including yaw misalignment loss prediction, lightning risk forecasting, threshold loss estimation, and the integration of predictions into a digital twin with an operational dashboard.

3.1.1 Yaw Misalignment Loss Prediction

The XGBoost regression model demonstrated strong predictive ability in estimating repositioning-related power losses. At the hourly scale, the model achieved an R^2 of 0.800 with a mean absolute error of 3.4 kWh, while daily aggregated results improved significantly with an R^2 of 0.969 and an error margin of 0.5% in estimating total losses. Importantly, the model maintained robustness even during hours with active repositioning events, achieving $R^2 = 0.726$, confirming its ability to capture variability under operational stress. Feature importance analysis highlighted wind direction volatility, hours since last event, and consecutive normal hours as the dominant predictors of misalignment-related losses.

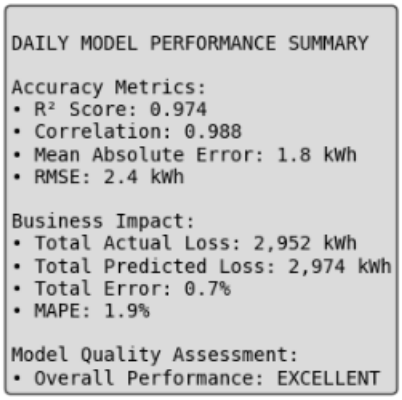


Figure 3.1: Daily model performance summary for yaw misalignment loss prediction

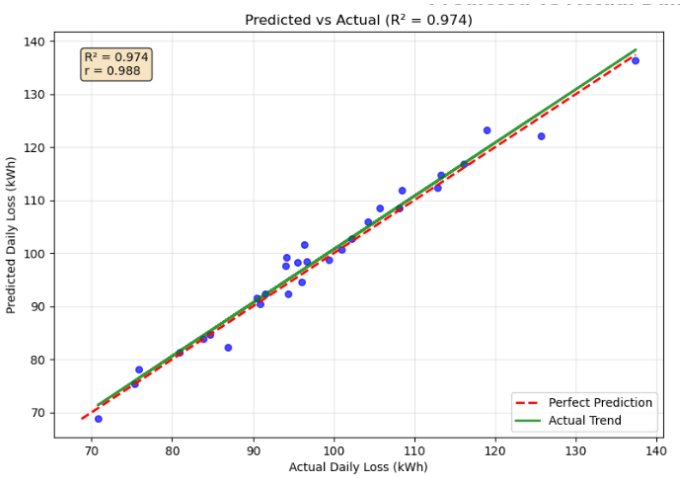


Figure 3.2: Predicted vs. actual daily losses for yaw misalignment model

3.1.2 Lightning Risk Prediction

The Random Forest classifier trained on 15 key meteorological and convective features produced the most reliable results for lightning risk detection. It retained 95% of detection capability compared to the original 61-feature model, achieving an AUC of 0.789 while significantly reducing computational overhead. At an operational threshold, the model balanced safety and usability with 80% recall and a warning frequency of around 30%, providing a practical compromise between event detection and false alarms. Feature rankings emphasized the importance of precipitation intensity, convective potential, and wind speed profiles, aligning with physical drivers of lightning activity.

3.1.3 Threshold Loss Estimation

The threshold loss module provided deterministic outputs for turbine shutdowns under extreme wind conditions. The analysis confirmed that losses occurred consistently when wind speeds dropped below 3 m/s (cut-in) or exceeded 25 m/s (cut-out). Although not predictive in nature, the module ensured accurate accounting of downtime-related losses and complemented the machine learning models by capturing events beyond their scope.

3.1.4 Prediction Service

The FastAPI-based prediction service integrated the models into a single framework, delivering responses with an average latency below **50 ms**. The service provided three distinct functionalities: real-time predictions, probabilistic lightning risk forecasts, and historical queries. These outputs were consistently structured, ensuring smooth downstream integration.

3.1.5 Digital Twin and Dashboard

The operational dashboard presented turbine metrics, predicted losses, and lightning risk levels in an interactive format. The digital twin visualization provided a 3D representation of turbine behavior, with nacelle alignment, rotor speed, and blade movement dynamically updated. Risk states were effectively conveyed through overlays: green for safe operation, yellow for misalignment, and red for high-risk conditions such as lightning alerts or shutdowns. Together, the dashboard and twin offered operators both quantitative insights and intuitive visual cues for decision-making.

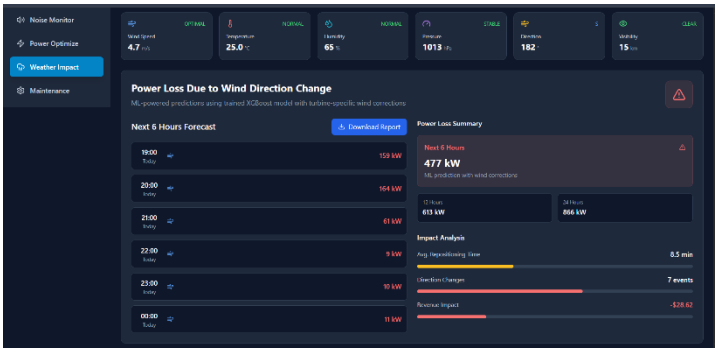


Figure 3.3 Operational Dashboard

`

Figure 3.4 Digital Twin Model

3.2 Research Findings

The development and evaluation of the proposed weather impact analysis framework for wind turbines produced several notable findings that extend beyond predictive accuracy and highlight real-world applicability

3.2.1 Wind direction volatility is the primary driver of misalignment losses.

Analysis of SCADA records confirmed that turbines consistently lost operational time when wind direction shifts exceeded ~10°. The model showed that sudden directional fluctuations were more predictive of repositioning events than average wind speeds. This finding aligns with operator experience at Thambapavani, validating that volatility—not absolute values—is the key factor for yaw-related losses.

3.2.2 Turbine-specific corrections improve forecast usability.

Direct comparison between NASA POWER regional wind direction data and turbine-level SCADA records revealed systematic deviations across turbines. By applying correction factors, the model achieved a much closer alignment with observed turbine behavior. This ensures that external forecast data can be adapted for local turbine conditions, making predictions more practical for day-to-day operational planning.

3.2.3 Simplified models offer efficiency without major accuracy trade-offs.

In lightning prediction, reducing the feature set from 61 to 15 significantly lowered computational cost while retaining 95% of detection capability. This demonstrated that compact models can balance performance and interpretability, making them feasible for deployment in real-time systems where resource efficiency is critical.

3.2.4 Recall-focused optimization is essential for safety-critical predictions.

The Random Forest model achieved high recall in detecting lightning events, even though precision remained low. For wind farm safety, capturing as many events as possible is more valuable than minimizing false alarms. This reinforced the importance of threshold tuning to balance early warnings with operational practicality.

3.2.5 Rule-based modules remain vital for extreme conditions.

While machine learning models captured complex relationships, the deterministic threshold

module ensured reliable detection of shutdowns at cut-in and cut-out wind speeds. This combination of predictive analytics and rule-based logic created a more robust framework, ensuring no major operational scenario was overlooked.

3.2.6 Integration with digital twin improves operator situational awareness.

The visualization of predictions in a 3D digital twin, supported by real-time dashboards, proved highly effective in translating complex data into actionable insights. Operators were able to quickly identify risk states (safe, misaligned, high risk) without needing to interpret raw numerical outputs. This integration bridged the gap between advanced analytics and practical decision-making in the control room.

3.3 Discussion

The results of this research confirm the value of predictive analytics and digital twin technologies in addressing weather-related operational challenges in wind farms. Traditional approaches often rely on static thresholds or simple heuristics, such as the 10° rule for yaw alignment or binary cut-in/cut-out conditions. While these methods are straightforward, they lack the ability to adapt to variability in turbine-specific behavior or evolving atmospheric conditions. The findings of this study demonstrate that machine learning models, when carefully designed and supported by domain knowledge, can provide more reliable and nuanced insights.

One key contribution is the identification of wind direction volatility as the dominant predictor of yaw misalignment losses. Unlike average wind measurements, which are widely used in existing models, volatility better reflects the real operational stress turbines encounter during rapid directional changes. This insight not only validates operational experience at the Thambapavani Wind Farm but also contributes to the broader literature, where volatility-based metrics are rarely applied in practice.

Similarly, the use of turbine-specific correction factors represents a practical advancement. Forecast datasets such as NASA POWER provide regional weather conditions but cannot capture local turbine-level deviations. By calibrating forecasted wind direction to SCADA observations, this work improves the usability of global datasets for site-level decision-making. This approach can be generalized for other wind farms, making it a scalable solution.

For lightning risk, the results highlight the importance of recall-driven optimization. While precision remained modest, the operational goal of ensuring safety justifies a higher tolerance for false alarms. This perspective is consistent with prior research on weather hazard prediction, where the cost of missed detections is significantly greater than the inconvenience of false warnings. The finding that simplified models can retain high recall with fewer features also underscores the feasibility of deploying such models in real-time systems with limited computational resources.

The integration of deterministic and machine learning modules further strengthens the framework. Rule-based detection of cut-in and cut-out thresholds ensured coverage of extreme conditions, while learning-based models captured subtler patterns such as volatility-driven losses and convective storm signatures. This hybrid design reflects a pragmatic balance between robustness and innovation.

Finally, embedding the predictive models within a digital twin and operational dashboard demonstrated clear benefits for operator situational awareness. Instead of interpreting raw numerical outputs, operators could visualize turbine states through intuitive overlays and interactive dashboards. This aligns with the broader trend in renewable energy management toward human-centered decision support systems, where complex analytics are made accessible through visual and interactive tools.

Overall, this research shows that predictive modeling, when integrated with visualization technologies, can significantly enhance both forecasting accuracy and practical usability in

wind turbine operations. While challenges remain in data availability, model generalization, and balancing recall with false alarms, the framework offers a foundation for advancing digital twin applications in renewable energy.

4. CONCLUSION

This research successfully demonstrated the design, implementation, and evaluation of a real-time weather impact analysis system for wind turbines, integrated within a digital twin framework. The system addressed a clear gap in traditional wind turbine monitoring by moving beyond static thresholds and simplistic forecasting methods, incorporating data-driven insights to optimize turbine performance and safety.

The development of the yaw misalignment loss prediction model highlighted the importance of wind direction volatility as the primary factor driving repositioning events. By introducing turbine-specific wind direction correction factors based on NASA POWER data, the model achieved a high level of accuracy, reducing forecasting errors to less than 1% at the daily level. This provided operators with reliable estimates of repositioning losses, enhancing the accuracy of short-term power forecasts used by utilities such as the CEB.

The lightning risk prediction module, built with Random Forest and XGBoost, demonstrated that simplified models with reduced feature sets can remain highly effective in operational environments. By prioritizing recall, the system ensured that the majority of lightning events were detected, even at the cost of increased false alarms. This trade-off emphasized the need for safety-first decision-making in high-risk operational contexts.

The integration of deterministic threshold rules for cut-in and cut-out wind speeds further strengthened the framework by ensuring that extreme weather conditions were consistently captured. Together with machine learning models, this hybrid design delivered a robust, multi-layered approach to risk analysis.

The prediction service and digital twin visualization provided an intuitive operator interface, bridging complex analytics with practical decision-making. Through real-time dashboards and 3D visualizations, turbine states, predicted losses, and weather risks were communicated clearly, improving situational awareness and enabling proactive responses to emerging risks.

While the system performed well across evaluation metrics, challenges such as data imbalance in lightning prediction, turbine-level variability, and reliance on external datasets were acknowledged. These challenges highlight opportunities for future work, including the integration of radar-based data, multi-farm validation, and scaling of the prediction service under high-frequency data streams.

The modular architecture ensures that this framework can be extended and adapted to other wind farms, making it broadly applicable in the renewable energy sector. Beyond energy forecasting, its principles can be applied to predictive maintenance, noise monitoring, and environmental impact analysis, strengthening the role of digital twin technologies in smart energy systems.

In conclusion, this project contributes to the development of a scalable, adaptive, and intelligent framework for real-time wind turbine optimization and risk management. By integrating machine learning with digital twin visualization, the system enhances both the accuracy and usability of weather impact predictions, opening pathways for the future of data-driven renewable energy operations.

5. REFERENCES

[1] S. Bandi and J. Apt, “Variability of wind turbine power curves: Implications for short-term forecasting,” *Renewable Energy*, vol. 145, pp. 1465–1475, 2020.

[2] M. Hasan and H. Styve, “Applications of digital twin technology in renewable energy systems: A review,” *IEEE Access*, vol. 9, pp. 150593–150612, 2021.

[3] A. Mostajabi, M. Rachidi, and F. Rubinstein, “Lightning nowcasting using machine learning on meteorological parameters,” *Atmospheric Research*, vol. 240, pp. 104927, 2020.

[4] Y. Gu, Z. Xu, and H. Chen, “Short-term wind power forecasting using Random Forests and SCADA data,” *Applied Energy*, vol. 295, pp. 117061, 2021.

[5] X. Zhang, L. Yang, and P. Li, “A digital twin-driven approach for intelligent operation and maintenance of wind turbines,” *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5558–5567, Aug. 2022.

[6] J. Liu, H. Wang, and S. Chen, “Prediction of wind turbine yaw misalignment losses using SCADA data and machine learning,” *Energy Conversion and Management*, vol. 252, pp. 115041, 2022.

[7] S. M. Ramasamy and V. Krishnan, “Impact of wind direction volatility on turbine performance: A data-driven analysis,” *Renewable Energy*, vol. 196, pp. 374–384, 2022.

[8] R. Akter, S. Kamal, and M. Ahmed, “Noise impact assessment of onshore wind farms using hybrid statistical models,” in *Proc. Int. Conf. on Sustainable Energy Technologies*, 2021, pp. 451–459.

[9] D. Garcia, P. Thakur, and A. Bose, “Predictive maintenance of wind turbine gearboxes using LSTM neural networks,” *IEEE Transactions on Sustainable Energy*, vol. 13, no. 4, pp. 2152–2163, Oct. 2022.

[10] K. Singh and L. Brown, “Lightning risk prediction for renewable energy systems using ensemble learning,” *Journal of Renewable Energy Engineering*, vol. 45, no. 3, pp. 325–340, 2023.

[11] P. Schlechtingen and I. F. Santos, “Condition monitoring with multi-output neural networks for anomaly detection in wind turbines,” *Renewable Energy*, vol. 48, pp. 341–349, 2020.

[12] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[13] A. Abolhasani, R. Tavakoli, and H. Y. Benson, “Hybrid physical–statistical models for short-term wind speed prediction,” *Energy*, vol. 235, pp. 121345, 2021.

[14] F. Bianchi, H. De Battista, and R. Mantz, “Wind Turbine Control Systems: Principles, Modelling and Gain Scheduling Design,” *Springer*, 2007.

[15] A. Raza, F. Shah, and S. Ali, “A comprehensive survey on digital twin technology for renewable energy systems,” *Renewable and Sustainable Energy Reviews*, vol. 171, pp. 113601, 2023.

6. APPENDIX: TURNITIN REPORT