Predictive Maintenance for Wind Turbines – Detailed Report

Introduction

Wind turbines are critical components in renewable energy production. Unexpected failures lead to downtime, revenue loss, and increased maintenance costs. Predictive maintenance aims to anticipate turbine failures before they occur, minimizing unplanned downtime and improving operational efficiency.

This project uses SCADA sensor data from wind turbines to:

- Predict failures as early as possible
- Keep false alarms low
- Provide interpretable results for operators

Two approaches are implemented and compared:

- **Supervised Learning (Random Forest Classifier)**
- **Unsupervised Learning / Clustering (K-Means)**

The final goal is to provide a user-friendly system where operators can upload weekly or monthly SCADA datasets and receive risk assessments for each turbine, along with actionable recommendations.

Data Analysis & Feature Selection (WP1)

Dataset Exploration

The datasets include sensor readings from multiple turbines:

- **Rotor speed** – indicates turbine rotation rate
- **Generator speed** – shows electrical generation performance
- **Wind speed** – environmental factor affecting turbine load
- **Temperature** – monitors mechanical and electrical component health
- **Power output** – turbine efficiency

These parameters provide both environmental conditions and turbine operational metrics, which are critical for detecting deviations that may indicate failure.

Failure Event Information

A separate CSV file (event_info) provides:

- Start and end timestamps of failure events
- Turbine ID associated with each event

We use this information to label each sensor reading as either:

- 1 = Failure
- 0 = Normal

This labeling is essential for supervised learning, allowing the model to recognize patterns preceding failures.

Feature Selection

We use all numeric sensor readings and generate lag features (previous 3 timestamps) to capture temporal trends. Selected features include rotor speed, generator speed, wind speed, temperature, and power output.

Rationale:

- Sensor readings are directly correlated with turbine health.
- Lag features allow early detection before a catastrophic failure.
- Tree-based models (Random Forest) are robust to outliers, ensuring extreme readings do not skew predictions.

Data Preprocessing

- Missing values filled with column means
- Scaling using StandardScaler for model efficiency
- Outliers handled naturally by tree-based models

Modeling Approaches (WP2)

Supervised Learning

- **Target:** label (1 = failure, 0 = normal)
- **Model:** Random Forest Classifier
- **Features:** Sensor readings + lag features
- **Output:** Failure probabilities per turbine reading

Advantages:

- Predicts failures several timestamps in advance
- Allows threshold adjustment to reduce false positives
- Provides feature importance for interpretability

Why it’s preferred:

- Gives exact failure probability, allowing operators to take precise preventive actions
- Minimizes false alarms, crucial for operational decision-making

Unsupervised Learning / Clustering

- **Method:** K-Means clustering
- **Output:** Cluster-based risk scores highlighting anomalous readings
- **Purpose:** Detect potential failures even when labels are rare or unknown

Notes:

- Cannot provide exact failure probabilities
- Useful as a complementary exploratory tool to highlight unusual turbine behavior

Evaluation & Results (WP3)

Supervised Model Metrics

- Precision: High → few false alarms
- Recall: Most failures detected early
- F1-score: Balanced performance
- Lead time: Early prediction using lag features

Unsupervised Model Metrics

- Detects anomalous clusters compared to actual failures
- Highlights unusual patterns but is less precise

Comparison

- Supervised Random Forest predicts exact failure probabilities → preferred for operational decisions
- Unsupervised K-Means highlights anomalies but lacks precision → complementary use

Recommendation

- Use **Random Forest** for routine monitoring and predictive maintenance
- Use **K-Means** for exploratory anomaly detection

Future improvements:

- Include more historical data and lag features
- Experiment with deep learning (LSTM, autoencoders) for sequential pattern detection
- Deploy interactive dashboards for real-time SCADA data analysis
- Gather operator feedback to refine thresholds

Workflow / Process Flow Diagram

Dataset Collection

↓

Data Preprocessing (missing values, scaling, lag features)

↓

Model Training

Supervised Learning → Failure Probabilities

Unsupervised Learning → Cluster-based Risk Scores

↓

Evaluation & Comparison (Precision, Recall, F1, Lead Time)

↓

Operational Recommendation

↓

Predict safe/unsafe turbines, low false alerts

Explanation:

1. Users provide SCADA data
2. Preprocessing ensures clean and normalized input
3. Two models analyze turbine performance
4. Supervised model predicts exact risk, unsupervised detects anomalies
5. System recommends actions with clear probability and safety indicators

Output Visualizations

- Confusion Matrix – shows supervised model performance
- Feature Importance – highlights top predictors of failure
- Failure Dashboard – visualizes turbine risk over time

Conclusion

- Supervised Random Forest is **best for operational predictive maintenance** because it provides interpretable risk probabilities, minimizes false alarms, and allows early interventions.
- Unsupervised clustering complements the approach by detecting unusual patterns not yet labeled as failures.
- The system can be extended for **interactive use**, allowing users to upload weekly/monthly SCADA data for immediate safety assessment.

Future Enhancements

- Deploy interactive dashboards for real-time SCADA monitoring
- Experiment with deep learning models for better sequential pattern detection
- Continuous improvement through operator feedback

- Deploy the system in real-world wind farms for live predictions