

Universidade Federal de São Paulo - UNIFESP

Instituto de Ciência e Tecnologia - ICT

Bacharelado em Ciência e Tecnologia / Engenharia de Computação

Predição de Falhas em Turbinas Eólicas Utilizando Modelos de Aprendizado Profundo e Dados Meteorológicos Exógenos

Projeto de Pesquisa na modalidade Iniciação Científica, submetido à
Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

Aluno: Thiago Solé Gomes Heleno

Orientador: Marcos G. Quiles

Coorientador: Mateus Giesbrecht (FEEC/Unicamp)

Vigência da Solicitação 01/06/2025 - 31/05/2026

São José dos Campos, 22 de abril de 2025

Resumo

A energia eólica desponta como peça-chave na transição para uma matriz energética mais limpa e sustentável. No entanto, manter as turbinas eólicas operando de forma eficiente e contínua representa um desafio complexo. Um grande obstáculo reside na dificuldade em identificar, com antecedência, defeitos em componentes mecânicos e sistemas eletrônicos – uma capacidade essencial para diminuir custos com reparos e evitar paradas não planejadas na geração de energia. As abordagens tradicionais de manutenção preventiva, muitas vezes baseadas em análises manuais ou modelos físicos simplificados, frequentemente falham em detectar padrões intrincados que podem levar a falhas. Tais limitações podem acarretar diagnósticos imprecisos ou reações tardias a problemas em desenvolvimento. Este projeto propõe o desenvolvimento de modelos preditivos para falhas em turbinas eólicas, por meio da integração de dados operacionais de sistemas SCADA (Supervisory Control and Data Acquisition) com variáveis meteorológicas externas. Investigaremos arquiteturas de aprendizado profundo, como redes LSTM (Long Short-Term Memory), CNN1D (Convolutional Neural Networks 1D) e Transformers, para analisar padrões temporais e buscar correlações entre condições climáticas e anomalias operacionais. Os dados virão de bases públicas (como repositórios SCADA, CPTEC/INPE e OpenWeatherMap) e passarão por tratamento para remoção de outliers e sincronização temporal. Nos experimentos, compararemos modelos treinados apenas com dados da turbina com modelos que incluem também dados meteorológicos, avaliando o desempenho com métricas de classificação (Acurácia, F1-Score, etc.) e erro de previsão (RMSE, MAE, etc.). Com esta pesquisa, esperamos contribuir para a manutenção preditiva em energia eólica, ajudando a reduzir custos de paradas não programadas e a aumentar a confiabilidade dos sistemas.

Abstract

Wind energy stands out as a key element in the transition towards a cleaner and more sustainable energy matrix. However, keeping wind turbines operating efficiently and continuously presents a complex challenge. A major obstacle lies in the difficulty of preemptively identifying defects in mechanical components and electronic systems – an essential capability for reducing repair costs and avoiding unplanned downtime in energy generation. Traditional preventive maintenance approaches, often relying on manual analysis or simplified physical models, frequently fail to detect intricate patterns that can lead to failures. Such limitations may result in inaccurate diagnoses or delayed reactions to developing problems. This project proposes the development of predictive models for wind turbine failures by integrating operational data from SCADA (Supervisory Control and Data Acquisition) systems with external meteorological variables. We will investigate deep learning architectures, such as LSTM (Long Short-Term Memory) networks, 1D CNNs (Convolutional Neural Networks), and Transformers, to analyze temporal patterns and seek correlations between weather conditions and operational anomalies. Data will originate from public databases (like SCADA repositories, CPTEC/INPE, and OpenWeatherMap) and will undergo processing for outlier removal and temporal synchronization. In the experiments, we will compare models trained solely on turbine data with models that also include meteorological data, evaluating performance using classification metrics (Accuracy, F1-Score, etc.) and prediction error metrics (RMSE, MAE, etc.). With this research, we hope to contribute to predictive maintenance in wind energy, helping to reduce costs from unscheduled shutdowns and increase system reliability.

Conteúdo

1	Introdução	5
2	Justificativa	6
3	Revisão Bibliográfica	6
4	Objetivos	8
5	Metodologia Proposta	10
5.1	Atualização do Referencial Teórico	10
5.2	Bases de Dados	10
5.3	Tratamento dos Dados	11
5.4	Desenvolvimento dos Modelos Preditores	12
5.5	Experimentos e Avaliação dos Resultados	13
5.5.1	Experimentos sem Dados Exógenos	13
5.5.2	Experimentos com Dados Exógenos	13
5.5.3	Avaliação dos Resultados	13
6	Cronograma	14
7	Sobre o Candidato	16
8	Infraestrutura Computacional	16

1 Introdução

O aprendizado de máquina (ML) vem ganhando destaque como ferramenta essencial para melhorar a confiabilidade e a eficiência de sistemas industriais complexos, como é o caso das turbinas eólicas, trazendo novas perspectivas para as estratégias de manutenção preditiva [1, 2]. Modelos de ML conseguem analisar padrões temporais em dados operacionais (vibração, temperatura, rotação, etc.) e integrar variáveis externas, como as condições meteorológicas, prevendo assim falhas com maior precisão [3]. Indo além da simples detecção de anomalias, técnicas avançadas como as redes neurais profundas possibilitam modelar as interações dinâmicas entre componentes mecânicos e fatores ambientais, o que ajuda a antecipar cenários críticos que poderiam levar a paradas custosas [4]. Essas abordagens não apenas diminuem custos operacionais, mas também favorecem a sustentabilidade energética ao minimizar desperdícios.

Contudo, construir modelos preditivos robustos esbarra em desafios importantes ligados à disponibilidade e qualidade dos dados. Apesar das bases públicas de dados SCADA (como a Wind Turbine SCADA open data [5]) e meteorológicos (CPTEC/INPE [6], OpenWeatherMap [7]) oferecerem informações valiosas, problemas como outliers, dados faltantes e dessincronização temporal são comuns [8]. Outro ponto é que a coleta de dados em condições operacionais extremas (tempestades, falhas catastróficas) é difícil, limitando a capacidade dos modelos de generalizar para eventos raros [9]. Tais obstáculos acabam comprometendo a escalabilidade e a confiabilidade das soluções propostas.

Diante disso, propomos neste projeto desenvolver e comparar modelos preditivos de falhas em turbinas eólicas, combinando dados operacionais e variáveis meteorológicas externas. Com base em uma revisão da literatura sobre abordagens multimodais em ML, implementaremos arquiteturas como LSTM, CNN1D e Transformers. Elas serão treinadas em dois cenários principais: (1) usando apenas dados da turbina e (2) dados da turbina combinados com dados meteorológicos. Como resultado, esperamos estabelecer protocolos de pré-processamento e avaliação (usando métricas como Acurácia, F1-Score, RMSE, MAE) que orientem a incorporação de dados exógenos em sistemas de predição de falhas, de modo a contribuir para avanços na gestão sustentável de parques eólicos.

2 Justificativa

Antecipar falhas em turbinas eólicas é fundamental para diminuir os custos com manutenções corretivas e evitar interrupções na geração de energia. Ao integrar dados meteorológicos aos modelos preditivos, podemos identificar padrões mais complexos, como o efeito de rajadas de vento ou de variações de temperatura no comportamento da turbina. Além disso, o uso de bases de dados públicas e ferramentas de código aberto (como TensorFlow e PyTorch) favorece a acessibilidade e a reproduzibilidade do projeto, um ponto importante frente às demandas por transparência e sustentabilidade no setor energético. Adicionalmente, a pesquisa contribui para o debate acadêmico sobre abordagens multimodais em aprendizado de máquina, trazendo insights práticos para aplicações em ambientes dinâmicos.

3 Revisão Bibliográfica

A energia eólica ocupa posição central na matriz energética global, porém sua viabilidade econômica está atrelada à operação contínua e eficiente das turbinas. Paradas não planejadas em componentes críticos geram custos elevados de operação e manutenção (O&M), além da perda de produção [10]. O monitoramento de condição (Condition Monitoring - CM) e a detecção precoce de falhas tornam-se, assim, essenciais para otimizar a manutenção e garantir a confiabilidade [11]. Com mais turbinas sendo instaladas globalmente, a necessidade de estratégias de manutenção eficazes se intensifica, refletida no aumento significativo da capacidade instalada anualmente, como visto em 2020 [12].

Embora existam métodos tradicionais de CM, a grande disponibilidade de dados operacionais de baixo custo dos sistemas SCADA (Supervisory Control and Data Acquisition) trouxe novas possibilidades para o monitoramento [12]. Os dados SCADA, mesmo não sendo ideais para CM e talvez menos precisos que sistemas dedicados, ajudam a detectar falhas incipientes, representando uma alternativa economicamente interessante [12, 13]. A aplicação de Inteligência Artificial (IA), e em especial do aprendizado de máquina (Machine Learning - ML), sobre esses dados vem sendo amplamente usada com bons resultados para identificar padrões anômalos que sinalizam falhas iminentes [13].

Diferentes algoritmos de ML, como k-Nearest Neighbors (kNN), Extreme Gradient Boosting (XGBoost) e Redes Neurais Artificiais (RNAs), foram testados com sucesso no diagnóstico de falhas usando dados SCADA, sendo que alguns conseguem prever falhas com semanas de

antecedência [13]. Ultimamente, o aprendizado profundo (Deep Learning - DL) tem ganhado espaço. Redes Neurais Convolucionais (CNNs), por exemplo, foram usadas para analisar representações de dados SCADA (como gráficos de radar) para detectar falhas em conversores, atingindo boa precisão [14]. Modelos híbridos, combinando CNN e LSTM (Long Short-Term Memory), também vêm sendo explorados para capturar características locais e dependências temporais dos dados, com resultados promissores [3].

Um complicador conhecido na análise de dados SCADA é a grande influência das condições ambientais e operacionais. Variações no clima (vento, temperatura, etc.) influenciam o desempenho e as leituras dos sensores, o que pode mascarar falhas ou gerar alarmes falsos. Por isso, integrar e tratar corretamente os dados meteorológicos é importante para normalizar o comportamento esperado da turbina e tornar os modelos de detecção mais robustos [15]. De fato, a combinação de dados SCADA e meteorológicos tem sido explorada com sucesso em outras aplicações relevantes na área de energia eólica, como na previsão de formação de gelo nas pás utilizando aprendizado de máquina [16] e na previsão de geração de energia por meio de redes neurais profundas [17]. Podemos aplicar técnicas de análise de correlação e redução de dimensionalidade a esses dados auxiliares antes de integrá-los aos modelos.

Apesar desses avanços, ainda existem desafios significativos. A qualidade dos dados SCADA brutos muitas vezes sofre com ruídos e erros, o que exige cuidado no pré-processamento e na limpeza de dados anômalos antes do treinamento [12]. Um dos desafios mais críticos, talvez, seja o desbalanceamento natural dos dados: falhas são eventos muito mais raros que a operação normal. Isso pode enviesar os modelos de ML, resultando em baixa performance na detecção justamente da classe minoritária (as falhas) [18]. Existem diferentes formas de lidar com isso, como abordagens no nível dos dados (sobreamostragem/subamostragem) ou no nível do algoritmo (ajuste de custos, algoritmos específicos). As estratégias no nível do algoritmo costumam ser preferidas pois evitam a necessidade de rotular mais dados, o que é caro [18]. Há estudos que propõem algoritmos, como certas configurações de KNN, que são naturalmente mais robustos ao desbalanceamento, o que evita o uso de técnicas explícitas de balanceamento [19]. A falta de bases de dados públicas, extensas e bem rotuladas ainda é um obstáculo para o desenvolvimento e a validação de novas técnicas [18].

Nossa proposta de iniciação científica busca avançar nesse cenário, investigando a aplicação de algoritmos de aprendizado de máquina para detectar falhas em turbinas eólicas, usando dados públicos SCADA junto com dados meteorológicos. Nosso foco será avaliar o impacto des-

ses dados do tempo na detecção de falhas e comparar o desempenho dos modelos LSTM, CNN e Transformer. Com isso, objetivamos aprimorar a detecção precoce de anomalias e, assim, contribuir para sistemas de monitoramento mais eficazes.

4 Objetivos

O propósito deste trabalho é desenvolver e comparar modelos preditivos de falhas em turbinas eólicas, utilizando dados SCADA e meteorológicos, visando aprimorar as estratégias de manutenção preditiva. Detalhamos a seguir os objetivos gerais e específicos:

Objetivos Gerais

- Fornecer formação e habilidades necessárias para o desenvolvimento da pesquisa: Proporcionar ao aluno as bases teóricas e práticas em aprendizado de máquina, análise de séries temporais e sistemas de energia eólica para conduzir a investigação proposta.
- Contribuir para a melhoria das habilidades de programação: Desenvolver a capacidade de implementar algoritmos eficientes e robustos para análise de dados operacionais de turbinas eólicas e dados meteorológicos, com foco na linguagem Python e suas bibliotecas científicas (e.g., Pandas, Scikit-learn, TensorFlow/PyTorch).
- Desenvolver habilidades sociais para o trabalho em grupo: Promover a colaboração e o trabalho em equipe com orientadores e potencialmente outros pesquisadores, fundamentais para a pesquisa científica.
- Aprimorar as habilidades de escrita científica: Capacitar o aluno na redação de relatórios técnicos e artigos científicos de alta qualidade, comunicando adequadamente os métodos e resultados da pesquisa.

Esses objetivos contribuirão positivamente para o desenvolvimento profissional e acadêmico do candidato, ampliando suas chances de sucesso tanto no mercado de trabalho em áreas de ciência de dados e engenharia, quanto na academia.

Objetivos Específicos

- Estudo sobre aprendizado de máquina necessário ao desenvolvimento da pesquisa: Explorar os conceitos fundamentais e avançados de aprendizado de máquina aplicáveis à análise de séries temporais e detecção de falhas/anomalias em sistemas de engenharia, com foco em turbinas eólicas.
- Estudo de técnicas de tratamento de dados SCADA e meteorológicos: Analisar e aplicar métodos para extração, limpeza, imputação de dados faltantes, tratamento de outliers e sincronização temporal de dados provenientes de múltiplas fontes.
- Estudo de ferramentas para manipulação e modelagem de dados de séries temporais: Investigar e utilizar bibliotecas e frameworks computacionais (e.g., Pandas, NumPy, Scikit-learn, TensorFlow/PyTorch) para o processamento e modelagem eficaz dos dados do projeto.
- Desenvolvimento e implementação de modelos de aprendizado de máquina para detecção de falhas: Implementar e treinar modelos como LSTMs, 1D-CNNs e/ou Transformers para identificar padrões anormais nos dados SCADA que indiquem falhas incipientes.
- Avaliação experimental do impacto de dados exógenos: Conduzir experimentos para comparar o desempenho dos modelos de detecção de falhas treinados apenas com dados SCADA versus modelos treinados com a adição de dados meteorológicos, utilizando métricas de avaliação apropriadas (e.g., Acurácia, Precisão, Revocação, F1-Score, AUC-ROC).
- Escrita de relatórios e artigos para divulgação dos resultados obtidos: Documentar a metodologia, os experimentos realizados e os resultados alcançados, visando a elaboração do relatório final de iniciação científica e, potencialmente, a submissão de um artigo para conferência ou periódico científico.

Desta forma, os objetivos específicos contribuirão para a formação do aluno em técnicas de pré-processamento de dados, modelagem de séries temporais com aprendizado de máquina e avaliação experimental, visando o desenvolvimento de habilidades importantes para a realização deste projeto de pesquisa e para a continuidade da sua carreira acadêmica ou profissional na área de dados e engenharia.

5 Metodologia Proposta

Aqui, detalhamos como pretendemos conduzir o projeto para alcançar os objetivos, cobrindo a revisão da literatura, a seleção e o tratamento dos dados, o desenvolvimento dos modelos e a avaliação experimental.

5.1 Atualização do Referencial Teórico

Manteremos a revisão da literatura como atividade contínua ao longo de todo o projeto. O objetivo é acompanhar os avanços em detecção de falhas em turbinas eólicas, modelos de ML para séries temporais (LSTM, CNN, Transformers), tratamento de dados SCADA e integração de dados externos, garantindo que o projeto utilize as melhores práticas e técnicas disponíveis. Buscaremos também, continuamente, por novas bases de dados públicas que possam ser úteis ao estudo. Essa etapa é vital tanto para a qualidade da pesquisa quanto para a formação do aluno.

5.2 Bases de Dados

Para desenvolver este projeto, utilizaremos bases de dados operacionais de turbinas eólicas disponíveis publicamente, além de dados meteorológicos das localidades ou períodos correspondentes. Nossa fonte principal para os dados operacionais será o repositório de dados SCADA abertos compilado em [5]. Esse repositório reúne diversos conjuntos de dados com variáveis importantes (velocidade do rotor, potência gerada, temperaturas, ângulos de pitch, etc.), geralmente amostradas em intervalos regulares (como 10 minutos). Usar dados públicos não só facilita a reproduzibilidade da pesquisa como também permite comparações com outros trabalhos da área.

Além disso, incorporaremos dados meteorológicos (externos) à análise. Isso nos ajudará a contextualizar a operação das turbinas e, potencialmente, melhorar a detecção de falhas. Esses dados trarão informações como velocidade e direção do vento em diferentes alturas, temperatura do ar, umidade e pressão atmosférica. Para obter esses dados, consideramos fontes como o Centro de Previsão de Tempo e Estudos Climáticos (CPTEC/INPE) [6] e APIs como a do OpenWeatherMap [7]. Ao combinar dados SCADA e meteorológicos, poderemos investigar como as condições ambientais influenciam o desempenho e a "saúde" das turbinas.

5.3 Tratamento dos Dados

Os dados brutos das fontes SCADA e meteorológicas precisam passar por um pré-processamento cuidadoso antes de alimentarem os modelos de aprendizado de máquina. Essa etapa é fundamental para assegurar a qualidade e a consistência dos dados. O tratamento seguirá estas etapas:

1. **Extração e Conversão:** Extrairemos os dados de suas fontes originais (logs, bancos de dados, APIs) e os converteremos para um formato tabular padrão, como CSV, usando bibliotecas como Pandas, para facilitar o acesso pelos modelos.
2. **Limpeza de Dados:** Aplicaremos técnicas para identificar e tratar dados anômalos (*outliers*), possivelmente gerados por erros de sensores ou transmissão. Podemos usar métodos estatísticos (Z-score, IQR) ou baseados em limites operacionais para remover ou ajustar esses pontos.
3. **Imputação de Dados Faltantes:** É comum que dados SCADA e meteorológicos contenham valores ausentes. Empregaremos estratégias para preencher essas lacunas, desde métodos simples (média, mediana) até mais avançados como interpolação temporal (linear, spline) ou modelos preditivos (kNN Imputer, MICE), disponíveis em bibliotecas como Scikit-learn.
4. **Sincronização Temporal:** Precisamos garantir que os dados SCADA e meteorológicos estejam alinhados no tempo. Para isso, os dados serão agregados ou interpolados para uma resolução temporal comum (e.g., 10 minutos, 1 hora), de forma a assegurar que cada registro da turbina tenha as informações meteorológicas do mesmo instante ou intervalo.
5. **Seleção de Variáveis (Feature Selection):** Poderemos usar análise de correlação e/ou técnicas baseadas em modelos para selecionar as variáveis SCADA e meteorológicas mais relevantes para prever falhas. Isso busca reduzir a dimensionalidade e, potencialmente, melhorar o desempenho e a interpretabilidade dos modelos.
6. **Normalização/Escalonamento:** Como as variáveis selecionadas podem ter diferentes escalas e unidades, aplicaremos técnicas de escalonamento (como MinMaxScaler) ou normalização (como StandardScaler) do Scikit-learn para evitar que variáveis de maior magnitude dominem o aprendizado.

Com isso, os dados estarão prontos para a divisão em conjuntos de treinamento, validação e teste, etapas essenciais para desenvolver e avaliar os modelos.

5.4 Desenvolvimento dos Modelos Preditores

Neste projeto, vamos explorar modelos de aprendizado de máquina avançados, que são especialmente adequados para lidar com a natureza sequencial e temporal dos dados SCADA de turbinas eólicas, usando frameworks como TensorFlow/Keras ou PyTorch. Nosso principal objetivo aqui é desenvolver modelos que consigam detectar padrões anormais indicativos de falhas em estágio inicial. As arquiteturas que pretendemos investigar são:

- **Redes Neurais Recorrentes (RNNs), especificamente Long Short-Term Memory (LSTM):** LSTMs são boas para capturar dependências de longo prazo em sequências, algo essencial para entender como o estado de uma turbina evolui. Elas conseguem modelar a dinâmica temporal das variáveis SCADA e identificar desvios sutis do normal [20]. Exploraremos arquiteturas uni e bidimensionais, além de camadas empilhadas (stacked LSTMs).
- **Redes Neurais Convolucionais Unidimensionais (1D-CNNs):** Apesar das CNNs serem famosas pelo processamento de imagens, as 1D-CNNs funcionam bem na extração de características locais em dados sequenciais. Elas conseguem identificar padrões temporais (motifs) relevantes nos sinais SCADA que podem indicar falhas. Muitas vezes, são usadas junto com LSTMs (arquiteturas CNN-LSTM) para capturar tanto características locais quanto dependências de longo prazo [3].
- **Transformers:** Criados inicialmente para processamento de linguagem natural, os Transformers vêm mostrando bons resultados em diversas tarefas com séries temporais. Seus mecanismos de auto-atenção (self-attention) permitem ao modelo ponderar a importância de diferentes partes da sequência de entrada ao fazer uma predição, capturando relações complexas e de longo alcance sem as limitações das RNNs. Seu uso na detecção de falhas em turbinas é um campo de pesquisa ativo [21].

Treinaremos estes modelos para realizar tarefas de classificação (identificar estado normal vs. falha), buscando lidar com o desbalanceamento de classes usando técnicas adequadas (reponderação de classes, over/undersampling, funções de perda específicas). A escolha final dos modelos e suas configurações (hiperparâmetros) vai depender das análises exploratórias e dos resultados nos conjuntos de validação.

5.5 Experimentos e Avaliação dos Resultados

Para avaliar a eficácia dos modelos e o impacto dos dados meteorológicos, faremos um conjunto estruturado de experimentos computacionais.

5.5.1 Experimentos sem Dados Exógenos

Primeiro, treinaremos e avaliaremos os modelos (LSTM, 1D-CNN, Transformer) usando apenas os dados SCADA pré-processados. O objetivo é ter uma linha de base (baseline) de desempenho, baseada unicamente nas informações internas da turbina.

5.5.2 Experimentos com Dados Exógenos

Depois, faremos o mesmo incorporando os dados meteorológicos pré-processados e sincronizados (vento, temperatura, etc.) como entradas adicionais. A ideia é investigar se essas informações contextuais do ambiente ajudam os modelos a distinguir melhor entre variações normais de operação (causadas pelo clima) e anomalias que de fato representam falhas.

5.5.3 Avaliação dos Resultados

O ponto principal será comparar o desempenho dos modelos nos dois cenários (com e sem dados exógenos). Para medir o desempenho na tarefa de classificação, usaremos métricas padrão, calculadas sobre o conjunto de teste:

- **Acurácia (Accuracy):** Percentual geral de acertos.
- **Precisão (Precision):** Das vezes que o modelo previu falha, quantas eram realmente falhas? Importante para evitar falsos alarmes.
- **Revocação (Recall / Sensitivity):** Das falhas que realmente ocorreram, quantas o modelo detectou? Importante para não deixar falhas passarem.
- **Pontuação F1 (F1-Score):** Um balanço entre Precisão e Revocação, útil principalmente com dados desbalanceados.
- **Área sob a Curva ROC (AUC-ROC):** Mede quão bem o modelo separa as classes (normal vs. falha).
- **Matriz de Confusão:** Para ver em detalhes os tipos de erros e acertos.

Além disso, caso a formulação do problema envolva prever valores (para detectar anomalias baseadas no erro dessa previsão), calcularemos também métricas de regressão como Erro Médio Absoluto (MAE) e Raiz do Erro Quadrático Médio (RMSE).

Seguiremos as boas práticas de aprendizado de máquina: dividiremos os dados em conjuntos de treinamento (para ajustar o modelo), validação (para otimizar hiperparâmetros e evitar overfitting) e teste (para uma avaliação final imparcial). Se apropriado, avaliaremos a significância estatística das diferenças encontradas. Essa comparação nos ajudará a concluir sobre o valor agregado das informações meteorológicas para a detecção de falhas em turbinas eólicas com os modelos que propomos.

6 Cronograma

Para o período de 12 meses previsto para esta iniciação científica, planejamos as seguintes atividades principais:

A1 Atualização da Literatura: Atividade contínua para manter o referencial teórico atualizado sobre detecção de falhas em turbinas eólicas, aprendizado de máquina para séries temporais e identificar novas bases de dados e métodos relevantes.

A2 Estudo de Conceitos Básicos de Aprendizado de Máquina: Estudo inicial sobre os fundamentos de aprendizado de máquina necessários para o desenvolvimento do projeto (regressão, classificação, métricas de avaliação).

A3 Estudo de Pré-processamento e Modelos para Séries Temporais: Estudo aprofundado das técnicas de tratamento de dados SCADA e meteorológicos (limpeza, imputação, sincronização) e dos modelos de aprendizado de máquina selecionados (LSTM, 1D-CNN, Transformers) e sua aplicação em séries temporais.

A4 Coleta e Pré-processamento de Dados: Coleta dos dados SCADA de fontes públicas e dados meteorológicos das fontes selecionadas (CPTEC, OpenWeatherMap), seguida pela aplicação das técnicas de tratamento e pré-processamento definidas em A3.

A5 Implementação e Treinamento dos Modelos Preditivos (sem dados exógenos): Desenvolvimento do código, implementação e treinamento dos modelos de ML utilizando apenas os dados SCADA pré-processados.

A6 Implementação e Treinamento dos Modelos Preditivos (com dados exógenos): Adaptação do código e treinamento dos modelos de ML incorporando os dados meteorológicos como variáveis adicionais.

A7 Avaliação e Comparaçāo dos Resultados dos Modelos: Execução dos modelos nos conjuntos de teste, cálculo das métricas de desempenho (Acurácia, Precisão, Recall, F1, AUC-ROC, etc.) e comparação quantitativa dos resultados dos cenários com e sem dados exógenos.

A8 Análise Detalhada e Interpretação dos Resultados: Análise qualitativa dos resultados, interpretação do impacto das variáveis meteorológicas e dos diferentes modelos na capacidade de detecção de falhas.

A9 Preparação de Relatórios e Artigos: Compilação e redação do relatório final de iniciação científica e, se aplicável, preparação de um artigo científico sumarizando a metodologia, resultados e conclusões do projeto.

O Cronograma de desenvolvimento proposto, distribuído em bimestres ao longo de 12 meses, é apresentado na Tabela 1.

Atividade	Bimestres					
	1º	2º	3º	4º	5º	6º
A1 (Literatura)	•	•	•	•	•	•
A2 (ML Básico)	•	•				
A3 (Estudo Pré-proc/Modelos)	•	•	•			
A4 (Coleta/Pré-processamento)		•	•			
A5 (Implement./Treino S/ Exóg.)		•	•	•		
A6 (Implement./Treino C/ Exóg.)			•	•	•	
A7 (Avaliação/Comparaçāo)				•	•	•
A8 (Análise/Interpretação)					•	•
A9 (Relatórios/Artigos)				•	•	•

Tabela 1: Plano de Atividades do Projeto

7 Sobre o Candidato

O candidato cursa o Bacharelado em Ciência e Tecnologia, com ênfase em Engenharia da Computação, desde 2024. Já completou disciplinas relevantes para o projeto, como Lógica de Programação, Algoritmos e Estruturas de Dados I e Fenômenos Mecânicos. Atualmente, cursa disciplinas que aprofundam seu conhecimento, como Algoritmos e Estruturas de Dados II e Fenômenos do Contínuo. Seu Coeficiente de Rendimento (CR) atual é de 7.809, o que reflete seu comprometimento com os estudos. Por fim, é importante destacar que o candidato tem participado do grupo de pesquisa do Prof. Marcos Quiles desde Agosto/2024, como voluntário. Neste período, adquiriu as habilidades necessárias para dar início a este projeto de pesquisa.

8 Infraestrutura Computacional

Desenvolveremos este projeto utilizando a infraestrutura computacional disponível no Instituto de Ciência e Tecnologia (ICT) da UNIFESP, com a possibilidade de usar também recursos do Centro de Inovação em Novas Energias (CINE) se precisarmos de maior poder computacional.

Temos acesso à seguinte infraestrutura:

- **Estações de Trabalho e Servidores Locais:** Contamos com 5 estações de trabalho equipadas com GPU RTX 4060 no ICT/UNIFESP, adequadas para desenvolver código, pré-processar dados e treinar modelos de machine learning de complexidade moderada. Uma Workstation com dual GPU NVIDIA RTX A6000 (totalizando 96GB de VRAM) está disponível, o que é ideal para acelerar o treinamento de redes neurais profundas;
- **Acesso a Recursos de HPC:** Para modelos com grande número de parâmetros ou grande volume de dados, podemos explorar os recursos de computação de alto desempenho disponíveis no CINE, como o Cluster ENIAC instalado no IQSC/USP.

Essa combinação de recursos nos dá a capacidade técnica necessária para todas as etapas do projeto, desde a coleta e tratamento de grandes volumes de dados até o treinamento e a avaliação rigorosa de modelos complexos de aprendizado profundo.

Referências

- [1] SASINTHIRAN, A.; SAKTHIVEL, G.; RAGALA, R. A review of artificial intelligence applications in wind turbine health monitoring. *Wind Engineering*, 2024.
- [2] MURALIDHARAN, P. M.; THAKUR, G.; SHALINI, M.; SHARMA, V.; ABOOTHARMAHOODSHA-KIR; DHABLIA, A. A review on condition monitoring of wind turbines using machine learning techniques. In: . c2024. v. 540. p. 03003.
- [3] GEBREAMLAK, B. *Hybrid CNN-integrated LSTM for fault detection and diagnosis of wind turbines*. May 2024. Dissertação (Mestrado em Física) - School of Computing, University of Eastern Finland, May 2024. Accessed: 2025-04-05.
- [4] BABU, G. V. R.; PORLEKAR, S. B.; ALI, H. M.; VUYYURU, V. A.; AL ANSARI, M. S.; RAJ, I. I. Dynamic fault diagnosis in wind turbines: A GNN-LSTM approach, 2024.
- [5] sltzgs. Wind_Turbine_SCADA_open_data: A collection of open data for wind turbine scada systems. GitHub repository. Accessed: 2025-04-01.
- [6] Centro de Previsão de Tempo e Estudos Climáticos (CPTEC) / Instituto Nacional de Pesquisas Espaciais (INPE). CPTEC/INPE - centro de previsão de tempo e estudos climáticos. Website. Accessed: 2025-04-01.
- [7] OpenWeatherMap. Weather API. Website API Documentation. Accessed: 2025-04-01.
- [8] GONZALEZ, E.; STEPHEN, B.; INFIELD, D.; MELERO, J. J. On the use of high-frequency SCADA data for improved wind turbine performance monitoring. *Journal of Physics*, 2017.
- [9] CHOI, K.; YI, J.; PARK, C.; YOON, S. Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines. *IEEE Access*, v. 9, 2021.
- [10] TUMMALA, A.; VELAMATI, R. K.; SINHA, D. K.; INDRAJA, V.; SHEKHAR, V. A review on small scale wind turbines. *Renewable and Sustainable Energy Reviews*, v. 56, p. 1351–1371, 2016.
- [11] CRABTREE, C. J.; ZAPPALÀ, D.; TAVNER, P. J. Survey of commercially available condition monitoring systems for wind turbines. Technical report for supergen wind energy technologies consortium, Durham University, School of Engineering and Computing Sciences, 2011.
- [12] MA, J.; YUAN, Y. Application of SCADA data in wind turbine fault detection – a review. *Sensor Review*, v. 42, n. 6, p. 509–525, 2022.
- [13] NG, E. Y. K.; LIM, J. T. Machine learning on fault diagnosis in wind turbines. *Fluids*, v. 7, n. 12, p. 371, 2022.
- [14] XIAO, C.; LIU, Z.; ZHANG, T.; ZHANG, X. Deep learning method for fault detection of wind turbine converter. *Applied Sciences*, v. 11, n. 3, p. 1280, 2021.
- [15] SCHLECHTINGEN, M.; SANTOS, I. F.; ACHICHE, S. Wind turbine condition monitoring based on SCADA data using normal behavior models. *Applied Soft Computing*, v. 13, n. 1, p. 259–270, 2013.

- [16] KREUTZ, M.; ALLA, A. A.; VARASTEH, K.; OELKER, S.; GREULICH, A.; FREITAG, M.; THOBEN, K. Machine learning-based icing prediction on wind turbines. *Procedia CIRP*, 2019.
- [17] LIN, W.-H.; WANG, P.; CHAO, K.-M.; LIN, H.-C.; YANG, Z.; LAI, Y.-H. Wind power forecasting with deep learning networks: Time-series forecasting. *Applied Sciences*, 2021.
- [18] OLIVEIRA-FILHO, A.; COMEAU, M.; CAVE, J.; NASR, C.; CÔTÉ, P.; TAHAN, A. Wind turbine SCADA data imbalance: A review of its impact on health condition analyses and mitigation strategies. *Energies*, 2024.
- [19] FAZLI, A.; POSHTAN, J. Wind turbine fault detection and isolation robust against data imbalance using KNN. *IET Renewable Power Generation*, 2024.
- [20] AFRASIABI, S.; AFRASIABI, M.; PARANG, B.; MOHAMMADI, M.; KAHOURZADE, S.; MAHMOUDI, A. Two-stage deep learning-based wind turbine condition monitoring using SCADA data. *IEEE Access*, 2020.
- [21] GUO, J.; WANG, Z.; LI, H.; SOARES, C. G. A transformer-based fault detection framework for offshore wind turbines based on SCADA data. In: *Advances in Maritime Technology and Engineering*. Boca Raton, FL: CRC Press / Taylor & Francis, 2024. Accessed: 2025-04-05.