energies

MDPI

Review

Wind Turbine SCADA Data Imbalance: A Review of Its Impact on Health Condition Analyses and Mitigation Strategies

Adaiton Oliveira-Filho [1,*], Monelle Comeau [2], James Cave [1], Charbel Nasr [1], Pavel Côté [2] and Antoine Tahan [1,*]

[1] Department of Mechanical Engineering, École de Technologie Supérieure, Université du Québec, 1100 Rue Notre Dame O, Montreal, QC H3C 1K3, Canada
[2] Power Factors, 7005 Boulevard Taschereau, Brossard, QC J4Z 1A7, Canada
* Correspondence: adaitonfilho@icloud.com (A.O.-F.); antoine.tahan@etsmtl.ca (A.T.)

Abstract: The rapidly increasing installed capacity of Wind Turbines (WTs) worldwide emphasizes the need for Operation and Maintenance (O&M) strategies favoring high availability, reliability, and cost-effective operation. Optimal decision-making and planning are supported by WT health condition analyses based on data from the Supervisory Control and Data Acquisition (SCADA) system. However, SCADA data are highly imbalanced, with a predominance of healthy condition samples. Although this imbalance can negatively impact analyses such as detection, Condition Monitoring (CM), diagnosis, and prognosis, it is often overlooked in the literature. This review specifically addresses the problem of SCADA data imbalance, focusing on strategies to mitigate this condition. Five categories of such strategies were identified: Normal Behavior Models (NBMs), data-level strategies, algorithm-level strategies, cost-sensitive learning, and data augmentation techniques. This review evidenced that the choice among these strategies is mainly dictated by the availability of data and the intended analysis. Moreover, algorithm-level strategies are predominant in analyzing SCADA data because these strategies do not require the costly and time-consuming task of data labeling. An extensive public SCADA database could ease the problem of abnormal data scarcity and help handle the problem of data imbalance. However, long-dated requests to create such a database are still unaddressed.

Keywords: wind turbine; SCADA data; imbalanced data; normal behavior model; data augmentation techniques

Academic Editor: Francesco Castellani

Received: 4 November 2024
Revised: 17 December 2024
Accepted: 25 December 2024
Published: 27 December 2024

Citation: 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* **2025**, *18*, 59. https://doi.org/10.3390/en18010059

1. Introduction

The global Wind Turbine (WT) installed capacity is rapidly increasing, driven by government and private investments to decarbonize the energy sector. The technological maturity of WT design and manufacturing has been pushing down their Levelized Cost of Energy (LCOE), thus contributing to consolidating this renewable source in the global energy mix [1]. Enhanced Operation and Maintenance (O&M) strategies contribute to achieving high availability and reliability and an overall cost-effective operation. Manufacturers, operators, analysts, and maintenance practitioners work to ensure WTs fulfill their objectives correctly and in the long term [2].

Optimal O&M strategies include decision-making and planning guided by the continuous monitoring of the WT health condition. These strategies aim at limiting abnormal or faulty conditions and, as a result, minimize production losses and maintenance expenditures. This subject is within the scope of Performance Monitoring (PM) [3], Condition Monitoring (CM) [4], detection and diagnosis of abnormal conditions [5], Condition Based Maintenance (CBM) [6], and Prognosis and Health Management (PHM) [7]. These analyses

of a system's health condition involve identifying and characterizing abnormal, degraded, or faulty conditions.

The WT health condition analyses aim to provide actionable information from measured data, especially from the Supervisory Control and Data Acquisition (SCADA) system [8,9]. Modern WTs' SCADA system embeds dozens of sensors measuring geometrical, kinematic, thermal, and electrical variables [10]. SCADA data-based analyses of the WT health condition have an intrinsic and sometimes overlooked challenge: the imbalance of datasets, with significantly more data representing the healthy condition than abnormal or degraded states. This imbalance toward the healthy condition is related to the high availability and reliability of modern WTs, with typical availability ranging over 95% [11].

The SCADA data imbalance can hinder data-driven analyses, as highlighted in the review papers listed in Table 1. These reviews highlight that classical imbalance-unaware approaches tend to be biased toward the majority class (healthy data), leading to poor representation of the minority classes (degraded or faulty conditions). According to reviews from Table 1, the SCADA data imbalance is due to multiple factors. Moreover, addressing the imbalanced SCADA data is necessary to improve the accuracy of wind turbine data-based fault detection, diagnosis, and prognosis.

Table 1. Articles and reviews highlighting the imbalanced Wind Turbine (WT) Supervisory Control and Data Acquisition (SCADA) data *.

Ref.	Author, Year	Title	Mention to the Problem of SCADA Data Imbalance
[8]	Pandit and Wang, 2024	A comprehensive review on enhancing WT applications with advanced SCADA data analytics and practical insights.	(a) *The review recognizes the limitations caused by the imbalance between normal operation data samples and abnormal data samples, negatively impacting model accuracy.* (b) *While data-driven machine learning algorithms offer closer alignment with actual wind turbine operation, accuracy is limited due to the lack of public data and imbalanced datasets.*
[12]	Maldonado-Correa et al., 2024	Classification of highly imbalanced SCADA data for fault detection of WT generators.	*Imbalanced distributions exist across the board between abnormal and normal classes, leading to inaccurate failure diagnoses and predictions because these models tend to be biased toward the most widespread class.*
[7]	Cuesta et al., 2024	Challenges on prognostics and health management for WT components.	(a) [High imbalance] *between healthy and unhealthy data is one of the problems to be solved urgently in the research of equipment life prediction.* (b) *PHM often deals with a large volume of data gathered under healthy conditions and a limited number of data points that indicate faulty states.*
[13]	Ma and Yuan, 2023	Application of SCADA data in WT fault detection—A review.	(a) *Because of the imbalance of fault categories and the scarcity of fault data in SCADA data, it is challenging to extract fault features accurately.* (b) *The scarcity and imbalance of fault data make it infeasible to train classification models using SCADA fault data.* (c) *The process of labeling SCADA data in engineering practice is a very time-consuming and error-prone task, which can lead to an imbalanced number of labeling categories.*
[9]	Badihi et al., 2022	A comprehensive review on signal-based and model-based CM of WTs: Fault diagnosis and lifetime prognosis.	(a) *The distribution of SCADA data is generally imbalanced, and anomalous data mining is usually insufficient.* (b) [The data imbalance might imply] *poor CM performance since the data-driven models tend to be biased toward the majority class.*

Table 1. *Cont.*

Ref.	Author, Year	Title	Mention to the Problem of SCADA Data Imbalance
[14]	Nunes et al., 2021	Use of learning mechanisms to improve the CM of WT generators: A review.	(a) *It is challenging for the classifier to learn abnormal behavior when the representative part of the data consists of non-fault samples.* (b) This review discusses the suitability of multiple approaches to analyze imbalanced SCADA data.
[15]	Stetco et al., 2019	Machine learning methods for WT CM: A review.	(a) [Labeling of the training data] *is time-consuming, error-prone and likely to result in a set of labeled vectors with an imbalanced number of classes. This is a common issue in practice.* (b) *There are several ways that the problem of imbalanced classes has been addressed, including under-sampling, oversampling, SMOTE, and Tomek-links.*
[16]	Chen et al., 2019	Learning deep representation of imbalanced SCADA data for fault detection of WTs.	(a) *The phenomenon of data imbalance is ignored and the information of abnormal SCADA data is discarded in the process of feature extraction. This problem makes diagnosis results biased toward the majority class, while the ability of novelty detection is fairly weak.* (b) *There are few studies about learning deep feature representation based on imbalanced SCADA data.* (c) *Data imbalance brings many obstacles to fault diagnosis of WTs when it comes to massive SCADA data. However, this problem is usually ignored for simplifying analysis.*
[17]	Helbing and Ritter, 2018	Deep Learning for fault detection in WTs.	(a) *Imbalanced datasets arise from the fact that major faults that cause more than one day of downtime represent only 25% of all failures, but account for 95% of the downtime.* (b) *It is of predominant interest to predict the major faults that usually occur rarely in any one WT.*

* This is a non-exhaustive list of review articles. Direct citations are italicized.

Supervised and unsupervised models require different strategies when dealing with imbalanced data. Indeed, supervised learning requires a deeper characterization of the data before the model training, while unsupervised models integrate some level of representation learning [18]. Pandit et al. [19] highlight that most of the data-driven techniques for WTs are regression and classification, which are supervised learning models. Nevertheless, labeling complex data is costly, time-consuming, and influenced by experts' sensitivity. Labeling the SCAD data is particularly challenging due to its high dimensionality and the large number of WT operational states [20]. Emerging research on unsupervised learning models aims at overcoming the need for SCADA data labeling [21].

The research addressing the impact of SCADA data imbalance in data-based approaches is still limited and often overlooked. To our knowledge, no previous literature review has specifically addressed the problem of SCADA data imbalance in the context of WT health condition analysis. Filling this gap is the primary motivation for the present review. It also includes the following contributions:

- Characterization of WT SCADA data imbalance and its impact on various WT health condition analyses such as detection, CM, diagnosis, and prognosis;
- Review of the strategies used to deal with data imbalance in a general context;
- Presentation of previous WT health condition analyses. The reviewed papers are organized according to the strategies to address the SCADA data imbalance.

The present review is organized as follows: Section 2 presents the data imbalance problem in a general framework. Then, Section 3 characterizes specifically the SCADA data imbalance. Section 4 reviews the methods dealing with SCADA data imbalance. Section 6 summarizes the findings and outlines future directions in WT health condition analyses.

2. The Class Imbalance Problem Across Multiple Application Domains

The imbalance of datasets impacts diverse application domains. Strategies to deal with data imbalance constitute a very active research topic, mainly because of the recurrence of this condition in real-world problems and its impacts. Data imbalance has implications in tasks such as anomaly detection, classification, diagnosis, and prognosis [22–30].

2.1. Characterization of Imbalanced Datasets

Characterizing the severity of the imbalance is important for choosing suitable methods and correctly evaluating its performance. The poor performance of imbalance-unaware classification approaches is due to three main factors: (i) Lack of minority class data when the corresponding datasets are too small, (ii) overlapping between majority and minority classes, and (iii) minority classes with complex behavior [24].

The class imbalance can be in a two-class problem or a multi-class problem, as depicted in Figure 1b and Figure 1c, respectively. The latter eventually includes different levels of data imbalance, as depicted in Figure 1d. The imbalance ratio measures the level of class imbalance. It can be defined as the ratio of the number of data points in the majority class to those in the minority class (majority class number of points:minority class number of points) or the ratio between the datasets' duration (duration covered by the majority class dataset:duration covered by the minority class dataset) [24,26].

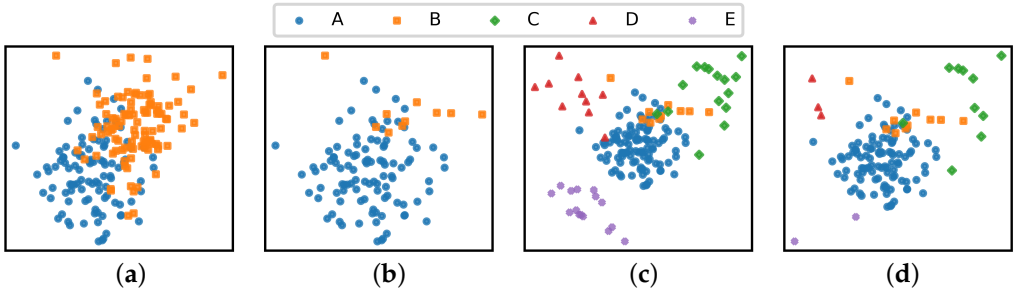

Figure 1. Configurations of datasets: (**a**) Balanced two-class datasets; (**b**) imbalanced two-class datasets, with imbalance ratio of 5:1; (**c**) imbalanced multi-class datasets, with imbalance ratio of 10:1 for each minority class; (**d**) imbalanced multi-class datasets, including imbalance within the minority classes, with an imbalance ratio of 25:5:5:1:1.

The literature distinguishes two levels of information in multiple class systems: within-class information and between-class information [16]. Within-class information refers to the distribution of data points belonging to each particular class, be it the healthy condition dataset or a degraded condition dataset. Such characterization involves understanding the system behavior within a given class. On the other hand, between-class information is about the relationships and differences between classes. The clustering of different datasets can give indications of relative dispersion and eventual intersections between the corresponding classes. The scarcity of the minority class can hinder the characterization of within-minority-class information. Assuming enough points are available for each dataset of interest, the issue of data imbalance impacts mainly the between-class information [16].

2.2. Categories of Strategies for Addressing Data Imbalance

Multiple strategies aim to address data imbalance, notably Normal Behavior Models (NBMs), data-level strategies, algorithm-level strategies, cost-sensitive learning, data augmentation techniques, and transfer learning. Table 2 describes the different classes of balancing strategies and their advantages and disadvantages. The difference between these strategies mostly relates to how they tackle between-class and within-class information [24].

Table 2. Comparison of strategies to address data imbalance across multiple application domains.

Category and References	Description	Advantages (A) and Disadvantages (D)
Normal Behavior Model (NBM) [31,32]	NBMs circumvent the problem of data imbalance by focusing on the characterization of the healthy condition. NBMs are based on one-class (binary) classifiers or regressive models.	(A) Requires healthy condition data only and admits unsupervised learning. (D) Limited outcome (healthy/unhealthy), with no further characterization of the unhealthy condition.
Data-level (External) [33,34]	Adapts sampling strategies to balance datasets, most often by oversampling the minority class, undersampling the majority class, or a combination of these.	(A) Versatile, conventional classifiers can be used after balancing the datasets. (D) May introduce bias into the representation.
Algorithm-level (Internal) [16]	Learns directly from the imbalanced data by prioritizing or forcing the learning of information from the minority class. This training requires appropriate definitions of the model architecture, loss function, and learning hyperparameters.	(A) Does not change the original data distribution since it works directly with imbalanced data. (D) More complex to build, tends to be problem-specific, and may require domain expertise and insight into why classical classifiers underperform.
Cost-sensitive Learning (Hybrid) [24]	Assigns different weights to classes to prioritize the minority class during learning. These include cost-sensitive neural networks and ensemble classifiers.	(A) Hybrid solution that can outperform data-level and algorithm-level models. (D) Requires defining a custom misclassification loss function.
Data augmentation techniques [35,36]	Uses available information to generate new data points. These include generative models based on the Variational Autoencoder and synthetic data created from simulation.	(A) Overcomes data scarcity; suitable for newly commissioned systems with limited data. (D) More complex and costly; requires modeling and validation before classification.
Transfer learning [5]	Adapts available knowledge from general or previous cases to new instances with small data sets.	(A) Overcomes data scarcity; can be progressively enhanced with new data. (D) More complex and costly to implement.

The choice of an adequate strategy to deal with the data imbalance requires assessing the properties of the interest datasets and delimiting the goal of the analysis. The possibility of labeling an imbalanced dataset is an important criterion. NBMs, data-level approaches, and cost-sensitive learning might require labeled datasets. An unsupervised approach should be prioritized if data labeling costs are impeditive for a given system or application. Section 4 of the present review analyzes the strategies addressing the SCADA data imbalance.

2.3. Performance Metrics for Imbalanced Data

The problem of data imbalance requires special attention when selecting classification performance metrics. Model training based on classical evaluation criteria such as accuracy can lead the model to ignore the minority class entirely. For instance, classification accuracy may be high even when the minority class is completely misclassified. For example, given a two-class dataset with a 1:100 imbalance ratio, a classifier would obtain a 99% accuracy by simply categorizing all instances as the majority class.

Modeling imbalanced datasets requires performance metrics focusing on the minority class, typically precision, recall, F1-score, and negative predictive value [37]. The classification task is usually performed on an extensive set of reference case studies to evaluate its performance. In the confusion matrix, each classification is compared against the actual class and counts as true positive TP, true negative TN, false positive FP, or false negative FN. The main performance metrics are presented below for context [37].

- Accuracy (acc):

$$acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

- Precision, also known as the positive predictive value (ppv):

$$ppv = \frac{TP}{TP + FP} \tag{2}$$

- Recall, also referred to as sensitivity or true positive rate (tpr):

$$tpr = \frac{TP}{TP + FN} \tag{3}$$

- The F_1-score balances precision and recall, and is generalized by the F_β score given by Equation (4), where $\beta \in [0, 1]$ weights the importance between precision and recall. The choice of $\beta = 1$ (F_1-score) gives equal importance to the two metrics.

$$F_\beta = \frac{(1 + \beta^2) \cdot ppv \cdot tpr}{\beta^2 \cdot ppv + tpr} \tag{4}$$

- Negative Predictive Value (npv), as given in Equation (5)

$$npv = \frac{TN}{TN + FN} \tag{5}$$

Other metrics can be found in the literature addressing imbalanced datasets. The Receiver Operating Characteristic curve (ROC) plots the True Positive Rate (y-axis) against the False Positive Rate (x-axis). The Area Under Curve (AUC) derives a scalar metric for the ROC plot [38]. The Geometric mean (G-mean) evaluates the balance between classification sensitivity and specificity [12]. The Matthews Correlation Coefficient (MCC) is a robust metric that takes all elements of the confusion matrix into account [12].

3. SCADA Data Imbalance and Its Impact on Wind Turbine Health Condition Analyses

Obtaining actionable insights from the raw SCADA data involves multiple steps, including data acquisition, preprocessing, and handling of its imbalanced nature.

3.1. SCADA Data Acquisition and Preprocessing

The SCADA system consists of multiple sensors measuring geometrical, kinematic, thermal, and electrical variables. The set of sensors includes anemometers, wind vanes, pressure sensors, RPM sensors, voltage sensors, current sensors, power sensors, and position sensors, as well as temperature sensors positioned in various locations: ambient temperature, nacelle, main bearing, gearbox, generator, generator bearings, cabinets, cooling systems, pitch systems, blades, and the tower. The SCADA data can be used to analyze the WT health condition because degradation modes on WT components have signatures in the SCADA measures. Table 3 indicates potentially informative measures for some kinds of WT degradation.

Table 3. Degraded conditions and interest SCADA variables.

Degraded Condition	Measures of Interest *	Refs.
Main bearing overtemperature	Main bearing temperature [°C] and rotor rotational speed [rpm]	[39,40]
Gearbox overtemperature	Gearbox bearings and gearbox cooling system temperature [°C]	[41]
Generator overtemperature	Generator bearings and generator cooling system temperature [°C]	[42]
Ice accretion on blades	Wind speed [m/s], active power [kW], ambient temperature [°C]	[43,44]
Yaw system misalignment	Yaw angle [°], nacelle position [°], wind direction [°]	[45]
Pitch system failure	Blade pitch angle [°], pitch motor current [A], actuator status	[46,47]
Rotor imbalance	Rotor speed [rpm], active power [kW], tower acceleration [m/s^2]	[48]
Hydraulic system failure	Hydraulic pressure [bar], hydraulic circuit temperature [°C]	[49]
Anemometer failure	Wind speed [m/s], active power [kW], wind farm wind speed [m/s]	[50]
Brake system degradation	Hydraulic pressure [bar], rotor rotational speed [rpm], brake status	[51]

* Wind speed [m/s] and active power [kW] are measures of interest for all degraded conditions; all conditions may use textual information from the SCADA log files.

The SCADA system aggregates measures acquired by its sensors as float-values time series with a 10 min time step. Each value is the mean of the measured signal over 10 min intervals. The 10 min aggregation industrial standard suits performance monitoring of WTs and spare storage and computational resources [3]. Eventually, the SCADA system also stores other 10 min statistics such as maximum value, minimum value, and standard deviation [52].

Moreover, the SCADA system produces log files listing messages about the WT status, including error messages, exceptions to nominal functioning, control protocol messages, fault codes, warnings, and alarms. These SCADA status messages are produced mostly by built-in threshold-based CM and safety protocols. For example, persistent overtemperature at critical components (e.g., main bearing, gearbox, generator, critical cooling systems) could trigger a shutdown protocol, and the SCADA system would record this succession of events in the SCADA log files [53]. In addition, O&M reports manually filled by maintenance practitioners are available for some wind farms. The existing entries can complement status information from the SCADA log files.

For a given WT, the availability of SCADA data depends on the sensors embedded within the SCADA system and is subjected to contractual constraints between the WT manufacturer and operators. The SCADA time series may be accessed by analysts directly from the SCADA system but are more likely provided once the data have flowed through different steps of the data pipeline. The SCADA system time series can present data quality issues due to sensor defaults and data transmission issues. In practice, a combination of preprocessing steps is implemented to eliminate these data quality issues and prepare the data for the targeted approaches. Different kinds of analysis require different preprocessing steps [9]. Selecting appropriate preprocessing steps is paramount, given it can influence the health analysis outcome [54].

SCADA data preprocessing steps include filtering out inconsistent values, filtering specific operational conditions, normalization, correlation analysis, feature engineering, data imputation, sliding overlapping time segmentation, and labeling [8]. Filtering out inconsistent physical measures, non-numerical values (*NaN* reading), and nonexistent entries (*None*) is among the first of the SCADA data preprocessing steps. Typically, the percentage of SCADA data concerned by this step remains below 3% [54]. Normalization techniques should be chosen according to the modeling approach. For example, Min-Max normalization is suitable for Deep Neural Networks (DNN) models. Statistics can be performed over large SCADA databases to estimate lower and upper bounds for the normalization. Table 4 illustrates such a definition for a large North American wind farm comprising over a hundred 1.84 MW-rated power WTs [55].

Table 4. SCADA features with Lower Bound (LB) and Upper Bound (UB) for a large wind farm.

Measure	Symbol	LB	UB
Wind speed	WS [m/s]	0	31
Rotor speed	n_{ROTOR} [rpm]	0	18
Power output	P [kW]	0	2000
Ambient temperature	T_{AMB} [°C]	−25	45
Nacelle temperature	T_{NAC} [°C]	−20	70
Bearing temperature	T_{BEA} [°C]	−20	70
Gearbox bearing temperature	$T_{GBX-BEAR}$ [°C]	0	100
Gearbox oil temperature	$T_{GBX\text{-}OIL}$ [°C]	0	100
Generator temperature	T_{GENi} [°C]	−10	140
Generator cooling temperature	$T_{GEN\text{-}COOL}$ [°C]	−10	120
Axial box temperature	$T_{AX\text{-}BOX}$ [°C]	0	60
Battery box temperature	$T_{BAT\text{-}BOX}$ [°C]	0	45

Correlation analysis and feature engineering allow for defining features with high informative power for a given analysis [56]. Works from diverse domains report that selecting highly informative features enhances the overall performance of DNN models [57].

Analyses based on time series often consider run-to-failure scenarios. In such cases, depending on the hypothesis for the degradation mode, data imputation can be used to complete missing data; sliding overlapping time segmentation allows for producing samples [58]. On the other hand, in analyses based on models trained with datasets regardless of the timeline, it is common practice to filter out data points corresponding to the WT at conditions that bring very little or no information, such as standby and shutdown modes [55]. The WT standby mode may be due to wind speed outside the operational interval or because of grid control requirements. The WT shutdown protocol can be triggered for reasons such as extreme weather conditions and inside-nacelle interventions [59].

Approaches comprising supervised or semi-supervised learning require labeling the datasets. Information from the SCADA log files and O&M reports can guide the labeling of datasets. Indeed, these textual data identify abnormal operating conditions with the respective time intervals. Data points and time series can be selected within the reported intervals.

3.2. Imbalanced SCADA Data

The SCADA data imbalance can be analyzed at three levels: imbalance between healthy and degraded datasets, imbalance between different degradation classes, and imbalance due to the scarcity of degraded data points in newly commissioned wind farms.

The predominance of the healthy condition data over degraded condition data is a positive outcome, as it indicates that wind turbines operate mainly in healthy conditions. The high cost of wind turbines makes preventive or systematic maintenance strategies better suited than curative repairs. In practice, this choice of more conservative maintenance strategies implies that any confident evidence of abnormal conditions would motivate the operators to take prompt action, e.g., curtail or shut down the WT, making run-to-failure instances rare.

SCADA data are also imbalanced within the minority classes, as most data points for cost-critical issues relate to only a few abnormal conditions [17,60]. Based on an extensive database with a total of 35,000 WT component failure events from over 13 years of operation of 1400 large onshore pitch-controlled WTs, Santelo et al. [61] estimated that 80% of maintenance costs are associated with just 20% of the components.

The database's acquisition period influences the class imbalance, with shorter acquisition periods potentially lacking data points for most of the degraded conditions. This shortcoming might limit some analyses on newly commissioned wind farms. The acqui-

sition of data within the wind farm lifetime allows for the characterization of diverse operation conditions, including much less frequent degradation modes. In practice, to analyze a newly commissioned wind farm, operators and analysts can overcome the scarcity or absence of data by analyzing the historical degradation of the specific WT model or a similar configuration [17]. For instance, the health condition analyses might target recurrent degradation conditions within the historical data, which indicates flaws in the WT model [62].

The combination of the factors mentioned in this section implies the imbalance ratio of SCADA datasets depends on each wind farm's characteristics and operational history. To illustrate the order of imbalance ratio, we considered SCADA data covering two years of operation of the abovementioned North American wind farm [55]. Table 5 gives the statistics of SCADA alarms concerning three critical components, the gearbox, the generator, and the main bearing. For this order estimation, the wind farm availability rate of 97% leads to the approximate average period in the healthy condition of $0.97 \times 365 \times 24$ h ≈ 8500 h. The imbalance ratio is estimated with respect to the periods, i.e., period of availability: period with active degradation SCADA alarm.

Table 5. Estimation of the order of SCADA data imbalance for an operating wind farm.

Attribute	Gearbox Degradation	Generator Degradation	Main Bearing Degradation
Frequency of SCADA Alarms (1/year/WT)	4.5	6.8	2.5
Period with Active SCADA Alarms (hours/year/WT)	23.6	17.3	2.9
Order of Imbalance Ratio	360:1	490:1	2930:1

In Table 5, the gearbox degradation includes the SCADA alarms of *Gearbox bearing overtemperature*, *Gearbox oil overtemperature*, *Gearbox oil overtemperature from thermal switch*, and *Gearbox oil pressure too low*. Generator degradation includes *Generator bearings overtemperature*, *Generator stator windings overtemperature*, *Generator over-speed*, *Generator brush wear shutdown*, and *Generator cooling air overtemperature*. Bearing degradation corresponds to the SCADA alarm *Shaft bearing overtemperature*.

4. Review of Strategies for Handling SCADA Data Imbalance

The present review investigates WT health condition analyses, focusing on papers addressing the imbalance of SCADA data. An exploratory approach with structured scoping principles was adopted to identify and compare strategies to address the SCADA data imbalance. The methodological workflow presented below aimed at balancing systematic search strategies with the flexibility to uncover citation connections in the recent literature. The selection of interest papers follows the criteria C1–C3:

C1. Query Structure:

("wind turbine" OR "wind energy") AND ("SCADA")
AND [data]("imbalance" OR "disbalance" OR "unbalance").

In this query, the specifier "[data]" ensures the selection of instances of the words "imbalance," "disbalance," or "unbalance" related to data, thereby excluding terms like "rotor imbalance" and "blade imbalance."

C2. Time Frame: This review considers papers published within the period (2019–2024), therefore, focusing on recent advancements.

C3. Exclusion Criteria: Review and conference papers were excluded to avoid redundancy and prioritize citation relation between papers.

The methodological workflow comprises two steps. First, a targeted bibliographic search step within Web of Science (WoS) and Google Scholar aims to select the initial corpus of papers. The second step uses the citation-oriented exploratory tool ResearchRabbit [63] to broaden and consolidate the corpus of papers. This tool exploits the citation links between papers (paper A cites paper B) to highlight the relation between works. Papers were screened within the ResearchRabbit categories of "Similar Work" and "Earlier Work" [63], and the ones with multiple citation links in the citation graph verifying the criteria C1–C3 were retained

The first step of the bibliographic research led to a primary corpus of 30 papers, 9 from WoS and 21 pertinent papers retained from Google Scholar. This corpus was then broadened and consolidated using ResearchRabbit, resulting in the selection of 56 papers reviewed in the present section. Figure 2 depicts the graph representation provided by ResearchRabbit.

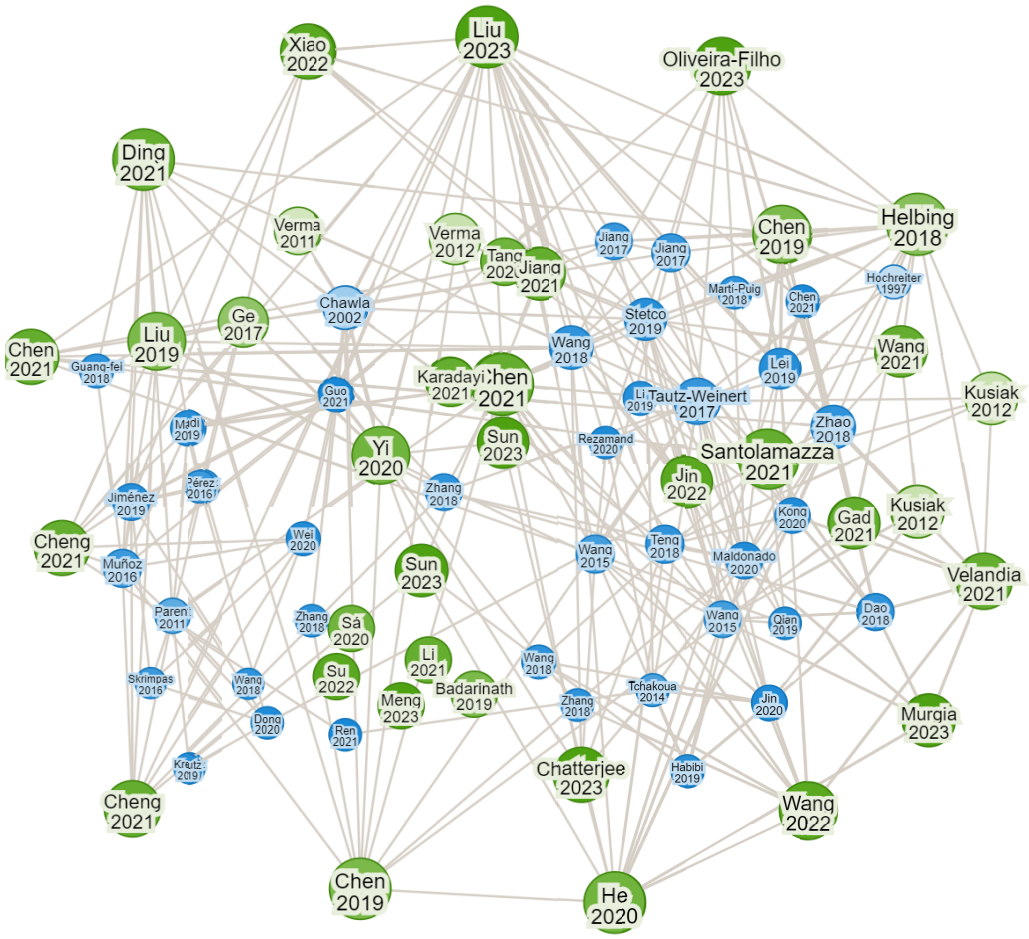

Figure 2. ResearchRabbit graph representation indicating the sets of selected papers (green) and similar papers (blue). Edges represent citation relationships [63].

This hybrid methodology is oriented by relevance and connectivity. It allowed for uncovering and comparing different strategies used in SCADA-based WT health condition analyses. However, this review does not comply with the PRISMA framework for systematic reviews and does not aim for exhaustive coverage [64,65].

The present review presents the papers within five categories: NBMs (Section 4.1), data-level methods (Section 4.2), algorithm-level methods (Section 4.3), hybrid methods (Section 4.4), and data augmentation techniques (Section 4.5). Section 5 discusses the strategies for managing the SCADA data imbalance.

4.1. Normal Behavior Models

Table 6 lists papers using NBMs on SCADA data-based WT health condition analyses.

Table 6. NBMs are based solely on healthy data, thus avoiding the problem of data imbalance.

Ref.	Author(s), Year	Focus	Description and Contributions
[66]	Murgia et al., 2023	Detection of temperature-related WT faults	Regression-based NBMs predicting temperature measures at multiple components are proposed. Healthy data are collected from WTs based on the O&M reports.
[67]	Bilendo et al., 2023	WT CM considering multiple abnormal conditions	An NBM targeting multiple SCADA measures is proposed. It is based on the heterogeneous stacked regression algorithm. Historical normal data are gathered using kernel density estimation (KDE) on the power curve representation.
[40]	Tutivén et al., 2022	WT main bearing fault diagnosis	An NBM for the main shaft temperature using a one-class Support Vector Machine (SVM) classifier is introduced. Healthy data come from 2 years of operation of a WT with no reported abnormal main bearing condition.
[68]	Yi and Jiang, 2020	Feature learning for blade ice accretion	Stacked auto-encoders (SAE) and sparse linear discriminant analysis (SLDA) are used to extract and project discriminative features from the SCADA data.
[69]	Wei et al., 2019	WT electric pitch system failure	The NMB is built using optimized relevance vector machine (RVM) regression. SCADA log files are used to filter SCADA data and remove interferential information.
[70]	Lebranchu et al., 2019	WT fault indexes at individual and wind farm levels	NBMs are proposed at the level of individual WTs and of the wind farm, and the fault indexes are based on the NBMs residues.
[71]	Saari et al., 2019	WT bearing faults detection and identification	The NBM uses the one-class support vector machine model. Healthy data correspond to 120 days of the vibratory data from the accelerometer mounted on the gearbox housing.

NBMs listed in Table 6 use various algorithms to analyze the SCADA data. It is worth mentioning that WT's normal or healthy condition comprises significantly heterogeneous operating conditions. For example, the active power can be a cubic or a constant function of the wind speed, and the nacelle temperature varies greatly under the influence of daily and seasonal variations. Consequently, creating NBMs for the overall healthy WTs can be challenging. Instead, most of the reviewed NBMs focus on the health condition analysis of a particular subsystem or component, thus limiting the healthy/unhealthy classification to the level of the component or subsystem. Regression-based NBMs model specific SCADA measures such as main bearing temperature [40] or gearbox bearing temperature [67]. The residue between NBM prediction and actual measures allows for detecting changes in the system's behavior. Alternatively, NBMs can target health indexes, with detection based on thresholds [70].

NBMs at the wind farm level assume that WTs from the same model behave similarly. In particular, their degradation follows similar patterns. This assumption might be coherent for most WTs from a large wind farm but may not hold for WTs after major repairs and component replacements [72].

A priori, NBMs would be trained solely on healthy data [73]. Nevertheless, in cases with severe imbalance toward the majority healthy class, such as the SCADA imbalance for some less frequent abnormal conditions, the NBM could eventually be trained with all data regardless of the condition provided that the modeling approach has low sensitivity to outliers [74].

4.2. Data-Level Methods

Table 7 presents an overview of data-level methods for dealing with SCADA data imbalance.

Table 7. Data-level methods for handling SCADA data imbalance.

Ref.	Author(s), Year	Focus	Description and Contributions
[12]	Maldonado-Correa et al., 2024	Multiple WT generator faults detection	Multiple oversampling techniques were analyzed to perform binary classification using high-frequency SCADA measures. Among SMOTE, SMOTE + Tomek, SMOTE + ENN, and ADASYN.
[75]	Fazli and Poshtan, 2024	WT fault detection and isolation	The operational dataset is labeled using the status and warning datasets. The labeled data are used for the supervised training of the K-Nearest Neighbors (KNN) classifier.
[76]	Li et al., 2023	Blade icing detection	The proposed CJBM method combines an improved clustering-based oversampling technique (γ MiniDPC-SMOTE) with LightGBM to balance data and improve icing prediction accuracy.
[77]	Jin et al., 2023	Blade icing detection	An approach combining downsampling of the normal data and upsampling of the icing data was introduced. The upsampling of the icing data combines the synthetic minority oversampling technique (SMOTE) with Wilson's edited nearest neighbor rule (ENN) technique. The ENN allows removing the misclassified nearest neighbor examples from the training set.
[78]	Chen et al., 2022	Blade icing detection	A method for oversampling of minority samples was proposed by combining the Adaptive Synthetic Sampling (ADASYN) method with the Sliding Window Upsampling (SWU) technique.
[79]	Jiang et al., 2022	Fault detection	Use of the downsampling method to reduce the effect of data imbalances.
[80]	Wang et al., 2022	Blade icing detection	A sliding window oversampling technique was used to create segments of data for normal and blade icing conditions. This technique might use overlapping segments.
[81]	Tian et al., 2021	Blade icing detection	Two approaches are explored: class-rebalanced loss function and data resampling procedure.
[82]	Jiang and Li, 2021	Multiple WT faults detection and localization	A Frequent principal fault detection and localization (FPFDL) approach is proposed using improved synthetic oversampling techniques combined with dependent wild bootstrap oversampling and 1D convolutional neural networks.
[83]	Ding et al., 2021	Blade icing detection	Proposes a PCC-based algorithm for measuring the degree of blade icing and an ensemble learning model to deal with the label missing problem and the class-imbalance problem.
[84]	Yi et al., 2020	Blade icing detection	Proposes a minority clustering SMOTE (MC-SMOTE) method that involves the clustering of minority class samples to improve the imbalance classification performance.
[85]	Liu et al., 2018	Blade icing detection	Use of the over-sampling technique SMOTE to tackle the SCADA imbalance issue.
[86]	Li et al., 2019	Multiple WT faults	Random undersampling of the majority class transformed the dataset, initially with an imbalance ratio of 47:1, into a balanced dataset with a 1:1 ratio between the majority and minority classes.

Data-level strategies derive from approaches that use undersampling of the majority class and oversampling of the minority class. They require explicit identification of the majority and minority classes, hence labeling.

4.3. Algorithm-Level Methods

Table 8 lists papers introducing algorithm-level approaches suitable for analyzing highly imbalanced SCADA data.

The algorithm-level methods listed in Table 8 are mostly DNNs designed to ensure or enhance learning from the minority class from the SCADA data through techniques involving appropriate loss functions (e.g., FL), attention mechanisms (e.g., TACNN and MT-STAN), ensemble learning (e.g., MK-FCNN), and hybrid architectures (e.g., CNN-RNN, 1D-CNN-SBiGRU, and MWGCN).

Table 8. Algorithm-level methods for handling SCADA data imbalance.

Ref.	Author(s), Year	Focus	Description and Contributions
[87]	Sun et al., 2023	Fault diagnosis	A matching contrastive learning method is proposed to extract spatial and temporal information from SCADA data.
[88]	Liu et al., 2023	Anomaly detection, blade icing detection	A triplet-Convolutional Deep Autoencoder (Conv DAE) is proposed to model normal data and acquire discriminative deep feature representation.
[89]	Sun et al., 2023	Fault diagnosis	A combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) is used to enhance recognition accuracy of feature classes.
[90]	Jiang et al., 2023	Blade icing detection	A novel multi-task temporal-spatial attention network (MT-STAN) is proposed. It integrates a spatiotemporal attention block for feature extraction and a multi-task learning module to enhance the discriminative ability and improve fault detection performance.
[91]	Man et al., 2023	Blade icing detection	The proposed method utilizes Focal loss to counter SCADA imbalance and XGBoost as a classifier for improved detection performances.
[92]	Lai et al., 2022	Blade icing detection	Classifier with Focal Loss (FL) is used to balance SCADA data. The Multiscale Wavelet-Driven Graph Convolutional Network (MWGCN) is proposed to extract features and determine implicit information of intervariable correlations in the SCADA data.
[93]	Tong et al., 2022	Blade icing detection	A semi-supervised extreme learning machine (ESS-ELM) is built based on the ELM algorithm to address the unlabeled and imbalanced SCADA data issue.
[94]	Cheng et al., 2022	Blade icing detection	Blockchain-empowered imbalanced federated learning model integrating cluster-based learning to address class imbalance.
[95]	Xiao et al., 2022	Blade icing detection	GMDH selective deep ensemble (GSDE) integrates cost-sensitive with focal loss deep neural networks to address data imbalance as well as improve fault detection performances.
[96]	Li et al., 2022	Blade icing diagnosis and prediction	A model combining the ReliefF algorithm for feature selection with the one-dimensional convolution and stacked bidirectional Gated Recurrent Unit (1D-CNN-SBiGRU) structure was proposed. It enhances the performance measured by a new weighted accuracy index.
[97]	Cheng et al., 2021	Blade icing detection	A deep class-imbalanced semisupervised (DCISS) model is introduced for estimating blade icing conditions. DCISS combines a prototypical network with semi-supervised learning (SSL).
[81]	Tian et al., 2021	Blade icing detection	Use of the sliding window (without overlapping) resampling algorithm as well as a class-balancing loss function.
[98]	Cheng et al., 2021	Blade icing detection	A temporal attention-based CNN (TACNN) is explored to identify important features from imbalanced sensor data.

Table 8. *Cont.*

Ref.	Author(s), Year	Focus	Description and Contributions
[5]	Chen et al., 2021	Fault diagnosis, blade ice accretion, gear cog belt fracture	The transfer learning algorithm TrAdaBoost is introduced to tackle the challenge of SCADA data imbalance. It assigns higher weights to misclassified samples, which enhances the model's sensitivity to underrepresented faulty data.
[99]	Pang et al., 2020	Multi-class fault diagnosis in WTs	A Multi-Kernel Fusion Convolutional Neural Network (MK-FCNN) is proposed to extract multiscale spatial correlations among different variables. It is succeeded by a Long Short-Term Memory (LSTM) model that learns the temporal dependence of the learned spatial features.
[100]	Tang et al., 2020	Pitch system fault detection	Method based on multi-class optimal margin distribution machine (MCODM) for accurately detect faults and to address imbalanced samples.
[101]	Sá et al., 2020	Fault detection	Proposed workflow combining the multi-objective optimization framework Non-dominated Sorting Genetic Algorithm II (NSGA-II) for feature and hyperparameter selection with Soft-Label and a Binary support vector machine (SVM) for fault detection.
[102]	Chen et al., 2019	Fault detection for variable speed WTs	A Long Short-Term Memory (LSTM) neural network is proposed to extract features of the fault signal
[16]	Chen et al., 2019	Blade icing detection	A fault diagnosis deep learning model is introduced. The approach uses the triplet loss function to preserve within-class information and between-class information.
[103]	Karim et al., 2019	Time series classification	Use of a squeeze and excitation block (multivariate time series classification model) to improve the accuracy of pre-existing LSTM methods.

4.4. Cost-Sensitive Learning and Hybrid Strategies

Cost-sensitive strategies attribute weights according to the data classification and often require labeling the datasets. These learning strategies can be seen as a combination of the data-level and the algorithm-level approaches, where the loss function is set differently for majority and minority classes [24]. Table 9 lists papers using cost-sensitive learning strategies to analyze imbalanced SCADA data.

Table 9. Cost-sensitive learning and hybrid strategies.

Ref.	Author(s), Year	Focus	Description and Contributions
[104]	Jiang et al., 2023	Blade icing detection	A spatiotemporal attention model is combined with a self-adaptive weight loss function. Classification is enhanced by adaptively assigning weights to data categories according to their numbers in divided batches.
[105]	Chatterjee, 2023	Multiple WT fault classification	Over and under-sampling are combined using adaptive SMOTE and Edited Nearest Neighbors (ASMOTE-ENN). The resulting multi-step approach reduces noise in the imbalanced datasets and obtains samples that precisely fall into the minority class.
[106]	Meng et al., 2023	Wind power prediction	The proposed method proactively addresses the imbalanced nature of samples by enhancing adaptive learning ability and significantly reducing prediction errors. This adaptive learning combines segment imbalance regression (SIR) with crisscross optimization.

Table 9. *Cont.*

Ref.	Author(s), Year	Focus	Description and Contributions
[107]	Tong et al., 2021	Blade icing detection	An adaptive weighted kernel extreme learning machine (AWKELM) algorithm is introduced, effectively improving fault detection performance under imbalanced data conditions. The adaptive weighting strategy considers sample distribution information in combination with a fixed weighting strategy.
[108]	He et al., 2020	Multi-class fault diagnosis	A SpatioTemporal Multiscale Neural Network (STMNN) model is introduced. It combines a deep echo state network for temporal feature extraction with a multiscale residual network for spatial feature extraction. Training with dynamically adjusted focal loss (FL) solves the data imbalance problem.

Among hybrid strategies, ensemble classification models combine multiple classifiers to obtain better accuracy compared to individual classifiers [24]. Examples of such strategies include AdaBoot [109,110] and Bagging [111]. Implementing hybrid strategies tends to be more complex than data-level or algorithm-level strategies taken alone.

4.5. Data Augmentation

Table 10 presents papers from a broad class of works, including generative models, synthetic data generated from simulation, and transfer learning.

Table 10. Generative models, synthetic data from simulation, and transfer learning.

Ref.	Author(s), Year	Focus	Description and Contributions
[112]	Wang et al., 2024	Fault diagnosis	A stacked capsule autoencoders is proposed to address the issues of inadequately labeled data and class imbalance, utilizing a prior knowledge-based convolution layer to optimize the initialization of capsules and improve spectral information learning.
[88]	Liu et al., 2023	Anomaly detection, blade icing detection	A data augmentation method is introduced. It is based on the Dynamic Time Warping-SMOTE Generative Adversarial Network (DTW-SMOTE GAN) model, which generates high-quality synthetic fault instances.
[113]	Oliveira-Filho et al., 2023	Multi-class detection and diagnosis	Two-steps approach: First, faulty data from multiple WTs are gathered to build fault-condition datasets. Second, a Variational Autoencoder-based data augmentation technique is used to generate new samples for each dataset.
[114]	Pujana et al., 2023	WT drivetrain digital twin modeling	A method to generate synthetic data combining a hybrid model with the statistical characterization of normal and faulty conditions. The generation of new failure scenarios is based on a random sampling of the respective probability distributions.
[115]	Su et al., 2022	Improved fault diagnosis method for WT gearboxes	Use of generative adversarial networks (GAN) to generate expanded samples that conform to the original distribution, amplifying imbalanced fault features, and providing more data space for diagnosis and classification.
[73]	Wang et al., 2022	A de-ambiguous CM scheme	The proposed De-Ambiguous CM Scheme with Transfer Layer (DCMT) integrates the least squares generative adversarial network (LSGAN) to augment health data and eliminate the effect of ambiguous data, improving the reliability of training datasets and enhancing early fault detection accuracy.

Table 10. *Cont.*

Ref.	Author(s), Year	Focus	Description and Contributions
[73]	Wang et al., 2022	De-ambiguous CM scheme	Least squares generative adversarial networks (LSGAN) to address data ambiguity in WT CM.
[116]	Jin et al., 2022	WT generator CM	An instance-based transfer learning approach is proposed for the CM of WT generator with insufficient data. It uses a weighting TrAdaBoost algorithm.
[117]	Chen et al., 2021	WT generator bearings CM	Use of deep convolutional generative adversarial networks (DCGAN) to generate synthetic data from healthy samples.
[118]	Velandia-Cardenas et al., 2021	Fault detection	SCADA data were preprocessed using principal component analysis (PCA). An ensemble method using a combination of random undersampling and the standard boosting procedure AdaBoost was compared against supervised ML methods fed with processed data using time-split and oversampling techniques.
[119]	Xu et al., 2020	Blade icing detection	A method combining fixed length undersampling and SMOTE oversampling techniques are introduced. The SVM model performance is enhanced using particle swarm optimization.

Generative models used to generate synthetic data for WT operation include Variational Autoencoders [120], Generative Adversarial Networks (GAN) [121], Wasserstein Generative Adversarial Networks (WGAN) [122], Least Squares Generative Adversarial Networks (LSGAN) [123], among others. The implementation of a transfer learning model for WTs may use data from other units of the same WT model, i.e., the same manufacturer and specifications [14]. The general deep learning model can be built for a WT model and then be trained or updated considering data from one specific WT unit [124–126].

5. Discussion

Most papers reviewed in Section 4 use specific strategies to deal with the SCADA data imbalance without further evaluation or comparison with alternative methods. The choice of one particular strategy can be linked to the characteristics of the SCADA data and the interest analyses.

The extent of SCADA data available is a determinant factor influencing the choice of the balancing strategy. NBMs only require healthy data and, in cases of severe imbalance, can use all available data regardless of the class, provided that the chosen model is not sensitive to outliers. Data-level, algorithm-level, and cost-sensitive strategies are suitable for imbalanced datasets, provided that enough minority-class data are available. Such data allow modeling within minority-class information. Generative models, simulation-based data generation, and transfer learning allow using information from other WTs of the same model, eventually from other wind farms. These data augmentation techniques are particularly suitable to overcome data scarcity, such as in newly commissioned wind farms.

The second factor is the possibility and cost of labeling the available data. Unsupervised learning strategies have the advantage of not requiring SCADA data labeling, which is time-consuming, costly, and prone to errors. NBMs can use unsupervised training and are easier to implement than the other modeling approaches. Data-level and hybrid strategies require labeling the SCADA database. When multiple degraded conditions are analyzed, it might be necessary to label multiple conditions, which is challenging. Algorithm-level strategies involve unsupervised learning, which makes them appropriate for analyzing imbalanced SCADA data without labels.

A third factor concerns the purpose of the analysis, as different data balancing strategies suit different kinds of WT health condition analysis. NBMs are suitable for anomaly detection and CM, but their binary characterization (normal behavior or not) limits their use for diagnosis and prognosis. Data-level, algorithm-level, and hybrid strategies are used in approaches aiming at the early detection of abnormal conditions, CM, diagnosis, and prognosis. Finally, data augmentation techniques are often combined with other models, which allows for WT health condition analyses even when data are scarce. Detection and CM are less sensitive to data imbalance than diagnosis since the latter involves distinguishing among various degradation states. The specificity of prognosis is the need for run-to-failure time series from the imbalanced SCADA database.

Forecasts for the wind energy sector suggest the WT global fleet is shifting toward larger WTs in the coming decades and anticipate significant growth in offshore installed capacity [127]. Advancements in wind turbine technologies, maturity of the wind energy sector, and scale factor may benefit strategies tackling the problem of SCADA data imbalance. Indeed, future-generation WTs are expected to include captors from enhanced SCADA systems and diverse Condition Monitoring System (CMS) captors. The CMS can provide various types of data, including vibratory measurements, lubricant or grease condition analysis, and strain gauge stress measurements. These characterizations favor physical-based CM approaches, potentially complementing or overcoming data-based approaches. At the system level, the availability of finer-scale measurements paves the way for digital twin modeling [114,128].

In closing this discussion, building and maintaining an extensive public SCADA database remains among the perspectives for the future of the wind energy sector. Such a database could ease the problem of abnormal data scarcity, thereby contributing to handling the SCADA data imbalance. However, only a few datasets have been made public to date, and these are limited to relatively short periods and a few classes of degradation or failure conditions [17].

6. Conclusions and Perspectives

Reviews highlighted in the Introduction (Table 1) state the importance of assessing and addressing the imbalance of SCADA data while implementing WT health condition analyses such as detection, CM, diagnosis, and prognosis. Nevertheless, the literature is limited in its appreciation of the impact of the SCADA data imbalance, and little attention is given to the specifics of balancing strategies.

This paper presented strategies to mitigate the imbalance of SCADA data within five categories: NBMs, data-level strategies, algorithm-level strategies, cost-sensitive learning, and data augmentation techniques. This review identified three key factors influencing the choice among the balancing strategies: (i) the extent of SCADA data availability, (ii) the possibility and cost of data labeling, and (iii) the purpose and level of the desired WT health condition analysis.

The present review suggests that unsupervised learning models were predominant among the targeted works—papers published from 2019 to 2024 explicitly mentioning the SCADA data imbalance. Unsupervised learning strategies motivate intense research activity and industrial interest because these models overcome the need to label the SCADA database, which is time-consuming, costly, and error-prone.

Some aspects of SCADA data analysis remain to be comprehensively addressed to fully assess the impact of SCADA data imbalance. The present review considered papers that explicitly mentioned the problem of SCADA data imbalance. A comprehensive review could be defined by relaxing this criterion to include other pertinent strategies for SCADA data balancing. The balancing strategies serve different kinds of health condition

analysis, which makes direct comparisons challenging. This review's authors believe the two following open questions are worth further investigation. The first question concerns the impact of feature selection. Analyzing the impact of the number and kind of SCADA measures selected as features could enhance the understanding of feature importance and optimize model performance. More complex models, such as graph neural networks, allow for integrating multiple subsystems or component variables under the assumption that a graph model is appropriate. The second recommended research question is whether analyzing multiple abnormal conditions (instead of a single degraded condition) can improve diagnostic accuracy. The multi-class modeling approach gathers information from different datasets in the same model. Addressing these aspects of SCADA-based analyses will potentially benefit the implementation of WT health condition analyses in operating wind farms.

Author Contributions: Conceptualization, methodology, and investigation: A.O.-F. and A.T.; Formal analysis: A.O.-F., M.C., P.C., J.C. and C.N.; Data curation: A.O.-F., J.C. and C.N.; Writing—original draft preparation: A.O.-F., J.C. and C.N.; Writing—review and editing: A.O.-F., J.C., C.N., M.C., P.C. and A.T.; Resources, supervision, and funding acquisition: A.T., M.C. and P.C.; Project administration: A.O.-F. and A.T.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) with the Alliance Grant number 580853–22.

Acknowledgments: The authors acknowledge Power Factors for providing the database used in the work.

Conflicts of Interest: Authors Monelle Comeau and Pavel Côté are employed by Power Factors. The remaining authors declare that this research was conducted without any commercial or financial relationships that could be perceived as a potential conflict of interest.

References

1. Lee, J.; Zhao, F. *Global Wind Report 2024*; Global Wind Energy Council (GWEC): Brussels, Belgium, 2024.
2. Bošnjaković, M.; Katinić, M.; Santa, R.; Marić, D. Wind turbine technology trends. *Appl. Sci.* **2022**, *12*, 8653. [CrossRef]
3. Gonzalez, E.; Stephen, B.; Infield, D.; Melero, J.J. Using high-frequency SCADA data for wind turbine performance monitoring: A sensitivity study. *Renew. Energy* **2019**, *131*, 841–853. [CrossRef]
4. Astolfi, D.; De Caro, F.; Vaccaro, A. Condition monitoring of wind turbine systems by explainable artificial intelligence techniques. *Sensors* **2023**, *23*, 5376. [CrossRef] [PubMed]
5. Chen, W.; Qiu, Y.; Feng, Y.; Li, Y.; Kusiak, A. Diagnosis of wind turbine faults with transfer learning algorithms. *Renew. Energy* **2021**, *163*, 2053–2067. [CrossRef]
6. Oh, S.Y.; Joung, C.; Lee, S.; Shim, Y.B.; Lee, D.; Cho, G.E.; Jang, J.; Lee, I.Y.; Park, Y.B. Condition-based maintenance of wind turbine structures: A state-of-the-art review. *Renew. Sustain. Energy Rev.* **2024**, *204*, 114799. [CrossRef]
7. Cuesta, J.; Leturiondo, U.; Vidal, Y.; Pozo, F. Challenges on prognostics and health management for wind turbine components. *J. Phys. Conf. Ser.* **2024**, *2745*, 012003. [CrossRef]
8. Pandit, R.; Wang, J. A comprehensive review on enhancing wind turbine applications with advanced SCADA data analytics and practical insights. *IET Renew. Power Gener.* **2024**, *18*, 722–742. [CrossRef]
9. Badihi, H.; Zhang, Y.; Jiang, B.; Pillay, P.; Rakheja, S. A comprehensive review on signal-based and model-based condition monitoring of wind turbines: Fault diagnosis and lifetime prognosis. *Proc. IEEE* **2022**, *110*, 754–806. [CrossRef]
10. Santiago, R.A.d.F.; Barbosa, N.B.; Mergulhão, H.G.; Carvalho, T.F.d.; Santos, A.A.B.; Medrado, R.C.; Filho, J.B.d.M.; Pinheiro, O.R.; Nascimento, E.G.S. Data-driven models applied to predictive and prescriptive maintenance of wind turbine: A systematic review of approaches based on failure detection, diagnosis, and prognosis. *Energies* **2024**, *17*, 1010. [CrossRef]
11. Pfaffel, S.; Faulstich, S.; Rohrig, K. Performance and reliability of wind turbines: A review. *Energies* **2017**, *10*, 1904. [CrossRef]
12. Maldonado-Correa, J.; Valdiviezo-Condolo, M.; Artigao, E.; Martín-Martínez, S.; Gómez-Lázaro, E. Classification of highly imbalanced supervisory control and data acquisition data for fault detection of wind turbine generators. *Energies* **2024**, *17*, 1590. [CrossRef]
13. Ma, J.; Yuan, Y. Application of SCADA data in wind turbine fault detection—A review. *Sens. Rev.* **2023**, *43*, 1–11. [CrossRef]

14. Nunes, A.R.; Morais, H.; Sardinha, A. Use of learning mechanisms to improve the condition monitoring of wind turbine generators: A review. *Energies* **2021**, *14*, 7129. [CrossRef]
15. Stetco, A.; Dinmohammadi, F.; Zhao, X.; Robu, V.; Flynn, D.; Barnes, M.; Keane, J.; Nenadic, G. Machine learning methods for wind turbine condition monitoring: A review. *Renew. Energy* **2019**, *133*, 620–635. [CrossRef]
16. Chen, L.; Xu, G.; Zhang, Q.; Zhang, X. Learning deep representation of imbalanced SCADA data for fault detection of wind turbines. *Measurement* **2019**, *139*, 370–379. [CrossRef]
17. Helbing, G.; Ritter, M. Deep learning for fault detection in wind turbines. *Renew. Sustain. Energy Rev.* **2018**, *98*, 189–198. [CrossRef]
18. Alloghani, M.; Al-Jumeily, D.; Mustafina, J.; Hussain, A.; Aljaaf, A.J., A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In *Supervised and Unsupervised Learning for Data Science*; Springer International Publishing: Cham, Germany, 2020; pp. 3–21. [CrossRef]
19. Pandit, R.; Astolfi, D.; Hong, J.; Infield, D.; Santos, M. SCADA data for wind turbine data-driven condition/performance monitoring: A review on state-of-art, challenges and future trends. *Wind Eng.* **2023**, *47*, 422–441. [CrossRef]
20. Zheng, M.; Man, J.; Wang, D.; Chen, Y.; Li, Q.; Liu, Y. Semi-supervised multivariate time series anomaly detection for wind turbines using generator SCADA data. *Reliab. Eng. Syst. Saf.* **2023**, *235*, 109235. [CrossRef]
21. Vásquez-Rodríguez, G.; Maldonado-Correa, J. Anomaly-based fault detection in wind turbines using unsupervised learning: A comparative study. *Iop Conf. Ser. Earth Environ. Sci.* **2024**, *1370*, 012005. [CrossRef]
22. Rezvani, S.; Wang, X. A broad review on class imbalance learning techniques. *Appl. Soft Comput.* **2023**, *143*, 110415. [CrossRef]
23. Megahed, F.M.; Chen, Y.J.; Megahed, A.; Ong, Y.; Altman, N.; Krzywinski, M. The class imbalance problem. *Nat. Methods* **2021**, *18*, 1270–1272. [CrossRef]
24. Galar, M.; Fernandez, A.; Barrenechea, E.; Bustince, H.; Herrera, F. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* **2011**, *42*, 463–484. [CrossRef]
25. Japkowicz, N.; Stephen, S. The class imbalance problem: A systematic study. *Intell. Data Anal.* **2002**, *6*, 429–449. [CrossRef]
26. Buda, M.; Maki, A.; Mazurowski, M.A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **2018**, *106*, 249–259. [CrossRef]
27. Ren, Z.; Lin, T.; Feng, K.; Zhu, Y.; Liu, Z.; Yan, K. A systematic review on imbalanced learning methods in intelligent fault diagnosis. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–35. [CrossRef]
28. Wu, Z.; Lin, W.; Ji, Y. An integrated ensemble learning model for imbalanced fault diagnostics and prognostics. *IEEE Access* **2018**, *6*, 8394–8402. [CrossRef]
29. Owusu-Adjei, M.; Ben Hayfron-Acquah, J.; Frimpong, T.; Abdul-Salaam, G. Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems. *PLoS Digit. Health* **2023**, *2*, e0000290. [CrossRef]
30. Yang, Y.; Khorshidi, H.A.; Aickelin, U. A review on over-sampling techniques in classification of multi-class imbalanced datasets: Insights for medical problems. *Front. Digit. Health* **2024**, *6*, 1430245. [CrossRef]
31. Zaher, A.; McArthur, S.; Infield, D.; Patel, Y. Online wind turbine fault detection through automated SCADA data analysis. *Wind Energy* **2009**, *12*, 574–593. [CrossRef]
32. Meyer, A. Multi-target normal behaviour models for wind farm condition monitoring. *Appl. Energy* **2021**, *300*, 117342. [CrossRef]
33. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
34. Mohammed, R.; Rawashdeh, J.; Abdullah, M. Machine learning with oversampling and undersampling techniques: overview study and experimental results. In Proceedings of the 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 7–9 April 2020; pp. 243–248. [CrossRef]
35. Tian, J.; Jiang, Y.; Zhang, J.; Luo, H.; Yin, S. A novel data augmentation approach to fault diagnosis with class-imbalance problem. *Reliab. Eng. Syst. Saf.* **2024**, *243*, 109832. [CrossRef]
36. Hsu, M.C.; Akkerman, I.; Bazilevs, Y. Finite element simulation of wind turbine aerodynamics: Validation study using NREL Phase VI experiment. *Wind Energy* **2014**, *17*, 461–481. [CrossRef]
37. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 1–54. [CrossRef]
38. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **1997**, *30*, 1145–1159. [CrossRef]
39. Cambron, P.; Tahan, A.; Masson, C.; Pelletier, F. Bearing temperature monitoring of a wind turbine using physics-based model. *J. Qual. Maint. Eng.* **2017**, *23*, 479–488. [CrossRef]
40. Tutivén, C.; Vidal, Y.; Insuasty, A.; Campoverde-Vilela, L.; Achicanoy, W. Early fault diagnosis strategy for WT main bearings based on SCADA data and one-class SVM. *Energies* **2022**, *15*, 4381. [CrossRef]
41. Bai, X.; Han, S.; Kang, Z.; Tao, T.; Pang, C.; Dai, S.; Liu, Y. Wind turbine gearbox oil temperature feature extraction and condition monitoring based on energy flow. *Appl. Energy* **2024**, *371*, 123687. [CrossRef]

42. Yan, J.; Liu, Y.; Meng, H.; Li, L.; Ren, X. Wind turbine generator early fault diagnosis using LSTM-based stacked denoising autoencoder network and stacking algorithm. *Int. J. Green Energy* **2024**, *21*, 2477–2492. [CrossRef]
43. Zhang, Y.; Kehtarnavaz, N.; Rotea, M.; Dasari, T. Prediction of Icing on Wind Turbines Based on SCADA Data via Temporal Convolutional Network. *Energies* **2024**, *17*, 2175. [CrossRef]
44. Ye, F.; Ezzat, A.A. Icing detection and prediction for wind turbines using multivariate sensor data and machine learning. *Renew. Energy* **2024**, *231*, 120879. [CrossRef]
45. Astolfi, D.; Pasetti, M.; Lombardi, A.; Terzi, L.; Girard, N.; Poncet, P.; Masson, J.; Dieudegard, T.; Castellani, F. A General Method For The Diagnosis Of Wind Turbine Systematic Yaw Error Based Solely On SCADA Data. *J. Physics Conf. Ser.* **2024**, *2767*, 042007. [CrossRef]
46. McKinnon, C.; Carroll, J.; McDonald, A.; Koukoura, S.; Plumley, C. Investigation of isolation forest for wind turbine pitch system condition monitoring using SCADA data. *Energies* **2021**, *14*, 6601. [CrossRef]
47. Zheng, Y.; Wang, C.; Huang, C.; Li, K.; Yang, J.; Xie, N.; Liu, B.; Zhang, Y. Hierarchical spatial–temporal autocorrelation graph neural network for online wind turbine pitch system fault detection. *Neurocomputing* **2024**, *586*, 127574. [CrossRef]
48. Mehlan, F.C.; Nejad, A.R. Rotor imbalance detection and diagnosis in floating wind turbines by means of drivetrain condition monitoring. *Renew. Energy* **2023**, *212*, 70–81. [CrossRef]
49. Elorza, I.; Arrizabalaga, I.; Zubizarreta, A.; Martín-Aguilar, H.; Pujana-Arrese, A.; Calleja, C. A sensor data processing algorithm for wind turbine hydraulic pitch system diagnosis. *Energies* **2021**, *15*, 33. [CrossRef]
50. Zhou, L.; Zhao, Q.; Wang, X.; Zhu, A. Fault diagnosis and reconstruction of wind turbine anemometer based on RWSSA-AANN. *Energies* **2021**, *14*, 6905. [CrossRef]
51. Entezami, M.; Hillmansen, S.; Weston, P.; Papaelias, M.P. Fault detection and diagnosis within a wind turbine mechanical braking system using condition monitoring. *Renew. Energy* **2012**, *47*, 175–182. [CrossRef]
52. Tautz-Weinert, J.; Watson, S.J. Using SCADA data for wind turbine condition monitoring—A review. *Iet Renew. Power Gener.* **2017**, *11*, 382–394. [CrossRef]
53. Zhang, D.; Tian, W.; Cheng, X.; Shi, F.; Qiu, H.; Liu, X.; Chen, S. FedBIP: A federated learning-based model for wind turbine blade icing prediction. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–11. [CrossRef]
54. Marti-Puig, P.; Blanco-M, A.; Cárdenas, J.J.; Cusidó, J.; Solé-Casals, J. Effects of the pre-processing algorithms in fault diagnosis of wind turbines. *Environ. Model. Softw.* **2018**, *110*, 119–128. [CrossRef]
55. Oliveira-Filho, A.; Zemouri, R.; Pelletier, F.; Tahan, A. System Condition Monitoring Based on a Standardized Latent Space and the Nataf Transform. *IEEE Access* **2024**, *12*, 32637–32659. [CrossRef]
56. Du, M.; Yi, J.; Mazidi, P.; Cheng, L.; Guo, J. A parameter selection method for wind turbine health management through SCADA data. *Energies* **2017**, *10*, 253. [CrossRef]
57. Cheng, R.C.; Chen, K.S. Ball bearing multiple failure diagnosis using feature-selected autoencoder model. *Int. J. Adv. Manuf. Technol.* **2022**, *120*, 4803–4819. [CrossRef]
58. Costa, N.; Sánchez, L. Variational encoding approach for interpretable assessment of remaining useful life estimation. *Reliab. Eng. Syst. Saf.* **2022**, *222*, 108353. [CrossRef]
59. Menezes, E.J.N.; Araújo, A.M.; Da Silva, N.S.B. A review on wind turbine control and its associated methods. *J. Clean. Prod.* **2018**, *174*, 945–953. [CrossRef]
60. Artigao, E.; Martín-Martínez, S.; Honrubia-Escribano, A.; Gómez-Lázaro, E. Wind turbine reliability: A comprehensive review towards effective condition monitoring development. *Appl. Energy* **2018**, *228*, 1569–1583. [CrossRef]
61. Santelo, T.N.; de Oliveira, C.M.R.; Maciel, C.D.; de A. Monteiro, J.R.B. Wind turbine failures review and trends. *J. Control. Autom. Electr. Syst.* **2022**, *33*, 1–17. [CrossRef]
62. Chestney, N.; Steitz, C. What Are the Issues with Siemens Gamesa's Wind Turbines? 2023. Available online: https://www.reuters.com/business/energy/what-are-issues-with-siemens-gamesas-wind-turbines-2023-06-23/ (accessed on 16 October 2024).
63. Cole, V.; Boutet, M. ResearchRabbit. *J. Can. Health Libr. Assoc.* **2023**, *44*, 43. [CrossRef]
64. Takkouche, B.; Norman, G. PRISMA statement. *Epidemiology* **2011**, *22*, 128. [CrossRef]
65. Moher, D.; Stewart, L.; Shekelle, P. Implementing PRISMA-P: Recommendations for prospective authors. *Syst. Rev.* **2016**, *5*, 1–2. [CrossRef] [PubMed]
66. Murgia, A.; Verbeke, R.; Tsiporkova, E.; Terzi, L.; Astolfi, D. Discussion on the suitability of SCADA-based condition monitoring for wind turbine fault diagnosis through temperature data analysis. *Energies* **2023**, *16*, 620. [CrossRef]
67. Bilendo, F.; Lu, N.; Badihi, H.; Meyer, A.; Cali, Ü.; Cambron, P. Multitarget normal behavior model based on heterogeneous stacked regressions and change-point detection for wind turbine condition monitoring. *IEEE Trans. Ind. Inform.* **2023**, *20*, 5171–5181. [CrossRef]
68. Yi, H.; Jiang, Q. Discriminative feature learning for blade icing fault detection of wind turbine. *Meas. Sci. Technol.* **2020**, *31*, 115102. [CrossRef]

69. Wei, L.; Qian, Z.; Zareipour, H. Wind turbine pitch system condition monitoring and fault detection based on optimized relevance vector machine regression. *IEEE Trans. Sustain. Energy* **2019**, *11*, 2326–2336. [CrossRef]
70. Lebranchu, A.; Charbonnier, S.; Bérenguer, C.; Prevost, F. A combined mono-and multi-turbine approach for fault indicator synthesis and wind turbine monitoring using SCADA data. *ISA Trans.* **2019**, *87*, 272–281. [CrossRef]
71. Saari, J.; Strömbergsson, D.; Lundberg, J.; Thomson, A. Detection and identification of windmill bearing faults using a one-class support vector machine (SVM). *Measurement* **2019**, *137*, 287–301. [CrossRef]
72. Soltani, M.; Kharoufeh, J.P.; Khademi, A. Structured replacement policies for offshore wind turbines. *Probab. Eng. Informational Sci.* **2024**, *38*, 355–386. [CrossRef]
73. Wang, A.; Qian, Z.; Pei, Y.; Jing, B. A de-ambiguous condition monitoring scheme for wind turbines using least squares generative adversarial networks. *Renew. Energy* **2022**, *185*, 267–279. [CrossRef]
74. Chesterman, X.; Verstraeten, T.; Daems, P.J.; Nowé, A.; Helsen, J. Overview of normal behavior modeling approaches for SCADA-based wind turbine condition monitoring demonstrated on data from operational wind farms. *Wind Energy Sci.* **2023**, *8*, 893–924. [CrossRef]
75. Fazli, A.; Poshtan, J. Wind turbine fault detection and isolation robust against data imbalance using KNN. *Energy Sci. Eng.* **2024**, *12*, 1174–1186. [CrossRef]
76. Li, S.; Peng, Y.; Bin, G. Prediction of wind turbine blades icing based on CJBM with imbalanced data. *IEEE Sens. J.* **2023**, *23*, 19726–19736. [CrossRef]
77. Jin, X.; Zhang, X.; Cheng, X.; Jiang, G.; Masisi, L.; Huang, W. A physics-based and data-driven feature extraction model for blades icing detection of wind turbines. *IEEE Sens. J.* **2023**, *23*, 3944–3954. [CrossRef]
78. Chen, W.; Cheng, L.; Chang, Z.; Wen, B.; Li, P. Wind turbine blade icing detection using a novel bidirectional gated recurrent unit with temporal pattern attention and improved coot optimization algorithm. *Meas. Sci. Technol.* **2022**, *34*, 014004. [CrossRef]
79. Jiang, G.; Fan, W.; Li, W.; Wang, L.; He, Q.; Xie, P.; Li, X. DeepFedWT: A federated deep learning framework for fault detection of wind turbines. *Measurement* **2022**, *199*, 111529. [CrossRef]
80. Wang, X.; Zheng, Z.; Jiang, G.; He, Q.; Xie, P. Detecting wind turbine blade icing with a multiscale long short-term memory network. *Energies* **2022**, *15*, 2864. [CrossRef]
81. Tian, W.; Cheng, X.; Li, G.; Shi, F.; Chen, S.; Zhang, H. A multilevel convolutional recurrent neural network for blade icing detection of wind turbine. *IEEE Sens. J.* **2021**, *21*, 20311–20323. [CrossRef]
82. Jiang, N.; Li, N. A wind turbine frequent principal fault detection and localization approach with imbalanced data using an improved synthetic oversampling technique. *Int. J. Electr. Power Energy Syst.* **2021**, *126*, 106595. [CrossRef]
83. Ding, S.; Wang, Z.; Zhang, J.; Han, F.; Gu, X.; Song, G. A PCC-Ensemble-TCN model for wind turbine icing detection using class-imbalanced and label-missing SCADA data. *Int. J. Distrib. Sens. Netw.* **2021**, *17*, 15501477211057737. [CrossRef]
84. Yi, H.; Jiang, Q.; Yan, X.; Wang, B. Imbalanced classification based on minority clustering synthetic minority oversampling technique with wind turbine fault detection application. *IEEE Trans. Ind. Inform.* **2020**, *17*, 5867–5875. [CrossRef]
85. Liu, J.; Qu, F.; Hong, X.; Zhang, H. A small-sample wind turbine fault detection method with synthetic fault data using generative adversarial nets. *IEEE Trans. Ind. Inform.* **2018**, *15*, 3877–3888. [CrossRef]
86. Li, Y.; Liu, S.; Shu, L. Wind turbine fault diagnosis based on Gaussian process classifiers applied to operational data. *Renew. Energy* **2019**, *134*, 357–366. [CrossRef]
87. Sun, S.; Hu, W.; Liu, Y.; Wang, T.; Chu, F. Matching contrastive learning: An effective and intelligent method for wind turbine fault diagnosis with imbalanced SCADA data. *Expert Syst. Appl.* **2023**, *223*, 119891. [CrossRef]
88. Liu, J.; Yang, G.; Li, X.; Wang, Q.; He, Y.; Yang, X. Wind turbine anomaly detection based on SCADA: A deep autoencoder enhanced by fault instances. *ISA Trans.* **2023**, *139*, 586–605. [CrossRef]
89. Sun, S.; Wang, T.; Chu, F. A multi-learner neural network approach to wind turbine fault diagnosis with imbalanced data. *Renew. Energy* **2023**, *208*, 420–430. [CrossRef]
90. Jiang, G.; Li, W.; Bai, J.; He, Q.; Xie, P. SCADA data-driven blade icing detection for wind turbines: An enhanced spatio-temporal feature learning approach. *Meas. Sci. Technol.* **2023**, *34*, 054004. [CrossRef]
91. Man, J.; Wang, F.; Li, Q.; Wang, D.; Qiu, Y. Semi-supervised blade icing detection method based on tri-XGBoost. *Actuators* **2023**, *12*, 58. [CrossRef]
92. Lai, Z.; Cheng, X.; Liu, X.; Huang, L.; Liu, Y. Multiscale wavelet-driven graph convolutional network for blade icing detection of wind turbines. *IEEE Sens. J.* **2022**, *22*, 21974–21985. [CrossRef]
93. Tong, R.; Li, P.; Gao, L.; Lang, X.; Miao, A.; Shen, X. A novel ellipsoidal semisupervised extreme learning machine algorithm and its application in wind turbine blade icing fault detection. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–16. [CrossRef]
94. Cheng, X.; Tian, W.; Shi, F.; Zhao, M.; Chen, S.; Wang, H. A blockchain-empowered cluster-based federated learning model for blade icing estimation on IoT-enabled wind turbine. *IEEE Trans. Ind. Inform.* **2022**, *18*, 9184–9195. [CrossRef]
95. Xiao, J.; Li, C.; Liu, B.; Huang, J.; Xie, L. Prediction of wind turbine blade icing fault based on selective deep ensemble model. *Knowl.-Based Syst.* **2022**, *242*, 108290. [CrossRef]

96. Li, Y.; Hou, L.; Tang, M.; Sun, Q.; Chen, J.; Song, W.; Yao, W.; Cao, L. Prediction of wind turbine blades icing based on feature Selection and 1D-CNN-SBiGRU. *Multimed. Tools Appl.* **2022**, *81*, 4365–4385. [CrossRef]
97. Cheng, X.; Shi, F.; Liu, X.; Zhao, M.; Chen, S. A novel deep class-imbalanced semisupervised model for wind turbine blade icing detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 2558–2570. [CrossRef] [PubMed]
98. Cheng, X.; Shi, F.; Zhao, M.; Li, G.; Zhang, H.; Chen, S. Temporal attention convolutional neural network for estimation of icing probability on wind turbine blades. *IEEE Trans. Ind. Electron.* **2021**, *69*, 6371–6380. [CrossRef]
99. Pang, Y.; He, Q.; Jiang, G.; Xie, P. Spatio-temporal fusion neural network for multi-class fault diagnosis of wind turbines based on SCADA data. *Renew. Energy* **2020**, *161*, 510–524. [CrossRef]
100. Tang, M.; Kuang, Z.; Zhao, Q.; Wu, H.; Yang, X. Fault detection of wind turbine pitch system based on multiclass optimal margin distribution machine. *Math. Probl. Eng.* **2020**, *2020*, 2091382. [CrossRef]
101. Sá, F.P.d.; Brandão, D.N.; Ogasawara, E.; Coutinho, R.d.C.; Toso, R.F. Wind turbine fault detection: A semi-supervised learning approach with automatic evolutionary feature selection. In Proceedings of the 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), Niterói, Brazil, 1–3 July 2020; pp. 323–328.
102. Chen, J.; Hu, W.; Cao, D.; Zhang, B.; Huang, Q.; Chen, Z.; Blaabjerg, F. An imbalance fault detection algorithm for variable-speed wind turbines: A deep learning approach. *Energies* **2019**, *12*, 2764. [CrossRef]
103. Karim, F.; Majumdar, S.; Darabi, H.; Harford, S. Multivariate LSTM-FCNs for time series classification. *Neural Netw.* **2019**, *116*, 237–245. [CrossRef]
104. Jiang, G.; Yue, R.; He, Q.; Xie, P.; Liu, Y. Imbalanced learning for wind turbine blade icing detection via spatio-temporal attention model with a self-adaptive weight loss function. *Expert Syst. Appl.* **2023**, *229*, 120428. [CrossRef]
105. Chatterjee, S. Highly imbalanced fault classification of wind turbines using data resampling and hybrid ensemble method approach. *Eng. Appl. Artif. Intell.* **2023**, *126*, 107104. [CrossRef]
106. Meng, A.; Xian, Z.; Yin, H.; Luo, J.; Wang, X.; Zhang, H.; Jiayu, R.; Li, C.; Zhen-bin, W.; Xie, Z.; et al. A novel network training approach for solving sample imbalance problem in wind power prediction. *Energy Convers. Manag.* **2023**, *283*, 116935. [CrossRef]
107. Tong, R.; Li, P.; Lang, X.; Liang, J.; Cao, M. A novel adaptive weighted kernel extreme learning machine algorithm and its application in wind turbine blade icing fault detection. *Measurement* **2021**, *185*, 110009. [CrossRef]
108. He, Q.; Pang, Y.; Jiang, G.; Xie, P. A spatio-temporal multiscale neural network approach for wind turbine fault diagnosis with imbalanced SCADA data. *IEEE Trans. Ind. Inform.* **2020**, *17*, 6875–6884. [CrossRef]
109. Schapire, R.E. The strength of weak learnability. *Mach. Learn.* **1990**, *5*, 197–227. [CrossRef]
110. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [CrossRef]
111. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
112. Wang, X.; Chen, H.; Zhao, J.; Song, C.; Zhang, Y.; Yang, Z.X.; Wong, P.K. Wind turbine fault diagnosis for class-imbalance and small-size data based on stacked capsule autoencoder. *IEEE Trans. Ind. Inform.* **2024**, *20*, 12694–12704. [CrossRef]
113. Oliveira-Filho, A.; Zemouri, R.; Cambron, P.; Tahan, A. Early detection and diagnosis of wind turbine abnormal conditions using an interpretable supervised variational autoencoder model. *Energies* **2023**, *16*, 4544. [CrossRef]
114. Pujana, A.; Esteras, M.; Perea, E.; Maqueda, E.; Calvez, P. Hybrid-model-based digital twin of the drivetrain of a wind turbine and its application for failure synthetic data generation. *Energies* **2023**, *16*, 861. [CrossRef]
115. Su, Y.; Meng, L.; Kong, X.; Xu, T.; Lan, X.; Li, Y. Generative adversarial networks for gearbox of wind turbine with unbalanced data sets in fault diagnosis. *IEEE Sens. J.* **2022**, *22*, 13285–13298. [CrossRef]
116. Jin, X.; Pan, H.; Ying, C.; Kong, Z.; Xu, Z.; Zhang, B. Condition monitoring of wind turbine generator based on transfer learning and one-class classifier. *IEEE Sens. J.* **2022**, *22*, 24130–24139. [CrossRef]
117. Chen, P.; Li, Y.; Wang, K.; Zuo, M.J.; Heyns, P.S.; Baggeröhr, S. A threshold self-setting condition monitoring scheme for wind turbine generator bearings based on deep convolutional generative adversarial networks. *Measurement* **2021**, *167*, 108234. [CrossRef]
118. Velandia-Cardenas, C.; Vidal, Y.; Pozo, F. Wind turbine fault detection using highly imbalanced real SCADA data. *Energies* **2021**, *14*, 1728. [CrossRef]
119. Xu, J.; Tan, W.; Li, T. Predicting fan blade icing by using particle swarm optimization and support vector machine algorithm. *Comput. Electr. Eng.* **2020**, *87*, 106751. [CrossRef]
120. Kingma, D.P. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
121. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*. https://proceedings.mlr.press/v70/arjovsky17a.html.
122. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein generative adversarial networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 214–223.
123. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Paul Smolley, S. Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2794–2802.

124. Zhang, T.; Chen, J.; Li, F.; Zhang, K.; Lv, H.; He, S.; Xu, E. Intelligent fault diagnosis of machines with small & imbalanced data: A state-of-the-art review and possible extensions. *ISA Trans.* **2022**, *119*, 152–171.
125. Yun, H.; Zhang, C.; Hou, C.; Liu, Z. An adaptive approach for ice detection in wind turbine with inductive transfer learning. *IEEE Access* **2019**, *7*, 122205–122213. [CrossRef]
126. Li, Y.; Jiang, W.; Zhang, G.; Shu, L. Wind turbine fault diagnosis based on transfer learning and convolutional autoencoder with small-scale data. *Renew. Energy* **2021**, *171*, 103–115. [CrossRef]
127. DNV. *Energy Transition Outlook 2024—A Global and Regional Forecast to 2050*; Technical Report; DNV: Hovik, Norway, 2024.
128. Branlard, E.; Jonkman, J.; Brown, C.; Zhang, J. A digital twin solution for floating offshore wind turbines validated using a full-scale prototype. *Wind Energy Sci.* **2024**, *9*, 1–24. [CrossRef]