TÉCNICO LISBOA

The Impact of Feature Causality on Normal Behaviour Models for SCADA-based Wind Turbine Fault Detection

Telmo Alexandre Silva Felgueira

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisor(s): Prof. Rui Manuel Gameiro de Castro

Examination Committee

Chairperson: Prof. Célia Maria Santos Cardoso de Jesus
Supervisor: Prof. Rui Manuel Gameiro de Castro
Member of the Committee: Prof. Maria Margarida Campos da Silveira

June 2019

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

“Think of your dreams and ideas as tiny miracle machines inside you that no one can touch. The more faith you put into them, the bigger they get, until one day they’ll rise up and taken you with them.”

William Kamkwamba, The Boy Who Harnessed the Wind

Resumo

Os custos de operação e manutenção de turbinas eólicas podem representar até 30% do custo da energia eólica. Tem havido investigação para monitorizar e detectar falhas incipientes em turbinas, a partir da recolha de grandes quantidades de dados gerados pelo seu sistema SCADA. Por sua vez, isto permite aos proprietários de parques eólicas adotarem uma manutenção preditiva, que pode reduzir consideravelmente os custos. Um dos métodos principais para monitorizar a condição de turbinas eólicas é a utilização de modelos de funcionamento normal. Mas não existe consenso relativamente a como diferentes variáveis de entrada afetam os resultados. Para além disso, a maior parte dos trabalhos utiliza redes neuronais, conhecidas por serem pouco interpretáveis e com altos custos computacionais. Este trabalho apresenta uma nova taxonomia baseada nas relações causais entre as variáveis de entrada e de saída. Com base nesta taxonomia, o impacto dos diferentes tipos de variáveis de entrada nos resultados do modelo é avaliado. Para tal, a detecção de falhas é formulada como um problema de classificação, em comparação com a abordagem padrão da literatura, que é por análise visual dos resíduos. Por fim, *gradient boosting machines* serão testadas como uma alternativa às redes neuronais. Os resultados deste trabalho indicam que os diferentes tipos de variáveis de entrada definidos na taxonomia afetam de forma diferente os resultados. Nomeadamente, os modelos com variáveis autoregressivas obtiveram os melhores resultados. Também foi demonstrado que a *framework* de detecção de falhas obteve resultados alinhados com os obtidos através de análise visual dos resíduos. Para além disso, também foi demonstrado que as *gradient boosting machines* obtiveram resultados competitivos com os de redes neuronais.

Palavras-chave: causal, NBM, turbina, falha, SCADA

Abstract

The operation and maintenance costs of wind turbines can account for up to 30% of the cost of wind energy. There have been recent efforts to monitor and detect incipient faults in turbines by harvesting the high amounts of data generated by their SCADA system. In turn, this enables the wind farm owners to employ a predictive maintenance strategy, which can considerably reduce costs.

One of the main methods for monitoring the condition of wind turbines is building normal behaviour models of the component. But there is a lack of consensus in how different input features affect the results. Besides that, most works use artificial neural networks, which are known to be black-box models with high computational costs. In this work, a new taxonomy based on the causal relations between the input features and the target is presented. Based on this taxonomy, the impact of different input feature configurations on the modelling and fault detection performance is evaluated. To this end, the detection of faults will be formulated as a classification problem, compared to the standard literature approach by visual residual analysis. Finally, gradient boosting machines will be tested as an alternative to artificial neural networks.

The results from this work indicate that the taxonomy-based input feature types differently affect model performance. The models with autoregressive features performed best overall. On the other hand, the use of simultaneity features completely eliminated the fault detection capabilities of the model. It was also shown that the fault detection framework obtained results aligned with those from visual residual analysis. Furthermore, it was demonstrated that gradient boosting machines obtain competitive results with those from artificial neural networks.

Keywords: causal, NBM, turbine, fault, SCADA

Contents

List of Tables

List of Figures

Acronyms

ACNBM Autoregressive Causal Normal Behaviour Model.

ANBM Autoregressive Normal Behaviour Model.

ANN Artificial Neural Network.

ASNBM Autoregressive Simultaneous Normal Behaviour Model.

AUC Area Under the Curve.

CM Condition Monitoring.

CMS Condition Monitoring Systems.

CNBM Causal Normal Behaviour Model.

DFIG Double-fed Induction Generator.

FN False Negative.

FP False Positive.

GBM Gradient Boosting Machine.

GPs Gaussian Processes.

KNN K-Nearest Neighbors.

LCOE Levelized Cost of Energy.

MAE Mean Absolute Error.

NBM Normal Behaviour Model.

O&M Operation and Maintenance.

PdM Predictive Maintenance.

RMSE Root Mean Squared Error.

ROC Receiver operating characteristic.

SCADA Supervisory control and data acquisition.

SD Standard Deviation.

SNBM Simultaneous Normal Behaviour Model.

SVM Support Vector Machines.

TN True Negative.

TP True Positive.

WT Wind Turbine.

Chapter 1

Introduction

1.1 Energy: Past and Present

In 2018, global energy-related CO_2 emissions reached a historic high of 33.1 gigatonnes. These emissions are caused by the burning of fossil fuels, mainly natural gas, coal and oil, which accounted for 64% of global electricity production in this same year [1]. Greenhouse gases like CO_2 are responsible for climate change which threatens to change the way we've come to know Earth and human life.

Fossil fuels are considered non-renewable resources because they don't renew themselves at sufficient rate for sustainable economic extraction in meaningful human time-frames. Based on the quantity of known fuel reserves and the current rate of production it is expected that fossil fuels will be extinguished by 2152 [2]. This is a controversial topic since the actual quantity of fuel reserves is unknown, but more important than asking how long fossil fuels will last is what would be the consequences of their continued use: In a scenario where there are no efforts to curb global warming, global average temperatures would be pushed by at least 4.0°C past industrial levels by 2100 [3]. This would have extremely profound impacts on the climate, such as the melting of glaciers and rising sea levels, the acidification of oceans and the increase in extreme weather events such as hurricanes, droughts and heat waves. [4]

Besides their impact on the climate, fossil fuels have also been fuelling geopolitical conflicts: Wars in the Middle East, the South-Sudan Civil War and more recently the South China Sea territorial disputes are just some examples of conflicts that have been fundamentally influenced by the fossil fuel industry [5, 6]. These conflicts have resulted in oil corporations and states fighting horrific wars over resource control - the Energy Wars.

Finally, it's also important to note that in developing countries over 1 billion people still lack electricity access, another 1 billion have unreliable supply and about 2.9 billion rely on traditional biomass use for heating and cooking. All this hinders important advances, including those related with health, education, gender equality, poverty eradication and ending hunger. Conventional power systems based on centralized power stations and transmission grids have not succeeded in reaching these scarcely populated areas, raising the need for new energy solutions to achieve universal electricity access [7].

1.2 Renewable Energy Revolution

Before the Industrial Revolution, almost all the energy used by humans came from renewable sources. The first milestone of mankind's utilization of energy was the mastery of fire, by burning biomass (mainly wood) for cooking, heating and light. After fire, humans managed to master water and air, using water-mills and windmills to meet their needs for crushing grains, tanning leather, smelting iron, sawing wood and so on. By the 18th century came the steam engine, this made the energy stocks accumulated in the earth's crust for hundreds of millions of years available to serve human needs. Due to their high energy density and being easily transportable, fossil fuels quickly replaced all other energy sources [8]. A clear example is that the percentage of fossil fuels as global primary energy consumption grew from 7% in 1860 to 59% by 1910 [9]. But as the 20th century ended, rising concerns over energy security, global warming, and eventual fossil fuel depletion led to an expansion of interest in all available forms of renewable energy - This marked the beginning of the renewable energy revolution.

Renewable energies are those that naturally replenish themselves on a human timescale, such as wind, hydro, geothermal and solar. In general, their lifecycle greenhouse emissions are much less significant when compared with fossil fuels, as can be seen in Table 1.1. Besides reducing negative impact on the climate, the transition towards a renewable energy system can also help solve other consequences of our fossil fuels dependency. Unlike fossil fuels, renewable energy sources are available in one form or another in most geographic locations. This abundance can help strengthen energy security and promote greater energy independence for most states, meaning that most of the oil and gas-related conflicts may decline [10]. Renewable resources also play a fundamental role in achieving universal electricity access. In fact, renewable-based mini grids could be the most cost-effective way to deliver access to more than a third of the 1.1 billion people across the world who still lack any electricity supply. Moreover, mini-grids have the ability to lay the foundation for development in rural areas, where more than four fifths of the world's unelectrified live, by powering economic activities such as agriculture, business and small industry. [11]

Table 1.1: Lifecycle emissions of selected power generation technologies (gCO2eq/kWh) - Source: [12].

Technology	Min	Median	Max
Coal – PC	740	820	910
Biomass – Cofiring with coal	620	740	890
Gas – combined cycle	410	490	650
Biomass – Dedicated	130	230	420
Solar PV – Utility scale	18	48	180
Solar PV – rooftop	26	41	60
Geothermal	6.0	38	79
Concentrated solar power	8.8	27	63
Hydropower	1.0	24	2200
Wind Offshore	8.0	12	35
Nuclear	3.7	12	110
Wind Onshore	7.0	11	56

These advantages coupled with years of policy support in key countries and momentum to mitigate climate change have helped promote the development of renewable energy. According to nearly every measure, renewable energy is gaining ground. Renewables provided an estimated 23.57% of all electricity generated in 2015, new markets are emerging in every region of the world and more jobs are to be found in the renewable energy sector than ever before [7].

Aligned with the need for the renewable energy revolution, the European Union has defined a target of 32% renewable share by 2030 for final gross energy consumption [13], Portugal has also set its owns goals of achieving a 40% share [14]. This final gross energy consumption includes energy spent on electricity, transport and heating. Focusing on the electricity part of energy consumption, wind energy is expected to contribute a significant part for the objectives of 2030. As can be seen in Figure 1.1, wind energy, both onshore and offshore, is expected to provide 18.7% of the produced electricity by 2030 in the EU. Wind energy is also expected to be the source with the highest share (24.8%) by 2040.

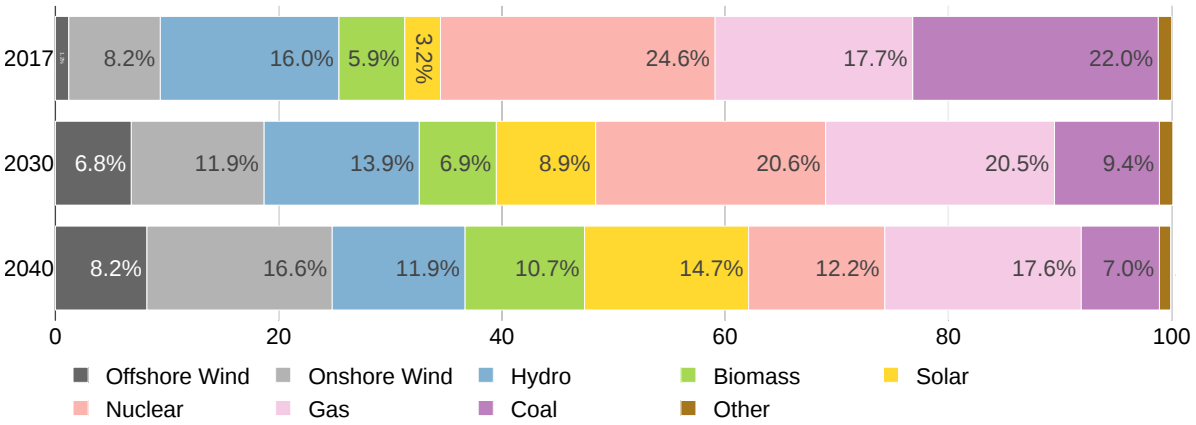


Figure 1.1: Breakdown of electricity produced in the European Union by source for 2017 and projected for 2030 and 2040 - Source: [15].

1.3 Wind Energy

The renewed interest in renewable energy by the late 20th century promoted technological advances in the field, such that by 1979 Danish manufacturers Vestas, Nordtank, Kuriant and Bonus ushered in the modern era of wind power with the mass production of large Wind Turbines (WTs) to produce electricity. From then on the fledgling commercial wind power industry started growing, reaching robust growth rates of around 20-40% per year by the beginning of the 21st century, as can be seen in figure 1.2. This growth was mostly driven by the ready availability of large wind resources, falling costs due to improved technology and also government policies such as tax credits, financial incentives, and priority access for renewable energy to the electricity grid. Specifically in Portugal, these government policies have been mainly in the form of feed-in-tariffs, which are long-term contracts in which the energy producer is offered compensation based on the generation cost of that technology.

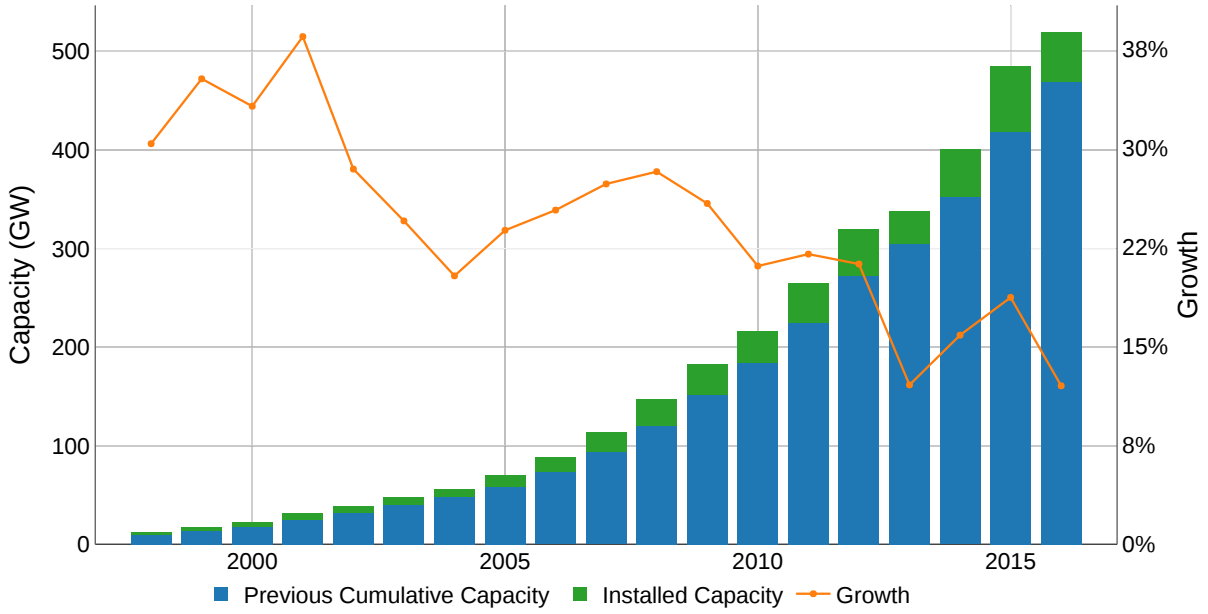


Figure 1.2: Global wind capacity (installed and cumulative) over time and corresponding growth - Source: [2].

1.3.1 Wind Energy Economics

The Levelized Cost of Energy (LCOE) is one of the most common methods to compare different power generating assets, it is equal to the average total cost to build and operate the asset over its lifetime divided by the total energy output over that same lifetime. As mentioned previously, the costs of wind installations have decreased considerably in the last years, as can be seen in Figure 1.3, the average onshore wind LCOE is now situated in the lower end of the fossil fuel cost range. Indeed, wind energy has become a competitive player in the energy market panorama of today. An illustrative example is the recent surge of non-subsidized wind projects, with various planned installations in different countries, such as 107MW planned to be installed in Finland, 300MW in Spain and 700MW in the Netherlands.

Although wind power is now one of the most competitive options for new power generation capacity, it is still crucial to continue lowering its costs. The main reason why is that now that wind energy is competitive, more and more governments are stopping the incentives that once propelled wind growth, in favor of more market-driven auctions. This has a very significant impact in the deployment of wind projects, as can be seen in figure 1.2 the years of 2004 and 2013 where the growth of wind energy clearly decreased correspond to years where the production tax credit expired in the United States, lowering the incentive of wind [17]. Besides that, there is still a high variability in wind energy costs within and between countries, as can be seen by the range of values in Figure 1.3, which coupled with the possibility of future development in fossil fuel technology means we can't take for granted wind energy competitiveness. Finally, as can also be seen in Figure 1.3, offshore wind is still barely competitive with fossil fuel technologies, making it all the more important to lower the costs. The previous arguments show that the need to lower the costs of wind are now as important as ever.

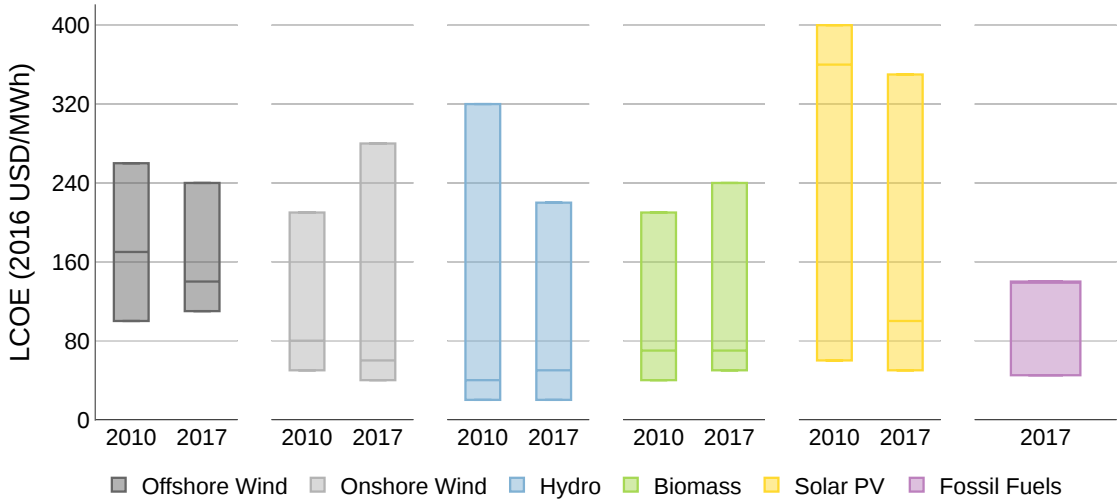


Figure 1.3: Maximum, average and minimum values of the LCOE by power technology for 2010 and 2017 - Source: [16]

1.3.1.1 Wind Energy Costs

Before understanding how to lower the LCOE of wind, it is first important to understand exactly how it is calculated by analysing the formula for the simplified LCOE according to [18] and reproduced below.

$$LCOE = \frac{I_t + C_{om} k_a}{h_a k_a} \tag{1.1}$$

$$k_a = \frac{(1+a)^n - 1}{a(1+a)^n} \tag{1.2}$$

Where:

- I_t: The total investment cost (€/kW) are assumed to all be taken at the initial time and correspond to the WTs costs, civil work, grid connection costs and more.
- C_{om}: The annual Operation and Maintenance (O&M) costs (€/kW/yr) are assumed to be constant throughout the years and correspond to the regular maintenance costs, repairs, insurance and more.
- h_a: The utilization factor is assumed to be constant throughout the years and corresponds to the number of hours the turbine would have to work at rated power to produce the same energy it actually produced.
- a: The discount rate is used to convert the values of payments or revenues made in different time instants to the same time instant.
- n: The lifetime of the turbine in years.

In first place, it's interesting to note how each of these individual parcels can help increase wind competitiveness. For example, increasing the utilization factor and the lifetime of the WT is a clear

advantage for Wind Farm Owners since they will produce more energy and thus get more revenue. On the other hand, reducing the O&M costs will not be advantageous for the Wind Farm Owner if he's not doing the O&M himself, in which case the benefits would be to the Independent Service Provider or the Original Equipment Manufacturer. These examples show the complex dynamics of the wind energy market and show why it's advantageous to use the LCOE, which combines all these factors and translates the costs for all the stakeholders involved. But since the LCOE takes into account different variables it is relevant to do a sensitivity analysis to gain intuition in how each of these individually affect the final result. This sensitivity analysis was done by using equations 1.1 and 1.2 and using as base values those described in [19] for a reference onshore wind project. For example, as can be seen in Figure 1.4, reducing the O&M costs and increasing the lifetime of the turbine both result in a significant decrease in the LCOE.

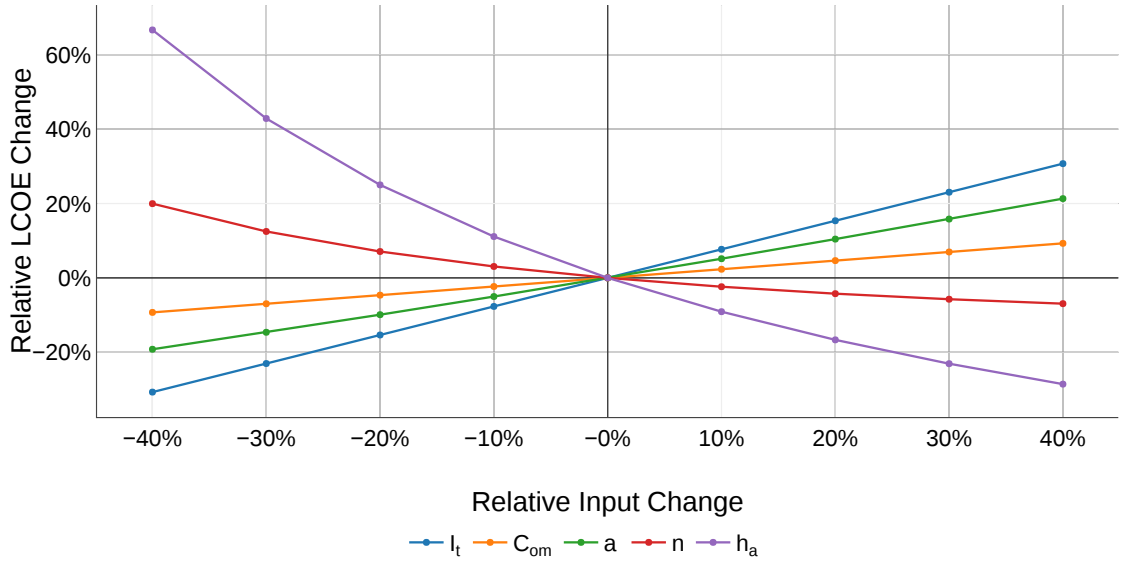


Figure 1.4: Sensitivity Analysis of LCOE for reference Onshore Wind Farm

Besides looking at the LCOE, there's also interest in understanding the specific costs of a wind project without taking into account the energy it will produce. For that we can take a look at Figure 1.5, where we can see that the O&M costs are particularly high for wind when compared to conventional fossil fuel plants, like a Gas Combined Cycle Plant. This happens because while generators in fossil fuel power plants operate in a constant, narrow range of speeds, WTs are designed to operate under a wide range of wind speeds and weather conditions. This means the stresses on components are significantly higher, thus increasing the probability of failures and consequently the maintenance costs. Besides that, wind farms are normally deployed in remote locations which naturally increases the costs of transportation and consequently the O&M costs. It's also important to note that for offshore turbines all these are aggravated by the difficulties in getting access to the WTs and also the more severe weather conditions. Finally, for turbines nearing their end of life these costs can extend up to 35% [20]. This is specially important since the global average turbine is just 6 years old, which means the O&M costs are expected to increase as turbines begin to age [20]. All these reasons make it extremely important to

understand how we can lower these maintenance costs.

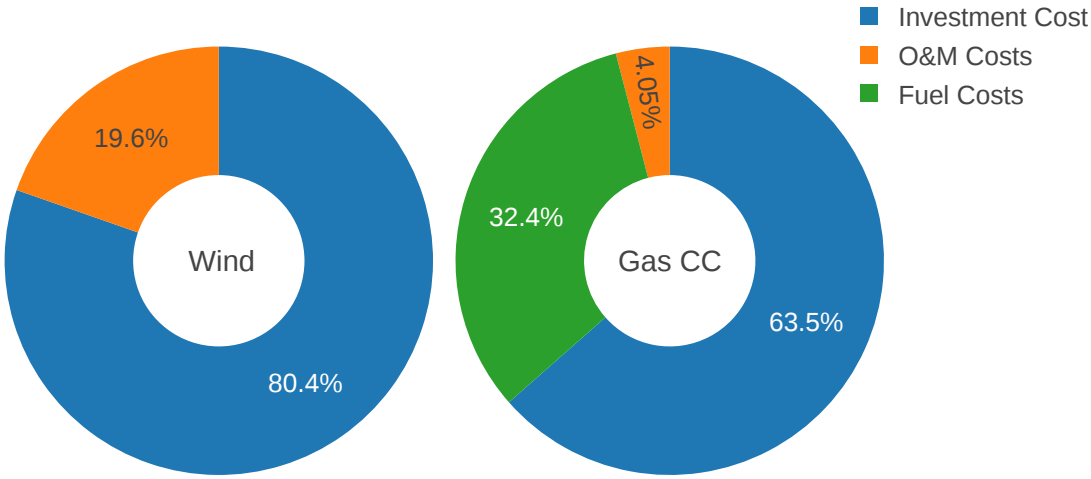


Figure 1.5: Cost breakdown for Wind and Gas CC projects - Source: [21]

1.3.2 Wind Turbine Maintenance

Maintenance of WTs has been mostly implemented in two forms: corrective maintenance and preventative maintenance. Corrective maintenance applies a run-to-failure approach, which can be attractive as it requires very little planning. But running until failure means there is no control for the operator in terms of when or how the WT goes off-line. This means that there will be downtime associated with the time until the maintenance action is taken, which could have been prevented. Besides that, the unscheduled nature of the maintenance means that the repairs may need to be done at an inconvenient time, for example of high power production or during winter time for offshore wind farms. Hence, although corrective maintenance represents the least expensive maintenance strategy to implement and operate, repair costs of running to failure far outweigh the benefits in most cases [20].

On the other hand, preventative maintenance involves servicing an asset in order to prevent unscheduled downtime. Maintenance is performed on a periodic basis at a pre-determined interval, usually estimated with component-specific historical knowledge. This can be an efficient strategy for components that wear out in a predictable and repeatable pattern. While this reduces the severity of failures and maintenance tasks compared to corrective maintenance strategies, the component life may not be fully exhausted when maintenance occurs, meaning that in a majority of cases the asset has been brought off-line unnecessarily [20].

This means that both corrective and preventative maintenance are far from ideal, hence why Predictive Maintenance (PdM) is being researched as a possible solution. In fact, it's expected that by 2025 new PdM strategies can reduce the LCOE of onshore wind by as much as 25% [22]. These are related with reduced downtime from unplanned maintenance, helping further increase the utilization factor by 18% [22], which could result in more than 15% decrease in the LCOE as Figure 1.4 shows. Besides that, new PdM strategies are also expected to significantly reduce O&M costs and help increase the

lifetime of WTs, which have a significative impact in the LCOE, as Figure 1.4 also shows. In fact, new PdM strategies are one of the main long-term research challenges in wind energy [23].

1.4 Predictive Maintenance

Before proceeding, it's important to note that the terms fault and failure have been used rather loosely in the literature. In this work we will follow the definition from [24]. To summarize, a fault is a condition of a machine that occurs when one of its components or assemblies degrades or exhibits abnormal behavior. A fault can in turn lead to a failure which is a termination of the ability of an item to perform a required function. Failure is an event as distinguished from fault, which is a state [25].

With this in mind, the objective of PdM is to optimize maintenance by making use of Condition Monitoring (CM) techniques on various components. CM involves continuously monitoring the health or condition of a piece of equipment with the objective of detecting incipient faults early and to determine any necessary maintenance tasks ahead of failure.

1.4.1 Wind Turbine Technology

To understand what type of failures actually occur in a turbine it's important to understand its technology, in this section a brief review will be given, focusing on the aspects that are most relevant for the rest of the work. It's important to note that there are many WT designs, but due to being the industry standard we're going to focus on the horizontal axis turbine with pitch control and high speed drive train. A scheme of this WT design is illustrated in Figure 1.6.

The WT works by using the rotor blades to convert the kinetic energy of wind into mechanical energy in the rotor hub, which is connected to the drivetrain that will convert the mechanical energy into electrical energy. As can be seen in Figure 1.6, the drivetrain is composed by the main shaft, the gearbox and the generator. The main shaft transfers the mechanical energy from the rotor hub to the gearbox, which will convert from the frequency of the rotor to the generator/grid frequency. Finally, the generator will produce power, which is mostly proportional to the wind speed, but due the control systems of the WT it is normally characterized by a power curve. A reference power curve is illustrated in Figure 1.7, and it's important to go into further detail on the regime of when the wind speed reaches rated speed. When this happens, it is necessary to waste this excess energy in order to avoid damaging the WT. In practice, this is done by using the pitch system, which controls the angle of the blades, and when the wind is past the rated speed it turns the blades out of the wind, thus lowering the input energy. Another fundamental control system in the WT is the yaw system, which controls the nacelle direction in order to align it with the wind direction, thus maximizing the share of energy in the wind that will be running through the rotor area.

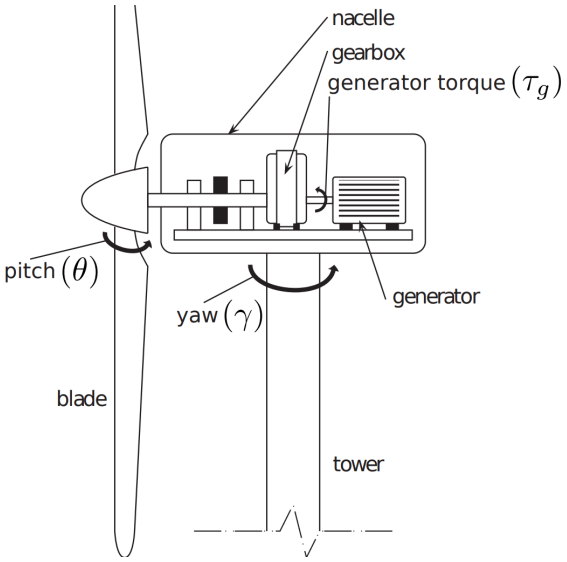


Figure 1.6: Horizontal Axis Wind Turbine with labelled main components. - Source: [26]

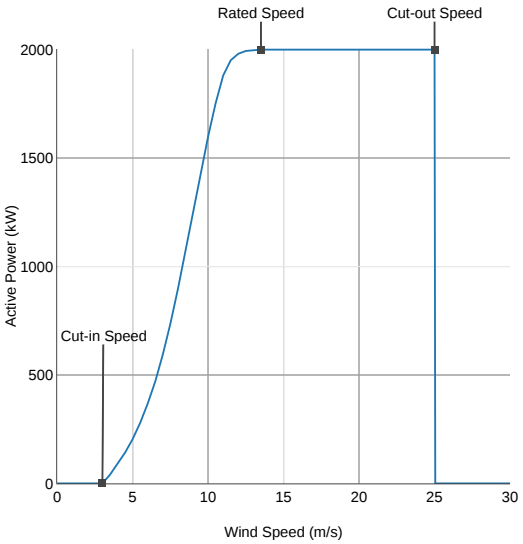


Figure 1.7: Reference Power Curve of a Wind Turbine

1.4.2 Wind Turbine Failures

Figure 1.8 shows the failure data of 350 modern offshore turbines [27]. As can be seen, the components with the highest rate of failures per turbine per year are the the pitch and the hydraulics system. In second place comes ”Other Components”, which includes ladders, hatches, lifts and door and nacelle seals. But it’s important to note that the failures in the previous components result mostly in minor repairs, on the contrary it can be seen that the generator and the gearbox have a significant amount of failures that led to major repairs and replacements. In fact, in [28] the authors performed an historical analysis of failure and downtime data of more than 4300 turbines. The results showed that the generator and gearbox were the most responsible components for downtime due to failures, which reduces the availability of the WT. Besides that, the authors in [27] also note that these are also some of the most expensive components to replace, with a gearbox replacement accounting to 230.000€ on average and 60.000€ for a generator replacement. For the previous reasons this work will focus on WT drivetrain failures, such as gearbox, generator and main bearing failures.

With this in mind, it’s now interesting to understand how can PdM reduce the LCOE of wind by predicting these drivetrain failures. In fact, if the failure of a component can be predicted months ahead, the planning of the repair or replacement is greatly improved as spare parts, equipment and personnel resources can be ordered in due time thus lowering the O&M costs. Besides that, due to this optimization, the downtime of the turbine is also kept at a minimum, which increases the utilization factor. Adverse weather conditions may also prevent repairs for extended periods of time, and it may be advantageous to react to a fault warning early to avoid long periods of downtime following a critical failure in a severe weather window. Finally, if faults are corrected at an early stage, secondary damage to equipment can be avoided, which can help prolong the lifetime of the turbine. But to actually implement a PdM strategy there must be a CM technique.

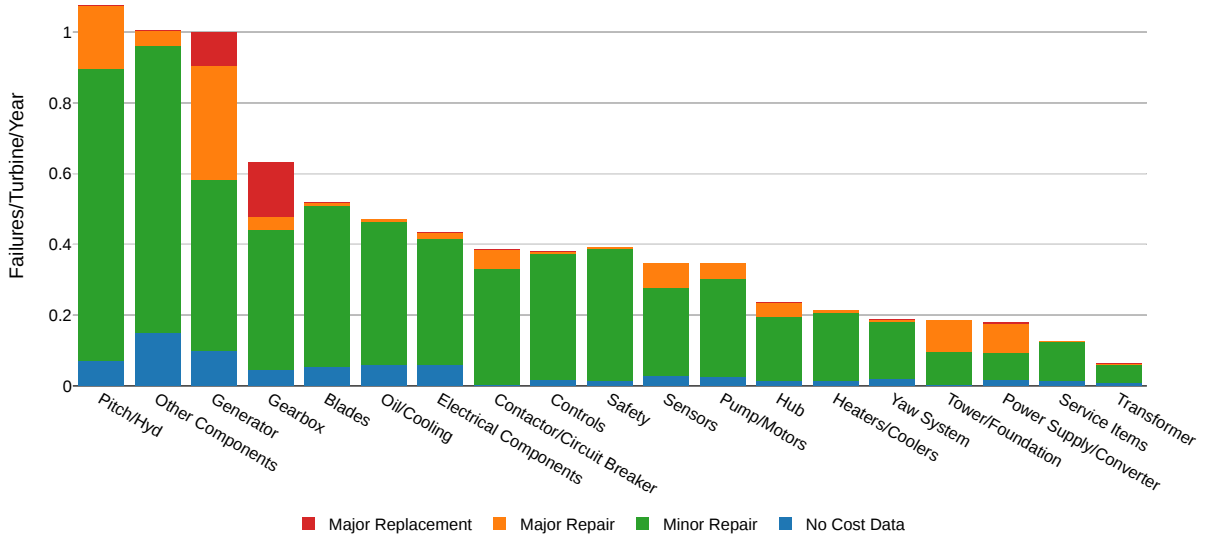


Figure 1.8: Failure rates of components of offshore turbines - Source: [27]

1.4.3 Drivetrain Condition Monitoring

In the wind industry, CM has traditionally been carried out by installing additional vibration, oil-particulate, or other sensors to the turbine. These systems are normally known as Condition Monitoring Systems (CMS). The sensors can be expensive to install and record data at very high frequencies, meaning there are high bandwidth and storage requirements. Although these systems have seen significant success in the oil and gas industries, the cost justification for wind turbines is not nearly as strong . Furthermore, due to the fact that wind turbines operate at relatively low and variable speeds, the signals are harder to interpret and have not been as successful at fault prognosis as in other industries [20].

However, all large utility scale WTs have a standard Supervisory control and data acquisition (SCADA) system principally used for performance monitoring. Such systems provide a wealth of data at normally 10 minute resolution, though the range and type of signals recorded can vary widely from one turbine type to another. As CM using SCADA data is a potentially low cost solution requiring no additional sensors, there's been a significant amount of research on how to best use this data for fault detection in WTs [29]. Having this in mind, in the next section the literature on SCADA-based WT drivetrain fault detection will be reviewed.

1.5 State of the Art

The SCADA system provides two types of data: alarms and signals. SCADA alarms can be used for condition monitoring, but as has been noted in the literature these alarms are currently too frequent for rational analysis [30]. This happens because SCADA alarms are simple thresholds to the signals, which is a very crude way to detect fault because the WT operates at varying operating regimes. For these reasons there's been research in developing more advanced data analysis methods for SCADA signal data of which a review will be done in this section.

The most common approach for SCADA-based CM is trough the method of bins for power curve

monitoring, defined in the standard IEC 61400-12-1 [31]. Briefly, the objective of this approach is to build a model of the power curve and monitor real time data to detect deviations from the model. More advanced methods to model the power curve have been researched, such as K-Nearest Neighbors (KNN) [32], Artificial Neural Networks (ANNs) [33], and Gaussian Processes (GPs) [34]. These have been successful in finding many types of faults related with underperformance of the WT such as pitch misalignment, yaw misalignment and blade icings. But in terms of detecting faults in the drivetrain of the WT the results haven not been as promising. The majority of the work that has been done has shown that sometimes there is some underperformance in the WT before the fault, but the high amount of false positives and the impossibility of performing fault diagnosis, since we're only looking at the output power, make this approach difficult [28, 35].

There has also been some research regarding the applicability of classification algorithms to predict failures. By using different sensors of the WT as features of a classification algorithm and the failures as labels, it's expected for the model to learn to predict failures. For example, in [36] and [37] the authors used Support Vector Machines (SVM) to predict generator faults. Although the results were promising, the unbalanced nature of the data resulted in less than ideal overall results. Besides that, the prediction window was at most 12 hours, which is very small for it to have an actual impact on PdM strategy. But the main disadvantage of this approach is that it requires labelled data to actually be able to train the classification algorithm in a supervised manner. As mentioned throughout this work, the existence of clear WT labels is extremely rare in the wind industry, which makes the classifier-based approach challenging.

Finally, the most promising method for SCADA-based drivetrain fault detection are Normal Behaviour Models (NBMs). They are being actively researched by both the research community and the industry. It should also be noted that, during the initial phase of the present work, different approaches that didn't use NBMs were tested. These approaches were based on unsupervised algorithms, which have the main advantage of not needing to assume that the training data belongs to a period of normal behavior. The first approach was regarding fault detection as a time series anomaly detection problem. Here, different anomaly detection algorithms were used, such as Seasonal Hybrid Extreme Studentized Deviate Test [38] and Matrix Profile [39]. The second approach used different statistical distances, such as Earth Mover's Distance [40] and Quadratic Chi-square Distance [41], to compare the distribution of temperatures between WTs to find anomalies. Nonetheless, their results in terms of actual fault detection were considerably worse when compared to NBM-based approaches. Hence why the focus of the thesis shifted towards the latter. The next section will provide a brief review of trending, which provided the fundamental reasoning behind NBMs.

1.5.1 Trending

The fundamental assumptions of NBMs come from a previous method that was used in the industry called trending. Trending is based on the fact that faults in a component can relate to changes in its efficiency. In turn, these changes in efficiency can be detected by changes in the thermodynamic

behaviour of the system. The authors of [42] and [43] demonstrated this for a rotating mechanical system of the wind turbine, such as the bearings of drivetrain components.

The first law of thermodynamics states that changes in the internal energy ΔU can derive either from heat supplied to the system Q or work done by the system W, as Equation 1.3 shows. Under quasi-stationary conditions, it can be assumed that the internal energy does not change, resulting in Equation 1.4.

$$\Delta U = Q - W \tag{1.3}$$

$$Q = W \tag{1.4}$$

The work done by the system can be described as the difference of kinetic energy E_k taken from and supplied to the system, as Equation 1.5 shows. Since these components are mainly transferring energy, the output energy can be described by the input energy multiplied by the efficiency of the bearing η, as shown in Equation 1.6. Since the kinetic energy is rotational energy, it can be described by Equation 1.7 with inertia I and angular speed ω. Thus resulting that the work done by the system can be described by Equation 1.8

$$W = E_{\mathrm{k,out}} - E_{\mathrm{k,in}} \tag{1.5}$$

$$E_{\mathrm{k,out}} = \eta E_{\mathrm{k,in}} \tag{1.6}$$

$$E_{\mathrm{k}} = \frac{1}{2} I \omega^2 \tag{1.7}$$

$$W = (\eta - 1) I \omega^2 \tag{1.8}$$

The heat flow can be described by Equation 1.9 with the temperature change ΔT and the heat transfer rate k for the material compound. Combining Equations 1.4, 1.8 and 1.9 it is possible to derive the temperature change in the bearing, as Equation 1.10 shows.

$$Q = -k \Delta T \tag{1.9}$$

$$\Delta T = (1 - \eta) \omega^2 \frac{I}{2k} \tag{1.10}$$

Since $\frac{I}{2k}$ is constant, it results that ΔT is a function of η and ω^2. It's also important to note that ω^2 is proportional to the produced power P. This results that monitoring the temperature measure by the sensor, T, to P ratio makes it possible to detect changes in the efficiency of the bearing η, and thus possible faults. In fact, many Wind Farm Owners, Original Equipment Manufacturers and

Independent Service Providers are already using SCADA temperatures and binning them by active power to monitor the health of the WT components. However, this simple method does not consider more complex thermodynamic relations that exist in the drivetrain of the WT, such as the seasonal effect of ambient temperature, the heat transfer between different components and the effect of the lubrication and cooling systems. Due to these challenges, there has been research in NBMs, which are able to model temperatures based on various input signals.

1.5.2 Normal Behavior Models

One of the most important aspects behind training a NBM are the input features. If these are carefully selected and the model is trained with data from a period when the turbine is healthy then it will learn the normal behaviour of the WT, being robust to different ambient temperatures and operational regimes. Afterwards, it is possible to compare the predictions of the NBM with the real measured values, to understand if the component is hotter or cooler than our model expected. The difference between the measured value and the prediction make it possible to detect anomalies in the component and possible faults. This process is summarized in Figure 1.9, where it should be noted that the differences between the predictions and the measured values are normally called the residuals.

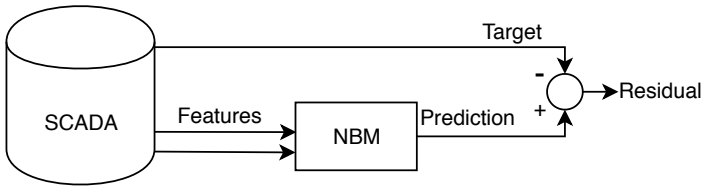


Figure 1.9: Normal Behaviour Model Testing diagram

One of the first works to apply NBMs to detect faults in the WT drivetrain was SIMAP [44]. This framework used ANNs to model various gearbox temperatures, such as the bearings and the cooling oil. In terms of the input features, the authors used lags of the target, thus including autoregressive properties into the model. The authors also used active power, nacelle temperature, and the cooling fan speeds. These features were selected using cross-correlation and impulse response analyses. Moreover, there were no results of a detailed case study to evaluate the quality of the NBM. Following works using similar feature selection and engineering techniques and also using ANNs showed more interesting results. For example, in [45] the authors found overheating problems in the gearbox cooling oil, detected almost six months before the failure of one turbine. In similar works [46] and [47] the authors were able to detect cracks in the gearbox body and the consequent oil leaks in gearboxes with four months in advance. In [48] the authors modelled the main shaft rear bearing temperature in direct-drive turbines and detected a fault three months ahead of failure.

All the mentioned works have used autoregressive features. In fact, [49] was work that assessed the impact of using autoregressive features. The authors used ANNs to model the generator bearing temperature and used as input features the generator stator temperature, nacelle temperature, active power

and the lagged target for the autoregressive model. The results showed that the temperature modelling performance of the autoregressive model was significantly better. But in terms of fault detection performance, the non-autoregressive model performed best, detecting faults up to five days earlier than the autoregressive model. Indeed, it's important to clarify that NBM can be evaluated in terms of how well they model the target temperature and in terms of how well they detect faults. The first corresponds to evaluating how low are the residuals during healthy periods of time, while the second corresponds to evaluating if there is an increase in the residuals before a failure. What the results from this work seem to indicate is that autoregressive features do increase the temperature modelling performance but worsens the fault detection performance. Other works such as [50] and [51] have reached similar results. But there have also been works that obtained the opposite results, where including autoregressive features improved both the temperature modelling performance of the model and the fault detection performance, such as [52, 53].

The first work to mention that features that are highly correlated with the target increase the temperature modelling performance, but in turn they may decrease the fault detection performance of the model was [54]. The authors tested the use of the gearbox oil temperature as a feature to model the gearbox bearing temperature, and it led to the model not being able to predict the failure. Afterwards, in [55] the authors compared between using features that are highly correlated with the target and not using them. The results seemed to indicate that using features that are highly correlated with the target resulted in lower fault detection performance, but the argument could not be fully assessed based on these results.

A summary of this literature review is presented in Table 1.2, where it is shown which input features each work used in NBMs for the gearbox bearing temperature. The algorithm used in each work is not represented in the table because all these works have used ANNs or ANN-based algorithms. This may be due to the existing domain knowledge of the non-linear relationships between the input features and the target. Since ANNs are known for their highly non-linear modelling capabilities, they are good candidates. Nonetheless, these works have also criticized their high computational costs and time-consuming optimization. This raises the need for other solutions. For example, Gradient Boosting Machines (GBMs) are also known for their highly non-linear modelling capabilities, while having considerably lower computational costs and being more robust to hyperparameter optimization [56]. In fact, GBMs have obtained excellent results in time series modelling challenges, as shown in works [57] and [58]. This makes them potential candidates for normal behavior modelling algorithms.

Regarding Table 1.2, it shows that there is a lack of consensus regarding which input features should be used, existing significant variation between works. Although some works have shown results which indicate that using features that are highly correlated with the target, such as the gearbox oil temperature, significantly reduces the fault detection performance of the model, these are still used in many works today such as [25, 50, 59, 60]. There are also conflicting opinions regarding the use of autoregressive features, which leads to some works using them and others not. The main reason behind this is the lack of consistent case studies that evaluate the impact of different features on both the temperature modelling and fault detection performances.

It's also important to note that, in most machine learning problems, if the model is well optimized

Table 1.2: Input features used in NBMs for Gearbox Bearing Temperature for selected works in the literature.

Work	Ambient Temp.	Rotor Speed	Active Power	Nacelle Temp.	Gearbox Oil Temp.	Autoregressive
[44]			X	X		X
[45]			X	X		X
[61]	X	X	X	X	X	
[62]		X		X	X	
[54]		X	X	X	X	X
[50]	X	X	X	X	X	X
[25]	X	X	X	X	X	
[63]	X					
[64]			X	X		
[60]		X	X		X	
[59]	X	X	X	X	X	X

and regularized, then the use of more features is expected to either lead to the same results or better. Indeed, in fault detection using NBMs this is not so trivial. To understand why, it's important to remind that the model that is being trained is the NBM, which minimizes the temperature modelling error. Indeed, it's expected that if the model has more features then the temperature modelling error should be the same or better. But the objective of building the NBM is to perform fault detection, and since the model is not being trained to minimize the classification error, then the impact that different features have on the fault detection performance is not trivial.

1.5.3 Fault Detection

There have also been works that used a classification-based approach towards WT fault detection, instead of using NBMs. In these classifiers the more features the model has the better its performance should be. But as the literature has shown, this approach has not been as successful. One of the main reasons behind this, is due to the lack of quality regarding groundtruth data of WT failures. Indeed, there is data of when the failure happened, but there is no information regarding when the fault state started, making it not trivial to formulate as a classification problem. Hence why the regression-based approach of building a NBM has been prevalent. NBMs also face difficulties with the nonexistence of groundtruth. Indeed, the evaluation of the temperature modelling performance of the NBM is coherent in the literature, but the evaluation of fault detection performance is not. This happens again due to the lack of clear groundtruth, which leads to the majority of the literature to evaluate fault detection results by visual inspection, observing the increase in the residuals before the failure. But this is problematic, because comparisons between different models will be highly subjective.

To perform fault detection using NBMs one has to analyse the residuals, which are the difference between the real values and the model predictions. As has been noted in [45], the raw residuals of the NBM are highly fluctuating, having several spikes, which makes it difficult to define a threshold to detect faults. The authors of [49] suggested using daily averages of the residuals, thus obtaining less fluctuating values, that represent how abnormal was the target temperature in each day. In [50], the authors suggested using rollings averages of the residuals, instead of resampling, which has the

advantage of not losing the resolution of the data while still smoothing the values.

Even though there have been significant advances in the post processing of the residuals, the process of detecting faults in the majority of the works is still similar. This process consists of doing a visual analysis of the residuals and observing if there is a significant increase before the occurrence of a failure. This method is both time-consuming and ambiguous, due to being performed in a manual approach. Besides that, many authors simply look at the residuals without actually inspecting the time series and understand why the predictions are differing from the true values.

There has been some work towards automating the process of fault detection. For example in [25], the authors formulated the detection of faults as a classification problem, but few details were given regarding what labels were considered. The clear definition of the classification problem and corresponding labels is extremely important, because it defines what is considered a True Positive (TP) and a False Positive (FP). The authors also used the Receiver operating characteristic (ROC) curve to evaluate the different models, which is questionable since ROC curves are not adequate for unbalanced classification problems such as fault detection, in which the quantity of fault states is notably smaller than the quantity of no-fault states. Indeed, there is no standard method to evaluate the fault detection performance of an NBM in the literature.

1.6 Objectives

Having in mind the challenges explored in the previous section, the main objectives of this work will be the following:

- Assess the use of GBMs as a low computational cost and higher interpretability solution for modelling the normal behaviour of the WT drivetrain temperatures.
- Develop an input feature taxonomy based on the causal relations between the features and the target to categorize different input feature types.
- Based on the developed taxonomy, evaluate the impact of the different feature types on the temperature modelling performance and the fault detection performance of the NBM.
- Formulate the detection of faults based on residuals as a classification problem and develop the corresponding evaluation framework. Compare the results from the framework with those from the visual residual analysis.

1.7 Thesis Outline

The main body of the thesis has the following structure:

- **Chapter 1** contextualizes the thesis, reviews the state-of-the-art and presents the motivation for the development of the present work.

- **Chapter 2** presents the methodology behind the feature taxonomy, GBMs and the fault detection framework.
- **Chapter 3** explores the data of the wind farm and the implementation of the data processing, feature configuration, temperature modelling and fault detection frameworks.
- **Chapter 4** presents the results in terms of temperature modelling and in terms of fault detection for the NBM with the different input feature selection methods.
- **Chapter 5** summarizes the findings from the thesis and presents possibilities for future work.

There are also several appendices, their content is described below:

- **Appendix A** contains extra case studies from the temperature modelling results.
- **Appendix B** contains extra residuals from the fault detection case studies.
- **Appendix C** contains the complete precision and recall results for the different models and predictive windows.
- **Appendix D** contains the paper that was accepted for presentation at the ICML workshop: ”Climate Change: How can AI help?”.

Chapter 2

Methodology

2.1 Feature Taxonomy

As mentioned in the Introduction, there have been works which indicated that using features that are highly correlated with the target increases temperature modelling performance but may decrease the fault detection performance of the model [54, 55]. For example, if the gearbox bearing temperature is being modelled, it's expected that when it gets hotter the gearbox oil temperature will also get hotter, due to heat transfer. Indeed, it is intuitive that the gearbox oil temperature is highly correlated with the gearbox bearing temperature and thus will be important for modelling. But the problem is that a state of fault is characterized by overheating in the gearbox bearing temperature, and again, due to heat transfer, the gearbox oil temperature will also be hotter than expected. This means that if the gearbox oil temperature is being used to model the gearbox bearing temperature, the model may model abnormal behavior and thus not be able to detect the incipient fault. Having this in mind, the present work hypothesizes that the decrease in fault detection performance is not due to the fact that the features are highly correlated with the target temperature, which is subjective since there is significant correlation between rotor speed and the gearbox bearing temperature. In fact, we suggest that what decreases the fault detection performance is using features that have an interdependence with the target, as in not only is the target dependent on them, they are also dependent on the target. This happens for all temperatures in the drivetrain, since there is heat transfer between all of them.

Although the ideas behind the previous hypothesis are intuitive, it's important to make it more objective by using a clearer nomenclature. For this reason, we will present a new taxonomy based on Econometric Causality [65] and Causal Inference [66], which distinguishes features based on their causal relations with the target. Basically, if the target is causally dependent of the features, they are considered causal features. On the other hand, if there is a causal interdependence between the feature and the target, as in they are dependent on each other, they are considered simultaneity features.

As was shown in the previous example, these causal relations can be assumed since we have domain knowledge of the physical system. The present work suggests the causal diagram presented in Figure 2.1, where the arrows represent causal relations. If the arrow is double-pointed, it means that

there is interdependence between the variables. It should also be clarified that mediators are also causal features. The only difference is that they are not the original cause, in fact they mediate the causal effect from other causal features. For example, the gearbox bearing temperature depends on the rotor speed, but the rotor speed depends on the wind speed. This means that the origin of the causal effect is the wind speed, thus meaning that this is a causal feature, while the rotor speed is a mediator of this causal relation. For the purpose of the work, mediators will be considered causal features, but there is interest in further exploring how this intricacies can affect the NBM both in terms of temperature modelling and fault detection performances. These causal relations of the different variables are also summarized in Table 2.1.

The introduced taxonomy clarifies some details from previous works. For example, usually the nacelle temperature was not considered a feature highly correlated with the target, and the works that criticized the use of highly correlated features actually used the nacelle temperature [54]. According to this taxonomy, both the nacelle temperature and the gearbox oil temperature are simultaneity features, which means that if the gearbox oil temperature does negatively affect fault detection performance, then it's expected that nacelle temperature also does.

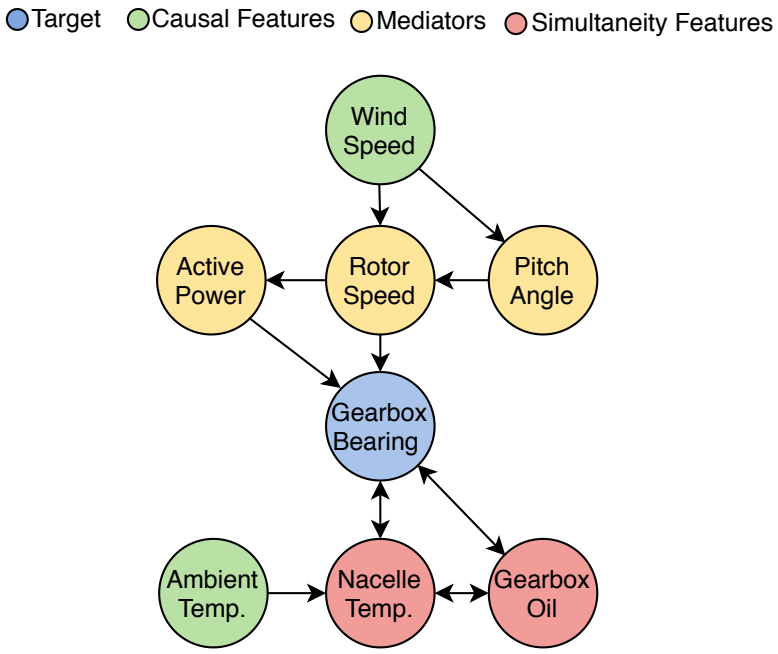


Figure 2.1: Causal diagram for the gearbox bearing temperature.

Regarding the use of autoregressive features, these are clearly causal features, since there is a temporal relation that prevents the future values from affecting past values. Nonetheless, as noted in the literature, the impact of these features in model performance is not consensual. For example, if there is a developing fault and the gearbox bearing is overheating, then the use of previous values that were already hotter than they should be to model the present temperature may result in the NBM modelling abnormal behavior. In fact, using autoregressive features in an NBM can also be seen as what is being

Table 2.1: Causal relations for the different variables.

Variable	Causal Relation
Gearbox Bearing Temperature	Target
Wind Speed	Causal
Ambient Temperature	Causal
Pitch Angle	Causal/Mediator
Active Power	Causal/Mediator
Rotor Speed	Causal/Mediator
Nacelle Temperature	Simultaneity
Gearbox Oil Temperature	Simultaneity

modelled is not the actual gearbox bearing temperature, but the temperature rate of change. Since the model has knowledge of the previous timestamps, it is expected to learn the normal rate of change of temperature. Thus, the model should be mostly insensitive to changes in the mean value of the target temperature. This is problematic if the fault is characterized by a change in the mean temperature for the given conditions, but if the fault is also characterized by an abnormal rate of change in the temperature then the autoregressive model should be able to detect it. This indicates that the impact of autoregressive features may depend on the type of fault, which may also explain the non-consensual results obtained in the literature.

In most machine learning problems, if the model is well optimized and regularized then with more features it is expected to either obtain the same results or better. In fault detection using NBMs this is not so trivial. To understand why, it is important to remember that the NBM minimizes the temperature modelling error. Therefore, it's expected that the more features the model has the temperature modelling error should be the same or better. But the objective of the NBM is to perform fault detection, and since the model is not being trained to minimize the classification error, then the impact that different features have on the results is not trivial. This shows why choosing a domain knowledge-based approach in terms of selecting the input features to the model is specially important in this type of problem.

2.1.1 Model Types

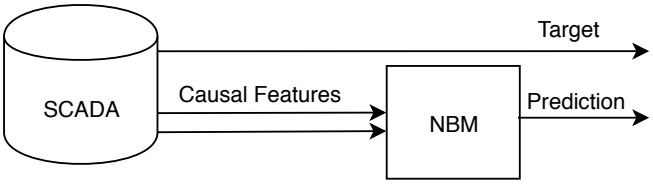
Based on the taxonomy that was previously presented to distinguish features based on their causal relations with the target, we will also define different model types. This formulation will help when evaluating the impact of the different input feature types on the model results. These models are the Causal Normal Behaviour Model (CNBM), Simultaneous Normal Behaviour Model (SNBM), Autoregressive Causal Normal Behaviour Model (ACNBM) and Autoregressive Simultaneous Normal Behaviour Model (ASNBM). Their differences are shown in Table 2.2. In Figure 2.2 are presented diagrams for two of the model architectures to clarify their differences.

2.2 Gradient Boosting Machines

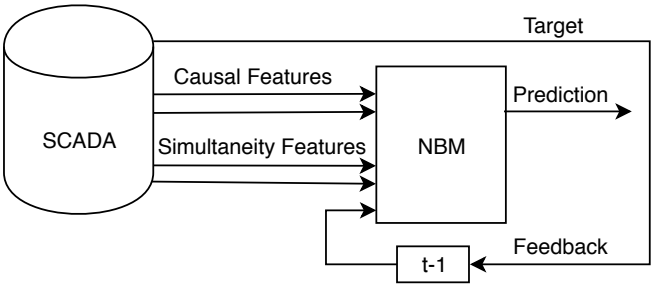
The NBMs in this work will be based in GBMs. GBMs are a machine learning technique which uses a prediction model in the form of an ensemble. This means that it combines multiple simple models into

Table 2.2: The input feature configuration used by each model type.

Model	Causal Features	Simultaneity Features	Autoregressive Features
CNBM	X		
SNBM	X	X	
ACNBM	X		X
ASNBM	X	X	X



(a) CNBM



(b) ASNBM

Figure 2.2: Schematic examples of some of the different NBM input feature types.

a single composite model. In boosting terminology, the simple models are called weak learners. In this work, as is standard for most problems, the weak learners will be decision trees.

In [67] the authors explain the fundamentals of GBMs, based on which a brief review will be presented in this chapter. Given the input features vector x, the ensemble $F_M(x)$ will be composed of M weak learners of the form $f_m(\mathrm{x})$, which will have the predictions $\hat{y}$ for the target y. This is formalized in equation 2.1. To understand the process of building the ensemble we can imagine that the starting weak learner simply predicts the mean of the observations, such that $f_0(x) = \overline{y}$. That means the residuals for that iteration will be given by Equation 2.2. Now, another weak learner, $f_1(x)$, will be added to the composite model, and ideally it would make $F_1(x)$ able to predict y as Equation 2.3 shows. For this to happen, the new weak learner must be equal to the residuals, as demonstrated in Equation 2.4. Of course that in practice the model would need more weak learners, and it would never truly be equal to the target values, but the main idea is that in GBMs the added models are trained on the residuals of the previous model. For the general case it can be summarized as in Equation 2.5, in which it's also been added the learning rate η, which is an hyper parameter used to prevent over fitting, so that each added weak learner has less of an effect on the composite model. It's also important to mention that it can be formally proven that GBMs perform gradient descent. The intuition behind this is related with the fact that Equation 2.6, which corresponds to the gradient descent position update equation, has similarities to Equation 2.5 that was previously explained.

$$F_M(x) = \sum_{m=0}^{M} f_m(x) = \hat{y} \tag{2.1}$$

$$r_0 = y - f_0(x) = y - \overline{y} \tag{2.2}$$

$$F_1(x) = f_0(x) + f_1(x) = y \tag{2.3}$$

$$f_1(x) = y - f_0(x) = r_0 \tag{2.4}$$

$$\hat{y}_m = \hat{y}_{m-1} + \eta r_{m-1} \tag{2.5}$$

$$y_t = y_{t-1} - \eta \nabla f\left(y_{t-1}\right) \tag{2.6}$$

2.2.1 Temperature Modelling Evaluation

To evaluate the temperature modelling performance of the NBMs the approach that will be used in this work is the one that is standard in the literature. The models will be evaluated on the test set, where periods of fault were removed. The motivation behind this is that when evaluating temperature modelling we are interested in how well the NBM models the temperature during normal behavior. This means that fault periods would not be relevant to this evaluation. In fact, since periods of fault are characterized by higher residuals, they would actually influence the regression metrics and introduce some subjectivity in the results.

As is standard in the literature, the regression metrics that will be used in this work are the Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE) and the Standard Deviation (SD). These metrics will be defined in function of the residuals, which are described in Equation 2.7. The regression metrics are defined in function of the residuals in Equations 2.8, 2.9 and 2.10. The reason behind using different metrics is because they all provide different insights. For example, it is known in the literature that the MAE is more robust to outliers than the RMSE [68]. Also, as noted in [62], the use of the SD of the residuals is also a meaningful regression metric for temperature modelling performance since it is expected that the residuals of an healthy turbine have an approximately normal distribution.

$$r_i = \hat{y}_i - y_i \tag{2.7}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |r_i|}{n} \tag{2.8}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} r_i^2}{n}} \tag{2.9}$$

$$\mathrm{SD} = \sqrt{\frac{\sum_{i=1}^{n} |r_i - \bar{r}|^2}{n}} \tag{2.10}$$

The majority of the literature compares the temperature modelling results between different works, but it's important to remind that this comparison is not free of subjectivity. Indeed, in the area of using NBMs for WT fault detection there is no standard dataset to compare the performance of different models. The reason behind this is due to the inherent confidentially of the majority of data in this area. Even though all works use different data, it is still possible to compare their results since all the data comes from the SCADA system, thus meaning that the difficulty of modelling the temperature of the components should be the same independent of the WT, apart from rare exceptions.

2.3 Fault Detection Framework

The process of detecting faults using NBMs starts with building the NBM and then analysing the residuals to understand if it was possible to predict the failure beforehand. Normally this is done by visual residual analysis which provides a qualitative score. In this work, a more automatic approach which provides a quantitative score is proposed. The overall process is presented in Figure 2.3.

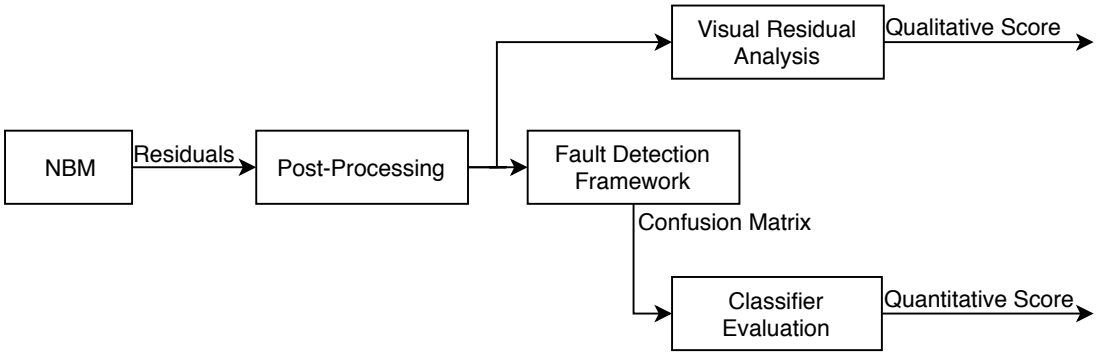


Figure 2.3: Diagram of the fault detection methods.

2.3.1 Residual Post-processing

As the literature has shown, the raw residuals that can be calculated from the implemented NBMs are highly fluctuating. For this reason, different post-processing techniques have been explored to make the residuals more insightful, the most used one being the rolling mean of the residuals. The rolling mean RM for the timestamp j, window W and signal x is described by Equation 2.11. The main advantage of using the rolling mean is that it smooths the values of the residuals but still has the same frequency as the original data, on the contrary to the resample transformation. This will be the method used in this work.

$$RM_j = \frac{1}{W} \sum_{i=0}^{W-1} x_{j-i} \tag{2.11}$$

2.3.2 Fault Detection Framework

For the visual residual analysis, the post processed residuals are sufficient to provide a qualitative analysis. But for the fault detection framework one must obtain alarms from the residuals. Indeed, this can be done by simply using thresholds, in which all the values above the threshold are alarms. But in the context of predictive maintenance it's not necessary to know if there is an alarm at every sample of the signal. In fact, for this work knowing if there is an alarm each day is sufficient. For this reason, the alarms that will be used in the framework will be generated from a daily mean resample of the post processed residuals. To understand the labels with which the alarms will be compared, one must first formulate the detection of faults as a classification problem.

2.3.2.1 Classification Problem Formulation

To develop an evaluation framework for fault detection, one must first formulate it as a binary classification problem where there are two labels: fault and no-fault. But these labels must be defined based on the data available, which comes from the maintenance logs of the wind farm owners, and contains only the date of failure. Indeed, there is no information regarding the fault state of the component, only the date of when it failed. Having this in mind, it was defined with the wind farm owners that for the type of failures discussed in this work it can be assumed that a fault state would be present at most 90 days before the failure. This means that any alarm triggered earlier than 90 days before the failure will be considered a FP. To be considered a TP, the alarm must be triggered less than 90 days before the failure, but also more than Δt days before the failure. The reason for Δt is that predicting a failure on the same day it happens is not necessarily relevant. In this work, Δt will be a variable which will allow us to compare how early different models are able to predict the failure.

Figure 2.4 presents a schematic example of the previously described fault detection classification problem formulation. Taking this example, it's important to note that the number of alarms triggered in the prediction window is not relevant, they are all aggregated as one TP. The main reason for this, is that if the aggregation is not done, then four alarms for the same failure would count as much as four detected failures with one alarm each. This clearly isn't what is intended of the framework, since one alarm should be enough to motivate an inspection, and detecting four failures with one alarm weights more than detecting one failure with four alarms. Finally, it's also important to note that alarms triggered less than Δt days before the failure are not considered FPs, since there is indeed a fault state, it simply isn't relevant, so they are considered True Negatives (TNs).

The output of the fault detection framework will be a confusion matrix, which contains the total number of TP, FP, TN and False Negative (FN). The confusion matrix for the schematic example in Figure 2.4 is presented in Table 2.3. It should be noted that for this example it's being considered that the time range is one year, which corresponds to 365 days. In these total number of samples, there are three false positives, one true positive, zero false negatives and 361 true negatives.

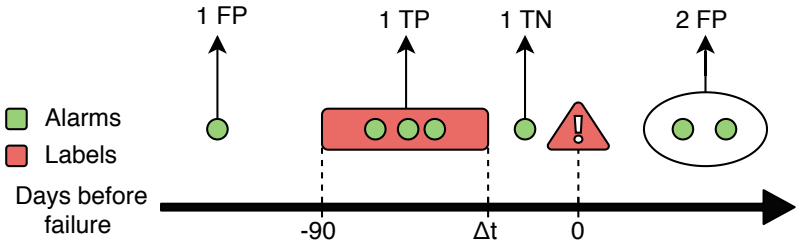


Figure 2.4: Schematic example of the fault detection classification problem formulation.

Table 2.3: Example Confusion Matrix

	Predicted: No Fault	Predicted: Fault
Label: No Fault	TN = 361	FP = 3
Label: Fault	FN = 0	TP = 1

2.3.3 Classifier Evaluation

As previously mentioned, the majority of the literature only does visual residual analysis, but there have been works that evaluated the results as a classification problem. For example, in [25] the author used the ROC curve to evaluate the performance of each model. But it's also important to remind that this problem is clearly an unbalanced one, since the number of no-faults labels is notably higher than the number of fault labels. From this results that the use of the ROC curve and other more common metrics such as accuracy are not adequate, since they are not robust to unbalanced data. On the other hand, metrics such as precision and recall are significantly more robust to unbalanced data [69]. The definition of precision and recall is presented in Equations 2.12 and 2.13, and it should be noted that they have an interpretable meaning. Precision corresponds to how many of the triggered alarms were true, and recall corresponds to how many of the failures were detected. For example, for the previous confusion matrix one can calculate the corresponding precision and recall: $P = 0.25$ and $R = 1$. It's also important to mention the Area Under the Curve (AUC), which corresponds to the area of the precision and recall curve, thus allowing to compare different precision and recall curves using just one metric. Since the precision and recall curve is a discrete function, the AUC can be calculated by Equation 2.14, where k corresponds to the index of the different thresholds of the framework.

$$P = \frac{tp}{tp + fp} \tag{2.12}$$

$$R = \frac{tp}{tp + fn} \tag{2.13}$$

$$AUC = \sum_{k=1}^{N} P(k)\Delta R(k) \tag{2.14}$$

Chapter 3

Data and Implementation

3.1 Data

Before proceeding to the data description it should also be noted that all the work developed in this thesis was done using Python 3. Regarding data processing, the main library used was Pandas [70]. In terms of data visualization, to produce the figures in the present work, Plotly [71] was used. Libraries used for other tasks will be referenced throughout the next sections.

The data used in this work comes from a wind farm composed of 16 turbines, from the beginning of 2008 to the end of 2013. The turbines have a rated power of 1.5MW and their generator is of type Double-fed Induction Generator (DFIG). The collected data are SCADA signals with ten minute resolution, the relevant signals related with the operation of the WT and the drivetrain are shown in Table 3.1. It should be noted that all the signals were recorded with a 15 minute frequency. Also, during the years of 2012 and 2013 there were a total of nine failures in the drivetrain of the WTs, related with the gearbox bearing. For these reasons, this will be the component for which a NBM will be trained, with the objective of predicting the corresponding failures.

Table 3.1: SCADA signals related with the operation of the WT and the drivetrain

Sensor	Unit
Active Power	W
Rotor Speed	rpm
Wind Speed	m/s
Pitch Angle	deg
Ambient Temperature	
Nacelle Temperature	
Main Bearing Temperature	
Gearbox Bearing Temperature	^{o}C
Gearbox Oil Temperature	
Generator Bearing Temperature	
Generator Phase Temperature	

3.1.1 Input Features

The WT operates under various regimes, the objective is to use input features that allow the NBM to model them. As was shown in the State-of-the-art, temperature changes in the bearings are highly dependent of the rotor speed, so this is an essential input feature for the model. But since DFIGs have control strategies for the rotor speed [72], it means that rotor speed alone can't completely explain the regimes of the WT. This is supported by Figure 3.1, where it's possible to see that the rotor speed stops increasing at 16 rpm while the active power continues increasing. This is due to the control system of the DFIG. This regime can have influences in the drivetrain temperatures, and to capture it it's necessary to use the active power as a feature. The blade pitch angle is used in the WT control system, thus being an essential feature to capture the full mode of operations. Both the rotor speed and pitch angle are only mediators, indeed the original causal feature is the wind speed, thus meaning it might also contain relevant information for the model. Finally, another important feature which is responsible for most of the false positives in simple models is the ambient temperature, those that don't use it have a significant higher amount of alarms in the summer.

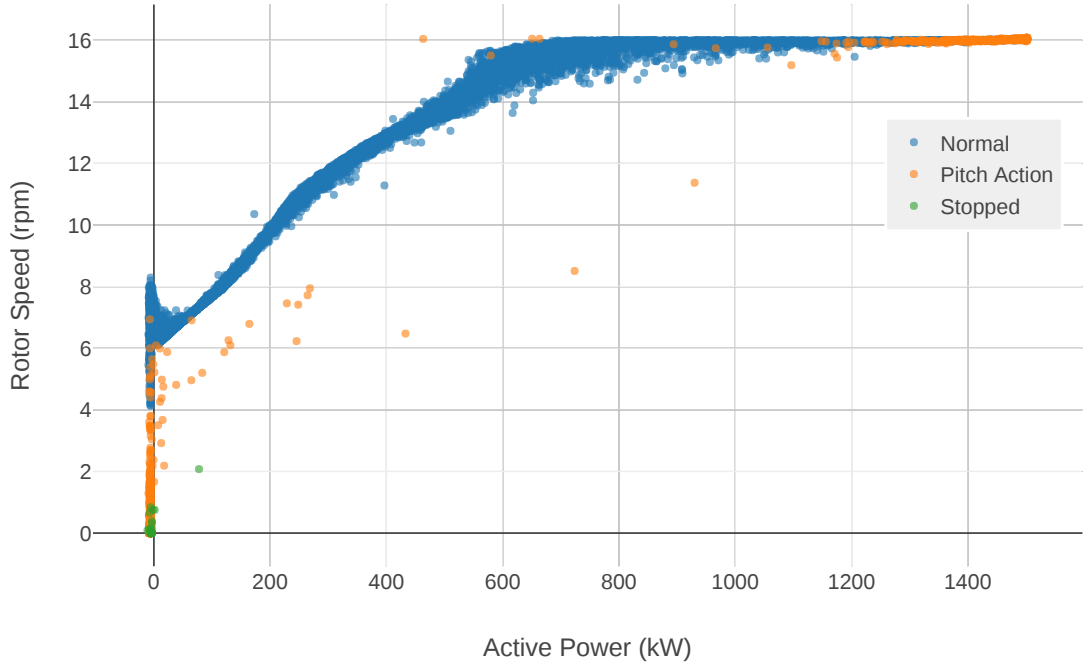


Figure 3.1: Scatter plot of active power and rotor speed during different operation regimes of a WT.

Regarding the simultaneity features, since the ones most used in the literature are nacelle temperature and gearbox oil temperature, these will be the ones tested in this work. So, having in mind the features that were selected an the model types defined in the State-of-the-art, the models that will be tested in this work are presented in Table 3.2.

Table 3.2: The defined models and the corresponding input features.

Model	Causal Features	Nacelle Temp.	Gearbox Oil Temp.	Autoregressive
CNBM	X			
SNBM1	X	X		
SNBM2	X		X	
ACNBM	X			X
ASNBM1	X	X		X
ASNBM2	X		X	X

3.2 Normal Behaviour Model

As noted in the Methodology, the NBMs will be implemented using GBMs, of which there are various open source implementations. The most efficient ones are XGBoost [73], LightGBM [74] and CatBoost [75]. Most of the benchmarks between these three implementations [76, 77] reach the same result: there is no clear winner, it highly depends on the problem at hand. CatBoost is known to have better results when there are categorical features, which are not used in this work. Between LightGBM and XGboost, the former is known to have a better computational performance. For these reasons, LightGBM will be used in this work. Regarding the evaluation of the temperature modelling results, we will use the scikit-learn library [78] . This library provides efficient implementations of the RMSE, MAE and SD. The next sections will provide a more thorough explanation on how the models were trained and optimized.

3.2.1 Training

The dataset was divided as presented in Figure 3.2. The training set was composed from the beginning of 2008 to the end of 2010. The validation set consists of 2011 and the test set of 2012 and 2013. Periods with faults will be removed from the training and validation datasets. It's also important to mention that although the WTs are all of the same model, in terms of the actual working of the WT there are always differences. For example, in Figure 3.3 is presented the distribution of the gearbox bearing temperature for different turbines over the training set, and although they have similar shapes, some are hotter and some are colder. This variation is expected, since it's not feasible for complex machines such as WTs to work exactly the same even if they are the same model. Due to these differences, in this work one NBM will be trained for each turbine.

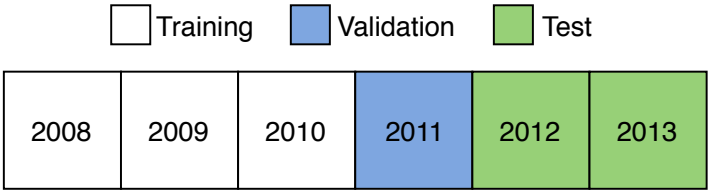


Figure 3.2: Dataset division for training, validating and testing.

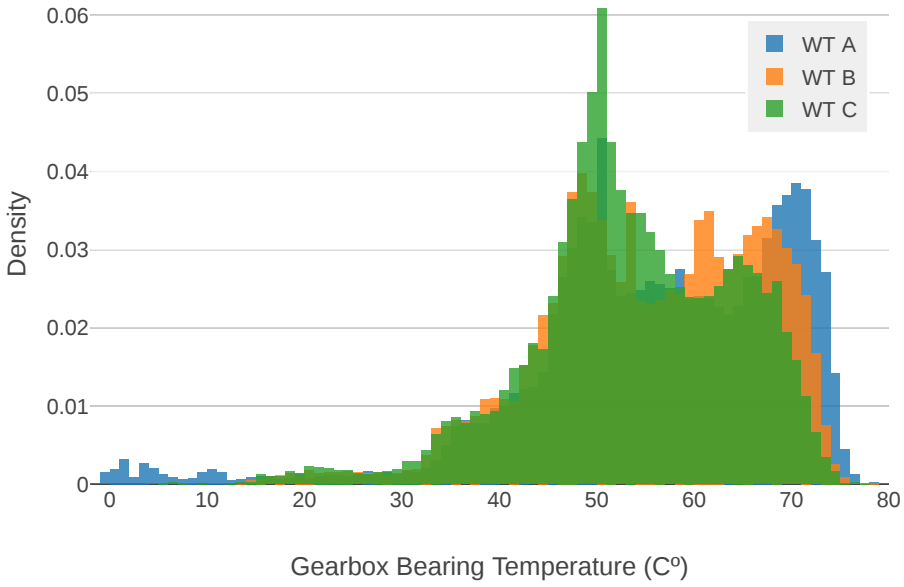


Figure 3.3: Gearbox bearing temperature distribution for different turbines on the training set.

3.2.2 Optimization

When comparing different models it's always important to guarantee that they are being compared fairly, in which hyperparameter optimization plays an important role due to its essential role in preventing overfitting. In fact, GBMs are significantly more robust to hyper parameter optimization than ANNs, which is why no exhaustive optimization will be performed. In this work, the number of estimators will be tuned by early stopping, and the rest of the parameters will be standard values that are advised in the literature. The early stopping works by defining a maximum number of estimators and the early stopping rounds, which correspond to how many rounds must pass without the validation loss decreasing for the training to stop. The values used in these work were 1000 and 100, respectively. An example of the training and validation losses for the CNBM is presented in Figure 3.4. In this case, the number of estimators was defined as 224, since by 324 estimators the validation loss had not significantly decreased. The number of estimators for the other models, as well as the other parameters that were defined based on best practice are presented on Table 3.3.

Table 3.3: Hyperparameters for each model with number of estimators obtained by early stopping.

Model	Number of Estimators	Learning Rate	Max Depth	Number of Leaves
CNBM	224	0.1	8	200
SNBM1	232			
SNBM2	149			
ACNBM	305			
ASNBM1	247			
ASNBM2	218			

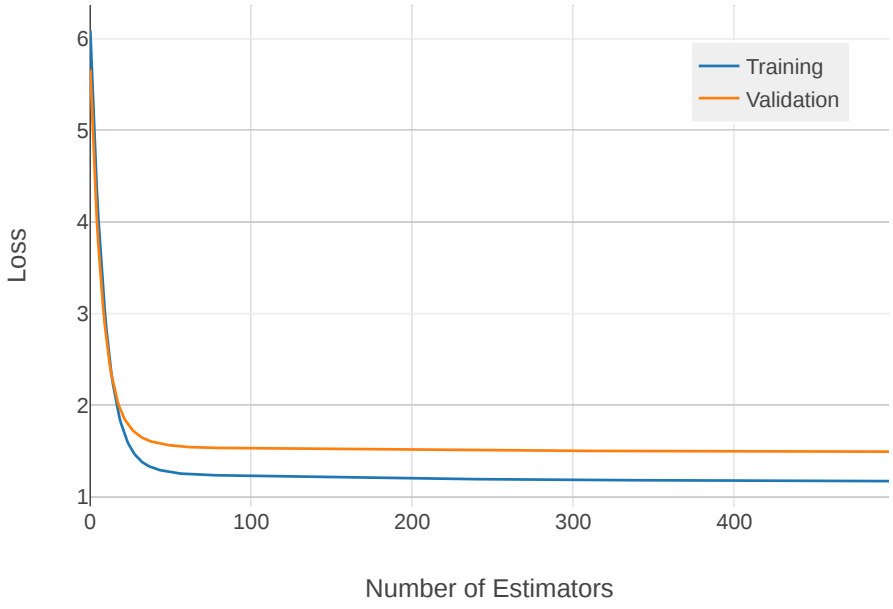


Figure 3.4: Training and validation losses as number of estimators increases for the CNBM.

3.3 Fault Detection

This section will explore the implementation of the fault detection framework and the libraries used for processing the residuals. As noted in the Methodology, the residuals must first be processed by using a rolling mean. This was done using the rolling function in Pandas, with a window of 6 hours, which will be used throughout the work. To give a better understanding of the impact of using the rolling mean, an example is presented in Figure 3.5. This figure shows the residuals for the ACNBM during a given failure, with and without the rolling mean function applied. As can be seen, the post processed residuals are more interpretable, having a clear increase before the failure and decrease after the maintenance is performed. On the other hand, the raw residuals are notably more noisy, not being trivial to identify a change in the residuals characteristic of a fault state. This will be particularly helpful when performing the visual residual analysis.

3.3.1 Fault Detection Framework

The fault detection framework proposed in this work requires the implementation of several functions. The function that uses the failure date to generate the classifier labels is presented in Algorithm 1. Note that $shift$ is a Pandas function that shifts the index by the desired number of periods. It was also necessary to develop the function that generates the alarms, based on the residuals. This function is presented in Algorithm 2. Note that $resample$ is a Pandas function that changes the frequency of the input data, in this case from a ten minute frequency to a daily frequency, and then applies the mean. The alarm vector will be binary, with ones when the threshold is above the residuals and zeroes when it's below. An example is given on Figure 3.6, where the resampled residuals of the ACNBM during a gearbox failure are presented. Two example thresholds are also presented on the Figure, to illustrate

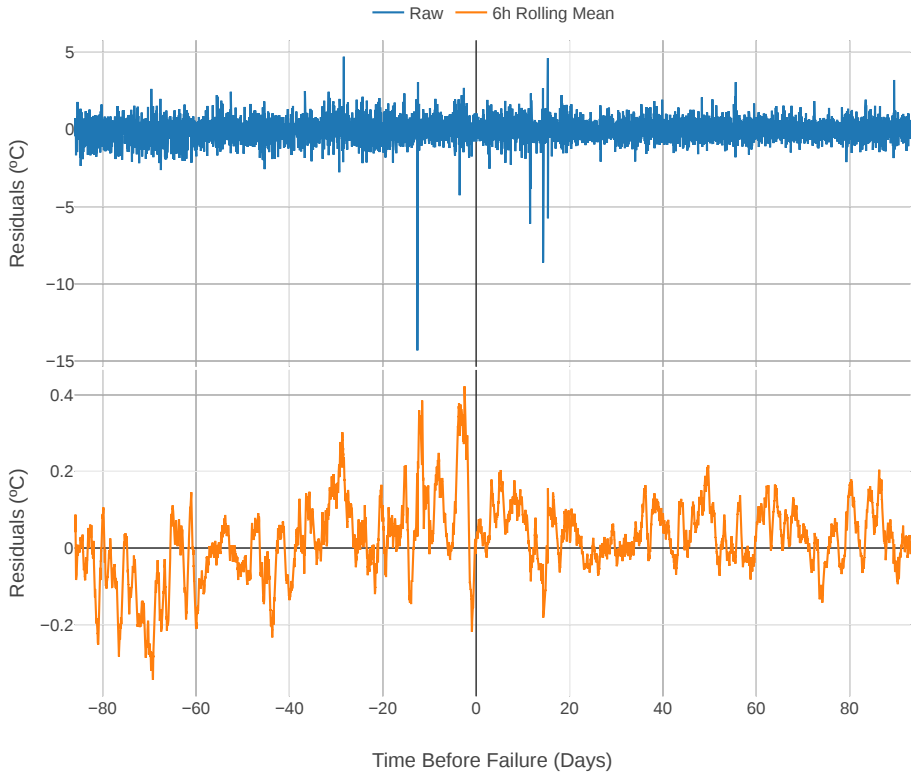


Figure 3.5: Gearbox bearing temperature residuals for the ACNBM for a given failure with and without post processing.

that all values above the threshold would be alarms.

Algorithm 1: label_generator function.

Data: prediction window Δt; failure f;
$l = f$;
for $i = \Delta t$ **to** 90 **do**
  $l = l + l.shift(i)$
end
return l

Algorithm 2: alarm_generation function.

Data: residuals r; threshold t;
$r = r.resample(24H).mean()$;
$a = r > t$;
return a

Indeed, the previous functions already produce alarms and labels that make it possible to evaluate the fault detection performance of different models. But as was motivated in the Methodology, we only want to consider the various alarms within the prediction window as one TP. This requires an extra function presented in Algorithm 3, which implements those subtleties. Note that this function already outputs the confusion matrix of the corresponding alarms and labels, by using the $confusion\_matrix$ function from Pandas. An example of the confusion matrices for the case of Figure 3.6, for thresholds A and B, is presented in Tables 3.4 and 3.5, respectively.

Finally, making use of the previous functions the evaluation framework was implemented as shown in

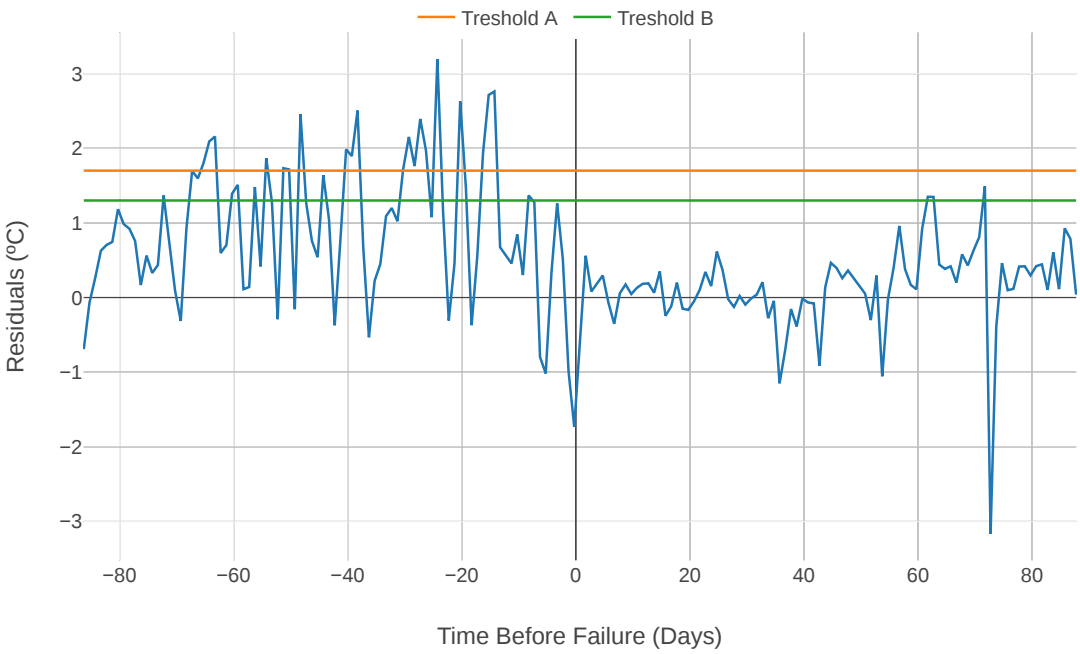


Figure 3.6: Gearbox bearing temperature residuals with 24H resample for the ACNBM.

Algorithm 3: cm_calculator function.

Data: alarms a; labels l; failure f
$x = AND(a, l)$;
$y = 0$;
if $x.sum() > 0$ **then**
  $y = f$;
end
$pred$ = a - x - f;
$pred[pred \neq 1] = 0$;
$pred = pred + y$;
cm = $confusion\_matrix(f, pred)$;
return cm

Table 3.4: Confusion matrix for threshold A.

	Predicted: No Fault	Predicted: Fault
Label: No Fault	TN = 227	FP = 0
Label: Fault	FN = 0	TP = 1

Table 3.5: Confusion matrix for threshold B.

	Predicted: No Fault	Predicted: Fault
Label: No Fault	TN = 224	FP = 3
Label: Fault	FN = 0	TP = 1

Algorithm 4. The confusion matrices are calculated for each combination of prediction window, threshold and turbine. Afterwards, the precision and recall scores are calculate using the $precision$ and $recall$

functions from scikit-learn. It's also relevant to mention the precision and recall values for the previous example. For threshold A we obtain $P = 1$ and $R = 1$. For threshold B we obtain $P = 0.25$ and $R = 1$. But it should be noted that they have different prediction windows, threshold A has a 65 days prediction window, while threshold B has a 73 days prediction window. This example illustrates the interest in using varying prediction windows, since they may result in different performances.

For the final results, with the total 16 turbines, the thresholds values ranged from -2 to 4, with incremental steps of 0.1. The prediction windows were 1, 10, 20, 30, 40, 50, 60, 70 and 80 days. Also, this process will be done for each of the six models. This results in a total of 51840 combinations, which can be considerably time-consuming to compute. Having this in mind, the evaluation framework was implemented making use of Dask [79], which provides parallel and distributed functionalities to Pandas. To take full advantage of Dask's distributed capabilities, the results were calculated using a cluster computing framework in the cloud, with a total of 30 workers with eight cores each, thus allowing 240 tasks to be performed in parallel and obtaining considerably better computational time performance.

Algorithm 4: evaluation_framework function.

Data: vector of prediction windows ΔT; vector of thresholds T; vector of turbines WT; vector of residuals R; vector of failures F

$scores = [\,]$;
for i ***in*** ΔT **do**
  for j ***in*** T **do**
    for k ***in*** WT **do**
      $r = R[k]$;
      $f = F[k]$;
      $l = label\_generator(i, f)$;
      $a = alarm\_generation(r, j)$;
      $cm = cm\_calculator(a, l, f)$;
      $pr = precision(cm)$;
      $rc = recall(cm)$;
      $scores.append([pr, rc])$;
    end
  end
end
return l

Chapter 4

Results and Discussion

4.1 Temperature Modelling

This section will begin by exploring some case studies that show the temperature modelling performance of the implemented models. The objective is to have a first understanding on how well the NBMs are able to model the gearbox bearing temperature during various operating regimes of the WTs. Afterwards, a more objective evaluation will be discussed, using the regression metrics previously described to analyse how the models developed in this work compare between themselves and how they compare against those in the literature.

4.1.1 Case Studies

In Figure 4.1 is presented a period of time where a healthy turbine worked under different regimes, such as high and low power production, as can be seen by the active power signal, and also during braking, as can be seen by the pitch angle signal. Regarding the predictions of the models, in this case study the CNBM and the ACNBM are being evaluated, and as can be seen both models are able to follow the true signal during the majority of time, even for the different turbine loads. But it's important to note that the predictions of the CNBM are significantly worse when the blades pitch to around 95º to stop the turbine. For example, during the morning of July 31st there are two periods of time where the WT stopped by using the pitch angle brake system, and the CNBM predicts that the temperature of the gearbox bearing would be lower than what it really is. This is problematic, since it would lead to false positives during the fault detection, because the real temperature is below the predicted one, not due to the existence of a fault but due to the model not learning the braking regime.The reason behind this may be due to the fact that the WT being stopped by the pitch brake system is not a common event, which means it is under represented in the dataset. This hypothesis can be supported by the histogram presented in Figure 4.2, where it's possible to see that the amount of data where the pitch angle is at around 95º is significantly under represented, comparing to when the turbine is operating normally, where the pitch angle is at around 0º. This means that the models wouldn't be able to learn this behavior as well as the most represented ones. On the other hand, the ACNBM is able to predict the temperature with much

less error in this regime. This may be due to the fact that the ACNBM has autoregressive features, which considerably simplify the prediction task, thus explaining the notably better results during this regime, independently of it being under-represented. Another possible explanation is that the braking of the turbine is exactly the time when it is harder to predict the temperature behavior from causal features, since it results in the stopping of the rotating components, making it the regime in which these causal features provide the least predictive power to the model. From this, results that in this regime the target temperature mostly depends on its previous values, hence why the autoregressive models performs better.

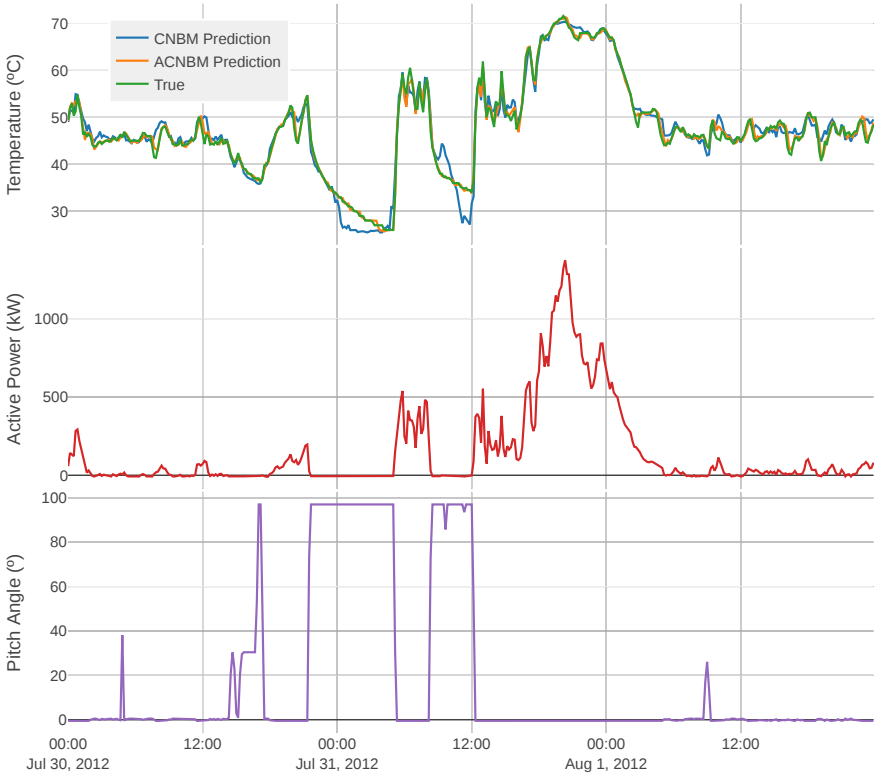


Figure 4.1: Gearbox bearing temperature predictions for the CNBM and ACNBM against the true values, during high and low loads and braking.

In Figure 4.3, a similar case study is presented, also comparing the CNBM and ACNBM. Again, both models follow the true signal for the majority of time, but during certain regimes the CNBM predicts with a significantly higher error. As it can be seen, during the night of January 31st the WT started braking and the CNBM predicts lower temperatures than the real ones, as in the previous case. But more interesting is in the night of February 1st, where the WT is not braking but it's also not producing, which means there is not enough wind speed for it to produce, and again the CNBM predicted the temperature with signifcant error. This means that besides braking, the CNBM also has a higher difficulty modelling certain shutting downs of the WT. Besides that, there's also another regime which the CNBM clearly hasn't learned as well: pitch action. As can be seen during the night of January 28th, the WT started increasing the pitch angle due to high wind speeds, and the CNBM predicts lower temperatures than the real ones. This same situation happens in February 1st when the WT starts increasing the pitch angle.

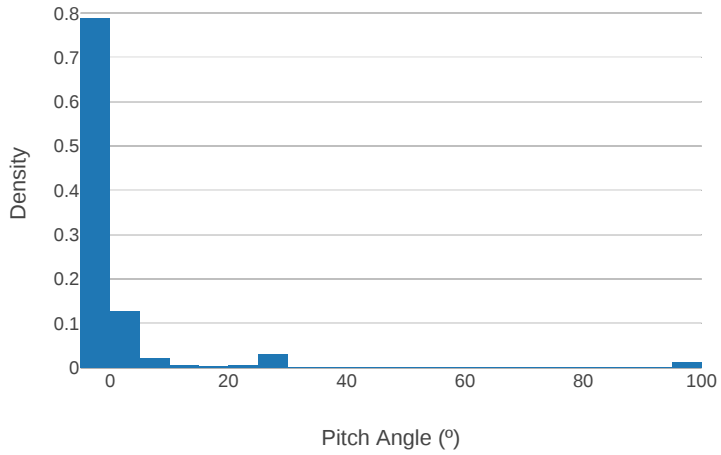


Figure 4.2: Histogram of the pitch angle for the complete training set.

The reason behind the CNBM not learning these regimes should be the same that was hypothesized in the previous case, indeed as can be seen in 4.2, the regime of when the pitch angle is being actuated is clearly under represented against the periods when the WT is working normally, thus leading to the model not learning these behaviors as well as the other ones. Again, this is highly problematic, since it may lead to false positives during fault detection.

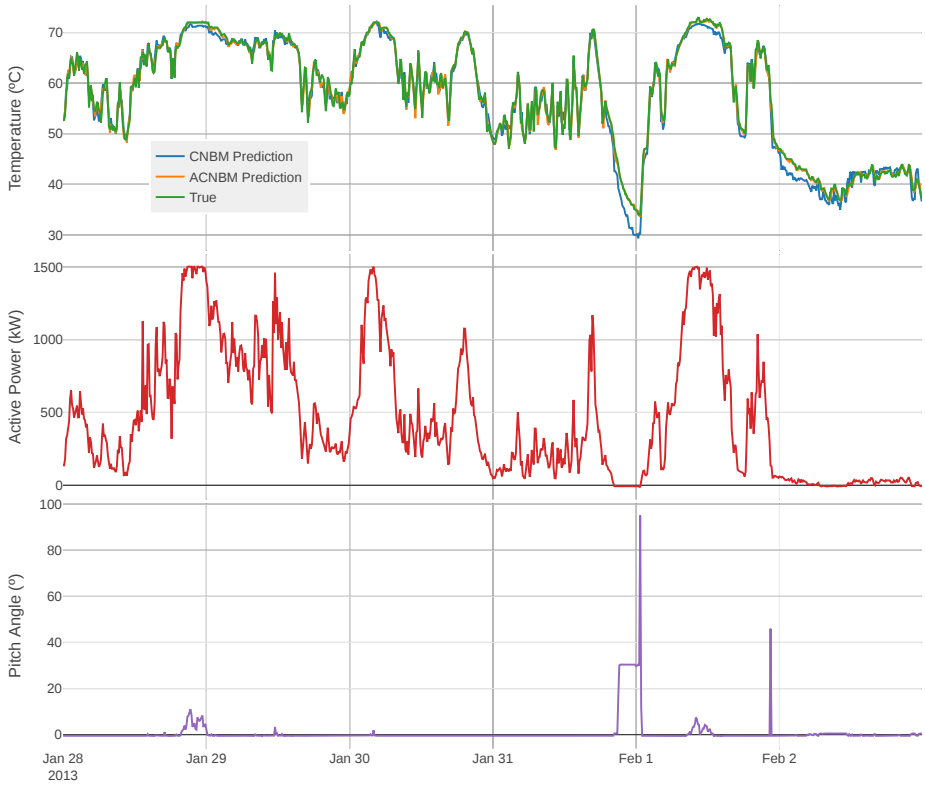


Figure 4.3: Gearbox bearing temperature predictions for the CNBM and ACNBM against the true values, during high and low loads, braking and pitch action.

Another case study is presented in Figure 4.4 which compares the SNBM1 and the SNBM2. It's

possible to see that both models are able to capture the general behavior of the WT and the temperature predictions follow the true signal during the majority of time. In this case it's interesting to note that both models obtain worse results when the turbine is braking, such as during the morning of July 21st, similarly to the CNBM in the previous case. In fact, none of these models have autoregressive features, hence why they have difficulty in following the true signal. Nonetheless, it should be noted that the SNBM2 obtains better results during these regimes, which indicates that the gearbox oil temperature contains relevant information for the model to capture the temperature decreasing behavior. It's also interesting to note that during July 20th there are two periods of time where, although the WT is not producing, the blade pitch angle is at around 30º, which is a rare situation, and it is exactly during these periods that the SNBM1 has a significantly higher error predicting the target temperature. What is interesting is that the SNBM2 is able to follow the true signal during these periods, meaning that again the gearbox oil temperature contains enough information regarding the gearbox bearing temperature for the model to significantly perform better. Indeed, simpler models like CNBM and SNBM1 perform similarly to the other models during the majority of time, but are significantly worse in specific regimes such as when the WT is braking, shutting down or during pitch action. Finally, it's also worth mentioning that although these are illustrative examples, they do generalize for the majority of the dataset, as can be seen in the extra case studies contained in Appendix A.

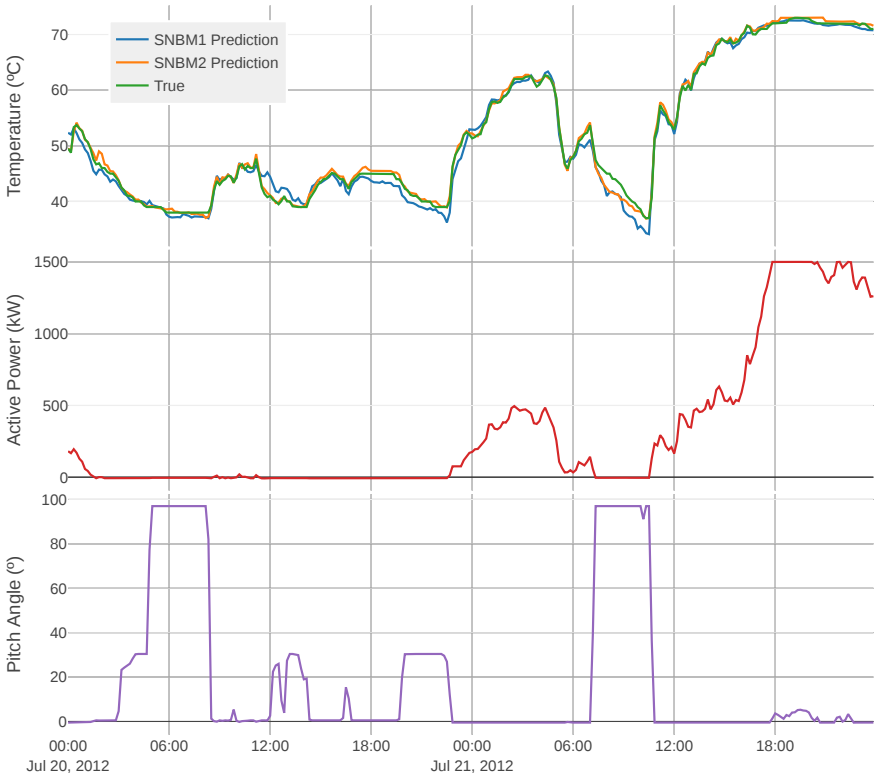


Figure 4.4: Gearbox bearing temperature predictions for the SNBM1 and SNBM2 against the true values for different operating regimes.

4.1.2 Evaluation

The previous case studies provided a general intuition on how the different models perform according to the different regimes of the turbine, but to objectively evaluate the temperature modelling performance it's important to use regression performance metrics. As is standard in the literature, these results were calculated on known healthy periods of the WTs. For example, the results corresponding to one of these healthy periods from the test set of one WT are presented in Figure 4.5 as a box plot of the MAE. It's interesting to note that the CNBM and the SNBM1 have similar performance, which means that the nacelle temperature doesn't seem to provide significant information to increase the predictive power of the model. On the other hand, the SNBM2 has a much lower median MAE and also less spread, which indicates that the gearbox oil temperature indeed increases the temperature modelling performance of the model.

It should also be noted that the autoregressive models have significantly better results than their non-autoregressive counterparts, this is due both to them capturing better the overall behavior of the target signal but also due to them performing better on certain regimes, such as braking and pitch action, as was previously observed. Also worth mentioning is that the SNBM2 has a slightly lower median MAE and significantly less spread than the ACNBM, which indicates that the gearbox oil temperature provides more information regarding the target temperature, then the lagged values of the target. Finally, the ASNBM2, which is the most complex model, is the one that obtains the best performance. This is to be expected for well optimized and regularized models, if a model has more features it is expected to perform either similarly or better. In this case, the extra features were relevant and improved the performance of the model.

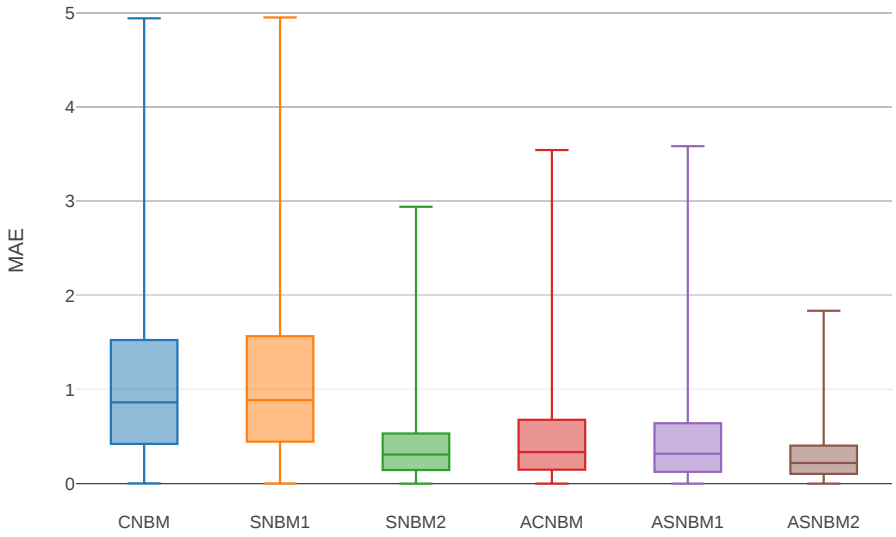


Figure 4.5: Box plot of the MAE for the different models.

The results for all the healthy periods of the total WTs are presented in Table 4.1 for the train and test sets, in terms of MAE, RMSE and SD. In general, these results follow the trend presented in the

previous illustrative example with the ASNBM2 obtaining the best performance on all metrics for both the train and test sets. These results also confirm the assumption from the previous example, that the gearbox oil temperature provides more predictive power to the model than the autoregressive feature, hence why the SNBM2 has a better performance than the ACNBM. These results also confirm the point previously raised, that autoregressive features and highly target correlated simultaneity features such as the gearbox cooling oil temperature increase the temperature modelling performance of the model.

It's also important to compare these results with the ones reported in the literature. To this end, in Table 4.2 are presented the results from the major works in the literature, with the corresponding input features used for the NBMs. It should also be noted that all the works presented used ANNs to build the NBM. Also, due to the inexistence of a standard dataset in this area, comparisons between works are always susceptible to some degree of subjectivity. But as it was previously motivated, the datasets are all from SCADA so the major part of the difference in the results should originate from the input features and the algorithm used. Having this in mind, it's interesting to note that [63] is the only model to only use causal features and the results obtained are significantly worse than the ones obtained by the CNBM in this work, which also only uses causal features. Regarding SNBMs, the results from [25] and [62] are also worse than the ones obtained in this work with the SNBM2. Finally, in terms of autoregressive models, the ASNBM2 obtained better results than [54] but worse results than [50]. In general, these results indicate that, with appropriate feature engineering, GBMs can obtain similar results to ANNs, thus being a potential alternative for significantly lower computational costs. Furthermore, tree-based methods such as GBMs have increased model interpretability, which is important for industry related applications, such as predictive maintenance, since there is skepticism regarding black-box models.

Table 4.1: Regression evaluation metrics for the different models on the train and test sets.

Model	Train Results (ºC			Test Results (ºC)		
	MAE	RMSE	SD	MAE	RMSE	SD
CNBM	1.05	1.63	1.25	1.36	1.75	1.11
SNBM1	1.03	1.61	1.23	1.31	1.74	1.15
SNBM2	0.39	0.56	0.40	0.52	0.73	0.51
ACNBM	0.55	0.85	0.65	0.63	0.94	0.69
ASNBM1	0.51	0.80	0.62	0.60	0.91	0.68
ASNBM2	**0.33**	**0.48**	**0.35**	**0.41**	**0.59**	**0.42**

Table 4.2: Features used in the NBMs and corresponding results for works in the literature.

Work	Input Features				Test Results (ºC)		
	Causal	Nacelle Temp.	Gearbox Oil Temp.	Autoregressive	MAE	RMSE	SD
[63]	X				2.15	2.93	2.88
[61]	X	-	-	-	0.663	-	-
[25]	X	X	X		-	1.22	-
[62]	X	X	X		-	-	**1.3**
[45]	X	X		X	-	1.23	-
[54]	X	X	X	X	**0.44**	0.77	-
[50]	X	X	X	X	-	**0.31**	-

4.2 Fault Detection

The main objective of building NBMs is to perform fault detection. Initially, some case studies will be presented in which a visual inspection of the residuals will be done to understand how each model performs during a state of fault. Afterwards, a more objective evaluation will be done based on the previously described evaluation framework to compare the overall fault detection results of all models.

4.2.1 Case Studies

The results presented in Figure 4.6 show the post-processed residuals of each model before and after a failure has occurred and the corresponding maintenance is performed. The models CNBM, SNBM1, ACNBM and ASNBM1 clearly show an increase in the residuals previously to the failure and a decrease after the corresponding maintenance. Indeed, these models detect a fault state previously to the failure. On the other hand, the models SNBM2 and ASNBM2, which contain the gearbox oil temperature feature, show no increase nor decrease during the fault state, which indicates that this fault resulted in an overheating of the gearbox bearing and in consequence the gearbox oil temperature also increased. This means that using the gearbox oil temperature as an input feature to the model resulted in a leak of information regarding the fault state of the component, thus making the model predict abnormal behavior and not detect the fault.

Another interesting result is that the use of the nacelle temperature as an input feature in SNBM1 and ASNBM1 doesn't seem to make a significant difference against their non-simultaneity counterparts CNBM and ASNBM. This means that, in terms of fault detection, the most interesting comparison is between the causal model CNBM and its autoregressive version ACNBM, which show an increase in the residuals at least 70 days before the failure. What is interesting is that the residuals seem to indicate that the CNBM could have actually predicted the failure even earlier, having spikes up to 160 days before the failure. This motivated a more careful inspection of the residuals which resulted that, in fact, these residual spikes are related with the regimes that the CNBM did not learn. This can be seen in Figure 4.8, where it's clear to see that the residual spikes around 120 days before the failure are due to the WT braking. Another example is the spike at around 160 days before the failure which is due to the WT increasing the pitch angle, as can be seen in Figure 4.9. This was expected, since the maximum predictive window should be 90 days, according to the established groundtruth. They also confirm the previous assumption that the regimes not learned by the CNBM lead to false positives. It's also relevant to look at the time series to understand how the models are detecting the fault state. The temperature predictions for the CNBM and ACNBM are presented in Figure 4.10 at around 70 days before the failure. As can be seen, the differences between the predictions and the real values happen during varying loads and they are not related with the regimes not learned by the CNBM. It's also interesting to look at Figure 4.11, which shows the same time period but for models SNBM1 and SNBM2, indeed the SNBM1 can detect the fault, while the SNBM2 completely follows the true temperatures, although they are higher than they should be, due to the leak of information provoked by the gearbox oil temperature,

Regarding the regimes not learned by the CNBM, it should be reminded that they happen mostly

when the turbine is shutting down. This makes them the least relevant for fault detection since it's when the rotating components are stopped, and thus the temperature variations are not necessarily related with efficiency changes, which are the basis of using NBMs for fault detection. Thus, a simple approach to improve the results, as was already proposed in [25], is to apply an active power filter. The majority of these regimes happened when the WT had active power below 200kW, so this was the value used, below which the residuals are filtered out. In Figure 4.7 are presented the results with the active power filter. It's interesting to note that the majority of models remained similar, but the residual spikes of the CNBM at 140 days and 120 days before the failure disappeared, indicating they were indeed false positives related with the turbine shutting down. On the other hand, the false positives at around 160 days before did not disappear, as expected, since these are related with the model not fully learning pitch action, which is not as straightforward to filter out, as it would also remove time periods that may be indicative of faulty behavior.

Another case study is presented in Figures 4.12 and 4.13, where the residuals without and with the active power are respectively shown. Again, the models SNBM2 and ASNBM2 don't show residual increase before the failure. On the other hand the ACNBM and the ASNBM1 show a clear increase in the residuals around 40 days before the failure. The CNBM and SNBM1 only seem to show an increase in the residuals around 30 days before the failure, but once the active power filter is applied it seems that the fault could be predicted around 60 days before, sooner than the autoregressive models, although with some false positives. To test this hypothesis it's important to understand what triggered these residual spikes by inspecting Figures 4.14 and 4.15. In fact, the signals seem to indicate that both of these spikes are related with the CNBM not completely learning pitch action. What is important to note is that the first spike was detected before the maximum predictive window, while the second was already within the window. This raises an important question, which is if the the residual spike 60 days before the failure is simply an error of the CNBM or if it is indicative of a developing fault. This example shows the inherent difficulty in evaluating fault detection by visual inspection, since it can be highly subjective. This example also illustrates the need for a more objective fault detection evaluation approach. In the evaluation framework developed in this work, the spike 60 days before would be considered a true positive since it is within the established predictive window, but it's important to note that all other residual spikes due to model errors present in healthy turbines would be taken into account as false positives. This also highlights the importance of not looking only at the periods of fault, analysing the false positives during states of no fault is also important to guarantee a robust fault detection system.

Summarizing, the results indicate that the nacelle temperature does not impact the performance of the models. On the other hand, the gearbox oil temperature completely eliminates the fault detection capabilities of the models. Furthermore, the autoregressive models are able to predict the failures with few false positives, while the causal model predicts failures but with some false positives, since it does not capture certain regimes. This model seems to improve when using an active power filter, since it removes some of the regimes not learned by the model while not filtering the regimes that may be indicative of faulty behavior. Although these are just illustrative examples, the overall results do generalize for the rest of the failures, which have their residuals presented in Appendix B.

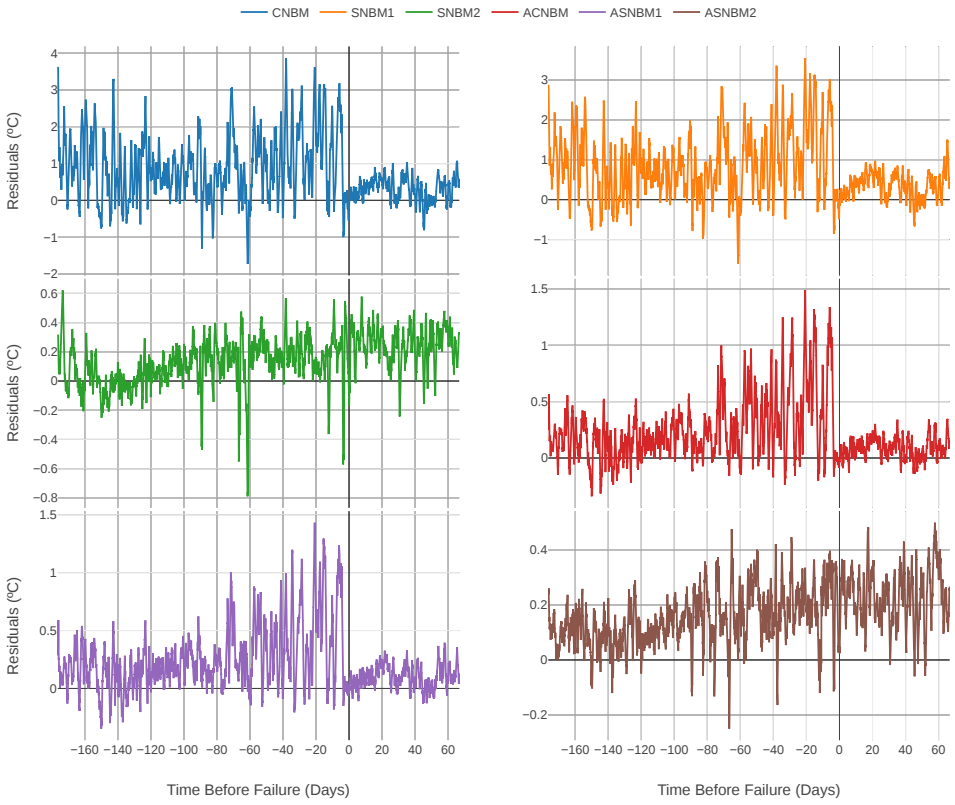


Figure 4.6: Post-processed residuals of different models for Failure A

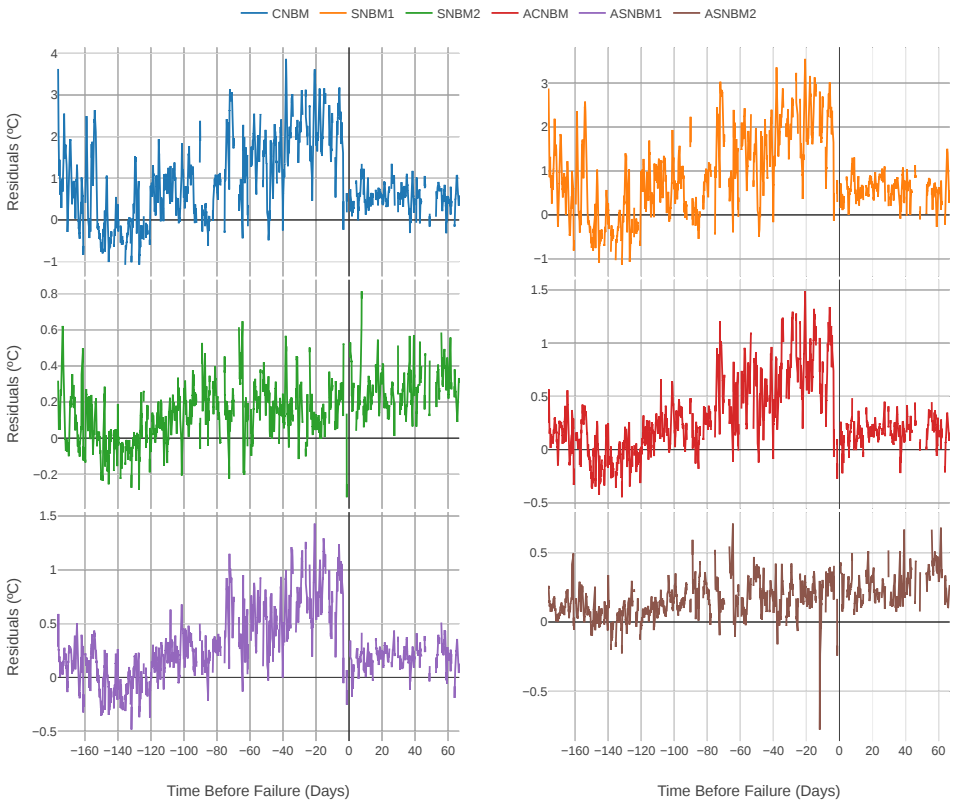


Figure 4.7: Post-processed residuals of different models with active power filter for Failure A

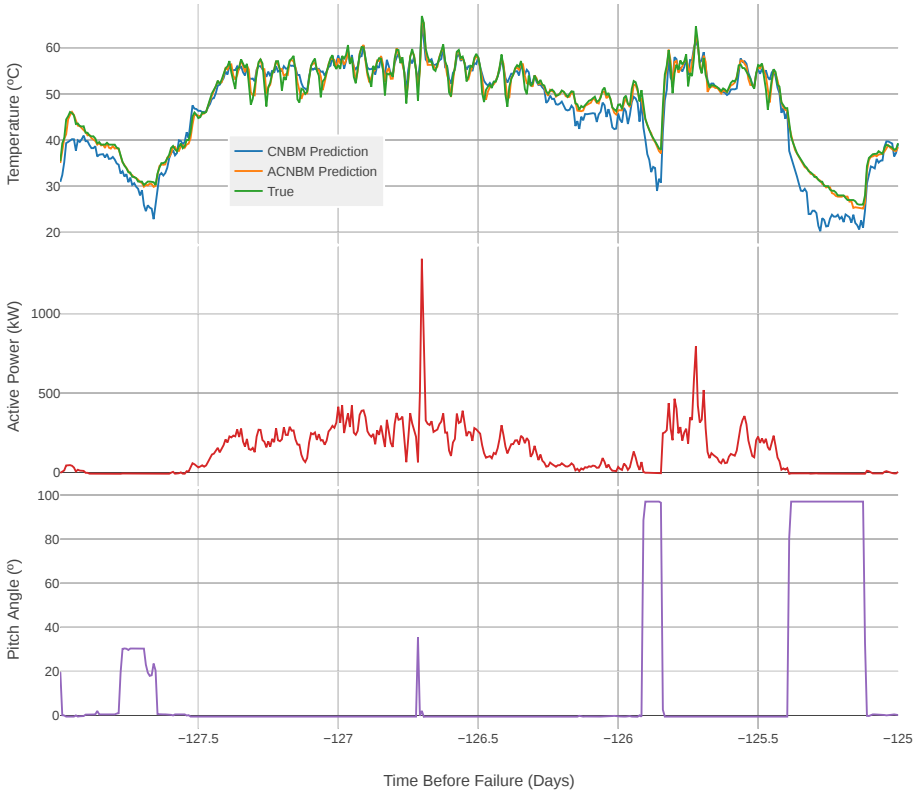


Figure 4.8: Temperature predictions for the CNBM and the ACNBM during the first FP

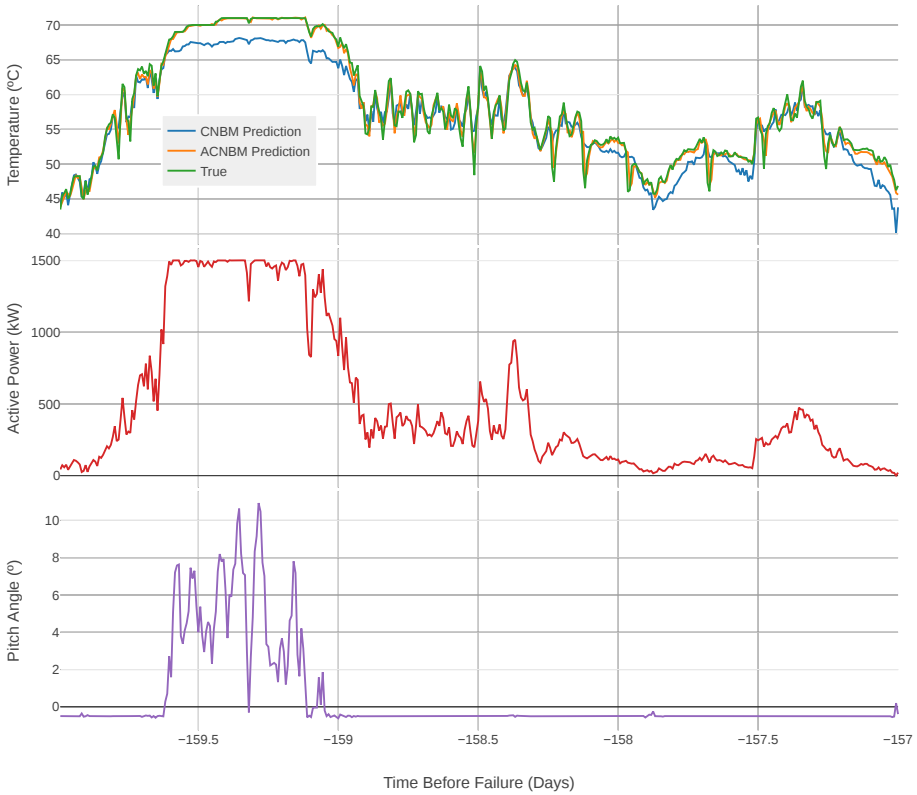


Figure 4.9: Temperature predictions for the CNBM and the ACNBM during the second FP

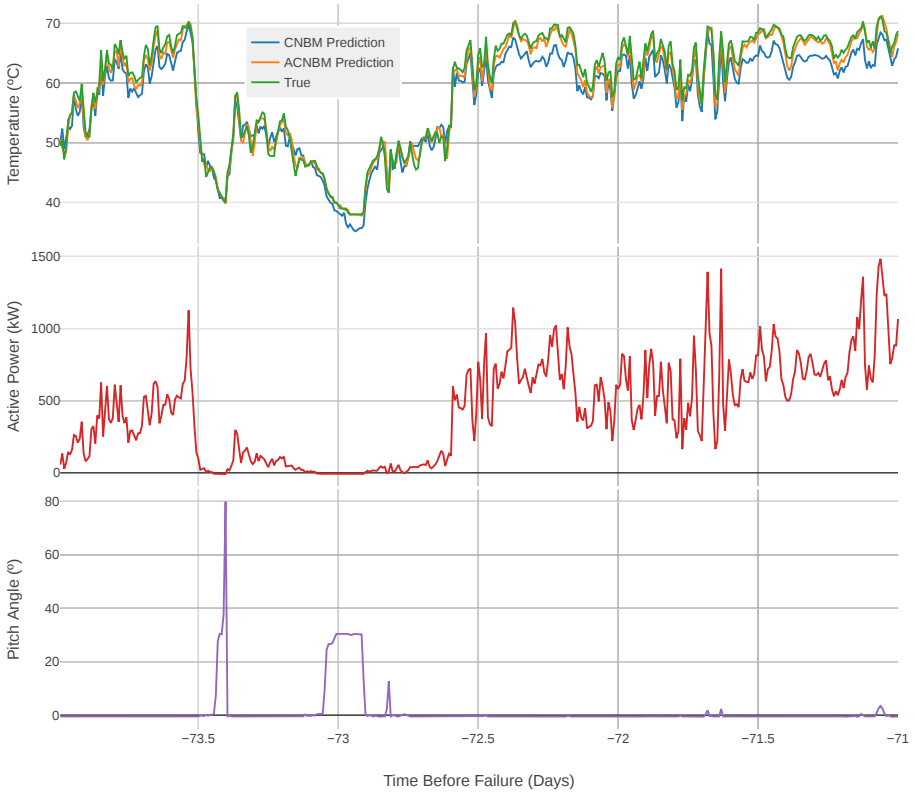


Figure 4.10: Temperature predictions for the CNBM and the ACNBM before Failure A

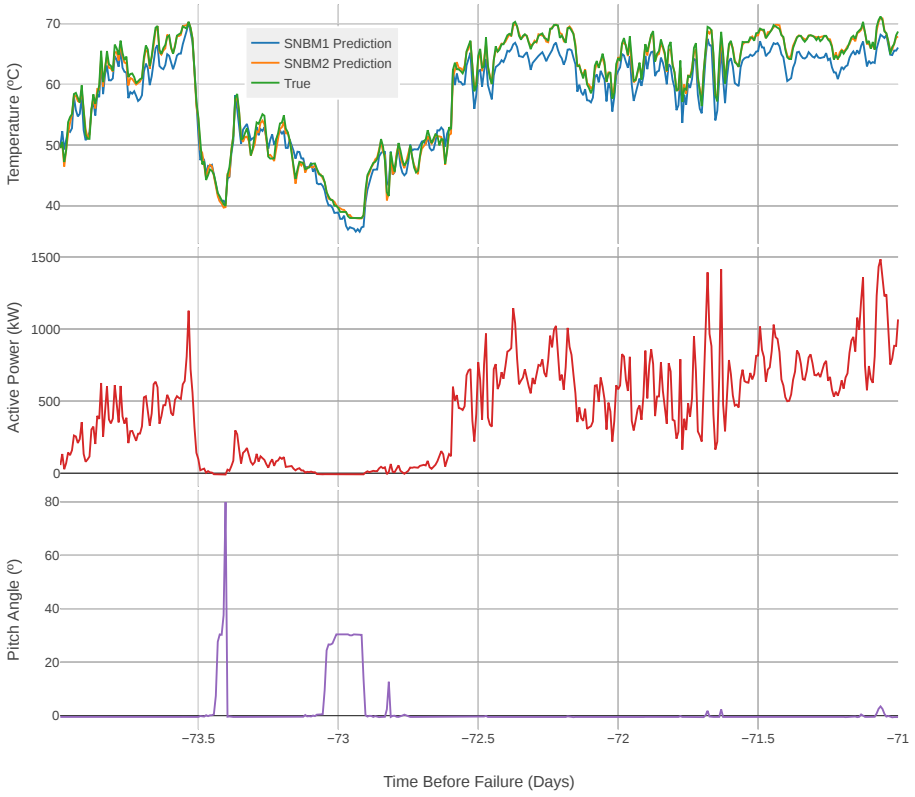


Figure 4.11: Temperature predictions for the SNBM1 and the SNBM2 before Failure A

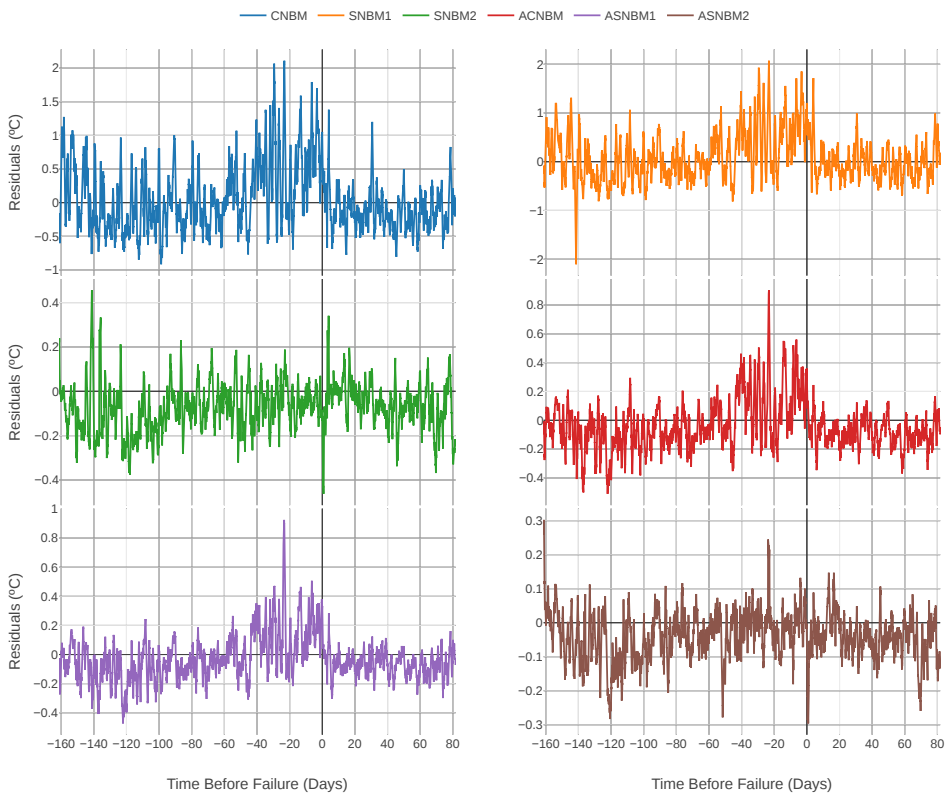


Figure 4.12: Post-processed residuals of different models for Failure B

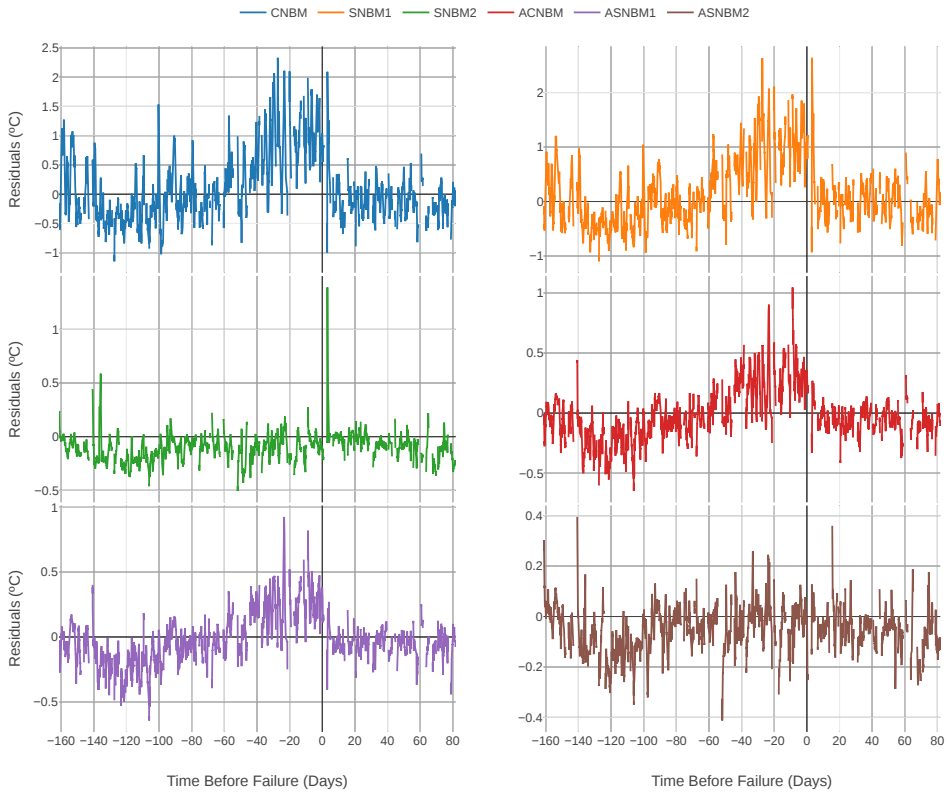


Figure 4.13: Post-processed residuals of different models with active power filter for Failure B

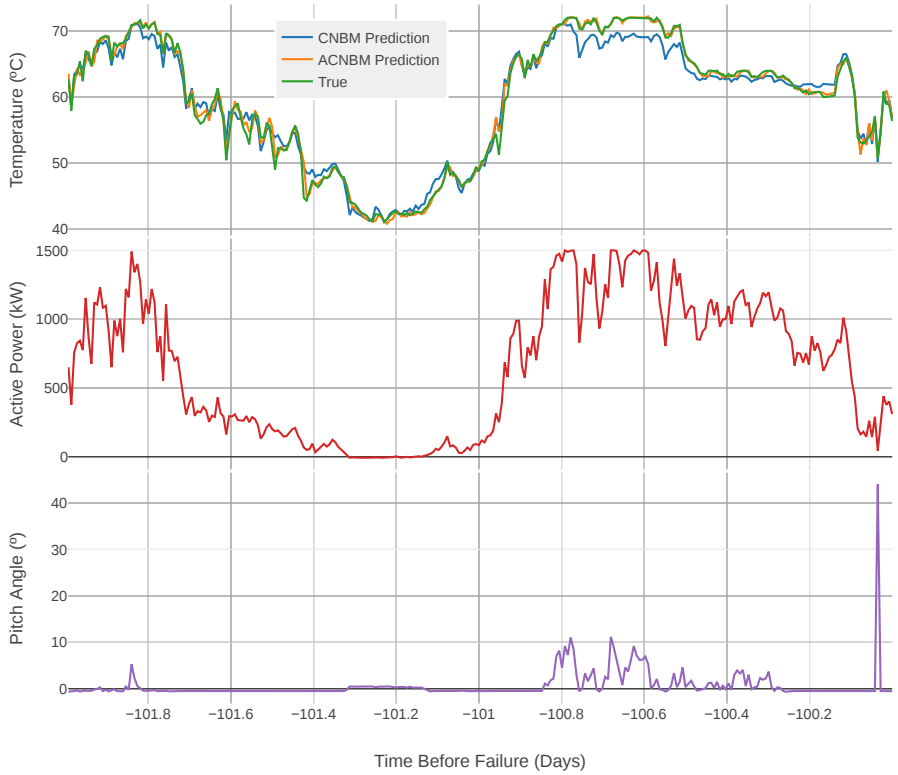


Figure 4.14: Temperature predictions for the CNBM and the ACNBM before Failure B

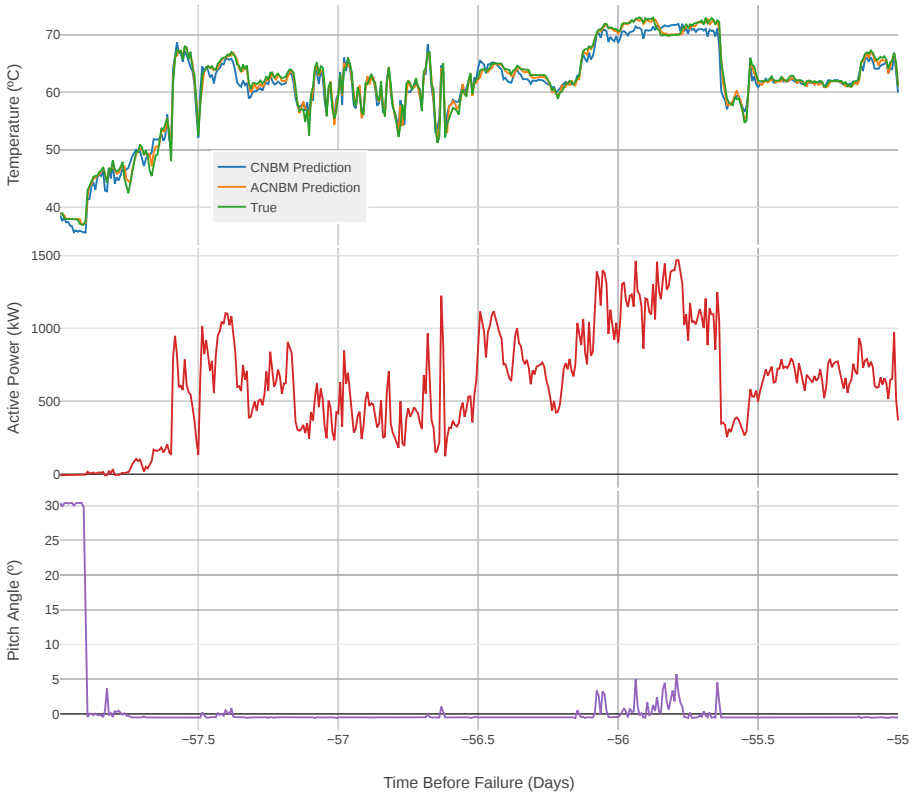


Figure 4.15: Temperature predictions for the CNBM and the ACNBM before Failure B

4.2.2 Evaluation

The majority of the literature work evaluates the results by visual inspection of the residuals, as performed in the previous section. But this is problematic, since it results in a subjective, manual and time consuming evaluation. Having this in mind, an evaluation framework was developed, based on the formulation of the detection of faults as a classification problem. The results will be presented in terms of precision and recall curves for fixed predictive windows and in terms of AUC for varying predictive windows.

Before proceeding to the results, it's important to explain the baseline to which the models developed in this work were compared. The detection of faults can be seen as an anomaly detection problem, where anomalies are the highest values in the target temperature distribution. Having this in mind, the baseline considered in this work is to set thresholds on the target temperature values and obtain the corresponding precision and recall. This means that the baseline reflects the performance of the detecting faults only looking at the target temperature, not taking into account all the input features that model the normal behavior of the WT as it is done in the NBMs.

4.2.2.1 Precision and Recall Curves

In first place, the fault detection performance of the different models was evaluated for a short prediction window, so Δt is equal to 10 days. The corresponding precision and recall curves are presented in Figure 4.16. The most straightforward result is that the ACNBM and the ASNBM1 obtain the best results, having higher precision for all recall scores. In fact, the ASNBM1 is able to obtain both recall and precision equal to 1, which means that the model was able to detect all faults with zero false positives at least 10 days in advance. It's also important to note that the models with the gearbox oil temperature obtain the worse results by a notable margin, which was expected since as was seen previously their residuals didn't show any failure-related predictive power. Moreover, the CNBM obtains results slightly better than the baseline but significantly worse than the ACNBM, which confirms the visual residual analysis, the causal model has more false positives thus lowering the precision. It's also worth mentioning that the CNBM and SNBM have similar results, again showing that the nacelle temperature doesn't seem to have a notable effect on the model both in terms of fault detection. Having in mind the previous assumption that the CNBM would benefit from using the active power filter, it was also applied for this evaluation and the results are presented in Figure 4.17. Indeed, the results for the CNBM and the SNBM clearly improve, being now very close to those of the autoregressive models, thus confirming the results from the visual residual analysis.

It's also important to evaluate the performance of the models for longer prediction windows, in this case Δt is equal to 50 days. The precision and recall curves are presented in 4.18 and, again, the ACNBM and the ASNBM1 obtain the best results. However, in this case the performance of the ASNBM1 is considerably worse than the ACNBM. This is an important result since it shows that the nacelle temperature not only does not improve the fault detection performance of the model, it actually seems to decrease it for longer prediction windows. This may be due to the fact that it is a simultaneity feature

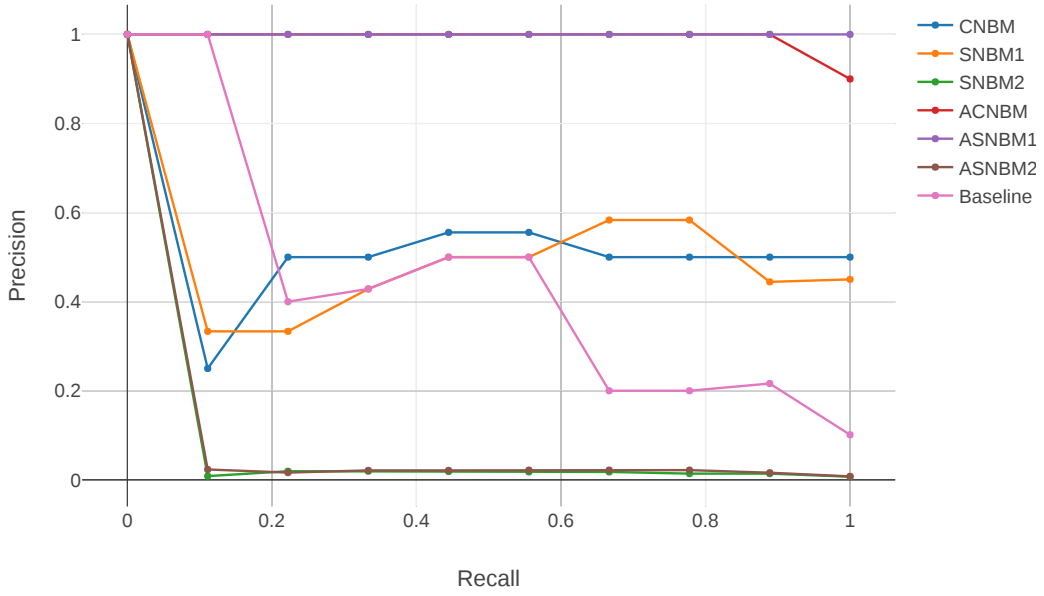


Figure 4.16: Precision and recall curves for the different models with $\Delta t = 10$ days.

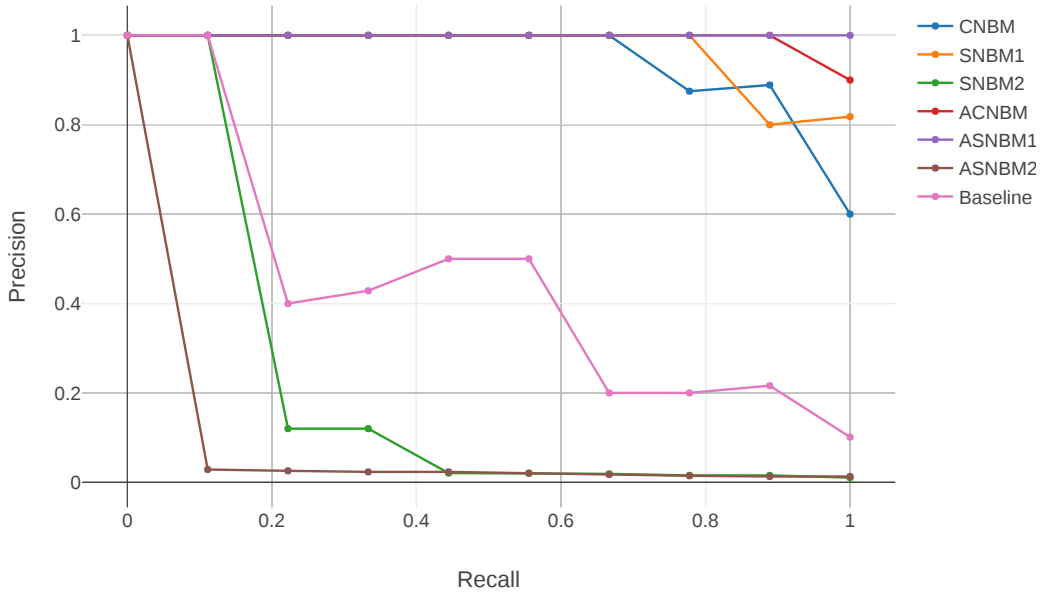


Figure 4.17: Precision and recall curves with active power filter for the different models with $\Delta t = 10$ days.

and, although it doesn't leak as much information regarding a gearbox fault state as the gearbox oil temperature, it does leak some information, which aggravates for longer predictive windows since it's when the temperature increases are more subtle. Also, this time the CNBM and SNBM1 performed considerably better than the baseline even without the filter, since the baseline is only based on the values of the target temperature it makes sense that it is less robust to the increase of the predictive window, on the contrary to the NBMs. Regarding the models with the gearbox oil temperature, they performed considerably worse than the baseline again. The precision and recall curves, already with the active power filter, are presented in 4.19. As in the previous case, the performance of the CNBM and the

SNBM1 clearly improved, but it's interesting to note that the ASNBM1 also improved while the ACNBM performed slightly worse. This is an important result, which shows that, sometimes, the active power filter may filter some fault-related behaviors thus reducing the fault detection capabilities of the models.

Finally, it's important to note that the precision and recall curves should be analysed from a high level perspective, since, as can be seen, they are not completely stable, existing considerable drops of precision. One of the main reasons behind this is the low amount of failures being analysed, it would be expected that with more failure data the results would be more consistent and the conclusions more robust. Nonetheless, various overall conclusions can still be derived: indeed, the results from the evaluation framework are aligned with those from the visual residual analysis, and also clarify some assumptions such as the negative impact of the FPs in the CNBM fault detection performance and also the performance increase due to the active power filter. It was also shown that although sometimes the filter can negatively impact the performance, in general it does improve the results. Again, these conclusions seem to apply for the rest of the predictive windows, whose precision and recall results are presented in Appendix C.

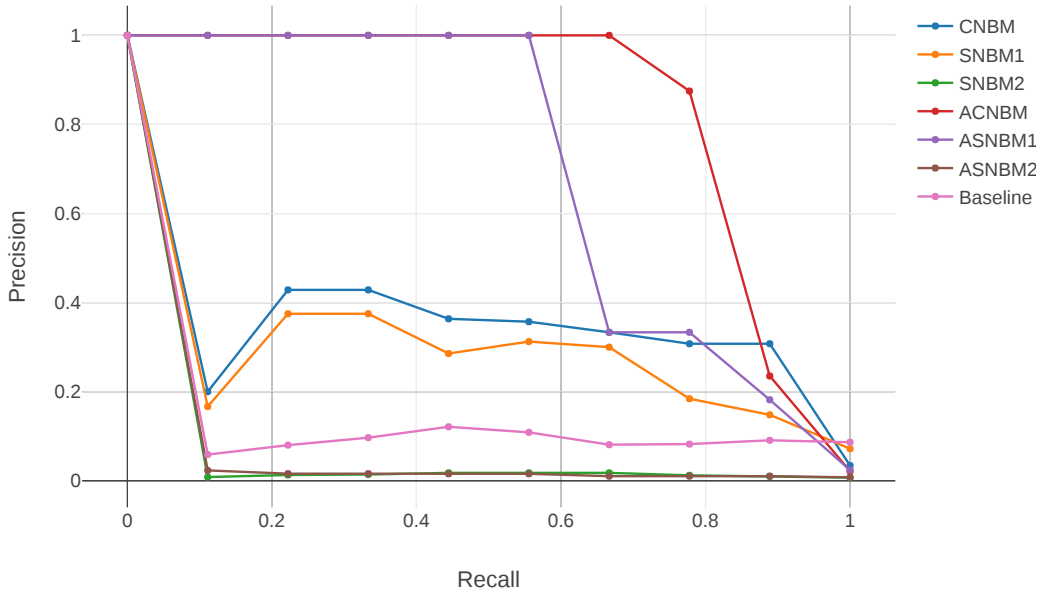


Figure 4.18: Precision and recall curves for the different models with $\Delta t = 50$ days.

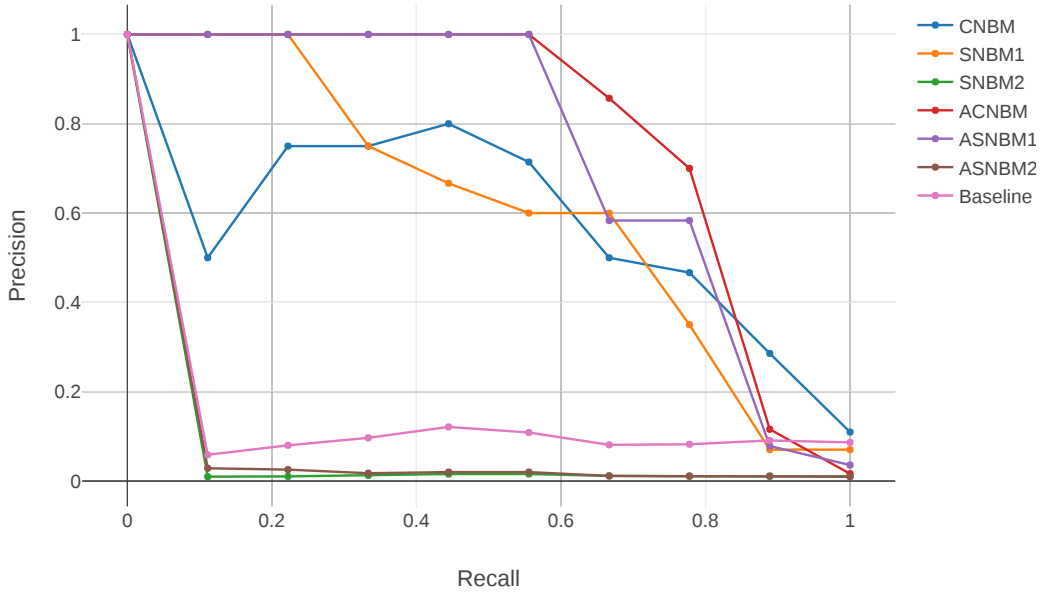


Figure 4.19: Precision and recall curves with active power filter for the different models with $\Delta t = 50$ days.

4.2.2.2 Area Under the Curve

The AUC summarizes the precision and recall curve into a value, which is important since sometimes it's not obvious which model is best because for each recall a different model has the best precision. In Figures 4.20 and 4.21 is presented the evolution of the AUC for the different models over the various predictive windows without and with the active power filter respectively. These results are also presented in Table 4.3. Regarding the results without active power filter, they seem to confirm the conclusions from the cases where $\Delta t = 10$ and $\Delta t = 50$, indeed the ACNBM is the model that obtains the best performance. Also, the CNBM is significantly worse than its autoregressive counterpart, but considerably better than the baseline. It's also interesting to note that the models with the nacelle temperature obtain worse results, confirming the indication that this feature can decrease the fault detection quality of the model. Finally, the models with the gearbox oil temperature clearly obtain the worse results. Regarding the results with the active power filter, they also confirm the conclusions previously made, it's clear that the performance of the CNBM and the SNBM1 are improved, while the autoregressive models are not significantly affected.

Table 4.3: AUC results without and with active power filter for the different models and predictive windows.

Model	AUC							AUC with Active Power Filter						
	10	20	30	40	50	60	70	10	20	30	40	50	60	70
CNBM	0.51	0.48	0.46	0.43	0.36	0.28	0.20	0.95	0.93	**0.92**	**0.86**	0.59	0.44	0.35
SNBM1	0.49	0.44	0.43	0.40	0.29	0.26	0.18	0.97	0.92	0.88	0.80	0.62	0.46	0.33
SNBM2	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.20	0.20	0.19	0.18	0.07	0.07	0.07
ACNBM	**1.00**	**1.00**	**0.93**	**0.88**	**0.85**	**0.56**	**0.45**	0.99	**0.99**	0.91	0.83	**0.80**	0.54	**0.43**
ASNBM1	**1.00**	**1.00**	0.89	0.84	0.71	0.51	0.41	**1.00**	0.98	0.86	0.77	0.75	**0.58**	**0.43**
ASNBM2	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.07	0.07	0.07	0.07	0.07	0.07	0.07
Baseline	0.44	0.39	0.35	0.17	0.14	0.14	0.12	0.44	0.39	0.35	0.17	0.14	0.14	0.12

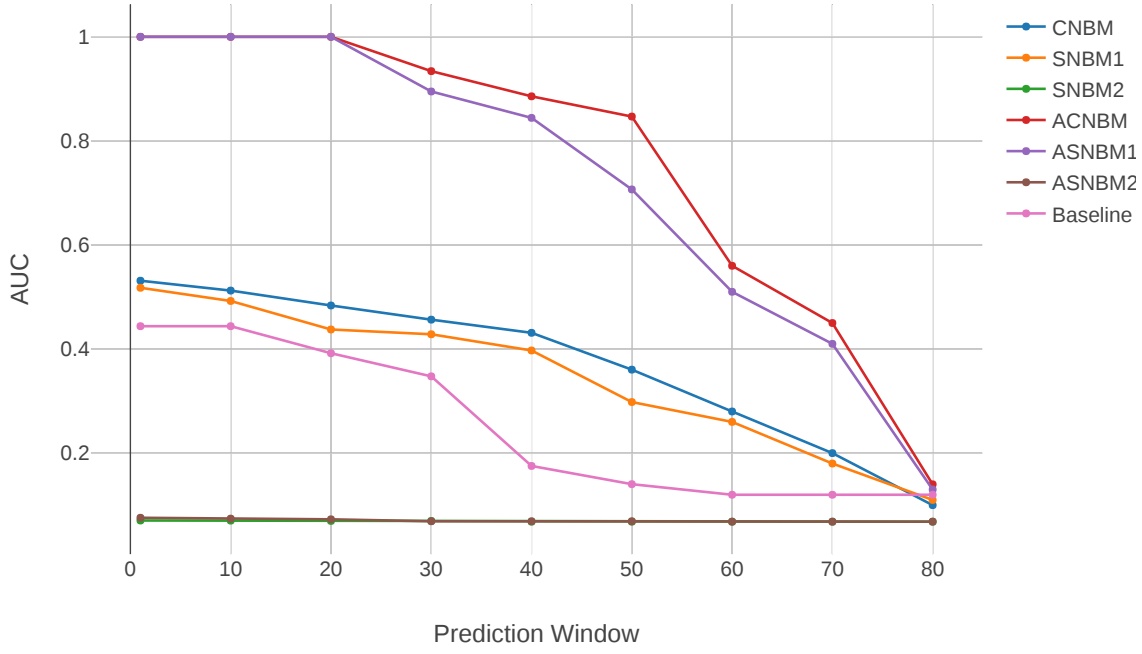


Figure 4.20: Evolution of the AUC over the prediction windows for the different models.

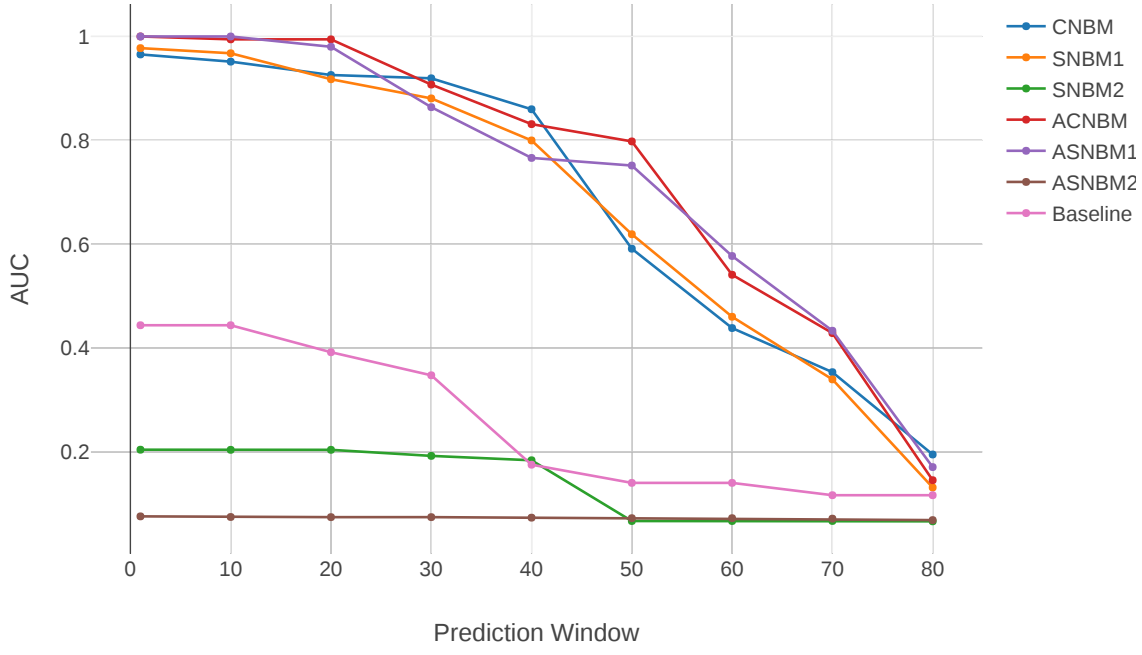


Figure 4.21: Evolution of the AUC over the prediction windows with active power filter for the different models.

Chapter 5

Conclusions

The majority of the literature uses ANNs for building NBMs. In this work, it was shown that GBMs obtain competive temperature modelling results. This is relevant because GBMs are known to have lower computational costs and also higher model interpretability. The latter is particularly important in this type of industry application, where there is skepticism regarding black-box solutions. It was also developed a taxonomy to categorize input features into different types based on their causal relation with the target temperature. This allowed to evaluate how different input feature affect the performance of the model. In terms of temperature modelling performance these are several main conclusions:

- Causal models are able to model the gearbox temperature during the majority of time. But the model has significant error during certain regimes, due to both the under representation of this regimes in the dataset and also to the fact that these regimes are when the target temperature least depends on the causal features from a domain knowledge perspective;
- The addition of autoregressive features notably improves the performance of the model. Indeed, with the autoregressive features the model is able to model the gearbox bearing temperature during those regimes that the causal model couldn't capture as well;
- The addition of the nacelle temperature feature doesn't have a significant improvement in the model. On the other hand, the gearbox oil temperature notably increases the model performance, even more than the autoregressive features.

The detection of faults was formulated as a classification problem and an evaluation framework was developed using classification metrics for unbalanced datasets. The results from this framework confirmed those obtained by visual inspection, indicating that this is a good alternative which is more objective and involves less manual work. The main conclusions were the following:

- Causal models are able to predict all the failures with better precision than the baseline with predictive windows of up to 70 days;
- The addition of autoregressive features notably increases the fault detection performance, obtaining better results than the causal model and the baseline for all predictive windows. Indeed, the

model with the autoregressive features was able to predict all failures without false positives up to 20 days before the failure;

- The use of the gearbox oil temperature completely eliminated the fault detection capabilities of the model. The nacelle temperature also seemed to decrease the fault detection performance of the model. This means that although simultaneity features can improve the temperature modelling performance of the model, they decrease the fault detection performance.

It should also be mentioned that the results regarding simultaneity features are aligned with the works in the literature that showed that highly correlated features, such as gearbox oil temperature, resulted in lower fault detection performance. But this work also showed that the origin of the problem is not in the fact that the features are highly correlated, but due to the simultaneity nature of their causal relation. Furthermore, for the type of failures evaluated in this work the use of autoregressive features did not reduce the fault detection performance of the model. In fact, it helped the fault detection performance by reducing the number of false positives from the causal model, by modelling the regimes that the previous had more difficulty capturing. In the literature, there were conflicting results regarding the use of autoregressive features, but the results from this work do align most with the ones that reported better results using them. It's also important to note that the cause of these conflicting results may be due to the type of failures analysed in different works, which have different datasets and thus different failures. In fact, the impact of autoregressive features on fault detection performance may be highly dependent on the type of failures that one is trying to detect. Furthermore, this work also showed that besides using autoregressive features to improve the fault detection performance of causal models, one can also use post-processing techniques on the residuals. The results showed that the active power filter considerably increased the fault detection performance of the model, due to filtering out the regimes that this model didn't learn as well.

Finally, it's important to remind that the method of evaluation of some of these works was based on visual residual analysis, which as was shown in this work has several disadvantages, such as not taking into account the existence of false positives. This again highlights the importance of the evaluation framework that was developed in this work based on the formulation of the detection of faults as classification problem, which was aligned with the results from the visual residual analysis but has the advantages of being more objective and less time-consuming.

5.1 Future Work

Several conclusions were derived from the results of this work. Having this in mind, they also leave opportunities for further work, and also various points of improvement:

- The results from this work indicated that GBMs are a potential alternative to ANNs. This assessment can be further tested by performing a direct comparison between the two algorithms on the same dataset;

- Several conclusions were derived in terms of the impact of feature causality on the GBMs. It would be relevant to assess if these conclusions generalize to the use of ANNs;
- As shown in the results, there are certain regimes that the causal model had difficulty learning, possibly due to their under representation in the dataset. Techniques to handle unbalanced datasets, such as under-sampling, may be investigated to improve this;
- Another way to improve the results from the causal model is to investigate other post processing techniques in the residuals. In this work, an active power filter was applied. But filters based on the pitch angle may lead to better filtering out the regimes not learned by the causal model;
- This work only had data of nine failures. With more failures it would be possible to separate them into a training and test dataset, where the former would serve the purpose of optimizing the residual threshold. With the set threshold, the test set results would give a better approximation of real time performance.

Bibliography

[1] I. E. Agency. *Global Energy & CO2 Status Report*. International Energy Agency, 2018.

[2] BP. Statistical review of world energy. Online: `https://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy.html` (16/04/2019).

[3] K. B. Tokarska, N. P. Gillett, A. J. Weaver, V. K. Arora, and M. Eby. The climate response to five trillion tonnes of carbon. *Nature Climate Change*, 6, 2016.

[4] D. J. Wuebbles, D. W. Fahey, K. A. Hibbard, D. J. Dok-ken, B. C. Stewart, and T. K. Maycock. *Climate science special report: fourth national climate assessment,. Volume 1*. U.S. Global Change Research Program, Washington, D.C., USA, volume 1 edition, 2017.

[5] P. Pham. Why is tension rising in the south china sea? Online: `https://www.forbes.com/sites/peterpham/2017/12/19/why-is-tension-rising-in-the-south-china-sea` (16/04/2019).

[6] O. Francis. Peace at last for south sudan? Online: `https://www.bloomberg.com/news/articles/2018-10-16/peace-at-last-for-south-sudan-that-may-depend-on-the-oil-price` (16/04/2017).

[7] IRENA. *REthinking Energy*. International Renewable Energy Agency, 2017.

[8] K. B. Kalimeris. *Revisiting the Energy-Development Link*, chapter A Brief History of Energy Use in Human Societies. Springer, 1^{st} edition, 2006.

[9] V. Smil. *Energy Transitions: Global and National Perspectives*. ABC-CLIO, 2016.

[10] I. E. Agency. *A New World: The Geopolitics of the Energy Transformation*. International Energy Agency, 2019.

[11] I. E. Agency. *Energy Access Outlook*. International Energy Agency, 2017.

[12] Schlömer S., T. Bruckner, L. Fulton, E. Hertwich, A. McKinnon, D. Perczyk, J. Roy, R. Schaeffer, R. Sims, P. Smith, and R. Wiser, 2014: Annex III: Technology-specific cost and performance parameters. In: Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Edenhofer, O., R. Pichs-Madruga, Y. Sokona, E. Farahani, S. Kadner, K. Seyboth, A. Adler, I. Baum, S. Brunner, P. Eickemeier, B. Kriemann, J. Savolainen, S. Schlömer, C. von Stechow, T. Zwickel and J.C. Minx (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

[13] European Comission. *The Revised Renewable Energy Directive*, 2018.

[14] Governo de Portugal. *Green Growth Commitment*, 2014.

[15] D. GL. *Energy Transition Outlook*. DNV GL, 2018.

[16] IRENA. Renewable energy cost database. Online: `http://resourceirena.irena.org/gateway/dashboard/?topic=3&subTopic=1065` (16/04/2019).

[17] R. Wiser, E. Lantz, T. Mai, J. Zayas, E. DeMeo, E. Eugeni, J. Lin-Powers, and R. Tusing. Wind vision: A new era for wind power in the united states. *The Electricity Journal*, 2015.

[18] R. Castro. *Uma introdução às energias renováveis: eólica, fotovoltaica e mini-hídrica*. IST PRESS, 2011.

[19] T. J. Stehly, P. C. Beiter, D. M. Heimiller, and G. N. Scott. 2017 cost of wind energy review. 2018.

[20] K. Leahy. *Data analytics for fault prediction and diagnosis in wind turbines*. PhD thesis, University College Cork, 2018.

[21] Lazard. *Levelized Cost of Energy Analysis 12.0*, 2018.

[22] IRENA. *The Power to Change: Solar and Wind Cost Reduction Potential to 2025*, 2016.

[23] G. A. M. van Kuik, J. Peinke, R. Nijssen, D. J. Lekou, J. Mann, J. N. Sørensen, and K. Skytte. Longterm research challenges in wind energy – a research agenda by the european academy of wind energy. *Wind Energy Science*, 2016.

[24] I. O. for Standardization. *ISO 13372:2012 Condition monitoring and diagnostics of machines – Vocabulary*, 2012.

[25] M. Bach-Andersen. *A Diagnostic and Predictive Framework for Wind Turbine Drive Train Monitoring*. PhD thesis, Technical University of Denmark, 2017.

[26] S. Boersma, B. M. Doekemeijer, P. M. O. Gebraad, P. A. Fleming, J. Annoni, A. K. Scholbrock, J. A. Frederik, and J. van Wingerden. A tutorial on control-oriented modeling and control of wind farms. In *2017 American Control Conference (ACC)*, 2017.

[27] J. Carroll, A. McDonald, and D. McMillan. Failure rate, repair time and unscheduled O&M cost analysis of offshore wind turbines. *Wind Energy*, 2016.

[28] E. Gonzalez and J. J. Melero. Wind turbine component fault detection by monitoring its performance using high-resolution SCADA data. In *30th Conference on Condition Monitoring and Diagnostic Engineering Management*, 2017.

[29] J. Tautz-Weinert and S. Watson. Using SCADA data for wind turbine condition monitoring - a review. *IET Renewable Power Generation*, 2017.

[30] W. Yang, P. Tavner, C. Crabtree, Y. Feng, and Y. Qiu. Wind turbine condition monitoring: technical and commercial challenges. *Wind Energy*, 2012.

[31] I. E. Comission. IEC 61400-12-1 wind turbines – part 12-1: Power performance of electricity-producing wind turbines. 2011.

[32] A. Kusiak, H. Zheng, and Z. Song. On-line monitoring of power curves. *Renewable Energy*, 2009.

[33] R. Bi, C. Zhou, and D. M. Hepburn. Applying instantaneous SCADA data to artificial intelligence based power curve monitoring and WTG fault forecasting. In *2016 International Conference on Smart Grid and Clean Energy Technologies (ICSGCE)*, 2016.

[34] R. Pandit and D. Infield. Gaussian process operational curves for wind turbine condition monitoring. *Energies*, 2018.

[35] E. Gonzalez and J. J. Melero. Statistical evaluation of SCADA data for wind turbine condition monitoring and farm assessment. *Journal of Physics: Conference Series*, 2018.

[36] K. Leahy, R. L. Hu, I. C. Konstantakopoulos, C. J. Spanos, and A. M. Agogino. Diagnosing wind turbine faults using machine learning techniques applied to operational data. In *2016 IEEE International Conference on Prognostics and Health Management (ICPHM)*, 2016.

[37] K. Leahy, R. Hu, I. C. Konstantakopoulos, C. J. Spanos, A. M. Agogino, and D. T. J. O'Sullivan. Diagnosing and predicting wind turbine faults from SCADA data using support vector machines. *International Journal of Prognostics and Health Management*, 2018.

[38] J. Hochenbaum, O. Vallis, and A. Kejariwal. Automatic anomaly detection in the cloud via statistical learning. *CoRR*, 2017.

[39] C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. A. Dau, D. F. Silva, A. Mueen, and E. Keogh. Matrix profile i: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 2016.

[40] H. Ling and K. Okada. An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE transactions on pattern analysis and machine intelligence*, 2007.

[41] O. Pele and M. Werman. The quadratic-chi histogram distance family. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision – ECCV 2010*, 2010.

[42] Y. Feng, Y. Qiu, C. Crabtree, H. Long, and P. Tavner. Use of SCADA and CMS signals for failure detection diagnosis of a wind turbine gearbox. 2011.

[43] Y. Feng, Y. Qiu, C. Crabtree, H. Long, and P. Tavner. Monitoring wind turbine gearboxes. *Wind Energy*, 2013.

[44] M. Cruz Garcia, M. Sanz-Bobi, and J. del Pico. SIMAP: Intelligent system for predictive maintenance: Application to the health condition monitoring of a windturbine gearbox. *Computers in Industry*, 2006.

[45] A. Zaher, S. McArthur, D. Infield, and Y. Patel. Online wind turbine fault detection through automated SCADA data analysis. *Wind Energy*, 2009.

[46] R. Mesquita, J. Carvalho, and F. Pires. Neural networks for condition monitoring of wind turbines gearbox. *J. Energy Power Eng.*, 2012.

[47] R. F. Mesquita Brandão, J. A. Beleza Carvalho, and F. P. Maciel Barbosa. Intelligent system for fault detection in wind turbines gearbox. In *2015 IEEE Eindhoven PowerTech*, 2015.

[48] Z.-Y. Zhang and K.-S. Wang. Wind turbine fault detection based on SCADA data analysis using ANN. *Advances in Manufacturing*, 2014.

[49] M. Schlechtingen and I. Santos. Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection. *Mechanical Systems and Signal Processing*, 2011.

[50] M. Bach-Andersen, B. Rømer-Odgaard, and O. Winther. Flexible non-linear predictive models for large-scale wind turbine diagnostics. *Wind Energy*, 2016.

[51] J. Tautz-Weinert and S. Watson. Condition monitoring of wind turbine drive trains by normal behaviour modelling of temperatures. In *Conference for Wind Power Drives (CWD 2017), Aachen, Germany, 7th-8th March 2017*, 2017.

[52] D. Karlsson. *Wind Turbine Performance Monitoring using Artificial Neural Networks With a Multi-Dimensional Data Filtering Approach*. PhD thesis, Chalmers University of Technology, 2014.

[53] Y. Cui, P. Bangalore, and L. Bertling Tjernberg. An anomaly detection approach based on machine learning and SCADA data for condition monitoring of wind turbines. In *2018 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, 2018.

[54] P. Bangalore, S. Letzgus, D. Karlsson, and M. Patriksson. An artificial neural network based condition monitoring method for wind turbines, with application to the monitoring of the gearbox. *Wind Energy*, 2017.

[55] J. Tautz-Weinert. *Improved wind turbine monitoring using operational data*. PhD thesis, Loughborough University, 2018.

[56] C. Krauss, X. Anh Do, and N. Huck. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the sp 500. *European Journal of Operational Research*, 2016.

[57] S. Ben Taieb and R. Hyndman. A gradient boosting approach to the kaggle load forecasting competition. *International Journal of Forecasting*, 2013.

[58] J. Robert Lloyd. GEFCom2012 hierarchical load forecasting: Gradient boosting machines and gaussian processes. *International Journal of Forecasting*, 2013.

[59] L. Colone, M. Reder, N. Dimitrov, and D. Straub. Assessing the utility of early warning systems for detecting failures in major wind turbine components. *Journal of Physics: Conference Series*, 2018.

[60] Y. Zhao, D. Li, A. Dong, D. Kang, Q. Lv, and L. Shang. Fault prediction and diagnosis of wind turbine generators using SCADA data. *Energies*, 2017.

[61] A. Kusiak and A. Verma. Analyzing bearing faults in wind turbines: A data-mining approach. *Renewable Energy*, 2012.

[62] M. Schlechtingen, I. Santos, and S. Achiche. Wind turbine condition monitoring based on SCADA data using normal behavior models. part 1: System description. *Applied Soft Computing*, 2013.

[63] J. Tautz-Weinert and S. J. Watson. Comparison of different modelling approaches of drive train temperature for the purposes of wind turbine failure detection. *Journal of Physics: Conference Series*, 2016.

[64] P. Mazidi, D. Mian, L. Bertling, and M. A. Sanz Bobi. Health condition model for wind turbine monitoring through neural networks and proportional hazard models. *Journal of Risk and Reliability*, 2017.

[65] J. J. Heckman. Econometric causality. *International Statistical Review*, 2008.

[66] J. Pearl. An introduction to causal inference. *The international journal of biostatistics*, 2010.

[67] J. H. Terrence Parr. How to explain gradient boosting. Online: `https://explained.ai/gradient-boosting/` (01/07/2019).

[68] T. Chai and R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)?– arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 06 2014.

[69] J. Davis and M. Goadrich. The relationship between precision-recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning, ACM*, 2006.

[70] W. Mckinney. Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, 2010.

[71] P. T. Inc. Collaborative data science, 2015. URL `https://plot.ly`.

[72] S. Engelhardt, I. Erlich, C. Feltes, J. Kretschmann, and F. Shewarega. Reactive power capability of wind turbines based on doubly fed induction generators. *IEEE Transactions on Energy Conversion*, 2011.

[73] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[74] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30*. 2017.

[75] L. O. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin. Catboost: unbiased boosting with categorical features. In *NeurIPS*, 2018.

[76] A. Anghel, N. Papandreou, T. P. Parnell, A. D. Palma, and H. Pozidis. Benchmarking and optimization of gradient boosted decision tree algorithms. arXiv:1809.04559, 2018.

[77] A. V. Dorogush, V. Ershov, and D. Kruchinin. Why every GBDT speed benchmark is wrong. *CoRR*, 2018.

[78] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.

[79] M. Rocklin. Dask: Parallel computation with blocked algorithms and task scheduling. In *SCIPY 2015*, 2015.

Appendix A

Extra Normal Behavior Modelling Case Studies

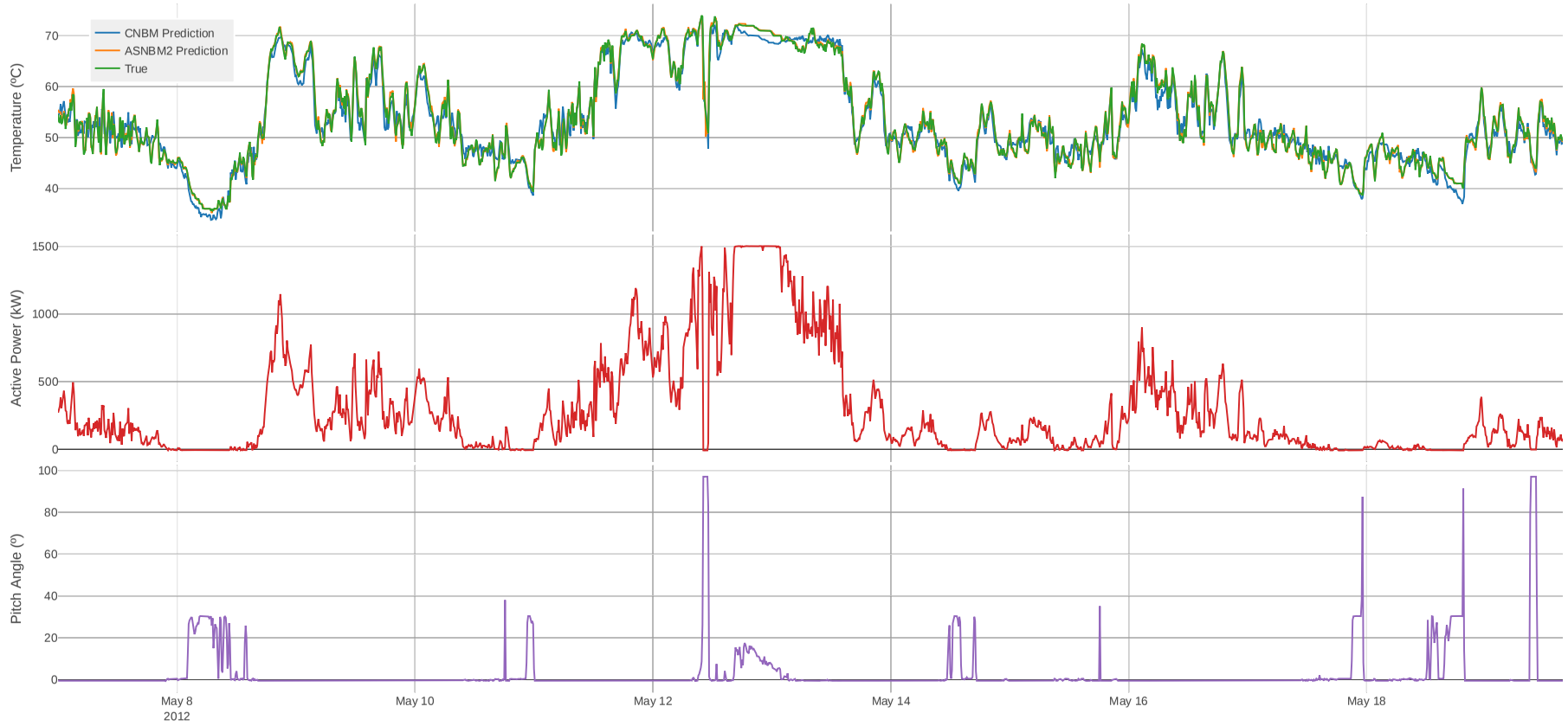


Figure A.1: Gearbox bearing temperature predictions for the CNBM and ASNBM2 against the true values for different operating regimes of an healthy turbine.

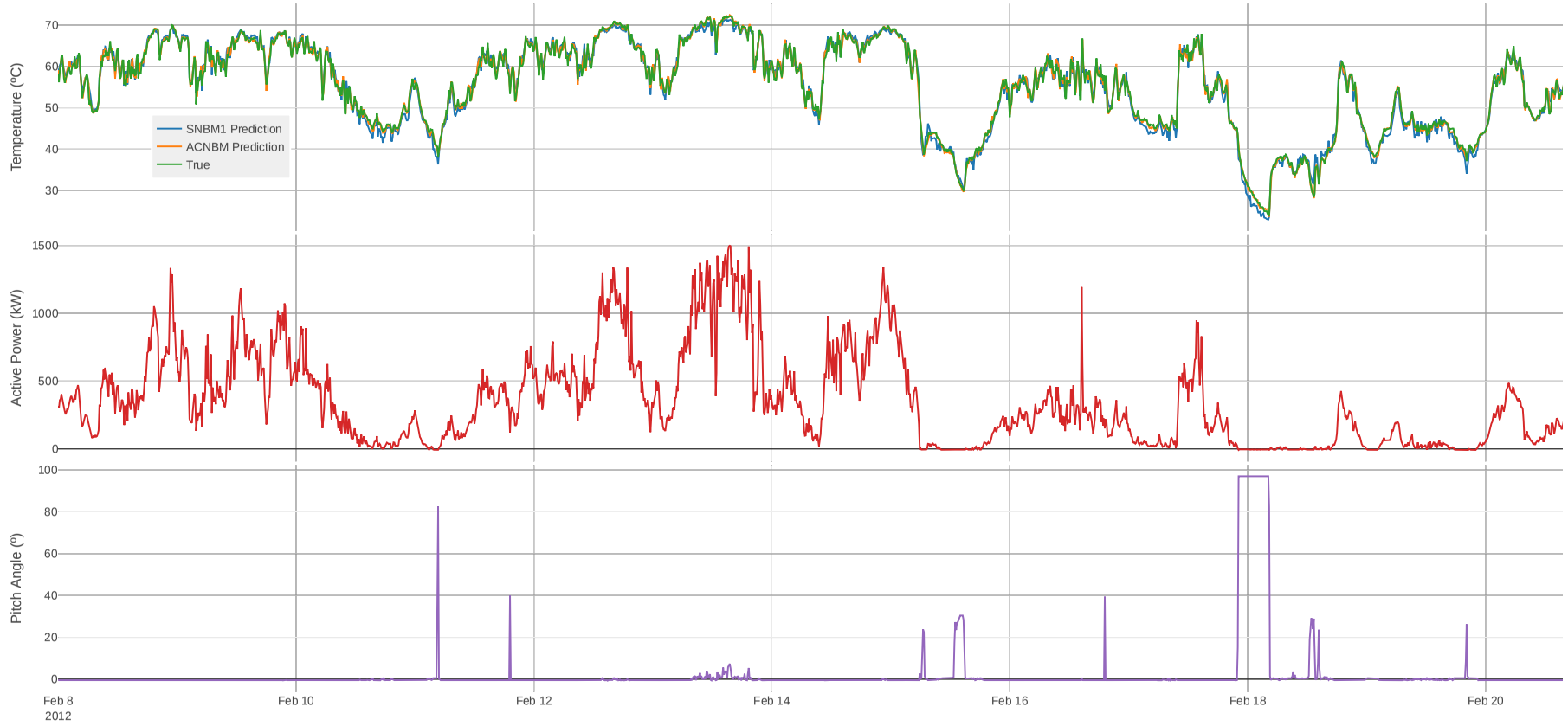


Figure A.2: Gearbox bearing temperature predictions for the SNBM1 and ACNBM against the true values for different operating regimes of an healthy turbine.

Appendix B

Extra Residuals Case Studies

In this appendix the results for the visual inspection of the residuals for the other failures in the dataset will be presented. The following blank space will be left on purpose so that the pairs of images from the same failure with and without active power filter are presented on the same page.

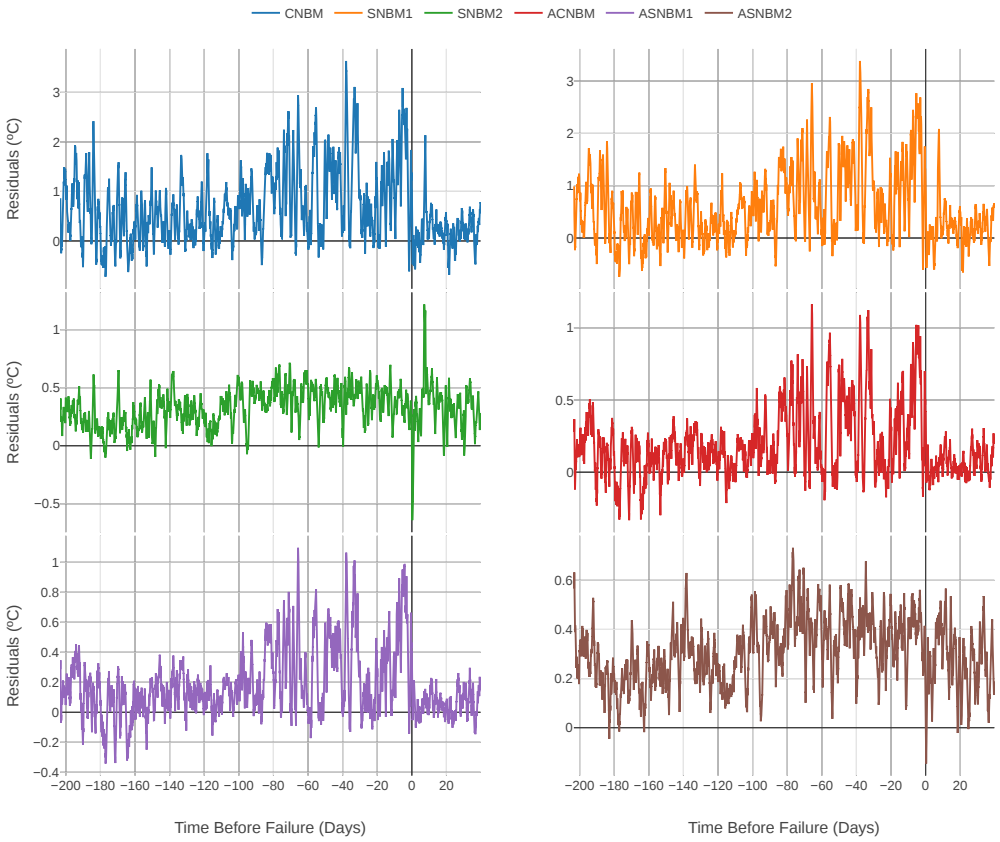


Figure B.1: Post-processed residuals of different models for Failure C

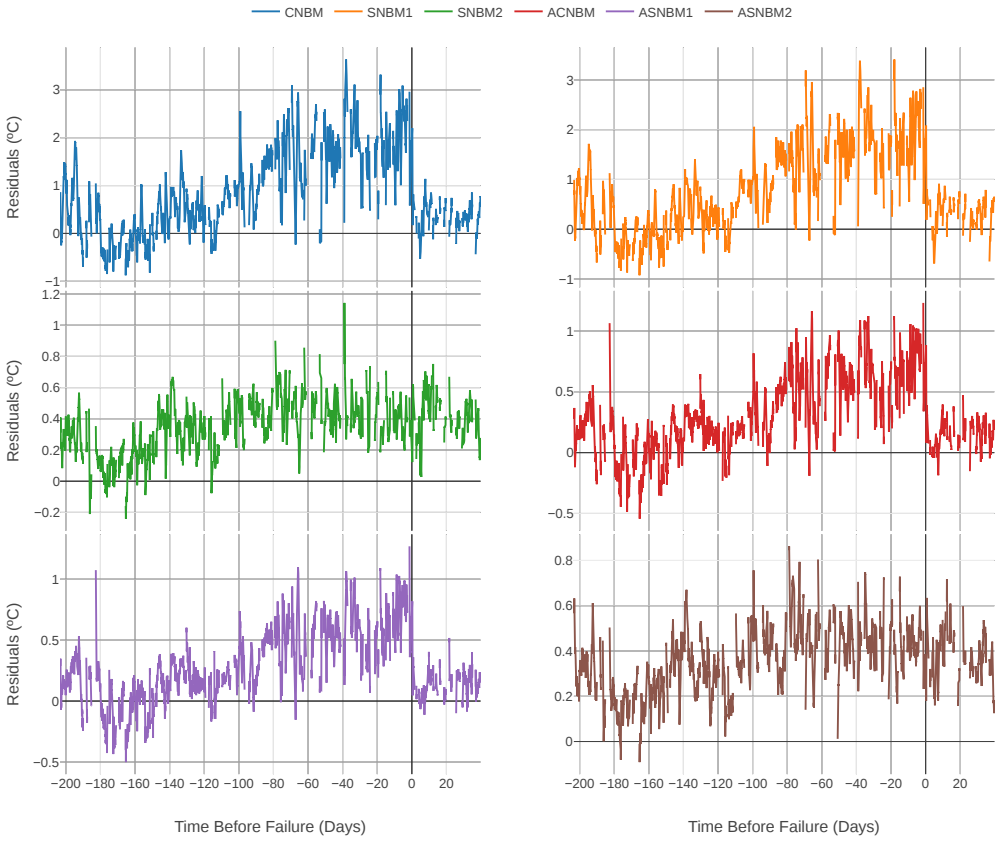


Figure B.2: Post-processed residuals of different models with active power filter for Failure C

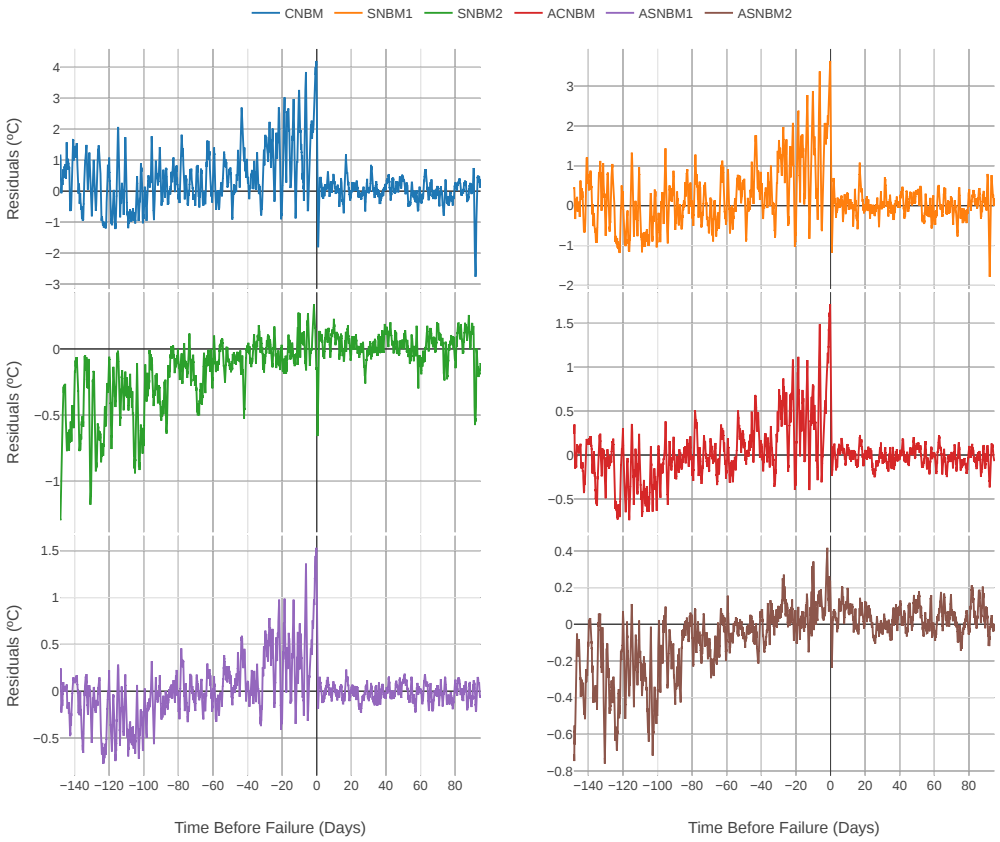


Figure B.3: Post-processed residuals of different models for Failure D

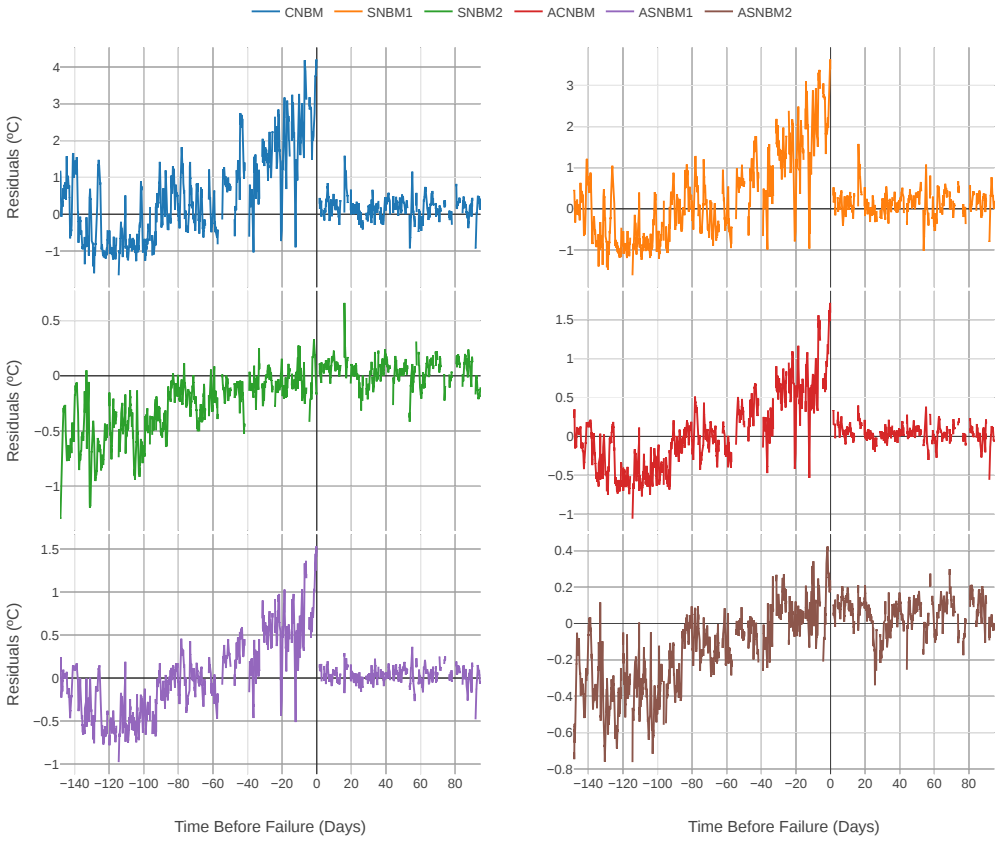


Figure B.4: Post-processed residuals of different models with active power filter for Failure D

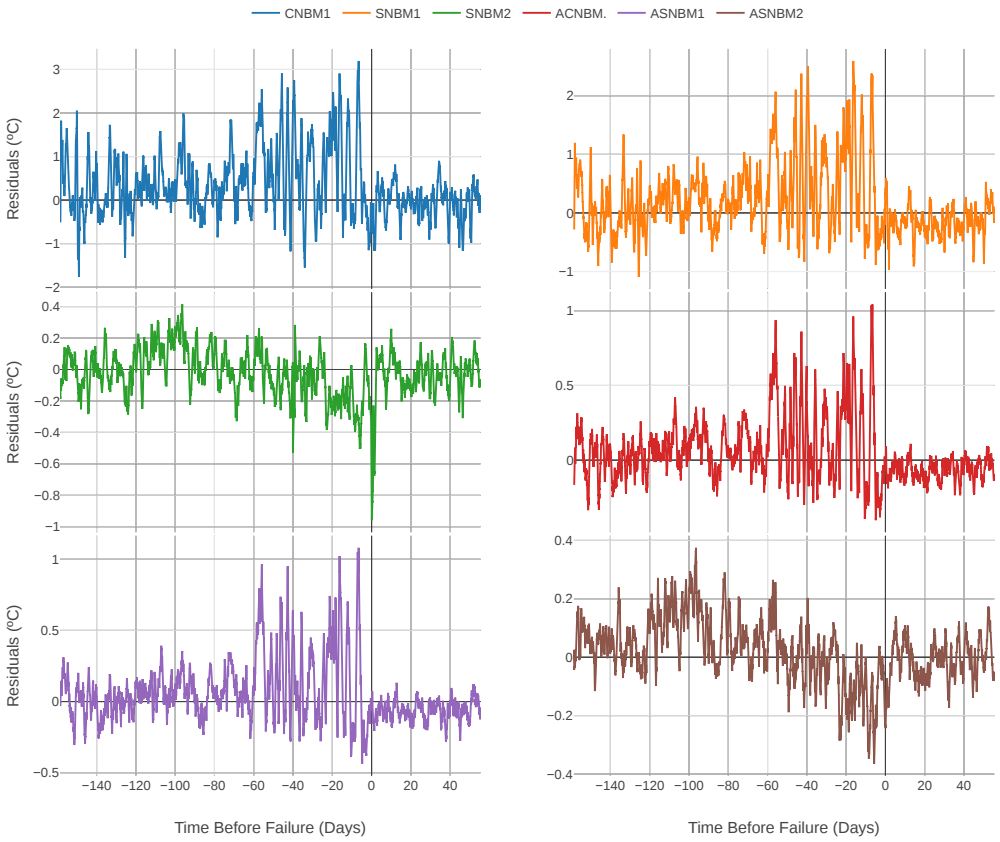


Figure B.5: Post-processed residuals of different models for Failure E

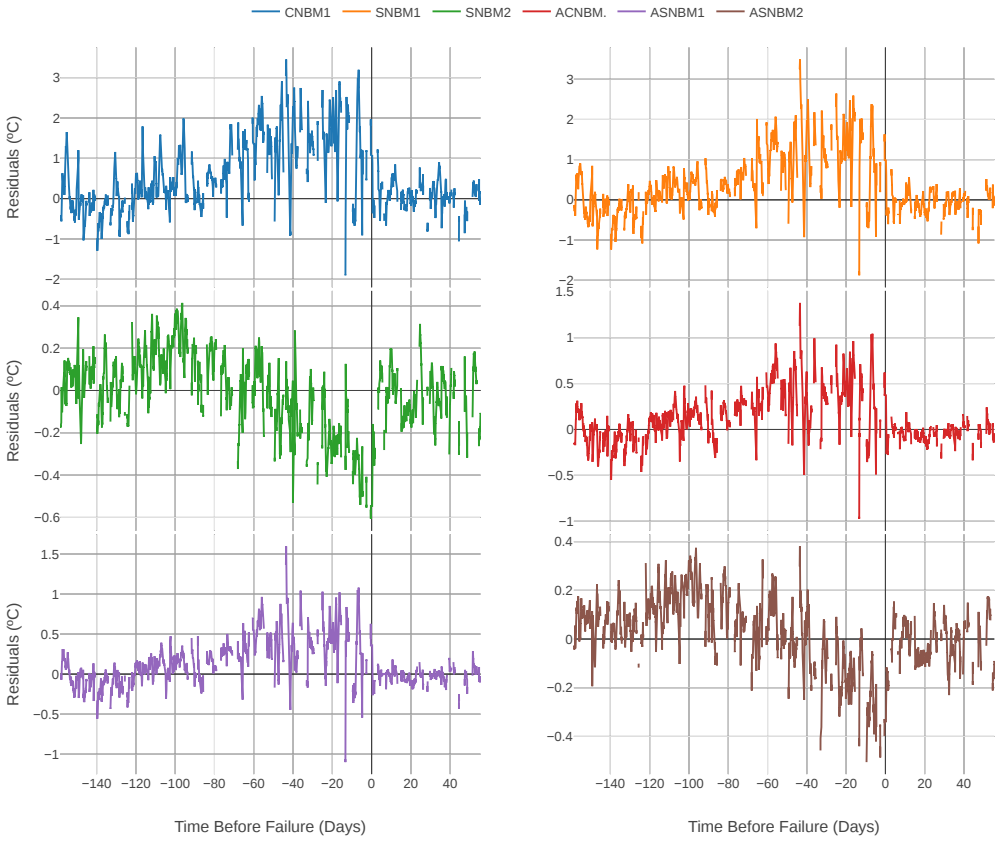


Figure B.6: Post-processed residuals of different models with active power filter for Failure E

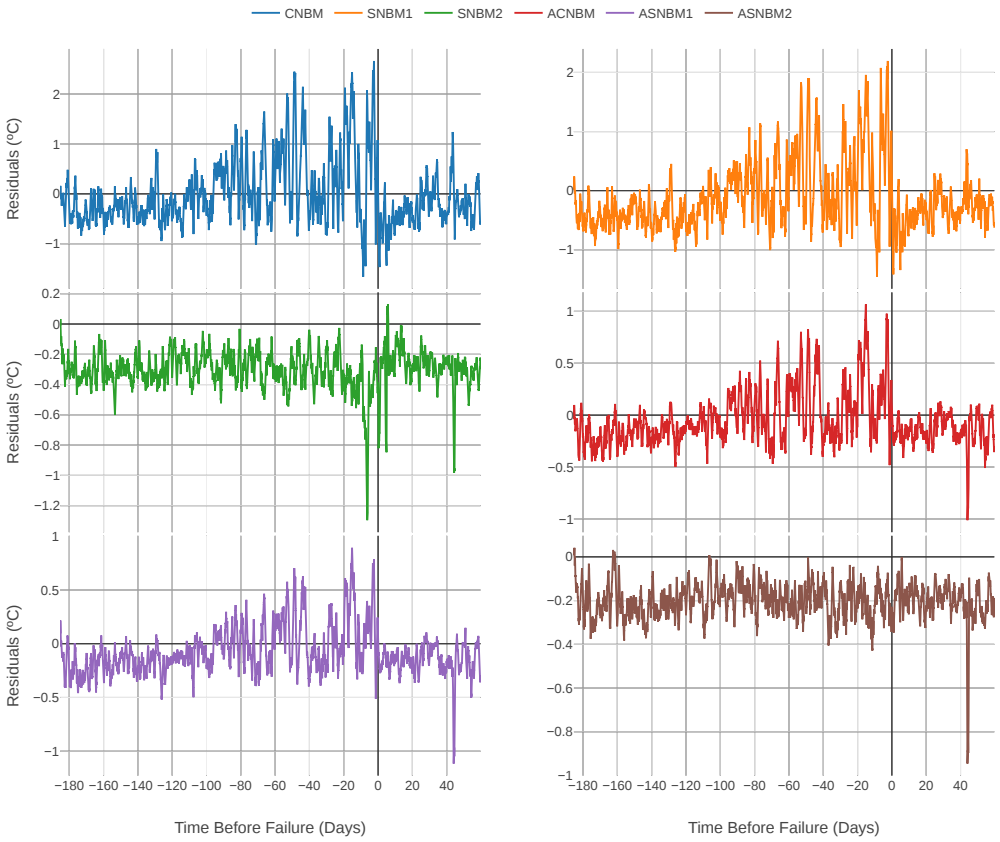


Figure B.7: Post-processed residuals of different models for Failure F

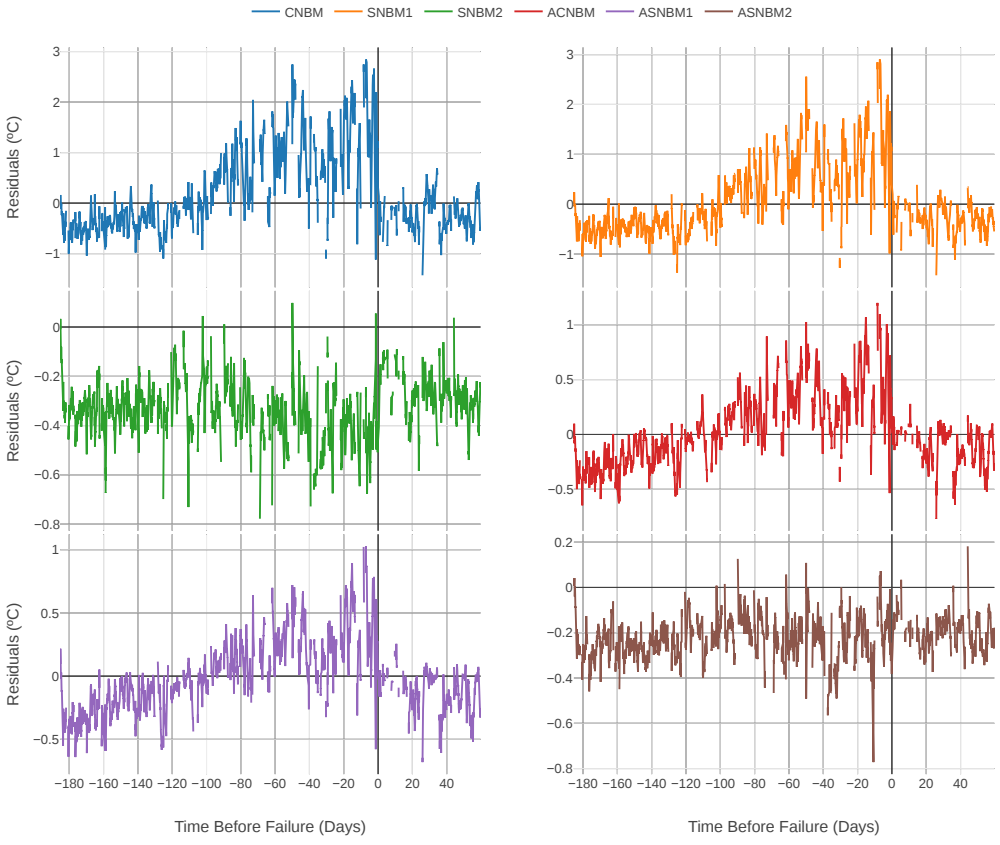


Figure B.8: Post-processed residuals of different models with active power filter for Failure F

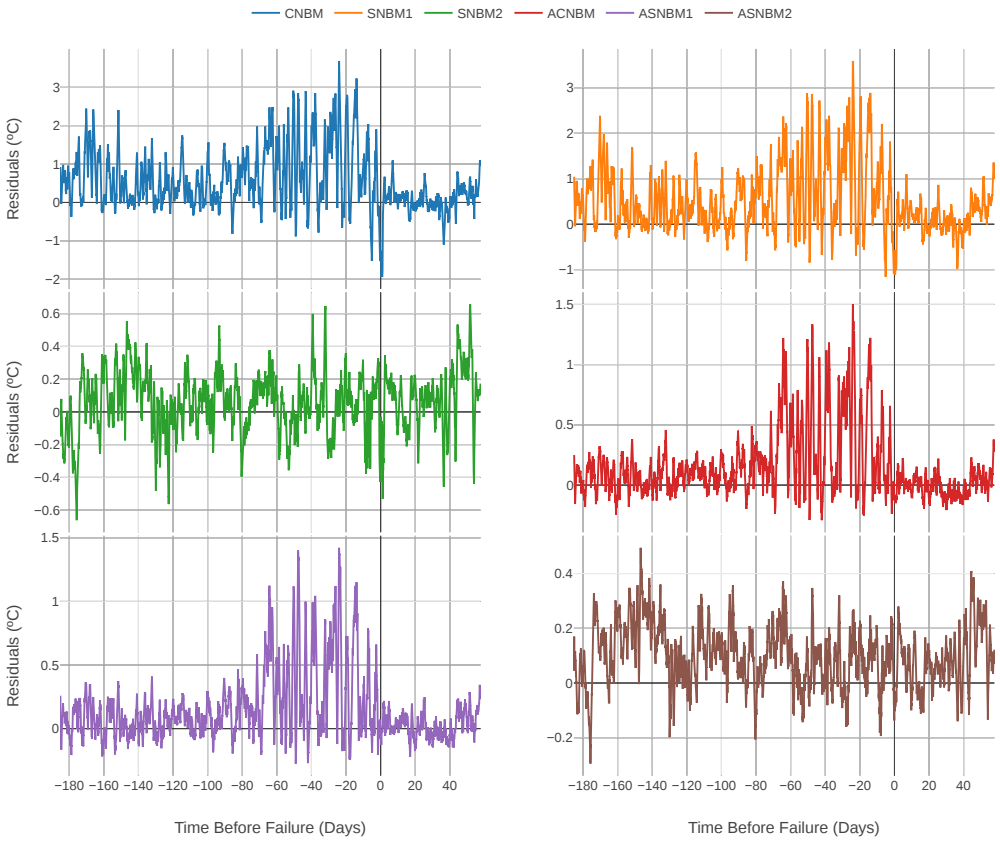


Figure B.9: Post-processed residuals of different models for Failure G

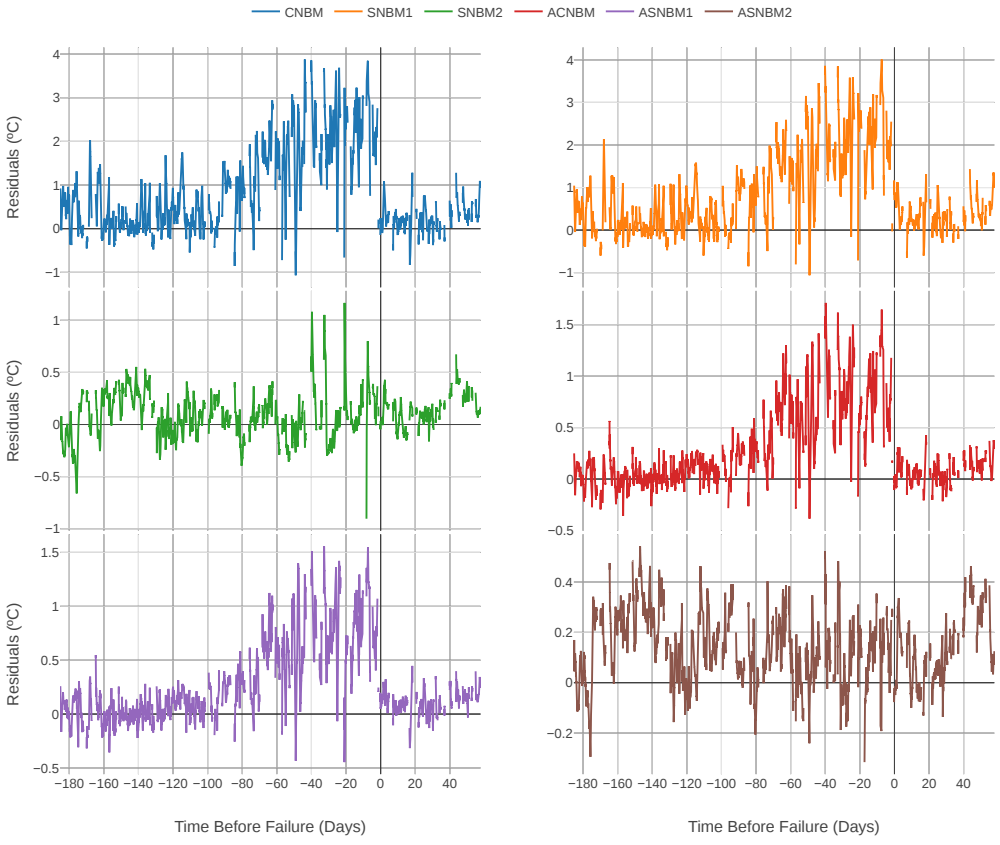


Figure B.10: Post-processed residuals of different models with active power filter for Failure G

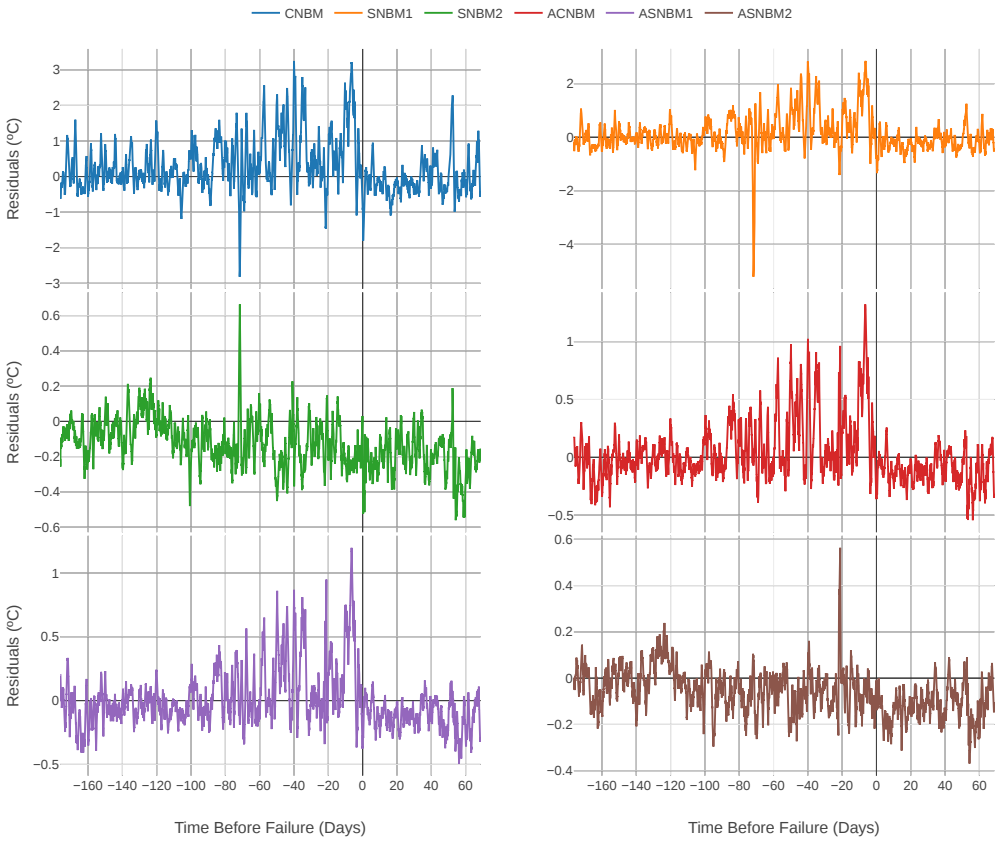


Figure B.11: Post-processed residuals of different models for Failure H

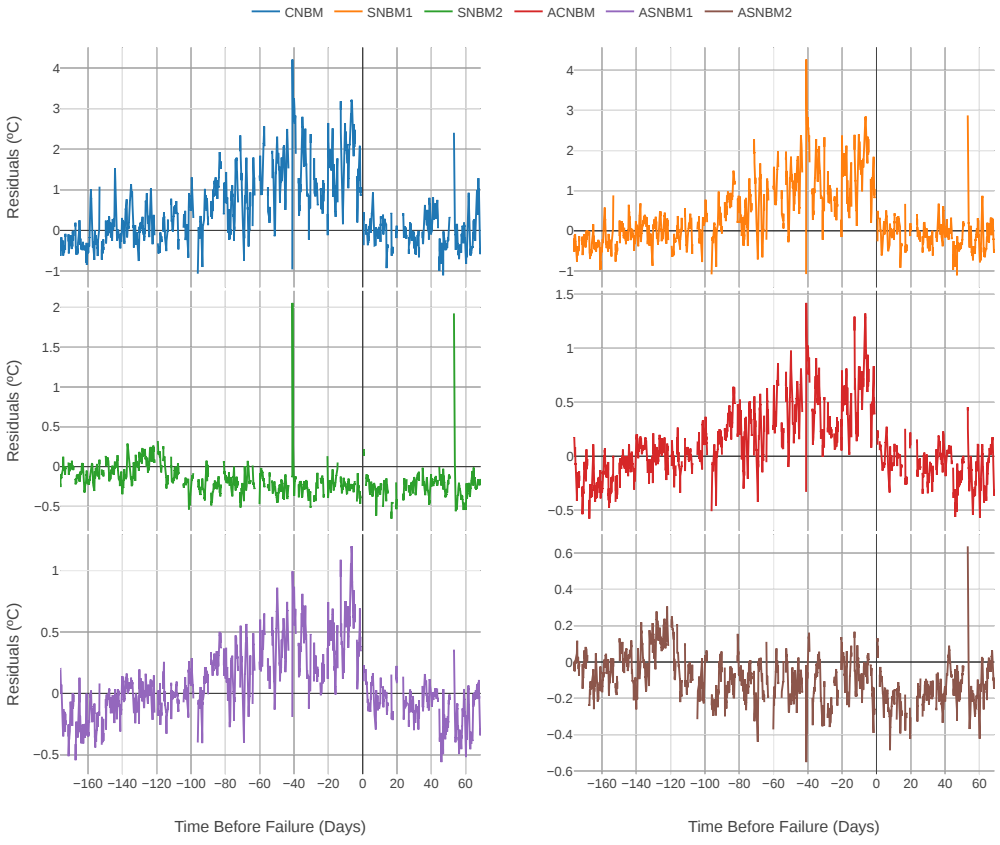


Figure B.12: Post-processed residuals of different models with active power filter for Failure H

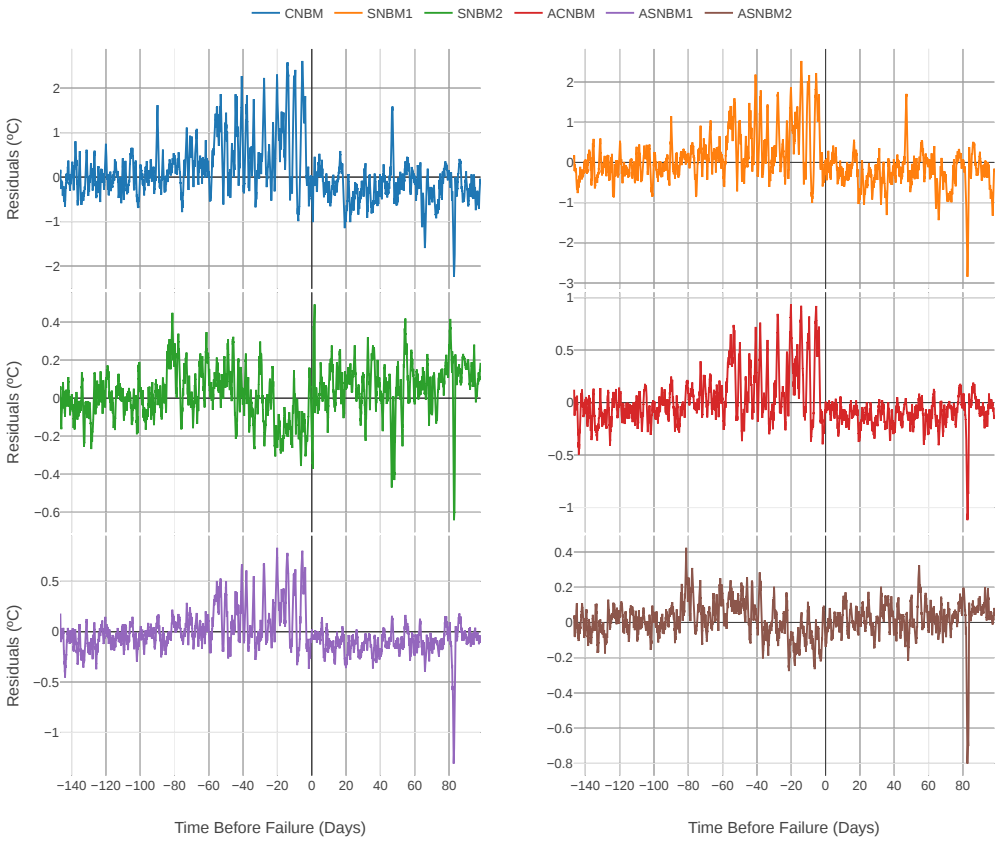


Figure B.13: Post-processed residuals of different models for Failure I

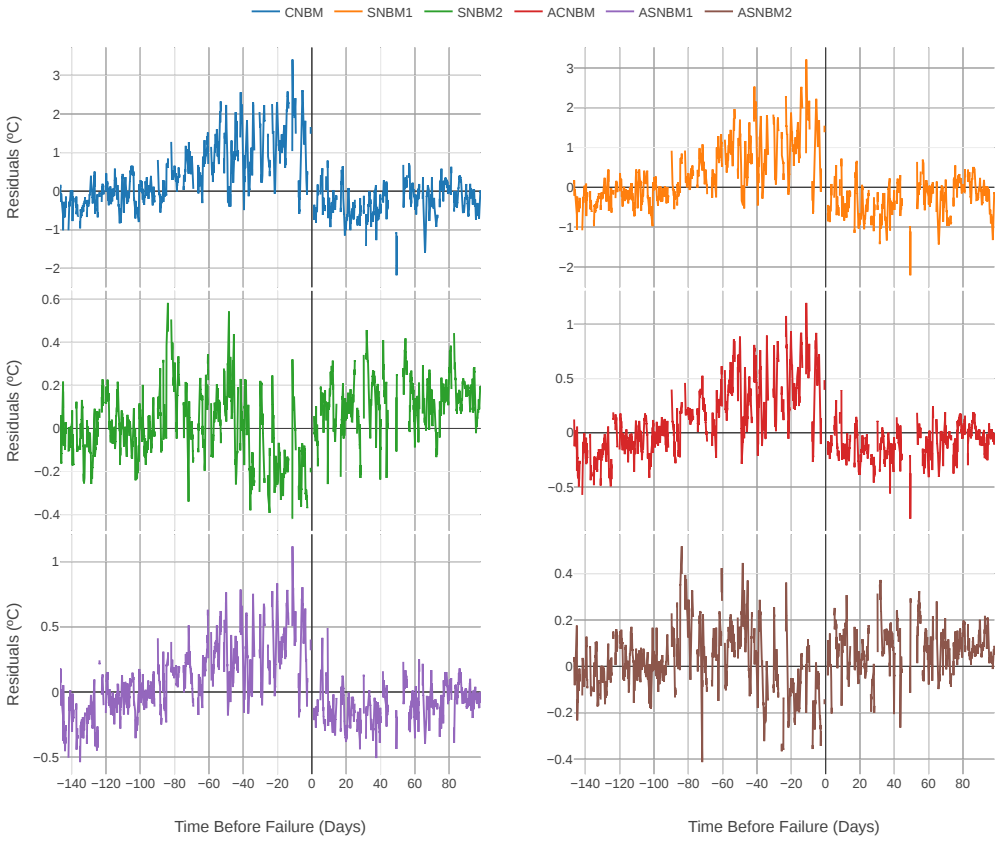


Figure B.14: Post-processed residuals of different models with active power filter for Failure I

Appendix C

Precision and Recall Results

Table C.1: Precision and recall values for the different models with $\Delta t = 10$

	0.00	0.11	0.22	0.33	0.44	0.56	0.67	0.78	0.89	1.00
CNBM	1.0	1.00	1.00	1.00	1.00	1.00	1.00	0.88	0.89	0.60
SNBM1	1.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.82
SNBM2	1.0	1.00	0.12	0.12	0.02	0.02	0.02	0.02	0.02	0.01
ACNBM	1.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.90
ASNBM1	1.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ASNBM2	1.0	0.03	0.03	0.02	0.02	0.02	0.02	0.01	0.01	0.01
Baseline	1.0	1.00	0.40	0.43	0.50	0.50	0.20	0.20	0.22	0.10

Table C.2: Precision and recall values for the different models with $\Delta t = 20$

	0.00	0.11	0.22	0.33	0.44	0.56	0.67	0.78	0.89	1.00
CNBM	1.0	1.00	1.00	1.00	1.00	1.00	0.86	0.88	0.80	0.60
SNBM1	1.0	1.00	1.00	1.00	1.00	1.00	0.75	0.80	0.80	0.82
SNBM2	1.0	1.00	0.12	0.12	0.02	0.02	0.02	0.02	0.02	0.01
ACNBM	1.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.90
ASNBM1	1.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.64
ASNBM2	1.0	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.01
Baseline	1.0	1.00	0.40	0.43	0.44	0.18	0.18	0.19	0.16	0.10

Table C.3: Precision and recall values for the different models with $\Delta t = 30$

	0.00	0.11	0.22	0.33	0.44	0.56	0.67	0.78	0.89	1.00
CNBM	1.0	1.00	1.00	1.00	1.00	1.00	0.88	0.88	0.80	0.45
SNBM1	1.0	1.00	1.00	1.00	1.00	1.00	0.78	0.78	0.67	0.41
SNBM2	1.0	1.00	0.08	0.05	0.02	0.02	0.02	0.02	0.02	0.01
ACNBM	1.0	1.00	1.00	1.00	1.00	1.00	1.00	0.88	0.73	0.13
ASNBM1	1.0	1.00	1.00	1.00	1.00	1.00	1.00	0.62	0.62	0.09
ASNBM2	1.0	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.01
Baseline	1.0	1.00	0.40	0.38	0.15	0.18	0.18	0.19	0.11	0.10

Table C.4: Precision and recall values for the different models with $\Delta t = 40$

	0.00	0.11	0.22	0.33	0.44	0.56	0.67	0.78	0.89	1.00
CNBM	1.0	1.00	1.00	1.00	0.80	0.83	0.88	0.88	0.80	0.11
SNBM1	1.0	1.00	1.00	1.00	1.00	0.83	0.78	0.78	0.25	0.12
SNBM2	1.0	1.00	0.04	0.03	0.02	0.02	0.02	0.01	0.01	0.01
ACNBM	1.0	1.00	1.00	1.00	1.00	1.00	1.00	0.70	0.25	0.06
ASNBM1	1.0	1.00	1.00	1.00	1.00	1.00	0.58	0.58	0.21	0.04
ASNBM2	1.0	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.01
Baseline	1.0	0.25	0.08	0.12	0.12	0.15	0.13	0.09	0.09	0.09

Table C.5: Precision and recall values for the different models with $\Delta t = 50$

	0.00	0.11	0.22	0.33	0.44	0.56	0.67	0.78	0.89	1.00
CNBM	1.0	0.50	0.67	0.75	0.40	0.35	0.35	0.26	0.11	0.11
SNBM1	1.0	1.00	0.67	0.60	0.56	0.56	0.11	0.07	0.06	0.04
SNBM2	1.0	0.01	0.01	0.01	0.02	0.02	0.01	0.01	0.01	0.01
ACNBM	1.0	1.00	1.00	1.00	0.57	0.23	0.23	0.23	0.12	0.01
ASNBM1	1.0	1.00	1.00	1.00	0.44	0.39	0.39	0.39	0.08	0.01
ASNBM2	1.0	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.01
Baseline	1.0	0.06	0.08	0.10	0.12	0.11	0.08	0.08	0.09	0.09

Table C.6: Precision and recall values for the different models with $\Delta t = 60$

	0.00	0.11	0.22	0.33	0.44	0.56	0.67	0.78	0.89	1.00
CNBM	1.0	0.50	0.67	0.33	0.33	0.31	0.23	0.18	0.10	0.05
SNBM1	1.0	1.00	0.67	0.43	0.24	0.07	0.05	0.04	0.04	0.04
SNBM2	1.0	0.01	0.01	0.01	0.02	0.02	0.01	0.01	0.01	0.01
ACNBM	1.0	1.00	1.00	0.50	0.33	0.17	0.12	0.12	0.12	0.01
ASNBM1	1.0	1.00	1.00	0.31	0.31	0.31	0.19	0.19	0.08	0.01
ASNBM2	1.0	0.03	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01
Baseline	1.0	0.04	0.05	0.07	0.06	0.06	0.08	0.08	0.08	0.09

Table C.7: Precision and recall values for the different models with $\Delta t = 70$

	0.00	0.11	0.22	0.33	0.44	0.56	0.67	0.78	0.89	1.00
CNBM	1.0	0.33	0.25	0.20	0.10	0.10	0.09	0.09	0.08	0.05
SNBM1	1.0	0.20	0.20	0.06	0.06	0.04	0.04	0.04	0.04	0.03
SNBM2	1.0	0.01	0.01	0.01	0.02	0.02	0.01	0.01	0.01	0.01
ACNBM	1.0	0.25	0.11	0.11	0.09	0.09	0.09	0.05	0.01	0.01
ASNBM1	1.0	0.17	0.21	0.21	0.14	0.14	0.07	0.07	0.01	0.01
ASNBM2	1.0	0.01	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01
Baseline	1.0	0.04	0.05	0.07	0.05	0.06	0.06	0.06	0.06	0.06

Table C.8: Precision and recall values for the different models with $\Delta t = 80$

	0.00	0.11	0.22	0.33	0.44	0.56	0.67	0.78	0.89	1.00
CNBM	1.0	0.33	0.25	0.06	0.06	0.06	0.06	0.06	0.02	0.02
SNBM1	1.0	0.20	0.20	0.06	0.03	0.02	0.03	0.03	0.02	0.02
SNBM2	1.0	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
ACNBM	1.0	0.25	0.08	0.08	0.08	0.08	0.02	0.02	0.01	0.01
ASNBM1	1.0	0.17	0.21	0.21	0.05	0.05	0.01	0.01	0.01	0.01
ASNBM2	1.0	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Baseline	1.0	0.04	0.05	0.04	0.05	0.05	0.05	0.05	0.05	0.05

Appendix D

ICML Workshop Paper

The Impact of Feature Causality on Normal Behaviour Models for SCADA-based Wind Turbine Fault Detection

Telmo Felgueira [1 2] **Silvio Rodrigues** [2] **Christian S. Perone** [2] **Rui Castro** [3]

Abstract

The cost of wind energy can be reduced by using SCADA data to detect faults in wind turbine components. Normal behavior models are one of the main fault detection approaches, but there is a lack of consensus in how different input features affect the results. In this work, a new taxonomy based on the causal relations between the input features and the target is presented. Based on this taxonomy, the impact of different input feature configurations on the modelling and fault detection performance is evaluated. To this end, a framework that formulates the detection of faults as a classification problem is also presented.

1. Introduction

In 2018, global energy-related CO_2 emissions reached a historic high of 33.1 gigatonnes. These emissions are caused by the burning of fossil fuels, mainly natural gas, coal and oil, which accounted for 64% of global electricity production in this same year (IEA, 2018). Greenhouse gases like CO_2 are responsible for climate change which threatens to change the way we have come to know Earth and human life. For the previous reasons, there has been a global effort to shift from a fossil fuel based energy system towards a renewable energy one. In fact, it is expected that by 2050 wind energy will represent 14% of the world's total primary energy supply (DNV-GL, 2018).

The operation and maintenance costs of Wind Turbines (WTs) can account for up to 30% of the cost of wind energy (EWEA, 2009). This happens because while generators in fossil fuel power plants operate in a constant, narrow range of speeds, WTs are designed to operate under a wide range of wind speeds and weather conditions. This means that stresses on components are significantly higher, which increases the number of failures and consequently the maintenance costs.

There have been recent efforts to monitor and detect incipient faults in WTs by harvesting the high amounts of data already generated by their Supervisory Control and Data Acquisition (SCADA) systems, which, in turn, enables the wind farm owners to employ a predictive maintenance strategy. In fact, it is expected that by 2025 new predictive maintenance strategies can reduce the cost of wind energy by as much as 25% (IRENA, 2016). One of the main methods for monitoring the condition of WTs is building Normal Behaviour Models (NBMs) of the component temperatures. The fundamental assumption behind the use of NBMs is that a fault condition is normally characterized by a loss of efficiency, which results in increased temperatures. By using SCADA data to build a model of the temperatures of the WT components, one can calculate the residuals, which are the difference between the real values measured by the sensors and the predicted values by the model. These residuals can be used to detect abnormally high temperatures that may be indicative of an incipient fault.

Multiple works (Zaher et al., 2009; Mesquita et al., 2012; Brandao et al., 2015) have reported good results using NBMs to predict WT failures, being able to predict failures in WT components months in advance. In these works the authors used as features active power, nacelle temperature and lagged values of the target temperature, thus including autoregressive properties into to the model, to predict the temperatures of various components. In (Schlechtingen & Santos, 2011) and (Bach-Andersen et al., 2016) the authors obtained an important result: although the use of autoregressive features resulted in better temperature modelling performance it also resulted in worse fault detection performance. Another important result was obtained in (Bangalore et al., 2017) and (Tautz-Weinert, 2018), which indicated that using features that are highly correlated with the target also increased the modelling performance but decreased the fault detection performance of the model. Nonetheless, this type of features are still used in many works today, such as (Bach-Andersen et al., 2016; Bach-Andersen, 2017; Colone et al., 2018; Zhao et al., 2017; Tautz-Weinert & Watson, 2016; Zhao et al., 2018; Mazidi et al., 2017). There are also conflicting opinions regarding the use of autoregressive

[*]Equal contribution [1]Department of Electrical and Computer Engineering, Instituto Superior Tecnico, Lisbon, Portugal [2]Jungle.ai, Lisbon, Portugal [3]INESC-ID/IST, University of Lisbon, Portugal. Correspondence to: Telmo Felgueira <telmo.felgueira@jungle.ai>.

features, with some works using them and others not. The main reason behind this is the lack of consistent case studies that evaluate the impact of different features on both the temperature modelling and fault detection performances. It should also be noted that in NBMs it's not trivial that the more features the model has the better its fault detection performance will be. This happens because the model is being trained to minimize the temperature modelling error and not the fault detection one. Having this in mind, this work will present a new feature taxonomy to distinguish different input feature types. Then, the impact of these input feature types on the temperature modelling and fault detection performances will be evaluated.

Finally, evaluating the fault detection performance of different models is not as trivial as evaluating their temperature modelling performance. In fact, there is no standard in the literature regarding how to evaluate fault detection performance. This happens because of the inherent nature of the fault detection problem, in which there is rarely groundtruth. Indeed, there is data of when the failure happened, but there is no information regarding when the fault state started, making it not trivial to formulate as a classification problem. Hence why the majority of the literature evaluates the fault detection results by visual inspection, observing the increase in the residuals before the failure. This is problematic, because comparisons between different models will be highly subjective. Having this in mind, this work will also present a formulation of the detection of faults as a classification problem.

2. Methods

2.1. Data and Training

In this work a dataset composed of 15 turbines during a 6 year period will be used. This data corresponds to SCADA signals with 10 minute resolution. During the year of 2012 there was a total of 5 failures related with the Gearbox IMS Bearing. For these reasons, this will be the component for which an NBM will be trained, with the objective of predicting the corresponding failures.

The models will be trained with data from the beginning of 2007 to the end of 2011 and tested on data from 2012. Periods with faults will be removed from the training data so the model does not learn abnormal behaviour. The models will be implemented with Gradient Boosting Decision Trees (GBDT), which work by iteratively combining weak decision trees into a strong ensemble learner. In terms of implementation, LightGBM (Ke et al., 2017) will be used due to its high computational performance. In terms of optimization, the year of 2011 will be used as a validation set when choosing the number of trees for each model by early stopping. Note that no exhaustive hyperparameter optimization was performed, so all models will use the same hyperparameters besides the number of trees.

2.2. Feature Taxonomy

In the present work we hypothesize that what causes a decrease in fault detection performance is not using input features highly correlated with the target, but using those whose sensors are physically close to the target sensor. If there is an increase in the temperature of a faulty component, the physically close components will also get hotter due to heat transfer. Thus, using physically close components as features to the model may leak information regarding the fault state of the target, making it unable to detect abnormal behaviour. These ideas can be clarified by using appropriate nomenclature. Based on Econometric Causality (Heckman, 2008), we will distinguish features based on their causal relations with the target. If the target is causally dependent of the features, they are causal features. On the other hand, if the target depends on the features but the features also depend on the target they are simultaneity features. Such causal relations are assumed based on the domain knowledge of the physical system.

Based on the taxonomy previously presented, different models will be defined based on their input feature configuration. The simplest model that will be tested is the Causal Normal Behaviour Model (CNBM), which only uses causal features. These are determined based on domain knowledge and will be: rotor speed, active power, pitch angle, wind speed and ambient temperature. All these features characterize the operation regimes of the WT, these are causal features because the gearbox IMS bearing temperature depends on their values, but their values are not dependent on it. For example, variations in the ambient temperature influence the gearbox IMS bearing temperature, but the influences of the latter on the ambient temperature can be disregarded.

On the other hand, simultaneity features will be chosen based on Pearson Correlation, which is a standard first approach for regression problems. The highest correlated feature with the gearbox IMS bearing temperature is the gearbox HSS bearing temperature, which is a simultaneity feature because there is heat transfer between the two sensors, thus meaning that their values are mutually causally dependent. Having this in mind, the Simultaneous Normal Behaviour Model (SNBM) will use all the features from the CNBM plus gearbox HSS bearing temperature. Two more models will be tested, which correspond to the autoregressive versions of the previously described models: Autoregressive Causal Normal Behaviour Model (ACNBM) and Autoregressive Simultaneous Normal Behaviour Model (ASNBM).

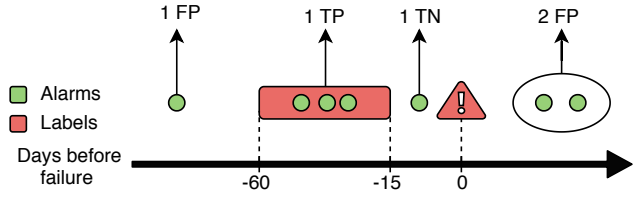


Figure 1. Schematic example of fault detection formulated as a classification problem.

2.3. Fault Evaluation Framework

To develop an evaluation framework for fault detection, one must first formulate it as a binary classification problem where there are two labels: fault and no-fault. Since there is no information regarding the fault state of the component, only the date of failure, it was defined with the wind farm owners that for the failures studied in this work it can be assumed that a fault state would be present at most 60 days before the failure. It was also defined that for the alarms to be useful they should be triggered at least 15 days before the failure. This means that to be considered a True Positive (TP) the alarm must be triggered between 60 and 15 days before the failure. Figure 1 presents a schematic example of the previously described problem formulation. Taking this example, it is important to note that the number of alarms triggered in the prediction window is not relevant, they are all aggregated as 1 TP. The main reason for this, is that if the aggregation is not done, then 4 alarms for the same failure would count as much as 4 detected failures with 1 alarm each. This clearly is not what is intended of the framework, since 1 alarm should be enough to motivate an inspection, and detecting 4 failures with 1 alarm outweighs detecting 1 failure with 4 alarms. Finally, it is also important to note that alarms triggered less than 15 days before the failure are not considered False Positives (FPs), since there is indeed a fault state, it simply is not relevant, so they are considered True Negatives (TNs).

3. Results

In terms of temperature modelling, the models were evaluated on periods of turbines that are known to be healthy. The results, presented in Table 1, indicate that the use of simultaneity features indeed improves the modelling performance, since SNBM obtains better results than CNBM. The use of autoregressive features also improves the modelling performance, since ACNBM and ASNBM obtain better results than their non-autoregressive counterparts. This results make sense, since there are certain regimes of the turbine that are difficult to model without simultaneity nor autoregressive features, such as the turning off of the turbine as noted in (Bach-Andersen, 2017).

Table 1. Regression error metrics for the training and test sets of each model.

Model	Training		Test	
	MAE	RMSE	MAE	RMSE
CNBM	1.48	2.14	1.80	2.62
SNBM	0.87	1.26	1.01	1.41
ACNBM	1.03	1.57	1.14	1.67
ASNBM	0.83	1.22	0.96	1.38

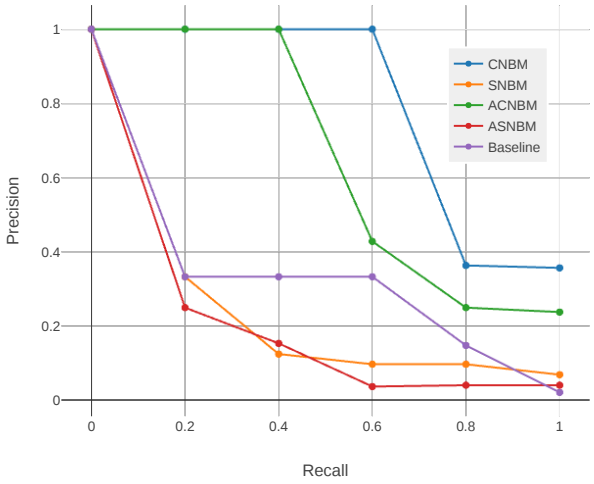


Figure 2. Precision and Recall curves for the different models.

In terms of fault detection, a baseline was defined that consists of setting different thresholds on the distribution of the target temperature and obtaining the corresponding precision and recall. For the models, also different thresholds were applied in the residuals to obtain the different values of precision and recall. The results are presented in Figure 2. As can be seen, the CNBM which obtained the worst modelling performance obtains the best fault detection performance. Also, note that the models with the simultaneity feature are significantly worse than the baseline.

4. Conclusions

An evaluation framework to formulate fault detection as a classification problem was presented. This hopes to contribute to the development of a standard approach for fault detection performance evaluation. Furthermore, a taxonomy regarding the causal relations of the different input feature types was presented, which hopes to make the discussion on how different features affect the performance of models clearer. Finally, it was demonstrated that although autoregressive and simultaneity features increase the modelling performance they decrease the fault detection capabilities of the model. This is an important contribution since the majority of works today still use these types of features.

References

Bach-Andersen, M. *A Diagnostic and Predictive Framework for Wind Turbine Drive Train Monitoring*. PhD thesis, Technical University of Denmark, 2017.

Bach-Andersen, M., Rmer-Odgaard, B., and Winther, O. Flexible non-linear predictive models for large-scale wind turbine diagnostics. *Wind Energy*, 2016.

Bangalore, P., Letzgus, S., Karlsson, D., and Patriksson, M. An artificial neural network based condition monitoring method for wind turbines, with application to the monitoring of the gearbox. *Wind Energy*, 2017.

Brandao, R., Carvalho, J., and Maciel-Barbosa, F. Intelligent system for fault detection in wind turbines gearbox. 2015.

Colone, L., Reder, M., Dimitrov, N., and Straub, D. Assessing the utility of early warning systems for detecting failures in major wind turbine components. *Journal of Physics: Conference Series*, 2018.

DNV-GL. *Energy Transition Outlook*. DNV GL, 2018.

EWEA. *The Economics of Wind Energy*, 2009.

Heckman, J. J. Econometric causality. *International Statistical Review*, 2008.

IEA. *Global Energy & CO2 Status Report*. International Energy Agency, 2018.

IRENA. *The Power to Change: Solar and Wind Cost Reduction Potential to 2025*, 2016.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. In *NIPS*, 2017.

Mazidi, P., Mian, D., Bertling, L., and Sanz Bobi, M. A. Health condition model for windturbine monitoring through neural networks and proportional hazard models. *Journal of Risk and Reliability*, 2017.

Mesquita, R., Carvalho, J., and Pires, F. Neural networks for condition monitoring of wind turbines gearbox. *J. Energy Power Eng.*, 2012.

Schlechtingen, M. and Santos, I. Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection. *Mechanical Systems and Signal Processing*, 2011.

Tautz-Weinert, J. *Improved wind turbine monitoring using operational data*. PhD thesis, Loughborough University, 2018.

Tautz-Weinert, J. and Watson, S. J. Comparison of different modelling approaches of drive train temperature for the purposes of wind turbine failure detection. 2016.

Zaher, A., McArthur, S., Infield, D., and Patel, Y. Online wind turbine fault detection through automated scada data analysis. *Wind Energy*, 2009.

Zhao, H., Liu, H., Hu, W., and Yan, X. Anomaly detection and fault analysis of wind turbine components based on deep learning network. *Renewable Energy*, 2018.

Zhao, Y., Li, D., Dong, A., Kang, D., Lv, Q., and Shang, L. Fault prediction and diagnosis of wind turbine generators using scada data. *Energies*, 2017.