

Article

A Cost-Aware Deep Learning Framework for Gearbox Fault Detection in Wind Turbines

Morten Lindberg Larsen ^{1,*}, Rasmus Døvnborg Frederiksen ^{1,2}, Alex Elkjær Vasegaard ¹ and Peter Nielsen ¹

¹ The AI4OR Group, Department of Materials and Production, Aalborg University, 9220 Aalborg, Denmark; aev@mp.aau.dk (A.E.V.)

² Siemens Gamesa Renewable Energy, Assensvej 11, 9220 Aalborg, Denmark

* Correspondence: morten4545@gmail.com

Abstract

The maintenance operations in the wind sector (particularly offshore) are extremely costly, and thus technologies for optimization of maintenance need to take the high risks related to these costs into account. On the topic of predictive maintenance, any false prediction can lead to high cost incursions, and even a few false predictions can invalidate the predictive maintenance strategy. The costs and benefits of correct and false predictions are not uniform or symmetric, leading the conventional performance metrics to misalign the perceived model performance with the actual business impact. This problem leads us to introduce a cost-aware deep learning framework that accounts for the costs and benefits of predictions. This framework is achieved by loss functions designed to adjust the parameters according to the cost–benefit and not to optimize model accuracy. For the case study of wind turbine generator gearbox failure prediction, several cost-aware loss functions are tested, along with a novel fine-tuning approach that combines them. The case study shows that using cost-aware expected-value loss functions improves cost–benefit performance by an average of 30–57% compared with models trained with conventional loss functions.

Keywords: predictive maintenance; wind turbine gearbox faults; deep neural networks; SCADA data; cost-aware machine learning

1. Introduction

Wind energy is a key driver of the global and European transition to renewable energy sources, with installed capacity reaching 272 GW in Europe by the end of 2023, with 238 GW onshore and 34 GW offshore, meeting 20% of the continent's electricity demand [1]. Countries such as Denmark and Ireland supplied up to 56% and 36% of their electricity demand from wind, respectively. Despite a record 16.2 GW of new installations in Europe in 2023, the region must build 33 GW annually to meet its 2030 target of 425 GW.

This interest in increasing total wind energy capacity has led to wind turbine generators (WTGs) being increasingly installed offshore, where higher energy potential and space allow larger WTGs. These locations are difficult and costly to access for maintenance, while the higher energy yields make maintenance even more relevant. Operation and maintenance (O&M) expenses can account for up to 30% of the total life cycle costs of an offshore wind project [2], making predictive maintenance and condition monitoring critical to reducing downtime and optimizing performance.

Among all failure modes of the WTG, there are several critical components and sources for failures; blades, electrical, gearbox, and so on. While the gearbox failure events



Academic Editor: Tonio Sant

Received: 24 November 2025

Revised: 19 December 2025

Accepted: 24 December 2025

Published: 27 December 2025

Copyright: © 2025 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

are not always the most costly on average, it is particularly susceptible to failure with frequent low-impact failure events and a few high-impact failures. This leads us to use such gearbox events as the best candidate for cost-aware deep learning as differentiation between such failures is necessary. Although deep neural networks have shown strong performance in predicting faults from Supervisory Control and Data Acquisition (SCADA) data, two major challenges remain [3]: the extreme class imbalance between normal and fault conditions, and the disconnect between model accuracy and the real-world cost impact of misclassifications. In practice, different types of faults lead to vastly different business consequences. For example, misclassifying an oil pump failure might be a far more costly endeavor than misclassifying a temperature warning. Thus, models that optimize only for accuracy may fail to minimize operational costs.

The remainder of this paper is structured as follows. Section 2 provides background on fault prediction in wind turbines and cost-aware machine learning. Section 3 describes the data, the model architecture, and the proposed loss functions. Section 4 presents the experimental results and evaluation metrics. Finally, Section 5 offers a discussion of the findings and their implications for predictive maintenance in the wind energy sector, before drawing a conclusion in Section 6.

Background and Related Work

Loss functions are central to assessing machine learning model performance. However, in many applications, standard loss formulations are insufficient because they do not reflect the true costs associated with different types of error. Two common situations motivate the use of cost-sensitive learning methods: imbalanced datasets [4] and asymmetric misclassification costs [5]. In the case of class imbalance, the events of interest may be rare, and a model trained with a conventional loss will tend to favor the majority class simply because this minimizes the average error. In other settings, misclassification errors carry unequal consequences, meaning that certain mistakes should be penalized more heavily than others. In both scenarios, practitioners seek mechanisms to incorporate these preferences directly into the learning process. A widely used approach is to modify the loss function—either by introducing class-specific weights or by designing custom loss functions that reflect the relative importance of different prediction outcomes [5–7].

Within specific domains and fields, there exists a set of requirements that are imposed on machine learning solutions, which necessitate these cost-sensitive learning schemes. Examples of such domain-specific problems are readily found, such as with satellite imaging, as in [8] where α —balanced binary cross entropy and $F\beta$ loss functions are used to specifically improve classification recall at the expense of precision, and in [9] where a scaled version of the mean squared error is used to fine-tune regression predictions. Physics-informed custom loss functions are used to ensure that predictions comply with physical laws, as in [10], or to improve the accuracy of predictions of physical systems, as in [11,12]. These physics-informed loss functions are also used in our field of research to estimate long-term fatigue in WTG [13]. In medicinal studies, often unbalanced data sets are the reason for using cost-sensitive learning [14,15], but the asymmetric consequences of wrongful diagnosis are also a good reason for cost-sensitive learning [16,17].

Although the above examples showcase a variety of rationales for cost-sensitive learning in many fields, the most relevant reason in relation to this paper is purely monetary, as the notion of long-term profits is quite uncertain, fuzzy, and vague. Thus, maximizing this involves many other temporal, operational, political, and climate considerations [18]. Wrongful predictions in general can be costly and have wide-ranging consequences. This is mirrored in finance, where the consequences of wrongful predictions are significant, and long-run profit-seeking is the overall goal [19–21]. Often, such machine learning

problems are solved with a custom cost matrix describing the cost–benefit of each correct and false classification of the model. This can be used to weigh the loss function during training to properly bias the model towards the most optimal predictions according to these cost–benefits. Such solutions require the investigation and formulation of this cost matrix, which, in itself, can become a substantial challenge.

Our aim is to apply similar cost-sensitive learning to the training of prediction models for WTG failure events. Machine learning approaches to WTG failure prediction are numerous, with reviews available in [22–24] with the general topic explored well in [25]. Specifically on the topic of prediction and detection of gearbox failures, there are successful models presented in [26–30]. Special inspiration for the base deep learning architecture for this paper is found in [26,31–33]. While these sources are presenting a variety of methods for prediction of failure events for WTGs, the majority of these approaches are limited by using a standard loss functions which optimize the training for accuracy over costs.

The objective of cost optimization in the wind service sector is often used to measure the effectiveness of strategies and services, such as maintenance optimization [34], inventory management [35], and WTG configurations [36]. As such, cost-optimization is central to the continued improvement of the wind sector. This leads us to search for references within the wind sector where such cost-matrices are created with actual cost–benefits. However, we are only able to find a limited number of cost-sensitive learning methods with respect to imbalanced data sets [37–39], where the cost matrix is created from the class sample rate.

This leads us to the following research gaps.

1. The available research on cost-sensitive machine learning solutions for failure prediction in the wind sector is limited.
2. None of the research on cost-sensitive machine learning solutions for failure prediction in the wind sector is conducted using deep learning models.
3. None of the available research on cost-sensitive machine learning solutions for failure prediction in the wind sector has found a cost matrix based on actual failure costs.

We propose to work on these gaps in tandem. We will introduce a deep learning recurrent neural network model that predicts gearbox failures. With a more extensive dataset than what is available in many other sources on this topic, the issue of imbalanced classes is not relevant in this paper. However, the cost-sensitive learning approaches remain relevant. Instead of creating a class-sample weight cost matrix, we propose using historical failure data from the WTGs to create a cost matrix that reflects the failure costs in monetary terms. This will lead to models that are not necessarily well-performing on conventional metrics but perform well in accordance with the profit from the implementation. See Figure 1 for an overview of the problem domain and the functionality of the solution.

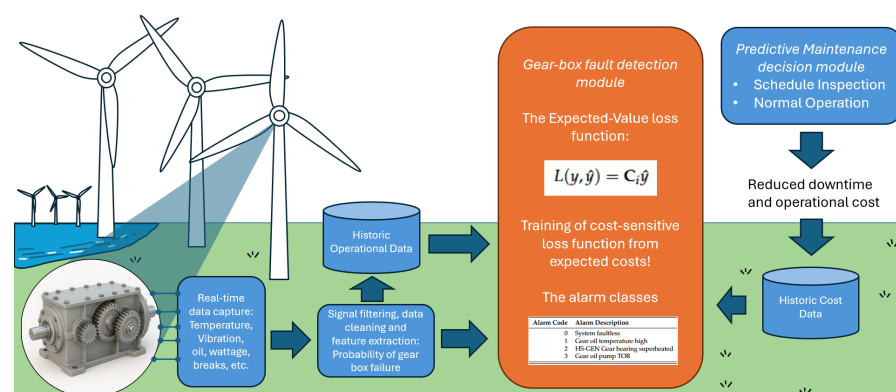


Figure 1. An illustrative representation of the predictive maintenance workflow for gearbox fault detection.

This paper addresses the above described challenges using a dataset comprising 12,000 samples with 106 feature time series. We formulate the problem as a cost-aware, multi-class classification task and introduce a novel loss function—the Expected Value Loss (EV Loss)—which enables direct cost-sensitive optimization using a domain-specific cost matrix based on the average turbine downtime associated with each type of fault. We evaluate multiple configurations of a deep neural network and show how aligning model training with operational costs leads to more economical and practical fault prediction systems.

2. Methodology

The base methodology of this paper relies on a conventional deep learning workflow. The data is pre-processed, split into train and test sets, a model is trained, and then evaluated. This workflow does not introduce new ideas or novelties by itself, as the central focus of this paper is the selection of the loss function in the model training phase. As such, the different loss functions scrutinized are presented below. The next section will introduce the training process of the case study in more detail, explain how the business costs are determined, and explain how the loss function weights are calculated based on these business costs.

2.1. Categorical Cross-Entropy

Categorical Cross-Entropy (CCE) is one of the most commonly used loss functions for multi-class classification problems. It is not cost-sensitive and is included in this paper as a benchmark against which we can compare the performance of other cost-aware loss functions. It can be expressed as follows.

$$L(y, \hat{y}) = - \sum_{i=1}^n y_i \log(\hat{y}_i), \quad (1)$$

where y_i is 1 if i is the true class and 0 otherwise, $\hat{y}_i$ is the predicted probability of class i , and n is the number of classes. CCE is commonly used because of its simplicity. Due to the slope of the logarithmic function, poor predictions will result in large losses and will, in turn, quickly lead the model to make adjustments to the parameter weights to achieve better predictions.

2.2. Class-Weighted Categorical Cross-Entropy

Class-Weighted Categorical Cross-Entropy (CWCCE) is a modification of CCE where a weight for each class is used to account for the importance of each class. In CWCCE, a weight vector $\mathbf{w}$ is used to determine a multiplier for each class, resulting in a loss function expressed as:

$$L(y, \hat{y}) = - \sum_{i=1}^n \mathbf{w}_i y_i \log(\hat{y}_i), \quad (2)$$

where $\mathbf{w}_i$ is the weight for class i . Although this loss function is cost-sensitive by multiplication of $\mathbf{w}$, it must be noted that this weight vector requires careful consideration in its construction. Even if the information on operational costs is readily available, it is not trivial to go from the actual costs related to each type of classification to the weights for each class to be used in CWCCE. In Section 3.2 we will explain how different weight vectors were derived for the case study of this paper.

2.3. Expected Value Loss Function

A problem with the CCE variants is that they only consider the predicted probability of the true class. In this paper, we propose a new type of loss function that considers all

predicted probabilities. This loss function will be referred to as EV loss and is formulated as follows:

$$L(y, \hat{y}) = \mathbf{C}_i \hat{y}, \quad (3)$$

where $\mathbf{C}_i$ is the i 'th row of the cost matrix $\mathbf{C}$ and i is also the true class. EV loss attempts to incorporate information on misclassification costs, which relate to the misclassification itself, as well as the opportunity cost depending on the true class. All this information is included in the cost matrix, and by evaluating the entire prediction score vector in the loss, the model has the ability to adjust according to a loss that includes all classes and costs. Furthermore, by multiplying these costs and the predicted probabilities, the loss function resembles the expected value calculations performed in economics.

As in the CWCCE, the construction of the cost matrix $\mathbf{C}$ is the main challenge of this approach and should be chosen to reflect the cost of each classification and misclassification; this way, the EV loss function would assign a loss that is linearly proportional to the expected value in all classes. Ultimately, by including all predicted probabilities in the loss function, the advantage is that EV loss will be able to train a more operationally efficient and cost-sensitive model. Additionally, it should also result in predicted probabilities that are already calibrated with respect to the cost, as opposed to predicted probabilities from models using CCE where predictions will often skew towards either 1 or 0 and require the use of a calibration model if they are to reflect the actual probabilities of an outcome and be used for business impact analysis.

2.4. Fine-Tuning Approach

Due to the poor accuracy performance of the EV loss function during testing, we theorized a fine-tuning approach using multiple loss functions during training to cover each loss function weaknesses. This is carried out by training a model using a loss function for a certain number of epochs and then switching to another loss function to further fine-tune the parameters of the model. In this paper, a fine-tuning model using CCE for 50 epochs, followed by EV loss for 50 epochs, is presented.

3. Case Study

The loss functions presented above will be evaluated on a dataset of WTG operational data and gearbox failure events. This data is collected from the WTG SCADA system. The operational data consists of a selection of time series variables from a large number of WTGs (all of the same design platform), with the objective of predicting and classifying gearbox fault events. The gearbox event data is also provided by the SCADA system with information on all gearbox faults, describing when it occurred, the exact alarm code and type, how long it lasted, etc. With predictions on such events, the information can be used for predictive maintenance. Illustrated in Figure 2 is a diagram of the experimental framework.

Each sample consists of 53 sensor readings taken at 10-min intervals over 5 h leading up to an event (either a fault or a non-event). For each sensor reading, the 10-min mean and standard deviation are provided, resulting in a total of 106 timeseries features. The gearbox failure events of interest are three different faults, which are described in Table 1. The training labels are the occurrences of one of these three faults, in addition to a normal operating state with no failures present. The samples for the failure events are collected so that the failure occurs at a random time within 10 min of the last sensor reading, ensuring that the prediction actually targets a future operational state of the WTG. Normal operating state samples are collected completely at random; however, it is ensured that there are no imminent gearbox failure events.

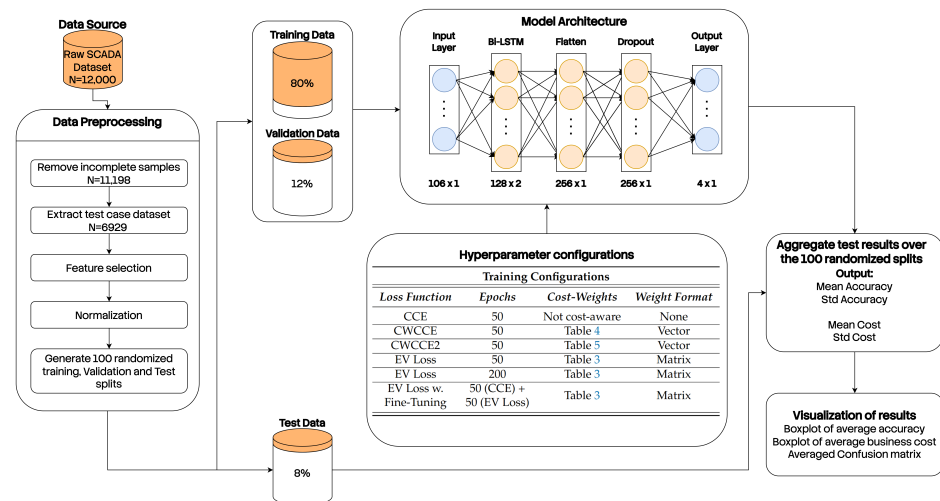


Figure 2. Diagram of the experimental framework for the case study.

Table 1. This table presents the alarm codes, their associated description, and the average downtime for each event.

Alarm Label	Alarm Description	Average Downtime	Samples
0	System faultless	0 s	4505
1	Gear oil temperature high	835 min	966
2	Gear bearing superheated	622.5 min	343
3	Gear oil pump outage	2517 min	1115

For the model training process, the dataset is split into train–validation–test sets, with 80%, 12%, and 8% of the samples, respectively. Since the focus of this paper is to test the effectiveness of various loss functions for cost-sensitive training, the train–validation–test split is repeated 100 times to generate 100 unique train–validation–test sets. This is done because it is unproven whether the outcome of the newly proposed EV loss function is consistent. By replicating the training on 100 sets, we can evaluate the consistency of each loss function.

3.1. Estimation of Classification Costs

Table 1 also presents the average downtime for each type of fault. This is based on 15 years of historical data collected from all WTGs, from which the training samples are collected, and consists of a total of 82,340 failure events. This average downtime for each of the gearbox failure types will be used as the basis for the loss function weights and cost matrices.

We make the following assumptions to simplify the process. It must be noted that, for this process to be industrially viable, some of the assumptions described above cannot be made, and more effort must be made in the formulation of these weight structures to make them as reflective of the true costs as possible. The most straightforward improvement is to change the minutes of downtime into actual cost in Euros. For the purpose of demonstrating the method, these assumptions and the use of minutes of downtime as costs are sufficient.

1. A correct classification will lead to the complete prevention of the failure event, and thus will result in no downtime. This is the overarching objective of predictive maintenance.
2. The misclassification of a fault as normal operating conditions leads to the occurrence of the failure event, resulting in downtime equal to the average downtime of the specific fault.

3. The misclassification of normal operating conditions as a fault leads to unnecessary maintenance. For this, the costs incurred are assumed as 50% of the average downtime caused by the incorrectly predicted fault.
4. The misclassification of a fault as another type of fault still leads to preventive measures; however, since an incorrect type of fault was predicted, it could result in the use of incorrect procedures. For this reason, the cost is assumed as 25% of the average downtime of the true failure event.

Based on these assumptions and with the average downtime per fault, we can construct a cost matrix for the 4 classes in the style of a confusion matrix, as presented in Table 2.

Table 2. Cost matrix—absolute costs. All numbers in seconds of downtime.

		Predicted			
		0	1	2	3
Actual	0	0.00	25,053.87	18,676.07	75,510.73
	1	50,107.74	0.00	12,526.94	12,526.94
	2	37,352.14	9338.03	0.00	9338.03
	3	151,021.47	37,755.37	37,755.37	0.00

In the context of predictive maintenance, the absolute costs are not the only measure of interest. The value created is compared to taking no action on the predictive information. As such, we need a cost matrix that reflects the costs and benefits relative to taking no action. This can be accomplished by subtracting the cost of taking no action from the cost matrix in Table 2. This leads to the cost matrix seen in Table 3. In this cost matrix, negative costs represent a monetary gain from the actions taken compared to taking no action.

Table 3. Cost matrix—relative to taking no action. Numbers are in seconds of downtime, and negative numbers are seconds of downtime saved when compared with taking no action.

		Predicted			
		0	1	2	3
Actual	0	0.00	25,053.87	18,676.07	75,510.73
	1	0.00	−50,107.74	−37,580.81	−37,580.81
	2	0.00	−28,014.10	−37,352.14	−28,014.10
	3	0.00	−113,266.10	−113,266.10	−151,021.47

3.2. CWCCE Weights

For CWCCE, we need to find an appropriate way to convert the cost of each type of classification or misclassification into a single weight for each class. To accomplish this, we took the mean of each cost column in Table 3 and standardized it by dividing by the smallest weight. This gives us the class weights presented in Table 4.

Table 4. Weight matrix used in CWCCE.

Class	Weight
0	2.13
1	1.34
2	1.00
3	4.04

The above weights do not take into account the frequency of each class in the dataset, which can lead to suboptimal results, as the frequency of uneven occurrences can result in an implementation that does not properly reflect the weights. By multiplying the class weights by the number of occurrences of each class and standardizing this time by dividing by the largest weight, we obtain the class weights presented in Table 5 (the model trained with these weights is hereafter referred to as CWCCE2).

Table 5. Weight matrix used in CWCCE2.

Class	Weight
0	1.0
1	0.135
2	0.0358
3	0.470

3.3. Model Training

The deep learning architecture and hyperparameter configuration have been determined through iterative testing with the CCE loss function to provide the best performance. To create the most comparable results, all further testing of the other loss functions will be performed with the same model architecture so that any resulting differences in performance are due to the loss functions specifically.

This deep learning architecture is based on a recurrent neural network approach, more specifically, a layer of bidirectional Long Short-Term Memory (LSTM) cells followed by a fully connected dense layer. This totals 240,644 trainable parameters. The complete architecture is presented in Figure 2. Table 6 consists of a brief overview of the hyperparameter configurations to be tested.

Table 6. Overview of the used loss functions and their configurations.

Loss Function	Training Configurations		
	Epochs	Cost-Weights	Weight Format
CCE	50	Not cost-aware	None
CWCCE	50	Table 4	Vector
CWCCE2	50	Table 5	Vector
EV Loss	50	Table 3	Matrix
EV Loss	200	Table 3	Matrix
EV Loss w. Fine-Tuning	50 (CCE) + 50 (EV Loss)	Table 3	Matrix

The tests with the loss function indicated as EV Loss 200 Epochs and EV Loss w. Fine-Tuning require a bit of explanation. During the testing of EV Loss, it became apparent that the results did not converge within the specified epochs, unlike the other loss functions. This might be due to the linear nature of the EV loss function. For this reason, we wanted to test whether it would be beneficial to train for additional epochs.

As for EV Loss w. Fine-Tuning, this is an attempt to achieve the best results from both cost-sensitive and conventional loss functions. The fine-tuning training will use the CCE loss function during the first of two training sessions to achieve a good baseline accuracy for the model, followed by fine-tuning this model with the EV Loss for an additional 50 epochs to achieve better cost performance.

4. Case Study Results

The training procedure is repeated 100 times for each loss function on a unique train–validation–test split. This allows us to evaluate the consistency of the results achieved by each loss function. In this section, these results are presented and evaluated.

4.1. Accuracy

The accuracy of the models on the test set is measured for each training iteration and plotted as a boxplot in Figure 3. This boxplot makes it clear that all variants of CCE will achieve high accuracy performance with low deviation across the 100 iterations.

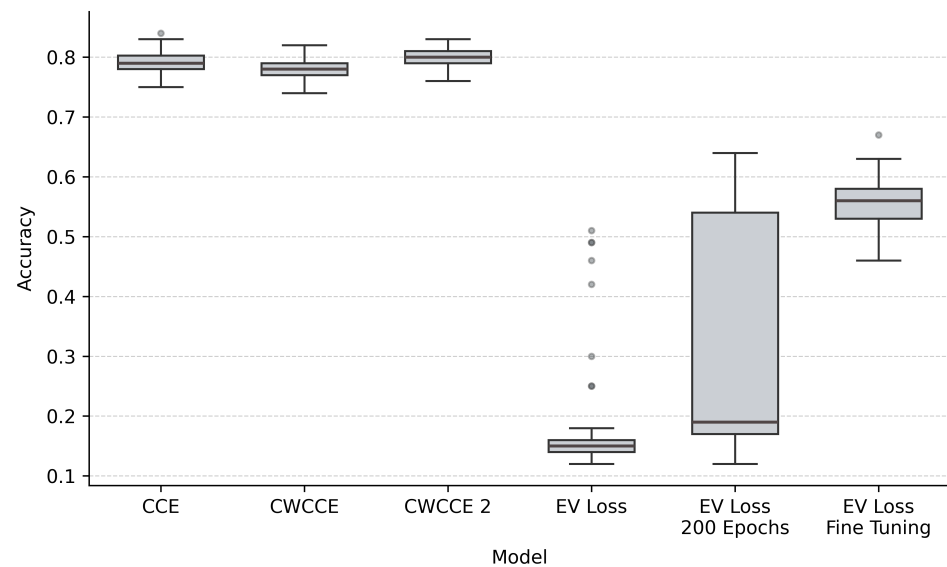


Figure 3. Boxplot of the accuracy for each model. Each model is trained over 100 iterations.

For models trained only with the EV loss, the accuracy is significantly worse. Within 50 epochs of training, there are some outliers with better accuracy than the bulk of iterations, and while this percentile of training iterations with better accuracy improves with 200 training epochs, this loss function seems to deliver quite varied results. This variability seems to be reduced by the fine-tuning approach, where the model first delivers good accuracy, then optimizes the model for cost efficiency which decreases the accuracy. The next section will investigate whether this improvement in costs is worth the decrease in accuracy.

Figure 4 shows the confusion matrices for these models, which help clarify the differences in the prediction behavior between the models. Note that these confusion matrices are the averages of all 100 iterations of the confusion matrices. A notable observation from the confusion matrices is that they clearly show the reason why EV loss is under-performing in accuracy compared to the CCE variants. EV loss heavily favors predicting class 2, while CCE most often predicts class 0, as this is the most common class. The reason EV loss predominantly predicts class 2 is likely because it is the class with the lowest expected cost when guessing a class in the absence of any data. Since EV loss optimizes for cost, it makes sense that it would default to guessing on class 2 in the absence of sufficient data, whereas CCE would default to guessing on the most common class to maximize accuracy.

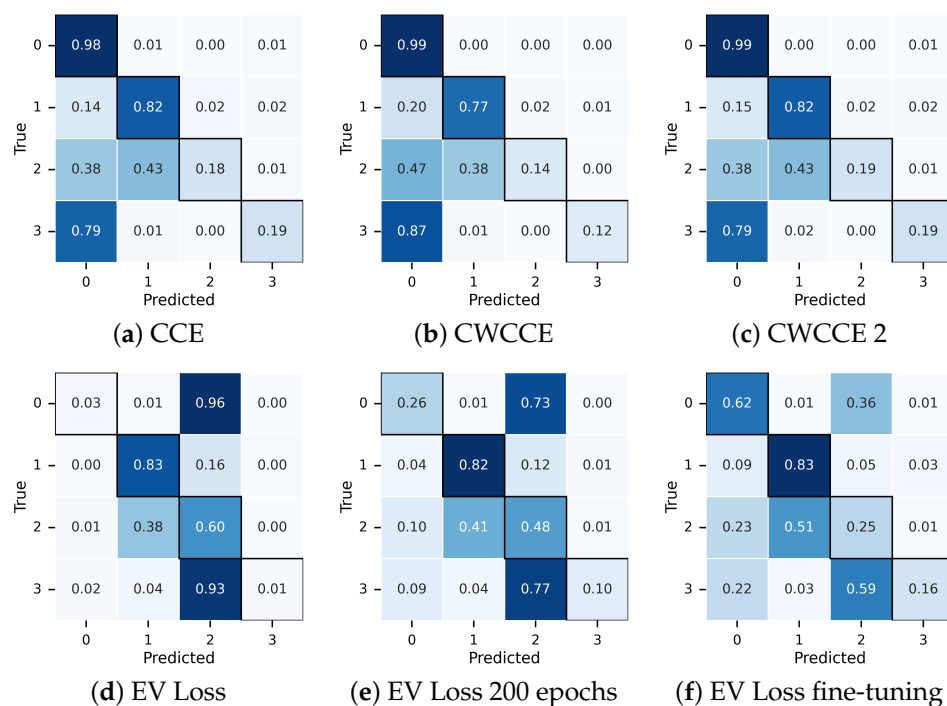


Figure 4. Confusion matrices for each loss function. Numbers are averaged over 100 training iterations and normalized by rows.

4.2. Cost Efficiency

Regarding the cost/benefit of each model, training with the cost-sensitive loss functions delivers a significant benefit, as presented in Figure 5. Training on loss functions CCE, CWCCE, and CWCCE 2 performs the worst. Although these models are highly accurate, the loss functions ignore costs and thus only reduce average downtime to 188, 155, and 189 min, respectively. EV Loss, although performing significantly worse in accuracy than the CCE variants, has a significantly lower average cost because the most costly predictions are the most impactful. Training for 200 epochs slightly improves the impact, although with higher variance between the training iterations. Lastly, the EV Loss w. fine-tuning model was the best-performing model, as it seems to mix the best of both approaches. Thus, EV loss variants reduce downtime by an average of 245, 266, and 296 min, respectively. This means an average improvement in downtime of 30%, 41%, or 57% compared to CCE for the EV loss, EV loss 200 epochs, and EV loss finetuning models, respectively.

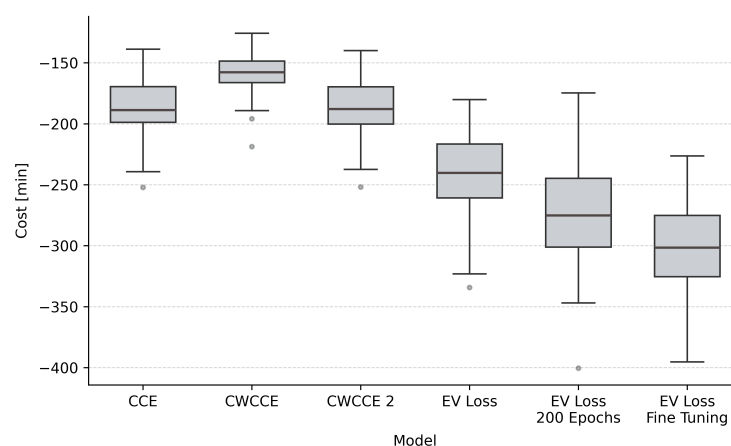


Figure 5. Boxplot of average downtime prevented (negative costs) in minutes.

Although the above results seem to indicate that EV loss functions create impactful models, the variation in model performance must be noted. From the boxplot whiskers, it is evident that over the many iterations, some of the models created are not efficient. The best performing CCE model even outperforms the worst EV loss model. This could indicate some fragility in the training process when using EV loss training, which should be considered for future use of this approach.

5. Discussion

The selection of suitable and rational weights is a key task in the process of training well-performing cost-aware models. Although EV loss simplifies this task compared to using CWCCE, the weight formulations used in this paper have room for improvements, and we cannot guarantee that there are more suitable weights that create results in favor of CWCCE compared to EV loss. It is highly probable that the advantages of EV loss are systematic and will transfer well with better weight representation, but more work and testing with more sophisticated and suitable weights are needed to ensure that the results presented in this paper are robust; however, it seems that for now, EV loss is the most sensible approach due to the simplified approach and better cost efficiency results.

As seen in the results, there was a substantial difference in the accuracy of the EV loss model with 50 epochs and the EV loss model with 200 epochs, which suggests that the EV loss model was not trained to exhaustion at 50 epochs. We suspect this is caused by the EV loss function having a slower learning rate due to its linear mapping. Other loss functions, such as CCE, use a non-linear transform to improve the learning rate of the model. For future work, it would be of interest to test various transforms of the predicted probabilities in the EV loss function to determine whether it is possible to improve the learning rate without losing the general properties of the EV loss function.

Further work on the creation of weights should take the direction of making sure that the weights are more realistic and better represent the actual costs and benefits of the maintenance performed in relation to the alarms predicted. The whole objective of creating cost-aware for predictive maintenance is to better control for the business impact following maintenance actions, however, the weights created for this paper merely represent the costs of the failure events as the model approach is the main focus. To make this approach more industrially viable, an in-depth analysis of maintenance actions and their failure mitigation potential is required to better represent this aspect of predictive maintenance. This will create a cost-aware model that more closely reflects the potential business impact.

Another consideration for improving the representation of the business impact is to examine non-linear mappings of the predictions scores used in EV loss and to calibrate the model prediction scores to more closely resemble probabilities. A transform of the prediction scores is often used in the loss function, like CCE which takes the logarithm of the prediction score. The EV loss function simply multiplies the prediction score with the associated classification cost, but to properly represent the expected costs, this prediction score must resemble probability, which is rarely the case with deep learning models. Such future work in the calibration of prediction scores [40] can help further solidify the expected value of the predictive maintenance performed with these models.

In this paper, we have only briefly tested the performance of the fine-tuning approach, but the results shown are very promising. With more focus on finding the right loss functions for the fine-tuning training process, there might exist variations of this process which significantly improve both the accuracy and cost performance. We see this as the most meaningful improvement to be made on the modeling work and an improvement that can bring great advantages to the predictive maintenance process. With the improved cost analysis discussed above, it is expected that misclassification costs will increase markedly,

as performing unnecessary maintenance not only is costly in itself, but the utilization of maintenance resources in this resource-restricted sector will also have significant opportunity costs. But by improving not only the cost performance, but the general accuracy as well, we believe this will create the most valuable models for predictive maintenance.

A clear next step for this approach is the actual operational implementation, considering the full loop of the maintenance operations and the continuous model training and deployment. Additionally this requires the full landscape of costs to be implemented and correspondingly consider costs related to already planned maintenance tasks, etc. In addition to this point, we also need to add a process of feature engineering and selection. In the current model, we have assumed that the model will identify trend-based features, longer temporal behavior, and physics-informed features on its own via combinations and patterns in the data, but it is expected that these behaviors requires larger amount of data and training (double descent behavior) and correspondingly to enforce these, we could present them via engineered features.

Lastly, all of the results presented in this paper are limited to the wind energy sector. The cost analysis and the data set trained on will vary considerably when this training process is applied to another sector. It is expected that the general dynamics of the presented results will transfer well into other domains, but without testing this cannot be guaranteed. We propose that other researchers with access to large datasets in similar industrial sectors test similar cost-aware and fine-tuning training processes for other datasets to do so. ensure the robustness and generalization of the results presented here.

6. Conclusions

In this paper, we introduce the novel EV Loss loss function for cost-sensitive multi-class classification problems that can directly account for the cost or benefit of each type of classification or misclassification. Using this extensive dataset containing 106 time series features and 12,000 samples of faults in the gearboxes of WTGs, we demonstrated that the EV Loss function was able to achieve better results than existing loss functions, such as CCE or CWCCE, when optimizing for cost. It did come at the cost of accuracy, however. The EV Loss model showed a significant reduction in accuracy compared to CCE and CWCCE. However, by combining EV Loss and CCE in a fine tuning model, we were able to achieve great results in terms of both accuracy and cost-sensitive performance. To validate the consistency of the results, we repeated the training and testing 100 times with random train–test–validation sets. Overall, the ability to directly account for the domain-specific costs in the EV Loss functions makes it nearly trivial to implement compared to CWCCE, which requires careful weight selection and can significantly improve cost-sensitive multi-class classification.

Author Contributions: Conceptualization, M.L.L., R.D.F., P.N. and A.E.V.; methodology, M.L.L.; software, M.L.L.; validation, M.L.L.; formal analysis, M.L.L.; investigation, M.L.L.; resources, P.N. and R.D.F.; data curation, R.D.F.; writing—original draft preparation, M.L.L.; writing—review and editing, M.L.L., R.D.F. and A.E.V.; visualization, M.L.L., R.D.F. and A.E.V.; supervision, R.D.F., A.E.V. and P.N.; project administration, A.E.V. and P.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Restrictions apply to the availability of these data, as it was acquired under a strict confidentiality agreement and cannot be shared further.

Conflicts of Interest: Author Rasmus Døvnborg Frederiksen was employed by the company Siemens Gamesa Renewable Energy during the work on this paper as part of an industrial PhD position. No direct funding was received for the creation of this work. The remaining authors declare that the

research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GW	Gigawatt
WTG	Wind turbine generator
O&M	Operations and maintenance
SCADA	Supervisory control and data acquisition
EV	Expected loss
CCE	Categorical cross-entropy
CWCCE	Class-weighted categorical cross-entropy
LSTM	Long short-term Memory

References

1. WindEurope. *Wind Energy in Europe: 2023 Statistics and the Outlook for 2024–2030*; WindEurope Intelligence Platform: Brussels, Belgium, 2024.
2. Gill, S.; Stephen, B.; Galloway, S. Wind turbine condition assessment through power curve copula modeling. *IEEE Trans. Sustain. Energy* **2012**, *3*, 94–101. [\[CrossRef\]](#)
3. Tautz-Weinert, J.; Watson, S.J. Using SCADA data for wind turbine condition monitoring—A review. *IET Renew. Power Gener.* **2017**, *11*, 382–394. [\[CrossRef\]](#)
4. Abhishek, K.; Abdelaziz, M. *Machine Learning for Imbalanced Data: Tackle Imbalanced Datasets Using Machine Learning and Deep Learning Techniques*; Packt Publishing Ltd.: Birmingham, UK, 2023; p. 344.
5. Elkan, C. The Foundations of Cost-Sensitive Learning. In *International Joint Conference on Artificial Intelligence*; Lawrence Erlbaum Associates Ltd.: Mahwah, NJ, USA, 2001.
6. Fernández, A.; García, S.; Galar, M.; Prati, R.C.; Krawczyk, B.; Herrera, F. Cost-Sensitive Learning. In *Learning from Imbalanced Data Sets*; Springer: Cham, Switzerland, 2018; pp. 63–78. [\[CrossRef\]](#)
7. Kukar, M.; Kononenko, I. Cost-Sensitive Learning with Neural Networks. In *Proceedings of the 13th European Conference on Artificial Intelligence (ECAI-98)*, Brighton, UK, 23–28 August 1998.
8. Liu, B.; Li, X.; Zheng, G. Coastal Inundation Mapping From Bitemporal and Dual-Polarization SAR Imagery Based on Deep Convolutional Neural Networks. *J. Geophys. Res. Ocean.* **2019**, *124*, 9101–9113. [\[CrossRef\]](#)
9. Ren, Y.; Li, X.; Xu, H. A Deep Learning Model to Extract Ship Size from Sentinel-1 SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5203414. [\[CrossRef\]](#)
10. Kodakkal, A.; Meethal, R.E.; Obst, B.; Wüchner, R. A finite element method-informed neural network for uncertainty quantification. In *Proceedings of the World Congress in Computational Mechanics and ECCOMAS Congress, Virtual*, 25–29 January 2021; Volume 800. [\[CrossRef\]](#)
11. McGowan, E.; Gawade, V.; Guo, W. A Physics-Informed Convolutional Neural Network with Custom Loss Functions for Porosity Prediction in Laser Metal Deposition. *Sensors* **2022**, *22*, 494. [\[CrossRef\]](#)
12. Barton, S.; Alakkari, S.; O’dwyer, K.; Ward, T.; Hennelly, B. Convolution network with custom loss function for the denoising of low SNR raman spectra. *Sensors* **2021**, *21*, 4623. [\[CrossRef\]](#)
13. de N Santos, F.; D’Antuono, P.; Robbelein, K.; Noppe, N.; Weijtjens, W.; Devriendt, C. Long-term fatigue estimation on offshore wind turbines interface loads through loss function physics-guided learning of neural networks. *Renew. Energy* **2023**, *205*, 461–474. [\[CrossRef\]](#)
14. Pham, T.C.; Luong, C.M.; Hoang, V.D.; Doucet, A. AI outperformed every dermatologist in dermoscopic melanoma diagnosis, using an optimized deep-CNN architecture with custom mini-batch logic and loss function. *Sci. Rep.* **2021**, *11*, 17485. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Boldini, D.; Friedrich, L.; Kuhn, D.; Sieber, S.A. Tuning gradient boosting for imbalanced bioassay modelling with custom loss functions. *J. Cheminform.* **2022**, *14*, 80. [\[CrossRef\]](#)
16. Yang, F.; Wang, H.Z.; Mi, H.; Lin, C.D.; Cai, W.W. Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. *BMC Bioinform.* **2009**, *10*, S22. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Freitas, A.; Costa-Pereira, A.; Brazdil, P. Cost-Sensitive Decision Trees Applied to Medical Data. In *Data Warehousing and Knowledge Discovery*; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin/Heidelberg, Germany, 2007; Volume 4654, pp. 303–312. [\[CrossRef\]](#)

18. Hillier, F.S.; Lieberman, G.J. *Introduction to Operations Research*; McGraw Hill Inc.: New York, NY, USA, 1995.
19. Petrides, G.; Moldovan, D.; Coenen, L.; Guns, T.; Verbeke, W. Cost-sensitive learning for profit-driven credit scoring. *J. Oper. Res. Soc.* **2022**, *73*, 338–350. [\[CrossRef\]](#)
20. Zhang, Y.; Zhao, P.; Li, B.; Wu, Q.; Huang, J.; Tan, M. Cost-Sensitive Portfolio Selection via Deep Reinforcement Learning. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 236–248. [\[CrossRef\]](#)
21. Dessain, J. Improving the Prediction of Asset Returns With Machine Learning by Using a Custom Loss Function. *Adv. Artif. Intell. Mach. Learn.* **2023**, *3*, 93. [\[CrossRef\]](#)
22. Fox, H.; Pillai, A.C.; Friedrich, D.; Collu, M.; Dawood, T.; Johanning, L. A Review of Predictive and Prescriptive Offshore Wind Farm Operation and Maintenance. *Energies* **2022**, *15*, 504. [\[CrossRef\]](#)
23. Joshua, A.; Sugumaran, V. Fault diagnostic methods for wind turbine: A review. *ARPJ. Eng. Appl. Sci.* **2016**, *11*, 4654–4668.
24. Barbosa, N.B.; Nunes, D.D.G.; Santos, A.Á.B.; Machado, B.A.S. Technological Advances on Fault Diagnosis in Wind Turbines: A Patent Analysis. *Appl. Sci.* **2023**, *13*, 1721. [\[CrossRef\]](#)
25. Kusiak, A.; Li, W. The prediction and diagnosis of wind turbine faults. *Renew. Energy* **2011**, *36*, 16–23. [\[CrossRef\]](#)
26. Zhu, Y.; Zhu, C.; Tan, J.; Tan, Y.; Rao, L. Anomaly detection and condition monitoring of wind turbine gearbox based on LSTM-FS and transfer learning. *Renew. Energy* **2022**, *189*, 90–103. [\[CrossRef\]](#)
27. Wang, J.; Loparo, K.A. Wind Turbine Gearbox Fault Detection Based on Sparse Filtering and Graph Neural Networks. *arXiv* **2023**, arXiv:2303.03496. [\[CrossRef\]](#)
28. Fu, J.; Chu, J.; Guo, P.; Chen, Z. Condition Monitoring of Wind Turbine Gearbox Bearing Based on Deep Learning Model. *IEEE Access* **2019**, *7*, 57078. [\[CrossRef\]](#)
29. Zhang, L.; Fan, Q.; Lin, J.; Zhang, Z.; Yan, X.; Li, C. A nearly end-to-end deep learning approach to fault diagnosis of wind turbine gearboxes under nonstationary conditions. *Eng. Appl. Artif. Intell.* **2023**, *119*, 105735. [\[CrossRef\]](#)
30. Jamil, F.; Verstraeten, T.; Nowé, A.; Peeters, C.; Helsen, J. A deep boosted transfer learning method for wind turbine gearbox fault detection. *Renew. Energy* **2022**, *197*, 331–341. [\[CrossRef\]](#)
31. Xiao, C.; Liu, Z.; Zhang, T.; Zhang, X. Deep Learning Method for Fault Detection of Wind Turbine Converter. *Appl. Sci.* **2021**, *11*, 1280. [\[CrossRef\]](#)
32. Wang, X.; Zheng, Z.; Jiang, G.; He, Q.; Xie, P. Detecting Wind Turbine Blade Icing with a Multiscale Long Short-Term Memory Network. *Energies* **2022**, *15*, 2864. [\[CrossRef\]](#)
33. Butte, S.; Prashanth, A.R.; Patil, S. Machine Learning Based Predictive Maintenance Strategy: A Super Learning Approach with Deep Neural Networks. In Proceedings of the 2018 IEEE Workshop on Microelectronics and Electron Devices, WMED 2018, Boise, ID, USA, 13 April 2018. [\[CrossRef\]](#)
34. Wang, J.; Zhao, X.; Guo, X. Optimizing wind turbine's maintenance policies under performance-based contract. *Renew. Energy* **2019**, *135*, 626–634. [\[CrossRef\]](#)
35. Schuh, P.; Schneider, D.; Funke, L.; Tracht, K. Cost-optimal spare parts inventory planning for wind energy systems. *Logist. Res.* **2015**, *8*, 4. [\[CrossRef\]](#)
36. Satymov, R.; Bogdanov, D.; Breyer, C. Global-local analysis of cost-optimal onshore wind turbine configurations considering wind classes and hub heights. *Energy* **2022**, *256*, 124629. [\[CrossRef\]](#)
37. Tang, M.; Zhao, Q.; Wu, H.; Wang, Z. Cost-Sensitive LightGBM-Based Online Fault Detection Method for Wind Turbine Gearboxes. *Front. Energy Res.* **2021**, *9*, 701574. [\[CrossRef\]](#)
38. Tang, M.; Chen, Y.; Wu, H.; Zhao, Q.; Long, W.; Sheng, V.S.; Yi, J. Cost-Sensitive Extremely Randomized Trees Algorithm for Online Fault Detection of Wind Turbine Generators. *Front. Energy Res.* **2021**, *9*, 686616. [\[CrossRef\]](#)
39. Tang, M.; Ding, S.X.; Yang, C.; Cheng, F.; Shardt, Y.A.; Long, W.; Liu, D. Cost-sensitive large margin distribution machine for fault detection of wind turbines. *Clust. Comput.* **2019**, *22*, 7525–7537. [\[CrossRef\]](#)
40. Zadrozny, B.; Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, 23–26 July 2002; pp. 694–699. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.