machines

MDPI

Article

Unveiling the Black Box: A Unified XAI Framework for Signal-Based Deep Learning Models

Ardeshir Shojaeinasab [1], Masoud Jalayer [2,3], Amirali Baniasadi [1] and Homayoun Najjaran [1,2,*]

[1] Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC V8P 5C2, Canada; ardeshir@uvic.ca (A.S.); amiralib@uvic.ca (A.B.)
[2] Department of Mechanical Engineering, University of Victoria, Victoria, BC V8P 5C2, Canada; masoud.jalayer@polimi.it
[3] Department of Management, Economics and Industrial Engineering, Politecnico di Milano, 20156 Milan, Italy
* Correspondence: najjaran@uvic.ca

Abstract: Condition monitoring (CM) is essential for maintaining operational reliability and safety in complex machinery, particularly in robotic systems. Despite the potential of deep learning (DL) in CM, its 'black box' nature restricts its broader adoption, especially in mission-critical applications. Addressing this challenge, our research introduces a robust, four-phase framework explicitly designed for DL-based CM in robotic systems. (1) Feature extraction utilizes advanced Fourier and wavelet transformations to enhance both the model's accuracy and explainability. (2) Fault diagnosis employs a specialized Convolutional Long Short-Term Memory (CLSTM) model, trained on the features to classify signals effectively. (3) Model refinement uses SHAP (SHapley Additive exPlanation) values for pruning nonessential features, thereby simplifying the model and reducing data dimensionality. (4) CM interpretation develops a system offering insightful explanations of the model's decision-making process for operators. This framework is rigorously evaluated against five existing fault diagnosis architectures, utilizing two distinct datasets: one involving torque measurements from a robotic arm for safety assessment and another capturing vibration signals from an electric motor with multiple fault types. The results affirm our framework's superior optimization, reduced training and inference times, and effectiveness in transparently visualizing fault patterns.

Keywords: condition monitoring; explainable artificial intelligence; trustworthy artificial intelligence; feature engineering; signal processing

Citation: Shojaeinasab, A.; Jalayer, M.; Baniasadi, A.; Najjaran, H. Unveiling the Black Box: A Unified XAI Framework for Signal-Based Deep Learning Models. *Machines* **2024**, *12*, 121. https://doi.org/10.3390/machines12020121

Academic Editor: Xiang Li

Received: 18 December 2023
Revised: 22 January 2024
Accepted: 25 January 2024
Published: 8 February 2024

1. Introduction

With the evolution from Industry 4.0 to 5.0, deep learning (DL) techniques have become even more integral, especially in the domain of condition monitoring (CM) for machinery and robots [1]. Considering the complexity of modern robotic mechanisms and their diverse roles, leveraging DL to predict and proactively manage potential malfunctions or ensure their safe movements is a game changer. However, as we leverage these sophisticated DL algorithms, a fundamental challenge emerges: trustworthiness. The inherently opaque nature of DL models, which are often termed 'black boxes', presents a hurdle in achieving transparency and reliability in their predictions, especially in critical domains like CM where the stakes are high [2]. As industries globally lean heavily on robots and machinery for production, ensuring the safe and consistent operation of these assets is paramount. A malfunction, if not pre-emptively addressed, can lead to substantial economic losses, compromised safety, and operational disruptions. Therefore, this paper endeavors to delve into the nuances of creating a DL-based CM system for robots and machinery that is not only efficient and predictive, but also transparent and trustworthy, bridging the gap between advanced AI capabilities and the indispensable human trust factor.

The historical trajectory of machine CM can be traced back to its rudimentary beginnings with simple circuit breakers designed to manage incoming power feeds. This was succeeded by the advent of multifunctional electronic relays, marking a pivotal evolution in monitoring capabilities. The introduction of Programmable Logic Controllers (PLCs) further revolutionized the field, enabling engineers to devise intricate protection mechanisms for machines. In today's context, buoyed by technological advancements that have democratized data acquisition costs and amplified computational prowess, a palpable shift in the industrial sector is taking place. Many plants are now fervently moving towards integrating data-driven systems into their machinery and robotic CM processes [3].

In recent years, the expanding domain of CM, especially within machinery and robotics, has greatly benefited from the advent and refinement of AI frameworks. Tailored to address intricate CM challenges, these AI-driven methodologies cast a wide net, encompassing a plethora of sensory data inputs. They proficiently integrate ultrasonic measurements [4], electrical currents [5], torque measurements on robotic arm joints [6], temperature profiles [7], and state-of-the-art techniques for wear debris detection [8]. Vibration and torque signal analyses, paramount aspects of these frameworks, have been harnessed extensively for both machinery and robotic systems, demonstrating their pivotal role and potential [9,10]. This confluence of diverse measurements fosters a more holistic understanding of equipment and robotic health, propelling advancements in predictive accuracy and operational longevity. L. Yang et al. propose the Twin Broad Learning System (TBLS) as a solution to the challenges faced by deep learning models that rely on extensive datasets for fault diagnosis in rotating machinery. The TBLS incorporates two non-parallel hyper-planes, which results in an improved generalization capability and diagnostic accuracy for overlapping fault patterns. Experimental results on benchmark datasets support the effectiveness and efficiency of the TBLS [11]. In another interesting study, Yuqing Zhou et al. introduce a semi-supervised fault diagnosis approach for rotating machinery, combining multi-scale permutation entropy with contrastive learning. This method significantly outperforms the benchmarks in gearbox and milling tool diagnosis, achieving high accuracy with limited labeled data [12].

While DL methodologies have demonstrated significant potential in enhancing CM systems, their 'black-box' character—being non-intuitive and opaque—poses challenges. This intrinsic nature often leaves human operators and management in a quandary, unable to decipher the underlying rationale behind the model's decisions. This intricacy frequently leaves stakeholders, spanning from on-ground operators to upper-tier management, grappling with the logic underlying the decisions proffered by these models. Recent scholarly discourse, as highlighted by Antwarg et al. [13], underscores this opacity as a notable impediment, potentially stymieing the trust and subsequent adoption of DL-driven CM systems among industry practitioners. Concurrently, the academic landscape has been enriched with a myriad of feature extraction methodologies, specifically tailored to signal-based CM paradigms. Yet, finding an optimal subset of features remains an academically contested area. This challenge is exacerbated by the increasing data dimensionality, which, in turn, amplifies the computational demands and intricacy of the accompanying DL architectures. It is against this backdrop that this paper introduces a novel framework, striving to surmount the aforementioned challenges. The novelty of our approach lies in its multifaceted capabilities:

(1) **Transparency in Decision Making**: While a plethora of publications have focused on the efficacy and accuracy of such models, only a handful of studies have attempted to unravel the intricacies of their decision-making processes. There have been concerted efforts to leverage more transparent models, such as decision trees, to enhance interpretability. Notably, ref. [14] showcased the potential of shallow-learning ensemble models, integrating XGBoost and Random Forests with model explainers to elucidate the underlying decision logic. On a similar tangent, the authors of [15] employed the Logical Analysis of Data as a pathway to achieve an interpretable machine learning technique that is specifically tailored to fault detection and diagnosis within intricate industrial chemical processes.

From a different perspective, to gauge the model's confidence level in its decisions, ref. [16] employed Bayesian models, particularly for monitoring machine signals in anomalous or unpredictable domains where misdiagnosis might occur in the absence of discernible symptoms. In their efforts to comprehensively model a robot's vibrational attributes, ref. [17] endeavored to quantify the uncertainty linked with eigenfrequency prediction, leveraging the precision of Monte Carlo uncertainty propagation.

Echoing a similar sentiment regarding the importance of confidence in outcomes, ref. [18] innovatively integrated Bayesian variational learning into Transformer architectures. Their approach instilled uncertainty into attention weights, paving the way for a probabilistic Bayesian Transformer specifically designed for dependable CM in rotating machines.

Our proposed framework therefore strives to surmount existing limitations. It tries to elucidate the specific patterns and signatures instrumental in the decision-making processes, particularly emphasizing their contributions to or against each classification category. Presented in a visually intuitive manner, this elucidation aims to render the decision-making process transparent and cogent for human operators and stakeholders. This explicability acts as a catalyst, engendering trust and confidence in AI-based CM models.

(2) **Optimal Feature Selection and Model Efficiency**: Utilizing specialized processing methods for sensory signals unveils informative signatures and patterns that are crucial to CM. In recent decades, a myriad of feature types have been proposed and implemented in the CM discourse. Notably, features like Kurtosis, Root Variance Frequency, Max Power Spectrum, Impulse Factor, and Crest Factor have been extensively employed as fundamental statistical attributes within machine-learning-based fault detection frameworks, providing pivotal insights into machinery states [19,20].

Moreover, the advent of sophisticated time–frequency analyses such as permutation entropy, multiscale permutation entropy, and multiscale entropy has been integral in reinforcing the analytical capabilities of both machine learning and DL-based CM frameworks [8,21,22]. Empirical Mode Decomposition (EMD) stands out as a notably effective, self-adaptive processing method that is adept at analyzing non-linear and non-stationary processes, despite its inherent challenges; these include mode mixing, end effects, interpolation problems, and complexity in selecting the optimal Intrinsic Mode Function (IMF) [23,24].

In parallel, an assortment of CM frameworks have incorporated grayscale diagrams and diagrams based on Fourier and wavelet transforms, each illuminating different facets of the machine's state, with their respective merits and demerits. Investigations into the reliability and efficacy of discrete Fourier transform (DFT), short-time Fourier transform (STFT), and continuous wavelet transform (CWT) have also been pivotal in defining rub detection parameters based on vibration signals [25]. In a concerted effort to amalgamate the benefits of diverse techniques, recent studies have explored the integrative application of these feature extraction techniques [26–28]. However, the field still lacks a unified consensus on the delineation of optimal feature engineering for machinery CM.

In light of the above considerations, our framework strives to amalgamate advanced signal processing techniques, focusing on harnessing the complementary strengths of the aforementioned methodologies to refine feature selection; this enhances model accuracy and reduces computational demands, leading to an expeditious and highly efficient CM framework that is suitable for machinery and robots.

(3) **Leveraging State-of-the-Art DL Architecture**: At the core of our framework lies the integration of the Convolutional Long Short-Term Memory (CLSTM) architecture, which has gained recognition as a leading approach in the field of deep learning. CLSTM exhibits exceptional capabilities in handling sequential data, making it highly suitable for tasks in condition monitoring (CM) [9,29].

Our research aims to bridge existing gaps in the CM literature by combining advanced signal processing techniques, state-of-the-art DL architectures, and a strong commitment to transparency and efficiency. Through this study, we seek to strengthen the foundations of a new era in machinery and robotic system diagnostics, characterized by both insightful

analysis and intuitive understanding. An earlier version of this work was presented at the 32nd International Conference on Flexible Automation and Intelligent Manufacturing in 2023 [30].

The rest of this article is organized as follows: An overview of Shapley values and XAI is provided in Section 2. Our framework is discussed in Section 3. Experiments and discussions are presented in Section 3. Finally, the conclusion is reached in Section 4.

2. Materials and Methods

2.1. Shapley Values

In straightforward models, the model itself stands as the optimal illustration; it embodies a clear and direct understanding. However, this is not applicable to more complex models like ensemble methods or deep networks. The inherent complexity of these models makes it difficult to use the original model as a direct explanation. Here, a better approach is to adopt a simpler, more understandable approximation of the original model as an explanatory tool [31]. This section delves into the progression from traditional Shapley Values to SHapley Additive exPlanation (SHAP), aiding in unraveling the intricacies of more complex models through a structured three-step approach that is noted in the classic Shapley values assessment [32].

In the initial phase, *Shapley regression values* function as markers of feature importance in linear models, especially when features are interdependent. This process necessitates retraining the model across all possible combinations of feature subsets, denoted as $S \subseteq F$, where F signifies the entire set of features. Each feature is assigned a value representing its influence on the predictions of the model. This is determined by contrasting predictions that are generated by two versions of the model: one that includes the feature (f_{SUi}) and another that does not (f_S). The discrepancy between predictions is gauged using the formula $f_{SUi}(x_{SUi}) - f_S(x_S)$, where x_S indicates the input features within set S. Given that the exclusion of a feature can affect others, it is crucial to assess these differences across all plausible subsets $S \subseteq F$. Shapley values then serve as indicators of feature contributions, encapsulating the average potential shifts, as displayed in Equation (1):

$$\phi_i = \sum_{S \subseteq F_i} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{SUi}(x_{SUi}) - f_S(x_S)], \quad (1)$$

In this scenario, h_x maps a value of either 1 or 0 back to the original input space, indicating whether an input is included or excluded in the model, respectively.

Next, *Shapley sampling values* aim to approximate the effects of removing a variable from the model by analyzing samples from the training dataset. This is achieved by applying sampling approximations to Equation (1), thereby avoiding the retraining of the model and reducing the computations to fewer than $2^{|F|}$. Although this method modifies the approach, the explanation model structure aligns with that of Shapley regression values, preserving its additive feature attribution property.

Lastly, a broader strategy, known as *quantitative input influence*, is introduced. While it extends beyond mere feature attributions, it includes a sampling approximation to Shapley values, similar to Shapley sampling values. This positions it as another additive feature attribution approach.

2.2. SHapley Additive exPlanation Values

To facilitate the practical application of traditional Shapley values to deep networks in an efficient timeframe, Lundberg and his team innovated the SHapley Additive exPlanation (SHAP), a groundbreaking method able to decipher the inner workings of DL models.

Essentially, SHAP values are formulated to delineate the critical role of the various features that influence the model's decision-making process for any specific input sample.

Accurately calculating SHAP values poses a significant challenge. Nonetheless, drawing from the current advancements in additive feature attribution methodologies, reliable approximations are achievable. In their work, Lundberg et al. devised two universal

approximation strategies [32]—the established Shapley sampling values and the pioneering Kernel SHAP. These strategies aim to diminish the computational time associated with SHAP value approximations and are grounded on the principles outlined in the traditional Shapley values. Specifically, Kernel SHAP optimizes the efficiency of sampling when deriving model-independent estimations of SHAP values. Additionally, the research is steered towards optimizing approximation techniques for specific model types, resulting in the creation of Linear SHAP and Low-Order SHAP for linear frameworks and the development of Max SHAP and Deep SHAP to expedite approximation processes under defined conditions.

These innovative techniques promise to yield SHAP value approximations in a manageable period. As illustrated in Figure 1, SHAP values have practical implications, particularly in illustrating the decision-making pathways of a Convolutional Neural Network (CNN) during the binary classification of visually similar animals belonging to diverse classes.

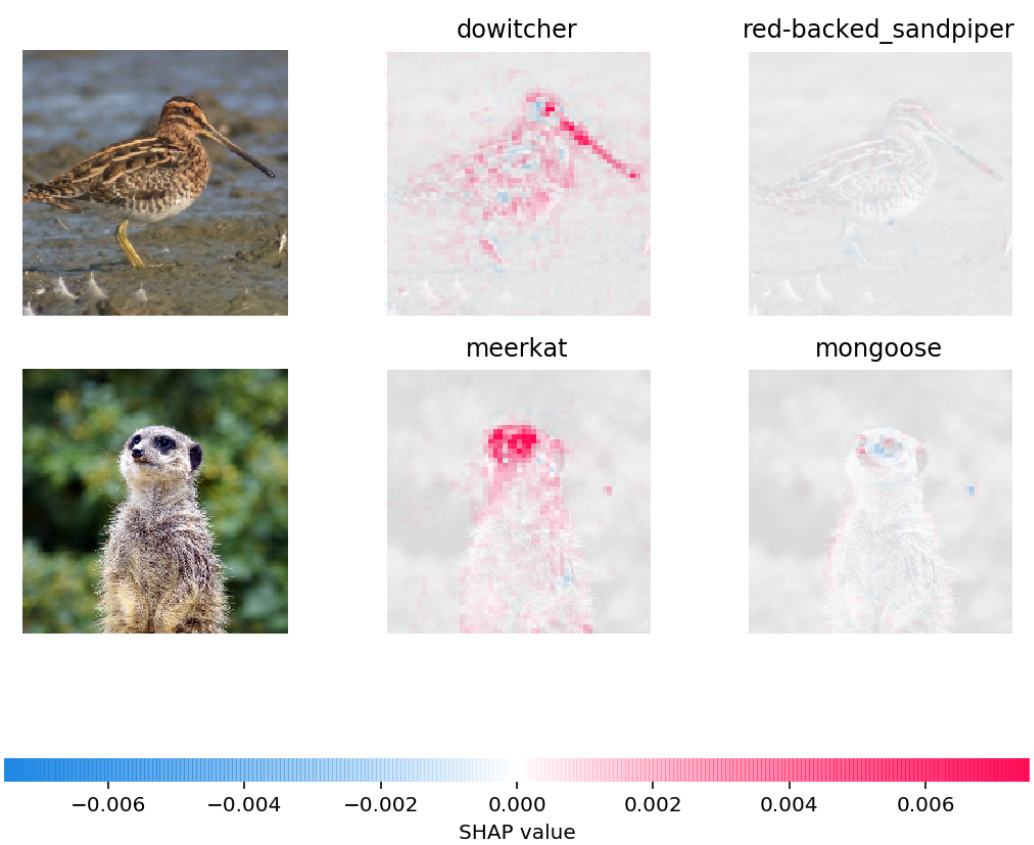

Figure 1. A demonstration of applying SHAP values in a computer vision task.

Figure 1 discerns the pixels within each image that significantly influence the model's classification outcome, either in a favorable or unfavorable manner. In this analysis, red pixels serve as positive influencers, guiding the model towards a specific classification, whereas blue pixels act as deterrents or negative indicators during the evaluation of a particular sample input.

3. The Proposed Model Explainer

In this study, we highlight the importance of efficient bearing fault detection as a cost-saving measure for manufacturers. To earn the trust of industry practitioners in using the DL model for detection tasks, we introduce a method that will help enhance reliability and ease of use. Figure 2 outlines the steps involved in our proposed framework for fault diagnosis using eXplainable Artificial Intelligence (XAI); these include the following:

1. **Feature Extraction**: In this initial step, various signal processing techniques, including time–frequency domains and Fourier and wavelet analyses, are used to uncover hidden patterns. This aims to increase both the accuracy and clarity of the DL model.
2. **Fault Diagnosis and Collision Detection**: This step involves creating and training a dual-directional Convolutional Long Short-Term Memory (CLSTM) structure using the extracted features from the first step.
3. **XAI-Based Model Optimization**: This part of the process utilizes SHAP values to remove less useful features, simplifying the data and making the model easier to implement in real-time settings, thus assisting practitioners in enhancing the model's performance.
4. **Signal-Based CM Interpretation System**: This final step aims to visually present the model's findings, helping operators grasp the diagnostic decisions made by the DL model.

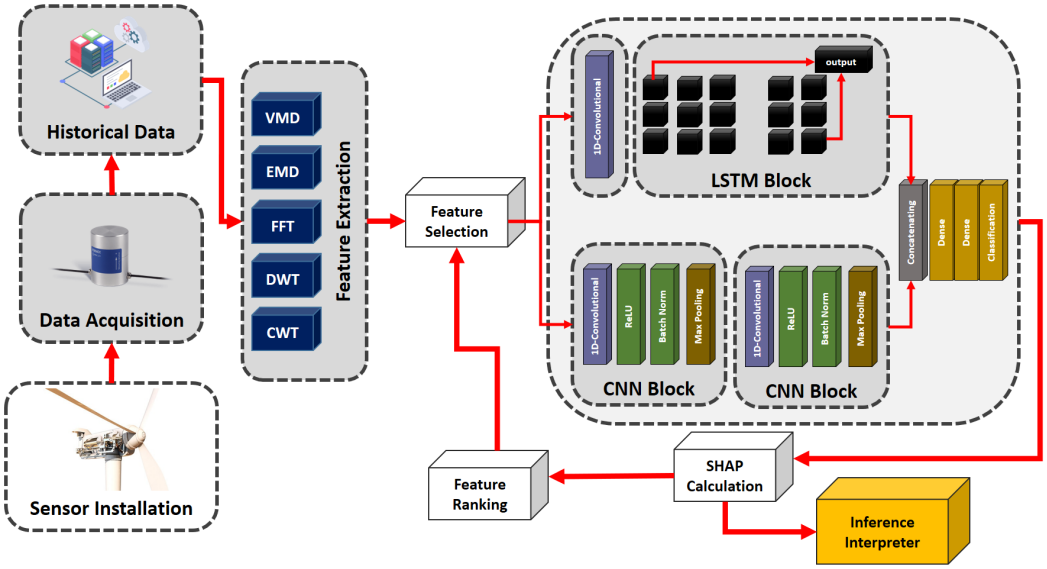

Figure 2. The suggested framework for XAI-based fault diagnosis.

Figure 3 showcases two visualizations proposed in our XAI framework; these were created using the CWRU dataset and focus on two different categories: Centered (@6:00) and Opposite (@12:00) Outer Race faults. These visuals assist operators in clearly identifying which parts of the signal are related to the detected class (indicated by red dots) and why the sample does not align with other classes (marked by blue dots). The operator can acquire insights into the operation of the DL model by examining a small set of varied samples that are explained using the ExAI framework for signals. It is important to note that the SHAP explainer's red and blue dots can only indicate the importance of patterns for the AI model's decision. It is the operator's responsibility to assess whether these patterns hold any significant relevance to the specific class.

In this paper, we use SHAP values, as discussed earlier, to illustrate the functioning of the detection model. To demystify the complex model, we propose a user-friendly visualization technique that enhances our understanding of how the model classifies data. Drawing inspiration from Lundberg's approach, but with a focus on signal data types, we present a method that uses color-coded indicators (red and blue dots) to convey how different parts of the signal influence the final classification decision. Here, red dots signal a higher likelihood of belonging to a certain class, while blue dots indicate a lower likelihood of being classified into a category.

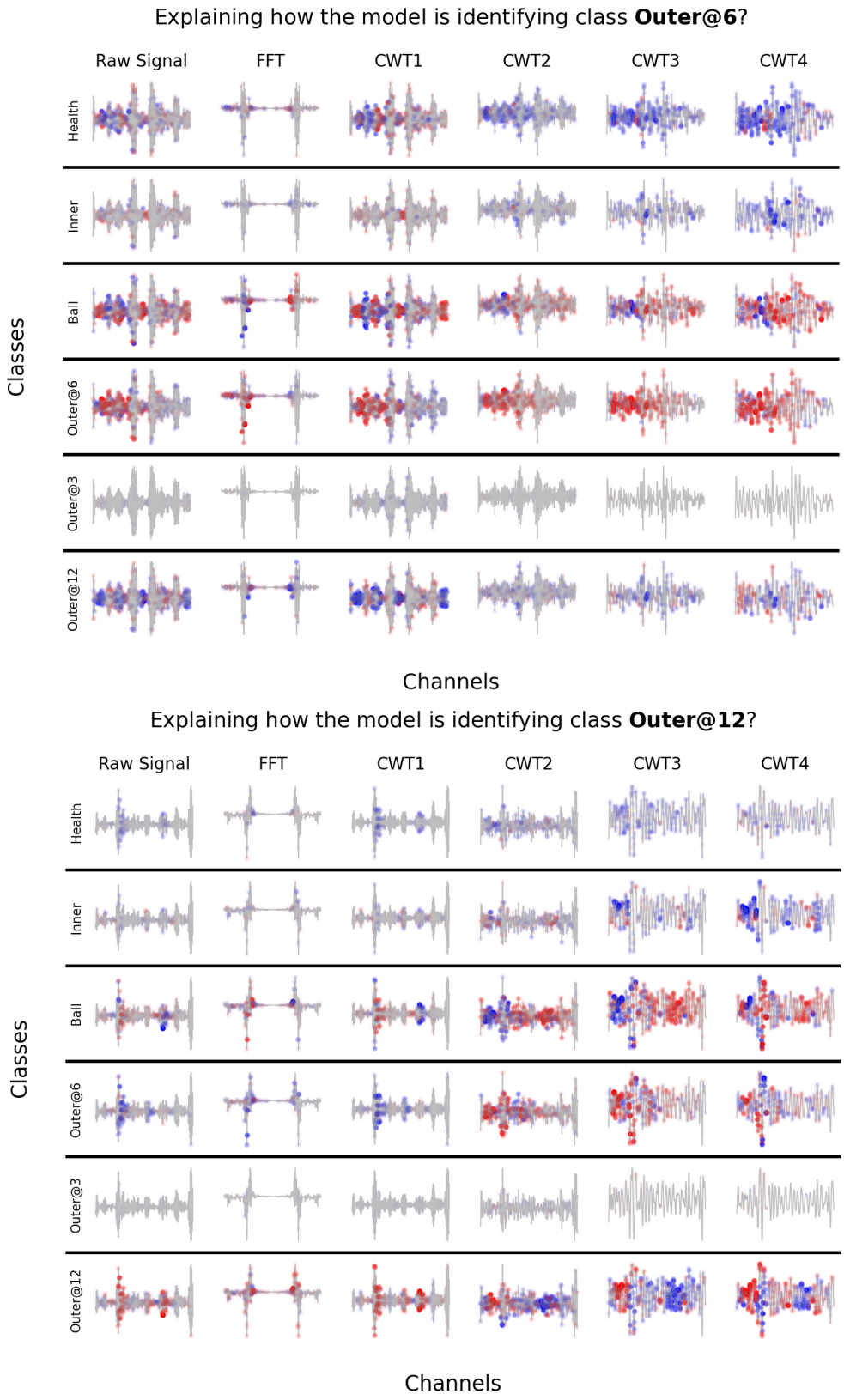

Figure 3. Visualization of SHAP values for fault diagnosis. The upper panel displays the original signal data, while the lower panel shows the synthesized signals generated by the model.

For a hands-on perspective and to facilitate reproducibility, the implementation of the proposed model is publicly available at our GitHub repository https://github.com/Ardeshir-Shon/SignalExplainer (accessed on 20 January 2024).

4. Experimental Results and Discussion

The proposed framework offers two key benefits. Firstly, it reveals the workings behind the DL model, explaining how the model makes its classifications. Secondly, it highlights the important features that significantly contribute to classification tasks and removes the unnecessary features that do not contribute to the output, thus making the model more efficient.

To assess the efficacy of the proposed framework, the authors employ two distinct datasets: one pertaining to fault detection task classification and the other focused on determining the status of robotic arms.

4.1. Dataset Descriptions

When emphasizing the importance of safety in robotic systems, understanding and monitoring their operational behavior becomes a focal point in advanced robotics research. A critical aspect of this is discerning the robot's interaction with its environment, such as detecting collisions with obstacles, understanding when it is being manually operated, or ascertaining when it is moving freely. To this end, [22], from the Technical University of Munich, conducted a comprehensive study where they recorded the torques for each joint of a seven-DoF KUKA arm. These signal instances span a time window of 1024 ms, collected at a 1 kHz sampling rate. Notably, any collisions or contact events are encapsulated around the 256 ms mark within these signal instances. For the purpose of this paper, and to focus on the most relevant data segments, we specifically extracted a 300 ms time window from each sample, ranging from 256 ms to 556 ms, when constructing our training and test sets. This ensured the efficient processing of the datasets, concentrating on the time frames that were most indicative of the robot's interactions.

This study employed the Case Western Reserve University (CWRU) bearing dataset and used the test stand shown in Figure 4, which consists of a motor, a torque transducer (/encoder), a dynamometer, and control electronics. The dataset included a healthy condition and five different fault types, i.e., three different outer race misalignments (outer@3, outer@6, and outer@12) along with the inner race and the ball faults. The faults were grouped according to their severity and range in diameter, from 0.007 inches to 0.040 inches. Additionally, the dataset was composed of four different motor speeds. However, for the sake of simplicity, this study only utilized one motor speed, 1797 RPM. The dataset was collected with a sampling rate of 12 kHz, with an accelerometer mounted on the drive-end of the machine. In the experiments, we took signal bursts of 800 timestamps, equal to 66.6 milliseconds, to generate some different datasets of approximately 25,500 signal bursts.

In this study, we also used a binary dataset, consisting of the torque values of a wind turbine with a 30 kW induction generator that were collected under rotor electrical unbalance (REU) conditions and in healthy conditions at varying loads and fault levels. The data were recorded for three different motor speeds of 1530, 1560, and 1590 rpm. The dataset contained some additional phase resistances equal to 0.099 Ω, 0.1485 Ω, and 0.198 Ω for one rotor phase to obtain 150%, 225%, and 300% REU, respectively. For the fault class, we only used the signals corresponding to 150% REU.

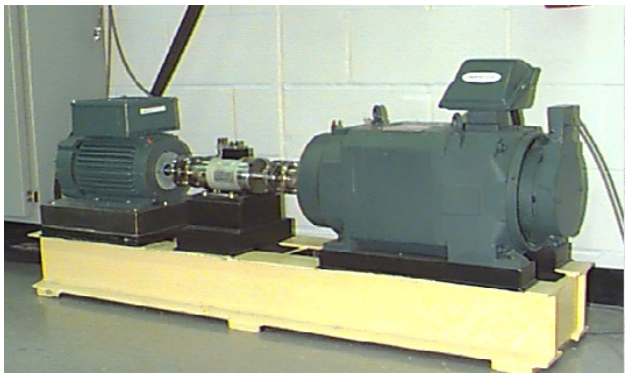

Figure 4. Two horsepower (**left**), a torque transducer and encoder (**center**), and a dynamometer (**right**) used to collect the dataset

4.2. Application I: Model Optimization/Feature Selection

To evaluate how well this framework works in feature selection, this study compared it with several top models: our proposed model (XAI-CLSTM), CLSTM [9], sdAE [33], a dfCNN [34], CAE-fft [35], and a CNN [36].

Tests using the CWRU dataset show that our framework improves diagnostic results across different burst lengths. As seen in Figure 5, for CWRU, XAI-CWRU has an f_1-score up to 7% better than that of other models. This difference is even clearer with shorter burst lengths, where there are fewer data to help figure out the signal and identify it. This speed in diagnosis is crucial when trying to spot machine problems quickly. The wind turbine dataset in Figure 6 shows similar results. However, for longer burst lengths, CLSTM performs a bit better. The other four models consistently scored lower than our proposed model on both datasets. In addition, the arm dataset illustrates that the accuracy remains unaffected when the XAI-CLSTM model is made lighter by employing more efficient feature selection using the proposed framework, as shown in Figure 7.

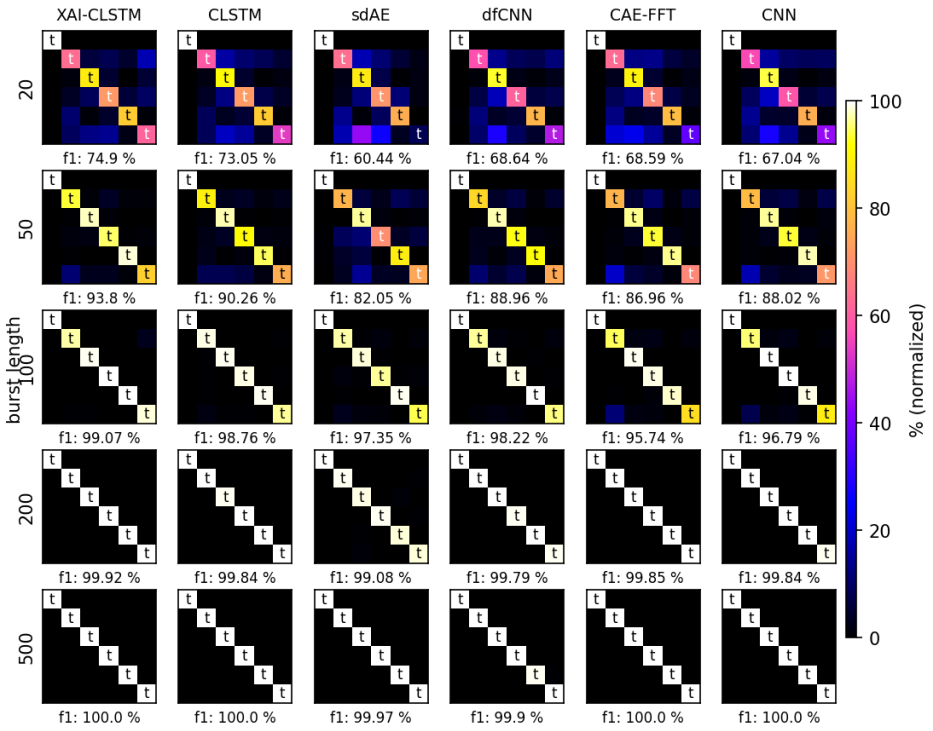

Figure 5. Classification performance comparison on CWRU dataset—t represents the correctly classified samples.

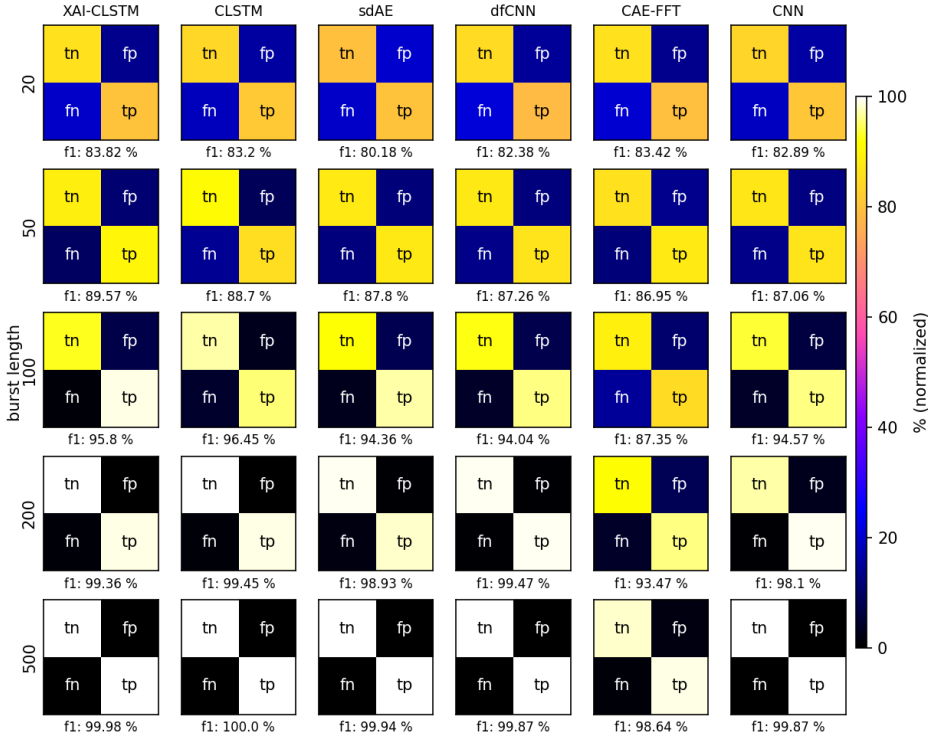

Figure 6. Classification performance comparison on wind turbine dataset—tn, fp, tp, and fn represent true negatives, false positives, true positives, and false negatives, respectively.

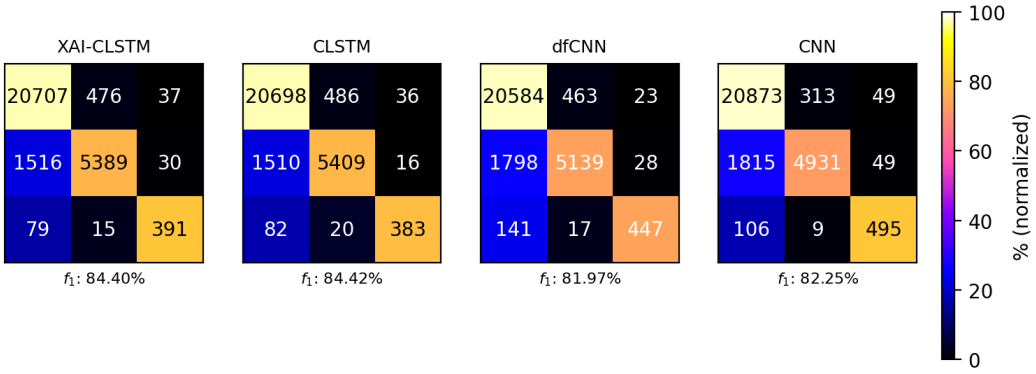

Figure 7. Classification performance comparison on seven-DoF KUKA arm dataset—t represents the correctly classified samples.

4.3. Application II: Evaluating Generative Models

The proposed XAI framework can be effectively applied to assess the synthetic samples made by GAN models. This paper includes a detailed evaluation of the Conditional Generative Adversarial Network (CGAN) introduced by [37]. Their work brings forth a unique data augmentation approach tailored to fault data synthesis.

Ahang et al. presented a CGAN variant designed to train on both regular and faulty data under a single condition. Using this trained network, they generated fault data from standard samples for motor speeds lacking corresponding fault data. For evaluation, we use a CLSTM classifier that is already fine-tuned on the CWRU dataset, which acts as our benchmark.

The evaluation process entails a comparison of the signals produced by the CGAN against ground truth signals. This allows for a detailed examination of common patterns and significant signal inflection points. This rigorous comparison ensures that the CGAN offers good generalization capabilities, mitigating concerns regarding mode collapse. At the same time, it certifies the CGAN's ability to produce diverse, high-quality samples. Such

samples not only align well with ground truth training, but also produce discernible class identifications. This is further confirmed as the SHAP values visualized by our framework distinctly highlight key signal features.

4.4. Application III: Model Explainer and Training

The XAI framework introduced in this manuscript serves another crucial function: elucidating the decision-making processes of models, particularly in discerning classes. This paper introduces a 'signal explainer' that visualizes SHAP values, illuminating the underlying rationale for class assignments.

To provide a tangible demonstration, we utilize the model and dataset presented by Zhang et al. Their research is centered on devising an online collision detection and identification system for robots that collaborate with humans. The architecture they propose consists of a signal classifier and an online diagnoser. This system monitors the robot's sensory signals, promptly detecting and classifying physical human–robot interactions [22].

In Figure 8, we illustrate the integration of Zhang et al.'s model with our XAI framework. By feeding the model with their dataset, we are able to provide deeper insights into how their online classifier discerns between torque signals. Specifically, the classifier segregates signals into three distinct categories: 'free', 'collision', and 'contact'; each of these indicates a different interaction status between the robot and its environment.

This interpretability afforded by the XAI framework empowers practitioners to comprehend essential signal patterns that are indicative of specific classes. As a result, there is enhanced trust in these online CM tools within the robotics industry. Moreover, this understanding paves the way for the more effective training of human operators, ensuring smoother human–robot collaboration.

In the evaluation of the signal eXplainer software using the seven-DoF KUKA dataset, a noteworthy observation emerges from Figure 8. Specifically, the software incorrectly classifies certain data points, leading to false positives. This misclassification is particularly evident in the lower panel of the figure, where the model confuses the 'contact' and 'collision' classes due to their similarity in the feature space. More precisely, torques 3, 5, and 6 suggest a 'collision' event, whereas torques 2 and 4 steer the model toward identifying it as 'contact'. This visualization elucidates the susceptibility of the model to false positives in such ambiguous scenarios.

Furthermore, the dataset poses an inherent limitation on the signal eXplainer's efficacy, particularly concerning the seven-DoF KUKA dataset. The dataset is imbalanced in terms of class distribution; for example, it lacks sufficient data points representing torque signals when the status corresponds to the 'free' class. This paucity of samples in the 'free' class renders the SHAP (Shapley Additive Explanation) values less reliable. Consequently, when the model encounters a data point from the 'free' class, the SHAP values may misrepresent its importance, biasing the model towards other classes. This is an intrinsic drawback that the current version of the model explainer is not equipped to address.

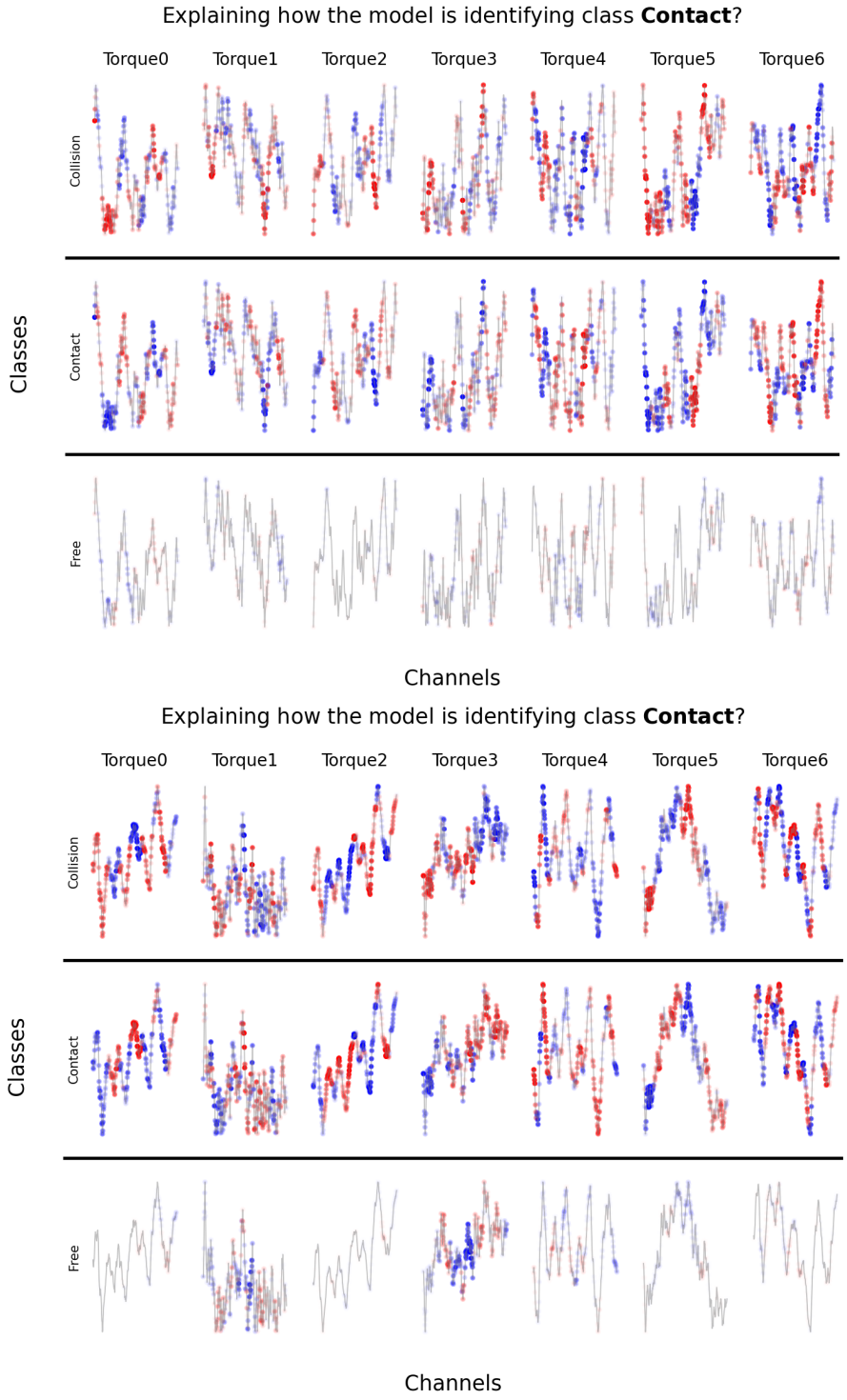

Figure 8. Output of the signal eXplainer software for the robotic arm status identification task. The upper panel displays the correct identification of the model, while the lower panel shows a false positive.

5. Conclusions

The critical importance of machinery and robot CM in ensuring operational efficiency and safety in the industrial sector cannot be overstated. As Industry 4.0 and 5.0 paradigms continue to shape the future of manufacturing and production, there is a pressing need for advanced CM methodologies. Signal-based CM methods, particularly those leveraging DL for fault diagnosis, have surged in prominence over recent years. However, despite

their capabilities, a significant challenge has persisted, namely the 'black-box' nature of such models; this often hinders their adoption in critical scenarios due to the opaque decision-making processes.

In order to fill this important gap, this paper presents an innovative method that aims to improve the performance of DL-based CM models and provide insight into their decision-making processes, thus promoting trust and transparency, as shown in Figure 9. Our proposed framework consists of a systematic four-step process, including (1) a comprehensive feature extraction phase, (2) advanced fault detection using a dual-path ConvLSTM architecture, (3) model optimization and feature refinement based on XAI principles, and (4) a dedicated module for interpreting inferences.

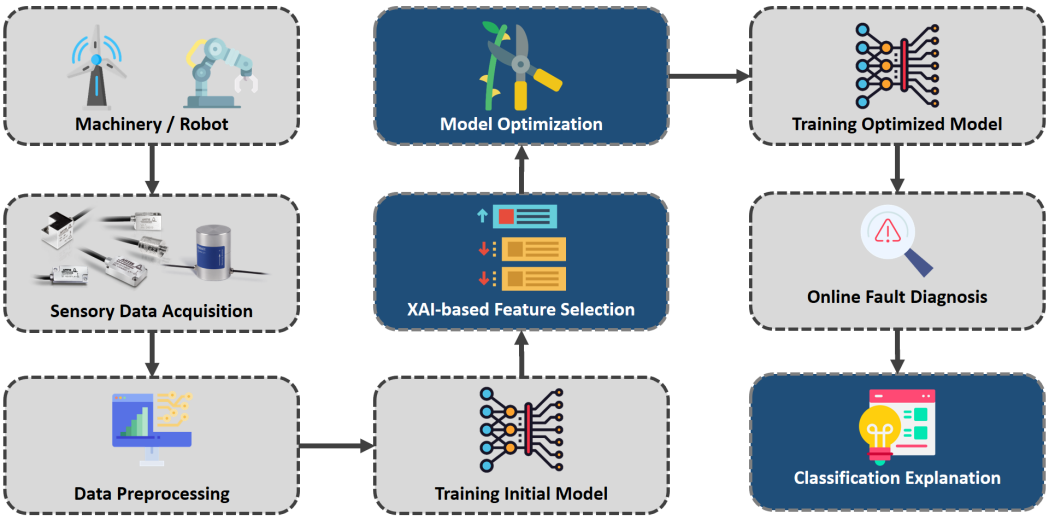

Figure 9. A visual outline of the suggested model.

Drawing from a rich set of signal processing techniques, we extracted a comprehensive set of features from the raw signals. These features, sourced from both the time and frequency domains and further enriched with Fourier and wavelet spectra, served as the backbone of our diagnosis system. Through the use of SHAP values, we were able to pinpoint and prioritize the features that played pivotal roles in the ConvLSTM model's decision making. This not only facilitated a cleaner and more streamlined input to our DL model, but also mitigated the risk of over-reliance on noise-prone or redundant features. The choice of a dual-path DL architecture—capitalizing on RNN-LSTM for temporal dependencies and a CNN for shift-invariant patterns—was pivotal in bolstering the robustness of our model, especially in the face of noisy data. Our crowning achievement, however, was in the application of SHAP values to elucidate the often complex reasoning of our network, revealing the key vibrational patterns and signatures linked to each diagnostic class.

In our study, we evaluated the performance of our XAI-ConvLSTM model compared to other algorithms in the field. By conducting thorough benchmarking on two different datasets, we found that our framework demonstrated a slightly superior performance in terms of both accuracy and computational efficiency. This advantage highlights the appropriateness of our approach for real-time applications, including resource-limited edge-computing environments.

In conclusion, the contributions of this paper have profound implications for the future of CM in the era of smart manufacturing. By ensuring a harmonious blend of performance, transparency, and efficiency, we hope to pave the way for safer and more reliable industrial ecosystems.

Future work will focus on domain adaptation and transfer learning associated with methods that aim to address model interpretability and improve the applicability of the proposed approach in different industrial scenarios.

Author Contributions: A.S. Conceptualization, Formal analysis, Investigation, Methodology, Algorithm, Coding, Validation, Visualization, Writing—Original Draft, Writing—Review & Editing, Supervision, Project administration. M.J. Conceptualization, Coding, Methodology, Validation, Visualization, Writing—Original Draft, Writing—Review & Editing, Supervision A.B. Writing—Review & Editing, Supervision. H.N. Writing—Review & Editing, Supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research received the financial support of NTWIST Inc., Edmonton, Canada and Natural Sciences and Engineering Research Council (NSERC) Canada under the Alliance Grant ALLRP 555220-20, and collaboration of Fraunhofer IEM, Düspohl Gmbh, and Encoway Gmbh from Germany in this research.

Data Availability Statement: All three datasets utilized in this study are openly accessible. The code can be found in the GitHub repository referenced in the main body of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Shojaeinasab, A.; Charter, T.; Jalayer, M.; Khadivi, M.; Ogunfowora, O.; Raiyani, N.; Yaghoubi, M.; Najjaran, H. Intelligent manufacturing execution systems: A systematic review. *J. Manuf. Syst.* **2022**, *62*, 503–522. [CrossRef]
2. Carvalho, D.; Pereira, E.; Cardoso, J. Machine learning interpretability: A survey on methods and metrics. *Electronics* **2019**, *8*, 832. [CrossRef]
3. Hofmann, D.; Phares, D. Remote monitoring and diagnostics of large rotating machinery. In Proceedings of the Cement Industry Technical Conference, 2003. Conference Record. IEEE-IAS/PCA 2003, Dallas, TX, USA, 4–9 May 2003; pp. 47–55. Available online: http://ieeexplore.ieee.org/document/1204708/ (accessed on 15 December 2023).
4. Li, F.; Xiao, J.; Huang, W.; Cai, S. Research on the Intelligent Obstacle Avoidance and Path Planning Strategy of UAV based on Multi-Sensor Fusion. In Proceedings of the 2022 IEEE International Conference On Advances In Electrical Engineering and Computer Applications (AEECA), Dalian, China, 20–21 August 2022; pp. 628–632.
5. Nguyen, V.; Case, J. Compensation of electrical current drift in human–robot collision. *Int. J. Adv. Manuf. Technol.* **2022**, *123*, 2783–2791. [CrossRef]
6. Heo, Y.; Kim, D.; Lee, W.; Kim, H.; Park, J.; Chung, W. Collision detection for industrial collaborative robots: A deep learning approach. *IEEE Robot. Autom. Lett.* **2019**, *4*, 740–746. [CrossRef]
7. Li, P.; Anduv, B.; Zhu, X.; Jin, X.; Du, Z. Across working conditions fault diagnosis for chillers based on IoT intelligent agent with deep learning model. *Energy Build.* **2022**, *268*, 112188. [CrossRef]
8. Azevedo, H.; Araújo, A.; Bouchonneau, N. A review of wind turbine bearing condition monitoring: State of the art and challenges. *Renew. Sustain. Energy Rev.* **2016**, *56*, 368–379. [CrossRef]
9. Jalayer, M.; Orsenigo, C.; Vercellis, C. Fault detection and diagnosis for rotating machinery: A model based on convolutional lstm, fast fourier and continuous wavelet transforms. *Comput. Ind.* **2021**, *125*, 103378. [CrossRef]
10. Safizadeh, M.; Latifi, S. Using multi-sensor data fusion for vibration fault diagnosis of rolling element bearings by accelerometer and load cell. *Inf. Fusion* **2014**, *18*, 1–8. [CrossRef]
11. Yang, L.; Yang, Z.; Song, S.; Li, F.; Chen, C.L.P. Twin Broad Learning System for Fault Diagnosis of Rotating Machinery. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 3510712. [CrossRef]
12. Zhou, Y.; Wang, H.; Wang, G.; Kumar, A.; Sun, W.; Xiang, J. Semi-Supervised Multiscale Permutation Entropy-Enhanced Contrastive Learning for Fault Diagnosis of Rotating Machinery. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 3525610. [CrossRef]
13. Antwarg, L.; Miller, R.; Shapira, B.; Rokach, L. Explaining anomalies detected by autoencoders using Shapley Additive Explanations. *Expert Syst. Appl.* **2021**, *186*, 115736. [CrossRef]
14. Cakiroglu, C.; Demir, S.; Ozdemir, M.; Aylak, B.; Sariisik, G.; Abualigah, L. Data-driven Interpretable Ensemble Learning Methods for the Prediction of Wind Turbine Power Incorporating SHAP Analysis. *Expert Syst. Appl.* **2024**, *237*, 121464. [CrossRef]
15. Ragab, A.; El-Koujok, M.; Poulin, B.; Amazouz, M.; Yacout, S. Fault diagnosis in industrial chemical processes using interpretable patterns based on Logical Analysis of Data. *Expert Syst. Appl.* **2018**, *95*, 368–383. [CrossRef]
16. Zhou, T.; Han, T.; Droguett, E. Towards trustworthy machine fault diagnosis: A probabilistic Bayesian deep learning framework. *Reliab. Eng. Syst. Saf.* **2022**, *224*, 108525. [CrossRef]
17. Busch, M.; Schnoes, F.; Elsharkawy, A.; Zaeh, M. Methodology for model-based uncertainty quantification of the vibrational properties of machining robots. *Robot. Comput.-Integr. Manuf.* **2022**, *73*, 102243. [CrossRef]
18. Xiao, Y.; Shao, H.; Feng, M.; Han, T.; Wan, J.; Liu, B. Towards trustworthy rotating machinery fault diagnosis via attention uncertainty in Transformer. *J. Manuf. Syst.* **2023**, *70*, 186–201. [CrossRef]
19. Shukla, S.; Yadav, R.; Sharma, J.; Khare, S. Analysis of statistical features for fault detection in ball bearing. In Proceedings of the 2015 IEEE International Conference On Computational Intelligence And Computing Research (ICCIC), Madurai, India, 10–12 December 2015; pp. 1–7.

20. Behzad, M.; Bastami, A.; Mba, D. Rolling bearing fault detection by short-time statistical features. *Proc. Inst. Mech. Eng. Part J. Process. Mech. Eng.* **2012**, *226*, 229–237. [CrossRef]
21. Wu, S.; Wu, P.; Wu, C.; Ding, J.; Wang, C. Bearing fault diagnosis based on multiscale permutation entropy and support vector machine. *Entropy* **2012**, *14*, 1343–1356. [CrossRef]
22. Zhang, Z.; Qian, K.; Schuller, B.; Wollherr, D. An online robot collision detection and identification scheme by supervised learning and bayesian decision theory. *IEEE Trans. Autom. Sci. Eng.* **2020**, *18*, 1144–1156. [CrossRef]
23. Huang, N.; Shen, Z.; Long, S.; Wu, M.; Shih, H.; Zheng, Q.; Yen, N.; Tung, C.; Liu, H. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. London. Ser. Math. Phys. Eng. Sci.* **1998**, *454*, 903–995. [CrossRef]
24. Lei, Y.; Lin, J.; He, Z.; Zuo, M. A review on empirical mode decomposition in fault diagnosis of rotating machinery. *Mech. Syst. Signal Process.* **2013**, *35*, 108–126. [CrossRef]
25. Silva, A.; Zarzo, A.; González, J.M.M.; Munoz-Guijosa, J. Early fault detection of single-point rub in gas turbines with accelerometers on the casing based on continuous wavelet transform. *J. Sound Vib.* **2020**, *487*, 115628. [CrossRef]
26. Sabir, R.; Rosato, D.; Hartmann, S.; Guehmann, C. LSTM based bearing fault diagnosis of electrical machines using motor current signal. In Proceedings of the 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, FL, USA, 16–19 December 2019; pp. 613–618.
27. Netsanet, S.; Zhang, J.; Zheng, D. Bagged decision trees based scheme of microgrid protection using windowed fast fourier and wavelet transforms. *Electronics* **2018**, *7*, 61. [CrossRef]
28. Jalayer, M.; Kaboli, A.; Orsenigo, C.; Vercellis, C. Fault detection and diagnosis with imbalanced and noisy data: A hybrid framework for rotating machinery. *Machines* **2022**, *10*, 237. [CrossRef]
29. Shi, J.; Peng, D.; Peng, Z.; Zhang, Z.; Goebel, K.; Wu, D. Planetary gearbox fault diagnosis using bidirectional-convolutional LSTM networks. *Mech. Syst. Signal Process.* **2022**, *162*, 107996. [CrossRef]
30. Jalayer, M.; Shojaeinasab, A.; Najjaran, H. A Model Identification Forensics Approach for Signal-Based Condition Monitoring. In *International Conference On Flexible Automation and Intelligent Manufacturing*; Springer Nature: Cham, Switzerland, 2023; pp. 12–19.
31. Belaid, M.K.; Mekki, D.E.; Rabus, M.; Hüllermeier, E. Optimizing Data Shapley Interaction Calculation from O (2^n) to O (tn^2) for KNN models. *arXiv* **2023**, arXiv:2304.01224.
32. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems. 2017; Volume 30. Available online: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (accessed on 20 January 2024).
33. Lu, C.; Wang, Z.; Qin, W.; Ma, J. Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification. *Signal Process.* **2017**, *130*, 377–388. [CrossRef]
34. Mao, W.; Feng, W.; Liu, Y.; Zhang, D.; Liang, X. A new deep auto-encoder method with fusing discriminant information for bearing fault diagnosis. *Mech. Syst. Signal Process.* **2021**, *150*, 107233. [CrossRef]
35. Shen, C.; Qi, Y.; Wang, J.; Cai, G.; Zhu, Z. An automatic and robust features learning method for rotating machinery fault diagnosis based on contractive autoencoder. *Eng. Appl. Artif. Intell.* **2018**, *76*, 170–184. [CrossRef]
36. Jia, F.; Lei, Y.; Guo, L.; Lin, J.; Xing, S. A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines. *Neurocomputing* **2018**, *272*, 619–628. [CrossRef]
37. Ahang, M.; Jalayer, M.; Shojaeinasab, A.; Ogunfowora, O.; Charter, T.; Najjaran, H. Synthesizing rolling bearing fault samples in new conditions: A framework based on a modified CGAN. *Sensors* **2022**, *22*, 5413. [CrossRef] [PubMed]