Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 180 (2021) 259–268

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Industry 4.0 and Smart Manufacturing

Explaining Learning Models in Manufacturing Processes

Claudia V. Goldman[a]* Michael Baltaxe[a] Debejyo Chakraborty[b] Jorge Arinez[b]

[a]*General Motors, Herzlyia, Israel*
[b]*General Motors, Warren, USA*

Abstract

The use of advanced machine learning (ML) models for manufacturing could potentially reduce the pre-production testing and validation time for new processes. Once we decide that ML is indeed a suitable tool to apply in smart manufacturing processes, the challenge lies in training, validating, and testing an ML model in a pre-production environment so that engineers can be confident that the model building effort can be successfully transitioned to actual production. This paper aims at explaining the in-works of a given in-situ classifier for predicting the quality welds in ultrasonic welded battery tabs. Predicting the quality of new samples cannot attain full certainty due to characteristics of the data the model was trained on (e.g., noisy or wrongly labeled). By developing explainable methods to such connectionist learning models (also known as black boxes), we show *why* the classifier outputs were predicted, making these predictions better understood and trustworthy.

Keywords: explainable AI; classifier learning systems; ultrasonic weld process monitoring

1. Introduction

Manufacturing engineers decide on parameter settings and machines calibrations in the plants based on statistical or other engineering common methods. Artificial Intelligence (AI) and machine learning (ML) solutions can improve these processes by speeding them up, by learning models of relevant data (e.g., [9,10,11]). AI methodologies are being introduced into industrial plants and processes through the Industry 4.0 and smart manufacturing trends. Machine to machine and human-machine interactions are studied to improve the effectiveness of production and planning processes [4]. Combining AI approaches into manufacturing control solutions seems to be of high promise of success where predictive models can support predictions for control of a plant or processes. The question is whether these

* Corresponding author. Tel.: +972-9-9720624; fax: +972-9-9720601.
E-mail address: claudia.goldman@gm.com

1877-0509
10.1016/j.procs.2021.01.163

advanced predictive models by themselves are enough, to be accepted and adopted by the manufacturing industry. Two important observations need to be analyzed. First, predictive models need to be monitored and adapted. In manufacturing facilities, changes in the environment or other settings can cause concept drifts, and consequently the accuracy of the learned models deteriorates. Second, the accuracy measure, computed today for quantifying the performance of learning models is not enough for quantifying the trust level humans should have in these models. Hence, even when advanced learning models can save time and effort, they might not be trusted.

With the advance in computing power, data availability and processing capabilities, learning algorithms in general, have become more complex. Consequently, the need to understand how these learning algorithms work, execute, predict and decide has become prevalent. Explainable AI is the area that has evolved to solve these issues, both in academia and in non-academic fields (see [8]) for an application in the medical domain). We believe that applying explainability methods can answer part of the issues of concern in smart manufacturing processes. Automated actions and predictions can be explained to the end users of these processes. Furthermore, explaining what a monitoring process encounters can assist in handling concept shifts. There are two main approaches in the explainable AI literature (for surveys see [2,13] (2020)): (1) Explainable systems: an AI system is built in a such a way that its behavior is interpretable. In this case, performance might be compromised to get added interpretability value. For example, a robot might not follow an optimal path, but it might follow a path that makes clear to its operator where it is headed [3,5]. Rudin [14] suggested using interpretable models instead of black boxes in domains at high stakes. Since concept drifts require the update of models repeatedly, a possible approach would be to develop transparent and interpretable models that are easy to understand by humans. Such interpretable models might make the human validation of these multiple models and their behaviors easier. (2) Explicatory systems: an explainable AI method applied to a given AI system provides explanations of its behavior. Machine learning (including deep learning applied to perception problems [6]) and planning [7,16]) are two main areas where explainable AI methods are being applied. For example, Lundberg et al. [8] have reported a solution based on decision trees learning models that could interpret global features from local decision trees.

In this paper, we opt for explicatory systems as our approach to explain predictive models in manufacturing. We assume that a classifier is given to us, therefore, our focus is not on building an interpretable model. Instead, our goal has been to provide explanatory methods to understand better the learning process of a classifier that predicted the quality of ultrasonic welding samples. We summarize the methods applied, together with the explanations we obtained. In particular, we have applied methods, developed for explaining perception systems learning from two-dimensional images, to data collected in a manufacturing process. Therefore, the explanations created result from visualization solutions (similarly as recommended in [4]). This can support the monitoring process when tracking concept drifts as mentioned previously.

Here, we report the performance of explainable AI algorithms run on historical data collected from ultrasonic welding (USW) processes. In the following sections we present the USW historical data (Section 2.1, see also [1]) and the classifier (Section 2.2) implemented to predict the quality of samples taken from the historical data. Section 3 describes two methods taken from the computer vision domain, applied here to explain the predictions of this classifier on the USW historical data (one dimensional (1D) data). Finally, sections 4 and 5 discuss the results and conclude this paper.

2. Ultrasonic welding quality prediction

Ultrasonic welding data that is used in this paper was obtained from battery tab welding of Volt Generation I battery at Brownstown Battery Assembly Plant and is the same one discussed in [1]. The ultrasonic welding setup was as shown in Figure 1.

A vibrating sonotrode compressed three cell tabs and a bus bar against a fixed anvil, establishing an electrical parallel connection. Three temporal signals were acquired at 100 kHz: applied power, welding horn translation along the plane perpendicular to the plane of the cell tabs, and the emitted acoustic signal. The system implemented at the Plant classified each weld into good and suspect quality. Only the suspect welds were then manually inspected to classify into a good and bad. The deployed algorithm took months of feature engineering and validation before it achieved its goal: miss no bad welds at the cost of misidentifying up to 30% good welds as bad.

The natural progression to eliminate feature engineering would be to use a black box AI technique such as deep learning. However, such black box methods lack the explainability property: it is not easy to determine from the neural network architecture and learning algorithm why a certain sample was predicted to have good quality or bad quality. The object of this study was to delve deep into a deep learning technique and help interpret its outcome. The interpretation was to be clear enough for a non-AI expert at the plant floor to relate to and take remedial action on the system. To demonstrate this concept, a classifier was built (not deployable, but enough to demonstrate the concept) and the outcome was explained.

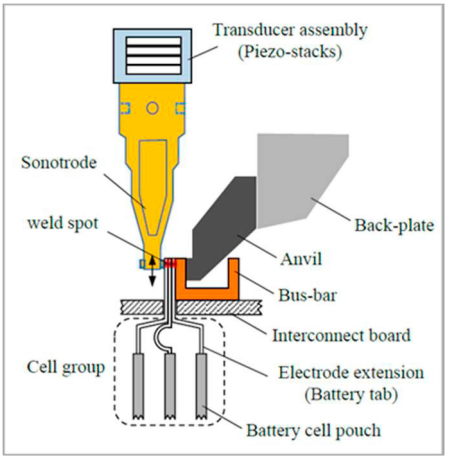

Fig. 1. Ultrasonic Welding Setup.

2.1. The ultrasonic welding historical data set

The historical data set comprised ten million data points, each including a set of 3 signals and a binary ground truth quality measure. The signals stored in the database are the following:

- Linear variable differential transformer (LVT): a measure of displacement of the horn during the welding process.
- Delivered power (PWL): the power delivered by the welder during the welding process.
- Acoustic signature (ASO): the acoustic signal produced during the welding process.

An automatic system analyzed the three signals and classified the quality of the welding. The welds that were marked as good quality (quality=0) were never inspected further and this label is the ground truth stored in the database. These welds were assumed to be good since the module passed the end-of-line functional / performance test, and over the years did not have a field return. The welds that were identified by the automatic process as bad quality were then manually investigated. If the operator found the weld to be good, then the good quality value (quality=0) was stored in the database as ground truth, otherwise the bad quality value (quality=1) was stored as ground truth.

This database eventually provided the ground truth and the signals for this study (Figure 3 shows an example of the shapes of these signals). Details on this could be found in [1].

2.2. Predicting the quality of a USW datapoint – the classifier design

A random subset of the data was divided into train (9587 samples) and test (2400 welding samples). A one-dimensional (1D) convolutional neural network (NN) classifier was trained. The three signals (PWL, LVT, ASO) were stacked to create a 3-channel input signal for the classifier. Two common performance measures for classifiers are: (1) the area under the ROC curve and (2) the classifier accuracy. The ROC curve depicts true positive rates vs. false positive rates. The area under the ROC curve is between 0 and 1. The larger the value of the area under the ROC,

the better the classifier (i.e., if we pick randomly a sample from the true positive set and another one from the true false samples, then the area under the ROC will equal the probability that the sample taken from the true positives will be closer to the highest value of the classifier output). Accuracy was calculated as the number of welding samples for which the ground truth equals the predicted value, divided by the total number of welding samples in the test set. Our classifier results were:

- Area under ROC curve: 0.77
- Accuracy: 0.78

The confusion matrix for the classifier is presented in Table 1. A confusion matrix describes the ratio of samples, classified for each class (the predicted label in the columns) considering the ground truth (as shown in the true labels in the rows in Table 1). For example, 81% of the welding samples with actual good quality were indeed classified as "good quality" welding samples. Only 19% of the true good quality welding samples were misclassified as having "bad quality". Usually, we would prefer 0% type II errors or false negatives (that is, the ratio of welding samples having true bad quality are all classified correctly). However, based on the data available to us for this project, the following confusion matrix shows the best classification we could attain. This is the classifier that we explain later in this report. The same methods implemented here to explain this classifier can be applied to other classifiers when successfully learned to attain 0% type II errors.

Table 1. Confusion Matrix for the Tested Classifier.

True Label / Predicted Label	Good Quality	Bad Quality
Good Quality	0.81	0.19
Bad Quality	0.37	0.63

3. Explainable AI applied to the USW classifier

In this section we discuss the applicability of two methods taken from the explainable AI literature. The methods implemented are mostly used when the objective is to explain black boxes learning models, used for classification tasks in computer vision. The goal, however, is general for a learning model that is not transparent to the user. Therefore, by applying such explainable methods we can understand the inner works of a connectionist model. We will show how we apply these methods in the USW welding domain, where the input data is provided as a 1-dimensional data vector and not as a two-dimensional image. Today, deep learning solutions can be highly accurate, meaning that for a test set of data, their prediction about the input category having learned from a training data set could be high.

However, this high-performance measure, although necessary to trust a learning model, is not enough. We believe that making the predictions of the learning classifiers understandable is a necessary step towards adopting the learning algorithms in the smart manufacturing community.

With this objective of increasing interpretability of such learning models in smart manufacturing, we describe the application of the following methods and the analysis of the obtained results in the next sub-sections:

1. Class Activation map
2. Contrastive gradient-based saliency maps

3.1. Method 1: interpretable model with Class Activation Maps

The Class Activation Map (CAM) method [18] works on convolutional neural networks whose last two layers are a global average pooling (GAP) layer followed by a single fully convolutional layer. This is the case of the classifier built for the USW data, which takes as input three signals stacked along the time axis (for time equals t, we have the values of the three signals).

Let f_k be the output of filter k in the last convolutional layer before the GAP and let w_k^c be the weight associated with filter k in the fully convolutional (output) layer for the class to explain c (c equals 0 when the class corresponds

to the "good quality" class, and 1 otherwise). The Class Activation Map (CAM) Mc is calculated as follows for the element x of the classifier input:

$$M_c\,(x) = \sum_k w_k^c\; f_k(x) \tag{1}$$

We can interpret this as follows. The weights w_k^c of the last fully connected layer govern the classifier outcome while the internal layers extract features f_k from the input. When features that correspond to the selected class are found at a particular location of x, these features $f_k(x)$ attain high values. The weights in the last layer provide a weighted sum of each pattern, yielding a final heat map, showing these high valued features at locations of x that support the classification decision towards class c.

Thus, a CAM can be interpreted as a heat map and visualized with several color maps. When we use a "temperature" color map, warm colors will show locations that provide evidence for a class defined as "1", while cool colors provide evidence for a class defined as "0". This is usually used in the computer vision domain to visualize regions of an input image that provide evidence towards the positive class, for example, elements in the scene that provide support for classifying the image as belonging to a given category. In our domain, it will highlight sections that provide evidence for either good or bad quality welding.

For the USW welding quality prediction CNN as reported here, we applied the CAM method. Figure 2 presents the output of the CAM method when a good quality welding sample is provided as input to the classifier. The figure shows in blue colors the representation of low activation scores, corresponding to the "good quality" class while the red colored graph represents high activation values corresponding to the "bad quality" class. A good quality welding sample was defined to have a score of 0, while a bad quality welding sample was defined to have a score of 1 (see the definition of the historic ultrasonic welding dataset above). Therefore, when we see a region with blue color, it will be understandable for a human that the welding sample has "good quality", on the contrary, red color will indicate a "bad quality" welding sample.

Figure 3 shows a comparison of results for good and bad quality welding samples. We can see that the color patterns shown by the explainable methods are different and can help understand what settings and patterns indicate whether the welding sample is going to be associated with a good or bad quality.

Figure 3 details two different welding samples (top and bottom). For each welding sample, we visualize three signals, which are aligned in time (i.e. the x axis is the same for the three signals). In each, the shape corresponds to each signal, while the color corresponds to the weighted activation of the last layer in the network (the output of the CAM method). Hence, for example, we can interpret the good quality of a welding sample by observing a blue colored peak in the PWL signal as shown in the upper left graph in Figure 3. Another example is the green colored bell shape of the acoustic signal in the bottom right graph of Figure 3, indicating a welding sample with bad quality.

CAM explains by means of visualization the importance of the input signals for good or bad quality predictions (all 3 signals were trained together PWL, AS0, LVT). We showed that the first part of the PWL signal determines the quality prediction. "Good" welding samples present blue-toned colors while "bad" welding samples get yellow to red colors. Patterns for visualizing good/bad quality predictions were also noted in the ASO signal.

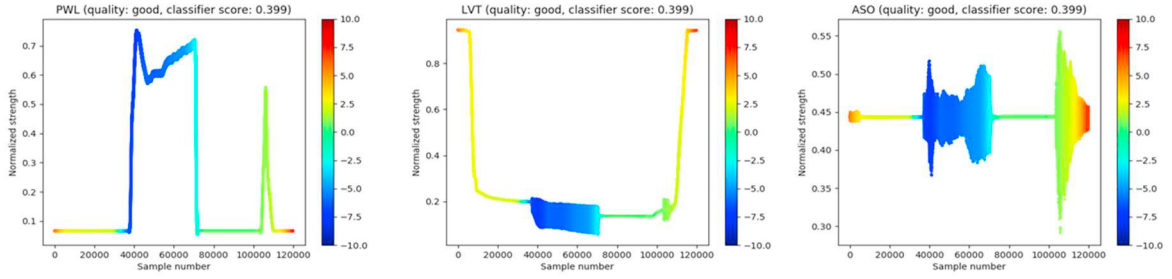

Fig. 2: A Good Quality Welding Sample

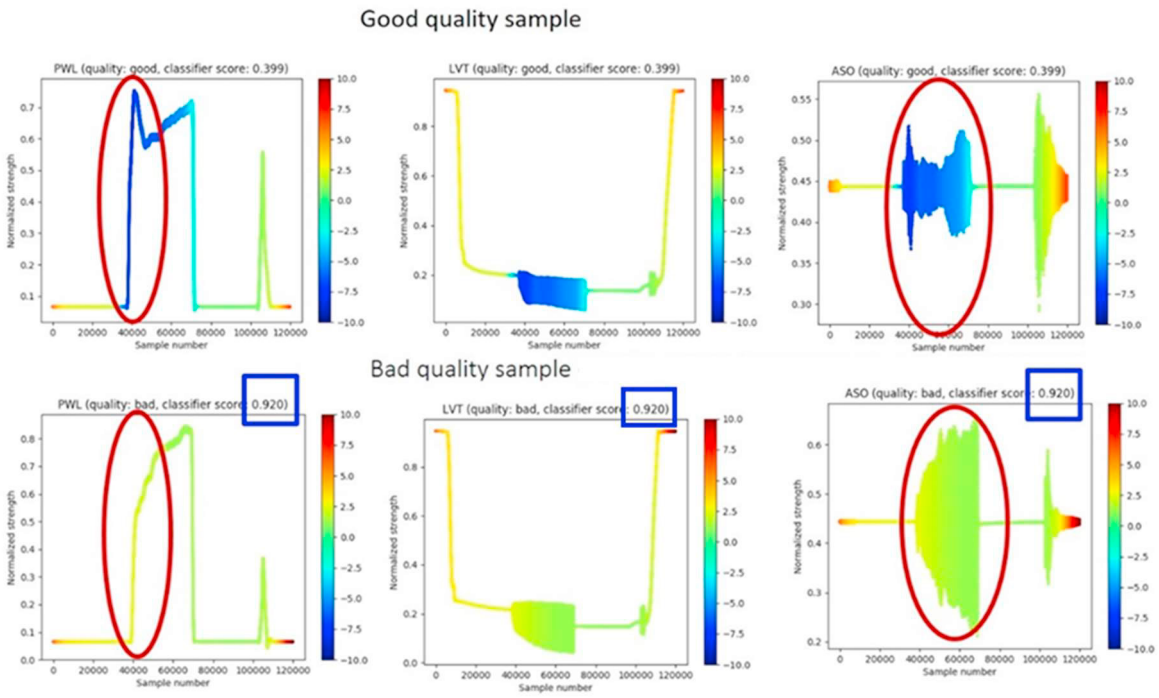

Fig. 3. Good Quality Welding Sample (top) and Bad Quality Welding Sample (bottom).

CAM is effective in determining features corresponding to each class (good / bad quality). An immediate result we observe from implementing CAM is that the quality of the classifier can be explained by looking at the speed at which the PWL steps up: when the PWL shows a strong step gradient, we can predict a good quality welding sample; but when the visualization shows a weak step gradient, we can predict a bad quality welding sample. We also see that the acoustic signal shape serves as another explanation for the classifier quality prediction. When this signal has a clear bell shape pattern, this is an explanation for bad quality welding samples. We should note that our implementation of stacking the three measurements into a 3-channel signal loses the information about the quantity responsible for a given classification, but it keeps the temporal information. An alternative implementation remains for future work, where we stack the signals sequentially along the temporal axis (in 1 channel). This would allow to identify both the important quantity and the temporal features. We believe this might also yield a less accurate but more interpretable classifier. Less accuracy will result from the loss of dependencies among the components of the signal. More interpretability will result from the focus directed to single parts of the signal.

3.2. Method 2: robust model with Contrastive Gradient-based Saliency maps

Evaluating the gradients along the layers of the learned model can help us better understand how the neural network captures the information along the learning process. Robust models are not sensitive to perturbations which is a very desirable characteristic of a classifier.

To measure the robustness of the classifier, we used contrastive gradient-based saliency maps [15]. This method calculates the saliency map for a specific input (I0) and class (c). The map is calculated as the gradient of neural network's output Sc(I) with respect to the input. For a traditional neural network, this gradient is calculated by backpropagation. Mathematically, the map (w) is calculated with the following expression:

$$w = \frac{\partial S_c}{\partial I}\Big|_{I_0} \tag{2}$$

Figure 4 presents an example of this saliency map. Strong red and blue colors represent regions where a small change to the input signal cause a great change in the output of the classifier. The color is related to the direction of the change in the output when a change in the positive direction occurs on the input. In our case, however, the direction is not important, only the magnitude.

Ideally, we would like to have only small magnitudes on the gradient since this means that the output of the classifier does not change with a small perturbation on the input, making it robust to noise and small variations on the input. In other words, we would like to have only green, cyan, yellow colors on the saliency map. In addition to understanding the behavior of the output, it is important to understand what the internal layers of the network are doing. This is important since it might shed light on the concepts learned by the network.

Our classifier is a convolutional neural network. This means that the network is composed of a set of convolution filters followed by a fully connected layer. The convolution filters can be interpreted as feature detectors. In other words, it is possible to think of an intermediate activation map as the response to a filter, which identifies if an interest feature is present or not. In this context we refer to an activation as Activation = convolution (signal, filter), where the filter is the interest feature.

The contrastive gradient-based saliency maps applied to manufacturing data explains the behavior of the internal activations of the network. We have found that shallow activations are very noisy and tend to "follow" the input signal.

As we go deeper in the network, activations focus on the pieces of the signal holding most of the information (see Figure 5). Keeping activations stable as the input signal changes slowly might be an indication of a stable classifier.

An interesting question is whether we can design a filter that is connected to the physics of the process or whether we can find the physics in the filters. This task remains for future work. For a learning model to be easy to explain, we would like it to be stable, i.e., we want small gradients through all the data space (i.e. all feasible input signals). Locations in data space that have large gradient magnitudes probably need more coverture in the training set. Guaranteeing an "average gradient magnitude" through all data space lower than a selected threshold might provide certainty of the classifier. Gradients can be calculated with respect to any activation, not only the output.

4. Discussion

This work shows how evaluations of two explainable AI methods (developed for computer vision) could be applied to better interpret the predictions learned in the manufacturing domain. These explanations could add value and justify smart manufacturing advanced technologies. For example, to explain how processes might be improved by reducing costs incurred during manual inspections (as an alternative to automated learning methods for inspection and prediction of quality). In case of a quality spill, for another example, a clear indication of the point of failure in the signal may provide clues to the fix. That is, if the initial pressure is consistently low across successive welds, one could infer that the engagement pressure is inadequate. A next step would be to apply the excitation backpropagation method to explain how intermediate layers capture information about the learning process [17]. Other related methods appear in [12].

The authors believe that explainable AI can be helpful in applying AI/ML methods and tools in the manufacturing environment where black-box solutions are not already adopted in solving real world problems. If we can develop ways to better describe why models give certain predictions and link them to what engineers physically observe, then this will facilitate adoption and implementation of such tools. The USW dataset provides an initial example for us to understand the explainable AI methods we started evaluating. There are other potential applications. Powertrain machining, body-shop dimensional quality, paint-shop quality, and other welding methods could benefit from descriptive classifiers, features, predictions, and more.

5. Conclusions

Predicting the quality of a manufacturing process output is valuable for advancing smart manufacturing. Today, machine learning models are scored with only accuracy scores, which are not enough to determine that these can be trusted even when this score is very high. We believe that developing methods to make learning models interpretable

and understandable to a larger audience of users, beyond those that develop them, is valuable to introduce these advanced algorithms in smart manufacturing processes. Methods such as the CAM method or contrastive gradient-based saliency map serve to increase the ability of development or even process control engineers to better visualize expected results beyond just simple numerical metrics. Such visualizations allow managers to develop intuition over time about the output of AI/ML derived classifiers. Potential areas of further investigation are to apply these methods to other manufacturing process control problems and assess the generalizability of these methods.

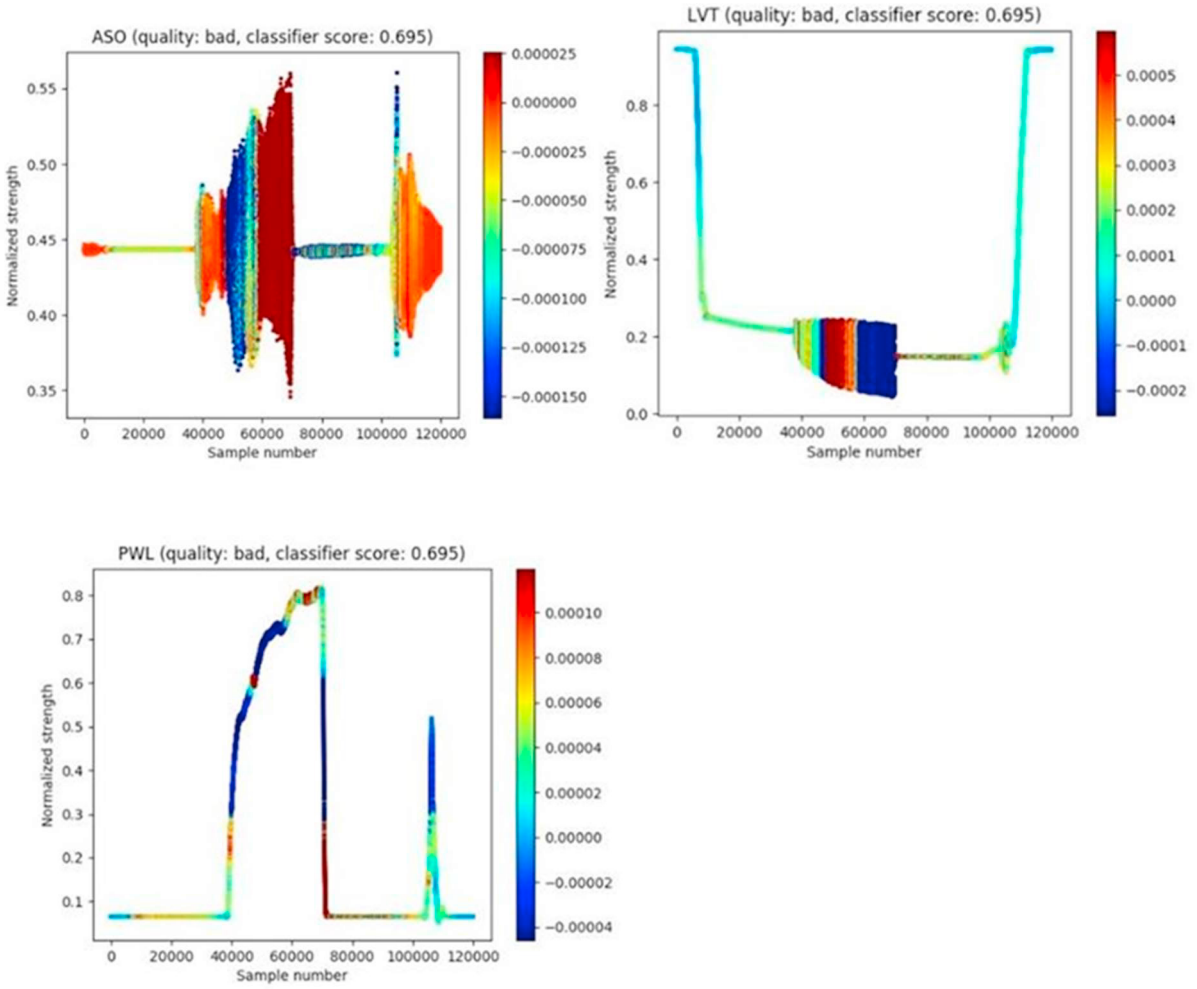

Fig. 4. Contrastive Gradient-based Saliency Map.

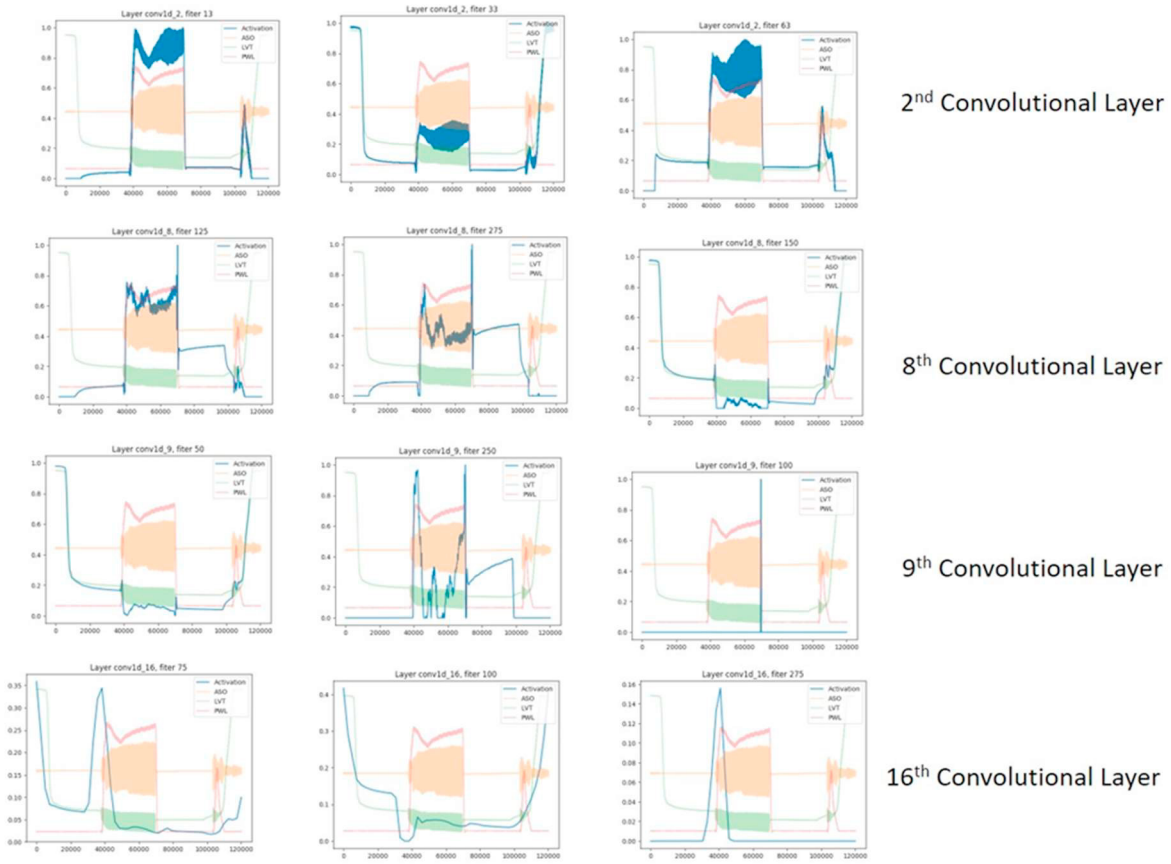

Fig. 5: Signal Activations Throughout the Network

References

[1] J. Abell et al. (2017). "Big Data-Driven Manufacturing—Process-Monitoring-for-Quality Philosophy", *Journal of Manufacturing Science and Engineering*.

[2] A. Adadi and M. Berrada, (2018). "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". *IEEE Access*, vol. **6**, pp. 52138-52160.

[3] J. Fisac et al. (2016). "Generating Plans that Predict Themselves". WAFR

[4] G. Kronberger et al. (2020). "Smart Manufacturing and Continuous Improvement and Adaptation of Predictive Models". *Procedia Manufacturing* **42** (2020) 528–531

[5] S. Huang et al. (2017). "Enabling Robots to Communicate Their Objectives. Robotics". *Science and System XIII*. July, MIT, MA.

[6] J. Kim and J. Canny. (2017). "Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention". *International Conference on Computer Vision*.

[7] A. Kulkarni et al. (2019). "Explicable Planning as Minimizing Distance from Expected Behavior". *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems.*

[8] S. Lundberg et al. (2020). "From local explanations to global understanding with explainable AI for trees". *Nature Machine Intelligence* volume **2**, pages56–67.

[9] McKinsey Global Institute. (2017). "What's now and next in analytics, AI, and automation". May. https://www.mckinsey.com/featured-insights/digital-disruption/whats-now-and-next-in-analytics-ai-and-automation

[10] McKinsey Global Institute. (2018). "Skill Shift Automation and the future of the workforce". May. https://www.mckinsey.com/

[11] McKinsey Global Institute. (2016). "The age of analytics: competing in a data driven world". December. https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world

[12] T. N. Mundhenk, B.Y. Chen, and G. Friedland (2020). "Efficient Saliency Maps for Explainable AI". https://arxiv.org/abs/1911.11293

[13] A. Rosenfeld & A. Richardson. (2020). "Why, Who, What, When and How about Explainability in Human-Agent Systems". *Proc. of the 19th International Conference on Autonomous Agents and MultiAgent Systems.*

[14] C. Rudin. (2019). “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead”. *Nature Machine Intelligence*, Vol 1, May 2019, 206-215
[15] K. Simonyan, A. Vedaldi and A. Zisserman. 2014. Deep inside convolutional networks: visualizing image classification models and saliency maps. Proceedings of *ICLR*.
[16] S. Sreedharan et al. (2019). “Model-Free Model Reconciliation”. *Proc. of the 28th International Joint Conference on Artificial Intelligence*. Pages 587-594
[17] J. Zhang et al. (2016). “Top-down Neural Attention by Excitation Backprop”. *Proceedings of ECCV*.
[18] B. Zhou et al. (2016). “Learning Deep Features for Discriminative Localization”. *Proceedings of CVPR.*