Journal of Industrial Information Integration 33 (2023) 100439

Contents lists available at ScienceDirect

Journal of Industrial Information Integration

journal homepage: www.elsevier.com/locate/jii

Full length article

Interpreting learning models in manufacturing processes: Towards explainable AI methods to improve trust in classifier predictions

Claudia V. Goldman [a,*], Michael Baltaxe [a], Debejyo Chakraborty [b], Jorge Arinez [b], Carlos Escobar Diaz [c]

[a] General Motors Global Research and Development, Herzliya Pituach, Israel
[b] General Motors Global Research and Development, Warren, MI, USA
[c] Amazon Flex Lab, Seattle, WA, USA

ARTICLE INFO

Keywords:
Explainable AI
Classifier learning systems
Ultrasonic weld process monitoring
Artificial intelligence quotient

ABSTRACT

Smart manufacturing processes, building upon machine learning (ML) models could potentially reduce the pre-production testing and validation time for new processes. Beyond calculating accurate and reliable models, one critical challenge would be for users of these models (plant operators, engineers and technicians) to trust these models' outputs. We propose to apply explainable AI methods to create trustworthy AI-based manufacturing systems. Consequently, these systems will be enriched with capabilities to explain their reasoning processes and outputs (e.g., predictions) automatically. This paper applies explainable AI methods to two problems in manufacturing: ultrasonic weld (USW) quality prediction and body-in-white (BIW) dimensional variability reduction. Class activation maps were computed to explain the effect of input signals and their patterns on the quality predictions of an ultrasonic weld yield by a neural network (good or bad). Contrastive gradient based saliency maps were also created to assess the robustness of this classifier. Furthermore, we explain a given connectionist network that predicts the dimensional quality of body-in-white framer points based on deviations in underbody points. Explaining these predictions help engineers understand which underbody points have more influence on deviations in the framer points. These two applications highlight the importance of explainable AI methods in the modern manufacturing industry.

1. Introduction

Studies and applications in the domain of industrial information integration have shown improvement in industrial processes in areas including agriculture, healthcare, automated factory, construction, and others (see [1,2] with applications such as robots, smart cities, smart energy application, smart healthcare, etc.). These applications are based on an information platform, and information technologies. In this paper, we want to emphasize the need for a higher-level layer on top of the information technology, whenever these technologies interact with their end users (e.g., customers, engineers, technicians, operators, etc.). The reader should note that the interpretability of the intelligent actions executed by the industrial information integration system might not be needed in all applications. For example, a 3D indoor map building system might not have to explain its user how it obtained the current coordinates of the system, while a user is interacting with it to understand the location of himself or another object [3].

This paper does focus on those systems where users and systems need to collaborate, particularly when these users are not those that created these system (e.g., the plant operator in a factory might need to interact with a system; although the operator is not necessarily the AI engineer that developed that system: a recommendation system or a machine learning model that predicts the quality of an element in the manufacturing process). For example, these could be systems with behaviors that are determined by complex machine learning models (i.e., these systems have been trained on large amount of data to eventually predict and recommend actions to their users). In such scenarios, these systems might need to explain their behaviors or the reasoning behind them (i.e., the why behind their outputs) for their users to trust them and consequently rely on them. For example, explanations might be needed when faults are detected (for example in a wind-energy related application [4] or in an industrial control ecosystem [5]) or in healthcare systems [6]. In particular, we are interested in smart manufacturing applications. Manufacturing plants are a wealth of data, and thus a fertile domain for data-driven analytics.

* Corresponding author.
E-mail address: claudia.goldman@gm.com (C.V. Goldman).

https://doi.org/10.1016/j.jii.2023.100439
Received 16 March 2021; Received in revised form 18 June 2022; Accepted 13 February 2023
Available online 28 February 2023
2452-414X/© 2023 Elsevier Inc. All rights reserved.

It is common knowledge today that artificial intelligence (AI) and machine learning (ML) solutions can improve and accelerate process parameter tuning and quality prediction, by learning the non-apparent structure within the data [7–9]. The advent of Industry 4.0 and 4th generation of non-destructive evaluation [10] (NDE 4.0) has enabled automatic data handling.

We are interested in maintaining a clear communication channel among the systems, based on information and their users. The essence of this clear communication is to create information-based systems that are trustworthy, well-accepted and used effectively by all their relevant users. The smart manufacturing process success depends on a trustful relationship between these advanced information systems and their users. Machine to machine and human–machine interactions are already being studied to improve the effectiveness of production and planning processes [11] and to improve interpretability of monitoring, diagnostics and prediction of smart industrial assets [12]. Yet, advanced machine learning based solutions for quality control and prediction are approached with, justly, caution and apprehension. Two recent reports review the applicability of explainable AI algorithms to different domains including industry and healthcare [13] and manufacturing in particular [14].

AI approaches are often treated as "black box" with little or no attempt to explain the results. Consequently, the AI results in manufacturing are received with skepticism. The solution by itself is inadequate. An explanation of "why" the solution works and the interpretation of the results is imperative for adoption in manufacturing. A reliable and well understood result is preferred over a fast one. Explainable AI addresses this gap between accurate intelligent solutions but hard to interpret by a human, and accurate intelligent solutions that are also understandable and interpretable by anyone of its users [15]. In other words, how humans have intelligence quotient, enterprises have knowledge quotient [16,17]. A successful AI needs to have a high AI quotient (AIQ) [18–20].

In the existing literature [21,22], explainable AI (XAI) has two distinct approaches, explainable systems and explicatory systems. An explainable system is inherently interpretable, even at the cost of performance. For instance in a robot navigation challenge, finding an explainable solution was preferred over an optimal one [23,24]. In a high stakes and complex environment, such as manufacturing, engineers would certainly agree with Rudin [25] that models which are interpretable by humans are more valuable than the obscure optimal ones.

An explicatory system on the other hand, attempts to explain the optimal obscure solutions. A couple well known examples that have attracted this approach are deep learning as applied to perception problems [26,27], and planning problems [28,29]. For example, Lundberg et al. [15] have reported a solution based on decision trees learning models that could interpret global features from local decision trees.

In this paper, we have opted for the explicatory system and demonstrated its applicability in two problems, ultrasonic weld (USW) quality prediction and body-in-white (BIW) dimensional variability reduction. Two connectionist networks were given, that output their predictions of relevant outputs in each domain. Our work applies explainability methods to these two domains resulting in automated and interpretable explanations of the computed predictions.

The USW quality prediction problem is discussed in Section 2. A neural network was built to classify good welds from bad ones. Then, the result of the classifier was explained using visualization techniques similar to [11]. This work, first communicated in [30], did not focus obtaining the best classifier. The primary objective was to explain the outcome. In Section 3, we have reconsidered that discussion, and in Section 4 extended our work to a dimensional problem of BIW.

BIW dimensional variability reduction is one of the most relevant challenges for multistage manufacturing quality control. Unchecked deviations in dimensions accumulate across successive assembly operations, amplifying the problem beyond recovery. Such defects are visible to the customer and at least impacts the perceived quality of the vehicle, if the vehicle is at all buildable. The recent advances in vision systems empowered by AI has enabled *in situ* assessment of dimensional quality. In Section 4 we have discussed a technique to predict dimensional deviations and attempted to understand what impacts such deviations. Such understanding augments human understanding and empowers the engineers on the floor to troubleshoot just-in-time and avoid quality spills.

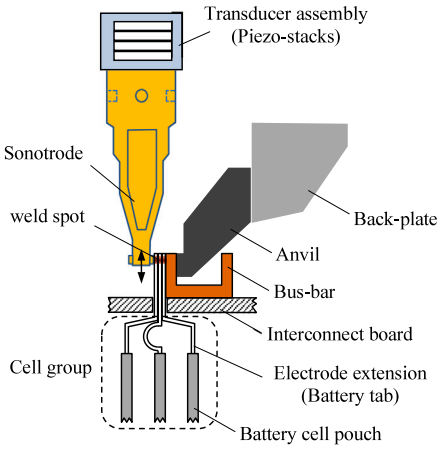


Fig. 1. Ultrasonic welding setup [31].

A consolidated understanding and the message is then provided in Section 5. That section also paves the path to research directions that would be invaluable to this industry.

This paper follows some notational convention that is designed to maintain consistent algebraic representation, aiding the readability. Throughout this document the domain of any algebraic entity discussed will be described as number sets: $\mathbb{R}$ is a set of real numbers, $\mathbb{N}$ is a set of natural numbers. Multidimensional numbers belong to multidimensional sets, for example, a three dimensional real number belongs to $\mathbb{R}^3$. Bold symbols are used to represent such data. Matrices are represented by upper case symbols.

2. Ultrasonic welding quality prediction

Ultrasonic welding of metal is a complex manufacturing process which cannot be directly verified. Surrogate mechanical and electrical functional properties are generally used to verify the quality, but that does not indicate the reliability or the stability of the manufacturing process. Monitoring of the process itself assures quality and goes beyond what the end of line functional check can ascertain. Such a process monitoring system was deployed to make the first generation Chevy Volt battery at Brownstown Battery Assembly Plant. The implementation philosophy and success of the method has been discussed in [31]. The original implementation entailed engineering knowledge and judicious design of classifier ensemble by subject matter experts. Those methods were carefully motivated and were very well understood from an engineering stand point. In this paper, we have investigated if XAI could provide some of that insight when applied to a "black box" deep learning technique.

Ultrasonic welding is a vibration welding method. In the above mentioned application, this welding was used to join three cell tabs to a bus bar (collectively referred to as the "weld target" hereafter) in a battery module using a setup schematically represented in Fig. 1. A sonotrode (horn) vibrating at about 20 kHz compresses the weld target against a fixed anvil, establishing an electrical parallel connection among the cells. The following three time series signals were acquired at 100 kHz for little over the duration of the weld.

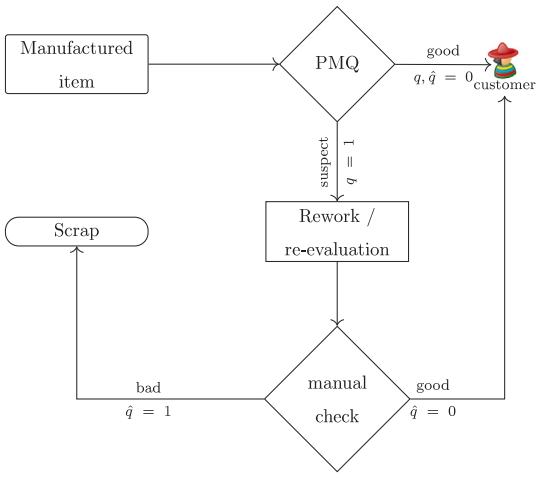


Fig. 2. Schematic representation of process monitor for quality implementation.

- 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.

Fig. 4 shows an example of the shapes of these signals.

The deployed system analyzed the three signals and classified the quality of the weld into "good" and "suspect" welds. This quality attribute $q \in \{0, 1\}$ is written in a database that contains all the information regarding these welds. The "good" ($q = 0$) welds did not go through any further inspection except an aggregate functional electrical test at the end of the assembly process. The "good" label thus became the ground truth $\hat{q} = 0$ that is available today. In case a functional test failed, and a weld was found to be a problem, the database was updated. The "suspect" ($q = 1$) welds went through secondary manual inspection to verify "good" ($\hat{q} = 0$) or "bad" ($\hat{q} = 1$). This classification thus became the final label. The "bad" welds were either repaired, and subsequently marked "good", or the module was scrapped altogether. Fig. 2 depicts the flow of this process, where PMQ stands for Process Monitoring for Quality and is a concept introduced in [31] that combines process monitoring (PM) with quality control (QC). The PMQ concept uses specific observable features and indicators derived from knowledge of a process along with the QC philosophy of predicting the fitness of the product. This database eventually provided the ground truth and the signals for this study (see [31] for further details on PMQ and corresponding data).

The system was designed to not miss any bad welds and was thus biased towards false rejects (calling good welds suspect). Up to a 30% false reject was acceptable.

The features of this system was from a hand crafted set, designed by subject matter experts. A natural next step would be to replace feature engineering with a black box method like deep learning. However, it is essential to explain why the quality was labeled the way it was. XAI tools are thus necessary to qualify such a system for plant floor use. To demonstrate if XAI could explain some aspects of the physical resemblance, a deep learning classifier was built (not deployable, but enough to demonstrate the concept) and the outcome was explained.

Out of the millions of welds available, a random block of 11,987 welds in contiguous time from a specific welder was selected for this simulation. A random subset of 9587 welds were chosen to train a one-dimensional (1D) convolutional neural network (CNN).

The input was created by combining PWL signal, LVT signal and ASO signal, $\boldsymbol{s}_\mathrm{p}, \boldsymbol{s}_\mathrm{v}, \boldsymbol{s}_\mathrm{a} \in \mathbb{R}^{n_s}$, respectively, into a 3-channel $S = [\boldsymbol{s}_\mathrm{p}\ \boldsymbol{s}_\mathrm{v}\ \boldsymbol{s}_\mathrm{a}]$;

Table 1
Ultrasonic weld quality classifier confusion matrix.

Truth ↓/predicted →	Good	Bad
Good	0.81	0.19
Bad	0.37	0.63

$S \in \mathbb{R}^{n_s \times 3}$, where $n_s \in \mathbb{N}$ is the number of samples in each signal. Here, $\boldsymbol{s}_\mathrm{p}, \boldsymbol{s}_\mathrm{v}, \boldsymbol{s}_\mathrm{a}$ are discrete time signals, i.e. $\boldsymbol{s}_\mathrm{p} = s_\mathrm{p}(\mathrm{x})$, $\boldsymbol{s}_\mathrm{v} = s_\mathrm{v}(\mathrm{x})$, and $\boldsymbol{s}_\mathrm{a} = s_\mathrm{a}(\mathrm{x})$, defined over all $\mathrm{x} = 1, \ldots, n_s$. Also, note that in the second index $\mathrm{y} = 1, 2, 3$ in $S(\mathrm{x}, \mathrm{y})$ correspond to $s_\mathrm{p}, s_\mathrm{v}, s_\mathrm{a}$, respectively.

The CNN had n_h layers, excluding the input layer. There were $n_h - 2$ convolutional layers containing $n_k^{[h]}$ filters in the hth layer ($h = 1, \ldots, n_h - 2$). The second last layer is a global average pooling (GAP) layer containing $n_k^{[n_h-1]}$ nodes. If the activation at any layer is denoted by a, then the activations at GAP layer is

$$a_k^{[n_{h-1}]} = \sum_\mathrm{y} \sum_\mathrm{x} a_k^{[n_{h-2}]}(\mathrm{x}, \mathrm{y})\ \forall k \in \{1, \ldots, n_k^{[n_{h-2}]}\}. \tag{1}$$

The final layer is the classification layer. Since there are only two classes "good" and "bad", $n_k^{[n_h]} = 2$. If the weights of any fully connected layer is w, the activation in this layer is

$$a_k^{[n_h]} = \sum_{t=1}^{n_k^{[n_{h-1}]}} w_t^{[n_{h-1}]} a_t^{[n_{h-1}]}\ \forall k = 1, \ldots, n_k^{[n_h]} \tag{2}$$

The obtained classifier (thresholded $a_k^{[n_h]}$) performance was evaluated by (1) the area under the receiver operating characteristic (ROC) curve $\alpha \in [0, 1]$, and (2) the classification accuracy $\eta \in [0, 1]$. The ROC curve [32] depicts true positive rates versus false positive rates. Perfect classification is achieved at $\alpha = 1$. Accuracy was calculated as the fraction of the welds for which the true quality $\hat{q}$ was correctly predicted, with the perfect value being $\alpha = 1$. This classifier achieved $\alpha = 0.77$ and $\eta = 0.78$ on 2400 test welds. The confusion matrix for the classifier is presented in Table 1. The correct classification rate for each class lies along the leading diagonal of a confusion matrix, and the off-diagonal entries represent the mistakes. For example, 81% of the good welds were classified as "good", only 19% of the good welds were misclassified as "bad". A perfect classifier with $\eta = 1$ would have an identity matrix as confusion matrix. The plant acceptance criteria was zero missed defects, but here we achieved 37%.

Since designing the best classifier was not the objective here, we accepted these results and moved on to exploring if XAI could explain these results.

3. Explainable AI applied to the USW classifier

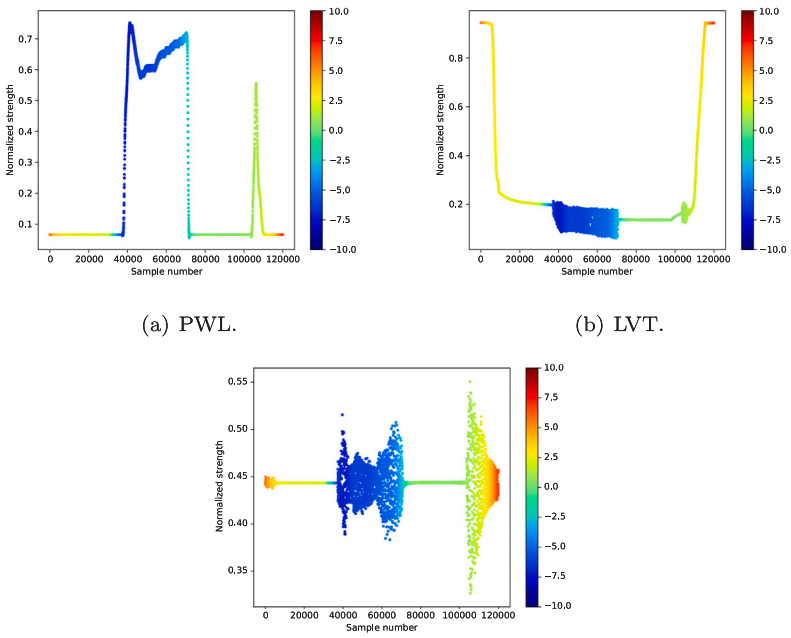
We have focused on two of the available XAI methods to explain our results on the USW data

1. class activation map (CAM), and
2. contrastive gradient-based saliency maps.

These are instrumental in explaining "black-box" learning methods used in computer vision, quite like the one used in Section 2. The objective is to understand the inner works of a connectionist model, which is otherwise opaque to the user. In the following, we would demonstrate how to adopt these methods for USW domain, which entails time-series data as opposed to image.

3.1. Interpretable model with class activation maps

The CAM method [33] works on convolutional neural networks whose last two layers are a global average pooling (GAP) layer followed by a single fully convolutional layer. The input was S as described in Section 2.



(a) PWL. (b) LVT.

(c) ASO.

Fig. 3. Signals from a good weld with classifier score 0.399.

Let us consider the final classification into

$$c = \begin{cases} 0, & \text{good quality} \\ 1, & \text{otherwise} \end{cases}$$

is obtained by thresholding the last layer activations $a_k^{[n_h]} > 0.5$. For any input S_x^{te} $\forall x \in \{1, \ldots, 2400\}$, CAM is calculated by combining Eqs. (1) and (2), CAM is calculated as

$$M_c^{\text{te}}(x)(\text{x}, \text{y}) \triangleq \sum_{t=1}^{n_{h-2}} w_t^{[n_{h-1}]} a_t^{[n_{h-2}]}(\text{x}, \text{y}). \tag{3}$$

Note that $n_{h-2} = n_{h-1}$ is an inherent constraint for this architecture. When features that correspond to the selected class are found at a particular location of x, these features $a_k(\text{x}, \cdot)$ 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 $S_x^{\text{te}}(\text{x}, \cdot) = [s_{\text{p}}^{\text{te}}(\text{x})\, s_{\text{v}}^{\text{te}}(\text{x})\, s_{\text{a}}^{\text{te}}(\text{x})]_x$ 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 weld.

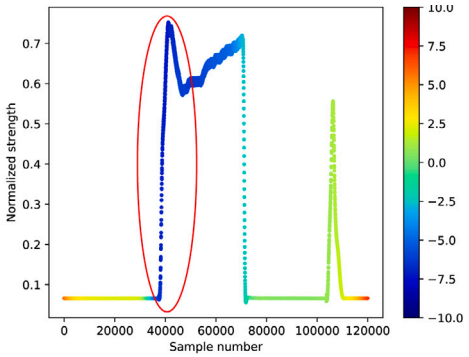
For the USW welding quality prediction CNN as reported here, we applied the CAM method. Fig. 3 presents the output of the CAM method when a good weld is provided as input to the classifier. The figure shows the representation of low activation scores corresponding to the "good" class in blue, while the red colored graph represents high activation values corresponding to the "bad" class. A good weld was defined to have a score of 0, while a bad weld 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 as "good" weld; on the contrary, red color will indicate a "bad" weld.

Fig. 4 shows a comparison of results for good and bad welds. 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 weld is going to be associated with a good or bad quality.

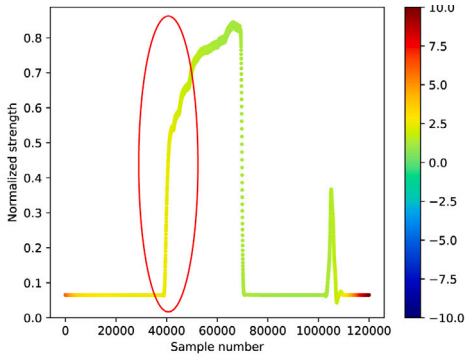
Fig. 4 details two different weld. For each weld, we visualize three signals, which are aligned in time (i.e. the x-axis is the same for the three signals). In each, the shape corresponds to each signal, while the color corresponds to the weighted activation of the last layer in the network (the output of the CAM method). Hence, we can interpret the good quality of a weld by observing a blue colored peak in the PWL signal in Fig. 4(a). Another example is the green colored bell shape of the acoustic signal in Fig. 4(f), indicating a weld with bad quality.

Through visualization, CAM explains the importance of the input signals for good or bad quality predictions. We showed that the first part of the PWL signal determines the quality prediction. "Good" welds present blue-toned colors while "bad" welds get yellow to red colors. Patterns for visualizing good/bad quality predictions were also noted in the ASO signal.

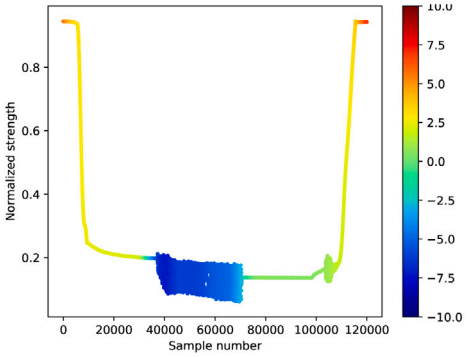
CAM is effective in determining features corresponding to each class (good/bad). An immediate result we observe from implementing CAM is that the quality of the classifier can be explained by looking at the speed at which the PWL steps up: when the PWL shows a strong step gradient, we can predict a good weld; but when the visualization shows a weak step gradient, we can predict a bad weld. We also see that the acoustic signal shape serves as another explanation for the classifier prediction. When this signal has a clear bell shape pattern, this is an explanation for bad welds. 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.



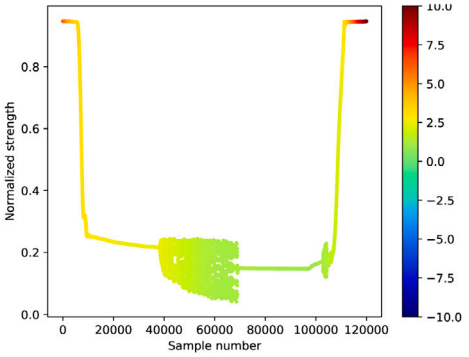
(a) Good PWL with classifier score 0.399.



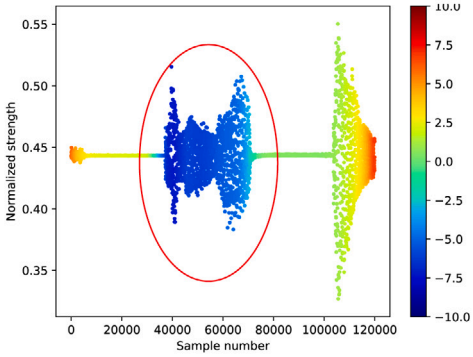
(b) Bad PWL with classifier score 0.920.



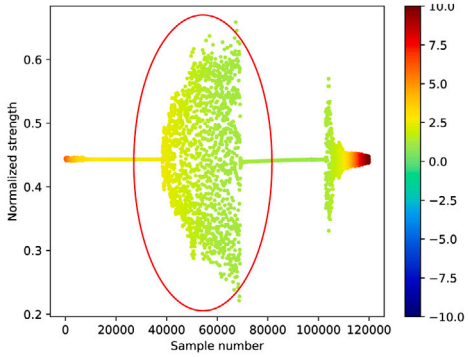
(c) Good LVT with classifier score 0.399.



(d) Bad LVT with classifier score 0.920.



(e) Good ASO with classifier score 0.399.



(f) Bad ASO with classifier score 0.920.

Fig. 4. Comparing signals from good and bad welds.

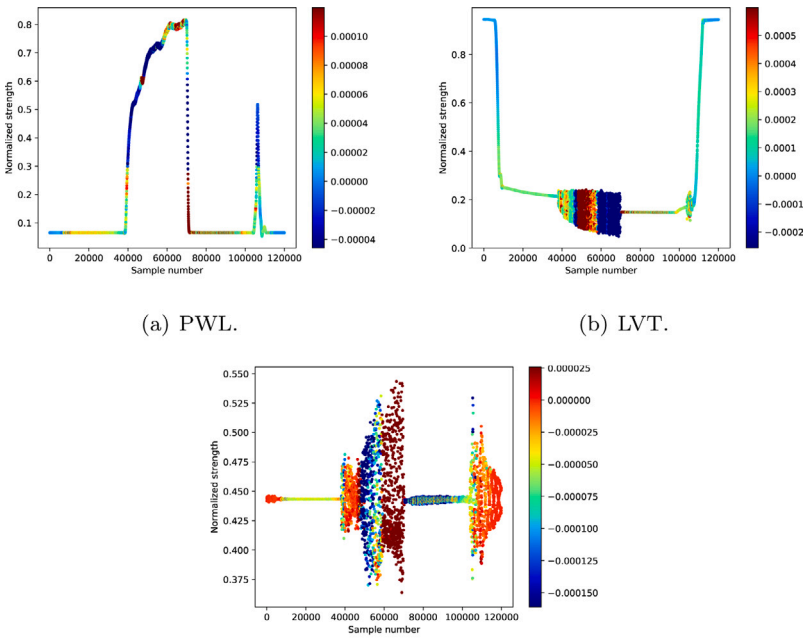
3.2. Robust model with contrastive gradient-based saliency maps

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 [34]. This method calculates the saliency map for a specific input S_x^{te} and class c. The map is calculated as the gradient of neural network's output $M_c^{\text{te}}(x)$ with respect to the input. For a traditional neural network, this gradient is calculated by backpropagation,

$$\dot{w}_k^{[n_h-1]} = \left.\frac{\partial M_c^{\text{te}}}{\partial S}\right|_{S_x^{\text{te}}}. \tag{4}$$

The value that a neural network outputs for a given class changes smoothly as a function of its inputs. In our case, the inputs are the three signals collected during the welding process (PWL, LVT, ASO) and the output for each class (good quality welding" or "bad quality welding") represents the probability that the class is the true label. Thus, as we slowly change the values in the input signals we expect to change the output in the neural network, even shifting from one class



(a) PWL. (b) LVT.

(c) ASO.

Fig. 5. Contrast gradient-based saliency map with classifier score 0.695.

to the other. By definition, the rate of change in the output with respect to the input is the gradient of the output with respect to the input and it represents how much will the output change with a small change in the input. In other words, these gradients represent how stable the classifier is. It is important to note that this gradient is calculated during the training process of the network, which is usually achieved through backpropagation of the gradients from the last layer of the network to the first (input) layer. Fig. 5 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 Fig. 6). Keeping activations stable as the input signal changes slowly might be an indication of a stable classifier. The network inputs are presented in pink, green and orange colors respectively. The graph in blue represents the activations of the nodes at the layer represented by the particular graph (e.g., Fig. 6(c)). As we dive deeper into the network, towards the deeper layers, the activations appear less noisy and better aligned with the informative features of the input signals (e.g., a peak is noticed in Fig. 6(l) where a peak is also noticed in the PWL pink signal).

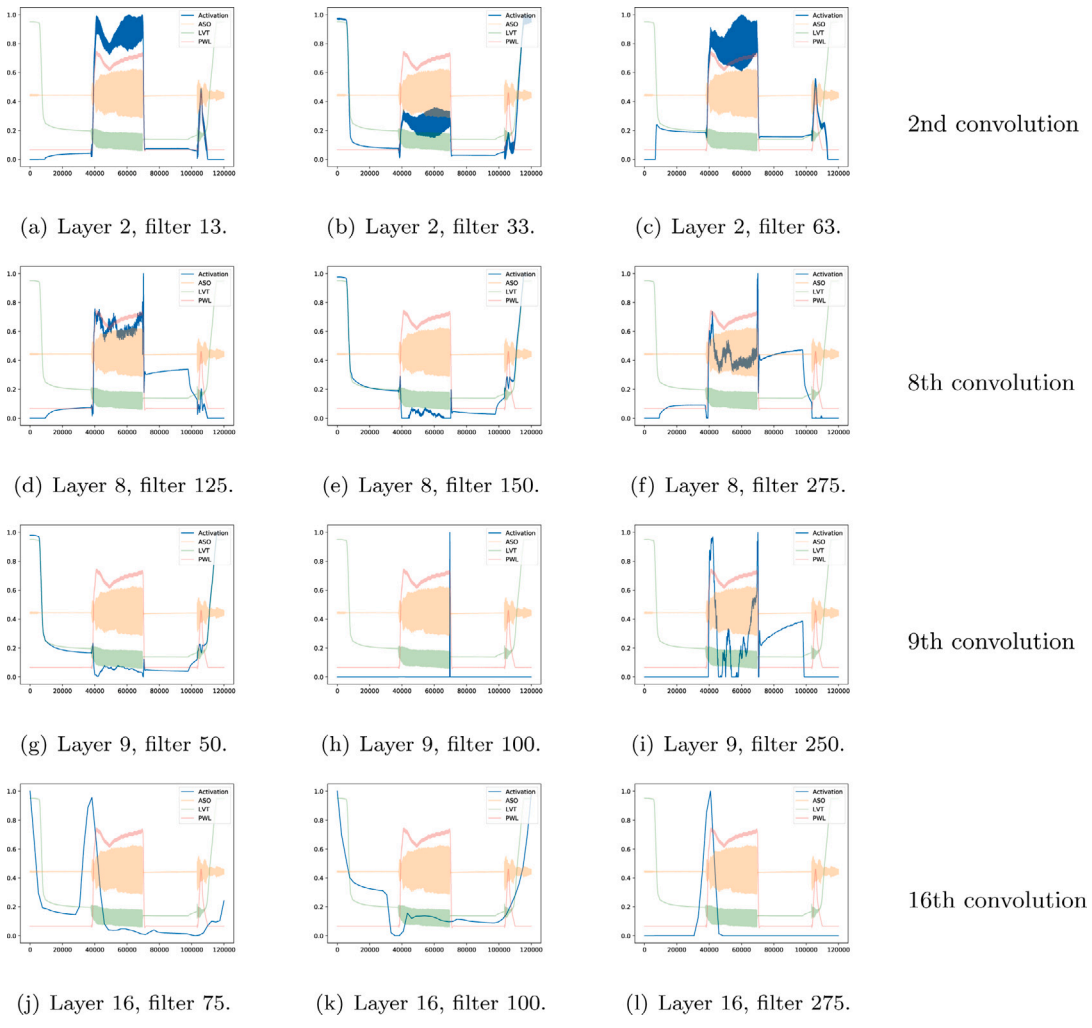
An interesting question is whether we can design a filter that is connected to the physics of the process or whether we can find the physics in the filters. This task remains for future work. For a learning model to be easy to explain, we would like it to be stable, i.e., we want small gradients through 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. Interpreting BIW (Body-In-White) dimension prediction

During the manufacturing process of the BIW, the underbody is staged in a holding fixture, Fig. 7(a). Sub-assemblies such as the engine compartment and floor pan are then clamped and welded together to produce the underbody. As the underbody further travels across various downstream stations in the body shop, inner and outer upper structures are welded in a framer station to produce the fully framed BIW shown in Fig. 7(b).

A three dimensional vision system scans the underbody and the framer to produce dimensional data $\boldsymbol{u}, \boldsymbol{f} \in \mathbb{R}^3$, respectively. Each dimension of the observed data $\boldsymbol{u}, \boldsymbol{f}$ are monitored using several univariate control charts (see Fig. 8) to identify out of tolerance dimensional measurement.

This traditional approach fails to establish inter-dimensional dependencies towards quality issues, and the degree of significance an underbody dimensional point $\boldsymbol{u}_i$ may have on a framer dimensional point $\boldsymbol{f}_j$ (for any $i, j \in \mathbb{N}$). In fact, till now there has not been a mechanism to predict if a framer point would be dimensionally accurate based on the accuracy of the underbody points. Using over five thousand BIW from a specific vehicle model, we have built a neural network based system to predict the accuracy of the framer points given the underbody



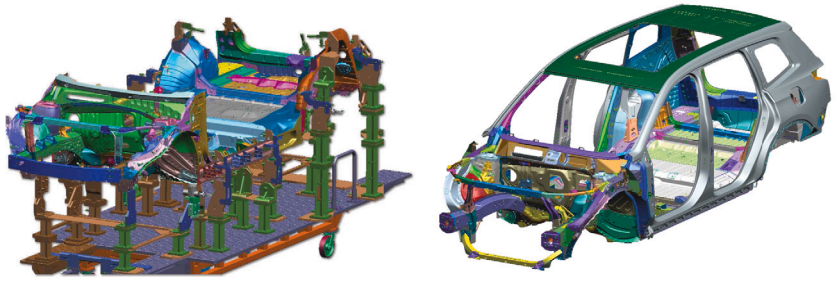
(a) Layer 2, filter 13. (b) Layer 2, filter 33. (c) Layer 2, filter 63.

(d) Layer 8, filter 125. (e) Layer 8, filter 150. (f) Layer 8, filter 275.

(g) Layer 9, filter 50. (h) Layer 9, filter 100. (i) Layer 9, filter 250.

(j) Layer 16, filter 75. (k) Layer 16, filter 100. (l) Layer 16, filter 275.

Fig. 6. Signal activations $a_k^{[h]}$ throughout the network, for layer h and filter k.



(a) Underbody mounted on a holding fixture. (b) Fully framed BIW.

Fig. 7. Body-in-white process.

points. Though the mechanics of this method are out of scope here, the relevance of the intermediate stages and outcome is the main focus.

Each underbody point $\boldsymbol{u}_i$ measured has a nominal value based on design $\tilde{\boldsymbol{u}}_i$, and a corresponding deviation from nominal $\boldsymbol{\delta}_{u_i} = \boldsymbol{u}_i - \tilde{\boldsymbol{u}}_i$, $\boldsymbol{\delta} \in \mathbb{R}^3$. Similarly, each measured point after the framer station $\boldsymbol{f}_j$ is also associated with its deviation from nominal $\boldsymbol{\delta}_{f_j} = \boldsymbol{f}_j - \tilde{\boldsymbol{f}}_j$. Each deviation $\boldsymbol{\delta}$ is defined by the deviation in x-, y-, and z- direction of the cartesian coordinate system, $\boldsymbol{\delta} = (\delta_{\mathbf{X}}, \delta_{\mathbf{Y}}, \delta_{\mathbf{Z}}) | \delta_{\mathbf{X}}, \delta_{\mathbf{Y}}, \delta_{\mathbf{Z}} \in \mathbb{R}$. Obviously, not all $\boldsymbol{\delta}_u$ are relevant, in predicting any $\boldsymbol{\delta}_f$, and would make the learning task over-complex. The RReliefF algorithm [35] was used to select the relevant ones.

Every underbody deviation $\mathfrak{d}_k \in \{\delta_{u_{i_{\mathbf{X}}}}, \delta_{u_{i_{\mathbf{Y}}}}, \delta_{u_{i_{\mathbf{Z}}}}, \forall i\}$ from a set $\mathbb{D} = \{\mathfrak{d}_k | k = 1, \ldots, m\}$ of all m underbody deviations was assigned a rank $r \in \mathbb{N}$ using the RRelieFF algorithm for every framer deviation $\mathfrak{f}_l \in \{\delta_{f_{j_{\mathbf{X}}}}, \delta_{f_{j_{\mathbf{Y}}}}, \delta_{f_{j_{\mathbf{Z}}}} \forall j\}$ from a set $\mathbb{F} = \{\mathfrak{f}_l | l = 1, \ldots, n\}$ of n framer deviation. The outcome was a matrix of ranks $\boldsymbol{R} \in \mathbb{R}^{m \times n}$, the values of which are shown in Fig. 9. The best rank is $r = 1$. Each row in $\boldsymbol{R}$ contains all the ranks of underbody point measurement dimensions corresponding to a framer point measurement dimension. The overall

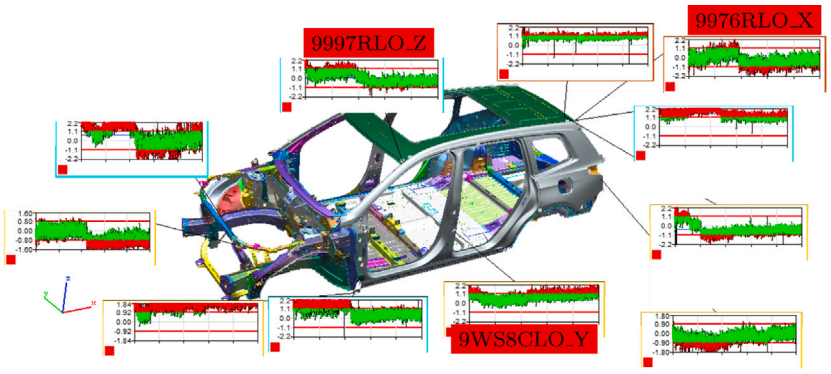


Fig. 8. Framer control charts (similar analyses are generated for underbody data).

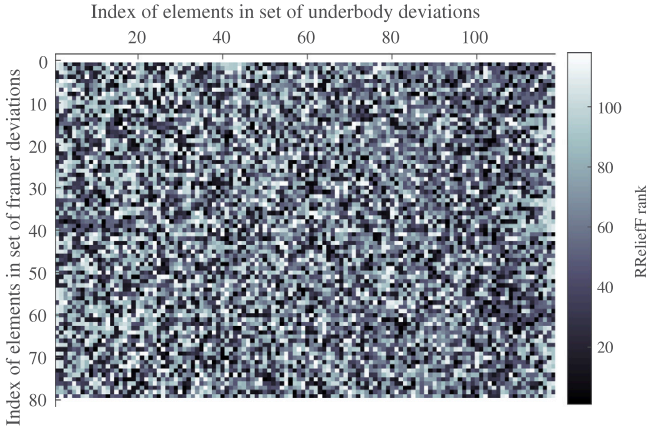


Fig. 9. Ranking of every element in $\mathbb{D}$ for every element in $\mathbb{F}$.

irrelevance[1] γ of an underbody point $\boldsymbol{u}_i$ to any framer point $\boldsymbol{f}_j$ is defined as the floored average of 2-norm of 3-dimensional ranks

$$\gamma_{i,j} \triangleq \left\lfloor C_1 \frac{1}{3} \sum_k \sqrt{\sum_l R(k\Delta_{j,k}, l\Delta_{i,l})^2} + C_0 \right\rfloor, \tag{5}$$

where Δ is a variant of Kronecker delta function such that the index k could pick all the dimensions of $\boldsymbol{f}_j$ and l could pick all the dimensions of $\boldsymbol{u}_i$. Here $C_1, C_0 \in \mathbb{R}$ are constants such that $\gamma \in \mathbb{N} \cap [1, 256]$. So, the smaller the irrelevance, the more relevant the point. In this dataset, not all x-, y-, z-directions were observed for every point. The missing dimensions were filled in as an average of the existing dimensions. For example, if for the jth framer point $u_{3_\mathbf{X}}$ (and thereby the corresponding rank $r_{3_\mathbf{X}}|j$) was the only dimension missing, then

$$r_{3_\mathbf{X}}|j = \frac{1}{2}(r_{3_\mathbf{Y}}|j + r_{3_\mathbf{Z}}|j)$$

was used. However, if two of the dimensions were missing, the available third dimension value was simply replicated.

Plotting these γ values in the true locations of all $\boldsymbol{u}$ renders the otherwise chaotic data from Fig. 9 meaningful. Let us pick three framer points from the control chart shown in Fig. 8, `9997RLO` from the driver side front roof corner, `9976RLO` the driver side back roof corner, and `9WS8CLO`. These points are demonstrated in Fig. 10(a). The irrelevance of the underbody points for these framer points are shown in Fig. 10. Figs. 10(b) and 10(c) show the physical orientation of some of the key underbody points. Some points (in red) from Fig. 10(b) and both the points from Fig. 10(c) are shown in the irrelevance plots using gray text. These points are common in all the irrelevance plots. The irrelevance points in black text are the top 5 relevant points, unique in each irrelevance plot. While Figs. 10(d)–10(f) shows the values calculated for three specific point according to Eq. (5), Fig. 10(g) is a global irrelevance of all the underbody points calculated by averaging the irrelevance over all points

$$\gamma_{i,\cdot} = \left\lfloor C_1 \frac{1}{3} \sum_k \sqrt{\sum_l R(k, l\Delta_{i,l})^2} + C_0 \right\rfloor.$$

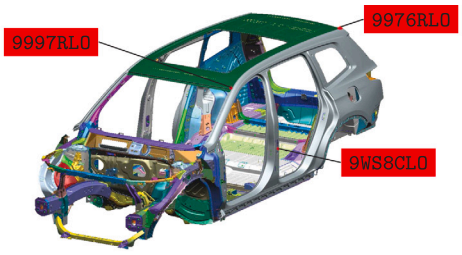
The orientation of the irrelevance plots are closely matched to the graphics of the BIW to draw visual relevance, which is the engineering insight we seek.

The obtained engineering relevance of the points provided the confidence to predict $\boldsymbol{\delta}_{f_j}$ from $\boldsymbol{\delta}_u$. Several neural networks were built to estimate each $\mathfrak{f}_l$ as $\hat{\mathfrak{f}}_l$ using a subset of the points $\hat{\mathbb{D}} \subset \mathbb{D}$. The first network used only the $\boldsymbol{\delta}_u$ corresponding to the best ranking underbody points, i.e. $\hat{\mathbb{D}}_1 = \mathbb{D}|r \in \{1\}$. The next network included $\hat{\mathbb{D}}_2 = \mathbb{D}|r \in \{1, 2\}$, then $\hat{\mathbb{D}}_3 = \mathbb{D}|r \in \{1, 2, 3\}$ and so on. Each network was a fully connected feed forward neural network with sigmoid activation on the only hidden layer, designed to predict any $\boldsymbol{\delta}_{f_j}$. A conceptual representation of a network is presented in Fig. 11, albeit the neural network structure is for demonstration purpose and does not represent the true number of nodes. The output of the network was then used to reconstruct all the predicted framer point dimensions $\{\hat{\boldsymbol{f}}_j = (\hat{f}_{j_\mathbf{X}}, \hat{f}_{j_\mathbf{Y}}, \hat{f}_{j_\mathbf{Z}}) \forall j\}$.

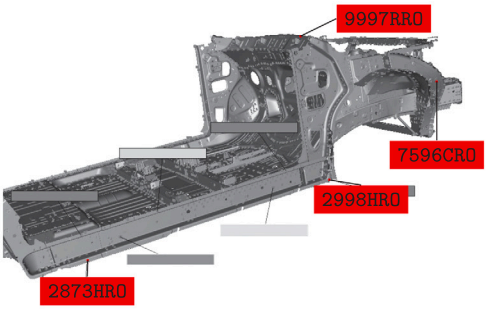
A 3-dimensional criterion [36] comprising Pearson's correlation coefficient $\rho \in [0, 1]$ and mean square error between each dimension of $\hat{\boldsymbol{f}}$ and $\boldsymbol{f}$, and number of input nodes (cardinality of $\hat{\mathbb{D}}$), was constructed. Occam's razor principle is implemented by penalizing solutions with higher number of input nodes.

In total, there were 9099 BIW that were used; 5459 were used for training, 1820 for validation, and the remaining 1820 for testing. The measure of success was ρ from the testing set. The outcome of each of

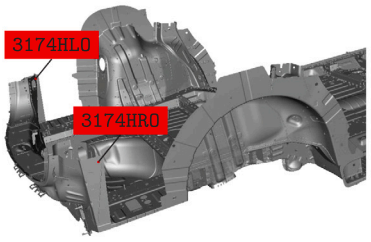
[1] The smaller the irrelevance, the more relevant the point is



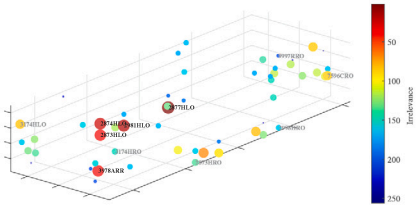
(a) Framer points 9997RLO, 9976RLO, and 9WS8CLO.



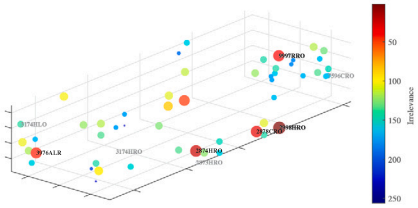
(b) Selective physical location of some underbody points.



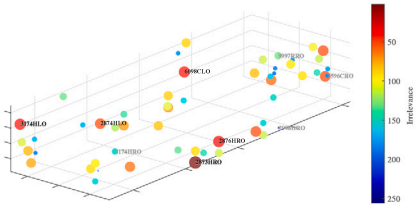
(c) Physical location of some rear underbody points.



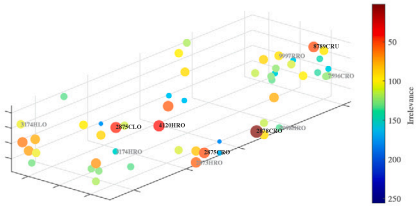
(d) Irrelevance for 9997RLO.



(e) Irrelevance for 9976RLO.



(f) Irrelevance for 9WS8CLO.



(g) Overall irrelevance.

Fig. 10. Irrelevance values in true physical space.

Table 2
Prediction performance on 79 framer points.

ρ	# of framer points
(0.9, 1.0]	7
(0.8, 0.9]	10
(0.7, 0.8]	13
[0.0, 0.7]	49

$n = 79$ networks, one for each $\mathfrak{f}_l$, is shown in Fig. 12 and the result is summarized in Table 2. The overall performance was fair with room for improvement.

5. Conclusion

The work here has demonstrated how XAI methods can help clarify otherwise "black box" methods to provide engineering relevance to machine learning results. Two applications were discussed here. One was ultrasonic welding quality prediction based on process signals, and the other was dimensional quality prediction of body-in-white framer points based on deviations in underbody points.

Real-time temporal signals from ultrasonic welding process of battery cell tabs were collected and modeled using convolutional neural network. The learned parameters of the network could then be used to derive engineering insight that was not otherwise obvious. Class activation maps and gradient based saliency maps techniques, that have been successful in vision applications, were adopted and applied to

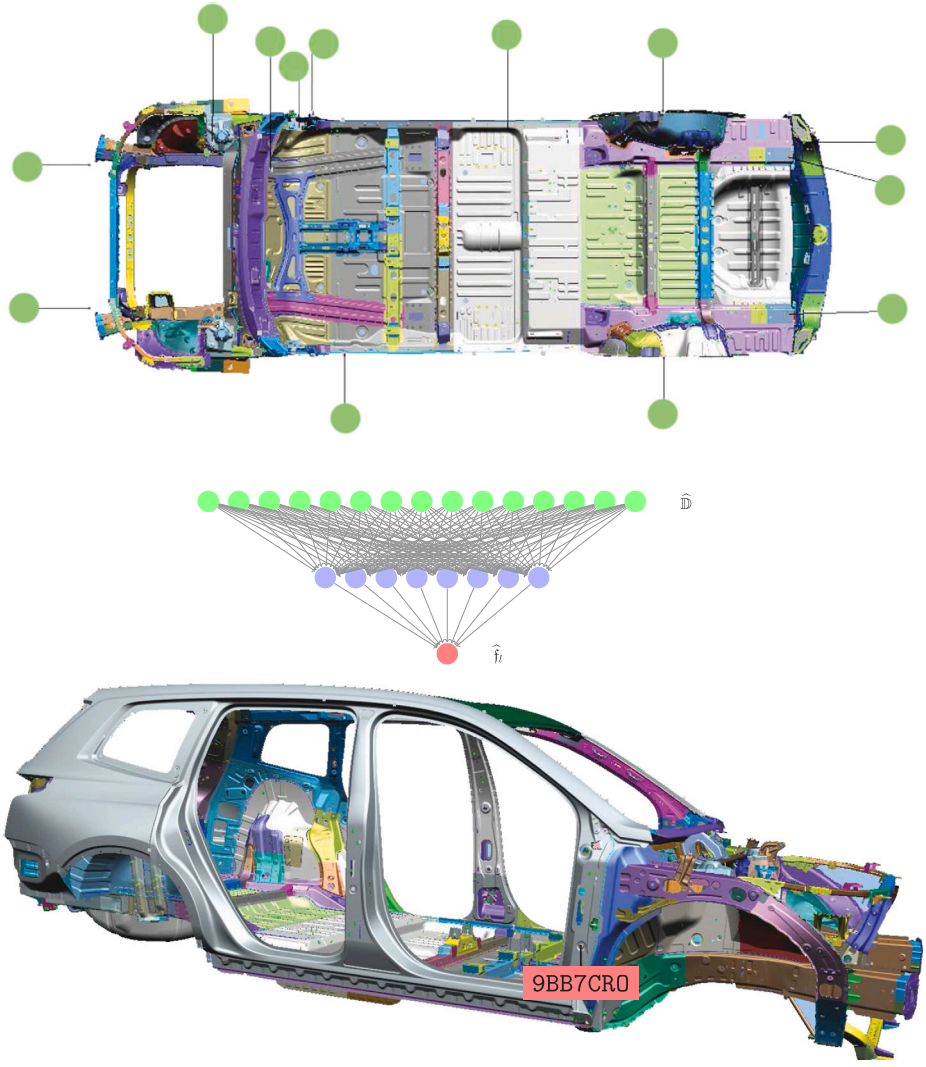


Fig. 11. Mapping tolerance in underbody points to a framer point 9BB7CR0.

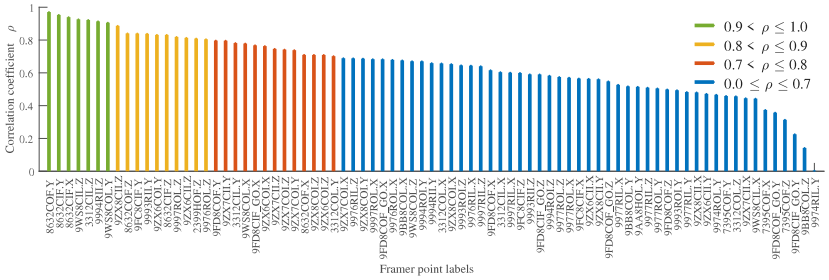


Fig. 12. Prediction results of each framer coordinate on the test set.

multichannel time series data. It could be inferred that quality of the final weld is indicated by the initial engagement pressure. A lower initial engagement pressure would result in an inadequate weld. A recommended next step is to apply the excitation backpropagation method to explain how intermediate layers capture information about the learning process [37]. Other related methods appear in [38].

The body-in-white application deployed feature selection followed by shallow neural network modeling. Deviations in selective underbody points were used to develop a causality relationship to specific framer points. The engineering interpretation was mainly to understand which underbody points had more influence on deviations in the framer points. Such interpretations could aid manufacturing operations in drawing focus to critical underbody points and could also be a feedback to design if a tighter tolerance to critical underbody points are not achievable in manufacturing.

The focus here has been on the explainability. The classifiers were developed to demonstrate explainability. In fact, this paper solicited the use of an interpretability score along with the accuracy as the right metric to select a classifier. Though these classifiers could be further improved through optimization of hyperparameters and further

training, the interpretation obtained was insightful and inline with general engineering understanding. Having a higher AI quotient is more valuable than pure classification accuracy. The authors believe that enriching AI through the development of such insights would make these methods significantly adoptable in the manufacturing industry.

As it is noticed in the surveys on industrial information integration, the processes in smart manufacturing are increasing in size and in coverage across many different disciplines (such as the sciences, automated factories, healthcare and transportation). Whenever, these processes include human interaction with manufacturing processes, it will be important for these automation solutions to interact transparently with their end users so as to build trustful interactions for the benefit of the effectiveness of these processes and outputs. Future work will focus on new manufacturing processes where learned classifiers (or other deep learning methods) can provide valuable insights. In parallel, new explainable AI methods could be applied as new domains start applying them to improve the human machine interactions.

CRediT authorship contribution statement

Claudia V. Goldman: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Project administration. **Michael Baltaxe:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Debejyo Chakraborty:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Resources, Writing – original draft, Visualization. **Jorge Arinez:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Project administration. **Carlos Escobar Diaz:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Resources, Writing – original draft, Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

The authors would like to thank Mitchell Washer, Lee Grimaldi, Jovan Keca, James Turner, and Greg Nirmalakumar from GM Manufacturing Engineering, and Michael Wincek from GM R&D for their input and support to this work.

References

[1] Y. Chen, Industrial information integration–A literature review 2006–2015, J. Ind. Inf. Integr. 2 (1) (2016) 30–64, http://dx.doi.org/10.1016/j.jii.2016.04.004.

[2] Y. Chen, A survey on industrial information integration 2016–2019, J. Ind. Integr. Manage. 5 (1) (2020) 33–163, http://dx.doi.org/10.1142/S2424862219500167.

[3] L. Zhao, Z. Fan, W. Li, H. Xie, Y. Xiao, 3D indoor map building with monte carlo localization in 2D map, in: 2016 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII), IEEE, 2016, pp. 236–240.

[4] M.B. Imani, M. Heydarzadeh, L. Khan, M. Nourani, A scalable spark-based fault diagnosis platform for gearbox fault diagnosis in wind farms, in: 2017 IEEE International Conference on Information Reuse and Integration (IRI), IEEE, 2017, pp. 100–107.

[5] Z. Huo, M. Mukherjee, L. Shu, Y. Chen, Z. Zhou, Cloud-based data-intensive framework towards fault diagnosis in large-scale petrochemical plants, in: 2016 International Wireless Communications and Mobile Computing Conference (IWCMC), IEEE, 2016, pp. 1080–1085.

[6] W. Glover, Q. Li, E. Naveh, M. Gross, Improving quality of care through integration in a hospital setting: A human systems integration approach, IEEE Trans. Eng. Manage. 64 (3) (2017) 365–376.

[7] What's now and next in analytics, AI, and automation, 2017, URL https://www.mckinsey.com/featured-insights/digital-disruption/whats-now-and-next-in-analytics-ai-and-automation.

[8] J. Bughin, E. Hazan, S. Lund, P. Dahlström, A. Wiesinger, A. Subramaniam, Skill shift: Automation and the future of the workforce, McKinsey Global Institute, 2018, URL https://www.mckinsey.com/featured-insights/future-of-work/skill-shift-automation-and-the-future-of-the-workforce.

[9] N. Henke, J. Bughin, M. Chui, J. Manyika, T. Saleh, B. Wiseman, G. Sethupathy, The age of analytics: Competing in a data-driven world, McKinsey Global Institute, 2016, URL https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world.

[10] D. Chakraborty, M.E. McGovern, NDE 4.0: Smart NDE, in: 2019 IEEE International Conference on Prognostics and Health Management (ICPHM), IEEE, 2019, pp. 1–8.

[11] G. Kronberger, F. Bachinger, M. Affenzeller, Smart manufacturing and continuous improvement and adaptation of predictive models, Procedia Manuf. 42 (2020) 528–531, International Conference on Industry 4.0 and Smart Manufacturing (ISM 2019).

[12] O.V. Vagan Terziyan, Explainable AI for industry 4.0: Semantic representation of deep learning, in: 3rd International Conference on Industry 4.0 and Smart Manufacturing, 2022.

[13] S.B. Mir Riyanul Islam, S. Begum, A systematic review of explainable artificial intelligence in terms of different application domains and tasks, Appl. Sci. 12 (3) (2022) 1353, http://dx.doi.org/10.3390/app12031353.

[14] G. Sofianidis, J.M. Rovzanec, D. Mladenić, D. Kyriazis, A review of explainable artificial intelligence in manufacturing, 2021, arXiv preprint arXiv:2107.02295.

[15] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell. 2 (1) (2020) 56–67.

[16] D. Schubmehl, D. Vesset, The Knowledge Quotient: Unlocking the Hidden Value of Information Using Search and Content Analytics, White Paper, International Data Corporation, 2014.

[17] N. Pal, S. Sundaresan, J. Ray, H. Bhargava, E. Glantz, M.W. McHugh, Knowledge Quotient™ (KQ): A Way to Measure the Knowledge Intensity of Your Team, Tech. Rep., The Penn State eBusiness Research Center, 2004.

[18] G. Klein, AIQ: Artificial intelligence quotient, 2020, URL https://www.psychologytoday.com/us/blog/seeing-what-others-dont/202007/aiq-artificial-intelligence-quotient.

[19] R.R. Hoffman, S.T. Mueller, G. Klein, J. Litman, Metrics for Explainable AI: Challenges and Prospects, Tech. Rep., Explainable AI Program, DARPA, 2018, URL https://arxiv.org/abs/1812.04608.

[20] G. Klein, R.R. Hoffman, S.T. Mueller, Scorecard for Self-Explaining Capabilities of AI Systems, Tech. Rep., Explainable AI Program, DARPA, 2020.

[21] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.

[22] A. Rosenfeld, A. Richardson, Why, who, what, when and how about explainability in human-agent systems, in: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '20, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2020, pp. 2161—2164.

[23] J.F. Fisac, C. Liu, J.B. Hamrick, S. Sastry, J.K. Hedrick, T.L. Griffiths, A.D. Dragan, Algorithmic foundations of robotics XII, in: K. Goldberg, P. Abbeel, K. Bekris, L. Miller (Eds.), Algorithmic Foundations of Robotics XII: Proceedings of the Twelfth Workshop on the Algorithmic Foundations of Robotics, Springer International Publishing, 2020, pp. 144–159, http://dx.doi.org/10.1007/978-3-030-43089-4_10.

[24] S.H. Huang, D. Held, P. Abbeel, A.D. Dragan, Enabling robots to communicate their objectives, Auton. Robots 43 (2) (2019) 309–326.

[25] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (5) (2019) 206–215.

[26] J. Kim, J. Canny, Interpretable learning for self-driving cars by visualizing causal attention, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969, http://dx.doi.org/10.1109/ICCV.2017.320.

[27] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, in: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2018, pp. 80–89.

[28] A. Kulkarni, Y. Zha, T. Chakraborti, S.G. Vadlamudi, Y. Zhang, S. Kambhampati, Explicable planning as minimizing distance from expected behavior, in: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2019, pp. 2075–2077.

[29] S. Sreedharan, A.O. Hernandez, A.P. Mishra, S. Kambhampati, Model-free model reconciliation, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 587–594, http://dx.doi.org/10.24963/ijcai.2019/83.

[30] C.V. Goldman, M. Baltaxe, D. Chakraborty, J. Arinez, Explaining learning models in manufacturing processes, Procedia Comput. Sci. 180 (2021) 259–268, http://dx.doi.org/10.1016/j.procs.2021.01.163, URL https://www.sciencedirect.com/science/article/pii/S1877050921002039, Proceedings of the 2nd International Conference on Industry 4.0 and Smart Manufacturing (ISM 2020).

[31] J.A. Abell, D. Chakraborty, C.A. Escobar, K.H. Im, D.M. Wegner, M.A. Wincek, Big data driven manufacturing — Process-monitoring-for-quality philosophy, J. Manuf. Sci. Eng. 139 (10) (2017) 101009–1–101009–12, http://dx.doi.org/10.1115/1.4036833.

[32] S.M. Kay, Fundamentals of Statistical Signal Processing Volume II Detection Theory, Prentice Hall PTR, 1998,

[33] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929, http://dx.doi.org/10.1109/CVPR.2016.319.

[34] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, in: Workshop At International Conference on Learning Representations, 2014.

[35] M. Robnik-vSikonja, I. Kononenko, Theoretical and Empirical Analysis of ReliefF and RReliefF, Mach. Learn. 53 (1) (2003) 23–69.

[36] C.A. Escobar, R. Morales-Menendez, Process-monitoring-for-quality–A model selection criterion for shallow neural networks, in: Annual Conference of the PHM Society, 11, 2019.

[37] J. Zhang, S.A. Bargal, Z. Lin, J. Brandt, X. Shen, S. Sclaroff, Top-down neural attention by excitation backprop, Int. J. Comput. Vis. 126 (10) (2018) 1084–1102, http://dx.doi.org/10.1007/s11263-017-1059-x.

[38] T.N. Mundhenk, B.Y. Chen, G. Friedland, Efficient saliency maps for explainable AI, 2020, arXiv:1911.11293.