Textile Surface Defects Analysis with Explainable AI

Ren Jun Soon
Department of Mechanical Engineering
National University of Singapore
Singapore
soonrenjun@u.nus.edu

Chee-Kong Chui
Department of Mechanical Engineering
National University of Singapore
Singapore
mpecck@nus.edu.sg

***Abstract*— The identification of surface level defects, such as those in manufacturing materials, is crucial in industrial processes. Deep learning methods have established their efficacy in defect detection. However, these black-box classification methods lack transparency, obscuring the rationale behind their classifications. This obscurity becomes a significant concern in safety-critical situations. Consequently, integrating explanatory mechanisms into these systems' classification results is essential. This paper proposes a novel defect detection system based on Explainable Artificial Intelligence (XAI). A conventional Convolutional Neural Network (CNN) is employed to process an image database of fabrics, showcasing various manufacturing defects. This CNN, achieving a classification accuracy exceeding 93%, subsequently undergoes an interpretive analysis. To elucidate the CNN's output, a statistical method grounded in the SHAP (SHapley Additive exPlanations) value corresponding to a feature in the examined image is employed. Additionally, an alternative explanatory approach utilizing an unsupervised neural network is explored. This entails the use of Self-Organizing Maps (SOMs) to classify images in the dataset, facilitating visualization of the data categorization.**

Keywords—explainable artificial intelligence, SHAP, style, self-organizing map

I. INTRODUCTION

The use of deep learning techniques like convolutional neural networks to classify manufacturing outputs has been well established. However, the interpretability of these systems' outputs remains a challenge, especially in safety-critical applications. The uncertainty arising from the lack of logic and trust in the outputs affects the usability of outputs of such classification systems. It is therefore necessary to explain the analysis of such outputs. One approach involves the use of Explainable AI (XAI), where deep learning outputs are explained using a variety of quantitative and qualitative metrics. This approach builds trust in the output of a deep learning classification system, but there are inherent accuracy limitations given how the explanations themselves are also derived or based on some form of machine or deep learning. In this paper, a XAI approach based on using SHAP values to interpret deep learning outputs is proposed. This is complemented with the use of Self-Organizing Maps (SOMs) that characterize the deep learning outputs, providing a visual interpretation to the output of a deep learning model, further helping in the understanding of the features contributing to an output.

For illustration purpose, we use the AITEX fabric dataset [1] which includes textiles classified as either defective or defect-free. Our deep learning analysis and subsequent interpretation will revolve around the classification of whether a textile has a defect or is defect-free, demonstrating the efficacy of our proposed method in a practical scenario.

II. LITERATURE REVIEW

XAI can be understood as making deep learning interpretable. Deep learning outcomes are given interpretations to allow an explanation of the deep learning outputs to be given [1]. XAI increases trust in deep learning outputs by providing the explanation behind the output of a deep learning output. There is real world value in doing so. Today, there are exploratory approaches towards the use of XAI in medicine to explain medical images, respiratory and cardiological data for diagnoses [2], [3]. This thereby reduces the reliance on human classification, as well as provide basis for insights that are not discernible by humans. In manufacturing and maintenance, XAI has been used in the justification of the predictive maintenance of critical equipment, process optimization, and risk analysis [4]–[6].

Contemporary XAI technologies focus on approaches that provide either quantitative or qualitative explanations. These are achieved through rule-based or example-based analyses [1], [2], [7], [8]. Rule-based analyses use statistical or engineering methods to interpret deep learning outputs. Example-based analyses are largely black box models that provide explanations of deep learning outputs through the further use of machine learning techniques [3], [9], [10]. There are other XAI approaches that explain the output derivation process (pre hoc interpretability), as well as approaches that explain and justify the correctness of outputs (post hoc interpretability) [11]–[13]. One notable post hoc interpretability approach within XAI is the use of Shapley Additive Explanations (SHAP) values [14], [15]. Compared to other approaches that use quantifiable metrics such as the Grad-CAM, the SHAP approach is one that is model agnostic, and is not prone to being disturbed or influenced by anomalies in the training dataset [16], [17].

SOMs, though not deep learning tools, are effectively used for visualizing multi-dimensional data in a low-dimensional grid. They simplify data by clustering similar inputs, making it easier to interpret [18], [19]. Whilst SOMs, by virtue of being a single-layer neural network are not capable of learning hierarchical relationships within data, their key advantage lies in their ability to process unsupervised data. There is no need for the preparation of an extensive training dataset, unlike in most deep learning algorithms. This makes it a good candidate for the generation of global explanations of classifications made by a deep learning system, with outputs produced being able to be quickly visualized [20].

A. SHAP Values

Similar to how specific figures qualify the determination of an outcome, the use of SHAP in XAI differs from other classification approaches that offer interpretable classifications such as 'robot imagination'. SHAP focuses on defining the importance of each feature (e.g. how variations in the fabric image captured can alter the eventual

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00248

classification output), rather than assessing feature characteristics (e.g. how a fabric defect shown in an image might affect the fabric's strength), as is typically done in techniques such as robot imagination[21].

In SHAP, the equations used in cooperative game theory are used to interpret the classifications provided by a system. They explain how an overall model prediction will change as a result of interactions with other variables (or features, in the context of an image), within the dataset. The key equation used, one that derives the classic Shapley value, can be represented as such as a weighted average of all possible differences between a model trained with a feature present and a model trained with the feature withheld [15]:

$$\phi_i = \sum_{S \subseteq F\{i\}} \frac{|S|!(|F|-|S|-1)!}{F!} [f_{s\cup\{i\}}(x_{s\cup\{i\}}) - f_s(x_s)] \quad (1)$$

where,

ϕ_i Shapley value

F set of all features

S features in set s

$f_{s\cup\{i\}}$ model trained with feature present

f_s model trained with feature withheld.

From this classic Shapley value derivation, one can then proceed to determine the SHAP value. This value is used to explain the output of a deep learning system. Reference [15] defines this SHAP value derivation as:

$$\phi_{SHAP} \approx m_{yi} f(y_i - E[y_i]) \quad (2)$$

where,

ϕ_{SHAP} SHAP value

f original model

$m_{yi}f$ $\dfrac{\phi_i(f,x)}{x_j - E[x_j]}$

m_{yi} SHAP derivation algorithm multiplier/constant

$E[y_i]$ Approximation of f_{y_i} given missing inputs in f_{y_i}.

By enabling the individual consideration of the influence one feature can have on the overall model predictions, SHAP values can help uncover patterns and provide quantifiable values to explain the rationale behind a model outcome [14].

Whilst SHAP values are not able to explain why a component is deemed to be physically failing, it can statistically explain why a component is defective, and even localize the area of the defect. In other words, a key limitation of the SHAP value in explaining the rationale behind a particular explanation is how it is able to only establish a relationship between a feature and the eventual classification without contextual justification [22].

B. Self-Organising Maps

The concept of the explainable self-organizing map is built upon the foundational structure of the self-organizing map. In a typical SOM, the reaction to a model's input is generalized and visually appears as a cluster or category of functions [19], [23]. Briefly, this process can be described as such in Fig. 1.

Explainable Self-Organizing Maps (SOMs) are said to provide both global and local explanations [19], [24]. Similar to the use of SHAP values in XAI, they yield post hoc interpretations of a model outputs. This is achieved via the use of different means and requires a good appreciation of the SOM training process described Fig. 1. Reference [19] describes an approach for determining local explanations of features found in a dataset by using the training process of the SOM to explain deep learning outputs. As highlighted in [24], this training process reveals the impact of specific variables on the outcome of a deep learning model.

The potential of SOM in providing explainable global explanations of phenomena is illustrated in [25], [26]. In [25], clusters are plotted against chemical concentrations in water to provide a visualization on predicted water quality. Interestingly, [25] also provides an analysis on the use of SHAP values as a secondary analysis to determine how different chemical compositions will affect water quality. This is also replicated in the work of [27].

The visual clustering approach of the SOM provides a highly visual, yet easily understandable method of post-hoc, deep learning outputs. Moreover, the SOM complements the outputs of a SHAP-based XAI approach by corroborating the statistical relationships that a SHAP analysis will bring about.

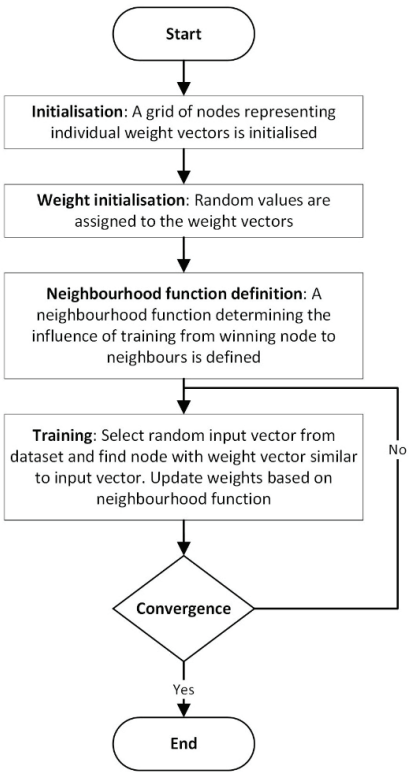

Fig. 1. Self-Organizing Map training process

[28] goes further to suggest how the visual clustering interpretability offered by SOMs can be further enhanced by

using various statistical regression techniques to analyze the classifications in the dataset. The strength of such an approach is that it provides a quantitative assessment of the rationale behind a classification.

This approach is mirrored in [29], where statistical or optimization functions are applied to the dataset points in a SOM. Similar to the statistical regression approach, the application of these functions allow the confidence level of a classification to be established arising from the quantitative values derived from the functions.

However, such statistical or mathematical approaches will require the identification of variables that will influence the classification results, thus requiring some form of supervision. Such SOM based XAI approaches thus lack appeal in how they compromise the inherent unsupervised nature of SOMs [28].

III. Experiments

Two experiments have been designed to establish the efficacy of applying both the SHAP and the explainable SOM approaches in the context of XAI. These experiments revolve around the use of either SHAP or SOMs as a means of interpreting the output of a convolutional neural network (CNN).

A. SHAP Approach

In this XAI approach, a CNN capable of performing the object classification task was customized. This CNN was established to be able to achieve an accuracy of up to 93%. The architecture used is a three-layer CNN with the input layer containing 64 filters of shape 2 x 2. The hidden layers have double the number of filters from the preceding layer, whilst the last layer, the sigmoid layer contains one filter for the binary classification of a defect having a defect or no defect. An illustration of the CNN is provided in Fig. 2.

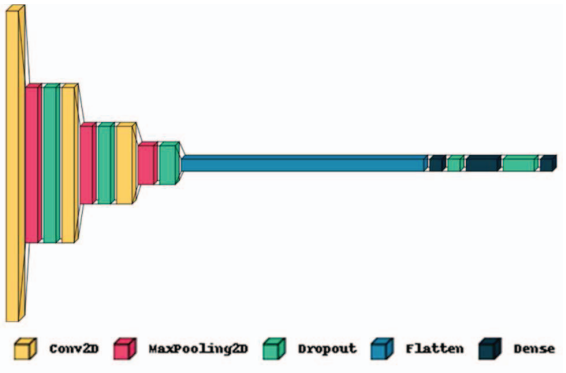

Fig. 2. CNN used for classification

Following this CNN classification, the SHAP values of the classifications were derived. These SHAP values were then superimposed onto the textile, illustrating points on the textile that either support or do not support the rationale for the classification. Such a graphical approach provides advantages over factor or rule based XAI approaches discussed in [4], [6], [16] by virtue of being able to inherently segment and isolate defective or anomalous areas of the textile that contribute to its eventual classification.

This was done by first creating a SHAP model based on the training data used in the customization of the CNN. The SHAP model was then used to provide SHAP values in the test dataset, before these SHAP values were superimposed on the original textile images. This SHAP model was derived based on Equations 1 and 2, with the SHAP values being the difference between the expected CNN model output based on the presence of features within a training dataset and the actual (ground truth) output.

Notably, in the test dataset used on this XAI model, all misclassified images were false negatives. That is, although a defect was present on the textile, the eventual classification was a defect-free classification. In these false negatives, the SHAP values obtained across the local features did not support the resulting wrongful classification. Positive SHAP values suggesting a correct classification were seen across local features of these wrongly classified textiles. Whilst contradictory, this suggests that there are underlying contextual and model influences that ultimately resulted in the wrong classification. These contextual and model influences will include the unintended aliasing of the features on the textile images as the images are being processed for the purpose of passing through the CNN.

This is an important characteristic of the SHAP approach towards interpretable classification outcomes. It serves as a secondary method of establishing the veracity of deep learning classification outputs. The SHAP values essentially quantifies the probability of the presence of a defect on the textile present on the textile, and its eventual influence of the CNN's classification. This is illustrated in Fig. 3. Red spots, indicating positive SHAP values, on the textile highlight points where the local features are identified to support the classification of the CNN. Conversely, blue spots on the textile, representing negative SHAP values, highlight points where local features do not support the classification of the CNN. Yet, there is an inadequacy in the form of how the SHAP values merely give a relative metric of probability. In other words, SHAP values across the different images within the dataset allude to how a textile sample more be more probable at being correctly classified compared to other textile samples in the model, but does not provide an absolute measure of if the textile is correctly classified.

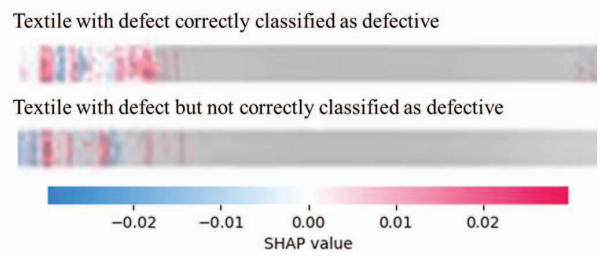

Fig. 3. SHAP values on correctly and wrongly classified textiles. The figure illustrates how an increased prominence of red spots can be linked to the increased probability of a textile being correctly classified as defective

B. Explainable SOM Approach

The presence of the false negatives despite SHAP values that do not support the classification suggests the need for a secondary approach to identify and manage explanations that are inconsistent with their explanations. The explainable

SOM approach attempts to address this by first creating a SOM of the image dataset. This SOM can classify if a textile is defective or defect free. It is represented in Fig. 4.

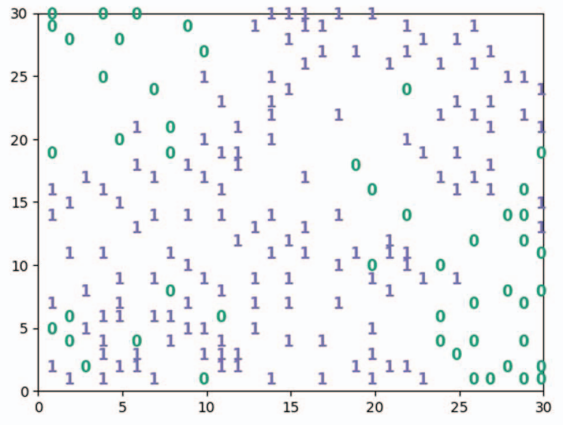

Fig. 4. SOM of AITEX textile dataset. Each number (1 or 0) represents a textile. Defect free textiles are labelled as '1'. Defective textiles are labelled as '0'.

There is global interpretability of the dataset, evident from the clusters of defective textiles centered across the top left-hand corner and bottom right-hand corner of the SOM. Supporting the false negatives seen on the CNN classification that was explained by SHAP, there were several defective textiles that resided within the no defect textile clusters. As described in [19], this is where the SOM approach in XAI can complement any CNN-based classification system by the introduction of multiple SOM plots of features within a dataset to establish the relationship between these features. By plotting and identifying the Euclidean distance between a neuron and its Best Matching Unit (BMU), across the different SOM plots, a relationship suggesting the influence of a feature within a dataset with respect to the eventual classification can be derived. In the context of images, such features can be introduced by the use of perturbations to the images (e.g., rotation, flipping). This will also provide the contextual justification that SHAP value explanations of classifications lack, given how the post data augmentation SOMs generated can provide such insights when the relationship between the actions done to augment the image and the influence on the eventual classification is established.

The approach used in [28] can also be adapted for application to boost the interpretability of the results. This was also done to quantify the strength of the classification. Briefly, following the classification of the dataset into defective and non-defective textiles, regression can be performed using a variety of variables. These variables are dependent on the SOM classification of the image, per the concept of the 'context aware representation' described in [28]. This will give an indicative quantitative confidence level to a classification. In the defect-free textile classification scenario, the number of blank pixels post thresholding, indicative of voids within the fabric, can be used as a variable to determine the strength of the classification. In the converse scenario, the strength of the image gradients, indicative of inconsistent fabric density, can used to quantify the confidence level of the classification.

The value add of such an approach will lie in how such an approach is computationally inexpensive compared to the earlier approach. Whilst it may require the presence of a human analyst to determine and derive the insights related to individual clusters and the empirical significance of various SOM based relationships, it would be free of data biases or overfitting that may be inherent in a SHAP based explanation method.

IV. Results and Discussion

The strength of both the SHAP and SOM approaches lie in how they are both able to explain and quantify the level of conviction of a classification output. The SHAP approach can perform this quantitatively, whereas the SOM approach performs this visually. The SOM approach can be augmented with means that provide an interpretable and quantitative level of conviction of a classification output.

In this, a key difference between the SHAP approach and the explainable SOM approach towards XAI lies in how the SOM approach is unable to perform semantic segmentation of images, unlike in the SHAP approach, where features that support or do not support a classification can be segmented for highlighting on the actual image.

The use of the SOM is one that is unsupervised and does not require any training datasets, unlike in the SHAP approach, where a separate dataset is required for the development of a model to generate the SHAP values required for explaining the basis of a classification. This gives the SOM approach the advantage of being able to produce explanations in small datasets.

Whilst the use of various statistical methods to quantify a SOMs classification adds interpretability to the classification, there is less automation in this XAI approach. This is given how individual variables need to be identified. Secondary means of establishing the values of these variables will also need development. Such development is context dependent. Thus, specific customization will need to be performed for every implementation of such an approach.

Through the process of identifying and isolating features trained within a particular dataset to derive the SHAP value, SHAP based XAI models are inherently more accurate when training datasets are larger. The use of SOMs as a method of explaining deep learning outputs is thus attractive when it complements SHAP based XAI approaches with small training datasets.

V. Conclusion

This paper has successfully demonstrated the application of XAI techniques in identifying surface level defects. Specifically, we focused on two approaches: the utilization of SHAP values and the implementation of SOMs. Our findings revealed that these approaches effectively complement each other, offering a more comprehensive understanding of the defect detection process. The SHAP value method provided insights into the importance and impact of different features in the classification decision, while SOMs offered a visual interpretation of the data, aiding in the identification of patterns and relationships within this dataset. The use of context dependent regression explanations further enhances the interpretability of SOM based classifications.

Our future work aims to expand the scope of these methodologies. We plan to apply both SHAP values and SOMs to more complex tasks, such as classifying various types of defects within the AITEX dataset. We anticipate not only enhancing the accuracy and efficiency of defect classification but also gaining deeper insights into the characteristics and nuances of different defect types. Our goal is to develop a more versatile and reliable XAI framework that can be applied to a wide range of materials and defect classifications, contributing significantly to the field of machine vision for manufacturing quality control.

REFERENCES

[1] B. H. M. van der Velden et al., 'Explainable artificial intelligence (XAI) in deep learning-based medical image analysis', *Med. Image Anal.*, vol. 79, p. 102470, Jul. 2022 [Online]. Available: 10.1016/j.media.2022.102470.

[2] Md. R. Hassan et al., 'Prostate cancer classification from ultrasound and MRI images using deep learning based Explainable Artificial Intelligence', *Future Gener. Comput. Syst.*, vol. 127, pp. 462–472, Feb. 2022 [Online]. Available: 10.1016/j.future.2021.09.030.

[3] F. Giuste et al., 'Explainable Artificial Intelligence Methods in Combating Pandemics: A Systematic Review', *IEEE Rev. Biomed. Eng.*, vol. 16, pp. 5–21, 2023 [Online]. Available: 10.1109/RBME.2022.3185953.

[4] A. de Jong, 'Automated Failure Diagnosis in Aviation Maintenance Using eXplainable Artificial Intelligence (XAI)', 2018.

[5] B. Hrnjica and S. Softic, 'Explainable AI in Manufacturing: A Predictive Maintenance Case Study', in *Advances in Production Management Systems. Towards Smart and Digital Manufacturing*, vol. 592, B. Lalic, V. Majstorovic, U. Marjanovic, G. von Cieminski, and D. Romero, Eds. Cham: Springer International Publishing, 2020, pp. 66–73[Online]. Available http://link.springer.com/10.1007/978-3-030-57997-5_8[Accessed: 11March2023].

[6] J.-R. Rehse et al., 'Towards Explainable Process Predictions for Industry 4.0 in the DFKI-Smart-Lego-Factory', *KI - Künstl. Intell.*, vol. 33, no. 2, pp. 181–187, Jun. 2019 [Online]. Available: 10.1007/s13218-019-00586-1.

[7] J. van der Waa et al., 'Evaluating XAI: A comparison of rule-based and example-based explanations', *Artif. Intell.*, vol. 291, p. 103404, Feb. 2021 [Online]. Available: 10.1016/j.artint.2020.103404.

[8] J. Zhou et al., 'Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics', *Electronics*, vol. 10, no. 5, p. 593, Mar. 2021 [Online]. Available: 10.3390/electronics10050593.

[9] M. Sayed-Mouchaweh, Ed., *Explainable AI Within the Digital Transformation and Cyber Physical Systems: XAI Methods and Applications*. Cham: Springer International Publishing, 2021[Online]. Available https://link.springer.com/10.1007/978-3-030-76409-8[Accessed: 11March2023].

[10] I. Ahmed et al., 'From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where', *IEEE Trans. Ind. Inform.*, vol. 18, no. 8, pp. 5031–5042, Aug. 2022 [Online]. Available: 10.1109/TII.2022.3146552.

[11] E. M. Kenny et al., 'Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies', *Artif. Intell.*, vol. 294, p. 103459, May 2021 [Online]. Available: 10.1016/j.artint.2021.103459.

[12] Z. C. Lipton, 'In machine learning, the concept of interpretability is both important and slippery.', *Mach. Learn.*

[13] A. Bennetot et al., 'Towards Explainable Neural-Symbolic Visual Reasoning'. arXiv, Oct. 22, 2019[Online]. Available http://arxiv.org/abs/1909.09065[Accessed: 25March2023].

[14] S. M. Lundberg et al., 'From local explanations to global understanding with explainable AI for trees', *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020 [Online]. Available: 10.1038/s42256-019-0138-9.

[15] S. Lundberg and S.-I. Lee, 'A Unified Approach to Interpreting Model Predictions'. arXiv, Nov. 24, 2017[Online]. Available http://arxiv.org/abs/1705.07874[Accessed: 10October2023].

[16] P. Linardatos et al., 'Explainable AI: A Review of Machine Learning Interpretability Methods', *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020 [Online]. Available: 10.3390/e23010018.

[17] R. R. Selvaraju et al., 'Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization', *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020 [Online]. Available: 10.1007/s11263-019-01228-7.

[18] C. Budayan et al., 'Comparing the performance of traditional cluster analysis, self-organizing maps and fuzzy C-means method for strategic grouping', *Expert Syst. Appl.*, vol. 36, no. 9, pp. 11772–11781, Nov. 2009 [Online]. Available: 10.1016/j.eswa.2009.04.022.

[19] C. S. Wickramasinghe et al., 'Explainable Unsupervised Machine Learning for Cyber-Physical Systems', *IEEE Access*, vol. 9, pp. 131824–131843, 2021 [Online]. Available: 10.1109/ACCESS.2021.3112397.

[20] G. Montavon et al., 'Explaining the Predictions of Unsupervised Learning Models', in *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek, Eds. Cham: Springer International Publishing, 2022, pp. 117–138[Online]. Availablehttps://doi.org/10.1007/978-3-031-04083-2_7.

[21] H. Wu et al., 'Put the Bear on the Chair! Intelligent Robot Interaction with Previously Unseen Chairs via Robot Imagination', in *2022 International Conference on Robotics and Automation (ICRA)*, Philadelphia, PA, USA, 2022, pp. 6276–6282[Online]. Available https://ieeexplore.ieee.org/document/9811619/[Accessed: 4November2023].

[22] D. Fryer et al., 'Shapley Values for Feature Selection: The Good, the Bad, and the Axioms', *IEEE Access*, vol. 9, pp. 144352–144360, 2021 [Online]. Available: 10.1109/ACCESS.2021.3119110.

[23] M. Alviano et al., 'Towards a Conditional and Multi-preferential Approach to Explainability of Neural Network Models in Computational Logic (Extended Abstract)'.

[24] J. Ables et al., 'Creating an Explainable Intrusion Detection System Using Self Organizing Maps'. arXiv, Jul. 15, 2022[Online]. Available http://arxiv.org/abs/2207.07465[Accessed: 24September2023].

[25] Md. Y. Mia et al., 'Analysis of self-organizing maps and explainable artificial intelligence to identify hydrochemical factors that drive drinking water quality in Haor region', *Sci. Total Environ.*, vol. 904, p. 166927, Dec. 2023 [Online]. Available: 10.1016/j.scitotenv.2023.166927.

[26] Á. J. García-Tejedor and A. Nogales, 'An open-source Python library for self-organizing-maps', *Softw. Impacts*, vol. 12, p. 100280, May 2022 [Online]. Available: 10.1016/j.simpa.2022.100280.

[27] O. Serradilla et al., 'Adaptable and Explainable Predictive Maintenance: Semi-Supervised Deep Learning for Anomaly Detection and Diagnosis in Press Machine Data', *Appl. Sci.*, vol. 11, no. 16, p. 7376, Aug. 2021 [Online]. Available: 10.3390/app11167376.

[28] S. Nguyen and B. Tran, 'XMAP: eXplainable mapping analytical process', *Complex Intell. Syst.*, vol. 8, no. 2, pp. 1187–1204, Apr. 2022 [Online]. Available: 10.1007/s40747-021-00583-8.

[29] D. Nagar et al., 'A novel data-driven visualization of n - dimensional feasible region using interpretable self-organizing maps (iSOM)', *Neural Netw.*, vol. 155, pp. 398–412, Nov. 2022 [Online]. Available: 10.1016/j.neunet.2022.08.019.