Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 217 (2023) 495–506

Procedia Computer Science

www.elsevier.com/locate/procedia

4th International Conference on Industry 4.0 and Smart Manufacturing

Causality-Aware Convolutional Neural Networks for Advanced Image Classification and Generation

Vagan Terziyan[a]*, Oleksandra Vitko[b]

[a] *Faculty of Information Technology, University of Jyväskylä, 40014, Jyväskylä, Finland*
[b] *Department of Artificial Intelligence, Kharkiv National University of Radio Electronics, 61166, Kharkiv, Ukraine*

Abstract

Smart manufacturing uses emerging deep learning models, and particularly Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs), for different industrial diagnostics tasks, e.g., classification, detection, recognition, prediction, synthetic data generation, security, etc., on the basis of image data. In spite of being efficient for these objectives, the majority of current deep learning models lack interpretability and explainability. They can discover features hidden within input data together with their mutual co-occurrence. However, they are weak at discovering and making explicit hidden causalities between the features, which could be the reason behind the particular diagnoses. In this paper, we suggest Causality-Aware CNNs (CA-CNNs) and Causality-Aware GANs (CA-GANs) to address the issue of learning hidden causalities within images. The core architecture includes an additional layer of neurons (after the last convolution-pooling and just before the dense layers), which learns pairwise conditional probabilities (aka causality estimates) for the features. Computations for these neurons are driven by the adaptive Lehmer mean function. Learned causalities are merged with the features during flattening and (via fully connected layers) influence the classification outcomes. Such causality estimates can be done for the mixed inputs where images are combined with other data. We argue that CA-CNNs not only improve the classification performance of normal CNNs but also open additional opportunities for the explainability of the models' outcomes. We consider as an additional advantage for CA-CNNs (if used as a discriminator within CA-GANs) the possibility to generate realistically looking images with respect to the causalities.

Keywords: causal discovery; causal inference; image processing; Convolutional Neural Network; Generative Adversarial Network.

* Corresponding author. *E-mail address:* vagan.terziyan@jyu.fi

10.1016/j.procs.2022.12.245

1. Introduction

Processes within Industry 4.0 are dealing with extensive data collection, data analysis and automatic decision-making on the basis of increasing amounts of data. Artificial Intelligence (AI) and particularly machine learning (ML) as advanced data analytics are essential for obtaining insights regarding production, better decision support, higher manufacturing quality and sustainability [1]. The complexity of ML models leads to unclear bases for the decisions they generate. Industry 4.0 experts are facing the increased requirements regarding fairness, accountability and transparency [2] of various decision-driven processes, especially in mission-critical scenarios and sensitive-use cases.

Deep learning algorithms are capable of capturing essential features from data. However, they are making decisions mainly on the basis of co-occurrence, association and correlation among the discovered features, which do not provide an insight into the actual reasons behind the decisions, i.e., hidden causal relationships [3]. "Correlation does not imply causation" – this is the major message from the great book by Pearl & Mackenzie [4]. In simple words, two variables y and x could be correlated (statistically dependent) and, therefore, seeing x allows predicting the value of y, but if y is not caused by x then setting the value of x won't affect the distribution of y. As noticed in [5], making the hidden causal mechanisms explicit is critically important for smart manufacturing. Being a popular trend nowadays, Explainable AI (XAI) is capable (to some extent) of understanding and interpreting the ML models. However, it is still weak in capturing and making explicit the causalities behind the behavior of the observed artifacts [6]. An excellent review on the state-of-the-art research within causal discovery (i.e., a step beyond the traditional statistical dependency) for the manufacturing domain is provided in [1]. According to the review, causal discovery is concerned with the problem of identifying causal relationships [7] from data for making explicit the corresponding causal structure [8] on the basis of discovered statistical artifacts (features, probability distributions, etc.). Such discovery identifies some pairs of causally related variables, features or patterns from the data, and, therefore, enables causal inference with the objective of identifying and numerically assessing causal effects based on a known causal structure.

As it is noticed in [9], efficient control over industrial processes requires real time monitoring and supervision using key performance indicators to enable awareness of the process flow. Causality learning in such cases is important for the prioritization of the influencing factors regarding the processes in order to provide smarter decision support. Hidden causalities within numerous human factors and human actions, which drive human behavior especially in stress or emergency [10], are also important to be made explicit for adequate simulations with the models of particular workers [11], which will result in optimized safety protocols and regulations handling industrial accidents.

Use of Bayesian Networks [12] for causality learning, in spite of the fact that they are specifically designed for the purpose and even have deeper (Bayesian Metanetworks) options for the architecture [13], is not always feasible as it requires a priori expert knowledge for causal structure construction before being able to learn corresponding conditional probabilities. Therefore, searching for a suitable (preferably on the basis of artificial neural networks) ML paradigm for causal structure learning based on data and without a priori knowledge is still an important objective for improving XAI for smart manufacturing. So far, the interpretability (extraction of relevant information from the ML model concerning relationships learned by the model [14]) and explainability (the ability to communicate this information via human-understandable language [15]) have been the main driving pillars of XAI [16]. As it is argued in [17], XAI needs human-level explainability in addition to model-agnostic methods (i.e., enabling transparency of deep learning models), and, therefore, more advanced explainability must also provide "causability" as causally understandable explanations and ensure that these explanations are reliably action-guiding [18]. Some hybrid models like neural-backed decision trees [19], which combine the neural networks (due to their high accuracy) with the decision trees (due to their interpretability) contribute to a smarter XAI, but still lack causability.

A special case would be to discover hidden causalities between objects presented in each single image from some image dataset. Specifics of images is that their representation does not include any explicit indications regarding features, patterns or objects; it introduces just the pixels for visual representation of a particular scene. Image datasets do not provide labels describing the objects' causal dispositions. Therefore, supervised ML as such cannot approach them. Also, unlike having the frame from a video, from a single image one may not see the dynamics of appearance and change of the objects in the scene. Therefore, a priori information as a hint for causality discovery is absent.

Lopez-Paz et al. [20] suggest approaching the problem with the "causal disposition" concept, which is more primitive than interventional causation (do-calculus) and causal graphs from Pearl's approach [4]. However, it could be the only way to proceed with the limited a priori information. It enables counting the number C(A,B) of images in

which the causal dispositions of artifacts A and B is such that B disappears if one removes A. One can assume then (aka rule of thumb) that the artifact A causes the presence of artifact B when C(A,B) is (sufficiently) greater than the converse C(B,A). In this way, any causal disposition induces a set of asymmetric causal relationships between the artifacts from an image (features, patterns, object categories, etc.) that represent (weak but better than nothing) causality signals regarding the real-world scene. The fundamental question, as pointed out in [20], would be to infer such an asymmetric causal relationship from the statistics observed in an image dataset.

In this paper, we suggest one way to compute such an asymmetric measure for possible causal relationships within image datasets and we include such computations as a component into a Convolutional Neural Network (CNN) architecture (which is known to be one of the most accurate so far generic models for image classification problems) to enable such "causality-aware" CNN or CA-CNN to classify images taking into account hidden causalities within them. We argue that CA-CNNs can be used with mixed data (images with other data) and also can work as a component within Generative Adversarial Networks (GANs) to enable generation of images with respect to causalities.

The rest of the paper is organized as follows: Section 2 discusses the intuition behind the approach; Section 3 suggests analytics for asymmetric causality estimates; Section 4 describes the architecture for the CA-CNN on the basis of the analytics and corresponding "causality map"; Section 5 extends CA-CNN to address mixed data (images plus other data); in Section 6, the architecture for "causality-aware" GAN (CA-GAN) is presented as a capability to discover "causal fakes" and generate images with respect to causal relationships; and we conclude in Section 7.

2. Causalities hidden in images (motivating scenario)

Let us start with the example, which is suggested in [20] as a show-case for causal disposition concept (see Fig. 1).

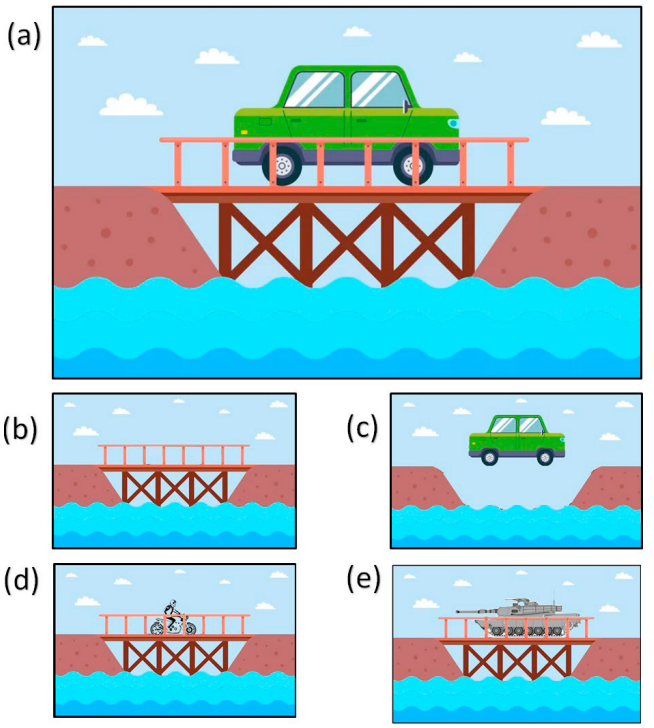

Fig. 1. Illustration of the "causal disposition" concept [20]: (a) original image of a car on a wooden bridge; (b) removal of the car from the scene keeps the image realistically looking; (c) removal of the bridge makes the scene inconsistent; (d) replacing the car with a bike is still a valid intervention; (e) replacing the car with the tank makes the consistency of the image questionable (i.e., too heavy vehicle for a wooden bridge).

Lopez-Paz et al. [20] consider two counterfactual questions regarding the scene from Fig. 1(a): "What would the scene look like if we were to remove the car?" and "What would the scene look like if we were to remove the bridge?"

The first intervention (Fig. 1(b)) does not change the consistency of the scene (an observer may see similar scenes among other images) while the second one (Fig. 1(c)) looks inconsistent (i.e., never seen among other images). Therefore, we may assume that the presence of the bridge has some effect on the presence of the car.

Let us add a couple of more complex interventions to answer the following questions regarding the scene from Fig. 1(a): "What would the scene look like if we were to replace the car with the bike?" and "What would the scene look like if we were to replace the car with the tank?" The first intervention (Fig. 1(d)) does not change the consistency of the scene (an observer may see similar scenes with similar "light" vehicles among other images) while the second one (Fig. 1(e)) looks inconsistent (i.e., never seen such "heavy" vehicles on such "light" bridges among other images). Therefore, we can make a more complex assumption here that the type of bridge has some effect on the type of vehicle located on it.

As a summary regarding these examples, we may admit that it would be possible and reasonable to get some weak causality signals from the individual images of some dataset just on the basis of statistics provided by ML without adding primary expert knowledge.

3. Analytics for the asymmetric causality estimates in images

Digital images are made up of pixels, and the features of an image (needed for supervised ML) are not explicit within the image representation but can be captured by the convolutional layers of CNNs. One may assume that the last convolutional layer outputs and localizes to some extent the object-like features.

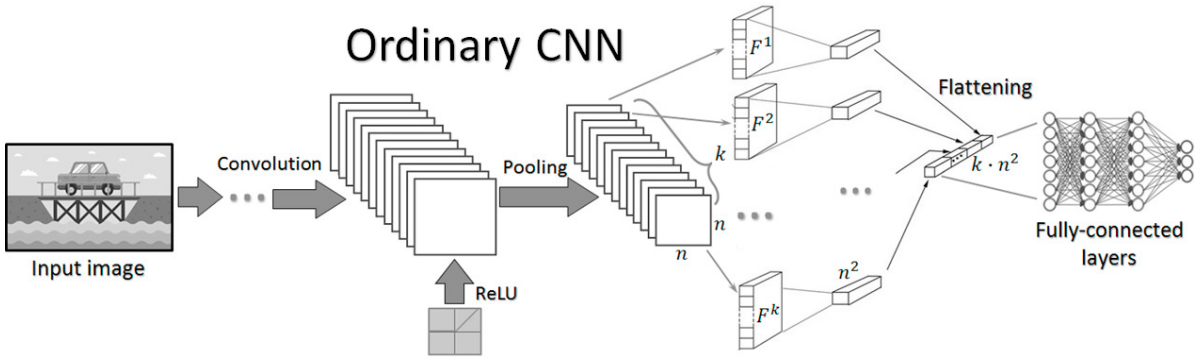

Fig. 2. Basic architecture of a typical Convolutional Neural Network

Fig. 2 illustrates a typical architecture of a CNN for image classification. It assumes that a deep neural network classifier (marked as fully-connected layers in the architecture) will get the features needed for classification not from the pixel representation of an input image directly but after several convolution + pooling layers, which capture these features from the image. Convolution layers summarize the presence of certain features in the image by systematically applying learned filters and producing a corresponding set of feature maps. These maps are sensitive to the location of the features in the image, therefore down-sampling (pooling) is applied aiming to get a summary (either average or maximal) of the presence of a particular feature within the batches (groups of adjacent pixels within the square shaped sub-regions) of the feature map. The pooling operation involves sliding a two-dimensional square filter (or kernel) over the feature map and summarizing the features lying within the batch covered by the filter. After the last pooling layer, we get k features $\mathbf{F^1}, \mathbf{F^2}, \dots, \mathbf{F^k}$ represented by $n \times n$ matrixes (feature maps). The presence of ReLU operations in the architecture guarantees that the feature maps contain only non-negative numbers as measures of the presence of a particular feature within a particular batch (location in the image). If you normalize these numbers to the interval $[0, 1]$ by dividing each of them to the maximal possible value of feature presence $MAX(\mathbf{F})$, we can interpret the feature maps' values as probabilities. For example, $F^i_{a,b} = 0.7$, could be interpreted as the value of presence (the probability of being present) of feature $\mathbf{F^i}$ in the location (batch) with coordinates (a, b) is estimated as 0.7 (70 %).

While the objectives of Lopez-Paz et al. [20] were to use the causal disposition concept to effectively distinguish between the features (which cause the presence of the particular object in the image and features that are caused by

the presence of the object) to get the causal structure of the world represented by some image dataset, we are going to measure and use this causal asymmetry for better image classification and generation.

Therefore, we are going to heuristically estimate the values for conditional probabilities regarding the pairs of features and use CNN to learn how these estimates influence image classification. As a generic schema, we will use the basic rule for conditional probability, which connects it with the joint probability:

$$P(\mathbf{F^i}|\mathbf{F^j}) = \frac{P(\mathbf{F^i},\mathbf{F^j})}{P(\mathbf{F^j})}. \quad (1)$$

There could be several ways to estimate the joint probability in formula (1) as each feature is represented by the matrix of normalized numbers (probabilities within particular locations). We suggest to do this as follows:

$$P(\mathbf{F^i}|\mathbf{F^j}) = \frac{\left(\max\limits_{l,r=1,n} F^i_{l,r}\right)\cdot\left(\max\limits_{l,r=1,n} F^j_{l,r}\right)}{\sum_{l,r=1}^{n} F^j_{l,r}}. \quad (2)$$

Formula (2) gives a number within $[0, 1]$ interval and it is a good estimate for conditional probability. It considers joint probability to be the maximal presence of both features in the image (each one in their own location). Formula (2) could be used to estimate asymmetric [because, in general case, $P(\mathbf{F^i}|\mathbf{F^j}) \neq P(\mathbf{F^j}|\mathbf{F^i})$] causal relationships between features $\mathbf{F^i}$ and $\mathbf{F^j}$. It is interesting that $P(\mathbf{F^i}|\mathbf{F^i})$ has also some sense other than simply $P(\mathbf{F^i}|\mathbf{F^i}) = 1$ with the traditional conditional probabilities, because it provides some information on the cause-effect of the appearance of the feature in one place of the image given the presence of the same feature within some other places of the image:

$$P(\mathbf{F^i}|\mathbf{F^i}) = \frac{\left(\max\limits_{l,r=1,n} F^i_{l,r}\right)^2}{\sum_{l,r=1}^{n} F^j_{l,r}}. \quad (3)$$

Here we also suggest another (more generic than formula (2)) option to interpret formula (1) by applying the generalized average function, particularly the Lehmer mean, which has useful properties (see, e.g., [21]) and could be controlled with a (trainable or adaptive) parameter as follows:

$$P(\mathbf{F^i}|\mathbf{F^j})_\alpha = \frac{LM_\alpha\,(\mathbf{F^i}\times\mathbf{F^j})}{LM_\alpha(\mathbf{F^j})}, \quad \text{where:} \quad (4)$$

- $\mathbf{F^i}\times\mathbf{F^j} = \{F^i_{11}\cdot F^j_{11}, F^i_{11}\cdot F^j_{12}, \dots, F^i_{11}\cdot F^j_{nn}, \dots, F^i_{nn}\cdot F^j_{11}, F^i_{nn}\cdot F^j_{12}, \dots, F^i_{nn}\cdot F^j_{nn}\}$ is a vector of n^4 pairwise multiplications;
- LM_α - Lehmer Mean with trainable parameter α, i.e.:
- $LM_\alpha(\mathbf{F^j}) = LM_\alpha\left(F^j_{11}, F^j_{12}, \dots, F^j_{nn}\right) = \frac{\left(F^j_{11}\right)^{\alpha+1}+\left(F^j_{12}\right)^{\alpha+1}+\dots+\left(F^j_{nn}\right)^{\alpha+1}}{\left(F^j_{11}\right)^{\alpha}+\left(F^j_{12}\right)^{\alpha}+\dots+\left(F^j_{nn}\right)^{\alpha}} = \frac{\sum_{p,q=1}^{n}\left(F^j_{pk}\right)^{\alpha+1}}{\sum_{l,r=1}^{n}\left(F^j_{lr}\right)^{\alpha}}$;
- $LM_\alpha(\mathbf{F^i}\times\mathbf{F^j}) = LM_\alpha\left(F^i_{11}\cdot F^j_{11}, F^i_{11}\cdot F^j_{12}, \dots, F^i_{nn}\cdot F^j_{nn}\right) = \frac{\sum_{p,q=1}^{n}\sum_{h,g=1}^{n}\left(F^i_{pk}\cdot F^j_{hg}\right)^{\alpha+1}}{\sum_{p,q=1}^{n}\sum_{h,g=1}^{n}\left(F^i_{pk}\cdot F^j_{hg}\right)^{\alpha}}$.

The option to use formula (2) to numerically estimate asymmetric causal relationships requires less computation than formula (4). However, formula (4) gives a certain flexibility (the Lehmer mean, depending on the parameter, is capable of producing the values between MIN and MAX across a simple average among the operands) and, therefore, if included into a neural network, the optimal parameter for formula (4) can be learned by backpropagation. Here an interesting bridge is foreseen with the SHAP-driven XAI framework [22], which assigns to each input feature a value of importance for a particular prediction. By learning the parameter in formula (4), we can potentially discover the SHAP importance not only for the individual features but also for their pairwise causal relationships.

4. Introducing Causality-Aware CNNs

The suggested causality estimates can be embedded into the basic CNN architecture, making it "causality-aware", i.e., CA-CNN as shown in Fig. 3. One may see that basic CNN is updated with the additional "causality map" $P(\mathbf{F}|\mathbf{F})$ of size $k \times k$ (k - number of features) just after the last pooling layer. The content of the causality map is computed either by formula (2) or (4). As it can be seen from the architecture, the features $\mathbf{F^1}, \mathbf{F^2}, \dots, \mathbf{F^k}$ represented by $n \times n$ feature maps are not only used to compute each of $P(\mathbf{F^i}|\mathbf{F^j})$ but also (as in traditional CNN) they go (after flattening) directly (as an input vector of size $k \cdot n^2$) to the fully connected layers, while the causality map after being computed provides additional k^2 inputs to the flattening vector. These additional inputs, which will be used for image classification, are (in a way) providers for the causality awareness. During the backpropagation learning of the corresponding weights and other parameters, CA-CNN will automatically discover which of the k^2 causalities are important as a (decisive) factor for the correct classification. This additional feature of CA-CNNs would be an especially important update for getting better classification accuracy for such image datasets, in which the logic behind the distribution of images among classes depends on the causal nature of the scenes represented in the images (likely case for various industrial image datasets, e.g., used for training classifiers for industrial diagnostics and predictive maintenance purposes).

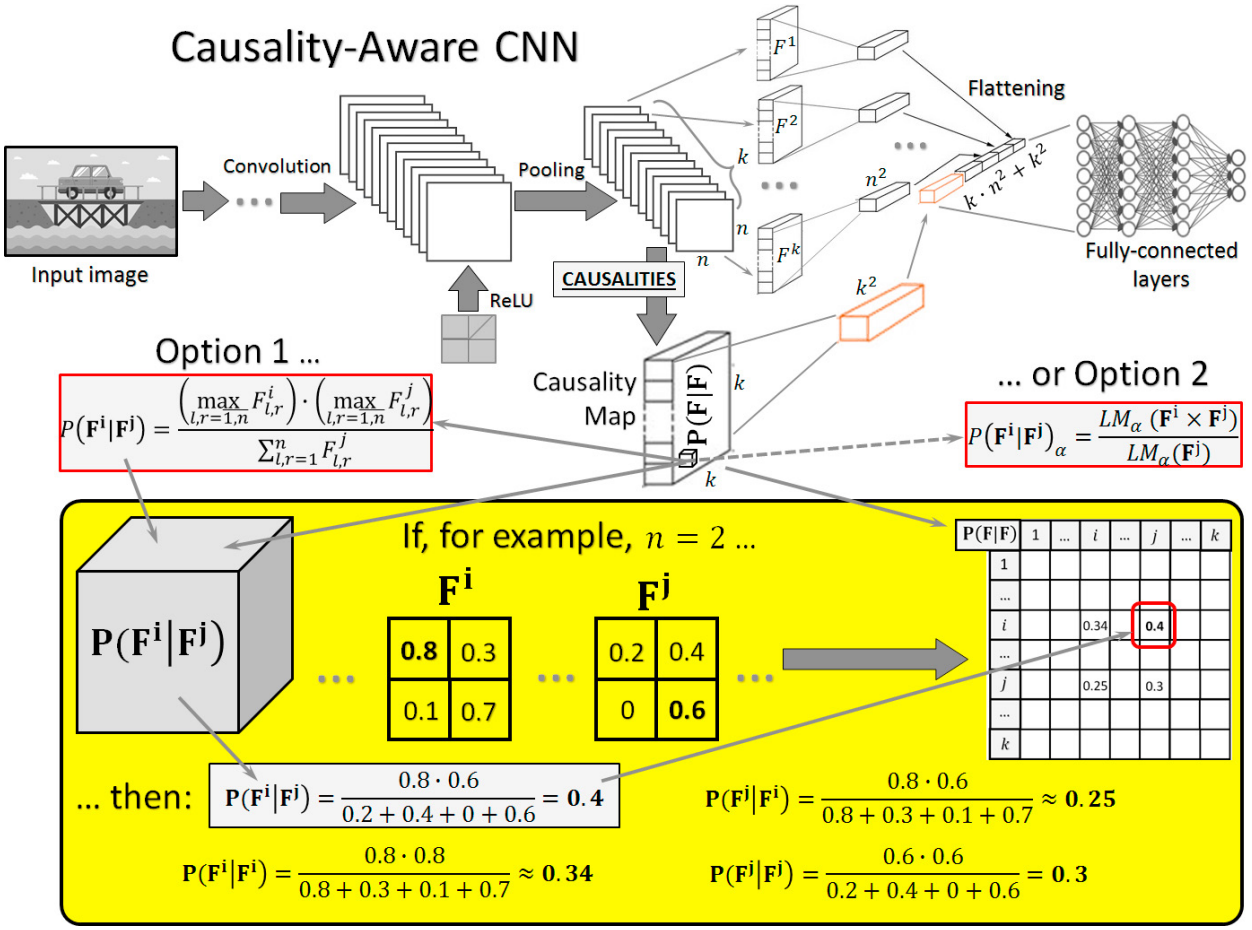

Fig. 3. The basic architecture of "causality-aware" CNN (CA-CNN). In addition to the feature maps after the last pooling layer, it also has the causality map containing estimates for pairwise causal relationship between the features. This causality map is computed by one of the two options as shown in the figure. Together with the feature maps themselves, the causality map produces additional inputs to the fully connected layers, which will be used for further image classification; and the weights for the corresponding additional connections (i.e., actual causality influences) will be learned by backpropagation the same way as other neural network parameters.

The intuition behind CA-CNNs is illustrated in Fig. 4 and Fig 5 on a simple example.

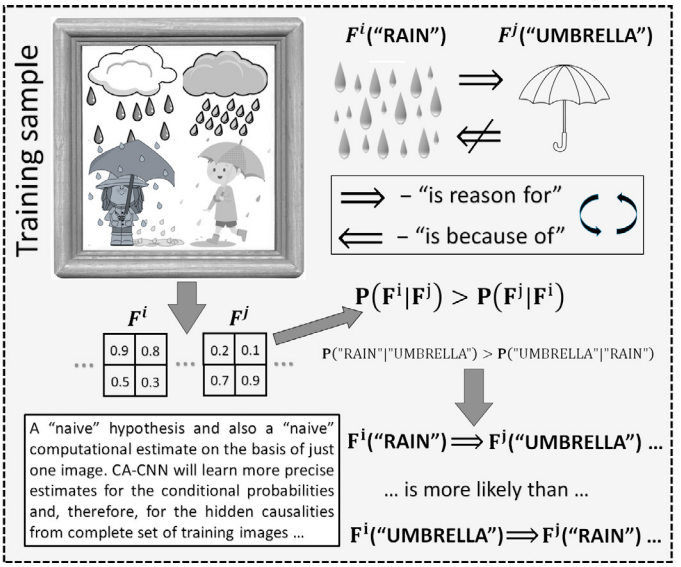

Fig. 4. An intuitive hypothesis (as a driver for the causality relationships discovery within the images) is illustrated by simple example.

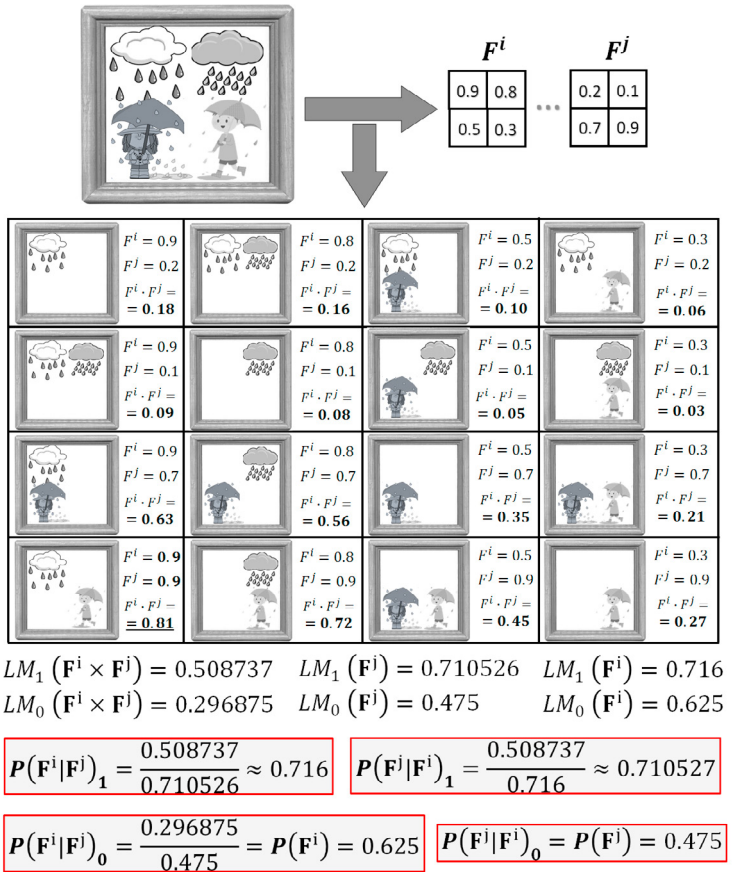

Fig. 5. The logic of causality map computation is illustrated (one may see the reasons behind each multiplication in the formula).

Consider a simple image as a training sample as shown in Fig. 4. Let us assume that this image is being processed by CA-CNN. Assume that, after the last pooling layer, we have several 2×2 feature maps, which correspond to the discovered features and provide a numeric estimate on the extent to which each of the features is represented in each of the 2×2 sectors in the image. Assume that (among the discovered features) we have feature $\mathbf{F^i}$ (let us name it "Rain") and feature $\mathbf{F^j}$ (let us name it "Umbrella"). The corresponding 2×2 feature maps are shown in Fig. 4. From these maps one may see that, for example, "Rain" is represented in the image sector with the coordinates (1,1) having the probability of presence equal to 0.9, while the presence of "Umbrella" in the same sector has the probability equal to 0.2. Due to certain prior knowledge, for us (humans) the likelihood of "Umbrella" appearance because of "Rain" is higher that the likelihood of "Rain" due to the "Umbrella". However, due to the fact that (after computing both $P(\mathbf{F^i}|\mathbf{F^j})$ and $P(\mathbf{F^j}|\mathbf{F^i})$ using either formula (2) or (4)), we get $P(\mathbf{F^i}|\mathbf{F^j}) > P(\mathbf{F^j}|\mathbf{F^i})$, from which the same (initially hidden from the neural network) assumption ("Rain" is likely the reason for "Umbrella" in the image) could be derived. An important point here is that, during learning, CA-CNN will get weights for each of such pairwise assumptions and will discover if they are valid to influence potential classification of the image.

Fig. 5 provides more insights into the computation regarding the "Rain" vs. "Umbrella" example. One may see that our pair of features $\mathbf{F^i}$ and $\mathbf{F^j}$ with each pair of sectors (i.e., $2 \times 2 \times 2 \times 2 = 16$ combinations altogether) give one multiplication as a component for joint probability (joint feature appearance). For example, the joint probability that the feature "Rain" is presented in the sector (1,2) and the feature "Umbrella" is present in the sector (1,1) is equal to $0.8 \cdot 0.2 = 0.16$ as shown in Fig.5. Further computations according to formula (4) with all the component joint probabilities are shown for two different parameter values ($\alpha = 1$ and $\alpha = 0$) for the Lehmer mean function. In both cases we get $P(\text{"Rain"}|\text{"Umbrella"}) > P(\text{"Umbrella"}|\text{"Rain"})$. Later, after AC-CNN is trained using all the image dataset, it will be clarified whether this estimated potential causal relationship between the "Rain" and "Umbrella" features (across all the training images) is important for image classification within this particular dataset.

5. Mixed data processing with the CA-CNNs

In diagnostics (both medical and industrial), it is often the case when each particular sample of data contains different numeric measurements with some categorical characteristics in addition to the images. It is known that the features discovered by a CNN, which is dealing with the image component of the data, can be merged with the outcome (features at certain level of abstraction) discovered by a multi-layer perceptron (deep fully-connected feedforward neural network) and then the merged features will go to the common fully-connected layers for the further classification. We suggest updating such a merging schema so that both channels (the convolutional channel for image data and the perceptron for other data) will have a shared causality map (see Fig. 6) so that the causal relationships will be discovered not just among the features taken from the convolutional channel but among the whole set of features from both channels. Assume that, in addition to the k features $\mathbf{F^1}, \mathbf{F^2}, \dots, \mathbf{F^k}$ from the convolutional channel (as was already described), we have also m features $\mathbf{f^1}, \mathbf{f^2}, \dots, \mathbf{f^m}$ represented by single values (not by the matrixes as in the case of feature maps). All the features are normalized to the [0, 1] interval. The heterogeneity of features requires slight modification of the computing schema for the mixed $(k+m) \times (k+m)$ causality matrix as follows:

$P(\mathbf{F^i}|\mathbf{F^j})_\alpha$ is computed as usual according to formula (4);

$$P(\mathbf{F^i}|\mathbf{f^j})_\alpha = \frac{LM_\alpha(\mathbf{F^i} \times \mathbf{f^j})}{\mathbf{f^j}}; P(\mathbf{f^i}|\mathbf{F^j})_\alpha = \frac{LM_\alpha(\mathbf{f^i} \times \mathbf{F^j})}{LM_\alpha(\mathbf{F^j})}, \quad (5)$$

where: $\mathbf{F^i} \times \mathbf{f^j} = \{F^i_{11} \cdot f^j, F^i_{12} \cdot f^j, \dots, F^i_{1n} \cdot f^j, \dots, F^i_{nn} \cdot f^j, F^i_{nn} \cdot f^j, \dots, F^i_{nn} \cdot f^j\}$ and $\mathbf{F^i} \times \mathbf{f^j} = \{f^i \cdot F^j_{11}, f^i \cdot F^j_{12}, \dots, f^i \cdot F^j_{1n}, \dots, f^i \cdot F^j_{n1}, f^i \cdot F^j_{n2}, \dots, f^i \cdot F^j_{nn}\}$ – are the vectors of n^2 pairwise multiplications;

$$P(\mathbf{f^i}|\mathbf{f^j})_\alpha = \frac{LM_\alpha(\mathbf{f^i}, \mathbf{f^j})}{LM_\alpha(\mathbf{f^j}, 1)}. \quad (6)$$

Resulting causality map will be used as the set of additional $(k+m)^2$ inputs merged with the actual features and provided to the fully-connected layers of the network for further classification as it is shown in Fig. 6.

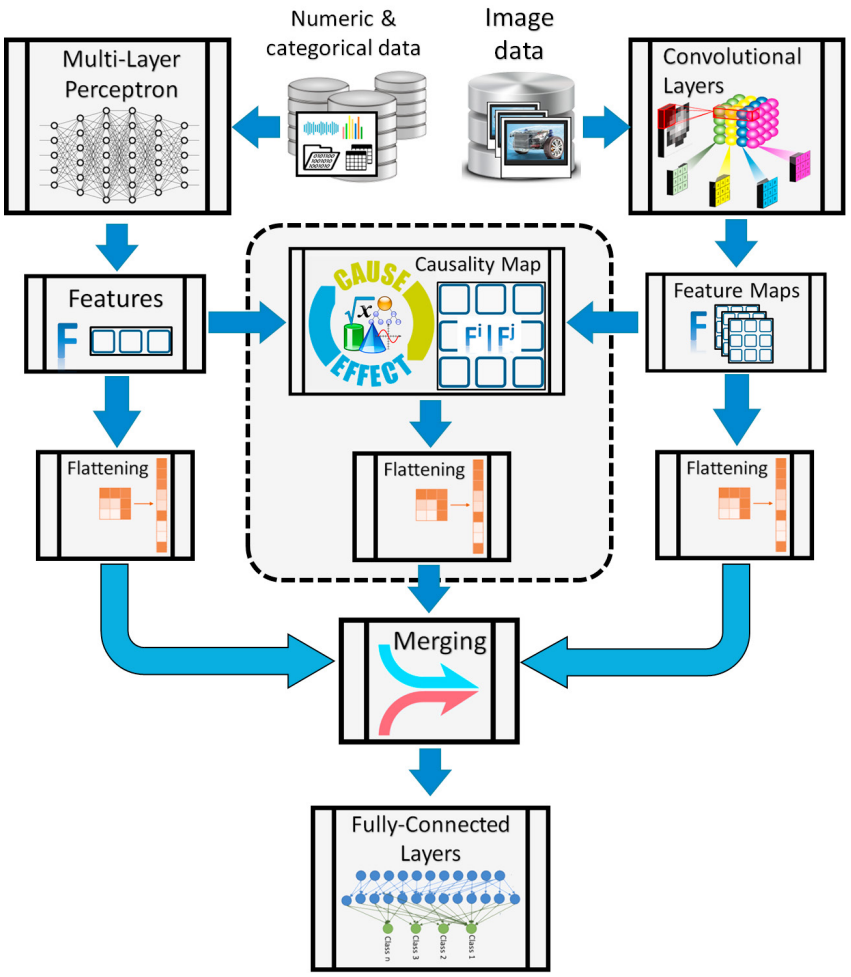

Fig. 6. An architecture, which illustrates the process of merging the features discovered from images by CA-CNN (i.e., from the feature maps and from the causality map) and from other than image sources of data. The causality map is constructed in a way to estimate pairwise causal relationships between features captured from both data channels: from the images and from other (numeric, categorical, etc.) data. The arrows here represent the ML process flow chains and visualize independent sub chains and the places where the processes are merged.

6. Generation of images with respect to causalities

Generative Adversarial Networks (GANs) [23] with many variations of their architectures [24] are known to be one of the most powerful ML tools for a wide range of applications, which involve image processing and generation [25]. The backbone idea behind GANs is synchronous adversarial training of two capabilities (competing neural networks): Discriminator, which separates generated (fake) images from the real ones; and Generator of realistically-looking images. Both networks are trained simultaneously until reaching some balance stage (i.e., generated images are too good to be distinguishable from the real ones). One approach to improve the quality of generated images would be the one which we suggested in [26]: if you enhance solely the architecture of the Discriminator (by some useful component), then one may expect that the Generator (while training to reach the balance with the stronger Discriminator) will also improve its performance in generating high quality images. Taking into account that Discriminator is basically a CNN, then updating of it with causality-awareness (via CA-CNN) will cause the Generator to improve its own generation performance with respect to causalities. Appropriate GAN architecture, which we call a Causality-Aware GAN (CA-GAN), is shown in Fig. 7 and an example of its work is shown in Fig. 8.

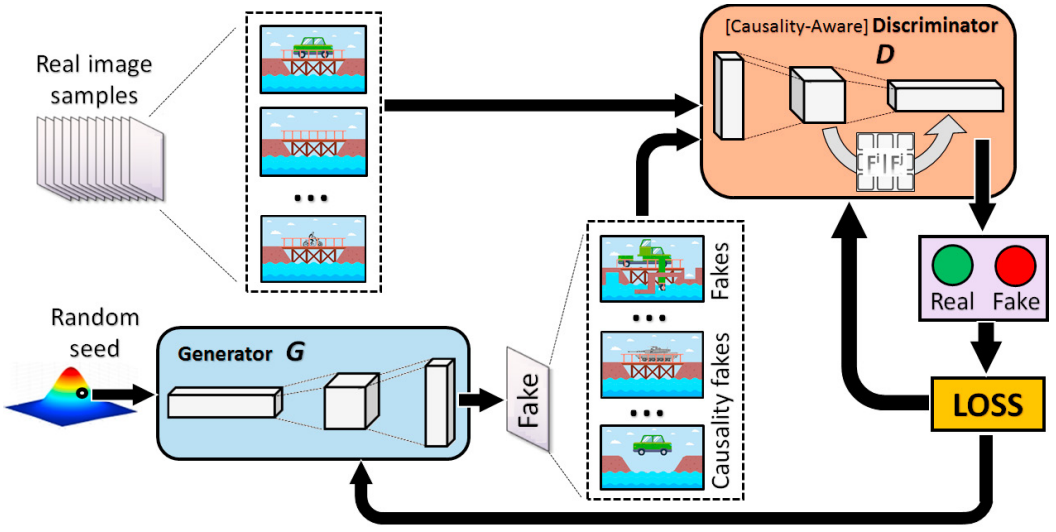

Fig. 7. An architecture of "causality-aware" Generative Adversarial Network (CA-GAN). The major modification here (comparably to the traditional architecture) is a Causality-Aware Discriminator, which is based on CA-CNN architecture and includes a causality map component for learning causal feature relationships. This update makes the Discriminator to be capable of recognizing not only general fakes (not realistically looking inputs) but also "causal fakes", i.e., images with realistically looking components but with inconsistent causal relationships among these components. Synchronously, the Generator learns to generate realistically-looking images with special respect to causalities within it.

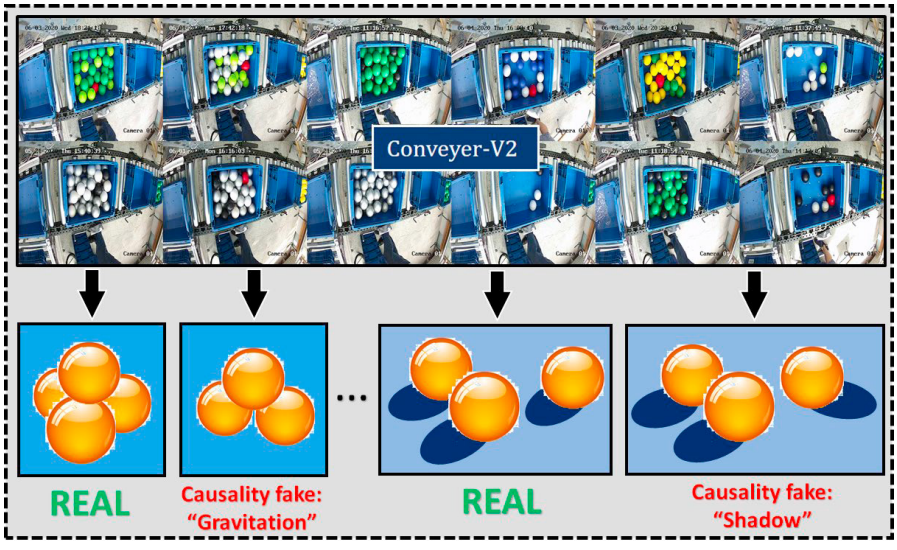

Fig. 8. An example of real-world simulations on a conveyor with a visual monitoring system. Collected observations have been used as an input to CA-GAN. During the training process, the Discriminator also acquired the capability to capture the causal fakes (examples are shown). After the training process, the Generator becomes capable of generating realistically-looking images (scenes) without causal inconsistencies.

Being a CA-CNN, the Discriminator will process images, taking into account causal relationships between the image features. This means that, if some causalities from the input image do not match with the feature relationships of the learned distribution of the images from the reality corpora, the Discriminator may consider such an image as a fake (aka "causality fake"), even if all the objects and backgrounds in the image look like the real ones. Previously, we provided examples of such fakes in Fig. 1(c) and Fig 1(e). Discovered causality fakes are provided as feedback

(loss) to the Generator network, which then updates its own generation skills accordingly and (sooner or later) will be capable of generating not just realistically-looking images but also images with realistic causal relationships.

In order to prove the concept of CA-CNNs by empirical evaluations for advanced image classification and CA-GANs for advanced image generation with the data from cyber-physical environments, we use the capacities of a logistics system and particularly of an interroll cassette conveyor (Fig. 8). Some preliminary experiments have been made there within the project IMMUNE: "Cyber-Defence for Intelligent Systems", which is a NATO SPS project (http://recode.bg/natog5511). The conveyor is used to simulate various scenes (cassette loads) and take images of them for further processing. 2198 such images have been taken by the cameras installed at the critical distribution points of the conveyer. Then CA-GAN is being trained to generate artificial images of the scenes. One may see the generated images at certain point of the training process in Fig.8. The monitoring of the training process shows that, at some point, the Discriminator starts to complain about "causality fakes" disclosing an important feedback to the Generator. This empirical study has discovered that about 10-12% of generated images classified as "fake" are actually causality fakes. In this way, the CA-GAN architecture enabled us to improve the quality of generated images comparably to our previous experiments [27] where images were generated as a "vaccine" for training digital immunity of the logistic system against adversarial attacks.

These were just preliminary experiments with relatively small numbers of training data and with more simulated rather than real scenarios. We believe, however, that the full hidden potential of CA-CNNs and CA-GANs is much higher and it is yet to be discovered with the more solid experiments involving huge volumes of training data.

7. Conclusions

In this paper, we suggested an additional component (causality matrix as a group of special neurons) to update the traditional CNN architecture towards causality-awareness within the image classification process. The resulting CA-CNN architecture is expected to distinguish between the classes of images, taking into account the causal relationships between the features from the images. The features itself are taken after the last pooling layer of the traditional CNN architecture and they are considered as the objects representing the scene shown in the image. The assumption is made that, in addition to the mutual appearance of each pair of such objects in the image, there might be a hidden causal relationship within each couple. Causality matrix computing is done to set-up (initialize) the hypothesis on potential causalities hidden within the input image, and the CA-CNN training process is a way to actually prove some of the hypothesis. We suggested two computation schemas for the causality matrix, with either more light or more heavy computations. The choice of the analytics to compute the initial values for the causality matrix is based on heuristics so that the estimates for conditional probabilities between the pairs of features could be used as a measure for potential causal relationship between the features. The preliminary experiments show that such estimates together with the corresponding weights and parameters trained by backpropagation actually make such a heuristics a reasonable one.

We show that the causality matrix as a shared component can be used to handle mixed data inputs, i.e., to combine image processing, which goes through CA-CNNs, with other data (numerical and categorical), which goes through, e.g., multi-layer perceptron. In this case, the causality matrix may contain causality relationship hypotheses regarding the pairs of features of different natures. We have updated corresponding analytics to handle such cases also.

We considered CA-CNN as a possible architecture of a Discriminator within a typical GAN architecture. Such a causality-aware Discriminator (being more powerful due to additional skills) forces its counterpart – the Generator – to respect causalities when generating images. Corresponding causality-aware GANs or CA-GANs could be useful for both purposes: for the discovery of "causality fakes" (realistically-looking images but with inconsistent causal relationships between their components); and for generating realistically-looking images with respect to causalities.

Preliminary experiments and simulations with relatively small numbers of data samples provide certain optimism on the potential of the suggested causality-aware architectures. More experiments with big data from industrial processes are foreseen to discover the full potential of the proposed architectures. Among future objectives we keep in mind also the XAI domain because our updated architectures have clear potential to improve the explainability of ML models. This study is a certain step towards explainability of CNNs, which is important for Industry 4.0 context. To get full benefit of our approach to Industry 4.0 needs, we will combine it in our future study with the self-attention mechanism in CNNs [28], which has proven its benefits for smart manufacturing applications.

References

[1] Vuković, M., and Thalmann, S. (2022). "Causal discovery in manufacturing: a structured literature review". *Journal of Manufacturing and Materials Processing*, **6(1)**: 10.

[2] Shin, D., and Park, Y. J. (2019). "Role of fairness, accountability, and transparency in algorithmic affordance". *Computers in Human Behavior*, **98**: 277-284.

[3] Kuehnert, C., Bernard, T., and Frey, C. (2011). "Causal structure learning in process engineering using Bayes Nets and soft interventions". In: *Proceedings of the 9th IEEE International Conference on Industrial Informatics* (pp. 69-74). IEEE.

[4] Pearl, J., and Mackenzie, D. (2018). "*The book of why: the new science of cause and effect"*. Basic books.

[5] Holzinger, A., Kieseberg, P., Weippl, E., and Tjoa, A. M. (2018). "Current advances, trends and challenges of machine learning and knowledge extraction: from machine learning to explainable AI". In: *Proceedings of the International Cross-Domain Conference for Machine Learning and Knowledge Extraction* (pp. 1-8). Springer, Cham.

[6] Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). "Explainable AI: a review of machine learning interpretability methods". *Entropy*, **23(1)**: 18.

[7] Eberhardt, F. (2017). "Introduction to the foundations of causal discovery". *International Journal of Data Science and Analytics*, **3(2)**: 81-91.

[8] Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H. (2020). "A survey of learning causality with data: problems and methods". *ACM Computing Surveys*, **53(4)**: 1-37.

[9] Amzil, K., Yahia, E., Klement, N., and Roucoules, L. (2020). "Causality learning approach for supervision in the context of Industry 4.0". In: *Proceedings of the International Joint Conference on Mechanics, Design Engineering & Advanced Manufacturing* (pp. 316-322). Springer, Cham.

[10] Nicoletti, L., & Padovano, A. (2019). Human factors in occupational health and safety 4.0: a cross-sectional correlation study of workload, stress and outcomes of an industrial emergency response. *International Journal of Simulation and Process Modelling*, *14*(2), 178-195.

[11] Longo, F., Nicoletti, L., & Padovano, A. (2019). Modeling workers' behavior: A human factors taxonomy and a fuzzy analysis in the case of industrial accidents. *International journal of industrial ergonomics*, *69*, 29-47.

[12] Pearl, J. (1995). "From Bayesian networks to causal networks". In: *Mathematical Models for Handling Partial Knowledge in Artificial Intelligence* (pp. 157-182). Springer, Boston, MA.

[13] Terziyan, V. (2005). "A Bayesian Metanetwork". *International Journal on Artificial Intelligence Tools*, **14(03)**: 371-384.

[14] Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). "Definitions, methods, and applications in interpretable machine learning". *Proceedings of the National Academy of Sciences*, **116(44)**: 22071-22080.

[15] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). "Causability and explainability of Artificial Intelligence in medicine". *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **9(4)**: e1312.

[16] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). "Explaining explanations: an overview of interpretability of machine learning". In: *Proceedings of the 5th IEEE International Conference on Data Science and Advanced Analytics* (pp. 80-89). IEEE.

[17] Chou, Y. L., Moreira, C., Bruza, P., Ouyang, C., and Jorge, J. (2022). Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications. *Information Fusion*, **81**: 59-83.

[18] Beckers, S. (2022). "Causal Explanations and XAI". *arXiv preprint arXiv:2201.13169*.

[19] Wan, A., Dunlap, L., Ho, D., Yin, J., Lee, S., Jin, H., ... and Gonzalez, J. E. (2020). "NBDT: neural-backed decision trees". *arXiv preprint arXiv:2004.00221*.

[20] Lopez-Paz, D., Nishihara, R., Chintala, S., Scholkopf, B., and Bottou, L. (2017). "Discovering causal signals in images". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6979-6987).

[21] Terziyan, V. (2017). "Social distance metric: from coordinates to neighborhoods". *International Journal of Geographical Information Science*, **31(12)**: 2401-2426.

[22] Lundberg, S. M., and Lee, S. I. (2017). "A unified approach to interpreting model predictions". *Advances in Neural Information Processing Systems*, **30**. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[23] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... and Bengio, Y. (2014). "Generative adversarial networks". *arXiv preprint arXiv:1406.2661*.

[24] Terziyan, V., Gryshko, S., and Golovianko, M. (2021). "Taxonomy of generative adversarial networks for digital immunity of Industry 4.0 systems". *Procedia Computer Science*, **180**: 676-685.

[25] Aggarwal, A., Mittal, M., and Battineni, G. (2021). "Generative adversarial network: An overview of theory and applications". *International Journal of Information Management Data Insights*, **1(1)**: 100004.

[26] Branytskyi, V., Golovianko, M., Malyk, D., and Terziyan, V. (2022). "Generative adversarial networks with bio-inspired primary visual cortex for Industry 4.0". *Procedia Computer Science*, **200**: 418-427.

[27] Golovianko, M., Gryshko, S., Terziyan, V., and Tuunanen, T. (2021). "Towards digital cognitive clones for the decision-makers: adversarial training experiments". *Procedia Computer Science*, **180**: 180-189.

[28] Liang, Y., Li, M., and Jiang, C. (2022). "Generating self-attention activation maps for visual interpretations of convolutional neural networks". *Neurocomputing*, **490**: 206-216.