A Human-Centric Evaluation Dataset for Automated Early Wildfire Detection from a Causal Perspective

Amelie Schmidt-Colberg
Fraunhofer FOKUS
amelie.schmidt-colberg@fokus.fraunhofer.de

Leonhard Löffler-Dauth
Fraunhofer FOKUS
leonhard.loeffler-dauth@fokus.fraunhofer.de

ABSTRACT

Insight into performance ability is crucial for successfully implementing AI solutions in real-world applications. Unanticipated input can lead to false positives (FP) and false negatives (FN), potentially resulting in false alarms in fire detection scenarios. Literature on fire detection models shows varying levels of complexity and explicability in evaluation practices; little supplementary information on performance ability outside of accuracy scores is provided. We advocate for a standardized evaluation dataset that prioritizes the end-user perspective in assessing performance capabilities. This leads us to ask what an evaluation dataset needs to constitute to enable a non-expert to determine the adequacy of a model's performance capabilities for their specific use case. We propose using data augmentation techniques that simulate interventions to remove the connection to the original target label, providing interpretable counterfactual explanations into a model's predictions.

Keywords

Wildfire detection, supervised learning, causality, evaluation

INTRODUCTION

Wildfires cause tremendous damage globally. They destroy ecosystems, harm human lives and property, and pollute the air. Approximately 463 million hectares of forest are burned annually worldwide (Shi and Touge, 2022). About 15% of the global greenhouse gas emissions into the atmosphere are resulting from wildfires. High temperature, dry soil and low air humidity increase the probability of their occurrence. Surface materials like grass, leaves and twigs usually ignite first, and once the fire spreads to the crowns of trees, it can be difficult or even impossible to control (Hirschberger, 2016). Therefore, early wildfire detection is of substantial use for firefighting. Traditionally, human observers detect wildfires (Guth et al., 2005). However, this is costly and inefficient as the area to be covered is vast. Technologies for automatic early wildfire detection are pivotal. Various AI-based solutions exist, e.g., detecting smoke or fire in images, evaluating satellite images, and analysing large temperature datasets.

This paper is exclusively concerned with machine learning (ML) models that use smoke and fire images in the visible light spectrum. While human intelligence can label such images with almost impeccable accuracy, the classification with AI often leads to false FPs and false FNs. A phenomenon that leads to misclassifications in ML models is the presence of objects that are coloured like fire. Numerous supervised learning models for fire detection on images achieve high classification accuracy, some over 99% (Khan and Khan, 2022; Lee et al., 2017; Oh et al., 2020). However, it is hard to compare the performances of these algorithms since there is no consistent performance evaluation standard. The high accuracies these algorithms achieve on their respective test data give little insight into whether they would succeed in real-world applications. It is difficult for non-expert end-users to decide if an algorithm is reliable and secure enough for their use case. We aspire to describe an evaluation dataset that allows for a comparison between fire detection models. Such an evaluation dataset should provide insight into whether a model is reliable and secure. It should also be interpretable for non-experts without requiring access to a model's training data or architecture.

Reviewing existing wildfire detection models, we identified a common issue. Some features learnt to predict the labels are not causally related to fire. Such non-causal features can lead to FPs and FNs when the model tries to predict labels for out-of-distribution (OOD) images. This paper describes performance issues that arise when a model does not fully learn causal features. We suggest building a general evaluation dataset that gives insight into whether causal features have been learnt by answering "counterfactual" questions of the following form. What would the model predict if ...

- … the fire's texture is changed to the texture of metal?
- … the fire is shaped like a teddy bear?
- … the fire's colour is pink?

Questions of this kind give insight into which high-level causal features the model uses for its predictions. We hope that the outlined general evaluation dataset can be built and then adapted as a benchmark, enabling the development of more stable models. The performance of the evaluation dataset would be easily interpretable. For instance, whether a model learns the colour of a fire, or its shape would be apparent by the predicted labels and the distinct content of the images. By linking data augmentation to "interventions", we claim that we can create such a dataset. However, creating it will be part of future research.

RELATED WORK

Image-based Fire Detection Technology

Modern technologies provide new solutions for the task of predicting and detecting wildfires. With the vast amount of data available, ML applications can be implemented to detect fires automatically. Input parameters for such ML applications include humidity, wind speed, temperature, season, time of day, the number of trees in an area and the composition of the forest surface (Calp and Kose, 2020; Castelli et al., 2015). However, this paper will focus on fire detection and ML algorithms that use images as training data. Images used to train models to come from different sources, including satellites, ground cameras and uncrewed aerial vehicles (UAVs) (Barmpoutis et al., 2020). Satellite images are utilised to identify environments from a considerable distance. This technology can cover a vast area (Giglio et al., 2016), but small fires are rarely visible in satellite images. These small fires are accountable for a significant proportion of the areas burned worldwide (Hu et al., 2021). In addition, the quality of satellite images is often affected by bad weather conditions, which can obscure vision.

Stationary Ground Cameras automatically recording images in the forest are a complementary option. Many sensors and cameras are set up on elevated points such as towers (Mohapatra and Trinh, 2022). Such images complement the satellite images since they are closer to the ground and less likely to be obscured by bad weather conditions. However, fog, dust, clouds, and smoke can obscure the flames, even from ground cameras. This can lead to errors in the localisation of the fire or a poor estimate of the size of the affected area. Therefore, helicopters or aeroplanes approximate the extent of a wildfire. This is expensive and poses a risk to the people in aerial vehicles. Using UAVs reduces such risks while being less costly. Real-world wildfire operations have already been conducted with single vehicles or a fleet of multiple vehicles. Experiments have shown that UAVs are useful in firefighting scenarios by bridging the gap between measurements from satellite- and ground-camera-based systems (Merino et al., 2012; Srinivas and Dua, 2019).

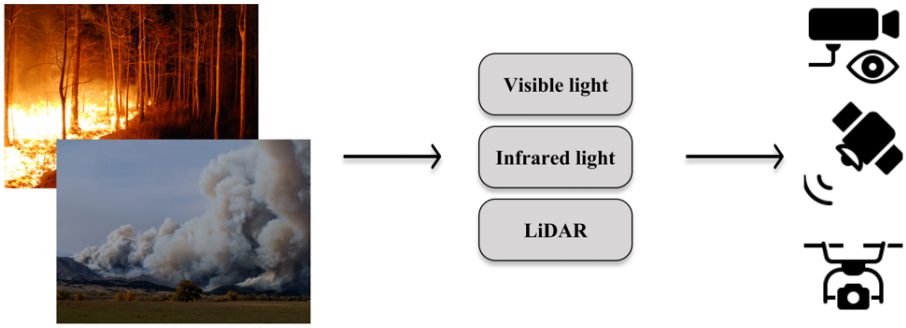

Figure 1. Images of smoke or fire are detected with satellites, ground cameras and UAVs. They record visible light, infrared light or utilise the LiDAR technology

All three technologies provide images a neural network can process to detect fire. However, not all pictures are recorded in a light spectrum visible to the human eye. Three categories of images are common:

- Images in spectra visible to the human eye.
- Infrared images, including heat signatures.
- Light Detection and Ranging (LiDAR). This remote sensing method uses light as a pulsed laser that reflects from smoke or other particles.

Automated Wildfire Detection Models

This paper will exclusively consider images in spectra visible to the human eye. Human-performed visual interpretation of photos is time-consuming and costly. On the other hand, automated understanding of images by ML models is inexpensive and can significantly reduce the time spent on evaluations (Labenski et al., 2022). However applicable ML models are, many models predict FPs and FNs, which would not happen to humans. Such images include shadows from clouds, dust, fog or even human equipment such as vehicles, clothing or road signs (Alkhatib, 2014). Further FNs we found in the literature are images of wildfires in the background (Khan, 2022). Some FPs include autumn foliage lamps and car lights (Khan and Khan, 2022; Park et al., 2022; Sun et al., 2021). The model described by Field Park et al. (2022) can correctly identify many of the mentioned FNs and FPs. In most cases, the model can differentiate light emitted by lamps or cities from the light emitted from a fire at night. The authors used standard data augmentation methods like brightness, saturation, rotation and mirroring. They also used generative adversarial networks (GAN) to augment images from day to night, change the seasonality to winter, and create additional synthetic wildfire images. Lu et al. (2022) introduced a diagonal swap of random origin as a data augmentation technique. This improved their model's ability to detect small fires. Data augmentation techniques all serve the same purpose: expanding the domain, a model is trained on, by increasing the training data.

Evaluating Wildfire Detection Models

A significant amount of research on fire detection models needs more insight into the domain in which the model was trained. Furthermore, many research papers do not provide examples of FPs or FNs. Consequently, evaluating the quality of such models is difficult. Moreover, models are typically evaluated using data that is similar to the training dataset. For instance, if a model is trained exclusively on daytime images, its performance will likely be poor when tested on night-time images. However, the average accuracy will likely be high if no night-time images are included in the evaluation dataset. Several approaches have been developed in the field of explainable AI (XAI) to address this limitation of performance score evaluation. XAI aims to understand and interpret AI systems to provide insights into their trustworthiness, fairness, and robustness. We refer to Linardatos et al. (2020) for an extensive review of XAI approaches.

Saliency maps and counterfactuals are example-based approaches within XAI that help identifying potential biases in a model's decision-making process. Saliency maps determine the most critical regions or features in an image used by a neural network to make a particular classification decision. This information can be used to explain why a specific image was classified in a certain way and helps identify potential biases in the network's decision-making process. Counterfactuals are used to explain why a neural network did not make a certain prediction[1]. They can be used to explore what-if scenarios and help identify improvement areas in the network's decision-making process.

However, XAI approaches have rarely been used in the literature reviewed, and only a few studies provided saliency maps. Wildfire detection models can only be evaluated based on the information shared by the developer or by thoroughly testing a model's prediction abilities. Thus, it can be challenging for non-experts to evaluate them. We therefore argue for the need of an evaluation benchmark dataset. If such a dataset exists, model performance can be compared, which would enforce a certain level of complexity in evaluation. We plan to develop an evaluation dataset that can indicate a model's trustworthiness by measuring the extent to which it has learned causal features. Performance on this dataset will be challenging but could encourage the research community to focus on models suitable for real-world applications and shift the focus from performance-score-focused to end-user-oriented development.

[1] See Peters et al. (2017) for a definition of counterfactual.

In the following sections, we argue that an ML model needs to perform well on OOD data to be safely applied in the real world. We connect "spurious correlations" and vulnerability to adversarial attacks to a lack of learnt causal features, which we identify as a requirement for reliable and secure models. Based on this, we propose data augmentation on images for a general evaluation dataset which measures to what extent a model has learnt causal features. These images can be seen as counterfactual explanations. The novelty is that we apply data augmentation on an evaluation dataset. We aim to develop a data augmentation technique from Pearl's notion of causality using structural causal models and generative networks. Our approach is distinct from other techniques for counterfactual explanations since other techniques are often not connected to Pearl's theory of causality (Chou et al., 2022).

METHODOLOGY

We conducted a comprehensive literature review, collecting 76 articles published between 2012 and 2022 on automated fire detection. Relevant search terms were used in Google Scholar to gather related articles. We identified papers within the collected articles introducing new fire or wildfire detection models. After assessing the relevance of the articles, 37 out of the initial 76 were retained for further analysis. For each selected article, we examined various criteria, including the following: performance score, data used, machine learning task, common FPs and FNs and other explanations such as saliency maps. We observed that most models are concerned with supervised learning tasks such as detection, classification, and segmentation, where each input has a label. The following definitions are adaptations from Shen et al. (2021).

Definition. Let $\mathcal{X}, \mathcal{Y}$ be sets of random variables with $Im\, X \subset \mathbb{R}^n$ and $Im\, Y \subset \mathbb{R}^m$ for every $X \in \mathcal{X}, Y \in \mathcal{Y}$ and some $n, m \in \mathbb{N}$. Let $P(X, Y)$ be the joint probability distribution of the elements in $\mathcal{X}$ and $\mathcal{Y}$. We call $\mathcal{X}$ the feature space and $\mathcal{Y}$ the label space. A *parametric model* is a map $f_\theta: \mathcal{X} \to \mathcal{Y}$ for learnable parameters $\theta \in \mathbb{R}^k$.

Definition. Let $\ell: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a loss function, that sums up the differences between the predicted labels and the ground truth of the labels. Partition the data from the distribution $P(X, Y)$ into a training set with distribution $P_{tr}(X, Y)$ and a test set with distribution $P_{te}(X, Y)$. A *supervised learning problem* is to find a parametric model f_θ^*, such that $\mathbb{E}[\ell(f_\theta^*(X), Y)]$ is minimal for the elements in the test set with distribution $P_{te}(X, Y)$ and the ground-truth $Y \in \mathcal{Y}$.

Publicly available wildfire models were tested on Hugging Face to identify additional cases of FPs and FNs. By combining the abovementioned approaches, we identified common FP and FN issues and noticed that evaluation criteria differed in complexity and explainability. Evaluation data was often simplistic and rarely tested on OOD data. Additional explanations, such as FPs/FNs or saliency maps, were also scarcely provided. This led us to ask: What should an evaluation contain to ensure to the end user that the model is reliable? We generally describe this aspect as a model's trustworthiness. We examined how to determine if a model would work in the real world, where it is likely to be presented with images distinct from the training distribution. This led us to identify the necessary evaluation conditions for a model to be considered trustworthy in real-life scenarios. In the subsequent sections, we will further discuss trustworthiness and explore how learning causal features indicates how well a model performs on unseen data. This discussion will lead us to argue that an evaluation dataset should consist of images that test whether a model has learned causal visual features.

TRUSTWORTHINESS

A machine learning model's reliability is commonly linked to an evaluation dataset's performance score (e.g., the number of correct predictions). However, model performance scores vary drastically depending on the dataset they were evaluated on. High accuracy on an evaluation dataset is meaningless if this dataset is too simple. Most models do not generalise well and perform poorly on unseen OOD data. The *general out-of-distribution problem* is a special case of a supervised learning problem in which the test distribution $P_{te}(X, Y)$ yields data that is significantly different from the data obtained from the training distribution $P_{tr}(X, Y)$.

A distribution shift can occur when the domain changes. A model trained on a specific domain of forest fire images may be able to detect fires from a similar domain with high accuracy but can still perform poorly on images from a different domain. Such dissimilar domains can provide data from unseen geographical regions or data collected at different times of the day. Different backgrounds, camera positions, image resolutions or weather can represent domain shifts. OOD data also includes objects or scenes which do not exist in the training data. The performance on OOD data is very important for a real-world application of AI models. Crowd-sourced images can differ vastly from those in the model's training data. For emergency operators to be able to rely on automated fire detection models, these models need to work well on unknown input data.

Consequently, we need a better way of evaluating a model's performance. Therefore, we link a model's reliability to whether it generalises well on OOD data. The FNs and FPs we found in the literature are examples of models failing to generalise on OOD data. The cause of a FP can be a model learning a correlation between the label and some of its relevant features, but not all of them. For example, autumn foliage is classified as fire because colour features are used over shape and texture.

Furthermore, FPs can happen because features not related to fire are learnt to predict labels. This is often called a spurious correlation. For example, a chipmunk in a forest environment might be classified as fire, but a chipmunk in a bathroom would not.

Additionally, we require models to be secure. We link a model's security to its robustness to adversarial attacks. Adversarial attacks are malicious attacks that use modified data to manipulate ML models in the attacker's interest (Rouani et al., 2019). Security is essential in the public sector because AI systems can be used to make important decisions, such as automatically activating emergency services. These decisions should not be affected by malicious adversaries. For instance, an adversary could alter the input image data in a forest surveillance system slightly and, thus, cause a false alarm and a costly firefighting operation. Similar examples of adversarial attacks can be found in (de Mello, 2020). For AI models to be applicable in the public sector, they should be generalisable and robust towards adversarial attacks to ensure their trustworthiness. To our best knowledge, current standard evaluation practices do not provide this insight. In the next section, we describe how causality has been inked to OOD generalisation and robustness towards adversarial attacks.

CAUSALITY

Structural Causal Models (SCM) describe a causal relationship with a function, e.g., $x = f_X(y, u_1)$. For an extensive review of SCMs, we refer to (Pearl, 2009; Peters et al., 2017). $X \rightarrow Y$ denotes a causal relationship, where X is the cause of Y. Consider the Directed Acyclic Graph (DAG) in Figure 2, depicting such a causal relationship. We use capital letters (e.g., X,Y,Z) for random variables and lowercase letters (x,y,z) as generic symbols for specific values taken by the corresponding random variables.

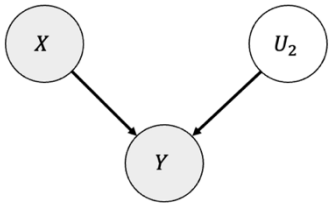

Figure 2. DAG depicting causal relationship between random variables

The following equations show an interpretation of the diagram using two unknown functions:

$$x = f_X(u_1)$$
$$y = f_Y(x, u_2) \qquad \text{Eq. (1)}$$

The u_i are the outcome of the exogenous variables U_i whose causal factors are unknown and kept unexplained, and x and y are observations from the endogenous variables X and Y, respectively. Eq. (1) suggests that if x is changed, so is y. Thus, the related random variables have a statistical and a causal relationship. Additionally, any variable missing from the right-hand side, like u_1 in f_Y, is said not to influence the result of the left-hand side, provided the other variables remain constant. In image classification, the image observations x_i are realisations of X, which is the effect of the target Y, so the causal relationship is $Y \rightarrow X$. Consequently, image recognition is an anti-causal problem because one tries to predict y from x, e.g., $p(y|x)$ (Schölkopf et al., 2012). This is the opposite of the causal direction, where x is caused by y. The cause of a target fire can be found in other exogenous factors, such as chemical processes leading to heat and light being emitted. This gives cause to an image x. The high-level features denoted by h_y are all effects of this process. If this cause-and-effect relationship can be learnt, one can recognise fire on an image in many contexts and differentiate its high-level features from those caused by other factors. As explained in the previous section, AI models rarely perform well on differentiating these cause-and-effect relationships.

Spurious correlations occur when two variables are correlated, but there is no causal relationship between them. Instead, the correlation is caused by a third confounding variable. For example, a confounding variable could be that most wildfire images were taken in one specific region during one season, e.g., in Australia.

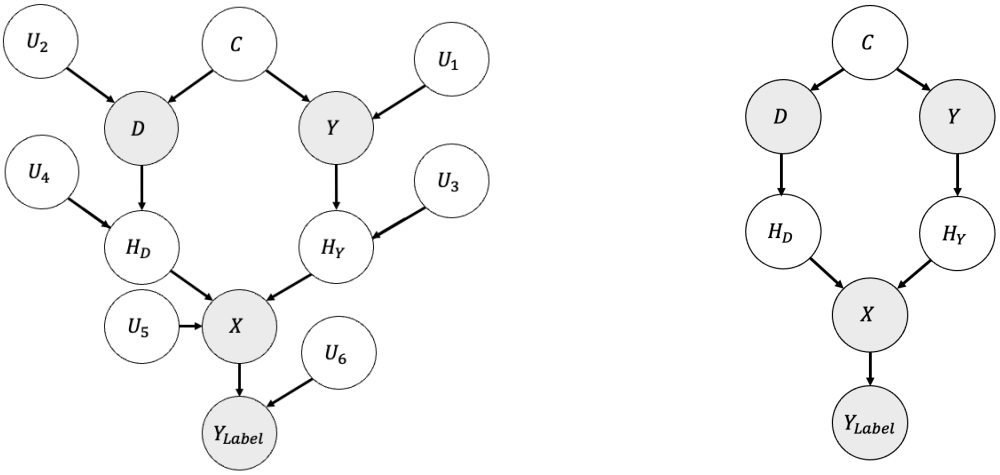

Figure 3. DAG depicting the domain generalisation problem (left) with exogenous variables U_i **(right)** simplified version with exogenous variables U_i omitted

Figure 3 depicts a DAG of a confounding variable. Here, D stands for a domain (e.g., a specific region, time-of-year, camera), Y is the target (e.g., fire), H_D are high-level features caused by D, and H_Y are high-level features caused by Y. The identifier X represents a set of images caused by H_D and H_Y. We used a similar DAG to the one presented in (Ilse et al., 2020) but included Y_{label} caused by X and exogenous random variables denoted by U_i. Here we assume that the target labels are produced by a human annotator who correctly identifies the effects of Y in an image. The following are the structural equations associated with the DAG in Figure 3.

$$
\begin{aligned}
y &= f_Y(c, u_1) \\
d &= f_D(c, u_2) \\
h_y &= f_{H_Y}(y, u_3) \\
h_d &= f_{H_d}(d, u_4) \\
x &= f_X(h_d, h_Y, u_5) \\
y_{label} &= f_{Human}(x, u_6)
\end{aligned}
\qquad \text{Eq. (2)}
$$

The existence of C leads to D and Y being no longer independent (Pearl, 2009). Moreover, since H_D is a child of D it is also spuriously correlated with Y. This means any model trained to predict Y from X can rely on both H_D and H_Y to predict Y_{label}. Unsurprisingly, a model trained on such data would predict the presence of fire when presented with an image of autumn foliage. If a model only uses h_y for its predictions, such FPs would not occur. Mahajan et al. (2022) also call causal features stable features. The authors have shown that they are reliable indicators for robustness towards several adversarial attacks while also generalising well on OOD data. It has been argued that DNN models are susceptible to adversarial attacks due to a lack of causal understanding (Kilbertus et al. 2018; Zhang et al. 2020). Consequently, we argue that learning causal features decreases the appearance of spurious correlations and increases robustness towards adversarial attacks while improving OOD performance. Thus, learning causal features makes a model more reliable and secure, according to our definition of reliability and security. As mentioned, predicting y from x is an anti-causal problem (Schölkopf et al. 2012).

Nevertheless, Kilbertus et al. (2018) have shown that for some deep learning algorithms (e.g. generative Models), it is possible to approximate the causal direction by doing an exhaustive search over samples of the input feature space of the anti-causal direction, but not for others (e.g. CNN). This remains a computationally hard problem. Talking about cause and effect is difficult in the case of conventional object detection, which heavily relies on

CNN-based models. Nevertheless, looking at fire detection from the causal perspective is beneficial. For instance, Sun et al. (2021) use intervention to suppress lamp disturbance. An intervention in a machine learning model is a change in the outcome of a random variable, such that the new outcome is determined. For example, consider a random variable X that yields "heads" or "tails", depending on some random input. An intervention could cause the outcome to be decisively "heads". An intervention is denoted using the $do()$ operator, e.g., $do(X = heads)$. Such an intervention deletes the edge to parent nodes in the corresponding DAG. In Eq. (2), if X is changed, it does not affect h_y or h_d. If the value of h_d is changed, it also changes the value of x. Therefore, if two variables are independent, we can do an intervention that renders the observational distribution independent of the domain $p(x, y|do(d)) = p(x, y)$.

Ilse et al. (2021) have shown that data augmentation can simulate such interventions. This implies that with the right data augmentation, one can simulate the interventional distribution $p(x, y|do(d))$. A model trained on such data learns more of h_y features and performs better on OOD data. This aligns with our observations that fire detection models trained with a vast set of domains or data augmentation techniques have fewer FPs and FNs.

FIRE EVALUATION DATASET

The research presented in this paper aims to describe standards which an evaluation dataset for wildfire detection models should meet. A model's performance on such a dataset could be meaningful for real-world applications. However, producing an evaluation dataset that meets the described standards is out of the scope of this paper but could be realised in future works. Knowing that learning causal features leads to more reliable models, we ask: "To what extent has the model learnt causal features?". We want to use interventions and the corresponding data augmentation simulation to answer this question. Consider the following two interventions in the data generation process from Figure 3 depicted by Figure 4:

$$do(H_Y = h_1) \quad \text{Eq. (3.1)}$$

$$do(H_D = h_2) \quad \text{Eq. (3.2)}$$

As stated earlier, an intervention removes the causal link to its parents. Since the causal features h_y have been intervened on, an image coming from 3.1. (e.g., $x_1 = f_X(h_d, h_1)$) will have fewer causal features than an image coming from 3.2 (e.g., $x_2 = f_X(h_d, h_2)$). When predicting the label y from x using causal features, the probability $p(Y_{label} = fire\ |X = x_i)$ should decrease for x_1 and remain unchanged for x_2. Based on Ilse et. al (2020) we argue that it is possible to develop data augmentation techniques that simulate the interventions above. Hereafter, we denote images simulating Eq. (3.1) with $aug_{Hy}(x)$ and images simulating Eq. (3.2) with $aug_{Hd}(x)$. Assuming that fire is present in an original image x, a model which relies mainly on causal features would predict the presence of fire in x and $aug_{Hd}(x)$ with the same probability and $aug_{Hy}(x)$ with a lower probability. Note that intervening on all high-level features in H_Y would change the label.

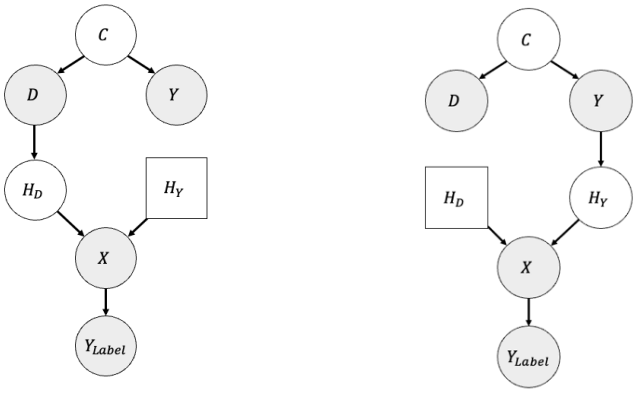

Figure 4. Simplified DAG where exogenous variables are omitted
(left) intervention on H_Y **(right)** intervention on H_D

Figure 5 is an original image of a forest fire without augmentation. Figure 6 contains examples demonstrating how data augmentation can simulate an intervention in the fire image generation process. For example, consider the shape of a fire. It could be ignited in a controlled object shaped a certain way which releases light. A fire that

is coloured in an unusual manner could result from adding certain chemicals. The textured fire image can be obtained by using a special camera with a reactive heat filter. However, all these examples could only be produced in a controlled environment and, statistically speaking, are very unlikely to occur in a wildfire scenario and hence would be gathered to obtain non-wildfire examples in the data generation process. Therefore, in all these cases, the label of the original image (Figure 6) would no longer be "wildfire".

We require the evaluation dataset to contain images with data augmentation simulating interventions on at least three visual features: shape, colour, and texture. Only images that have not been "intervened on" in such a way will retain their original label.

Figure 5. Original Wildfire Image

These images were created manually, and we leave the development of a model which generates such augmentations for future research. As stated earlier, a good performance on the interventional distribution $p\big(X\big|do\big(H_Y = h_y\big), H_D = h_d\big)$ is insufficient for evaluation; a model trained on a completely different task could perform well. Thus, an evaluation dataset also needs to contain images of fire with various backgrounds e.g., images drawn from the probability distribution $p\big(X\big|H_Y = h_y, do(H_D = h_d)\big)$. Figure 7 shows examples of such images: Fire caused by wood burning (our target value) in various backgrounds, angles, seasons, and time-of-day. Finally, to further test for OOD generalisation, consider images with the same or similar anti-causal effects as fire such as the images depicted in Figure 8.

Figure 6. Counterfactual images of fire simulating interventions on
(left) shape **(centre)** texture and **(right)** colour

Figure 7. Fire images with various domain changes

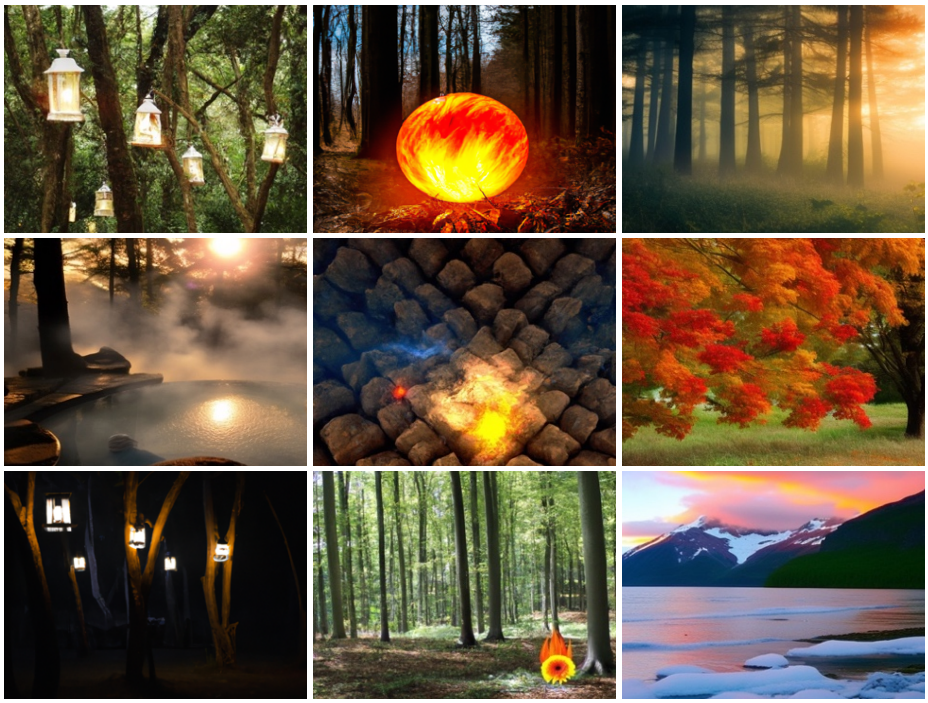

Figure 8. Images with anti-causal effects like fire

There are several advantages to an evaluation dataset as we described it. The performance of a model can be tested without requiring access to the prediction function. Since we augment high-level features, the results are interpretable even by non-experts. Possible observations are: “A model mainly uses colour features to make its predictions”, “The model does not consider shape or texture”, “Objects that illuminate their surroundings are common FPs”, and “A model fails to recognise small fires during daytime reliably”. People developing models can use this information to improve them. An end-user can use this information to decide whether the model is reliable and secure enough for their personal use case.

Furthermore, since we are specifically interested in the performance of a model on OOD data, the evaluation dataset must be created independently of a model’s training and testing data. This means that once such a dataset

has been created, it can be used to evaluate any model, making it possible to compare the performance between different models. Finally, if such a dataset is adapted as an evaluation benchmark, the research would be encouraged to develop models suitable for real-world application.

LIMITATIONS

Without knowledge about the causal structure of the data generation process and the high-level causal features, it will be challenging to define augmentation techniques that simulate the intervention $do(H_y)$. Additionally, there may be features h_d causing an image that would be mistaken for fire, even by a human. However, evaluating whether a model performs on par with human deduction ability is already of great value. From the end-user perspective, knowing this could be reassuring and offers more explanation of performance than merely reporting accuracy on test data.

Creating a data augmentation technique simulating the interventional distribution $p\big(X\big|H_Y = h_y, do(H_D = h_d)\big)$ is complex in the case of wildfire since there is no clear distinction between foreground and background (Sauer and Geier, 2021). We have not included the anti-causal effects of fire in our causal model but try to mitigate this by including images of objects with similar anti-causal effects in the evaluation dataset.

CONCLUSION

Current evaluation practices need to give more insight into whether a fire detection model can be applied reliably in the real world, where models are subjected to OOD data in the form of unseen objects, unusual events, or security threats. This can cause many FPs or FNs, potentially leading to false or missed alarms. Before an ML model is employed, the end-user should know how a model is making its predictions. With this information, one can decide how much a system is sufficient. We identified causal features to improve a model's OOD generalisation ability and robustness towards adversarial attacks. Based on this, we propose data augmentations derived from causal intervention. Those data augmentations can be used to measure to what extent a model has learnt causal features. From this, an evaluation dataset can be constructed and used independently of particular models and their training data. With such a dataset, one can compare and investigate model performance in terms of the ability to learn causal features. This indicates a model's reliability and security. In the next steps of our research, we plan to build a counterfactual generative model building on the work of Sauer and Geier (2021) and Ilse et al., (2021). We will use FASDD (Fire) and ImageNet (Non-Fire classes) datasets for this.

REFERENCES

Alkhatib, A. A. A. (2014). A Review on Forest Fire Detection Techniques. *International Journal of Distributed Sensor Networks*, 10 (3): 597368. https://doi.org/10.1155/2014/597368.

Sauer, A. and Geier, A. (2021). Counterfactual Generative Networks. *International Conference on Learning Representations (ICLR)*.

Barmpoutis, P., Papaioannou, P., Dimitropoulos, K.and Grammalidis, N. (2020). A review on early forest fire detection systems using optical remote sensing. *Sensors, 20*(22), 6442.

Calp, M. H. and Kose, U. (2020). Estimation of burned areas in forest fires using artificial neural networks. *Ingeniería Solidaria, 16*(3), 1-22.

Castelli, M., Vanneschi, L. and Popovič, A. (2015). Predicting burned areas of forest fires: an artificial intelligence approach. *Fire ecology, 11*(1), 106-118.

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

de Mello, F. L. (2020). A survey on machine learning adversarial attacks. *Journal of Information Security and Cryptography (Enigma), 7*(1), 1-7.

Giglio, L., Boschetti, L., Roy, D., Hoffmann, A. A., Humber, M. and Hall, J. V. (2016). Collection 6 modis burned area product user's guide version 1.0. *NASA EOSDIS Land Processes DAAC: Sioux Falls, SD, USA*.

Guth, P. L., Craven, T., Chester, T., O'Leary, Z. and Shotwell, J. (2005). *Fire location from a single osborne firefinder and a dem.* Paper presented at the ASPRS 2005 Annual Conference Geospatial Goes Global: From Your Neighborhood to the Whole Planet.

Hirschberger, P. (2016). Forests ablaze: causes and effects of global forest fires. *WWF: Berlin, Germany) Available at https://www.wwf. de/fileadmin/fm-wwf/Publikationen-PDF/WWF-Study-Forests-Ablaze. pdf [Verified 26 March 2020]*.

Hu, X., Ban, Y. and Nascetti, A. (2021). Uni-temporal multispectral imagery for burned area mapping with deep learning. *Remote Sensing, 13*(8), 1509.

Ilse, M., J. M. Tomczak, and P. Forré. (2021). Selecting Data Augmentation for Simulating Interventions. *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, M. Meila and T. Zhang, eds., 4555–4562. PMLR.

Khan, S., and A. Khan. (2022). FFireNet: Deep Learning Based Forest Fire Classification and Detection in Smart Cities. *Symmetry*, 14 (10): 2155. https://doi.org/10.3390/sym14102155.

Kilbertus, N., G. Parascandolo, and B. Schölkopf. (2018). Generalization in anti-causal learning. *arXiv preprint arXiv:1812.00524*.

Labenski, P., Ewald, M., Schmidtlein, S. and Fassnacht, F. E. (2022). Classifying surface fuel types based on forest stand photographs and satellite time series using deep learning. *International Journal of Applied Earth Observation and Geoinformation, 109*, 102799.

Lee, W., Kim, S., Lee, Y.-T., Lee, H.-W. and Choi, M. (2017). *Deep neural networks for wild fire detection with unmanned aerial vehicle.* Paper presented at the 2017 IEEE international conference on consumer electronics (ICCE).

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy, 23*(1), 18.

Lu, K., J. Huang, J. Li, J. Zhou, X. Chen, and Y. Liu. (2022). MTL-FFDET: A Multi-Task Learning-Based Model for Forest Fire Detection. *Forests*, 13 (9): 1448. https://doi.org/10.3390/f13091448.

Mahajan, D., S. Tople, and A. Sharma. 2022. "The Connection between Out-of-Distribution Generalization and Privacy of ML Models."

Merino, L., Caballero, F., Martínez-De-Dios, J. R., Maza, I. and Ollero, A. (2012). An Unmanned Aircraft System for Automatic Forest Fire Monitoring and Measurement. *Journal of Intelligent & Robotic Systems, 65*(1-4), 533-548. doi:10.1007/s10846-011-9560-x

Mohapatra, A. and Trinh, T. (2022). Early Wildfire Detection Technologies in Practice—A Review. *Sustainability, 14*(19), 12270.

Oh, S. H., Ghyme, S. W., Jung, S. K. and Kim, G.-W. (2020). *Early wildfire detection using convolutional neural network.* Paper presented at the Frontiers of Computer Vision: 26th International Workshop, IW-FCV 2020, Ibusuki, Kagoshima, Japan, February 20–22, 2020, Revised Selected Papers 26.

Park, M., D. Q. Tran, J. Bak, and S. Park. (2022). Advanced wildfire detection using generative adversarial network-based augmented datasets and weakly supervised object localization. International Journal of Applied Earth Observation and Geoinformation, 114: 103052. https://doi.org/10.1016/j.jag.2022.103052.

Pearl, J. 2009. Causality. Cambridge university press.

Peters, J., D. Janzing, and B. Schölkopf. (2017). Elements of causal inference: foundations and learning algorithms. The MIT Press.

Rouani, B. D., Samragh, M., Javidi, T., & Koushanfar, F. (2019). Safe machine learning and defeating adversarial attacks. *IEEE Security & Privacy, 17*(2), 31-38.

Schölkopf, B., D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. (2012). On Causal and Anticausal Learning. *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, 459–466. Madison, WI, USA: Omnipress.

Shen, Z., J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui. (2021). Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*.

Shi, K. and Touge, Y. (2022). Characterization of global wildfire burned area spatiotemporal patterns and underlying climatic causes. Sci. Rep. 12. In.

Srinivas, K.and Dua, M. (2019). *Fog computing and deep CNN based efficient approach to early forest fire detection with unmanned aerial vehicles.* Paper presented at the International Conference on Inventive Computation Technologies.

Sun, K., Q. Zhao, and X. Wang. (2021). Using knowledge inference to suppress the lamp disturbance for fire detection. *Journal of Safety Science and Resilience*, 2 (3): 124–130. https://doi.org/10.1016/j.jnlssr.2021.07.002.

Zhang, C., K. Zhang, and Y. Li. (2020). A causal view on robustness of neural networks. *Advances in Neural Information Processing Systems*, 33: 289–301.