World Academy of Science, Engineering and Technology
International Journal of Nuclear and Quantum Engineering
Vol:17, No:1, 2023

Resilient Machine Learning in the Nuclear Industry: Crack Detection as a Case Study

Anita Khadka, Gregory Epiphaniou, Carsten Maple

Abstract—There is a dramatic surge in the adoption of Machine Learning (ML) techniques in many areas, including the nuclear industry (such as fault diagnosis and fuel management in nuclear power plants), autonomous systems (including self-driving vehicles), space systems (space debris recovery, for example), medical surgery, network intrusion detection, malware detection, to name a few. Artificial Intelligence (AI) has become a part of everyday modern human life. To date, the predominant focus has been developing underpinning ML algorithms that can improve accuracy, while factors such as resiliency and robustness of algorithms have been largely overlooked. If an adversarial attack is able to compromise the learning method or data, the consequences can be fatal, especially but not exclusively in safety-critical applications. In this paper, we present an in-depth analysis of five adversarial attacks and two defence methods on a crack detection ML model. Our analysis shows that it can be dangerous to adopt ML techniques without rigorous testing, since they may be vulnerable to adversarial attacks, especially in security-critical areas such as the nuclear industry. We observed that while the adopted defence methods can effectively defend against different attacks, none of them could protect against all five adversarial attacks entirely.

Keywords—Resilient Machine Learning, attacks, defences, nuclear industry, crack detection.

I. Introduction

THE nuclear industry consists of complex infrastructures and processes. Since the beginning, the industry has been regulated by its comprehensive security and safety policies. The stringent security policies highlight the extreme and challenging environment of the industry and the reason behind the slow adaptation to new technological advancements. For example, nuclear facilities still use analogue systems that were first built in the 1950s [1]. However, in the meantime the other sectors have had major developments such that the potential benefits can no longer be avoided. For example, with the progress of AI and ML techniques, several tedious and time-consuming jobs can be handed over to intelligent machines from human operators including smart record-keeping, video surveillance, data processing, and automation. However, the consequences of attacks (e.g., targeted or indirect) in the industry can be fatal and can result in the loss of human lives and also finance. Although the diverse adversarial attacks on ML models may be unavoidable, however, with the adoption of resilient-by-design principle while building any system can defend from such attacks.

The nuclear industry comprises of different structures including, nuclear power plants, reactors, fuel facilities, uranium enrichment plants, spent fuel storage facilities, and power grid. With such multifaceted structures, the attack surface can also be diverse. This can provide opportunities to attackers to exploit vulnerabilities and weaknesses. For example, if the physical security is disabled, commando-like ground-based attacks on equipment could lead to a reactor core meltdown or widespread dispersal of radioactivity, and external attacks such as crash into a reactor, or cyber-attacks [2].

We analysed different methods proposed over the years to generate adversarial attacks and defence mechanisms against those attacks. As an exemplar, we focus on crack detection on concrete walls. In general, an inspection of any structure involves a visual examination by an inspector if the structure is accessible and small scale. However, for the large and sensitive infrastructures like nuclear facilities, automatic video inspection of the structure can help monitor the walls continuously. For this, a mechanical robot carries out a video recording of the structure, and after the video shooting, a human inspector will investigate the damages or anomalies by rigorous inspection of the captured video [3]. This process can be tediously filled with human errors because of monotonous and lengthy work patterns. It is reported that about 60% of major failures in the Nuclear Power Plant (NPP) are caused by human errors [4]. There has been an effort on eliminating human errors by focusing on automation in the industry. With the help of ML techniques, such tedious works can be automated and efficiently performed. However, the industry's strict policy and regulation have led to slow progress in maturing security requirements governing automated operations in these environments.

Since nuclear facilities are security and safety prone due to the involvement of high-risk factors like nuclear bombs and nuclear weapons. Thorough checks and rigorous evaluation of any automation systems should be carried out before entirely relying on them.

By analysing several adversarial and defensive ML techniques, our contributions are following:

- We conduct a comprehensive study of Resilient Machine Learning (rML) in the nuclear industry with a case study of detecting cracks on the concrete walls.
- We compare different ML techniques to generate adversarial examples for detecting cracks on the concrete walls.
- We analysed various defensive strategies to make the learning methods resilient against the adversarial attacks.

The paper is structured as follows. We present related work on detecting cracks on the walls and Adversarial Machine

A. Khadka, G. Epiphaniou and C. Maple are with the Secure Cyber Security Research Group (SCSRG), WMG, University of Warwick, United Kingdom (e-mail: anita.khadka@warwick.ac.uk, gregory.epiphaniou@warwick.ac.uk, cm@warwick.ac.uk).

Learning (aML) in crack detection models in Section II. In Section III, we present an overview of aML which comprises a threat model in machine learning and different types of adversarial attacks and finally methods to generate and assess adversarial examples. Section III-B describes how to defence the adversarial attacks. We present the utilised dataset in Section IV-A and experimental results in Section IV-C. Finally, we discuss the results and conclude the work in Section V.

II. Detecting Cracks on the Walls

There have been many works on adversarial attacks and defences on the ML models. However, the research on investigating aML in the nuclear industry is still in the infancy stage. The progression of application of ML methods have moved towards security sensitive areas, including autonomous driving, nuclear facilities and medical procedures where human lives can be at stake if something goes wrong. Therefore, the initiatives of research towards resilient ML have also been slowly emerging in such areas.

Even though there is a lack of research on injecting adversaries for crack detection on the walls of nuclear facilities, several works that have studied ML methods for crack detection for roads, building and also applying adversaries on them. Cracks in any infrastructure are dangerous, and in nuclear power plants, they can be catastrophic due to the involvement of hazardous elements like uranium and plutonium. If cracks are not spotted early, the consequences can be expensive. For instance, in 1996, a leaking valve caused an accident in the Millstone Nuclear Power Station in Waterford, CT, USA, which cost $254 million [5]. In 2008 Fukushima province, Japan, A 7.2 magnitude earthquake cracked reactor cooling towers and spent fuel storage facilities, spilling 19 litres of radioactive wastewater, and damaging the Tokyo Electric Power Company's No. 2 Kurihara Power Plant, which was worth $45 million damages. Likewise, in 2010, leaked radioactive tritium from deteriorating underground pipes cost $700 million at the Vermont Yankee Nuclear Power Plant in Vernon, VT, USA [5]. Therefore, regular inspections of components of the nuclear facilities are needed to ensure the safety and securities of the operations. However, due to the complex infrastructure of nuclear power plants, a direct inspection of every corner of the facilities is not feasible. A failure to detect cracks well in advance is one of the major causes of accidents [6].

Due to the lack of autonomous inspection, existing systems rely on human inspectors to detect cracks on the walls of nuclear facilities. However, the environment is hazardous and a direct inspection is not feasible and most NPP depend on remote video recordings. A typical system includes a robotic arm that manoeuvres a camera to record videos, followed by human operators inspecting these recordings to detect cracks [7]. This human-involved task is subjective, time-consuming, and tedious and sometimes error-prone [3]. Chen et al. [3] developed a ML technique to detect crack patches in each frame of the recorded video by combining naive Bayes and neural network based approach. Fig. 1 shows the schematic of their crack detection ML model.

Detecting cracks on the NPP's walls is a challenging task as these cracks can be small, or noisy patterns exist on the components' surfaces [8]. Any false positives can be catastrophic and it is estimated that 60% of nuclear facilities accidents were caused by human errors [4]. In recent years, research on applying ML techniques to inspect and detect cracks has been on a rising trend. Some works utilised conventional heuristic ML based methods for crack detection [9], [10]. While others have focused on deep learning-based methods [11]–[17]. The studies of detecting cracks are extended from concrete surface [13], [14], [17], [18] to metallic surface on NPP [3], [8] and road surfaces [12]. While the works on detecting cracks autonomously without interference from the human operator are progressing rapidly, the concern over safety and security over automation is also growing exponentially. As the infrastructures containing such cracks are vulnerable ones, the wrong diagnosis can have severely damaging consequences. It has been established that detecting damages, cracks on the surface of nuclear facilities is a crucial task. Any failure on such tasks can be expensive, economically and can also be risk to human lives. As shown in Fig. 2, when an adversary attacks the ML based systems that detect cracks on the wall of the NPP, the output could be compromised. For example, when there are cracks on the wall and the corrupted model can output a 'no cracks' label. Then the NPP will be carried out its work as normal and the consequence can be fatal. However, if there are not any cracks on the wall and the output result from the compromised model is 'cracks' then the NPP could be shut down and this can lead to economic loss.

The surge of application of ML techniques and implementation of autonomous behaviour on the nuclear facilities is increasing to achieve effective operations. The study of adversarial attacks on such learning techniques is needed because of the nature of the nuclear industry. The industry being security-sensitive and comprised of highly complex infrastructure, various types of adversarial attacks on the ML models need to be studied thoroughly before completely relying on that computational intelligence. It is equally imperative to find a defensive strategy on those adversarial attacks as well. Therefore, we investigate some of the current state-of-the-art attacks, and defensive mechanisms against those attacks are analysed, focusing on a case study of crack detection on the concrete surface.

III. Methodology

This section presents the methodology for attacking and defending against such attacks on a crack detection ML model following different methods in aML and rML that are adopted for our case study.

A. Adversarial Machine Learning (aML) Techniques

The adoption of ML techniques is widespread and in diverse areas including computer vision, speech recognition, natural language understanding. While the adoption of applying the ML technique has risen exponentially, the concerns over the security and safety of the applications where they have been

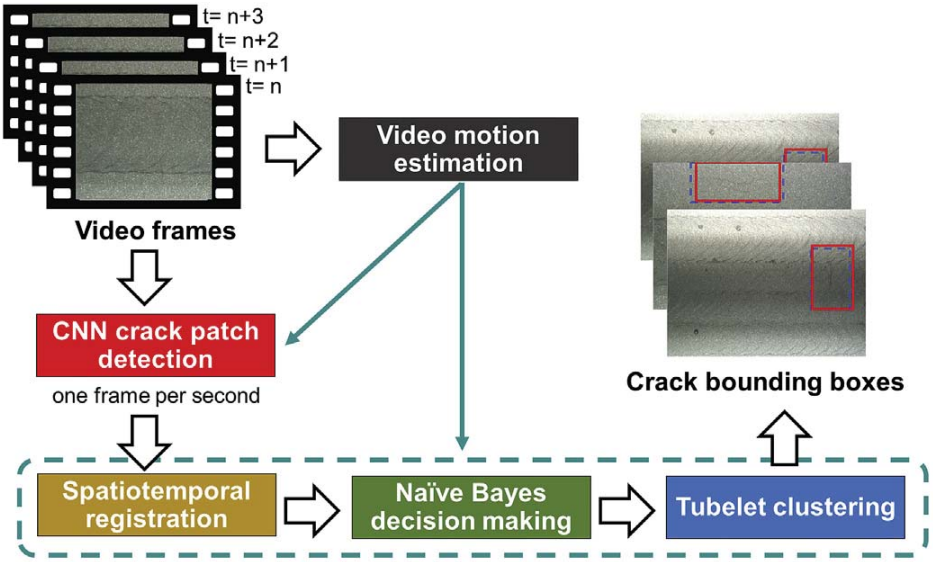

Fig. 1 ML-based crack detection model in the nuclear power plant [3]

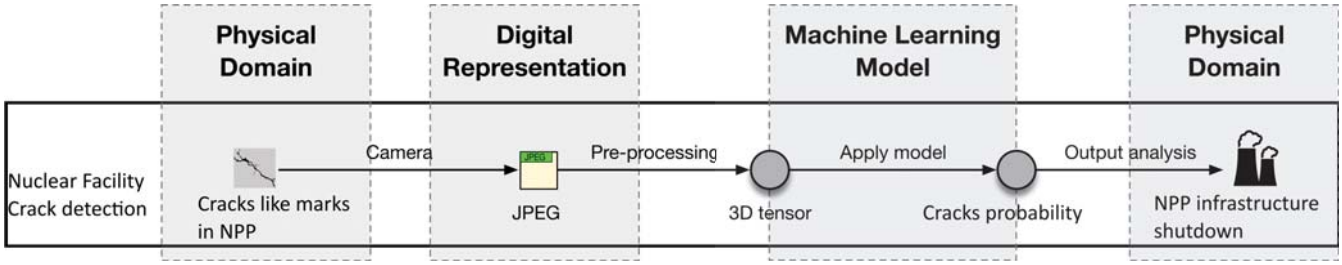

Fig. 2 ML attack surface on a crack detection model: Depending on the probability of detecting cracks, the action of shutting down of nuclear facility's infrastructure (e.g., NPP). This example is influenced by [19]

applied have also risen in the community. Since Szegedy et al. [20] showed that the deep neural networks could be tricked, the attention towards the aML has increased. This has led to the awareness of the security in ML techniques. For instance, different types of adversarial attack strategies exists affecting both data and model architecture to study their vulnerabilities.

White-box attacks and black-box attacks are two types of adversarial attacks that are widely studied. In white-box attacks, the adversary has full knowledge about the model and the data. This includes information about all the parameters such as features, model type, model architecture, values of all parameters, and trainable weights. On the other hand, the attackers do not have knowledge relating to models and data, except the input and output in black-box attacks.

Papernot et al. [21] explored black box attacks by training a deep neural network by crafting human imperceptible inputs. Biggio et al. [22] studied security on Support Vector Machine (SVM) learning methods by aiming to maximise the classification error from SVM by injecting well-crafted, adversarial label noise attacks . They flip the labels in the training data. Most of the adversaries in the classification tasks are providing wrong output, flipping the labels. [23], [24] presented that the cross-model transferability of adversarial data points between Deep neural networks (DNN)s. This implies launching an efficient attack through the use of surrogate models even though their training or neural network architectures are different.

The vulnerability of neural networks to adversarial examples were initially studied by [20]. According to Szegedy et al. [20], imperceptible adversarial perturbations are introduced to data to mislead ML classifiers. They influenced their work based on the well-known Limited Memory Broyden–Fletcher–Goldfarb–Shanno algorithm (L-BFGS) to solve the optimisation problem by detecting the smallest possible adversarial perturbation. The problem can be formalised as in (1):

$$\underset{x'}{argmin}\ f(x+x') = l\ subject\ to\ (x+x') \in D, \quad (1)$$

where x is the input example, which is correctly classified by f, is perturbed with x' to obtain the resulting adversarial example $x_* = x + x'$. The perturbed sample remains in the input domain D, however, it is assigned the target label l.

Following this optimisation based method, several methods were developed to generate adversarial attacks in the ML models. We now briefly discuss some notable aML methods that we have investigated to inject adversaries in our use case in the following section.

- **Fast Gradient Simple Method (FGSM)**: Goodfellow et al. [23] introduced the FGSM method to test the idea

that adversarial examples can be found using only a linear approximation of the target model. The loss function is linearised in the infinity norm L_∞ neighbourhood of an input object and finds the exact maximum of linearised function for generating adversarial samples, for which the following closed-form Equation (2) is used.

$$X_* = X + \epsilon * sign(\nabla x J(X, y)) \quad (2)$$

where ϵ is the magnitude of the perturbation.

- **Basic Iterative Method (BIM)**: This is an extension of the FGSM method, proposed by Kurakin et al. [25]. It is proposed to improve the performance of FGSM by running a small step size iterative optimiser multiple times, and clipping the intermediate adversarial samples after each step ensuring to be in the range of an original input. The formulation in the i^{th} iteration becomes:

$$X'_{t+1} = Clip(X'_t + \alpha \cdot sign(\nabla x J(\theta, x'_t, y))) \quad (3)$$

where α is the constant for controlling the magnitude of the perturbation, and clip function keeps the generated adversarial examples within the range of an original input.
- **Projected Gradient Descent (PGD)**: Madry et al. [26] proposed the Projected Gradient Descent (PGD) attack where they have explored the loss function by restarting the attack at different points with the vector norm of infinity L_∞ around the input examples. It is similar to BIM but with the loss function.
- **Carlini-Wagner Attack**: Carlini et al. [27] proposed adversarial examples generating algorithm which made the perturbation quasi-imperceptible by restricting their vector norms (l_2, l_∞, l_0). Their investigation focused on minimising the loss function that has smaller values on adversarial examples and higher on clean examples. With the exploration of several varieties of loss functions, they achieved the strongest L_2 norm attack, which is formulated as:

$$\min_{\delta} D(x, x+\delta) + c \cdot f(x+\delta) \; subject \; to \; x+\delta \epsilon [0, 1] \quad (4)$$

where δ denotes the adversarial perturbation, the distance metric with a vector norm states a success function f, that is, $f(x+\delta) \leq 0$ if the neural network prediction is the attack target and minimise the sum with a trade-off constant 'c'. The constant 'c' is chosen by modified binary search [28]. It has shown that the Carlini-Wagner (C&W) attack to be powerful but is expensive in terms of computation.
- **DeepFool**: Moosavi-Dezfooli et al. [29] proposed DeepFool to generate adversaries on the ML model. This method is designed to apply in the non-targeted case where the attacker can only ensure that the model classifies the adversarial example in a class different from the original ones. DeepFool is inspired by the fact that the separating hyperplanes in linear classifiers indicate the decision boundaries of each class; it aims to find the least distortion (in terms of euclidean distance) leading to misclassification by projecting the input example to the closest separating hyperplane.
- Others: There are other adversarial attacks that exist including Jacobian-based Saliency Map Attack (JSMA), Gaussian noise. For more details, we point the readers to the latest survey papers [30]–[32].

B. Resilient Machine Learning Techniques (rML)

The widespread continuous exploration and adoption of ML techniques have shown the importance of machine intelligence in human lives. However, with the rise of comfort using the technology, the concern of attacks on them also increases, such as cyber-attacks. As mentioned earlier, even with a small amount of perturbation in one of the advanced ML techniques, DNN, the algorithm is easily fooled. To tackle adversarial attacks, several defence techniques have been proposed. These include feature squeezing [33], novel model architecture using regularisation [34], adversarial training [35], the use of JPEG compression for pre-processing the input [36], [37], and neural fingerprinting [38]. These methods have exhibited success in mitigating aML attacks. In this section, we provide a brief description and discussion of some of the notable works.

- **Brute-Force adversarial training**: Adversarial training is a standard brute force approach in which a method generates adversarial examples and augments these perturbed data into the training set while training the targeted model [32]. The augmentation can be performed by feeding the model with both the original data and the adversary crafted data [25]. The primary objective of the adversarial training method is to increase model robustness by injecting adversarial examples into the training set [20], [23], [39], [40]. Some of the adversarial training techniques are FGSM adversarial training [23], BIM adversarial training, adversarial training ensemble, Adversarial logit pairing, and Generative Adversarial Network (GAN) [41].
- **GAN-based defence**: Lee et al. [42] used a GAN [41] to train an ML model that is robust against FGSM [23] like attacks. GAN automatically discovers and learns the regularities or patterns in the input data in such a way that the model can be used to generate new examples that plausibly could have been drawn from the original dataset [41]. [42] proposed to directly train the network along with a generator network that attempts to generate perturbation for that network. During its training, the classifier keeps trying to classify both the clean and perturbed data correctly. On the other hand, Shen et al. [43] used GAN and utilised the generator part of the network to rectify a perturbation and proposed Adversarial Perturbation Elimination GAN (APEGAN). The method eliminates the perturbation of the adversarial examples first and then feed it into the target model to increase the robustness.
- **Other strategies**: There are other defensive methods that have been proposed over the years, including deepcloak [44], Magnet [45]. For more information, we refer the reader to the works [30]–[32].

IV. Experiments

This section discusses our experimental study for adversarial attacks and defences in the network for detecting cracks on the concrete walls. Here, we introduce the used dataset for our experiment followed by the explanation of the proposed ML model and parameter settings. Then we present and discuss the results.

A. Dataset

Due to the safety and security policies regulation, nuclear industry is considered to be a closed industry. Most of the nuclear-related data are not readily available. Therefore, we utilised a dataset containing crack images of concrete walls, pavements, and roads. The dataset comprises of forty thousand cracked and uncracked images of concrete blocks and is available online [46]. For our purpose, we separated the dataset into three sets, training set, validation set and test set. While 80% of 40k total data is used to form the training set, of which 25% is used as the validation set. Finally, 20% of the entire data is used to evaluate the performance of the methods.

B. Crack Detection Model

We adopted a Convolution Neural Network (CNN) model by [47] to build a crack detection ML model. The structure of the model comprises of three convolution layers, the first with 16 filters and the subsequent other two containing 32 layers. Max pooling and dropout layers are added to all three layers. The model then has a flatten layer followed by a fully connected 64 nodes layers. The output layer follows the flattened layer. In addition, rectified linear unit (ReLu) and sigmoid functions are used as activation functions. ReLu was chosen for all the layers except the output layer where the sigmoid function was used.

C. Experimental Results

We analysed various methods for the adversarial attack on a ML model and how to defend those attacks by following various defence strategies. Once the ML model is created as mentioned in Section IV-B, we chose five adversarial attack generating techniques to perturb our crack detecting model, they are i) FGSM, ii) PGD, iii) BIM, iv) DeepFool v) Carlini-Wagner. In terms of making ML methods resilient against such attacks, we have chosen to investigate two defence mechanisms; they are: i) Adversarial training, and APEGAN. They are presented in Table I. For the experiment, we used 0.1 epsilon value to add noise in the model. Table I shows the accuracy results obtained from adversarial methods and also defensive methods.

The CNN model generated 98.99% of accurate result on detecting cracks on the concrete block. When adversarial attacks are inserted on the CNN model, the classifying accuracy of the model is reduced nearly by half. The most successful attack was generated by the DeepFool method which resulted in less than 20% of accurate classification by the crack detection model whereas the least successful attack are BIM and C&W which allowed more than 51% accurate classification. We believe the success of deepfool lies in the way it perturbs the object. It works on only needed minimal perturbation to fool the model, for which it efficiently approximates the decision space of the target classifier to identify such perturbation.

TABLE I Mean Accuracy of the CNN Model for Different Adversarial Methods

Attack Method		Accuracy (%) Defensive Method	
		Adv. Training	APEGAN
FGSM	50.55	93.5	**93.8**
PGD	43.1	**90.55**	86.05
CW	51.85	**98.31**	93.05
DeepFool	**17.9**	**48.75**	44.7
BIM	51.9	94.26	**94.65**

The base accuracy is 98.99%.

In terms of defensive mechanisms, both adversarial training and APEGAN methods managed to boost the resiliency of the system towards adversarial attacks which is evident with increment of the accuracy from 17.9% (– DeepFool attack) to 48.75% (–Adversarial training defence) and 44.70% (APEGAN defence) respectively. For other attacks, such as C&W, the accuracy of adversarial training has increased from 51.85% to 98.31% which shows promising result. More details on the other attacks and their subsequent defences are provided in Table I. While the result of adversarial training is slightly better than APEGAN overall, the defensive methods could not protect the model entirely.

V. Discussion and Conclusion

ML has been widely used in many complex applications including security risks such as autonomous driving, remote inspections in nuclear power plants. While they are widely used they are equally in danger of getting fooled as well. Therefore, there is an increasing interest in the research community to study the adversarial attacks and their ability to impact the ML models which can corrupt the decision processes. At the same time, protecting the models against such attacks has also been studying widely. In this paper, we study various adversarial attacks and analysed different attacks generating methods. We also investigate defensive mechanisms to protect the model against such attacks. We focus on five attacks and two defences. As an exemplar, we selected a safety-critical task in the nuclear industry of detecting cracks on the concrete walls. However, the methods presented in this work are equally applicable to other ML based systems for classification tasks.

We observed that there are several strong adversarial attacking methods proposed in the literature. Some attacks facilitate an adversary to query a trained ML model to predict whether or not a particular example was contained in the model's training dataset. Whereas some can corrupt the ML decision process by simply guessing the parameters and features based on the input and output data.

On the other hand, it is shown that none of the defensive methods can protect the ML model entirely. Although we only analysed two defensive strategies we noticed that both

adversarial training and APEGAN were able to defend against all five attacking methods.

In the future, we plan to extend the list of defensive methods and investigate novel defensive methods to protect the ML model entirely. We will also expand the use case from detecting cracks on concrete to metal, and bricks.

VI. Acknowledgment

The work presented has been funded by Grant **EP/R026084/1** Robotics and Artificial Intelligence for Nuclear (RAIN) through the Engineering and Physics Research Council (EPSRC).

References

[1] "Nuclear power reactors," https://world-nuclear.org/information-library/nuclear-fuel-cycle/nuclear-power-reactors/nuclear-power-reactors.aspx, 2022.

[2] J. F. Ahearne, A. V. C. Jr, H. A. Feiveson, D. Ingersoll, A. C. Klein, S. Maloney, I. Oelrich, S. Squassoni, and R. Wolfson, "The future of nuclear power in the united states," 2012.

[3] F.-C. Chen and M. R. Jahanshahi, "Nb-cnn: Deep learning-based crack detection using convolutional neural network and naïve bayes data fusion," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 5, pp. 4392–4400, 2018.

[4] *Advanced Control Systems to Improve Nuclear Power Plant Reliability and Efficiency*, ser. TECDOC Series. Vienna: INTERNATIONAL ATOMIC ENERGY AGENCY, 1997, no. 952. [Online]. Available: https://www.iaea.org/publications/5604/advanced-control-systems-to-improve-nuclear-power-plant-reliability-and-efficiency

[5] B. Sovacool, "A critical evaluation of nuclear power and renewable electricity in asia," *Journal of Contemporary Asia*, vol. 40, pp. 369 – 400, 2010.

[6] S. Suman, "Artificial intelligence in nuclear industry: Chimera or solution?" *Journal of Cleaner Production*, vol. 278, p. 124022, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959652620340671

[7] S. Cumblidge, M. T. Anderson, S. Doctor, F. Simonen, and A. Elliot, "An assessment of remote visual methods to detect cracking in reactor components," 2008.

[8] S. J. Schmugge, L. Rice, N. R. Nguyen, J. Lindberg, R. Grizzi, C. Joffe, and M. C. Shin, "Detection of cracks in nuclear power plant using spatial-temporal grouping of local patches," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–7.

[9] M. Gavilán, D. Balcones, O. Marcos, D. F. Llorca, M. A. Sotelo, I. Parra, M. Ocaña, P. Aliseda, P. Yarza, and A. Amírola, "Adaptive road crack detection system by pavement classification," *Sensors*, vol. 11, no. 10, pp. 9628–9657, 2011. [Online]. Available: https://www.mdpi.com/1424-8220/11/10/9628

[10] Y. Sari, P. B. Prakoso, and A. R. Baskara, "Road crack detection using support vector machine (svm) and otsu algorithm," in *2019 6th International Conference on Electric Vehicular Technology (ICEVT)*, 2019, pp. 349–354.

[11] Y. Xu, S. Li, D. Zhang, Y. Jin, F. Zhang, N. Li, and H. Li, "Identification framework for cracks on a steel structure surface by a restricted boltzmann machines algorithm based on consumer-grade camera images," *Structural Control and Health Monitoring*, vol. 25, no. 2, p. e2075, 2018, e2075 STC-16-0276.R1. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/stc.2075

[12] L. Zhang, F. Yang, Y. Daniel Zhang, and Y. J. Zhu, "Road crack detection using deep convolutional neural network," in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3708–3712.

[13] K. Chen, A. Yadav, A. Khan, Y. Meng, K. Zhu, and Q.-F. Liu, "Improved crack detection and recognition based on convolutional neural network," *Model. Simul. Eng.*, vol. 2019, Jan. 2019. [Online]. Available: https://doi.org/10.1155/2019/8796743

[14] S. Li and X. Zhao, "Image-based concrete crack detection using convolutional neural network and exhaustive search technique," *Advances in Civil Engineering*, vol. 2019, pp. 1–12, 2019.

[15] K. Gopalakrishnan, H. Gholami, A. Vidyadharan, Alok, Choudhary, and A. Agrawal, "Crack damage detection in unmanned aerial vehicle images of civil infrastructure using pre-trained deep learning model," 2018.

[16] F. Kucuksubasi and A. G. Sorguc, "Transfer learning-based crack detection by autonomous uavs," in *Proceedings of the 35th International Symposium on Automation and Robotics in Construction (ISARC)*, J. Teizer, Ed. Taipei, Taiwan: International Association for Automation and Robotics in Construction (IAARC), July 2018, pp. 593–600.

[17] J. J. Kim, A.-R. Kim, and S.-W. Lee, "Artificial neural network-based automated crack detection and analysis for the inspection of concrete structures," *Applied Sciences*, vol. 10, no. 22, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/22/8105

[18] Y.-J. Cha, W. Choi, and O. Büyüköztürk, "Deep learning-based crack damage detection using convolutional neural networks," *Computer-Aided Civil and Infrastructure Engineering*, vol. 32, no. 5, pp. 361–378, 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12263

[19] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, "Sok: Security and privacy in machine learning," in *2018 IEEE European Symposium on Security and Privacy (EuroS P)*, 2018, pp. 399–414.

[20] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2014.

[21] N. Papernot, P. Mcdaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, 2017.

[22] B. Biggio, B. Nelson, and P. Laskov, "Support vector machines under adversarial label noise," in *Proceedings of the Asian Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, C.-N. Hsu and W. S. Lee, Eds., vol. 20. South Garden Hotels and Resorts, Taoyuan, Taiwain: PMLR, 14–15 Nov 2011, pp. 97–112. [Online]. Available: http://proceedings.mlr.press/v20/biggio11.html

[23] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015.

[24] I. Rosenberg, A. Shabtai, L. Rokach, and Y. Elovici, "Generic black-box end-to-end attack against state of the art api call based malware classifiers," 2018.

[25] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," 2017.

[26] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2019.

[27] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," 2017.

[28] N. Papernot, F. Faghri, N. Carlini, I. Goodfellow, R. Feinman, A. Kurakin, C. Xie, Y. Sharma, T. Brown, A. Roy, A. Matyasko, V. Behzadan, K. Hambardzumyan, Z. Zhang, Y.-L. Juang, Z. Li, R. Sheatsley, A. Garg, J. Uesato, W. Gierke, Y. Dong, D. Berthelot, P. Hendricks, J. Rauber, R. Long, and P. McDaniel, "Technical report on the cleverhans v2.1.0 adversarial examples library," 2018.

[29] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," 2016.

[30] A. Kumar and S. Mehta, "A survey on resilient machine learning," 2017.

[31] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," *ArXiv*, vol. abs/1810.00069, 2018.

[32] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," 2018.

[33] D. Sgandurra, L. Muñoz-González, R. Mohsen, and E. C. Lupu, "Automated dynamic analysis of ransomware: Benefits, limitations and use for detection," 2016.

[34] H. Kannan, A. Kurakin, and I. Goodfellow, "Adversarial logit pairing," 2018.

[35] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2019.

[36] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, L. Chen, M. E. Kounavis, and D. H. Chau, "Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression," 2017.

[37] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, S. Li, L. Chen, M. E. Kounavis, and D. H. Chau, "Shield: Fast, practical defense and vaccination for deep learning using jpeg compression," 2018.

[38] S. Dathathri, S. Zheng, Y. Yue, and R. M. Murray, "Detecting adversarial examples via neural fingerprinting," 2019. [Online]. Available: https://openreview.net/forum?id=SJekyhCctQ

[39] C. Lyu, K. Huang, and H.-N. Liang, "A unified gradient regularization family for adversarial examples," *2015 IEEE International Conference on Data Mining*, pp. 301–309, 2015.

[40] U. Shaham, Y. Yamada, and S. N. Negahban, "Understanding adversarial training: Increasing local stability of supervised models through robust optimization," *Neurocomputing*, vol. 307, pp. 195–204, 2018.

[41] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.
[42] H. Lee, S. Han, and J. Lee, "Generative adversarial trainer: Defense to adversarial perturbations with gan," *ArXiv*, vol. abs/1705.03387, 2017.
[43] G. Jin, S. Shen, D. Zhang, F. Dai, and Y. Zhang, "Ape-gan: Adversarial perturbation elimination with gan," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3842–3846.
[44] J. Gao, B. Wang, Z. Lin, W. Xu, and Y. Qi, "Deepcloak: Masking deep neural network models for robustness against adversarial samples," 2017.
[45] D. Meng and H. Chen, "Magnet: A two-pronged defense against adversarial examples," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 135–147. [Online]. Available: https://doi.org/10.1145/3133956.3134057
[46] C. F. M. Ozgenel and A. G. Sorguç, "Performance comparison of pretrained convolutional neural networks on crack detection in buildings," in *Proceedings of the 35th International Symposium on Automation and Robotics in Construction (ISARC)*, J. Teizer, Ed. Taipei, Taiwan: International Association for Automation and Robotics in Construction (IAARC), July 2018, pp. 693–700.
[47] V. B. Rodrigues, "Concrete-crack-classification-model," https://github.com/vbrodrigues/Concrete-Crack-Classification-Model, 2018.

Dr Anita Khadka received her MSc. in Intelligent systems and Robotics from University of Essex and Ph.D in Computer Science from The Open university. She is currently a Research Fellow working in the area of Machine learning and their security in inherently complex domains like Nuclear and Space industries in University of Warwick. Her research interests include Machine learning, Resilient machine learning, and Data Science.

Dr Gregory Epiphaniou Currently holds a position as an Associate Professor of security engineering at the University of Warwick. His role involves bid support, applied research and publications. Part of his current research activities is formalised around cyber effects modeling, wireless communications with the main focus on crypto-key generation, exploiting the time-domain physical attributes of V-V channels and cyber resilience. He led and contributed to several research projects funded by EPSRC, IUK and local authorities totalling over £4M. He currently holds a subject matter expert panel position in the Chartered Institute for Securities and Investments. He acts as a technical committee member for several scientific conferences in Information and network security and served as a key member in the development of WS5 for the formation of the UK Cybersecurity Council.

Professor Carsten Maple leads the Secure Cyber Systems Research Group in WMG at the University of Warwick, where he is also the Principal Investigator of the NCSC-EPSRC Academic Centre of Excellence in Cyber Security Research. He is a co-investigator of the PETRAS National Centre of Excellence for IoT Systems Cybersecurity where he leads on Transport & Mobility and Warwick PI on the Autotrust project. Carsten has an international research reputation and extensive experience of institutional strategy development and interacting with external agencies. He has published over 250 peer-reviewed papers and is coauthor of the UK Security Breach Investigations Report 2010, supported by the Serious Organised Crime Agency and the Police Central e-crime Unit. Additionally he has advised executive and non-executive directors of public sector organisations and multibillion pound private organisations. Professor Maple is a past Chair of the Council of Professors and Heads of Computing in the UK, a member of the Zenzic Strategic Advisory Board, a member of the IoTSF Executive Steering Board, an executive committee member of the EPSRC RAS Network and a member of the UK Computing Research Committee, the ENISA CarSEC expert group, the Interpol Car Cybercrime Expert group and Europol European Cybercrime Centre.