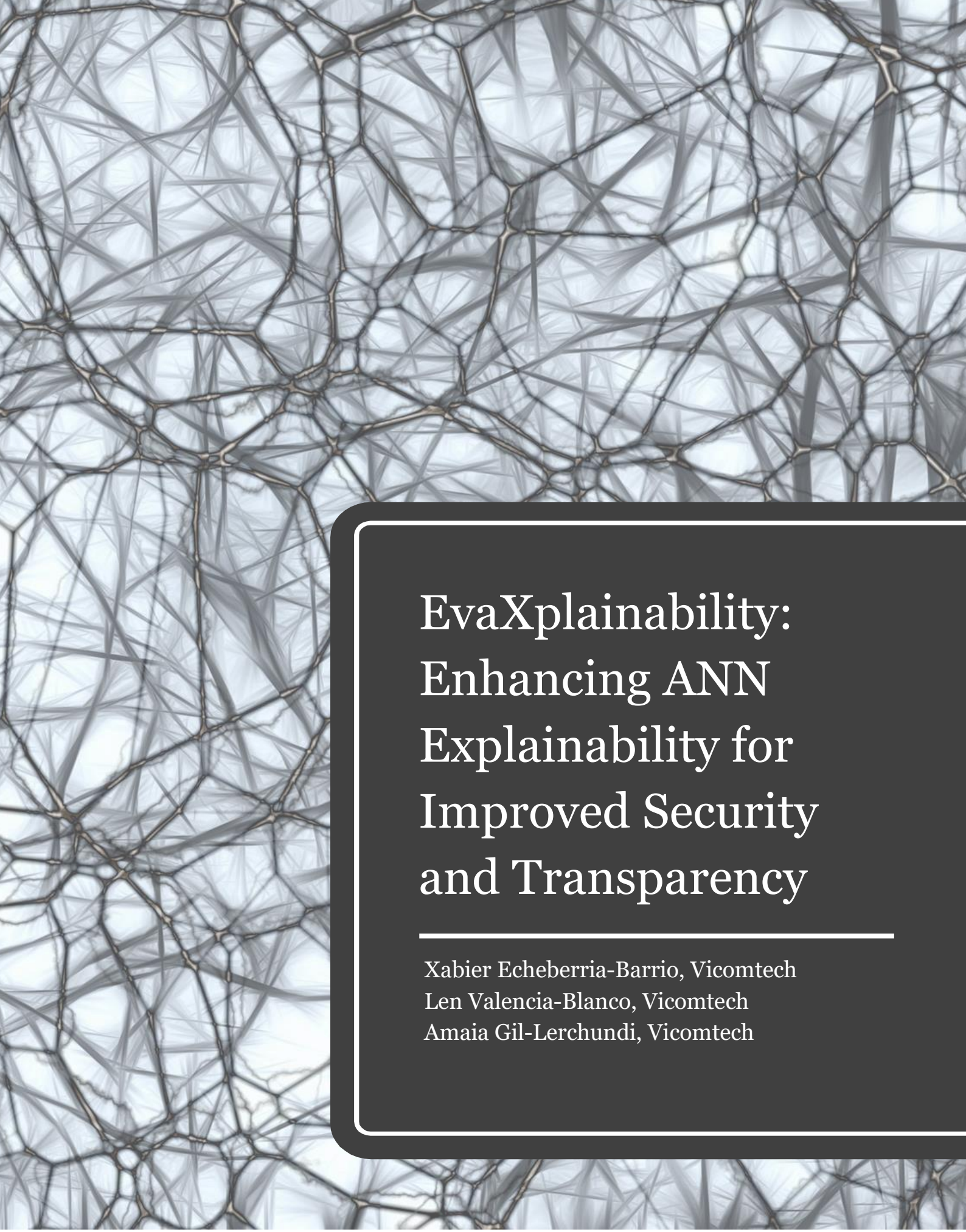

# EvaXplainability: Enhancing ANN Explainability for Improved Security and Transparency

Xabier Echeberria-Barrio, Vicomtech
Len Valencia-Blanco, Vicomtech
Amaia Gil-Lerchundi, Vicomtech

Co-funded by the European Union

This project has received funding from the European Union's Horizon Europe Research and Innovation Programme under the Grant Agreement No. 101073909.
The contents of this document represent the views of the authors only and remain their sole responsibility. The European Research Executive Agency and the European Commission are not responsible for any use of the included information.

# EvaXplainability: Enhancing ANN Explainability for Improved Security and Transparency

*Xavier Echeberria-Barrio, Len Valencia-Blanco, Amaia Gil-Lerchundi (Vicomtech)*

*The paper introduces the EvaXplainability tool as a response to the growing implementation of artificial neural network (ANN) technology in various fields. It addresses the challenges posed by the complexity and lack of transparency in ANNs by focusing on explainability methods. EvaXplainability monitors the neural behaviour of targeted ANNs, highlighting pivotal neurons and their impact on decision-making. The document discusses the current state of explainability in AI, emphasizing its role in transparency and safeguarding ANN system integrity. The tool's development within the ATLANTIS cybersecurity project, focusing on a deep fake detector, is outlined. Challenges, such as scalability and validation, are acknowledged, along with potential barriers related to data privacy regulations. The benefits of EvaXplainability include enhanced ANN security and robustness. The future outlook emphasizes a proactive approach to scalability and the tool's adaptability to seamlessly integrate with existing ANN systems, ultimately enhancing overall security and explainability.*

## 1. Introduction

Artificial Neural Network (ANN) technology is being implemented in several fields, directly affecting decisions in critical aspects of life. In particular, fields such as healthcare, automotive, cybersecurity, and so on, have suffered a huge enhancement due to this technology. However, AI technology has also brought new vulnerabilities and complications.

A key concern of the incorporation of ANNs in the decision-making process is that humans cannot easily understand how this technology works. Especially in critical applications, such as in healthcare, humans need to trust the reasoning process. However, nowadays, artificial neural network technology is typically a *black box*, i.e., the reason for its decision is usually unknown and hardly understandable. That is why, in the last years, researchers have been developing different methods to understand the decision-making of those *black boxes*, trying to convert them into *white boxes*. These methods are widely referred in the literature as AI Explanability methods [1].

This work presents the EvaXplanability tool, which enables the interpretation of the decision-making process of a targeted artificial neural network by monitoring its behavior at the neural level. Concretely, it shows how important the neurons are in the prediction and which are the most critical regarding impact on decision-making. In addition, this paper presents how EvaXplainability tool is applied in ATLANTIS project.

## 2. The Current State of Affairs in Explainability

In the ever-evolving landscape of artificial intelligence, the quest for enhancing the transparency of complex ANN models continues to be a critical pursuit. The ANN explainability has been approached through various methodologies, each tailored to unravel

different layers of complexity within these models. Techniques such as visualizing model behaviour, exemplified by tools like LIME [2] and SHAP [3], offer insights into the inner workings of AI, making it accessible to a broader audience. Simultaneously, endeavours like relating individual features to predictions using methods such as PDPs [4], DeepLIFT [5], and CAM [6] contribute to deciphering the intricate relationships that drive the decision-making within these models.

Ongoing research and development efforts are not only focused on refining existing explainability techniques but also on introducing innovative approaches. These advancements are particularly noteworthy in specific domains, such as image recognition, where the application of explainability methods to convolutional neural networks (CNNs) has opened new avenues for understanding the nuanced aspects of model predictions. Techniques like SUMMIT [7], among others [8], showcase the researchers' commitment to providing users with a clearer picture of how these models operate.

Furthermore, the role of explainability extends beyond mere comprehension; it has become a pivotal tool in safeguarding the integrity of ANN systems. By understanding the normal behaviour of a targeted ANN model, one can effectively identify and address anomalous patterns, potentially signalling instances of model corruption or manipulation. This dual role of explainability, both as a means of comprehension and a safeguarding mechanism, underscores its importance in ensuring the reliability and robustness of ANN in practical applications. The work of researchers, such as Echeberria-Barrio et al. [9], highlights the tangible benefits of incorporating explainability methods as integral components in the development and deployment of ANN models.

## 3. The Role of EvaXplainability

The EvaXplainability tool tries to identify the pivotal neurons of the targeted ANN to show its expected behaviour. This work follows a novel field in ANN technology, which tries to know their vulnerabilities and threats. Basically, cybersecurity in ANN is recent among researchers, and all the tools in the literature are in the early stages. Concretely, as mentioned in Section 2, explainability techniques are becoming important as helpful methods to understand and figure out the vulnerabilities within ANN models and detect threats.

The tool relies on the work presented by Echeberria-Barrio et al. in [9] and extends the behaviour analysis to bigger models with many more neurons, making these assessments accessible for more types of ANNs. Moreover, this tool aims to generate a neuron mask containing the most important neurons within the targeted ANN. This mask will reduce the number of neurons to be analysed in the targeted ANN, opening the possibility to analyse even bigger ANN models.

In addition, EvaXplainability is the first explainability tool to generate the graph representation of the targeted ANN with different attributes, concretely a Convolutional Neural Network (CNN) model. This approach is developed to understand the behaviour of the ANN models, filter the neurons to be analysed, help to detect the vulnerabilities and threats of ANN models, and make it interactive and visual all this information to the users.

# 4. The Research and Development Path in ATLANTIS

This tool is being used under a cybersecurity use case in the ATLANTIS project. As mentioned in Section 2, the explainability techniques are helpful for the security of ANN, and there are already studies demonstrating that they can help take a huge step in this field. In this sense, the tool explains why a targeted ANN model is getting attacked by an adversary. The target model is a deep fake detector developed in ATLANTIS, and the tool learns the expected behaviour of the model and its behaviour when it is suffering an evasion attack.

This tool starts generating an initial study about the targeted model's interaction and corresponding data. Figure 1 shows the process carried out in the presented method/analysis. This analysis takes normal and corrupted data to compute the behaviour of the targeted model in both cases. That behaviour is obtained by calculating a feature called impact, which corresponds to the difference between the input and the output of each neuron in the ANN model. Note that this feature is a neuronal attribute, i.e., the targeted ANN model obtains several impact values depending on the number of neurons within it (ANN impact mapping). Moreover, each sample gives a unique impact mapping, representing the behaviour of the ANN model in the corresponding sample. Therefore, the analysis generates normal and corrupted behaviours since it takes normal and corrupted data. Considering all the impact mappings, the neurons are compared to detect which neurons suffer the most modification regarding the impact attribute when the behaviour is normal and when not. At the end of the initial study, the tool knows an approach to the neurons most relevant in the model's decision-making.

Once the tool is initialized, it can receive new input to detect the most relevant neurons in its case. This process takes the most pivotal neurons seen in the initial study and intersects them with the most critical neurons in its case. Thus, the tool selects a subgroup of the essential neurons detected in the previous research. In the end, those are the highlighted neurons of the tool for the received sample.

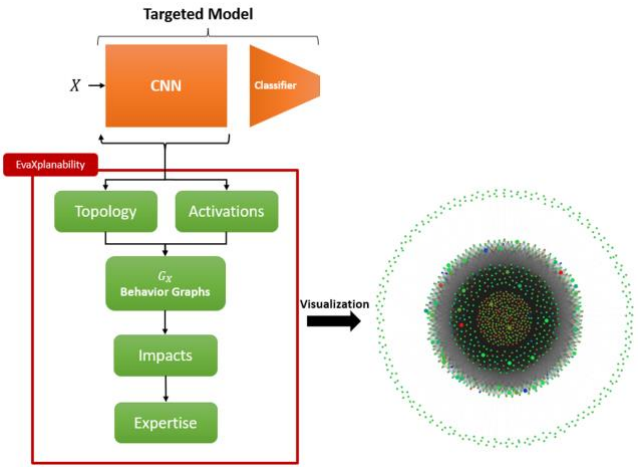


*Figure 1. The architecture of the EvaXplainability tool. Note that in the visualization, the impact is represented by colour and the expertise is represented by size.*

Eventually, this information is taken and fed to a visualization module, where it is displayed as a layered structure ordered from left to right, with the leftmost grouping corresponding

to the first layer of the neural network. For each layer, the belonging neurons follow a uniform distribution. Each neuron is represented by a circle with its expertise codified as the radius and its impact reflected as a colour in a hue gradient, with reds and yellows representing high negative impact, greens codifying low impact values, and blues highlighting high positive impact values.

All this development is in the early stages. The technologies are already implemented, and the prototype's first version exists. However, there are several points, such as the kernel filtering method and pixel highlighting by desired kernels, to validate and polish to generate a viable tool. The following steps in ATLANTIS will focus on the demonstration of the performance of the tool, and at the same time, the technologies inside will be enhanced in terms of time and scalability. This tool will be applied in ATLANTIS as protector of an ANN model trained to detect Fake News. Moreover, some issues can appear during the demonstrations that will be fixed in ATLANTIS. Finally, the visualization part of the tool needs more development to show the ANN's behaviour clearly and faithfully.

## 5. The Challenges and Barriers

The application of AI explainability tools, such as the EvaXplainability, involves several challenges. One of them is the scalability of the tools. Considering that ANN models, nowadays, are complex and huge, containing millions of neurons, the graphs generated by millions of neurons require high computational power. This research is trying to handle this problem. Concretely, detecting the critical/main kernels of the targeted model and studying only them. Another challenge is to demonstrate the successful performance of the tool. Some approaches are being used to confirm that the selected neurons contain the main information to detect the critical behaviour of the targeted ANN. The main approach consists of the comparison of the corrupted behaviour detection with all the kernels (nodes) and with the filtered kernels, and depending on the results, in case of the performance is not reduced considerable, the filtering can be considered successful. Furthermore, regarding data and its availability, the potential barriers are those that the actual ANNs are suffering to work with different datasets. Concretely, any regulation on private or personal data will affect that tool if it must be implemented in an ANN working with this type of data. In general, this situation can happen when the targeted ANN system's data is under any regulation; then, it will generate barriers at the EvaXplainability level.

Currently there exist a notably challenge in the complexity of growing neural networks. The sheer volume of data and intricate interplay of neurons pose obstacles to effective monitoring. Interpreting the significance of individual neurons in decision-making adds another layer of complexity. To address these challenges, continuous refinement of the EvaXplainability tool is essential.

## 6. The Benefits and Impact

The EvaXplainability tool serves as security measure for Artificial Neural Networks (ANNs), providing a comprehensive method to monitor and evaluate their behaviour, particularly by spotlighting pivotal neurons in decision-making. This not only enhances ANN security but

also yields benefits for artificial intelligence. A significant advantage is the tool's capacity to raise the robustness of ANNs. Understanding the role of pivotal neurons could allow developers to adjust network architecture, resulting in more accurate and resilient predictions, better equipped to avoid evasion attacks. Moreover, the tool contributes to the safety of ANNs by pinpointing critical neurons, deepening our understanding of decision-making processes, and enabling the identification and mitigation of vulnerabilities. This security approach is crucial for ensuring the reliability and safety of deployed AI systems.

## 7. Future Outlook

As highlighted earlier, scalability has surfaced as a notable challenge during the developmental phase of our tool. Acknowledging this challenge early on has prompted a proactive approach to ensure that the final prototype achieves the requisite level of scalability. This ongoing effort is crucial, particularly given the prevalence of contemporary models characterized by intricate structures housing a vast number of neurons. The complexity of these models emphasizes the importance of monitoring and addressing scalability concerns to guarantee the tool's effectiveness across a spectrum of applications.

Moreover, adaptability emerges as a distinctive strength of our technology. Acting as a parallel tool, it effortlessly integrates with existing Artificial Neural Network (ANN) systems. This intrinsic adaptability allows for a smooth assimilation into diverse AI environments without disrupting ongoing operations. It ensures that users can readily harness the benefits of our tool as an augmentative feature, enhancing the security and explainability aspects of their neural network systems.

Highlight that ongoing research should focus on improving scalability to handle larger and more complex networks. Additionally, enhancing interpretability through advanced visualisation techniques and explanatory methods will facilitate a clearer understanding of intricate neural dynamics.

## 8. Conclusions

The work introduces the EvaXplainability tool in response to the increasing implementation of artificial neural network (ANN) technology. The inherent complexity and lack of transparency in ANNs pose challenges, prompting the development of explainability methods. EvaXplainability aims to address this by monitoring the neural behaviour of targeted ANNs, spotlighting on pivotal neurons and their impact on decision-making. The paper discusses the current state of explainability in AI, emphasizing its role in enhancing transparency and safeguarding the integrity of ANN systems. The tool's development within the ATLANTIS cybersecurity project, targeting a deep fake detector that the use of the tool in ATLANTIS project is presented. Challenges, such as scalability and validation, are acknowledged, along with potential barriers related to data privacy regulations. The benefits of EvaXplainability include improved ANN security and robustness. The future outlook highlights the proactive approach to scalability and the tool's adaptability to seamlessly integrate with existing ANN systems, enhancing overall security and explainability.

## References

[1] Weiping Ding, Mohamed Abdel-Basset, Hossam Hawash, Ahmed M. Ali, Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey, Information Sciences, Volume 615, 2022, Pages 238-292, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2022.10.013.

[2] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[3] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

[4] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

[5] Shrikumar, A., Greenside, P., & Kundaje, A. (2017, July). Learning important features through propagating activation differences. In International conference on machine learning (pp. 3145-3153). PMLR.

[6] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2921-2929).

[7] Hohman, F., Park, H., Robinson, C., & Chau, D. H. P. (2019). Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. IEEE transactions on visualization and computer graphics, 26(1), 1096-1106.

[8] Haar, L. V., Elvira, T., & Ochoa, O. (2023). An analysis of explainability methods for convolutional neural networks. Engineering Applications of Artificial Intelligence, 117, 105606.

[9] Echeberria-Barrio, X., Gil-Lerchundi, A., Egana-Zubia, J., & Orduna-Urrutia, R. (2022). Understanding deep learning defenses against adversarial examples through visualizations for dynamic risk assessment. Neural Computing and Applications, 34(23), 20477-20490.

*Front cover image by Gerd Altmann via Pixabay.*
*https://pixabay.com/users/geralt-9301*