Research Report

Insights from Explainable Interactive Machine Learning in the Age of COVID-19

COVID-19 HAS AGAIN TIGHTENED ITS GRIP AROUND THE WORLD AND THE HEALTH SYSTEM. THIS ARTICLE GIVES AN INTRODUCTION TO EXPLAINABLE INTERACTIVE MACHINE LEARNING AND PROVIDES INSIGHTS ON HOW THIS METHOD MAY NOT ONLY HELP IN ENGINEERING MORE POWERFUL AI SYSTEMS, BUT ALSO HOW IT MAY HELP TO EASE THE BURDEN OF VIRAL STRAINS ON THE HEALTHCARE SYSTEM.

Oliver Hinz

Wolfgang Stammer

Benjamin M. Abdel-Karim

Christian Hügel

Kristian Kersting

Nicolas Pfeuffer

Patrick Schramowski

Andreas Bucher

Gernot Rohde

Introduction

The latest reports of the World Health Organization on the coronavirus pandemic (www.who.int) show alarming numbers of millions of new infections every week. Although these numbers may reflect the severity of the outbreak, they do not yet reveal the dimensions of the burden the virus has on the health system. Recent reports from the German center of intensive healthcare (www.intensivregister.de/#/reporting) show that, by November 2020, the amount of required beds and intensive care unit slots for COVID-19 patients has spiked from a low between May and October to unseen high numbers. These numbers are just an indication of how much healthcare workers on the forefront are exceedingly fighting for the lives of the patients in an ever-growing fear of an unmanageable situation (e.g., Lai et al., 2020).

Machine Learning as a Supportive Pillar for Healthcare

To support our healthcare system and workers in the fight against the pandemic, scholars from the field of computer science, information systems, and medicine continuously try to engineer machine learning (ML)-based clinical decision support systems (CDSS). In the course of the year 2020, a wealth of research projects and papers has been published, ranging from aggregated open datasets to support the development of CDSS against COVID-19, implementation of black-box (e.g., Chowdhury et al., 2020) and white-box approaches, as well as experiments on the usefulness of such systems versus humans and in human-machine hybrid constellations (e.g., Mei et al., 2020).

The pitfalls of ML-Based CDSS

Arguably, many studies have reported promising results of either novel or existing architectures, e.g., convolutional neural networks (CNN) in the detection of COVID-19 from X-rays (e.g., Chowdhury et al., 2020) or CT-Scans (e.g., Mei et al., 2020), or for human-hybrid constellations (e.g., Mei et al., 2020). Nevertheless, two evident problems of these CDSS solutions are (1) that physicians cannot correct or improve these systems based on their expertise, and (2) that many of these proposed systems come with a lack of transparency.

As a result, these systems are not only potentially overconfident, but also dangerously intransparent. Hence, the employment of such systems in time-pressing, stressful environments could make up a potentially dangerous combination.

Our Methodology: XIL to the Rescue

Against this background, we aim to remedy the shortcomings of current systems with the aid of a methodology called explainable interactive machine learning (short: XIL, see Figure 1).

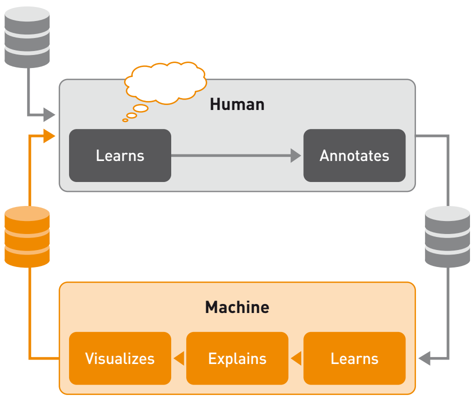

Figure 1: XIL Cycle (adapted from Hinz et al., 2020)

Grounded on the concepts of explainability (e.g., Selvaraju et al., 2017) and interactive machine learning, recent experiments with biologists had shown the potential of XIL in not only making black box systems, such as CNN, more transparent, but also its potential in helping experts to make the system more accurate and satisfying to use (Schramowski et al., 2020). Apart from the advantages in engineering, by augmenting ML systems with explainable AI features (xAI) and putting the human in the loop, XIL also has the potential to help humans discover and learn novel facts about the underlying task. Since knowledge about COVID-19 is still incomplete, XIL may serve as a valuable method for exploring and gaining novel knowledge. Our methodology consisted of several stages: first, we selected a suitable database for engineering a helpful CDSS for the case of COVID-19-based pneumonia.

In this matter, we oriented ourselves according to the diagnostic guidelines of the British society of thoracic imaging (Nair et al., 2020), and thus decided to build a CNN with the target of predicting COVID-19-based pneumonia from X-rays. Our selected database (Chowdhury et al., 2020) consisted of three classes: normal (X-rays without clear signs of pneumonia), viral pneumonia (pneumonia caused by other viral pathogens), and COVID-19-based pneumonia (pneumonia caused by the SARS-CoV-2 pathogen). As a foundational architecture, we used the ImageNet pre-trained AlexNet architecture (Krizhevsky et al., 2017). With initial fine-tuning of the classifier on our dataset, we, then, engage in several cycles of XIL to improve the classifier as well as to try to generate novel insights.

Let the XIL Cycle Begin

The XIL process has the advantage that, after each training cycle, the results and the corresponding explanations for classification can be inspected and can, then, be readjusted according to the human's expertise. Our team consisted of computer scientists, information systems researchers, as well as radiologists and pneumologists. In the case of our application area, namely the detection of COVID-19-based viral pneumonia, especially pneumologists and radiologists played a vital role in assessing potential errors that could bias the system and ultimately put its usefulness at risk.

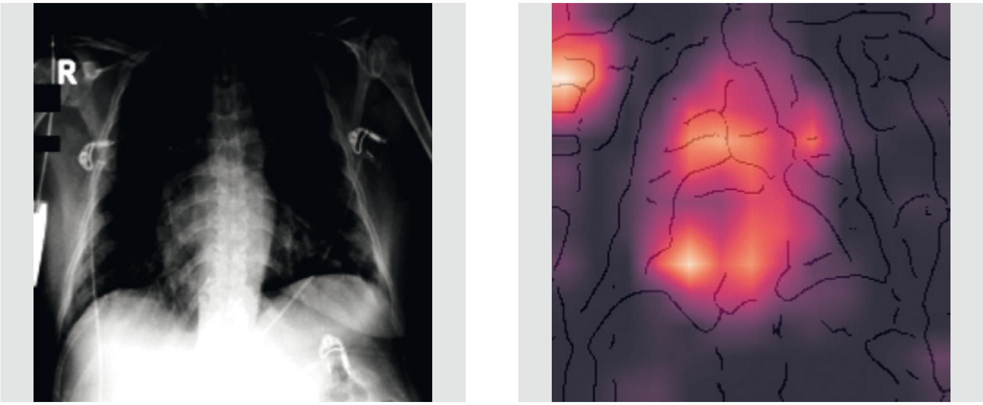

Figure 2: Original X-Ray on the Left, X-Ray with Grad-CAM Overlay on the Right (highlighted regions indicate a focus of the CNN on these salient areas)

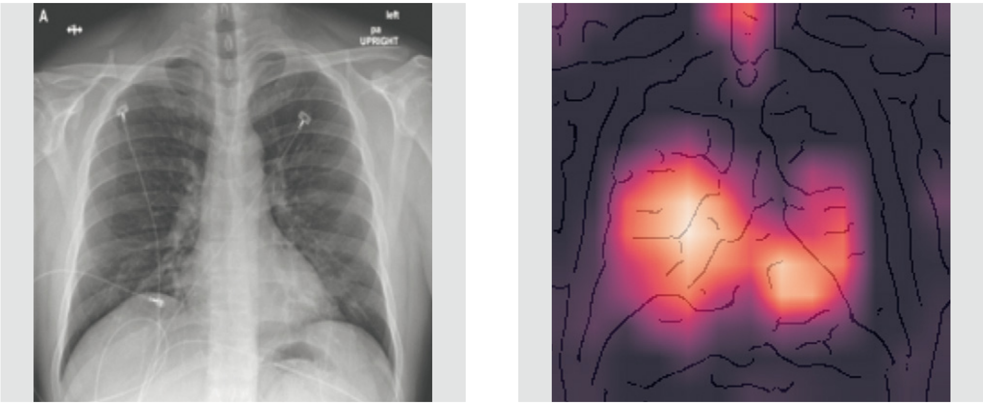

Figure 3: Original X-Ray on the Left, X-Ray with Grad-CAM Overlay on the Right (highlighted regions show the effect of annotations and penalization)

Cycle 1

At the beginning of Cycle 1, we train the CNN with the available X-ray images. Our resulting classifier showed high performance metrics, thus, indicating a very good classifier (accuracy: 91.06%, precision: 93.64%, recall: 92.53%). With the aid of explainability features, we are, then, able to provide explanations for each classified image in the test set. For generating the explainability, we applied the Captum library (see https://captum.ai/). We initially used Grad-CAM (e.g., Selvaraju et al., 2017), providing coarse results (see Figure 2), but, then, switched to Captum's VarGrad implementation to which we added a self-developed red/blue filter to provide a more precise explanation (e.g., Figure 4).

After thorough inspection of the test cases by the physicians and the rest of the team, we concluded that several obvious confounders (such as letters or medical instruments in the images, see the "R" in Figure 2) needed to be excluded or penalized in the model.

Cycle 2

We, thus, proceeded to refine the database of the model by letting crowdworkers from Mechanical Turk annotate important regions for the classification (torso, lungs, throat) in about 3,000 X-rays. In combination with the generated annotations, we fine-tuned the CNN again – now penalizing the classifier for unimportant regions (Figure 3 shows now that the algorithm focuses on the actual areas of interest). We again arrived at a very good model as indicated by the performance metrics (accuracy: 94.50%, precision: 96.30%, recall: 92.93%). After thorough inspection with the help of explanations in Cycle 2, the research team discovered another important issue: apparently, the used data were systematically biased. The labelled classes for the 'normal' and 'viral pneumonia' consisted of mainly pediatric images, while the 'COVID-19' class consisted of people of mixed age. This issue only became apparent, since the explanation features indicated that the algorithm looked at specific

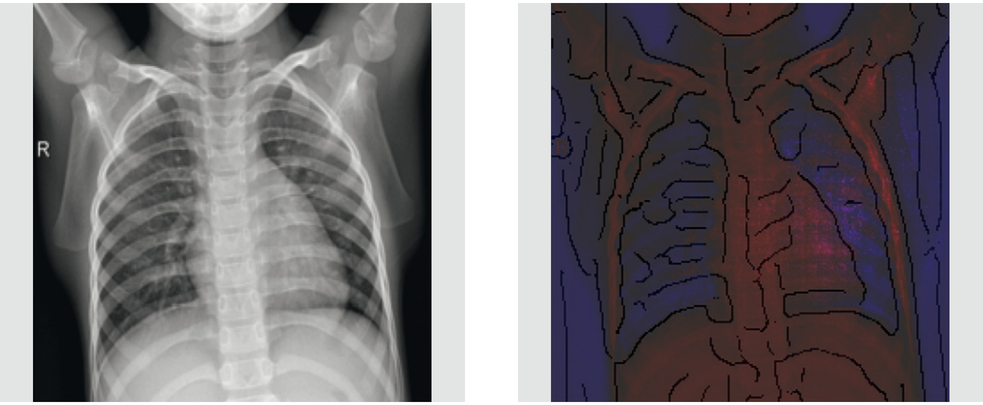

Figure 4: Original X-Ray on the Left, X-Ray with VarGrad Red/Blue Saliency Map Overlay on the Right (highlighted regions indicate a focus of the CNN on these salient areas)

skeletal characteristics of children; notably, we discovered that this issue, thus, persists in various published papers that used the same database but they were not able to recognize this potentially critical flaw.

Cycles 3a and 3b

The findings of XIL Cycle 2 lead us to subsequently revise the database in different ways, discarding the 'COVID-19' class in Cycle 3a to inspect explanations and accuracy of a binary classifier for viral pneumonia, as well as shifting to a different database in Cycle 3b (Wang et al., 2017). In doing so, we achieve a less accurate, but yet potentially more generalizable and meaningful model in this final cycle.

Summary of Results

To summarize our research efforts, we explored the utility of employing XIL in the construction of a mature CDSS for the case of classifying COVID-19-based viral pneumonia from X-rays. In our research efforts, we are able to show the benefits of XIL, namely a deep inspection of the classifier, not only providing accountability and transparency of the classifier, but also helping to uncover potentially dangerous confounders, which helps to finally arrive at realistic and promising models for practice. In a general perspective, our insights are not only important for the medical domain or for image data, but they foreshadow the potential of a promising novel methodology for every kind of data and domain. For example, XIL may also be highly helpful for the assessment of different kinds of algorithmic bias in ML-based systems to correct systems and may, thus, help organizations in fulfilling regulatory demands. In conclusion, we are confident that XIL may be further explored by science and adopted in practice to contribute to the usage of accurate, unbiased, and transparent ML-based systems.

References

Chowdhury, M. E.; Rahman, T.; Khandakar, A.; Mazhar, R.; Kadir, M. A.; et al.:
Can AI Help in Screening Viral and COVID-19 Pneumonia?
In: IEEE Access, 8 (2020), pp. 132665-132676.

Hinz, O.; Pfeuffer, N.; Stammer, W.; Schramowski, P.; Abdel-Karim, B. M.; et al.:
How Explainable Interactive Machine Learning Can Solve the Most Pressing Problems in Machine Learning – The Example of Diagnosing Viral Pneumonia.
In: Working Paper, 2020.

Krizhevsky, A.; Sutskever, I.; Hinton, G. E.:
ImageNet Classification with Deep Convolutional Neural Networks.
In: Communications of the ACM, 60 (2017) 6, pp. 84-90.

Lai, J.; Ma, S.; Wang, Y.; Cai, Z.; Hu, J.; et al.:
Factors Associated with Mental Health Outcomes Among Health Care Workers Exposed to Coronavirus Disease 2019.
In: JAMA Network Open, 3 (2020) 3, e203976, pp 1-12.

Mei, X.; Lee, H. C.; Diao, K. Y.; Huang, M.; Lin, B.; et al.:
Artificial Intelligence-Enabled Rapid Diagnosis of Patients with COVID-19.
In: Nature Medicine, 26 (2020) 8, pp. 1224-1228.

Nair, A.; Rodrigues, J. C. L.; Hare, S.; Edey, A.; Devaraj, A.; et al.:
A British Society of Thoracic Imaging Statement: Considerations in Designing Local Imaging Diagnostic Algorithms for the COVID-19 Pandemic.
In: Clinical Radiology, 75 (2020) 5, pp. 329-334.

Schramowski, P.; Stammer, W.; Teso, S.; Brugger, A.; Herbert, F.; et al.:
Making Deep Neural Networks Right for the Right Scientific Reasons by Interacting with Their Explanations.
In: Nature Machine Intelligence, 2 (2020) 8, pp. 476-486.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D.:
Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization.
In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); Venice, Italy, 2017, pp. 618-626.

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R. M.:
ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases.
In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Honululu (HI), US, 2017, pp. 2097-2106.