The black box problem of AI in oncology

Markus Hagenbuchner
University of Wollongong, School of Computing and Information Technology, Wollongong, Australia
E-mail: markus@uow.edu.au

Abstract. The rapidly increasing amount and complexity of data in healthcare, the pace of published research, drug development, biomarker discovery, and clinical trial enrolment in oncology renders AI an approach of choice in the development of machine assisted methods for data analysis and machine assisted decision making.

Machine learning algorithms, and artificial neural networks in particular, drive recent successes of AI in oncology. Performances of AI driven methods continue to improve with respect to both speed and precision thus leading to a great potential for AI to improve clinical practice. But the acceptance and a lasting breakthrough of AI in clinical practice is hampered by the black box problem. The black box problem refers to limits in the interpretability of results and to limits in explanatory functionality. Addressing the black box problem has become a major focus of research [1]. This talk describes recent attempts to addressing the black box problem in AI, offers a discussion on the suitability of those attempts for applications to oncology, and provides some future directions.

1. Introduction
Artificial Intelligence (AI) is a broad interdisciplinary field whose success is largely driven by advances in artificial neural networks (ANNs). Deep learning methods such as convolutional neural networks, generative adversarial networks, and graph neural networks are particularly influential ANNs that have driven AI to breakthroughs in numerous application domains.

Results produced by many traditional machine assisted non-parametric methods such as regression, decision trees, K-Nearest Neighbour, and rule based learning were found uncompetitive to those obtained by modern AI methods. AI is rapidly replacing traditional methods and has become one of the most important technologies that transform oncology in a wide range of applications such as:

- **Diagnosis**: Cancer imaging and detection, cancer recognition, image analysis, image segmentation, pathologic diagnosis, genotype-phenotype correlation, mutation detection and identification.
- **Prognosis and Prediction**: Toxicity of treatment, outcome prediction and survivability, cancer risk prediction.
- **Decision support**: Biomarker discovery, patient profiling, cancer management, risk modelling and prediction.
- **Treatment**: Optimal dose identification and energy deposition modelling in radiotherapy, patient journey optimization.
ANNs are increasingly applied in oncology for the purpose of assisting clinicians and patients in decision making processes. A long-standing problem with ANN, the black-box problem, inhibits wider adoption of AI in such decision making processes. It is important for a clinician and for patients to understand why a given machine response was made to be able to make founded and informed decisions. There is significant risk associated with methods that would require humans to blindly trust the result of a machine. Without interpretation facilities the suitability of ANNs for many decision support applications is limited [2].

ANNs are imperfect systems that can make errors. But due to the black-box problem we cannot understand why a particular error was made. An understanding of the factors that led to the error is crucial for the design and development of ANNs that avoid making subsequent errors of the same or similar nature. The back-box problem is particularly unhelpful in oncology since many processes in oncology require certification or underlie regulatory requirements. It is imperative to have transparency and interpretability for AI solutions to gain regulatory acceptance [3]. The black-box problem is not new. Relevant AI methods have their foundation in [4]. Technological limitations and knowledge gaps prevented the development of approaches that tackle the black-box problem until recently. The black-box problem is a hard problem that describes two inabilities:

1. The inability to explain what the values inside the model actually represent, and
2. The inability to explain reasons that led the model to produce a given output.

These problems are closely linked but differ profoundly. The first problem concerns the understanding of how a given model works whereas the second problem concerns the understanding of why a given model produced a particular result. Research and methods that address the first problem are called “Explainable AI” whereas “Explanatory AI” or “Interpretable AI” addresses the latter problem.

2. Related work
There is a distinction among models that are interpretable by design (such as regression, decision trees, K-Nearest Neighbour, rule based learning), and black box models (e.g. Support Vector Machines, Artificial Neural Networks) which need to be explained by means of external augmented techniques. An alternative categorization of these models are transparent models and ad-hoc explainability. Ad-hoc techniques are developed to augment models which are not readily interpretable by design. Ad-hoc explainable techniques can be classified into:

- Text explanations: Techniques that deal with explainability by learning to generate text explanations that help explain the results from a given model [6].
- Visual explanation: Techniques for ad-hoc explainability deal with visualising the black box model’s behaviour [7,8]. These visualisation approaches aid human interpretation by visualizing complex interactions among variables involved in the model.
- Local explanation: Techniques that create explanations representing smaller solution subspaces which are relevant for the whole model. These techniques aim at obtaining discernability characteristics to explain certain parts of the whole model [9].
- Explanations by simplification: A second model is developed based on a trained black box model. The second model aims at reducing the complexity of the black box model to help simplify the understanding of the functioning of the original black box model [10].
- Feature importance/relevance: Techniques for the computation of a relevance score or importance score for the input variables. These methods can reveal the sensitivity of an input variable on the output of a black box model [5].

Plug-in methods are developed to work with a variety of AI methods. For example:
LIME (Local Interpretable Model-Agnostic Explanations): Generates locally linear models around the predictions of a black box model to explain its functioning [9]. The method is a variant of explanations by simplification as well as of local explanations.

G-REX: Extracts rules from some AI methods [11]. This was further enhanced in [12] to explain complex AI models in a human-interpretable form. Limitations for these approaches are (i) AI expertise is needed to operate these methods, (ii) correlation between features are ignored which can lead to unrealistic explanations, (iii) the method is very sensitive to data variations. Small value changes can lead to radically different explanations.

SHAP (SHapley Additive exPlanations): This is an approach to feature relevance identification [5]. The method calculates additive feature importance scores for predictions with a set of desirable properties that the black box model lacked.

Other approaches that tackle the contribution of features to predictions as in SHAP are (i) coalitional Game Theory [13], (ii) local gradients [14], and (iii) the automatic STRucture IDentification method (ASTRID) that inspect which input attributes are exploited by a classifier [15]. Limitations affecting SHAP and its derivates are (i) the method is computationally expensive which makes it impractical to use in the presence of large number of instances, (ii) the techniques often ignore feature dependence and correlation (with the exception of the algorithm called TreeSHAP).

Methods that obtain visual explanations are (i) Sensitivity Analysis methods (data based, Monte-Carlo, cluster-based methods) and a novel input importance measure [8,16], and (ii) a modular ensemble technique which uses a dimension reduction technique and prototyping methods to discover correlations and importance of input features [7].

There have been several other attempts to improving explanations by means of modular ensemble methods: One of the earliest studies proposed to create a second, less complex model from a set of randomly selected samples from a set of data [17]. The Simplified Tree Ensemble Learner (STEL) is a more effective and recent approach for simplification [10]. The approach is similarly to [18] in that they suggest the creation of two models where one model is in charge of interpretation and the other of prediction by means of Expectation-Maximization. DeepSHAP, too, is an ensemble model which stacks multiple classifier systems in addition to Deep Learning models [19].

There are studies introduce explanations to deep learning models such as Deep Multi-Layer Networks (MLP), Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). Corresponding approaches target explainability by using external augmentation techniques for either local explanations, feature importance detection, or both.

MLP: Relatively little has been done to address explainability in MLP. DeepRED, a computationally expensive approach, uses a decompositional approach to rule extraction by engaging the decision tree method on neuron level for rule extraction [20]. An approach by [21] uses model simplification through a distillation method called Interpretable Mimic Learning. The algorithm uses gradient boosting trees and hence the approach is also computationally expensive. DeepLIFT is an approach to compute feature importance scores in an MLP by computing an interestingness measure known as LIFT which is widely used in Association rule Mining [22]. The approach is computationally efficient, but cannot detect relationships between inputs and the explanatory capabilities are limited.

CNN: Due to the popularity of CNNs numerous attempts have been made to address explainability. For example, Deconvnet uses the feature map from a selected CNN layer to reconstruct the maximum activations [23]. These reconstructions can reveal some insight about the most influential parts of an input image. A subsequent work demonstrated how a saliency map can be generated by iteratively occluding different region of an input image [23]. The approach is computationally very expensive but can significantly enhance the
explanability of a CNN. A different approach is to compute a loss for each filter in the first convolutional layer [1].

• RNN: RNNs are commonly used when processing temporal information (i.e. time series of data, data sequences). An approach to extract a specific propagation rule of the RNN uses the Long Short Term Memory (LSTM) and analyses the Gated Recurrent Units of the LSTM [24]. RETAIN (REverse Time AttentIoN) is an approach to detect influential past patterns by means of a two-level neural attention model [25]. These approaches introduce limited explanability but provide a good framework for future improvements.

There is thus a wide spectrum of approaches to achieving explanatory AI. Research is still in very early stages. Current approaches afflict constraints which prevent a wider adoption to oncology. Many existing methods are either not scalable, robust, or, most commonly, do not offer legible results. For example, decision trees provide legible explanation of results but the corresponding algorithms are computationally inefficient and are not robust to noise and outliers. SHap on the other hand is computationally efficient but does not provide legible explanations. SHap requires human experts to translate the numerical associations found by SHap.

3. The way ahead

Current research shows that explanatory AI is possible but further research is needed to address a gap in knowledge on how to obtain legible interpretations of results produced by a given AI model to oncologists and cancer patients. Key requirements for explanatory AI in oncology are to obtain:

1. Legible, human interpretable explanations.
2. Explanations that are suitable for the target audience: explanations should not require an AI expert for interpretation of explanations, explanations should be suitable for interpretation by an oncologist, or by cancer patients, or both.
3. Explanations provide meaningful information of either the logic involved, or the role of input attributes and their values, or both.
4. If used for automated or semi-automated decision-making, then the methods should also explain “legal or similarly relevant effects” on individuals.

These requirements address questions of ethics, accountability, safety and liability. Point 3 and 4 are legally required i.e. by the EU General Data Protection Regulation.

In addition to these requirements the following capabilities are needed to achieve full acceptance of AI in oncology:

5. Produce explanations in natural language.
6. Explanations need to be informative for a given user. Obvious explanations or explanations that are already known by a user should be avoided or suppressed.
7. Incorporation of user feedback mechanisms for obtaining cooperative argumentative AI.

Explanatory AI for oncology is within reach but research is needed. Collaborative research with domain experts will make AI an accepted tool and allow AI to make important contributions to advances in oncology.

Some early work has been conducted in this area. A research team at the University of Wollongong developed machine learning ensemble methods that combine an explanatory subsystem with given black box method [7]. They demonstrated their work as part of a proof-of-concept study. They showed that the approach can introduce an effective explanatory subsystems to AI which in turn can significantly enhance precision of results. They demonstrated this on breast cancer survivability prediction of a large set of patients (SEER dataset) where accuracy improved from 63.27% to 86.96%
while offering previously unknown insights into factors that influence survivability [7]. Though challenges of translating machine interpretable results into human interpretable results remain.

4. Conclusions
Significant advances in addressing the black-box problem in AI have been made in recent years. Challenges remain to render AI a valuable, permitted, and more widely accepted tool in oncology. Continued collaborative research engagements between oncologists, radiologists, and machine learning experts will allow AI to accelerate and drive advances in oncology.

5. Acknowledgements
The author acknowledges the funding received from the EIS seed grant (University of Wollongong) to carry out some of the presented work in this paper.

6. References
[1] Zhang Q and Zhu S C 2018 Visual interpretability for deep learning: a survey https://arxiv.org/abs/1802.00614
[2] Young S 2019 What the heck is Explainable (XAI) AI and why do we need it online https://becominghuman.ai/what-the-heck-is-explainable-xai-and-why-do-we-need-it-c1a37b64972c
[3] Sokalski M, Klous S and Chandrasekaran S 2019 Controlling AI, The imperative for transparency and explainability KPMG available online: https://assets.kpmg/content/dam/kpmg/br/pdf/2019/08/Controlling%20AI%20online.pdf
[4] Rosenblatt F 1961 Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms
[5] Lundberg S M and S-I Lee 2017 A unified approach to interpreting model predictions Advances in Neural Information Processing Systems 4765-4774
[6] Bennetot A, Laurent J L, Chatila R and Rodriguez N Diaz 2019 Towards explainable neural-symbolic visual reasoning NeSy Workshop IJCAI
[7] Shukla N, Hagenbuchner M and Win K T 2018 Breast cancer data analysis for survivability studies and prediction Computer Methods and Programs in Biomedicine 155 199-208
[8] Cortez P and Embrechts M J 2013 Using sensitivity analysis and visualization techniques to open black box data mining models Information Sciences 225 1-17
[9] Ribeiro M T, Singh S and Guestrin C 2016 Why should I trust you? Explaining the predictions of any classifier ACM Int. Conf. on Knowledge Discovery and Data Mining 1135-44
[10] Deng H 2019 Interpreting tree ensembles with intrees, Int. J. of Data Sci. and Anal. 7 277-87
[11] Konig R, Johansson U, and Niklasson 2008 L. G-rex: A versatile framework for evolutionary data mining IEEE Int. Conf. on Data Mining 971-74
[12] Su G, Wei D, Varshney K R and Malioutov D M 2015 Interpretable two-level boolean rule learning for classification Corr
[13] Strumbelj E and Kononenko I 2010 An efficient explanation of individual classifications using game theory J. of Machine Learning Research 11 1-18
[14] Robnik-Šikonja M and Kononenko I 2018 Explaining classifications for individual instances IEEE Trans. on Knowledge and Data Engineering 20 589-600
[15] Henelius A, Puolamäki K and Ukkonen A 2017 Interpreting classifiers through attribute interactions in datasets Proc. of the ICML Workshop on Human Interpretability in Mach. Learning
[16] Cortez P and Embrechts M J 2011 Opening black box data mining models using sensitivity analysis IEEE Sym. on Comp. Intelligence and Data Mining (CIDM) IEEE 341-348
[17] Domingos P 1998 Knowledge discovery via multiple models, Intel. Data Analysis 2 187-202
[18] Hara S and Hayashi K 2018 Making tree ensembles interpretable Proc. of the Twenty-First Int. Conf. on Artificial Intelligence and Statistics PMLR 84 77-85
[19] Chen H, Lundberg S and Lee S-I 2019 Explaining models by propagating shapley values of local components

[20] Zilke J R, Mencía E L and Janssen F 2016 DeepRED–rule extraction from deep neural networks Int. Conf. on Discovery Science Springer 457-473

[21] Che Z, Purushotham S, Khemani R and Liu Y 2016 Interpretable deep models for ICU outcome prediction AMIA Annual Symposium Proc. 371-380

[22] Shrikumar A, Greenside P and Kundaje A 2017 Learning Important Features through Propagating Activation Differences Proc. of the Int. Conf. on Mach. Learning 70 3145-53

[23] Zeiler M D and Fergus R 2014 Visualizing and understanding convolutional networks European Conf. on Comp. Vision 818-33

[24] Arras L, Montavon G, Müller K and Samek W 2017 Explaining recurrent neural network predictions in sentiment analysis Proc. of the 8th Workshop on Comp. Approaches to Subjectivity, Sentiment and Social Media Analysis 159-68

[25] Choi E, Bahadori M T, Sun J, Kulas J, Schuetz A and Stewart W 2016 Retain: An interpretable predictive model for healthcare using reverse time attention mechanism Adv. in Neural Information Processing Systems 3504-12