other words. We do this by comparing the network's activations on the original input to its activations when removing individual words. The resulting differences in activations show the interaction between different words of the input in different layers. The interactive tool we built allows users to enter a baseline input and directly perturbate this input by excluding words and observing the influence on activations through all layers of the network including the model's predictive output.

Figure 1 shows how the initial removal of a single noun affects the processing of the activations belonging to all other words. The words most strongly influenced are those that have a linguistic relationship with the removed word, like a preposition referencing a noun. Similarly, we analysed (amongst other perturbations) the effect of changing the word order in the source or target sentences revealing the impact of positional characteristics.

State-of-the-art models for many natural language processing tasks are artificial neural networks which are widely considered to be black box models. Improvements are often the result of untargeted parameter sweeps and architectural modifications. In order to efficiently and systematically improve such models a deeper understanding of their inner workings is needed. We argue that interactive exploration through input perturbation is a promising and versatile approach for inspecting neural networks' decision processes and finding specific target areas for improvement.

References:
[1] M. D. Zeiler, R. Fergus: "Visualizing and understanding convolutional networks", European Conference on Computer Vision, Springer, Cham, 2014.
[2] J. Yosinski, et al.: "Understanding neural networks through deep visualization", in International Conference on Machine Learning (ICML) Workshop on Deep Learning, 2015.
[3] S. Bowman, et al.: "A large annotated corpus for learning natural language inference", in Proc. of EMNLP, 2015.

Please contact:
Alexander Dür, TU Wien, Austria
+4369919023712,
alexander.duer@tuwien.ac.at

Personalisable Clinical Decision Support System

by Tamara Müller and Pietro Lió (University of Cambridge)

We introduce a Clinical Decision Support System (CDSS) as an operation of translational medicine. It is based on random forests, is personalisable and allows a clear insight into the decision making process. A well-structured rule set is created and every rule of the decision making process can be observed by the user (physician). Furthermore, the user has an impact on the creation of the final rule set and the algorithm allows the comparison of different diseases as well as regional differences in the same disease.

Neurodegenerative diseases such as Alzheimer's and Parkinson's impact millions of people worldwide. Early diagnosis has proven to greatly increase the chances of slowing the diseases' progression [2]. Correct diagnosis often relies on the analysis of large amounts of patient data, and thus lends itself well to support from machine learning algorithms, which are able to learn from past diagnosis and see clearly through the complex interactions of a patient's symptoms. Unfortunately, many contemporary machine learning techniques fail to reveal details about how they reach their conclusions, a property considered fundamental when providing a diagnosis. This is one reason why we introduce our personalisable CDSS that provides a clear insight into the decision making process on top of the diagnosis. Our algorithm enriches the fundamental work of Masheyekhi and Gras [1] in data integration, personal medicine, usability, visualisation and interactivity.

Our algorithm performs by extracting a rule set from a random forest, which is then minimised within several steps. The algorithm can be divided into three major steps. (1) Firstly, a random forest, an ensemble of decision trees, is created as the foundation of the algorithm. (2) Secondly, a set of rules is extracted from the random forest. (3) Thirdly, this rule set is reduced significantly. The user can influence step (3) by preferring as important considered features. The algorithm is implemented in Python and we trained it to predict whether patients are likely to suffer from Alzheimer's or Parkinson's disease, but it is a generic algorithm that can be applied to any kind of disease. Three main factors are taken into account during the reduction process: (a) the performance of individual rules, (b) the rules' transparency, and (c) the personal preferences of the user. The reduction leads to a new, smaller set of rules, with predictive performance comparable to the original set, but much easier to comprehend. A clear understanding of the process behind a diagnosis is crucial for both doctor and patient, and it is hoped that systems like this one will become increasingly prevalent as we continue to improve the state-of-the art in predictive medicine.

Mashayekhi and Gras [1] introduced two methods called RF+HC and RF+HC_CMPR which allow to extract a rule set from random forests. The main idea of their work is to reduce the number of rules dramatically and therefore increase the comprehensibility of the underlying model. We expanded this idea and added another personalisable layer to it to allow the user to add data driven and personal evidence to the algorithm.

We applied our algorithm to different data sets with a variety of data types like magnetic resonance images, bio-

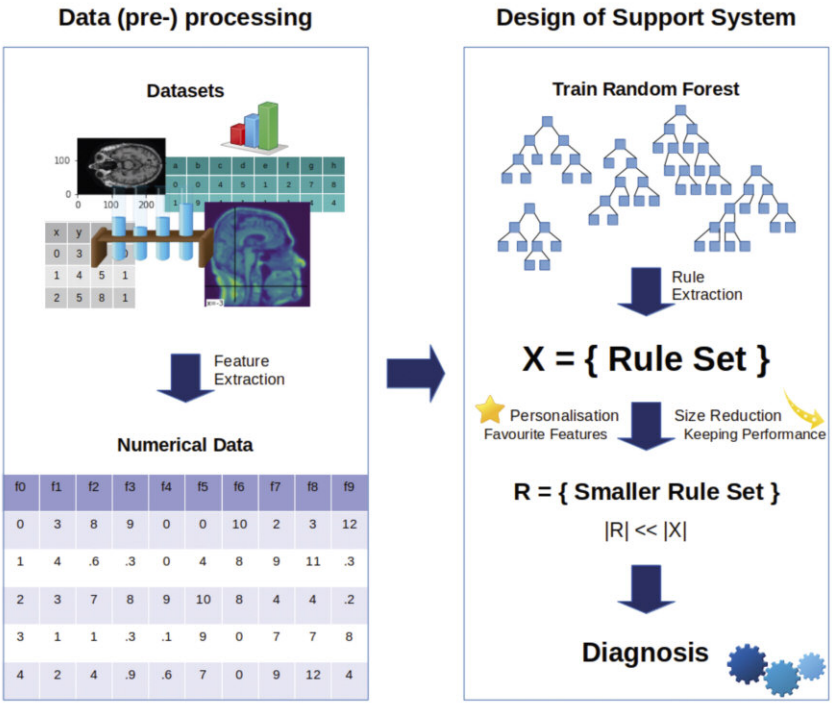

Figure 1: A visual overview of the algorithm.

medical voice measurements, drawings, demographic characteristics, etc. After the rule set is extracted from the forest, each rule is assigned with a score. This score depends on the rule's length, its performance on the training set, and whether it contains any preferred features. Based on this score, the weakest rules are eliminated to minimise the rule set. The application to Alzheimer's and Parkinson's data sets has shown that the reduction of the rule set combined with the consideration of preferred features, does not generally impair the performance. It sometimes even has a positive impact on the prediction, as setting preferred features can diminish the risk of over-fitting and take regional characteristics and expertise into account. Furthermore, a deliberately reduced set of rules is less likely to contain noisy rules [1]. The algorithm also reveals information about the importance of features, which allows to draw conclusions about diseases and their indicators. We achieved accuracies of up to 100% and one rule set could be reduced to 0.5% of the original rule set size without reducing its performance, e.g.

A graphical user interface allows the algorithm to be used easily, where data of a new patient can be added intuitively. Furthermore, all rules can be inspected by the physicians. It is also possible to show statistical distributions to get a better understanding of which features are more or less important in the decision making process. It allows to compare different diseases and detect regional differences like urban characteristics versus rural ones or international variations in diseases.

References:

[1] M. Mashayekhi, R. Gras: "Rule extraction from random forest: the RF+ HC methods", 2015.

[2] Alzheimer's Association: "2017 Alzheimer's disease facts and figures" Alzheimer's & Dementia 13, no. 4 (2017): 325-373.

Please contact:
Tamara Müller, Pietro Lió
University of Cambridge, UK
contact@tamaramueller.com
pl219@cam.ac.uk

Putting Trust First in the Translation of AI for Healthcare

by Anirban Mukhopadhyay, David Kügler (TU Darmstadt), Andreas Bucher (University Hospital Frankfurt), Dieter Fellner (Fraunhofer IGD and TU Darmstadt) and Thomas Vogl (University Hospital Frankfurt)

From screening diseases to personalised precision treatments, AI is showing promise in healthcare. But how comfortable should we feel about giving black box algorithms the power to heal or kill us?

In healthcare, trust is the basis of the doctor-patient relationship. A patient expects the doctor to act reliably and with precision and to explain options and decisions. The same accuracy and transparency should be expected of computational systems redefining the workflow in healthcare. Since such systems have inherent uncertainties, it is imperative to understand a) the reasoning behind such decisions and b) why mistakes occur. Anything short of this transparency will adversely affect the fabric of trust in these systems and consequently impact the doctor-patient relationship.

Current solutions for transparency in deep learning (used synonymously with AI) centre around the generation of heatmaps. These highlight high-impact image regions on deep learning decisions. While informative in nature, direct adaptation of such methods into healthcare is insufficient, because the actual reasoning patterns remain opaque and leave a lot of room for guesswork. Here, deep generative networks show promise by generating visual clues as to why a decision was made [1].

We believe transparency in image based algorithmic decision making can only be achieved if expert computer scientists and healthcare professionals (radiologists, pathologists etc.) closely collaborate in an equal-share environment.