Report 1 **Francisco Maria Calisto**
Prof. Dr. Jacinto Nascimento Date: 14/03/2018
Prof. Dr. Nuno Nunes

1 Problem Statement

The hereby **Report 2** will state a first essay for literature knowledge that might support our research. We evaluate several research work. Works from (i) **Classifier Predictors**; and (ii) . The following sections will explain each work.

2 Investigation/Research

In this section we describe our investigation and research over the proposal readings. Investigation throughout the examination of the reading facts, novel possibilities and results. Also, addressing the research reasoned conclusions of those readings.

2.1 Classifier Predictors

Many learning-based algorithms are able to solve problems that previously seemed completely impossible. They can look at an image and describe what they depict in a sentence, or even turn medical diagnosis into a much more accurate by giving doctor's a second opinion. Amazing new results keep appearing every single week. However, an important thing that we need to solve is that if we deploy these Neural Networks (NN) in a production environment, we would want to know if we are relying on a good or bad decision. The narrative is very simple: if we do not trust a classifier, we will not use it. And perhaps the best way of earning the trust of a human would be if the Artificial Intelligence (AI) could explain how it came to a given decision.

Strictly speaking, a NN can explain it to us, while showing us hundreds of thousands of neuron activations that are completely unusable for any sort of intuitive reasoning. That said, what is even more difficult to solve is that this explanation happens in a way that we can interpret. An earlier approach has use decision trees that described what the learner looks at and how it uses this information to arrive to a conclusion. The work presented with the title *Why Should I Trust You? Explaining the Predictions of Any Classifier* from Ribeiro et al. [3] new work is quite different. For instance, imagine that a NN would look at all the information we know about a patient and tell us that this patient likely has the flu. Also, in the meantime, it could tell us the fact that the patient has a headache and sneezes a lot contributed to the conclusion that he has the flu, but, the lack of fatigue is notable evidence against it. Our doctor could take this information and instead of blindly relying on the output, could make a more informed decision.

Other related works [2, 1] are describing an automated methodology for the analysis of unregistered medical imaging views in order to estimate the patient's risk of developing cancer. The main innovation behind this methodology lies in the use of Deep Learning (DP) models for the problem of jointly classifying unregistered medical imaging views and respective segmentation maps of lesions. This is a holistic methodology that can classify medical imaging exams, containing several modality of views and the segmentation maps, as opposed to the classification of individual lesions, which is the dominant approach in the field. The authors also demonstrate that the proposed system is capable of using the segmentation maps generated by automated detection systems, and still producing accurate results.

A big additional selling point is that these techniques are model agnostic, which means that it can be applied to other learning algorithms that are able to perform medical classification. For example, complementing medical imaging with medical text (e.g. patient's clinical history). It is also a possibility that an AI is only right by chance, and if this is the case, we should definitely know about that. And here, in this example, with the additional explanation, it is rather easy to find out that we have a bad model that looks at some modality of the medical image and acknowledge if it is a severe lesion or not.

The tests indicate that humans make significantly better decisions when they lean on explanations that are extracted by these techniques. The source code of this project is also available.

3 Final Evaluation

References

[1] Gustavo Carneiro, Jacinto Nascimento, and Andrew P Bradley. Automated analysis of unregistered multi-view mammograms with deep learning. *IEEE transactions on medical imaging*, 36(11):2355–2365, 2017.

[2] Maryellen L Giger, Nico Karssemeijer, and Julia A Schnabel. Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. *Annual review of biomedical engineering*, 15:327–357, 2013.

[3] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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*, pages 1135–1144. ACM, 2016.