Explainable AI: An Intuitive Analysis of Deep Learning in Medical Imaging and Biosignals.

Eleftherios Trivizakis[1,2] and **Kostas Marias**[1,3]

[1]Computational BioMedicine Laboratory (CBML), Institute of Computer Science (ICS), Foundation for Research and Technology-Hellas (FORTH)

[2]Medical School, University of Crete

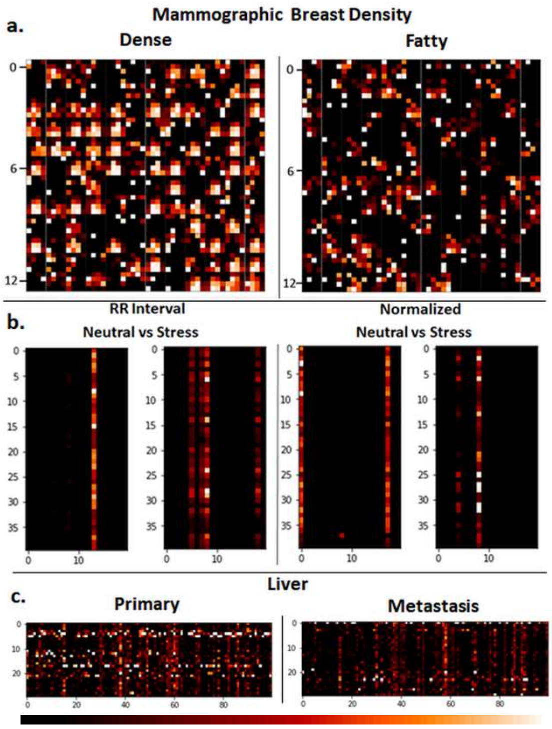
[3]Department of Electrical and Computer Engineering, Hellenic Mediterranean University

Presenting & Corresponding author: Eleftherios Trivizakis, email: trivizakis@ics.forth.gr

ABSTRACT

The introduction of state-of-the-art Deep Learning (DL) techniques in medicine provided outstanding performance in many cases better than expert clinicians. This initiated the debate on the explainability of these remarkable predictions of DL models. In particular, transparency of the inference process is a key feature for integrating AI in medical practice. DL architectures are characterized by the fully-automated convergence of millions of learnable parameters for matching the raw input data to their underlying information**.** The investigation of the low-level mechanisms with traditional methods can be demanding due to the extremely high complexity of the depended variables in the deep models.

An intuitive approach compatible with the medical practice may contribute to address this fundamental drawback in AI decision making process. The proposed meta-analysis of the inference process of the examined models can be achieved by visualizing features maps from the last convolutional layer [1] or activations from the neural-part [2-3] of the Convolutional Neural Network (CNN). The 2D CNN [1] was trained with mammographic images to perform breast density scoring (dense vs fatty – AUC 87%). The visualized feature maps from the last convolutional layer reveal different subsets of activated kernels (Fig. a, 3rd-6th line) for dense samples in contrast to fatty-labeled samples. For stress detection [2] in ECG time series the proposed deep network (AUC up-to 99.5%) demonstrates increased prediction confident visually evident by the distinct and consistent patterns (Fig. b) for the neutral and stress state respectively. The 3D model [3] predicts the liver cancer tissue (primary vs metastatic - AUC 80%) but although most of the subjects are well-differentiated, as their activations depict higher energy in Fig. c, the visualization is challenging due to the high-dimensionality of the feature vector after the fully-connected neural layer.



Our goal is to present novel ways of analyzing raw high-dimensional medical data in addition to demonstrating comprehensive visual representations of the AI's inference process towards clinical decisions interpretable and meaningful for expert clinicians.

REFERENCES

[1] Trivizakis E., Ioannidis GS, Melissianos VD, Papadakis GZ, Tsatsakis A., Spandidos DA, and Marias K. 2019. A novel deep learning architecture outperforming 'off-the-shelf' transfer learning and feature-based methods in the automated assessment of mammographic breast density. *Oncology Reports,* https://doi.org/10.3892/or.2019.7312.

[2] Giannakakis G., Trivizakis E., Tsiknakis M, and Marias K., 2019. A novel multi-kernel 1D convolutional neural network for stress recognition from ECG. *Eighth International Conference on Affective Computing and Intelligence Interaction (ACII),* **In press:** IEEE Xplore.

[3] Trivizakis E., Manikis GC, Nikiforaki K., Drevelegas K., Constantinides M., and Marias K. 2018. Extending 2-D Convolutional Neural Networks to 3-D for Advancing Deep Learning Cancer Classification with Application to MRI Liver Tumor Differentiation, *IEEE journal of biomedical and health informatics*, *23*(3), 923-930.

Eleftherios Trivizakis acknowledges the financial support of the Stavros Niarchos Foundation within the framework of the project ARCHERS ("Advancing Young Researchers' Human Capital in Cutting Edge Technologies in the Preservation of Cultural Heritage and the Tackling of Societal Challenges").