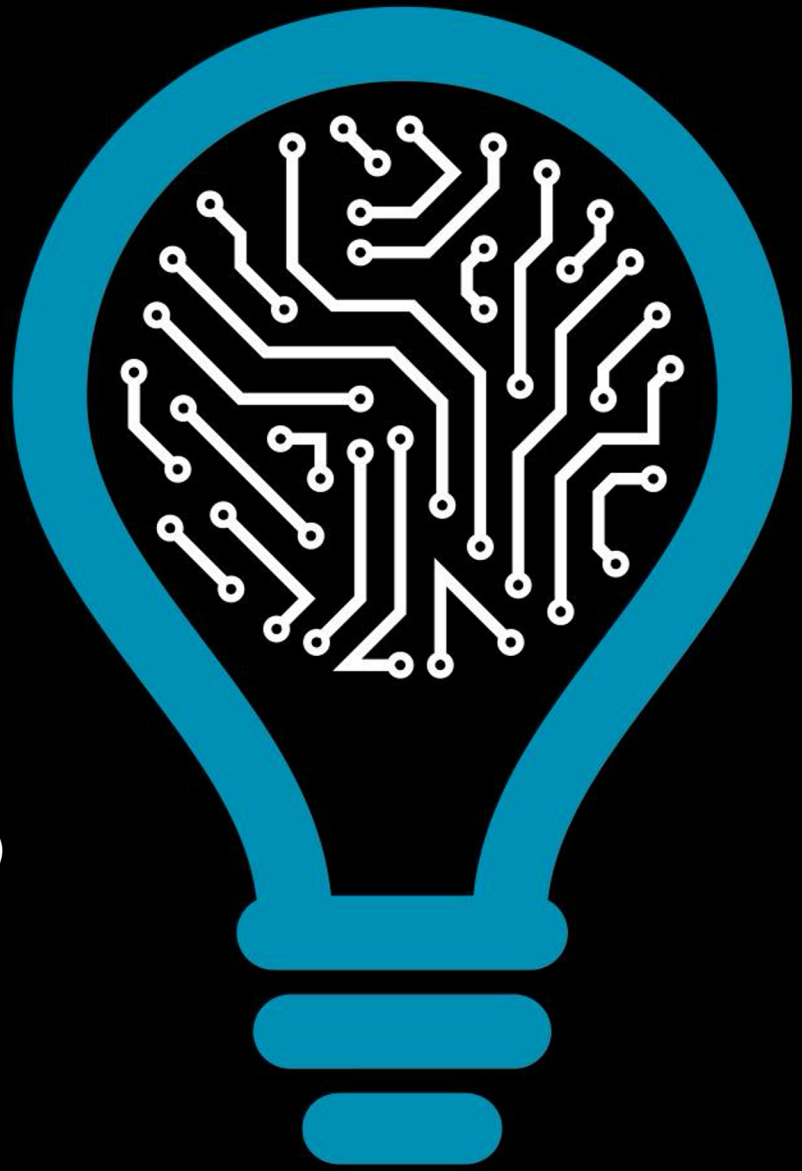
Interpretable Machine Learning and its Application to Medical Decision Support Systems

PhD Research Work | Supervisor: Jaime S. Cardoso

Group Meeting | AI Technology for Life | May 1, 2024
University of Utrecht, Netherlands

Tiago Filipe Sousa Gonçalves (tiago.f.goncalves@inesctec.pt)



INESCTEC
INSTITUTE FOR SYSTEMS AND COMPUTER ENGINEERING, TECHNOLOGY AND SCIENCE

INESCTEC

Outline

1. Introduction
2. Attention Mechanisms for Medical Image Analysis
3. Multi-modal Data Strategies for Medical Applications
4. Intrinsically Interpretable Models in Medical Context
5. Conclusion

INESCTEC

1. Introduction

Context

- **The increase of available computational power and the democratised access to a huge amount of data** has leveraged the development of novel artificial intelligence (AI) algorithms and their applications

- Deep learning techniques **have been challenging human performance** at some specific tasks such as cancer detection in biomedical imaging[1] or machine translation in natural language processing[2]

- However, most of these models work as black boxes (i.e., their internal logic is hidden to the user) that receive data and output results **without justifying their predictions in a human understandable way**[3]

Sources: [1] McKinney et al. "International evaluation of an AI system for breast cancer screening", [2] Belinkov and Glass "Analysis Methods in Neural Language Processing: A Survey", [3] Cynthia Rudin "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead"

Motivation

- The topic of explainable artificial intelligence (XAI) appeared intending to contribute to a more **transparent AI**[1]
  - There are three distinct strategies: **pre-, in- and post-model** methods

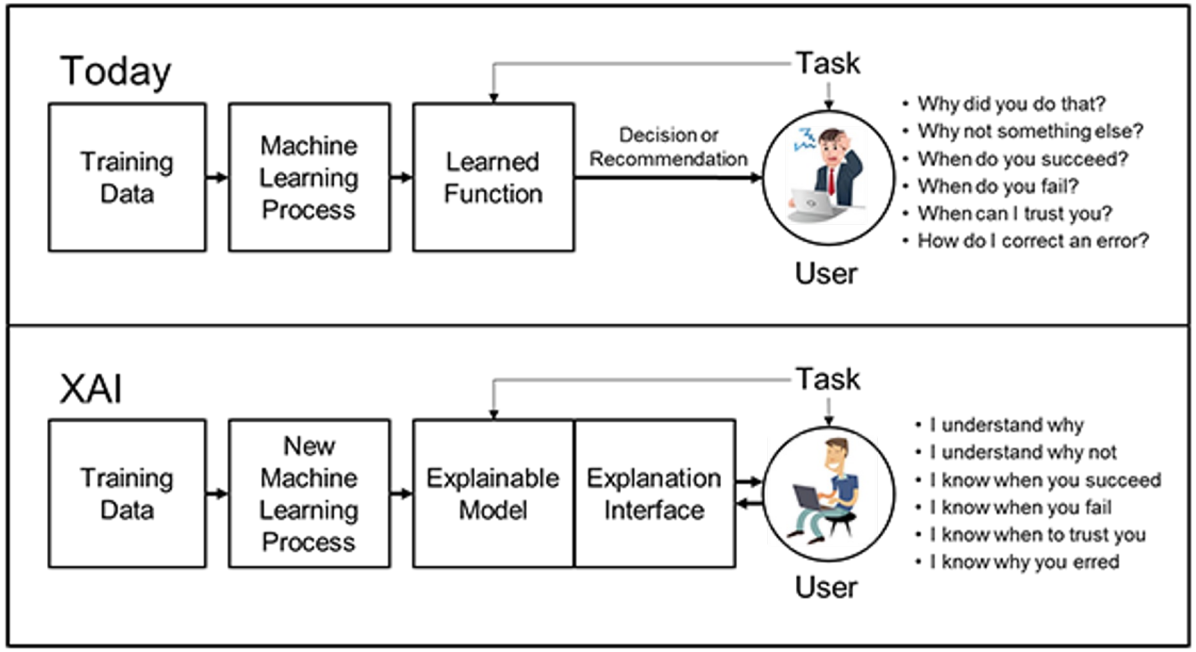


Figure: The concept of XAI, from the Defense Advanced Research Projects Agency (DARPA)[2]

Sources: [1] L. H. Gilpin et al. "Explaining Explanations: An Overview of Interpretability of Machine Learning", [2] https://www.darpa.mil/program/explainable-artificial-intelligence

Motivation

- The **generalised belief that complex models seem to uncover "hidden patterns"** actively contributed to the research and **development of post-model** methods

- There are **several drawbacks of exclusively investing** in a post-model strategy[1]:
  - **Explanations are just an approximation** to what the model computes
  - **Explanations may not provide enough detail** to understand what the model is doing

- It is fundamental to assess the quality of these explanations[1] and to dedicate more effort to pre- and in-model strategies
  - **Pre-model interpretability**: understanding the data distribution that we are dealing with will contribute to an increase of confidence with the posterior decisions and explanations[2]

  - **In-model interpretability**: since models that are inherently interpretable provide their explanations and are faithful to what the machine learning model actually computes[1]

Sources: [1] Cynthia Rudin "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead",
[2] Wilson Silva et al. "How to produce complementary explanations using an Ensemble Model"

INESCTEC

2. Attention Mechanisms for Medical Image Analysis

Introduction

- In AI systems, **some parts of the input data are more relevant than others** (e.g., in automatic translation systems, only a subset of words is relevant)[1]
  - In the deep learning context, the first successful implementations of attention mechanisms were accomplished with RNNs, which can learn and process data with a temporal component

- A possible taxonomy for the classification of attention mechanisms[1] proposes the following categories
  - **Number of Abstraction Levels:** single-, multi-level
  - **Number of Positions:** soft, hard, global, local
  - **Number of Representations:** single-, multi-representational, multi-dimensional
  - **Number of Sequences:** distinctive, co-attention, self-attention

Sources: [1] S. Chaudhari et al. “An Attentive Survey of Attention Models”

State of the Art

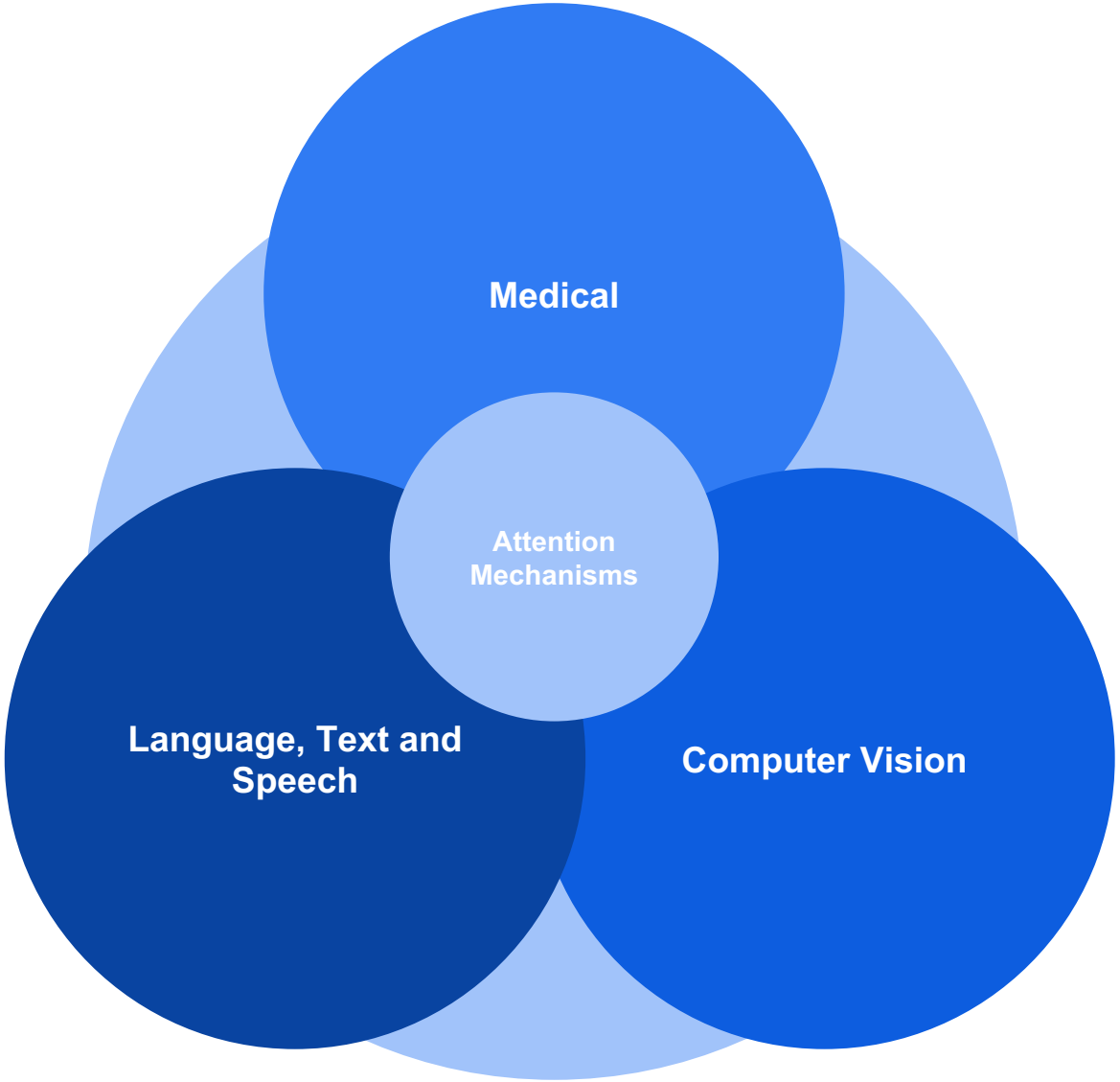


Figure: Overview of the State of the Art on Attention

State of the Art

Medical

- Most of the use-cases focus on **medical image segmentation or classification** using **different modalities** (e.g., computed tomography, magnetic resonance imaging, ultrasound[1], positron emission tomography)

- In automatic report generation, different attention methodologies (**contrastive**, **variational topic inference**) have been proposed to represent better the visual features of abnormal regions or to align image and language modalities in a latent space, thus improving the quality of the generated reports[2]

- The potential of Transformer-based architectures is also being explored in the medical context, as more recent methodologies on medical image segmentation are taking advantage of a hybrid use of the **Vision Transformer and the U-Net**[3]

Sources: [1] Aleksandar Vakanski et al. "Attention Enriched Deep Learning Model for Breast Tumor Segmentation in Ultrasound Images", [2] Fenglin Liu et al. "Contrastive Attention for Automatic Chest X-ray Report Generation", [3] Bingzhi Chen et al. "TransAttUnet: Multi-level Attention-guided U-Net with Transformer for Medical Image Segmentation"

Attention Mechanisms for Medical Applications

A survey on attention mechanisms for medical applications: are we moving towards better algorithms?[1]

- **Problem:** the impact of the presence or absence of an attention mechanism in the degree of interpretability of the models

- **Methodology:**
  - Models: DenseNet121, ResNet50, SEDenseNet121, SEResNet50, CBAMDenseNet121, CBAMResNet50, DeiT (Data efficient image Transformer)
  - Post-hoc xAI methods: DeepLIFT and LRP
  - Use cases: **diabetic retinopathy, pleural effusion in chest X-ray images, skin lesion**

Sources: [1] Tiago Gonçalves et al. "A survey on attention mechanisms for medical applications: are we moving towards better algorithms?"

Attention Mechanisms for Medical Applications

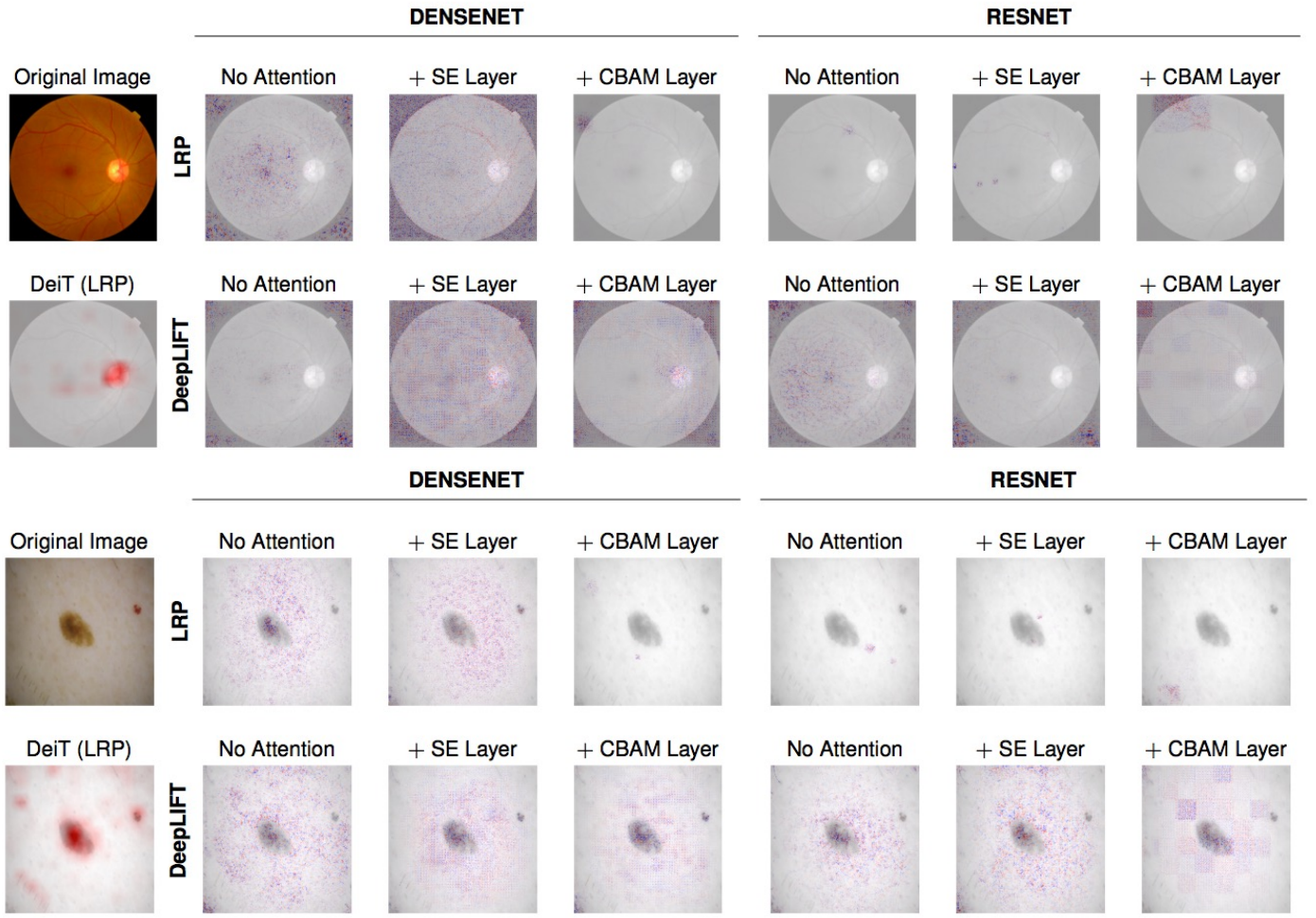


Figure: Examples of results from [1]

Sources: [1] Tiago Gonçalves et al. "A survey on attention mechanisms for medical applications: are we moving towards better algorithms?"

Attention-Driven Medical Image Retrieval

Computer-aided diagnosis through medical image retrieval in radiology[1]

- **Problem:** content-based medical image retrieval of lung conditions in thorax X-ray images
- **Methodology:**
  - Different approaches: **baseline deep neural network**, **attention-based deep neural network**, **interpretability-guided deep neural network**
  - We use **our models to obtain the feature representations** of each image and then a **similarity metric** (e.g., Euclidean distance) is used to get the most similar images

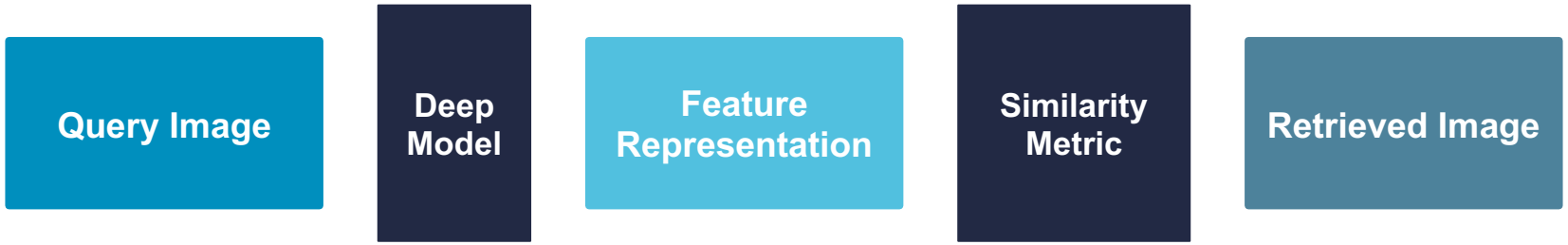
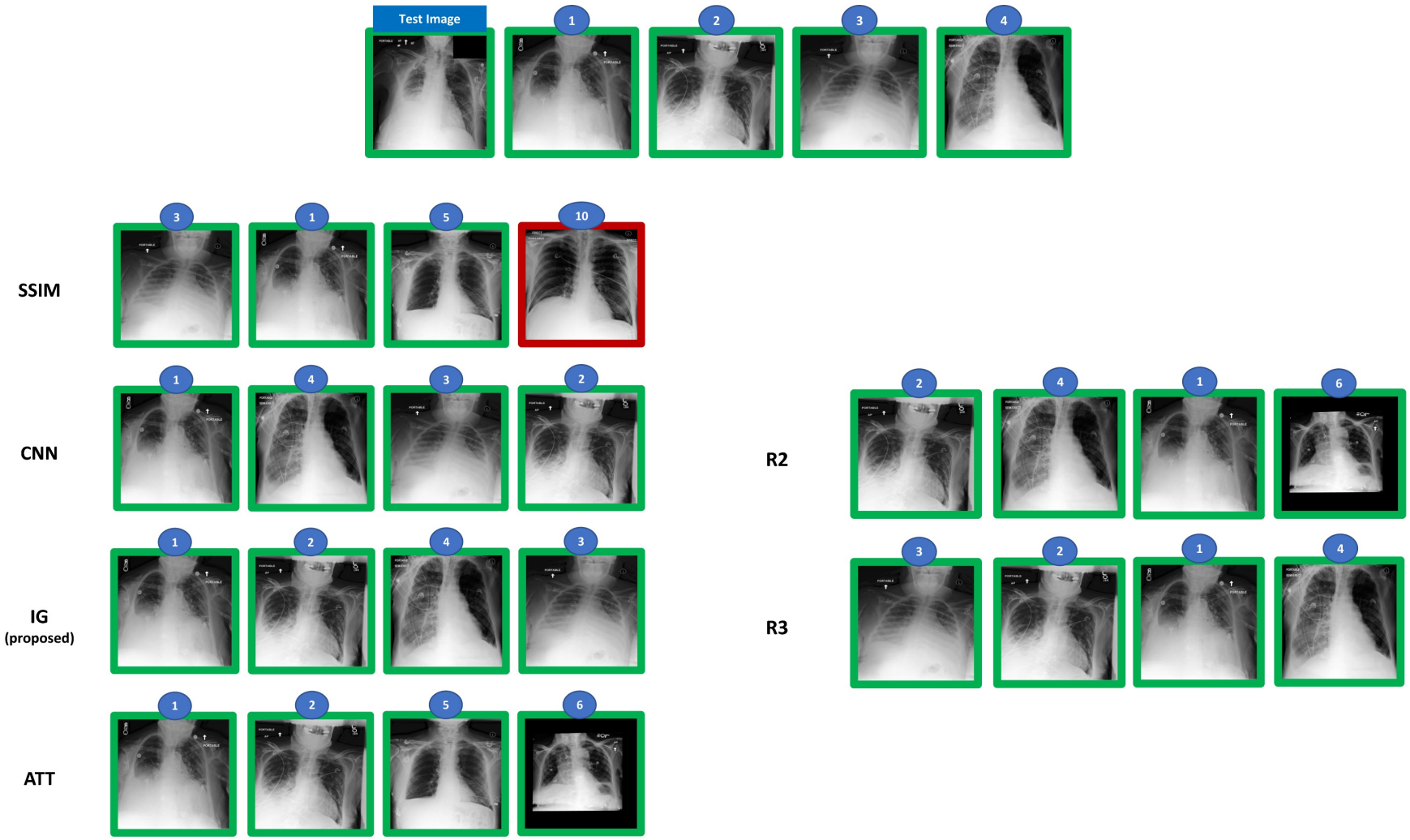


Figure: Medical Image Retrieval Process

Sources: [1] Wilson Silva et al. "Computer-aided diagnosis through medical image retrieval in radiology"

Attention-Driven Medical Image Retrieval



Sources: [1] Wilson Silva et al. "Computer-aided diagnosis through medical image retrieval in radiology"

INESCTEC

3. Multi-modal Data Strategies for Medical Applications

Introduction

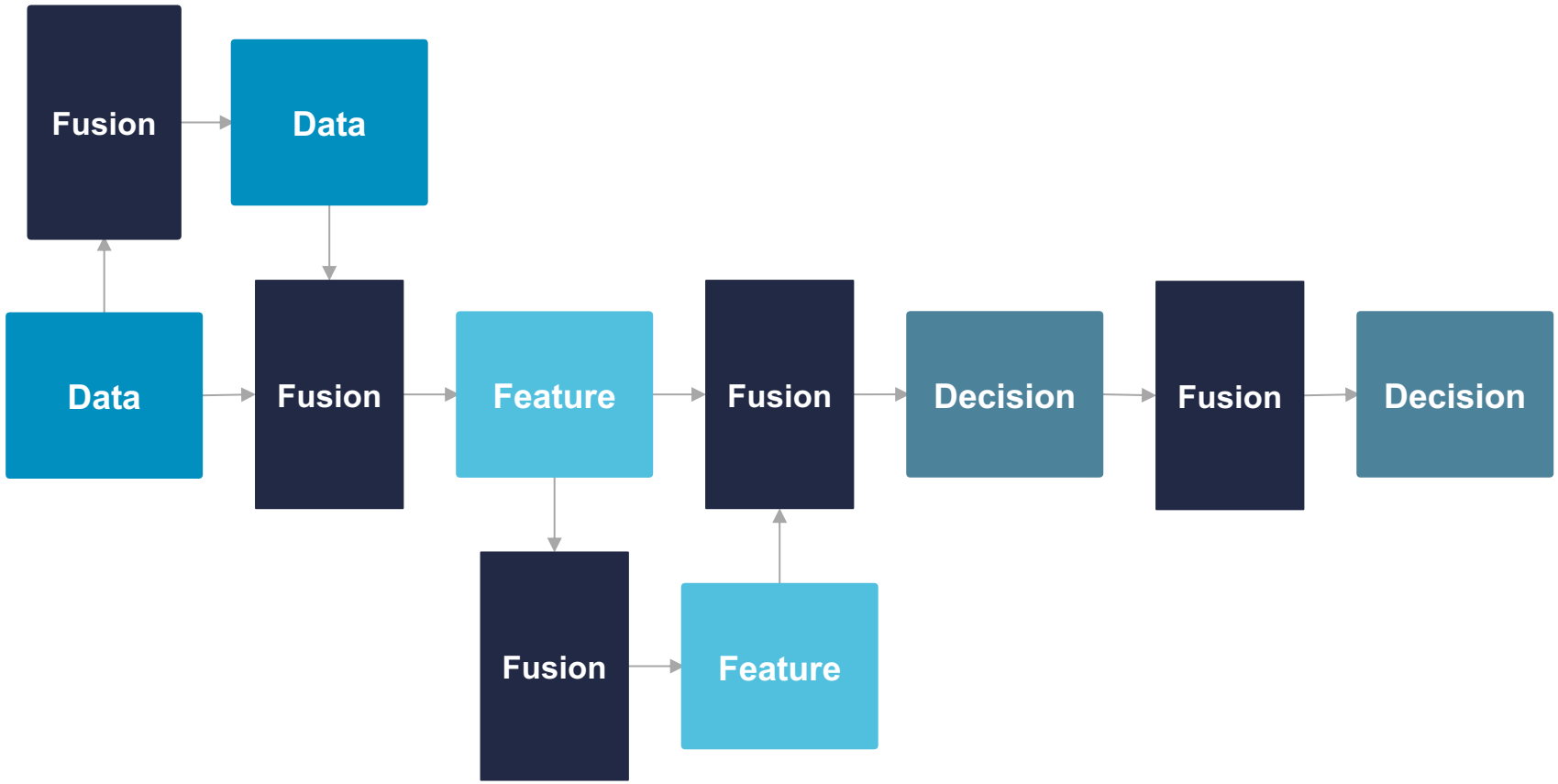


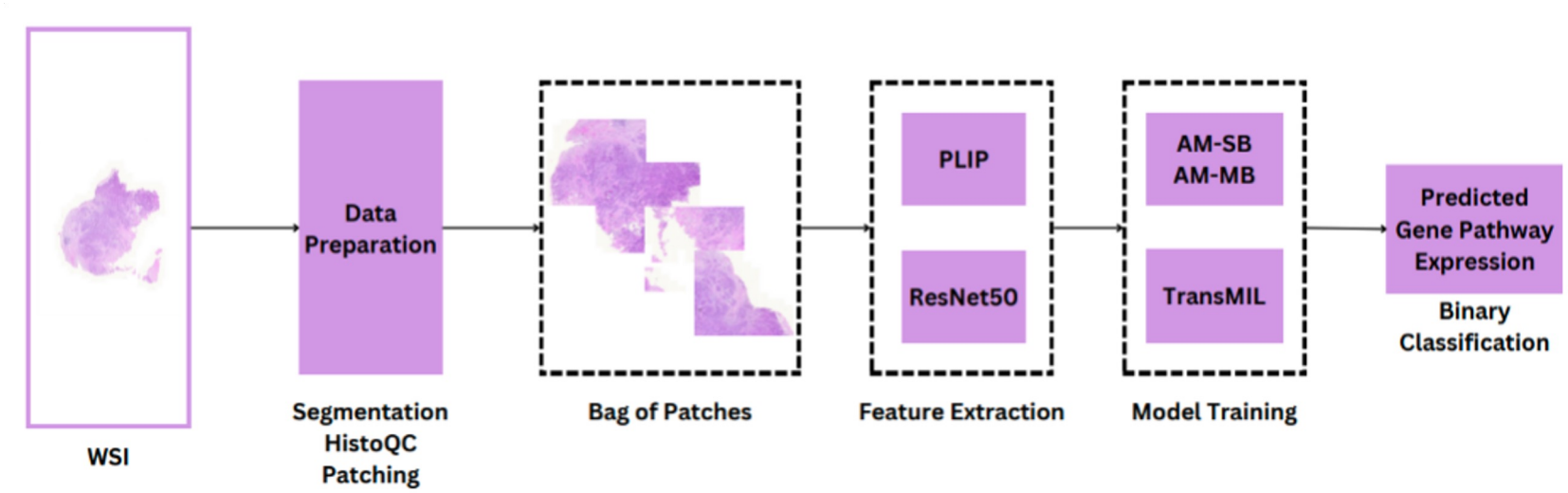
Figure: Different strategies of data fusion[1]

Sources: [1] R. C. Luo et al. "Multisensor fusion and integration: approaches, applications,and future research directions"

Deep Learning-based Prediction of Breast Cancer Tumor and Immune Phenotypes from Histopathology[1]

- **Problem**: can we use deep learning (DL) to predict several facets of the tumour microenvironment (TME) through recognizing biologically relevant features (e.g., tissue architecture, tumour-immune interface) within an individual H&E image?

- **Methodology**:



Sources: [1] Tiago Gonçalves et al. "Deep Learning-based Prediction of Breast Cancer Tumor and Immune Phenotypes from Histopathology"

Deep Learning-based Prediction of Breast Cancer Tumor and Immune Phenotypes from Histopathology[1]

About the Predictive Performance

Table 2: **AUROC results on the test set, obtained for the best model of each feature set.** The best-performing feature-extraction strategy is highlighted in bold. Models trained with PLIP-derived features generally perform better than ResNet-derived features, suggesting that using a feature extractor pre-trained on H&E-related data may improve performance.

Task	Features	
	ResNet50	PLIP
B-cell proliferation	0.7322	**0.7755**
T-cell mediated cytotoxicity	0.7564	**0.7770**
Angiogenesis	0.7053	**0.7435**
Epithelial-mesenchymal transition	**0.8110**	0.8082
Fatty acid metabolism	**0.6323**	0.6030
Glycolysis	0.7996	**0.8330**
Oxidative phosphorylation	0.6926	**0.7332**
Immunosuppression	0.8133	**0.8542**
Antigen processing and presentation	0.7599	**0.7924**
Cell cycle	0.7809	**0.7939**

About the Interpretability

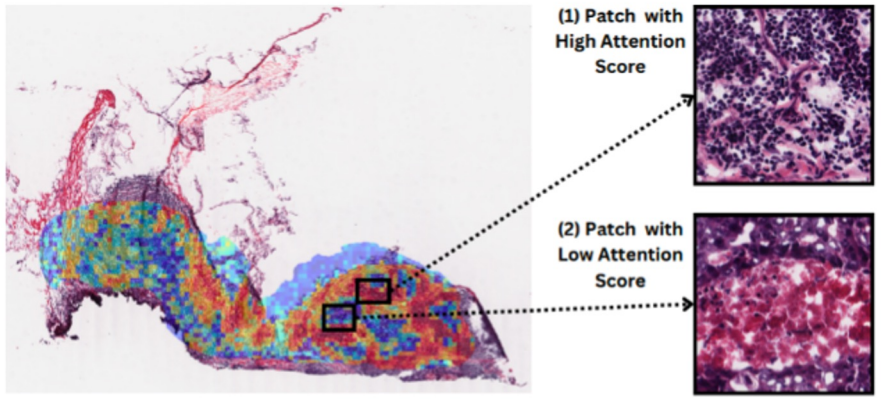
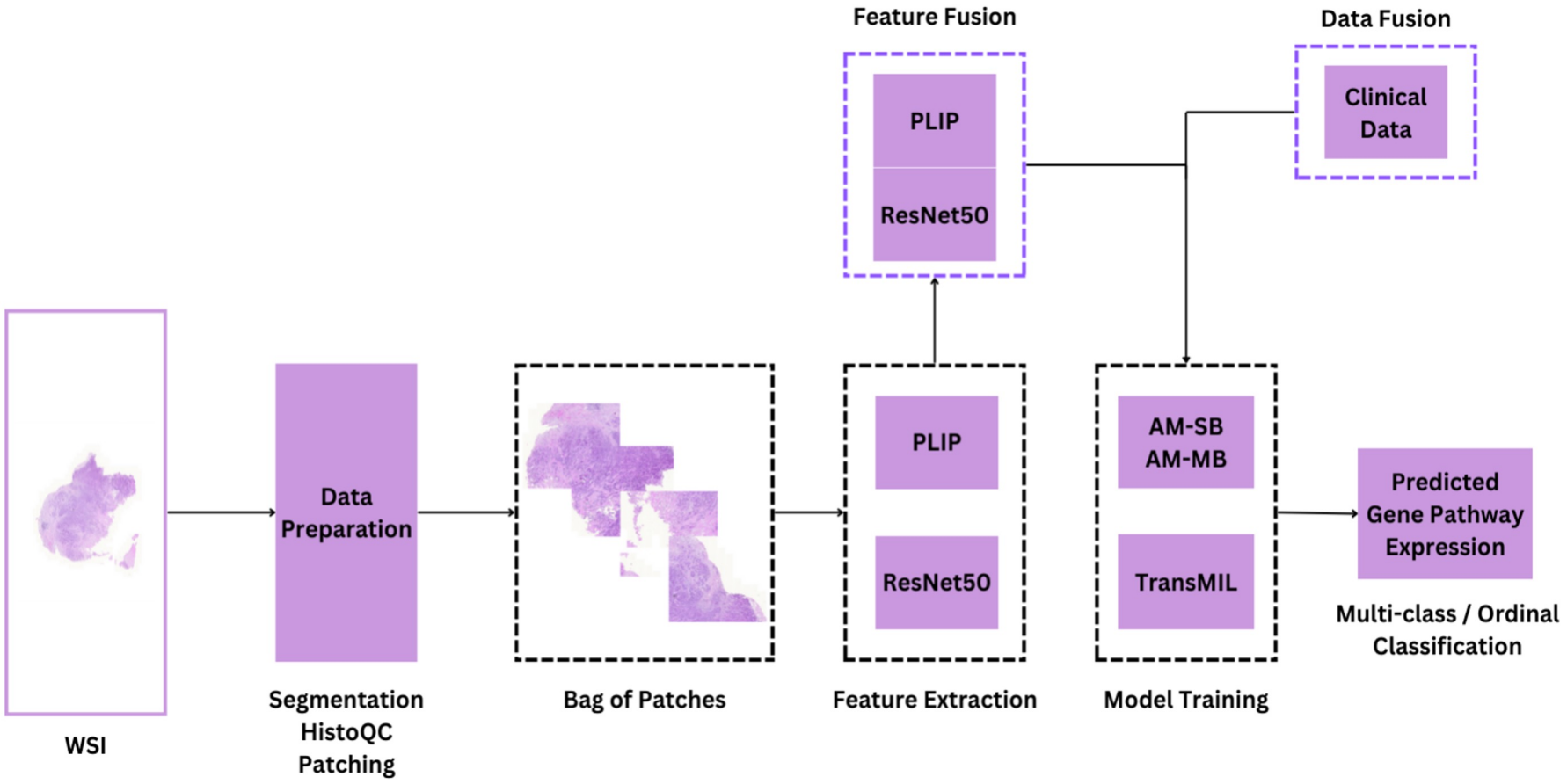


Figure 2: **Attention map obtained using the AM-SB architecture for a WSI predicted to have a high-degree of T-cell mediated cytotoxicity**. Red zones and blue zones indicate high and low attention scores, respectively. On the right, we provide two exemplar patches: 1) the high-attention patch illustrates abundant tumor-infiltrating lymphocytes without tumor cells, which are suggestive of high immune activity, and 2) the low-attention region demonstrated areas of tumor necrosis and minimal lymphocytes, consistent with low immune activity.

Sources: [1] Tiago Gonçalves et al. "Deep Learning-based Prediction of Breast Cancer Tumor and Immune Phenotypes from Histopathology"

Deep Learning-based Prediction of Breast Cancer Tumor and Immune Phenotypes from Histopathology[1]



Sources: [1] Tiago Gonçalves et al. "Deep Learning-based Prediction of Breast Cancer Tumor and Immune Phenotypes from Histopathology"

INESCTEC

4. Intrinsically Interpretable Models in Medical Context

Introduction

- The **black box behaviour of deep learning models does not help decision-makers** to have a **clear understanding of their inner-functioning**, thus preventing them to diagnose errors and potential biases or deciding when and how much to rely on these models[1]

- There has been a huge effort into the **development of post-model strategies** to explain the behaviour of black box models, however, the outputs of these algorithms are prone to **subjective evaluation**, **may be misleading**[2] or **fooled**[1]

- Besides, we agree that just being able to obtain explanations is not enough and that we rather need to take into account at the development stage that these methods must **respect specific constraints** that give them the **capability of generating human-understandable explanations** and make decisions based on such premises[3]

Sources: [1] Dylan Slack et al. "Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods", [2] Zachary C. Lipton "The Mythos of Model Interpretability", [3] Cynthia Rudin "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead"

State of the Art

- An interesting line of work focuses on **learning and optimising scoring systems**[1]: these linear classification models only require users to add, subtract and multiply a few small numbers to make a prediction
  - These models are **difficult to learn from data** because they need to be accurate and sparse, have co-prime integer coefficients, and satisfy multiple operational constraints

- Regarding **case-based explanations in deep learning**, we point to:
  - A deep learning architecture that highlights parts of the image and uses these **prototypes** to provide a score for the probability of the specific diagnosis for this image[2]

  - A deep learning architecture that learns to predict the **concepts**[3] and then uses these concepts to predict a label (i.e., **concept bottleneck definition**)

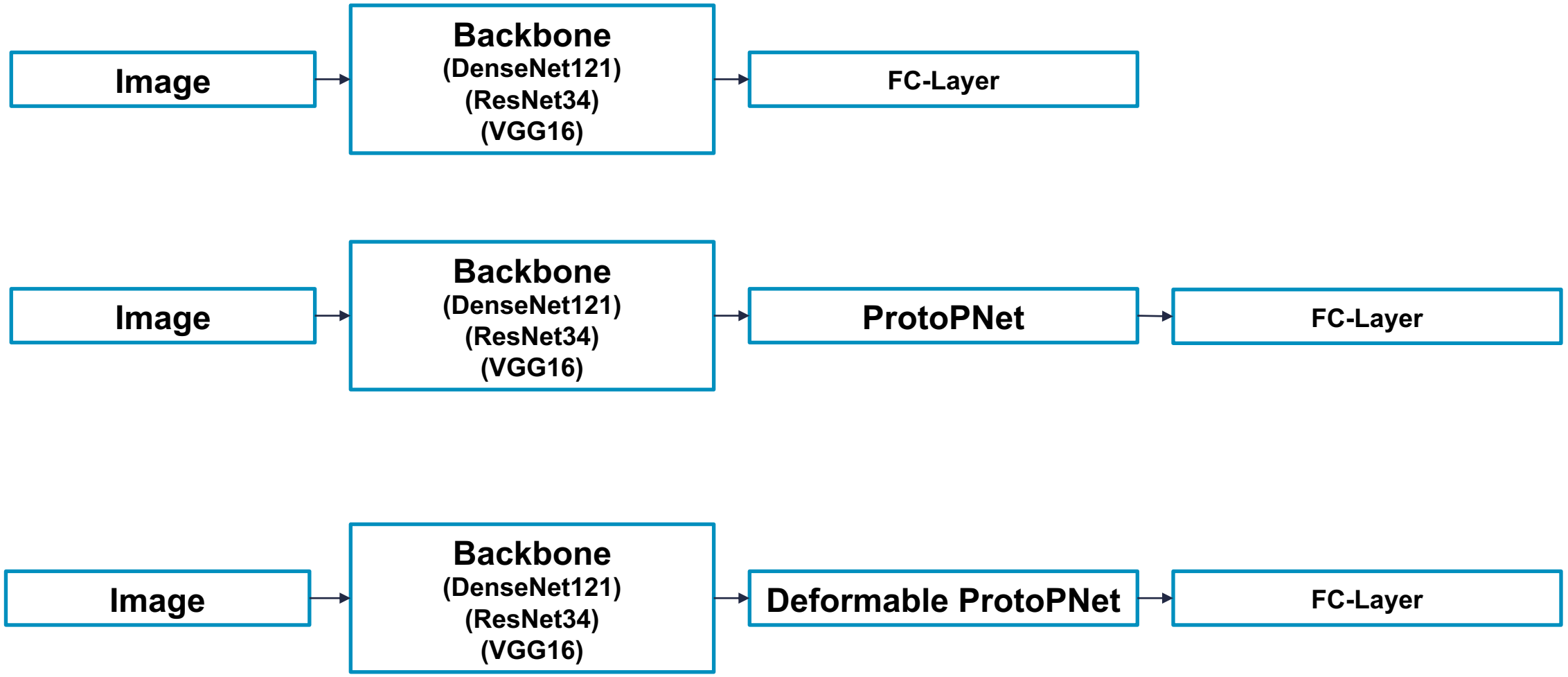
- Newer approaches take into account **causality, ethics and fairness**[4]

Sources: [1] B. Ustun and C. Rudin "Supersparse Linear Integer Models for Optimized Medical Scoring Systems", [2] Barnett et al. "IAIA-BL: A Case-based Interpretable Deep Learning Model for Classification of Mass Lesions in Digital Mammography", [3] Chen et al. "This Looks Like That: Deep Learning for Interpretable Image Recognition", [4] T. Wang and C. Rudin "Causal Rule Sets for Identifying Subgroups with Enhanced Treatment Effect"

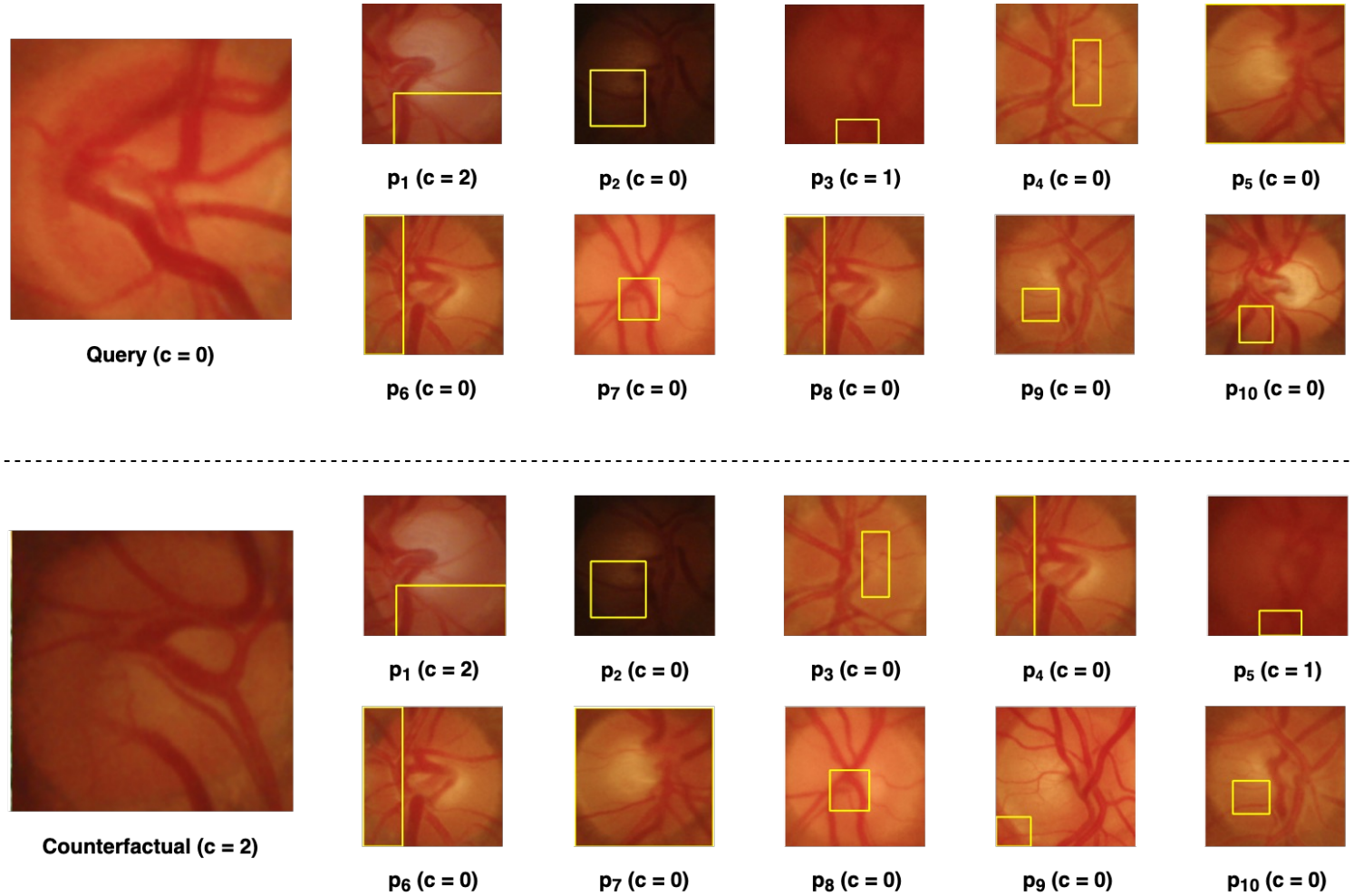
Explaining Counterfactuals with Prototypes

- **Problem**: generating counterfactual explanations using prototypical learning networks
- **Methodology**:
  - Train different neural networks with a prototype layer, ensuring that results are independent from the backbone
  - Use different datasets, to ensure results are independent from the data
  - Decide on the best approach to get the counterfactual label
  - Choose a proper latent space to perform the retrieval of the counterfactual examples
  - Use the learned prototypes to explain the decision

Explaining Counterfactuals with Prototypes



Explaining Counterfactuals with Prototypes



6. Conclusion

Responsible AI relies on fundamental principles

- **Responsible AI** is a framework that guides how we should address the challenges around artificial intelligence from both an **ethical, technical and legal** point of view[1]
  - We must resolve ambiguity for where responsibility lies if something goes wrong!

- This framework relies on fundamental principles[2]:
  - **Accountability**
  - **Interpretability**
  - **Fairness**
  - **Safety**
  - **Privacy**

Figure - Responsible AI (Image from [1])

Sources: [1] https://searchenterpriseai.techtarget.com/definition/responsible-AI, [2] https://towardsdatascience.com/what-is-responsible-ai-548743369729

We need to design ethical and fair algorithms

- To facilitate trust (and increase transparency) in AI algorithms it is important to **ensure a priori that these models are interpretable**, and understand how decisions are made in the clinical context

- On the other hand, it is important to understand what the algorithms are already learning and to **evaluate the quality of such explanations** (e.g., understand if the algorithms are extracting relevant features for the clinical context)

- A different dimension of the application of AI in sensitive domains such as healthcare is the **development of ethical and fair algorithms**[1]
  - This strategy is supported by the new European Union's General Data Protection Regulation (EU-GDPR)[2] which advises that these **algorithms should be able to explain their decisions for the sake of transparency**

Sources: [1] Irene Y. Chen et al. "Treating health disparities with artificial intelligence", [2] Margot E. Kaminski "The Right to Explanation, Explained"

While keeping an attentive eye on the technologies that are shaping our lives

- Many entities are already leveraging their data sources to **optimize their inner processes or to develop new services or products**, thus enabling them to achieve a substantial competitive advantage[1]

- In the healthcare context, systems and algorithms need to go through a continuous **pipeline of validation and error assessment**

- Hence, it is **reasonable to accept that these technologies may need to be calibrated** to the data sources of the institutions that are integrating them into their information systems and that these algorithms **may have a continuous learning policy over time**

- Moreover, to assure **transparency, accountability and accessibility**, new regulatory frameworks will have to be developed to allow model adaptations that enable optimal performance while **ensuring reliability and patient safety**[2]

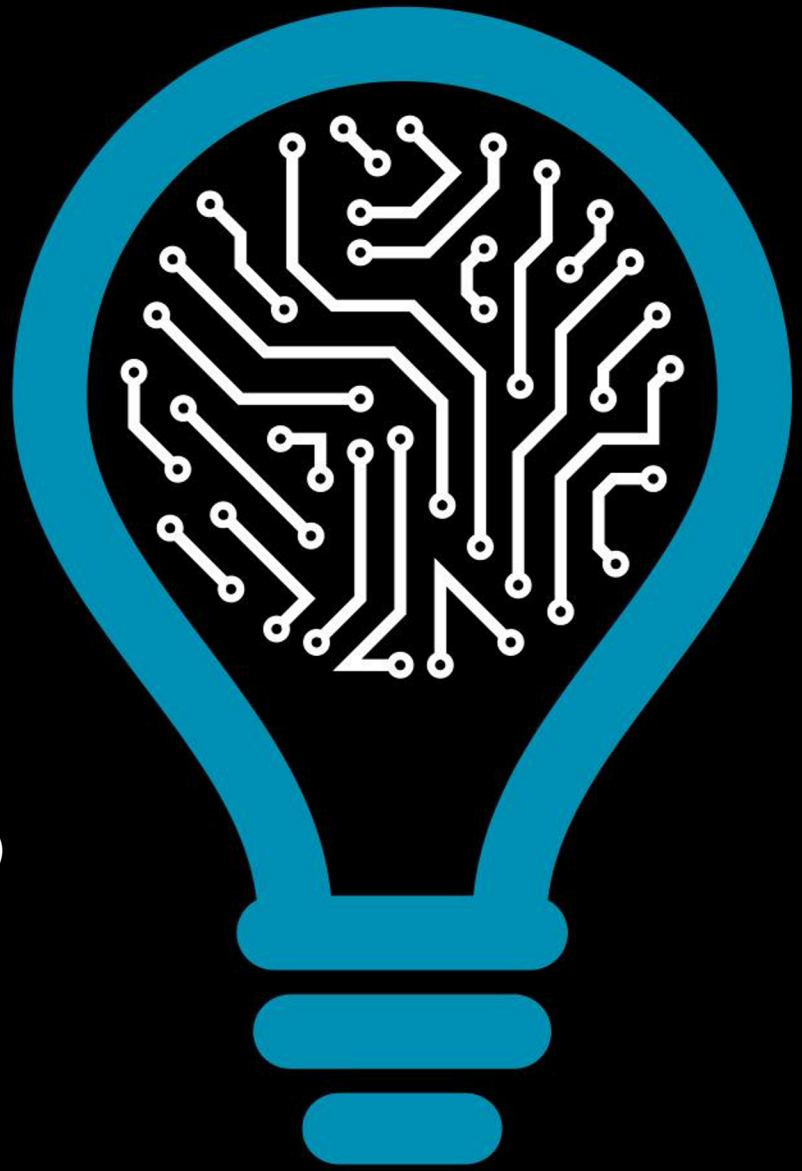
Sources: [1] Lucas Baier et al. "Challenges in the deployment and operation of machine learning in practice",
[2] Farhad Maleki et al. "Machine Learning Algorithm Validation: From Essentials to Advanced Applications and Implications for Regulatory Certification and Deployment"

Interpretable Machine Learning and its Application to Medical Decision Support Systems

PhD Research Work | Supervisor: Jaime S. Cardoso

Group Meeting | AI Technology for Life | May 1, 2024
University of Utrecht, Netherlands

Tiago Filipe Sousa Gonçalves (tiago.f.goncalves@inesctec.pt)



INESCTEC
INSTITUTE FOR SYSTEMS AND COMPUTER ENGINEERING, TECHNOLOGY AND SCIENCE