MEMBER OF
BASQUE RESEARCH
& TECHNOLOGY ALLIANCE

Universidad de Navarra

Explainability in Deep Convolutional Neural Networks and Visual Interpretability

Diego Borro

Computer Science PhD

Vision and Robotics research line at CEIT

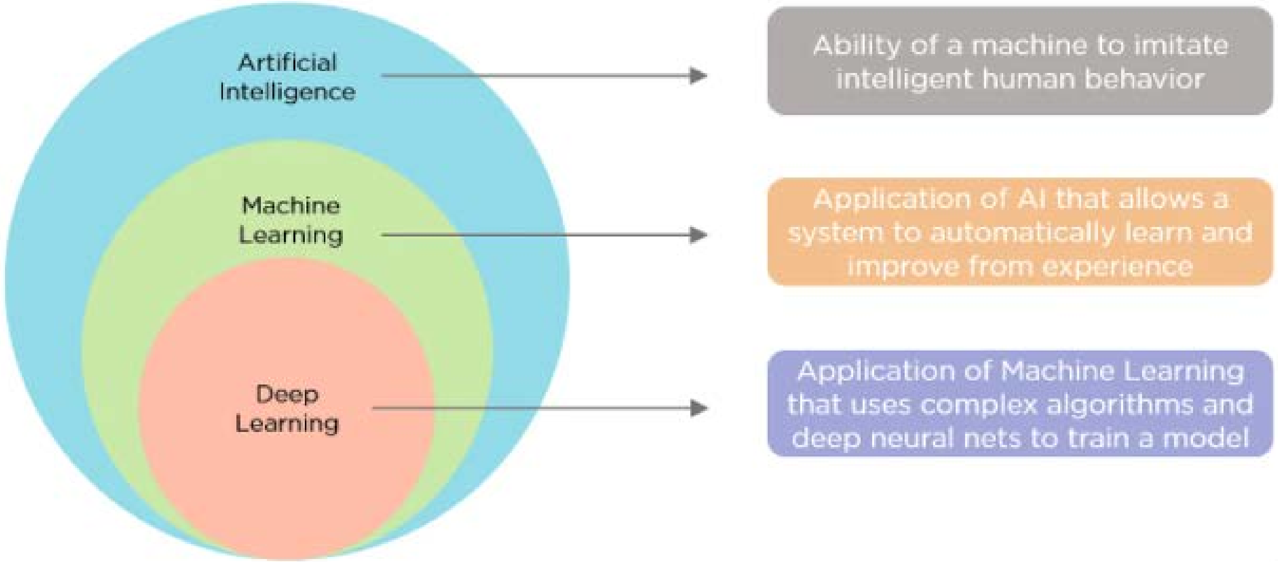
Research Professor at TECNUN

dborro@ceit.es

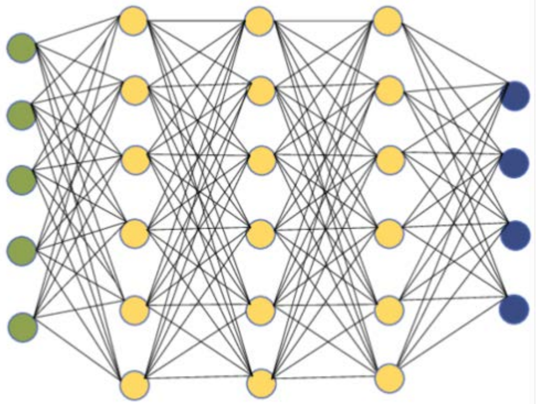
Index

1. Context and introduction
2. Explainable AI (XAI) and techniques
3. Real use cases
4. Conclusions and discussion

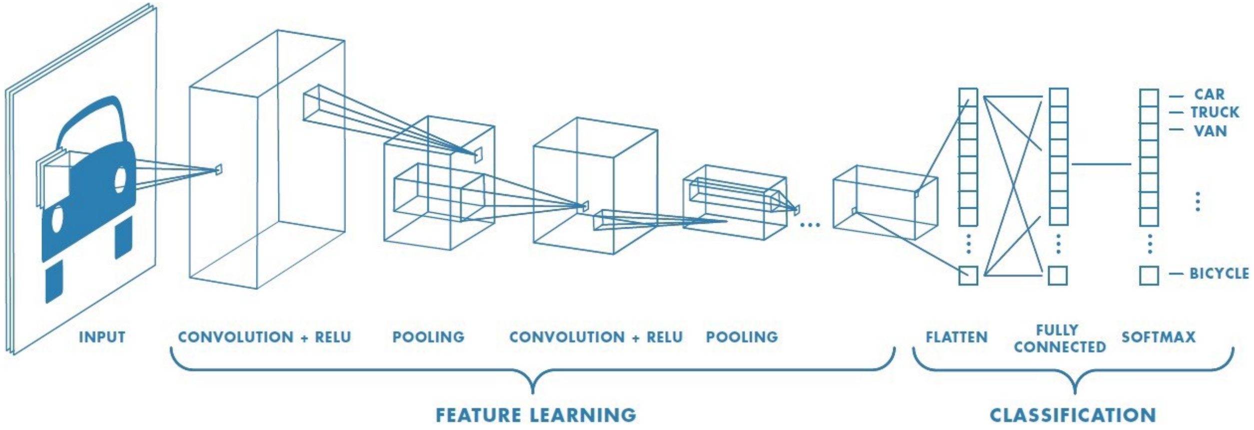
Deep Learning models



Dense Neural Networks

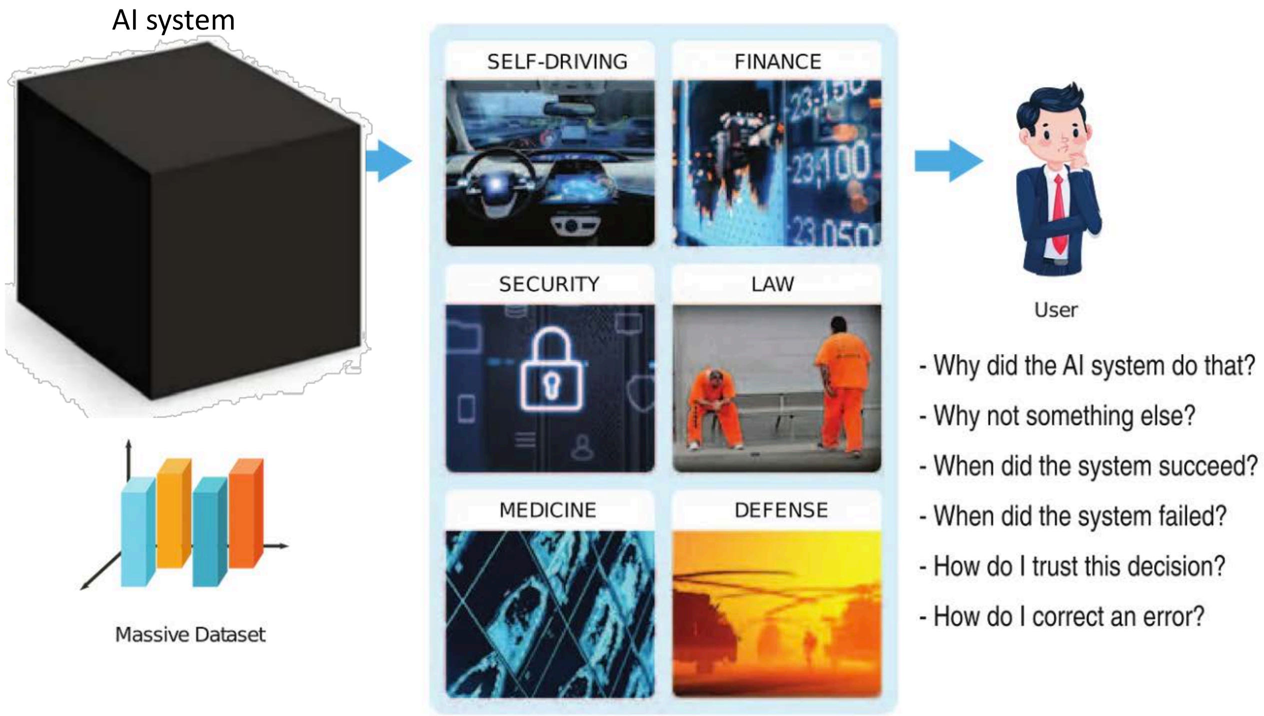


Convolutional Neural Networks

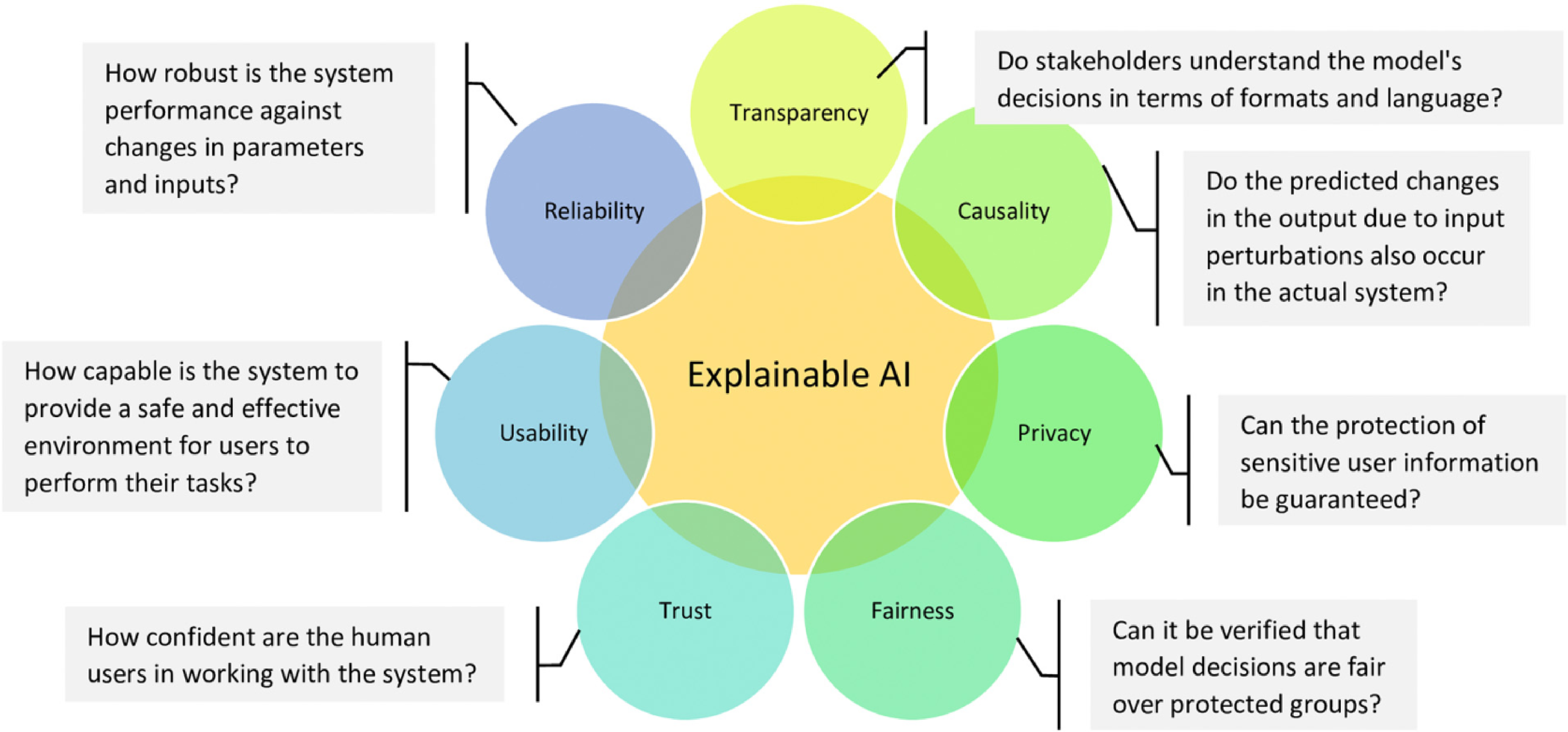


Deep Learning models = Black box models

- Deep learning models are far more complex to interpret than most machine learning models (opaque nature)
  - Many layers and parameters
  - Multiple types of non-linear activation functions
  - No well-defined criteria for choosing an architecture and hiperparameters (trial and error process)
  - Learning and reasoning are embedded in the behavior of thousands of simulated neurons, arranged in hundreds of interconnected layers
  - “Perfect” matching input-output but no direct evidence how



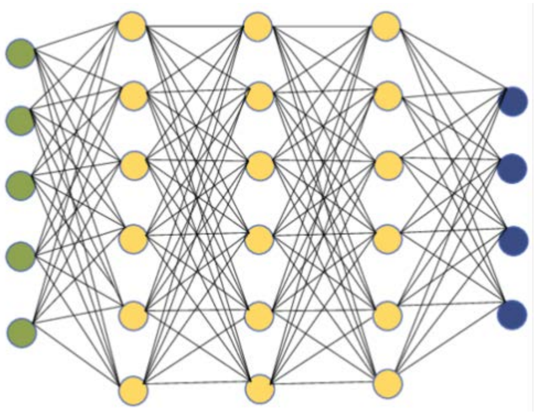
Goals of XAI


How robust is the system performance against changes in parameters and inputs?
Transparency
Do stakeholders understand the model's decisions in terms of formats and language?
Reliability
Causality
Do the predicted changes in the output due to input perturbations also occur in the actual system?
How capable is the system to provide a safe and effective environment for users to perform their tasks?
Explainable AI
Usability
Privacy
Can the protection of sensitive user information be guaranteed?
Trust
Fairness
How confident are the human users in working with the system?
Can it be verified that model decisions are fair over protected groups?

eXplainable AI

- **Explainable Artificial Intelligence (XAI)** is a concept that **explains decisions** made by machine learning models and provides justification in a **way interpretable by humans** [1]
- XAI are **tools to visualize and understand** how a complex model is making decisions, which can help "**explain**" these decisions in more intuitive terms

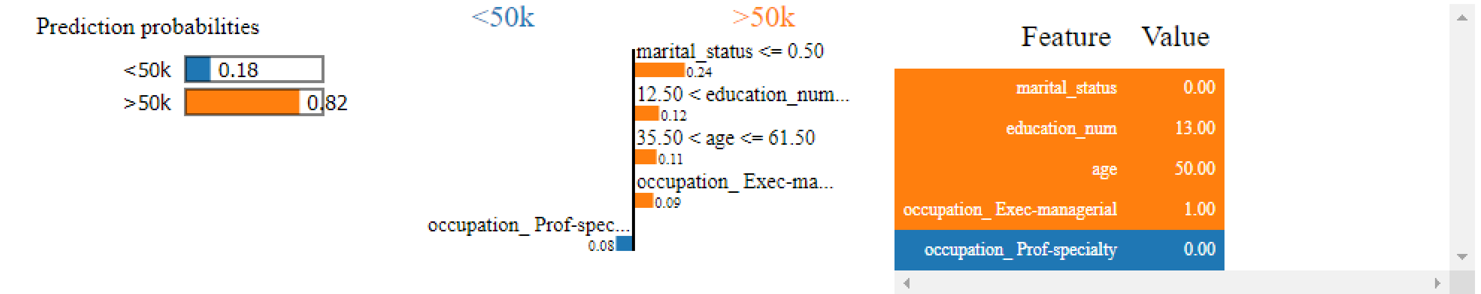
In a FNN (fully-connected neural network), neurons learn representation and patterns that is difficult to extract and present in a human-readable form



- LIME (Local Interpretable Model-Agnostic Explanations) [2]
- SHAP (SHapley Additive exPlanations) [3]

They try to understand the importance of features by seeing how predictions change when input features are perturbed, removed of changed (Bias detection!!)

```
marital_status                    Married
education_num                   Bachelors
hours_per_week                         40
fnlwgt                             167065
sex                                  Male
age                                    50
random                          0.0412146
workclass_ Private                      1
occupation_ Exec-managerial             1
race_ White                             1
Name: 23706, dtype: object
```

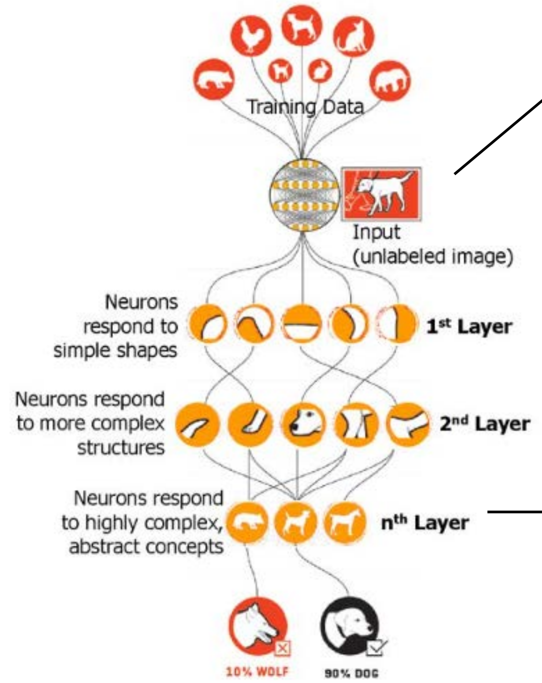


[1] S. Ali, et al., "Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence," Information Fusion, p. 101805, 2023.
[2] Ribeiro, MT., et al., "Why Should I Trust You?": Explaining the Predictions of Any Classifier", ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
[3] SM Lundberg, SI Lee. "A unified approach to interpreting model predictions", Advances in neural information processing systems, 2017.

eXplainable AI

- **Explainable Artificial Intelligence (XAI)** is a concept that **explains decisions** made by machine learning models and provides justification in a **way interpretable by humans** [1]
- XAI are **tools to visualize and understand** how a complex model is making decisions, which can help "**explain**" these decisions in more intuitive terms

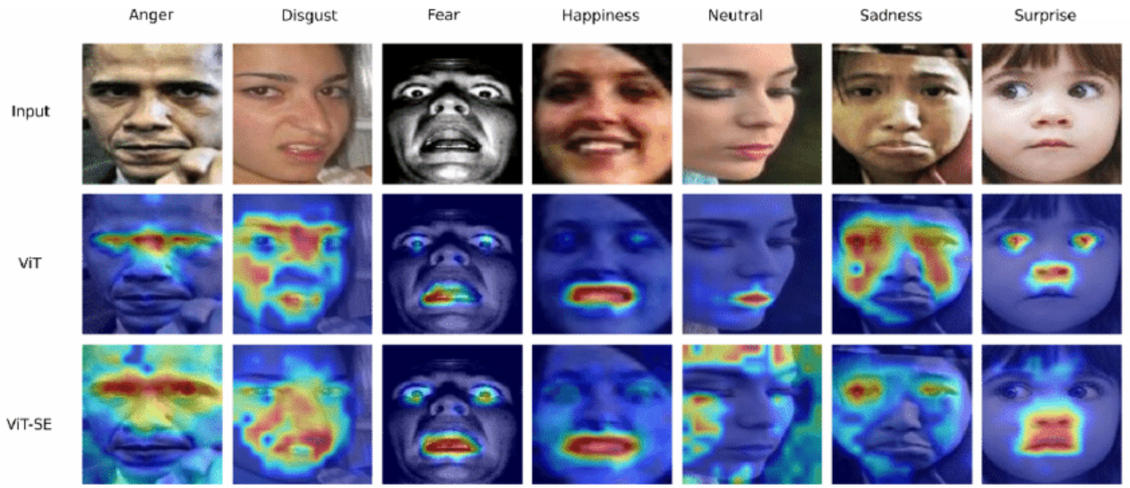
CNN are focused on image processing problems so, those patterns are images!!



Convolutional layers naturally retain the spatial information of the input data

Shapes and patterns are detecting at successive layers

Deeper representation in a CNN capture high-level abstracts or visual concepts



[4]

[4] M. Aouayeb, et. al. "Learning Vision Transformer with Squeeze and Excitation for Facial Expression Recognition", ArXiv, 2021.

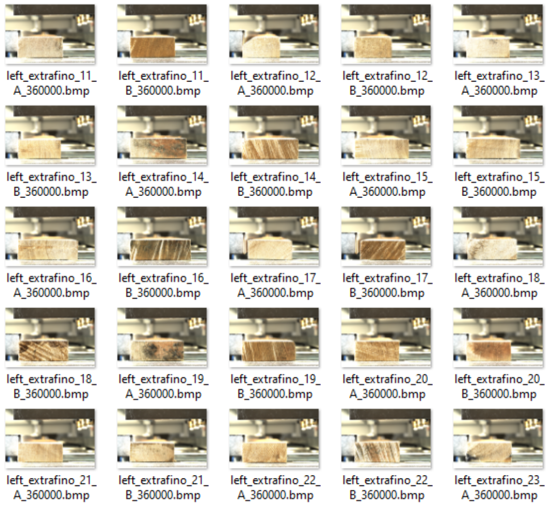
Some XAI techniques for CNNs

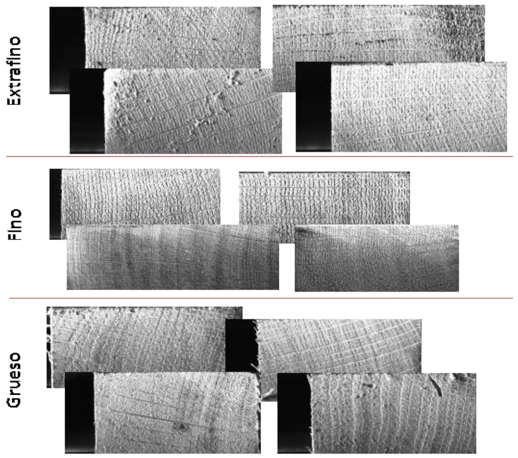
Name	Focus	Eq
Layer visualization	Last convolutional layer	$\sum_{i=1}^{n} FM_i$
Saliency maps [5]	Impact in the output respected to input changes (pixels)	$\nabla(L,x) = \frac{\partial L(y,\hat{y})}{\partial x}$ $L(y,\hat{y}) = -\sum_i y_i \log(\hat{y}_i)$
Grad-CAM [6]	Impact in the output respected to FM changes (high-level features)	$\mathcal{L}^c_{Grad-CAM} = ReLU(\sum_k \alpha^c_k \cdot A^k)$ $\alpha^c_k = \frac{1}{Z} \cdot \sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}}$
Attention maps [7]	Image areas where the model pays attention	$Attention(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$
Guided Backpropagation [8]	Impact in the output respected to positive input changes	$\nabla(L,x) = \left\lvert\frac{\partial L(y,\hat{y})}{\partial x}\right\rvert$ $L(y,\hat{y}) = -\sum_i y_i log(\hat{y}_i)$
Integrated Gradients [9]	Impact in the output respected to changes in N inputs (pixels)	$IG = \sum_{\alpha=0}^{1} \nabla(L,x)_\alpha$

[5] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," arXiv preprint arXiv:1312.6034, 2013.
[6] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that?" Nov. 2016.
[7] Alexey Dosovitskiy y col. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". En: CoRR abs/2010.11929 (2020). arXiv: 2010.11929. url: https://arxiv.org/ abs/2010.11929.
[8] J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: the all convolutional net", Proceedings of the International Conference on Learning Representations (ICLR 2015).
[9] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks", Proceedings of the 34th International Conference on Machine Learning (ICML'17), Vol. 70, pp. 3319-3328. August 2017.

Solving wood heterogeneous texture classification: A deep learning approach with cropping data augmentation

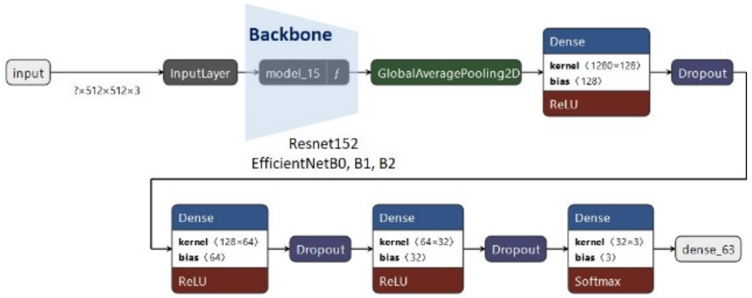
Data heterogeneity



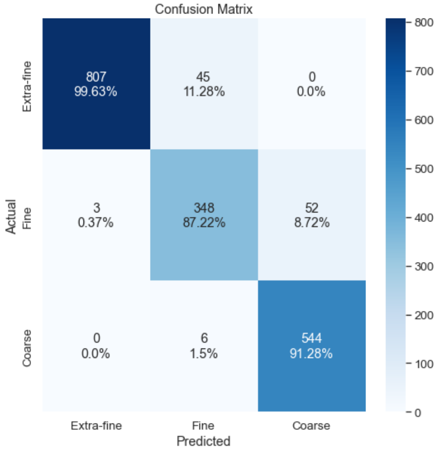


Deep Learning for crops classification

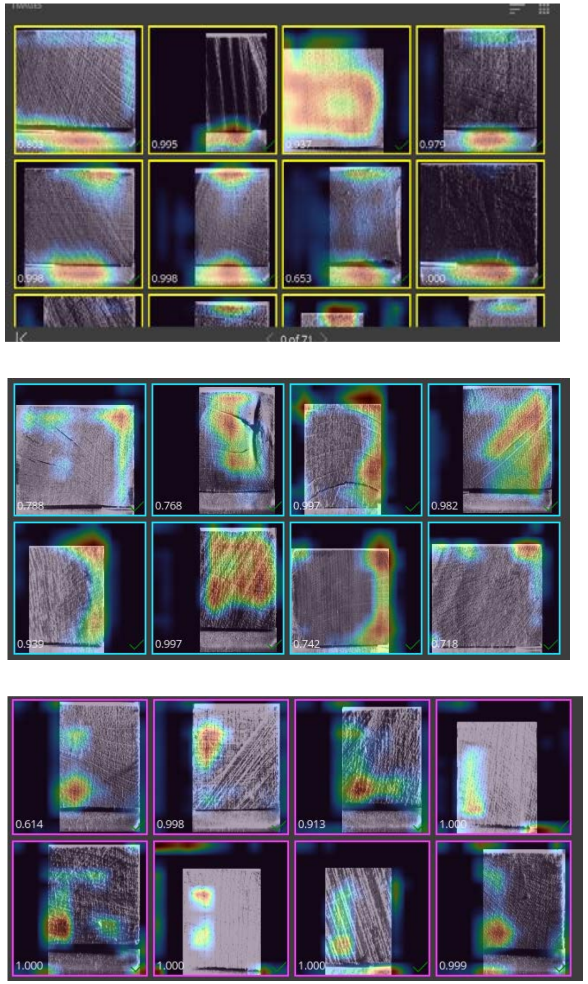
Model architecture



Confusion Matrix of classification

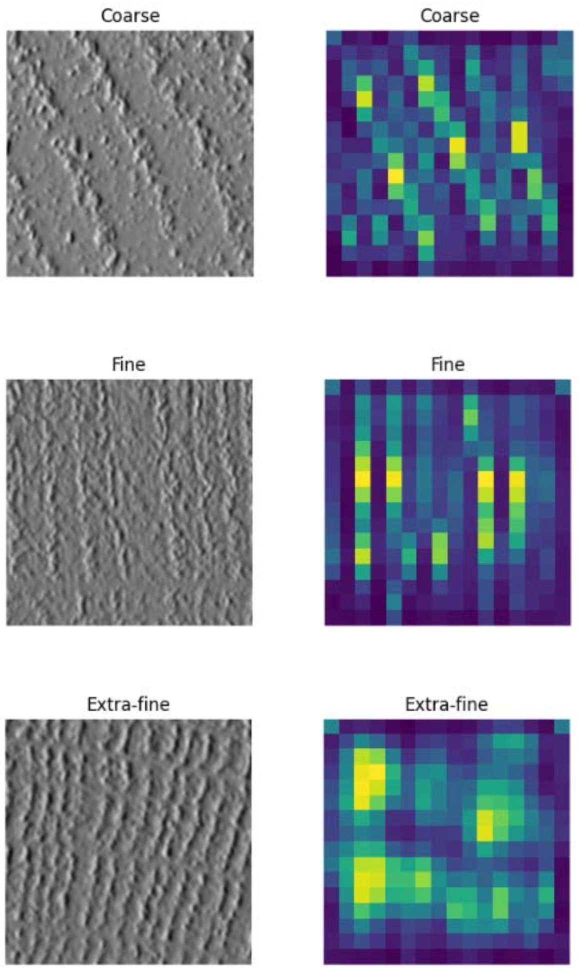


Heat maps

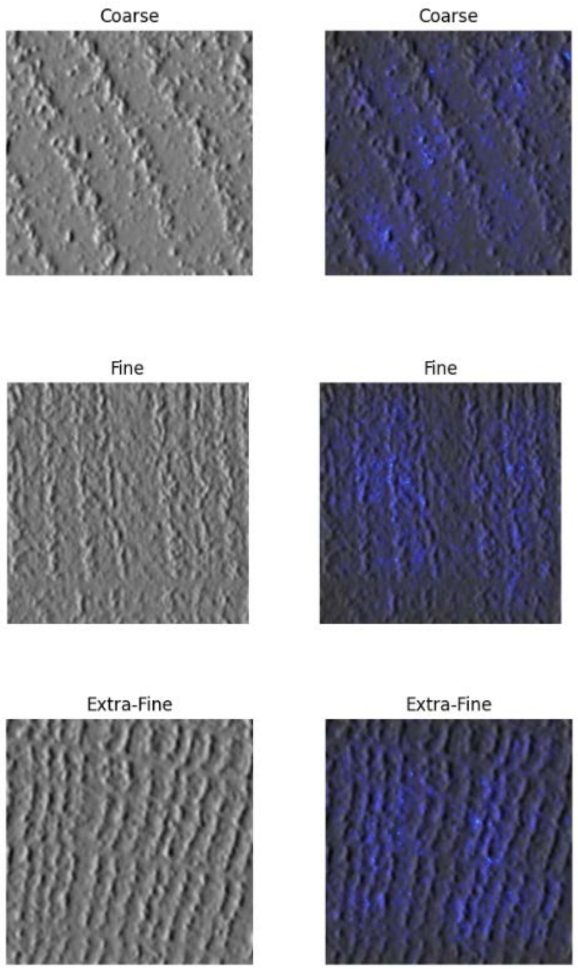


Solving wood heterogeneous texture classification: A deep learning approach with cropping data augmentation

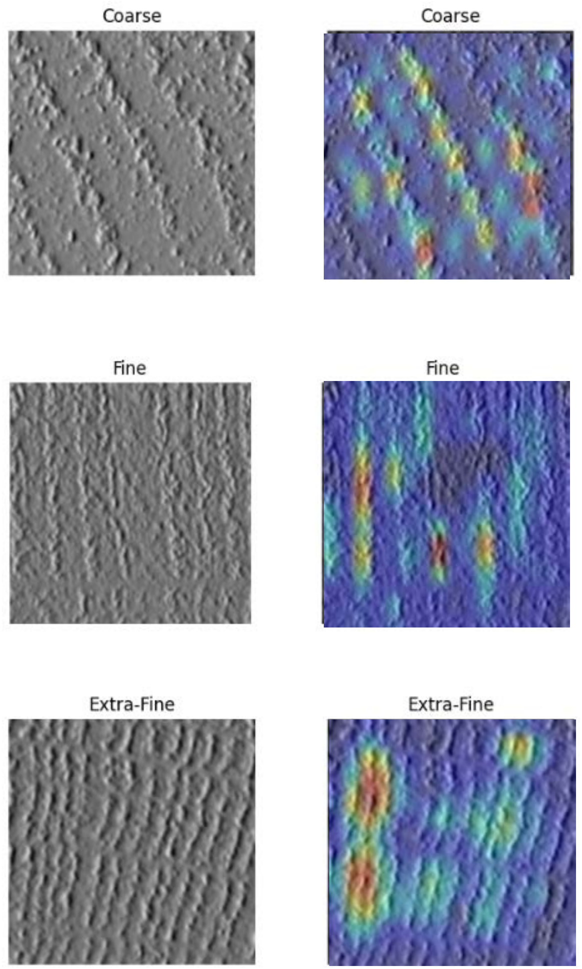
Layer visualization



Saliency maps



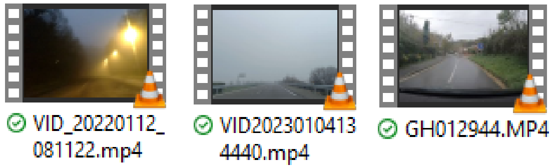
Grad-CAM



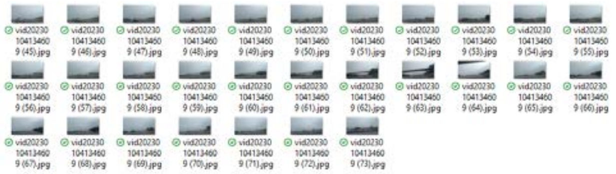
Classifying road fog scenes: A deep learning approach with data imbalance and complex images

Data acquisition and labelling

Original driving videos:



Frame extraction:

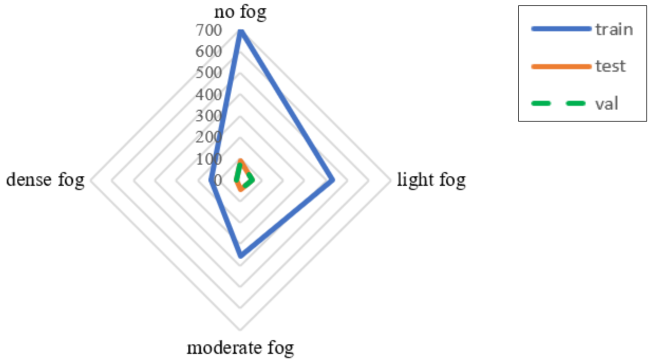


Manually label each frame:

- dense-fog
- moderate-fog
- ligh-fog
- no-fog

Dataset split, augmentation and Deep Learning

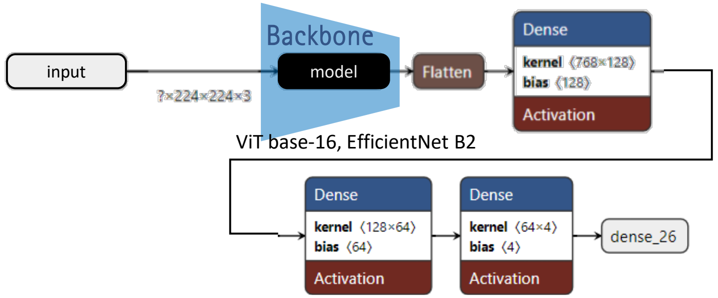
K-fold cross-validation to manage imbalance when splitting the dataset:



Dataset augmentation:

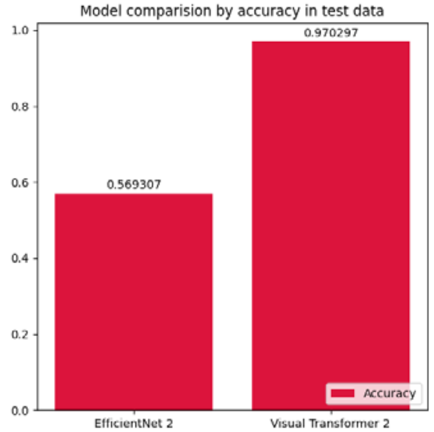
```
ImageDataGenerator(rescale= 1.0/255,        rotation_range=20,
                   horizontal_flip=True,    vertical_flip=True,
                   zoom_range=0.1,          fill_mode='reflect')
```

Model architecture:

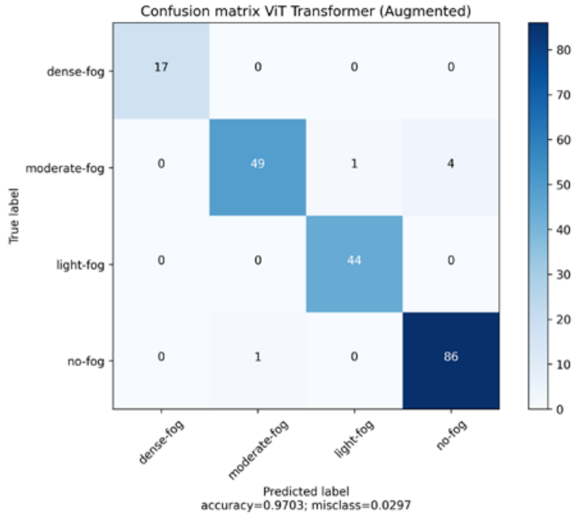


Different fog level classification

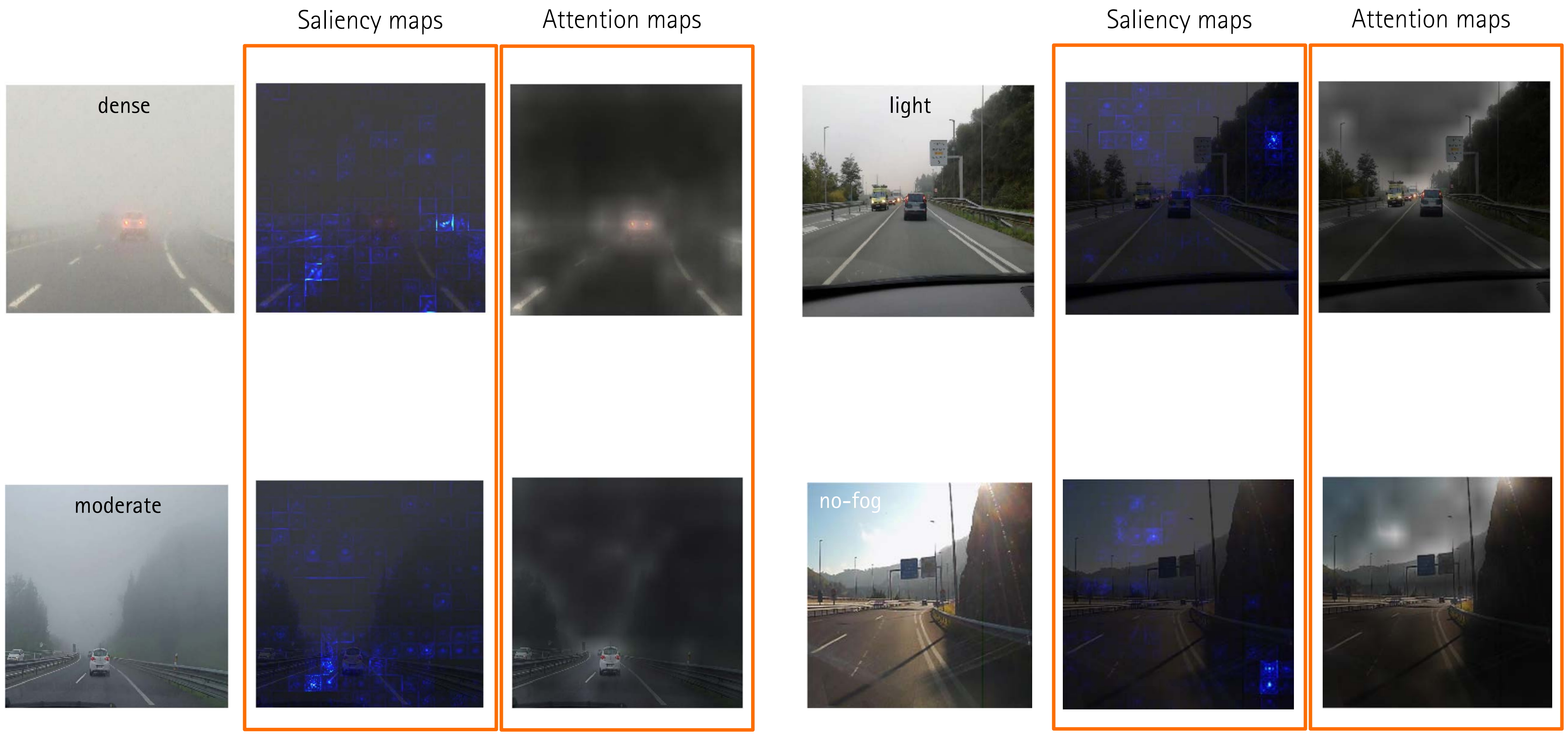
Models accuracy



Confusion Matrix with target images

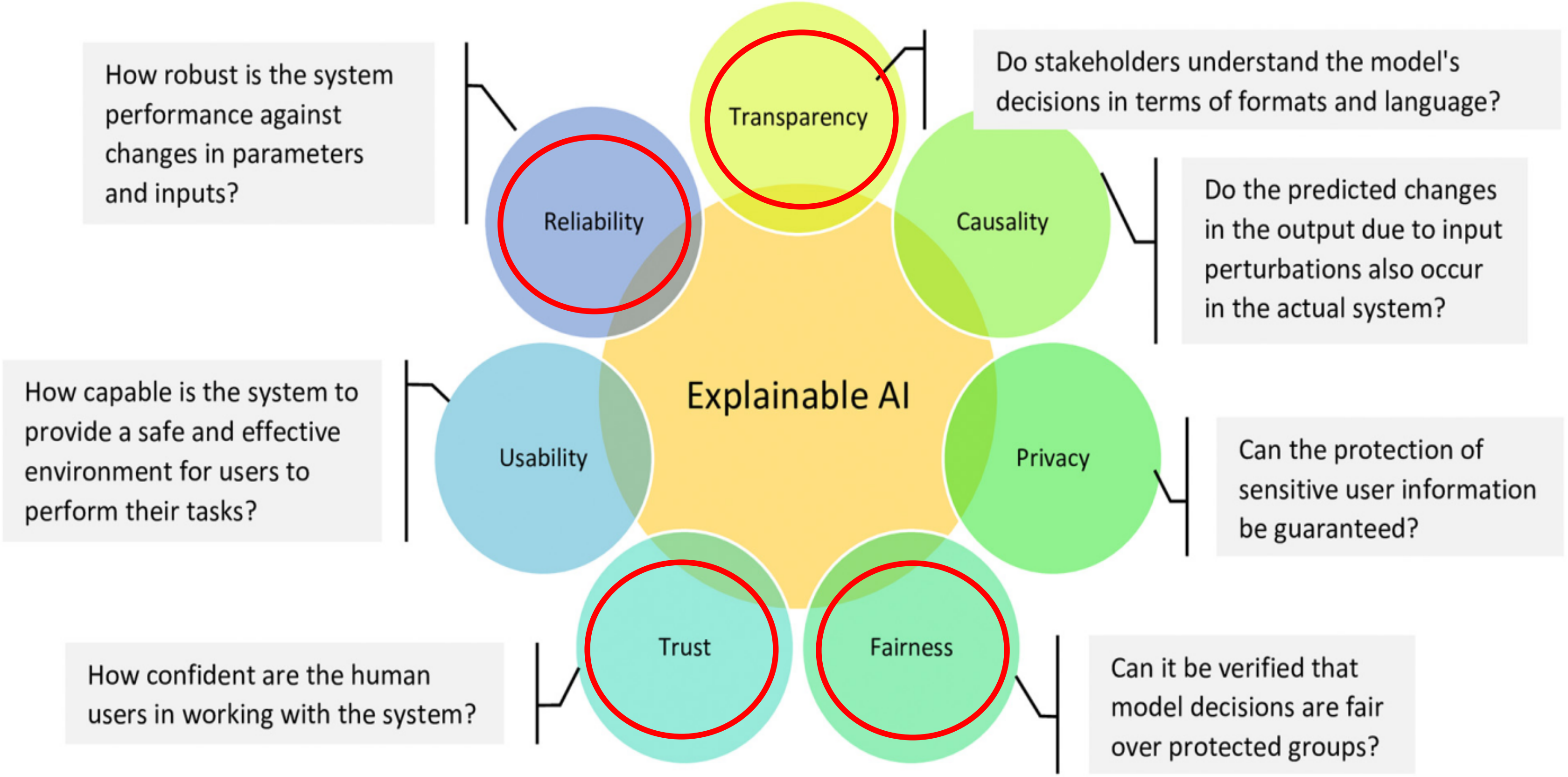


Classifying road fog scenes: A deep learning approach with data imbalance and complex images



Conclussions and discussion

- XAI for a **better understanding AI**



Conclussions and discussion

- XAI for a **better understanding AI**

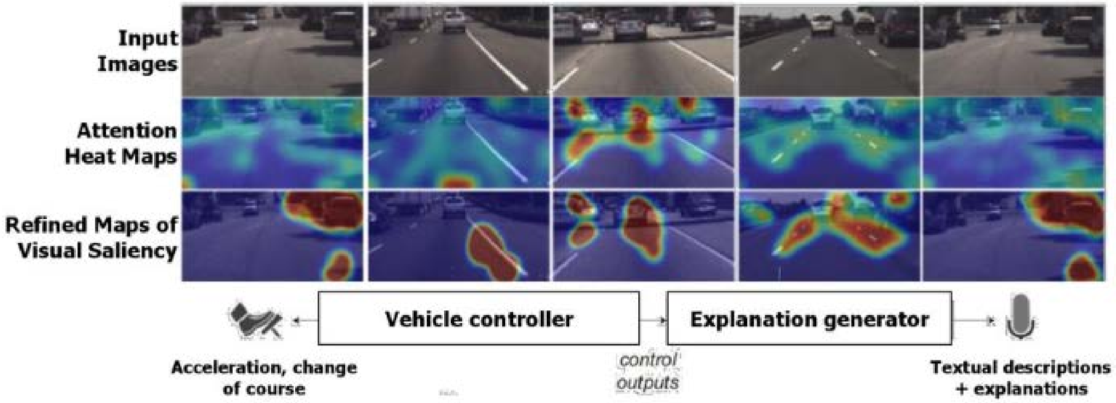
- **Not a general** XAI solution (like metrics)

- Depending on how XAI explanation is provided:
  - **Visual interpretability** methods: visual explanations and plots
  - **Textual explanations**, given in text form
  - **Mathematical or numerical explanations**

- XAI basis for **future authorities**?

UC Berkeley

Textual justification system embedded into refined visual attention models to provide appropriate explanation of the behavior of a deep neural vehicle controller



Without explanation: *"The car heads down the street"*

With explanation: *"The car heads down the street because there are no other cars in its lane and there are no red lights or stop signs"*

Examples of Action Description and Justification

Action Description	Action Justification
The car accelerates	**because** the light has turned green
The car accelerates slowly	**because** the light has turned green and traffic is flowing
The car is driving forward	**as** traffic flows freely
The car merges into the left lane	**to** get around a slower car in front of it

- Refined heat maps produce more succinct visual explanations and more accurately expose the network's behavior
- Textual action description and justification provides an easy-to-interpret system for self-driving cars

Kim, Rohrbach, Darrell, Canny, and Akata. Show, Attend, Control, and Justify: Interpretable Learning for Self-Driving Cars
Kim and Canny. Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention