

Abstract geometric lines in the top left corner, consisting of several overlapping, irregular polygons and lines that create a complex, layered effect.

# EXPLAINABLE AI

helps characterize model accuracy, fairness, transparency and outcomes in AI-powered decision making

# AGENDA

Introduction

BLACKBOX-AI vs Explainable-AI

DeepLearning approach with XAI

Future based Research

# WHAT IS XAI (EXPLAINABLE AI) ?

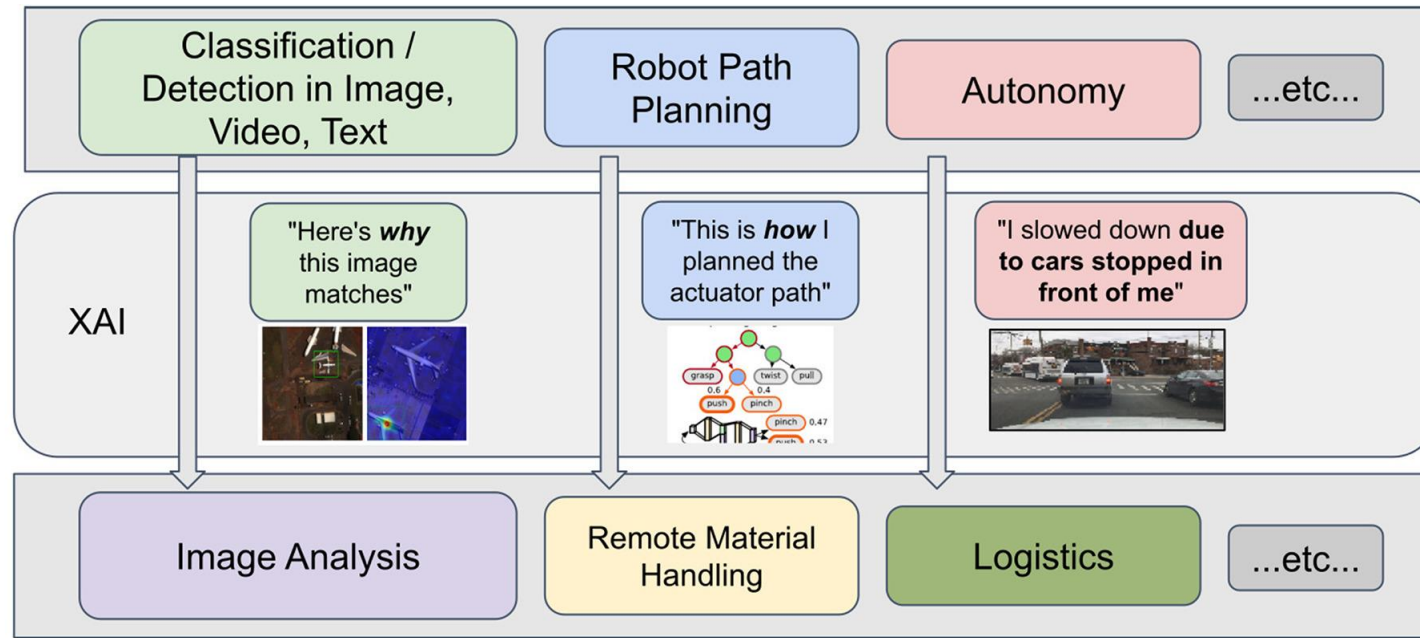


- ✓ Explainable AI, also known as XAI, is a subfield of artificial intelligence that focuses on the development of AI systems that can provide understandable and interpretable explanations for their actions and decisions.
- ✓ The goal of explainable AI is to create AI systems that are transparent and accountable, so that their decision-making processes can be understood and trusted by humans.

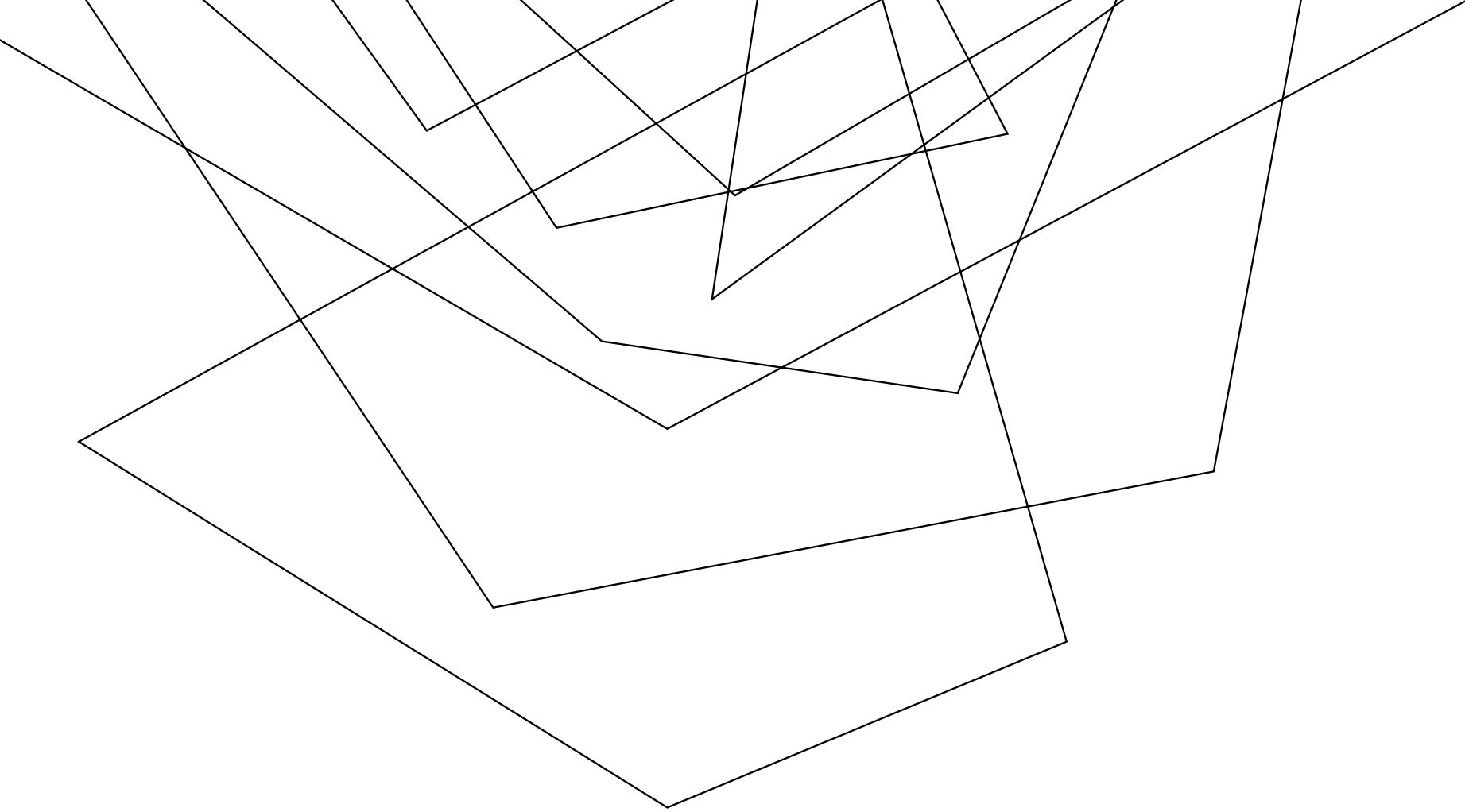
## AI Capabilities



User

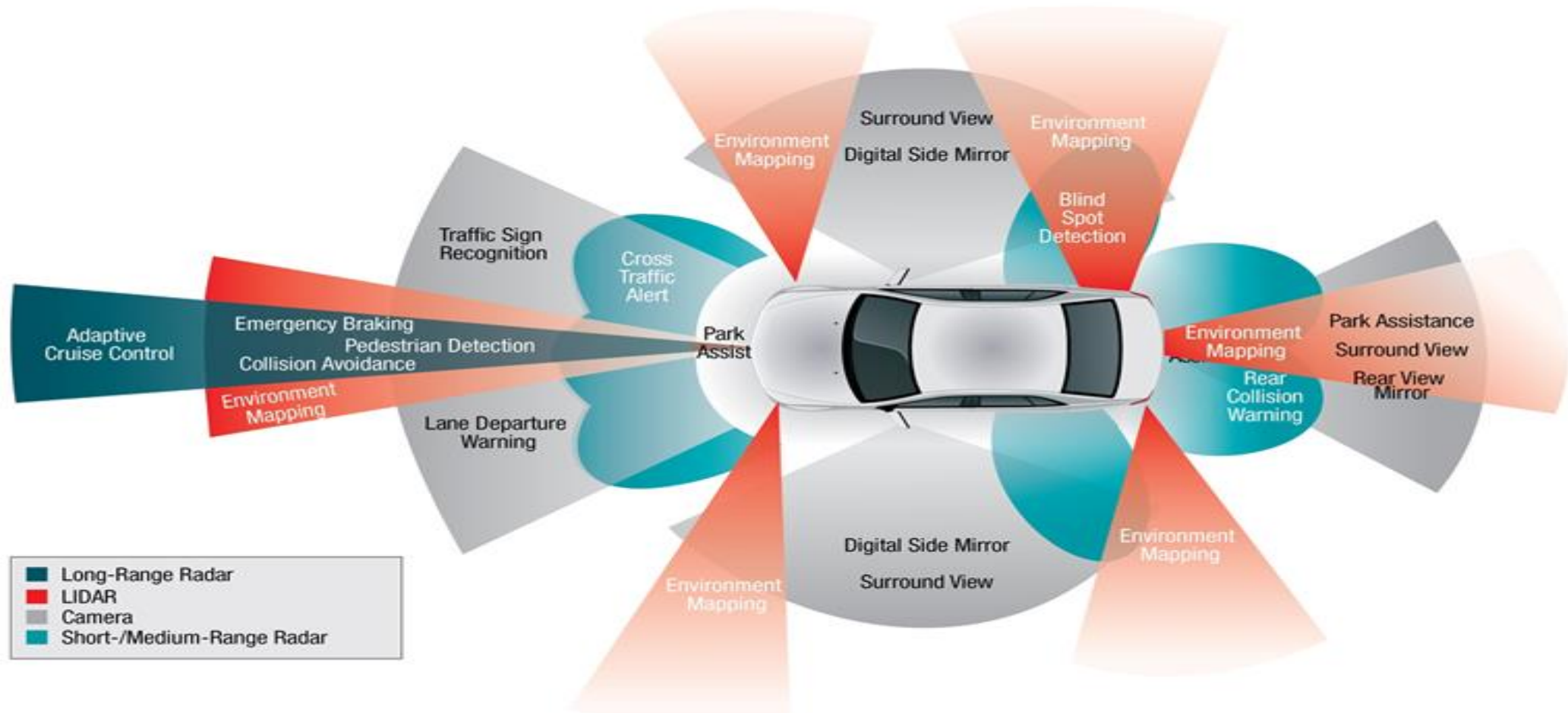


- ✓ One of the main challenges in developing explainable AI is that many modern AI systems, such as deep learning networks, are highly complex and operate on a large number of parameters, making it difficult to understand how they arrive at their decisions.
- ✓ As a result, there is a growing need for techniques and methods that can help make these systems more transparent and interpretable.

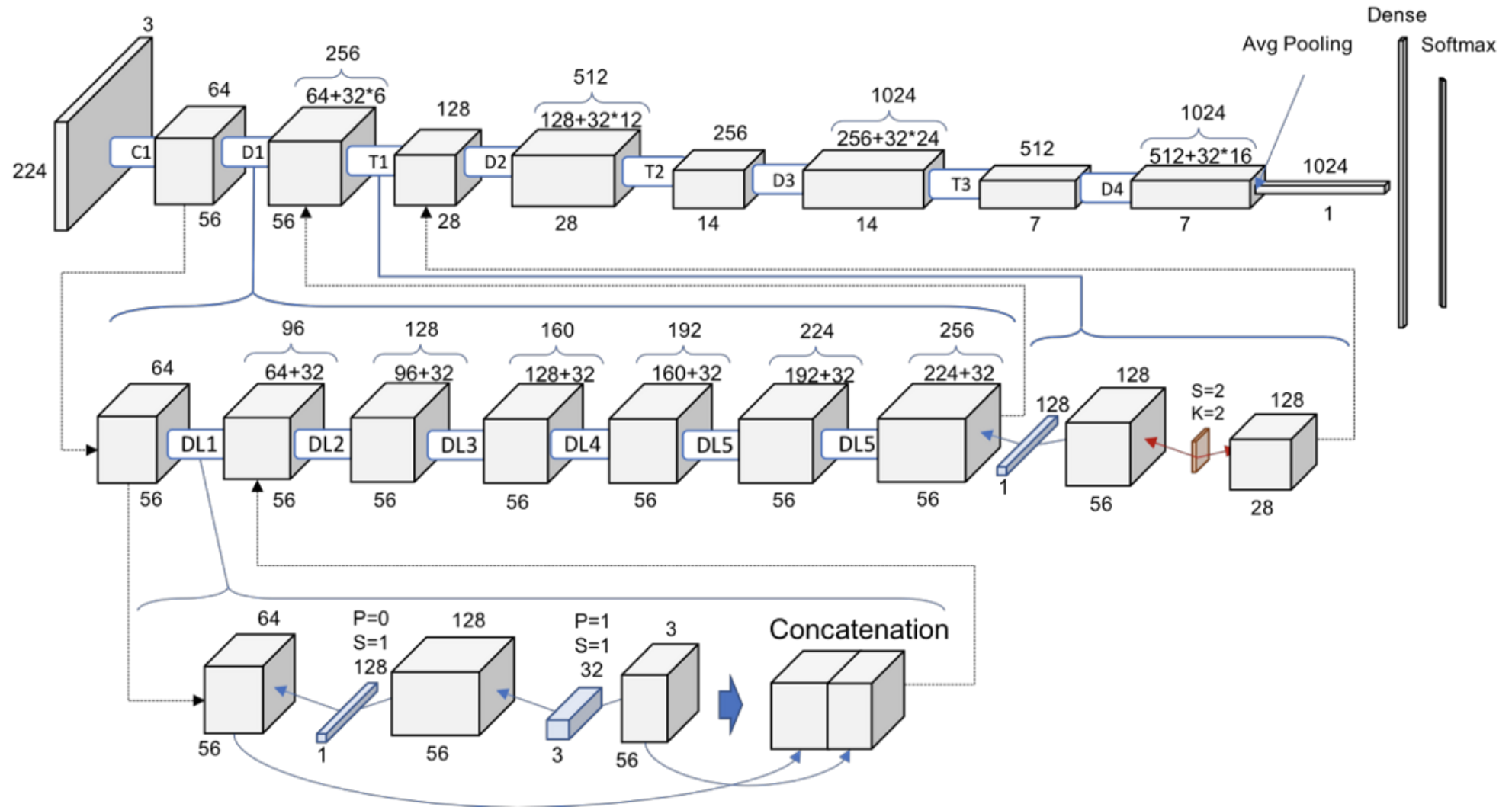


# **EXPLAINABLE AI IN AUTONOMOUS DRIVING**

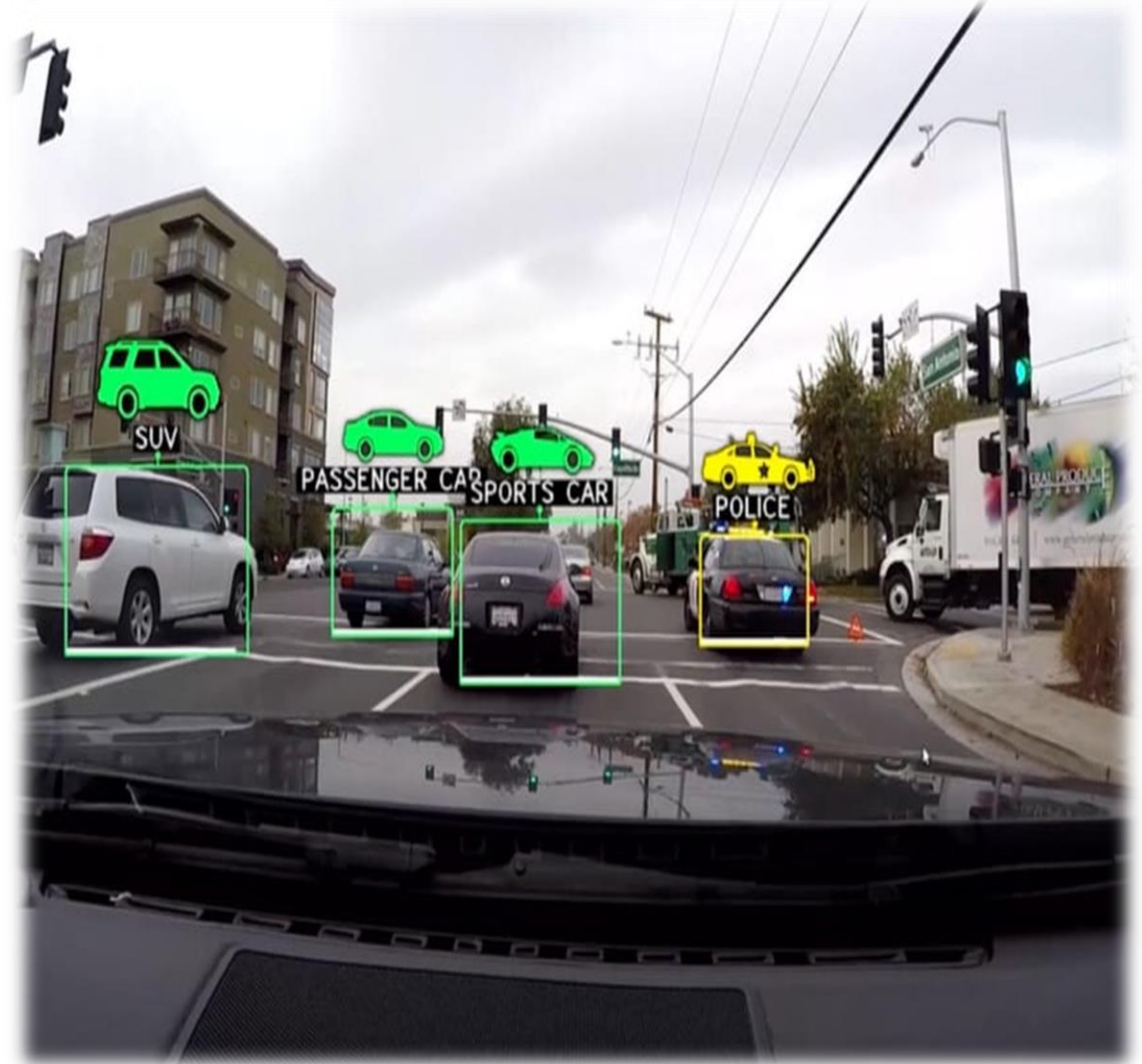
# FACTORS IN AUTONOMOUS DRIVING



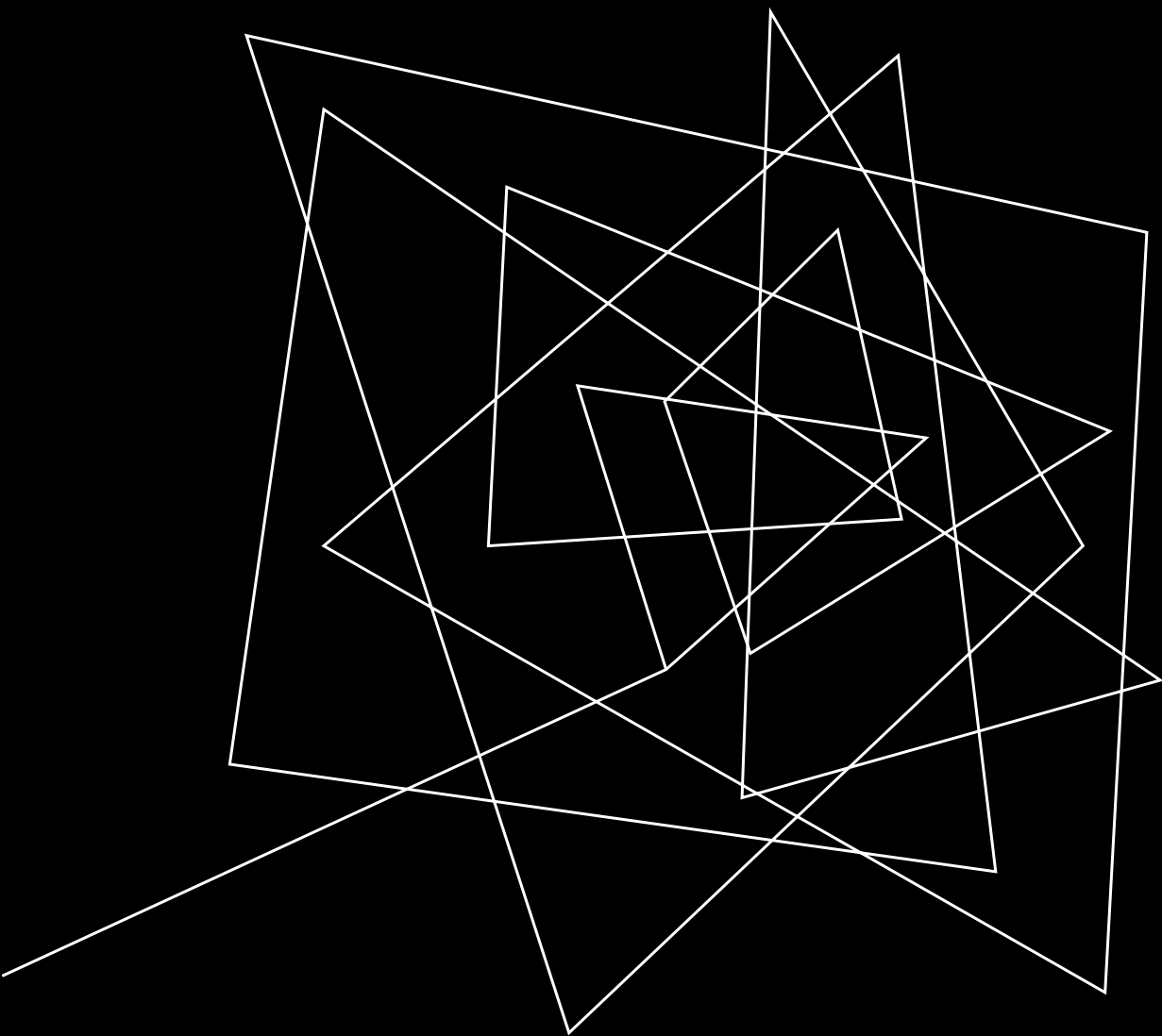
# COMPLEX ALGORITHMS REQUIRED



- ❑ By training a CNN on a large dataset of labeled images, it can learn to recognize and classify objects with high accuracy.
- ❑ However, the internal workings of a CNN can be complex and difficult to understand, which can make it challenging to explain the decisions it is making.
- ❑ To address this issue, researchers are developing techniques to make CNNs more interpretable and transparent, such as by visualizing the features that the network has learned to recognize or by providing explanations for the predictions it makes.



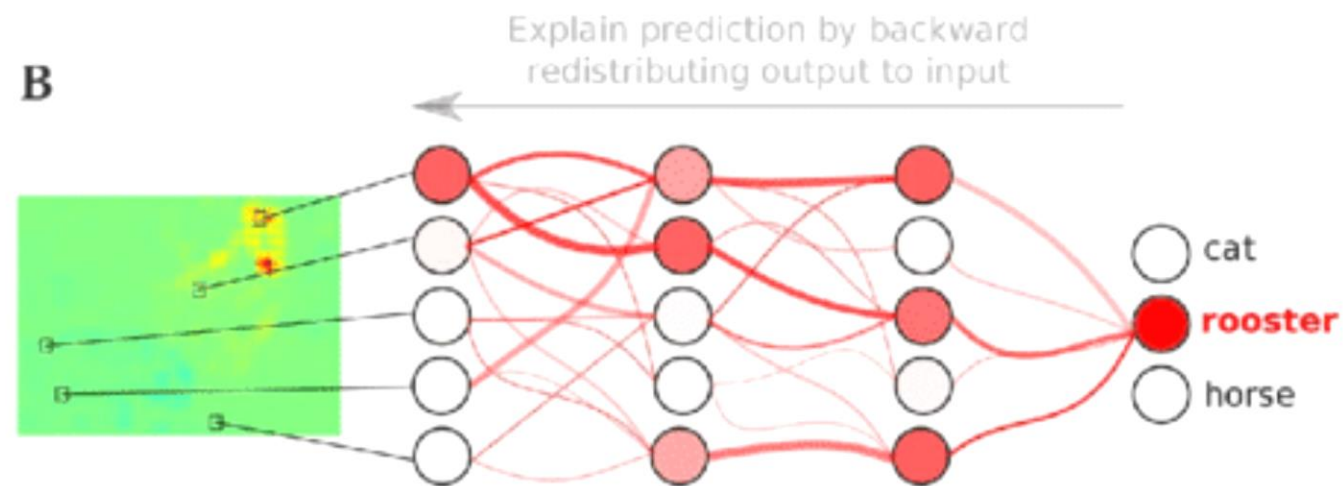
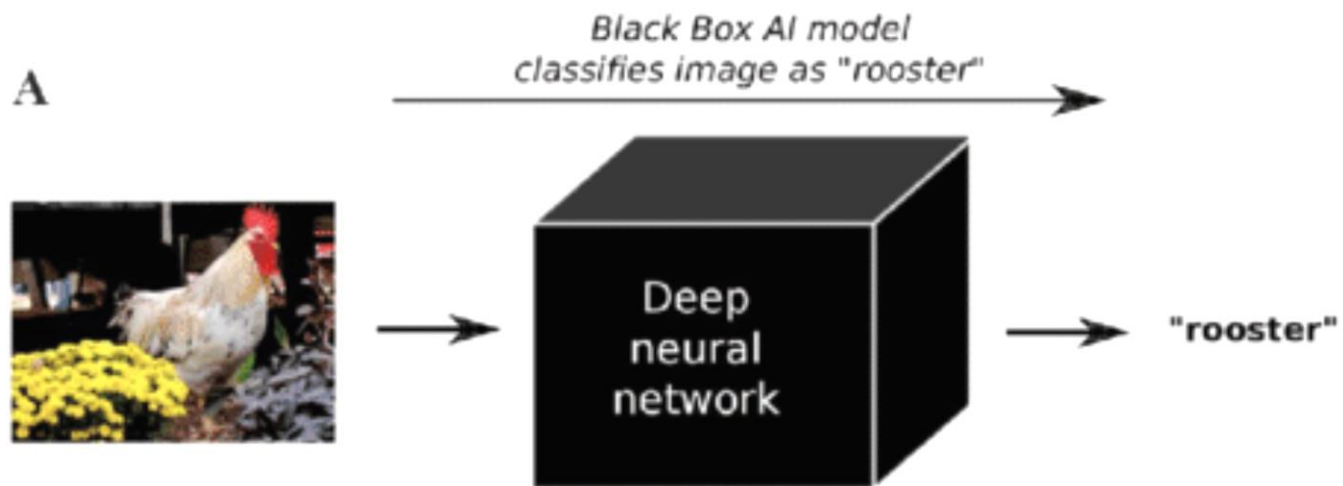




BLACKBOX AI

Vs

EXPLAINABLE AI



**C**

Image classified as "rooster" because of rooster's comb and wattles ✓

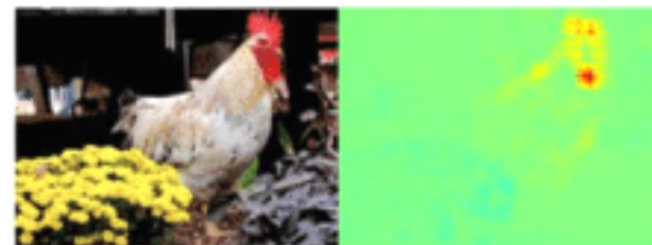


Image classified as "cat" because of cat's ears and nose ✓

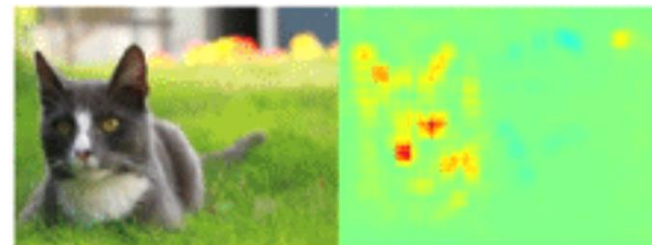
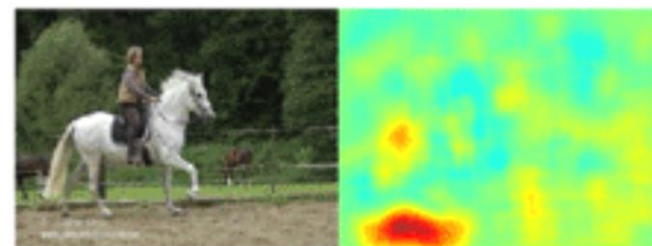
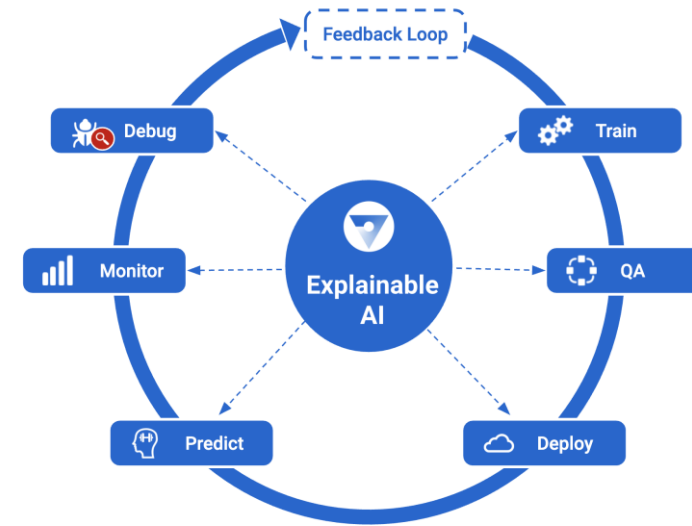


Image classified as "horse" because of a copyright tag ✗





## HOW EXPLAINABLE AI HELP'S IN DEEP LEARNING ?

What are the approaches Involved in XAI ?

# What is XAI Approaches involved in deep learning ?

## ➤ Feature-based approaches

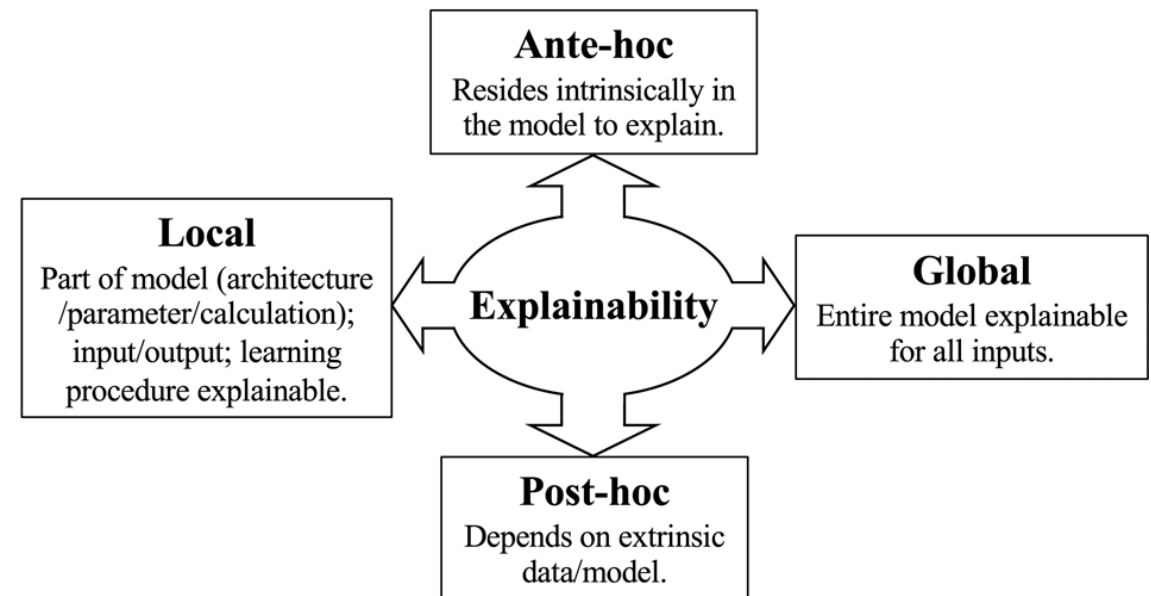
- ✓ Feature visualization
- ✓ Sensitivity analysis

## ➤ Rule-based approaches

- ✓ Decision tree Analysis
- ✓ Rule Lists

## ➤ Model-based approaches

- ✓ LIME
- ✓ SHAP



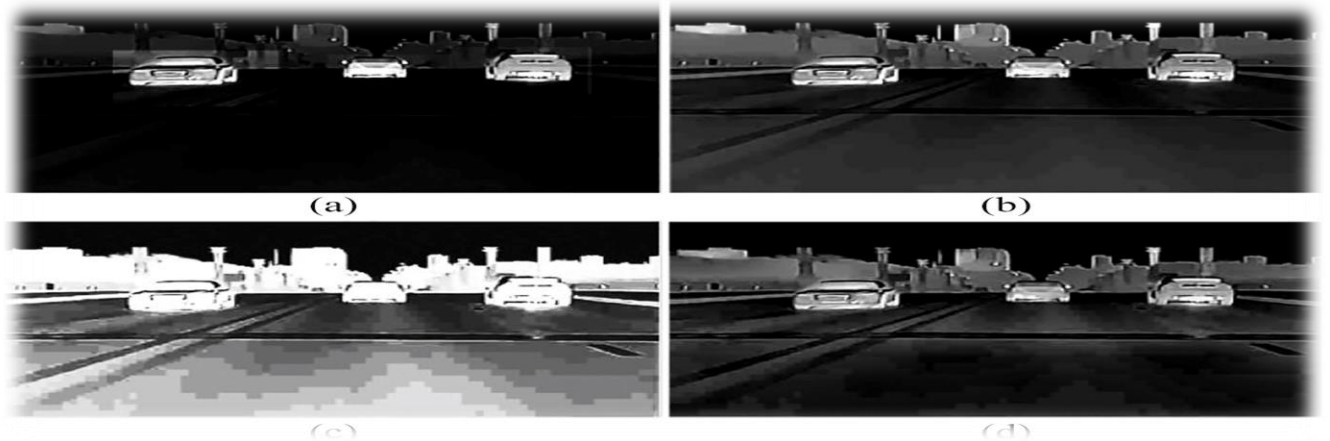
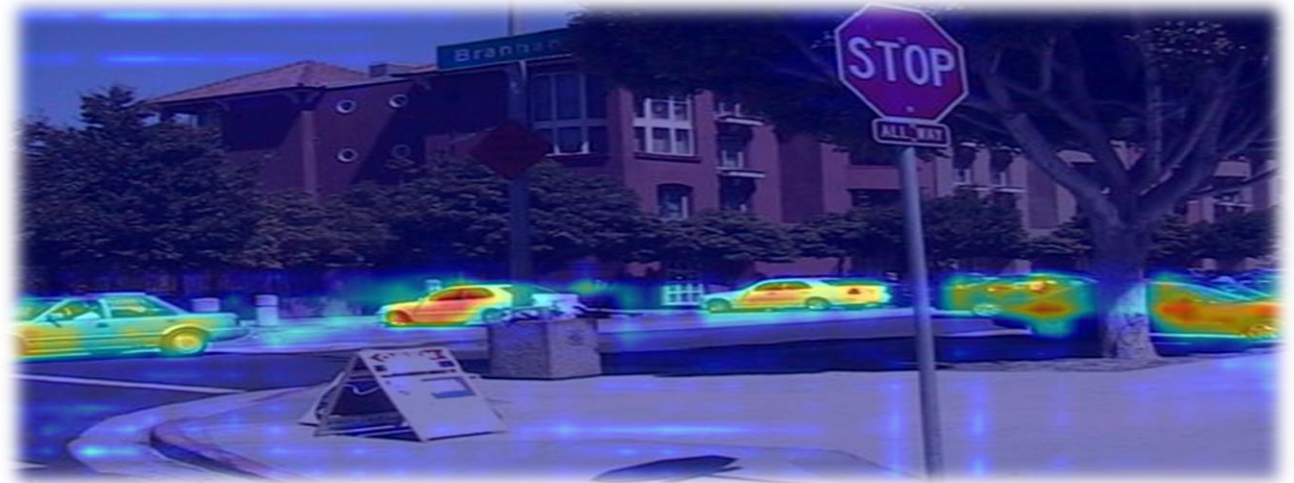


# **FEATURE BASED- APPROACHES**

# Feature visualization

It is an approach which involves feature extraction from dataset, which helps to identify the patterns and relationship among the features.

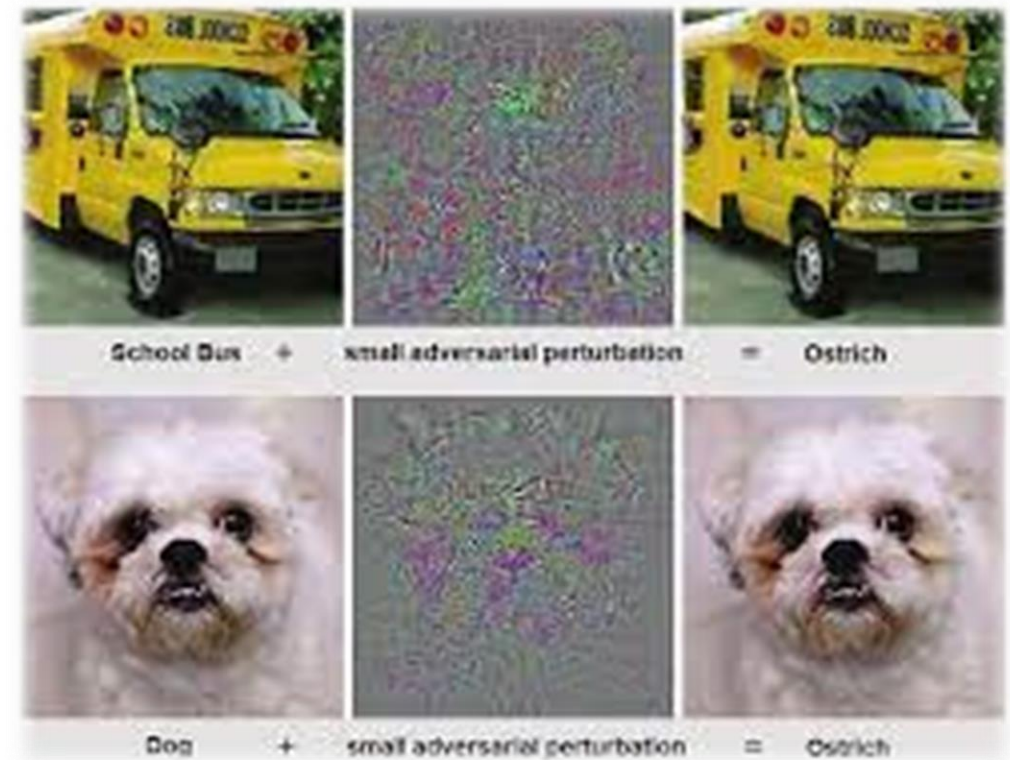
- **Class-Activation maps:** Activation maps show which features in the input data are most activated or "important" for making a prediction. This can help identify which features are most influential in the model's decision-making process.
- **Saliency maps:** Saliency maps highlight the regions in the input data that most influence the model's prediction. This can help identify the specific areas in an image or text that are most important for the model's decision.



# Sensitivity analysis

The approach involves the analysing the how model prediction's changes, when dependent variables varies. This can help identify the most important variables for making a prediction and identify any biases in the model.

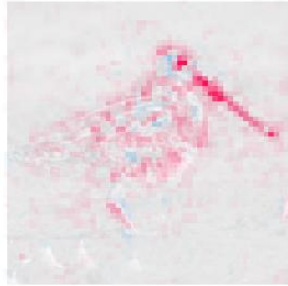
- ✓ **Perturbation analysis**: This involves making small changes to the input data and observing how the model's prediction changes. This can help identify which features are most influential in the model's decision-making process.







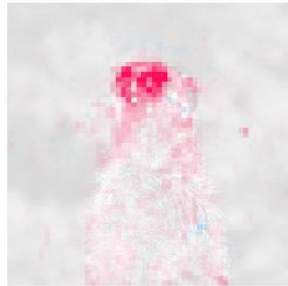
dowitcher



red-backed\_sandpiper



meerkat



mongoose



- ✓ **Shapley values**: These values measure the contribution of each feature to the model's prediction. This can help identify which features are most important for making a prediction and identify any biases in the model.





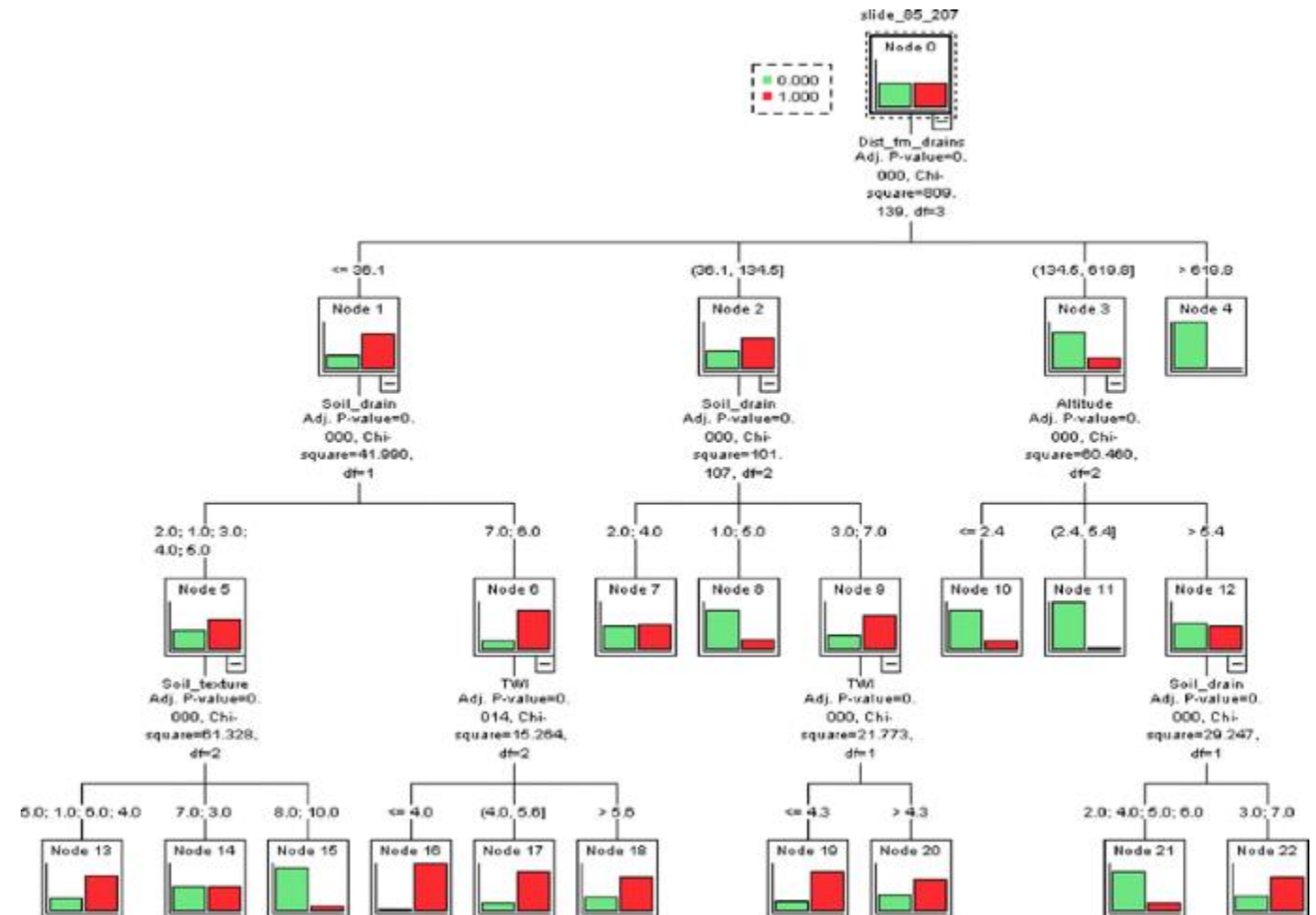


# **RULE BASED- APPROACHES**

# Decision Tree Analysis

decision tree analysis approaches in XAI for deep learning models are used to create clear, logical explanations for the decisions made by the model and improve its transparency and interpretability.

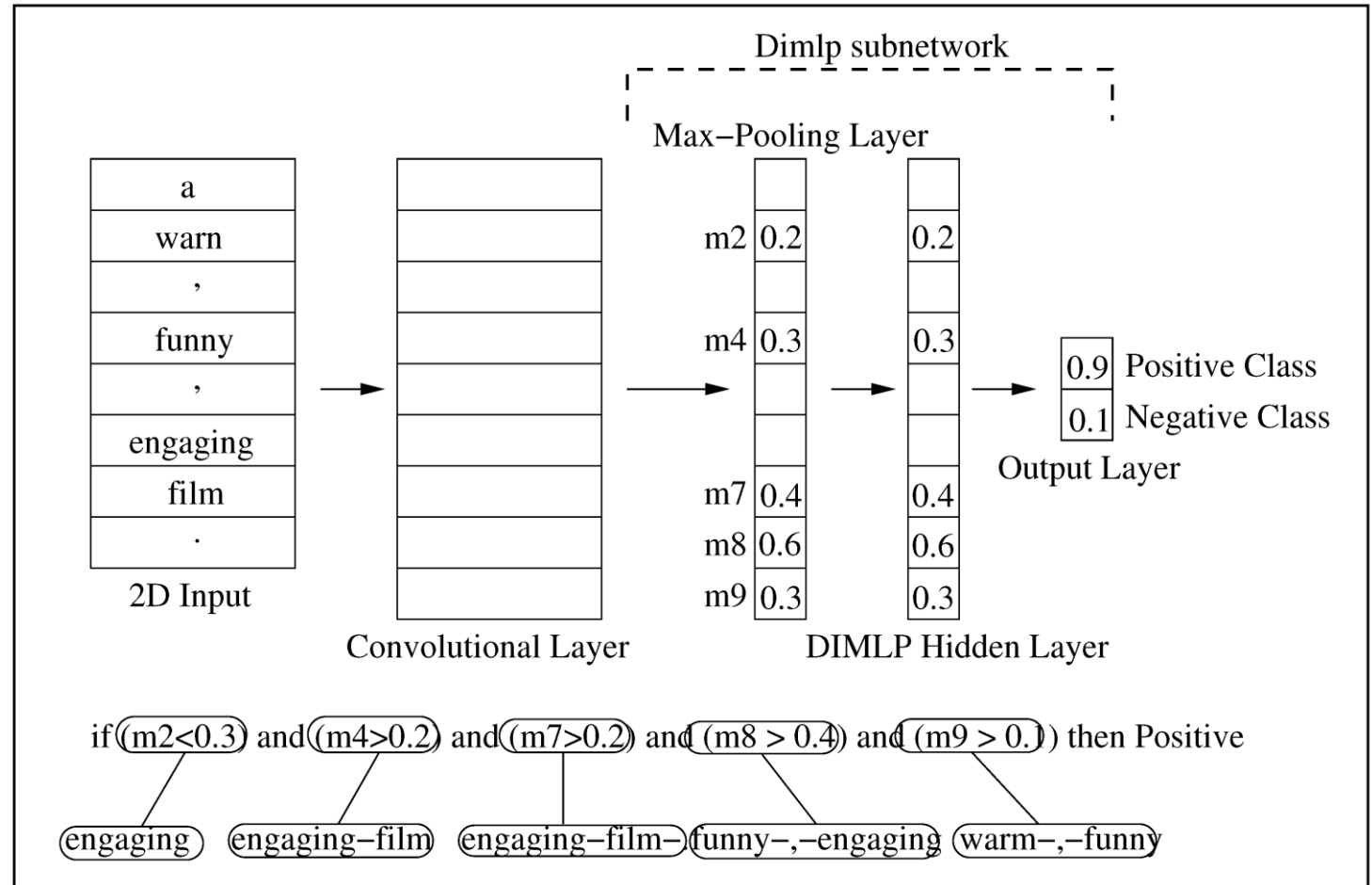
- **CHAID (Chi-squared Automatic Interaction Detection):** This approach is used for regression tasks and involves the use of the chi-squared statistical test to determine the statistical significance of the relationships between variables. It can also identify interactions between variables, allowing for a more robust and accurate model.

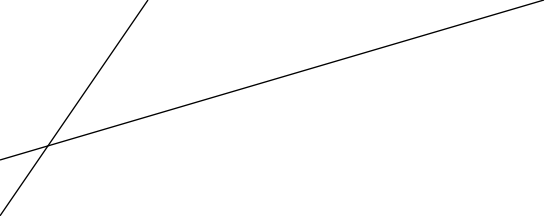


# RULE LISTS

A rule list is a set of logical statements (e.g., "if x is greater than y, then output z") used to represent the decision-making process of a model. Rule lists can be more compact than decision trees, but may be less intuitive to interpret.

- ✓ **Rule lists** can be used to explain the decision-making process of a deep learning model at a high level, but may not provide as much detail as other XAI techniques such as decision trees or gradient-based methods. They can be useful for understanding how the model is making predictions and identifying potential biases, but may not be as effective at explaining the model's decision-making process at a detailed level.

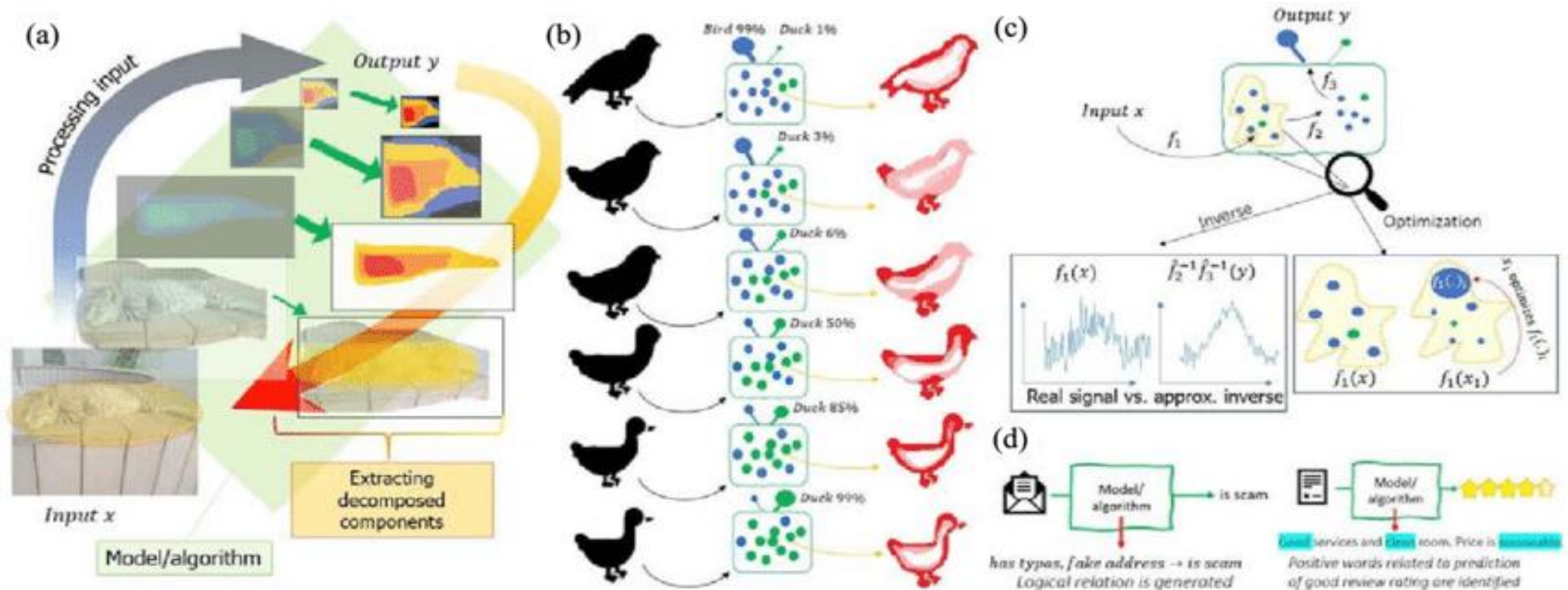




# **MODEL BASED- APPROACHES**

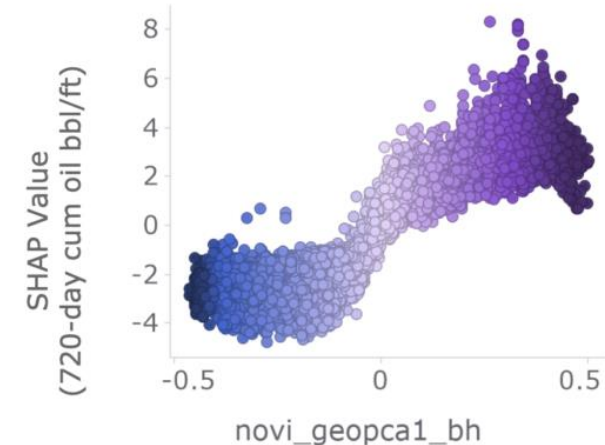
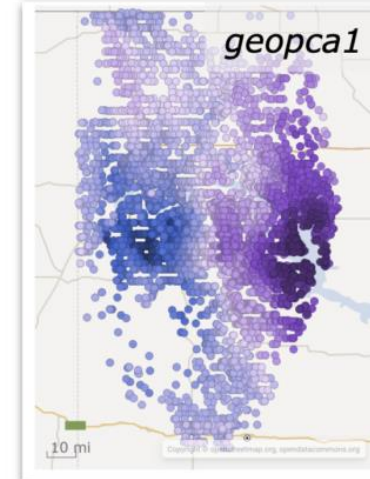
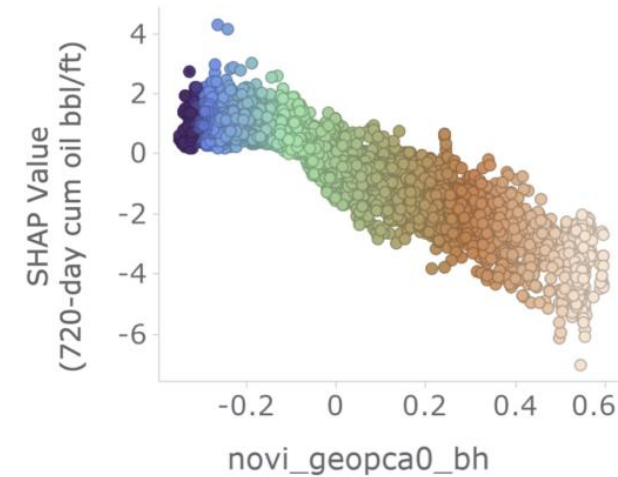
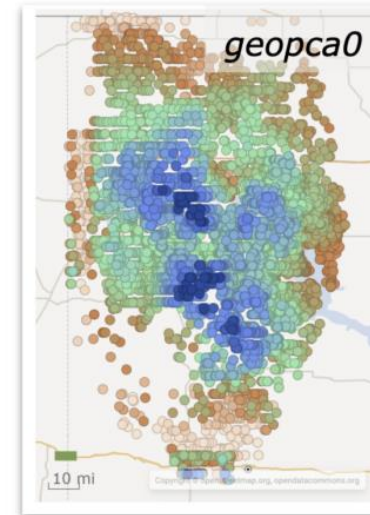
# LIME

(Local interpretable model-agnostic explanations) LIME is a model-agnostic approach that involves fitting a simple interpretable model to the local neighborhood around a single prediction made by a deep learning model. LIME can be used to explain the decision-making process of the deep learning model at a specific prediction, rather than globally.



# SHAP( Shapley Additive Explanations)

- AI has shown impressive results in areas such as natural language processing, computer vision, etc. Today, it takes just a few lines of code to implement state-of-the-art AI models, it's fascinating. However, as humans, how can we interpret the predictions that AI models make? How can we measure the importance that the model gives the data to be inferred?
- The interpretability of AI models is an active research area. Several alternatives have been proposed in recent years, one of it is SHAP (*Shapley Additive Explanations*).





## Pandemic Data



### Feature Engineering & Model Analysis



- *features selection*
- *gender analysis*
- *age-groups analysis*
- *days under treatment*
- *temporal evolution*
- *cluster-map of cases*
- *region-based analysis*

### Machine Learning with Hyperparameters

- *degree parameter*
- *c-regularization*
- *gamma ( $\gamma$ )*
- *kernel*



*Tuning*

### Deep Learning

### Autoencoders



ANN

Autoencoder-based  
SVM & LR

Predictions of Survival Chances of  
Quarantined Patients

*Performance Analysis &  
Prediction*



THANK YOU