XAI in Image Forensics

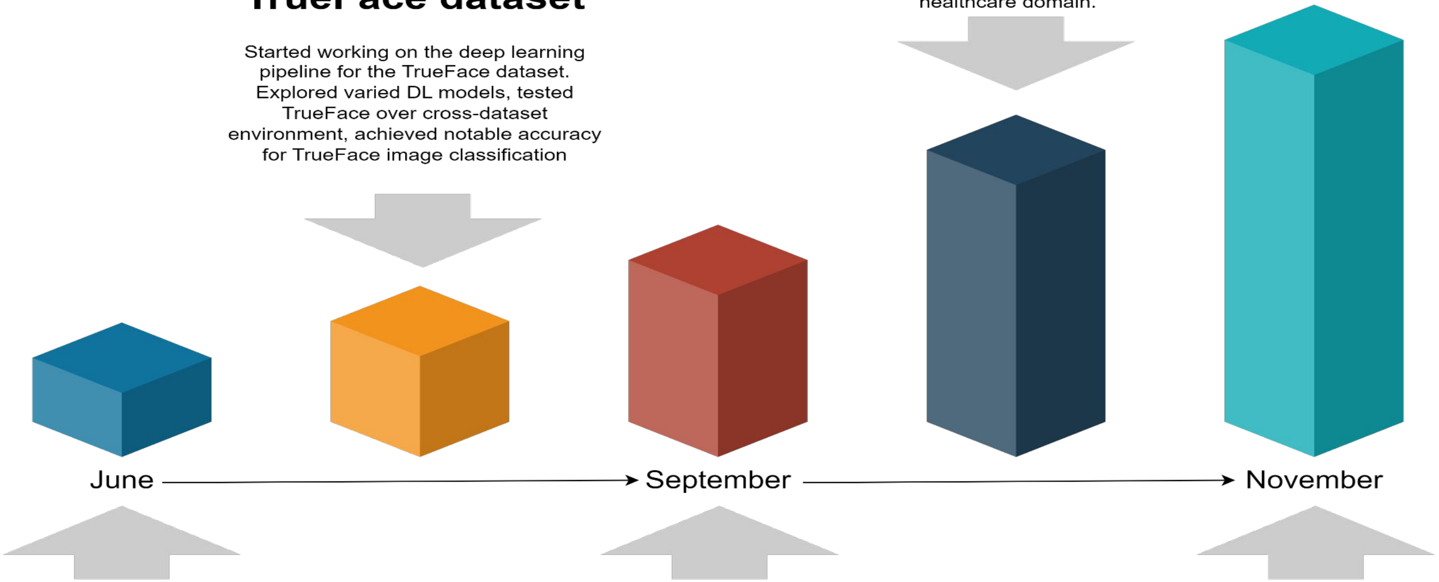
Under the supervision of Dr. Puneet Goyal

By Jadhav Abhilasha Sahebrao and Protyay Dey

Topics of Discussion

- Journey So Far
- What is XAI?
- XAI Techniques
- XAI in Image Forensics
- LIME
- SHAP
- GradCAM
- GradCAM++
- Visual Examples of XAI Techniques
- References

Journey So Far



Literature Survery: DeepFake

Started amidst of June, reading papers related to DeepFake image classification and Image Forensics. Also came across TrueFace Dataset.

Exploration of TrueFace dataset

Started working on the deep learning pipeline for the TrueFace dataset. Explored varied DL models, tested TrueFace over cross-dataset environment, achieved notable accuracy for TrueFace image classification

CVIP 2023 Paper Acceptance

Successfully submitted our work on TrueFace dataset (DeepFake Image Classification). The paper was accepted at CVIP 2023, IIT Jammu. Were inquisitive about how models were classifying images with high accuracy.

Literature Survey: XAI

Started reading about various XAI techniques, their classification, their usage in different scenarios. Implemented basic XAI techniques namely: CAM, Saliency Map. Also explored XAI in healthcare domain.

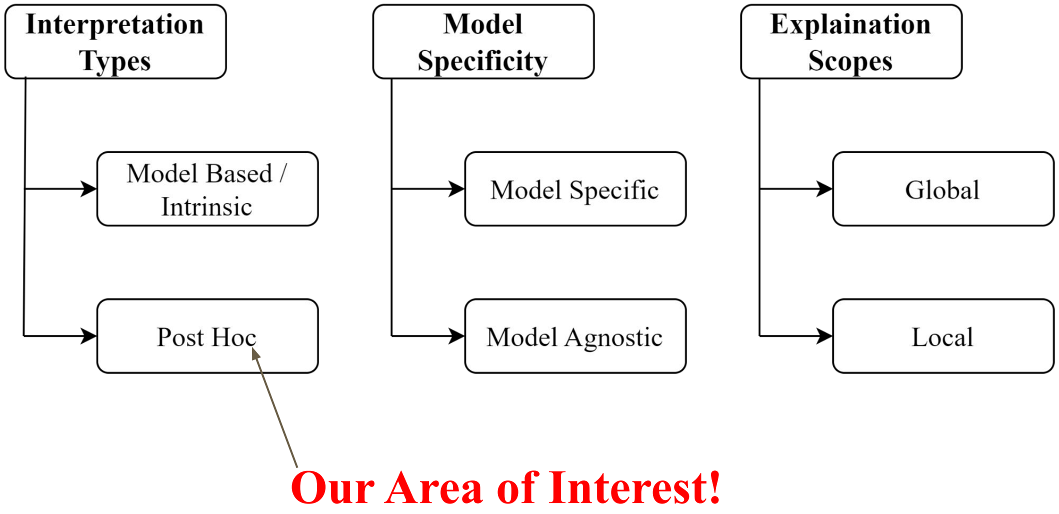
XAI in Image Forensics

Explored the domain of XAI in image forensics and Implemented GradCAM, GradCAM++, LIME on Intel Image Classification Dataset. Surveyed various XAI techniques applied in Image Forensic problems.

XAI Techniques

XAI techniques are based on three criterias:

- **Model-Based / Intrinsic versus Post Hoc**
- **Model-Specific versus Model-Agnostic**
- **Global versus Local** (i.e., the scope of the explanation).

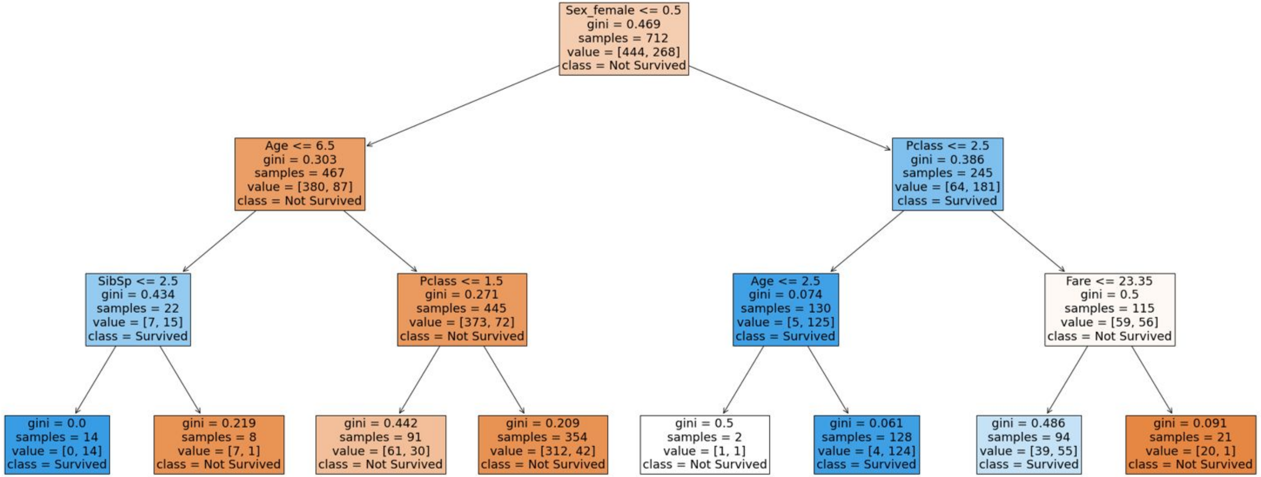


Source: Chaddad, Ahmad, et al. "Survey of explainable AI techniques in healthcare." Sensors 23.2 (2023): 634. [9]

Model-Based / Intrinsic Explanation

Model Nature

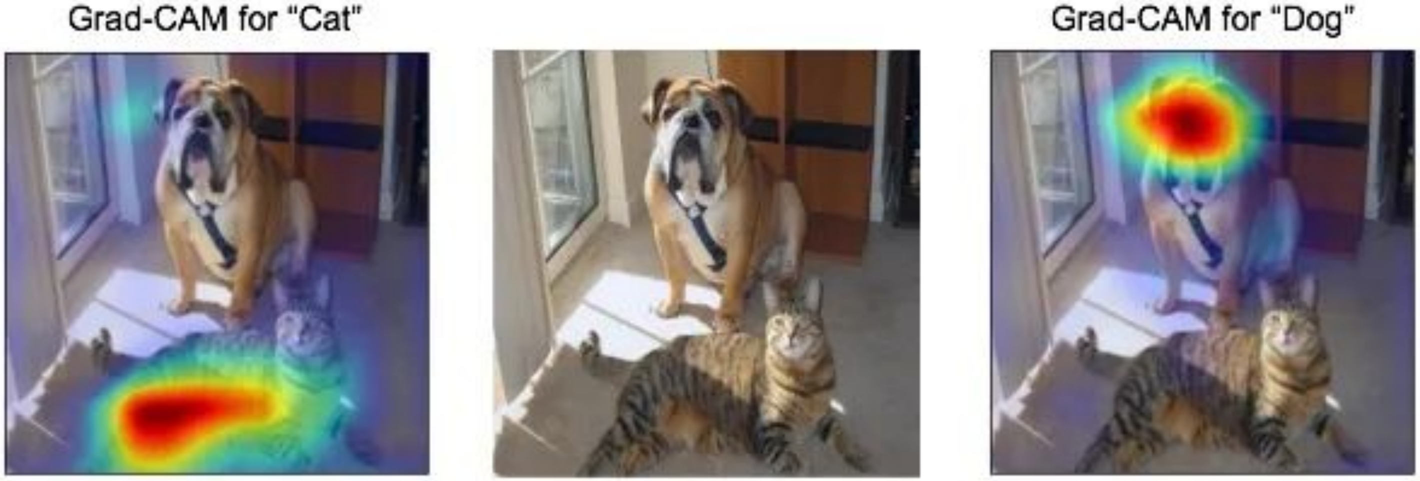
- Model-based / Intrinsic explanations are **based on traditional machine learning models**, such as linear regression or support vector machines, which are relatively simple and interpretable.



Source: https://www.shedloadofcode.com/blog/understanding-explainable-ai-for-classification-regression-and-clustering-with-python/ [23]

Post Hoc Explanation

- **Employed after the training of a complex model**.
- These methods do not enforce the model to be inherently explainable during training but they **analyze the model's behavior after it has been trained**.



Source: https://www.kaggle.com/datasets/puneet6060/intel-image-classification/data [24], https://medium.com/@mohamedchetoui/grad-cam-gradient-weighted-class-activation-mapping-ffd72742243a [26]

Post Hoc Explanation

- Post hoc explanations are **used with complex models**, particularly deep neural networks, which have thousands to millions of weights.
- These models are **inherently complex**, non-sparse, and not suited for direct human simulation and reasoning.
- Post hoc explanation methods include **analyzing learned features, feature importance,** and **interactions** after the model is trained.
- Visual explanations through techniques like **saliency maps** are used to highlight the most influential regions of an input, providing insights into the model's decision-making.

Model-Specific

- **Model Specific**: Tailored to specific classes or types of models.
- Use attributes and techniques that are specific to a particular model class, **such as certain neural network architectures**.
- Optimized to work with the characteristics and features unique to the chosen model type.
- Restricts the choice of models.
- Leads to **less flexible and adaptive explanation strategy** due to the focus on a particular model class.

Model-Agnostic

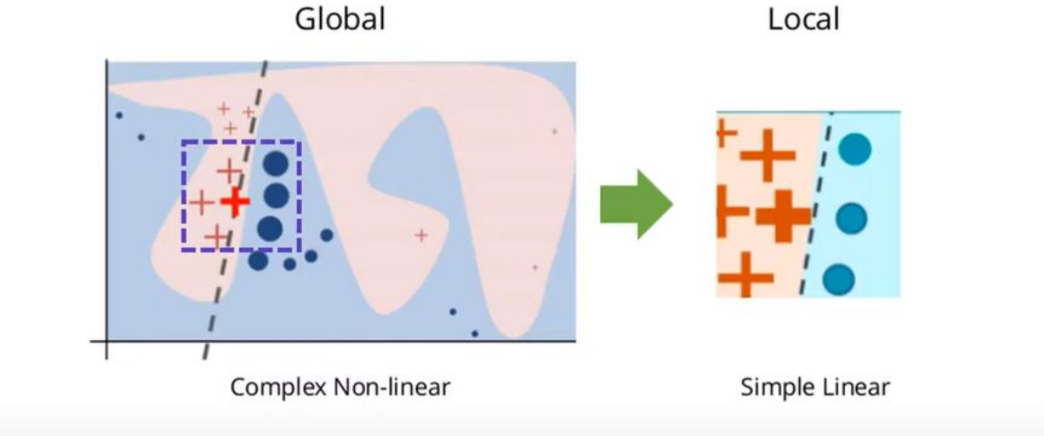
- **Model-agnostic**: Operates independently of the model choice and architecture.
- Not specific to any particular model class but focus solely on the **input** and **output** of the model in a generic way. These are designed to work with a wide variety of models, making them versatile and adaptable.
- Involves **perturbing** the input data to observe how changes in the input affect the model's output.
- Naturally **PostHoc**

Global Explainability

- **Global-scoped Explainability**: XAI techniques aim to provide insights and explanations that cover the behavior of the AI model across the **entirety** of its **dataset** or **decision space** rather than focusing on individual predictions or instances.
- **L**ayer-wise **R**elevance **P**ropagation (LRP), SHAP(**SH**apley **A**dditive ex**P**lanations)

Local Explainability

- **Local Explainability**: Focuses on **individual predictions**.
- LIME(**L**ocal **I**nterpretable **M**odel Agnostic **E**xplanations), SHAP(**SH**apley **A**dditive ex**P**lanations)



Source: https://censius.ai/blogs/global-local-cohort-explainability [25]

XAI in Image Forensics

Badhrinayaran et. al [12] focuses on detecting **Deepfake videos** using **XceptionNet** [17] trained on **FaceForensics++** [18] dataset.

It emphasizes interpretable models via XAI methods like **LRP** and **LIME**[3].

With a **balanced** dataset of real and fake images, the model achieved **94.33%** test accuracy.

XAI techniques, particularly LIME [3], effectively captured crucial regions in the images, confirming the model's attention.

LRP localized manipulation areas, identifying key features like the nose and mouth.

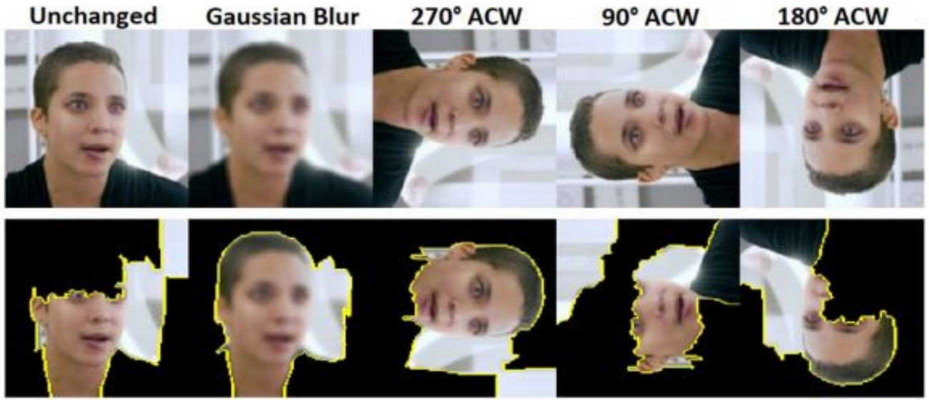


Figure 4. 1st Row shows the input along with its perturbations and 2nd row shows the LIME descriptions of these inputs. **[12]**

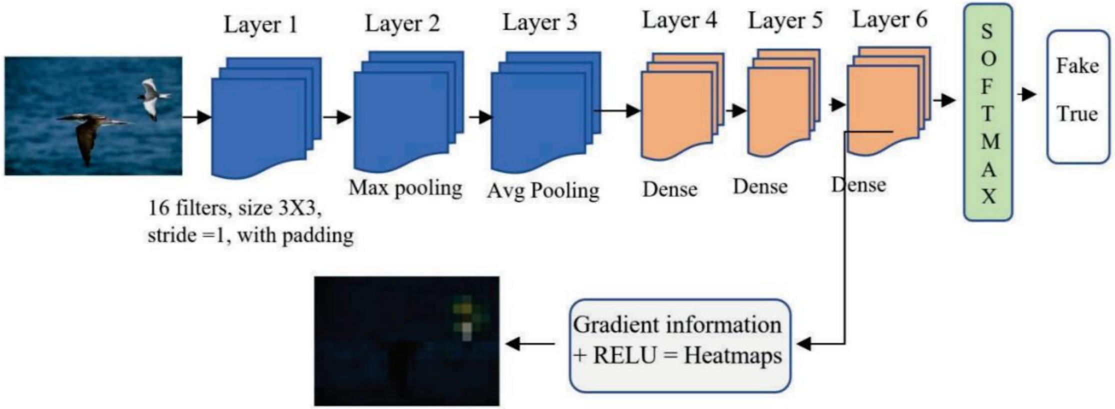
Source: Malolan, Badhrinarayan, Ankit Parekh, and Faruk Kazi. "Explainable deep-fake detection using visual interpretability methods." 2020 3rd International Conference on Information and Computer Technologies (ICICT). IEEE, 2020. [12]

XAI in Image Forensics

Bhuvanesh and Dilip [11] discusses about the detection of faithful images using a DL model. The images contain **Copy move** and **spliced** images.

Employing **high-pass** filters during weight initialization improves the detection of altered image elements.

The researchers employed **GradCAM[1]** to enhance interpretability, enabling precise localization of manipulated regions within the images.



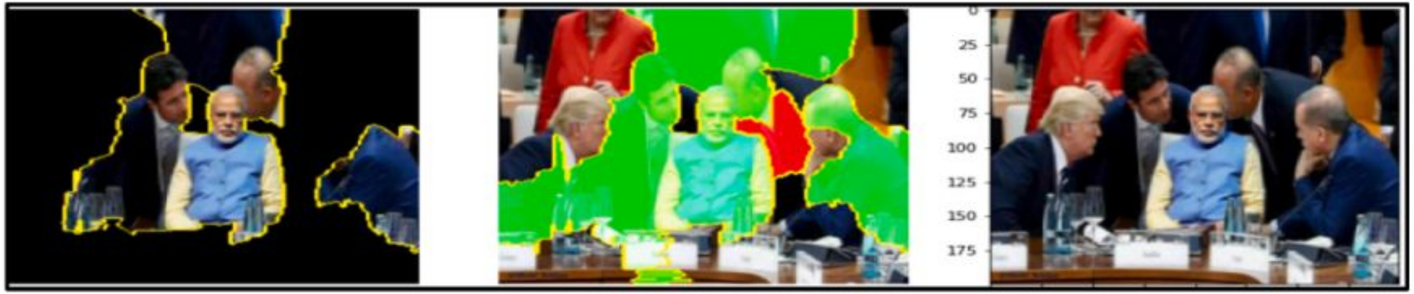
Source: Singh, Bhuvanesh, and Dilip Kumar Sharma. "Image forgery over social media platforms-A deep learning approach for its detection and localization." 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, 2021. [11]

XAI in Image Forensics

Bhuvanesh and Dilip [10] proposed a **custom CNN embedded with an attention mechanism**, the paper focuses on detecting forged images circulated on microblogging platforms.

Leveraging **high-pass filters** during weight initialization **enhances identification of tampered features.**

They utilized **LIME [3]** for **interpretability and pinpointing tampered regions.**



Source: Singh, Bhuvanesh, and Dilip Kumar Sharma. "SiteForge: Detecting and localizing forged images on microblogging platforms using deep convolutional neural network." Computers & Industrial Engineering 162 (2021) [10]

XAI in Image Forensics

Hasan et al. [8], conducted a study addressing the rising concern of **Deepfake** media by employing advanced AI methods.

They used different CNN models on a dataset(deepfake from Kaggle) containing **70,000** real(FFHQ) and **70,000** fake images(GAN) to distinguish between **authentic** and **Deepfake** content.

Among tested models, **InceptionResNetV2[19]** achieved **99.87%** accuracy.

To ensure reliability, they used **L**ocal **I**nterpretable **M**odel-Agnostic **E**xplanations [3]for Explainable AI, confirming the model's efficacy.

Source: Abir, W. Hasan, et al. "Detecting Deepfake Images Using Deep Learning Techniques and Explainable AI Methods." Intelligent Automation & Soft Computing. (2023): 2151-2169. [8]

XAI in Image Forensics

Kuchumova et al. [13], introduces a **Steganalysis** methodology called **STEG-XAI [13]**.

It detects **image steganography** and also explains the **model's detections** and extracts specific **signatures** of steganography algorithms.

used **EfficientNet-B0** [14] architecture and accompanied by XAI tools like LIME [3] and Grad-CAM [1], the model is trained on a **ALASKA2**[15] dataset of **UERD** modified **JPEG** images. achieved **AUC** of **0.944**.

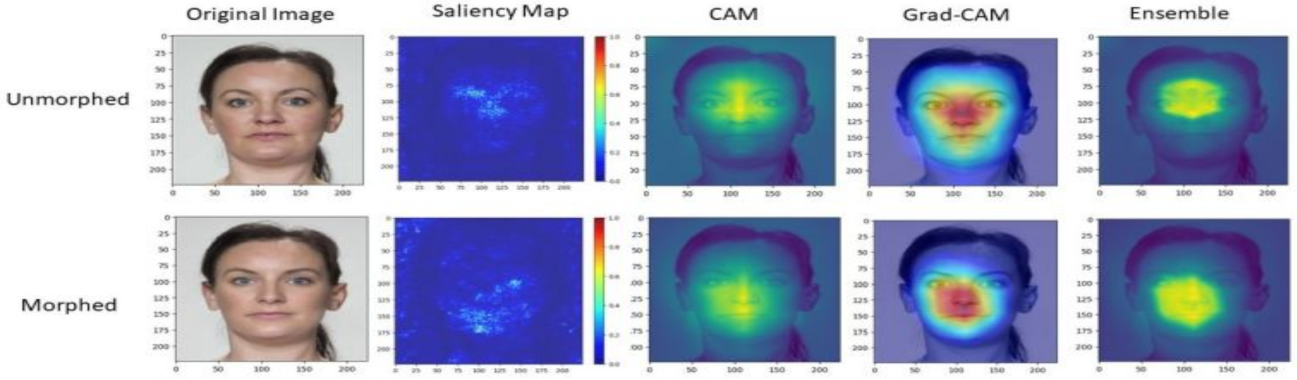
Source: Kuchumova, E., Martínez-Monterrubio, S. M., & Recio-Garcia, J. A. (2023). STEG-XAI: explainable steganalysis in images using neural networks. Multimedia Tools and Applications, 1-18 [13]

XAI in Image Forensics

With increasing attacks on biometric systems using morphed images, Rudresh et al. [16] have used **EfficientNet**[14] for Morphing Attack Detection. This work introduces **Ensemble XAI**, combining **Saliency**, **Class Activation**, and **Gradient-CAM**[1], to explain EfficientNet-B1's decisions on morphed vs. genuine images.

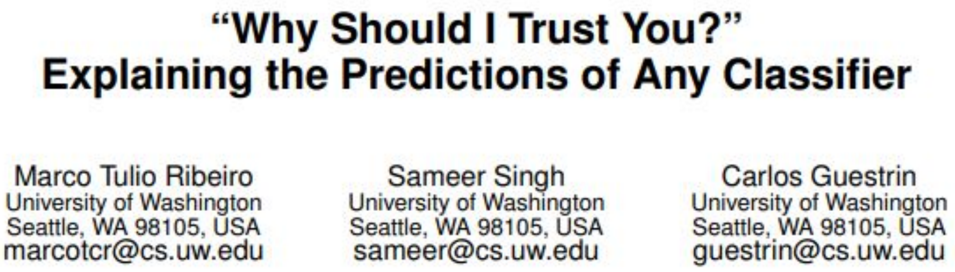
They used three datasets:

- **Face Research Lab London**(FRLL)[20] (102 faces),
- **Wide Multi Channel Presentation Attack**(WMCA)[21] (72 individuals)
- **Makeup Induced Face Spoofing**(MIFS)[22] (107 transformations).



Source: Dwivedi, Rudresh, et al. "An Efficient Ensemble Explainable AI (XAI) Approach for Morphed Face Detection." arXiv preprint arXiv:2304.14509 (2023).[16]

Local Interpretable Model Agnostic Explanations

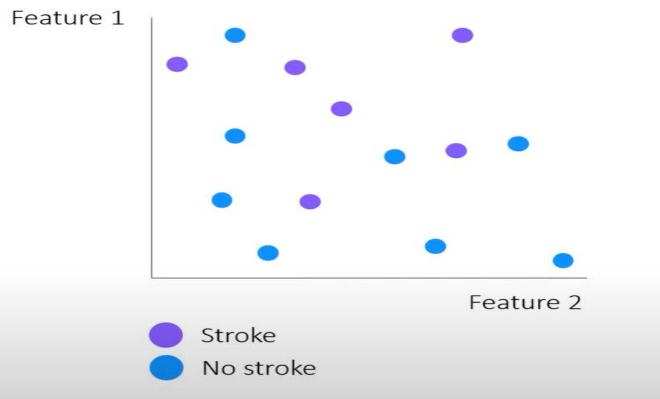


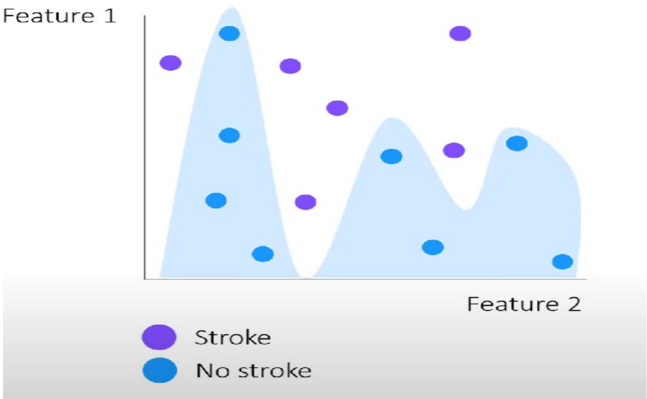
“Why Should I Trust You?”
Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu





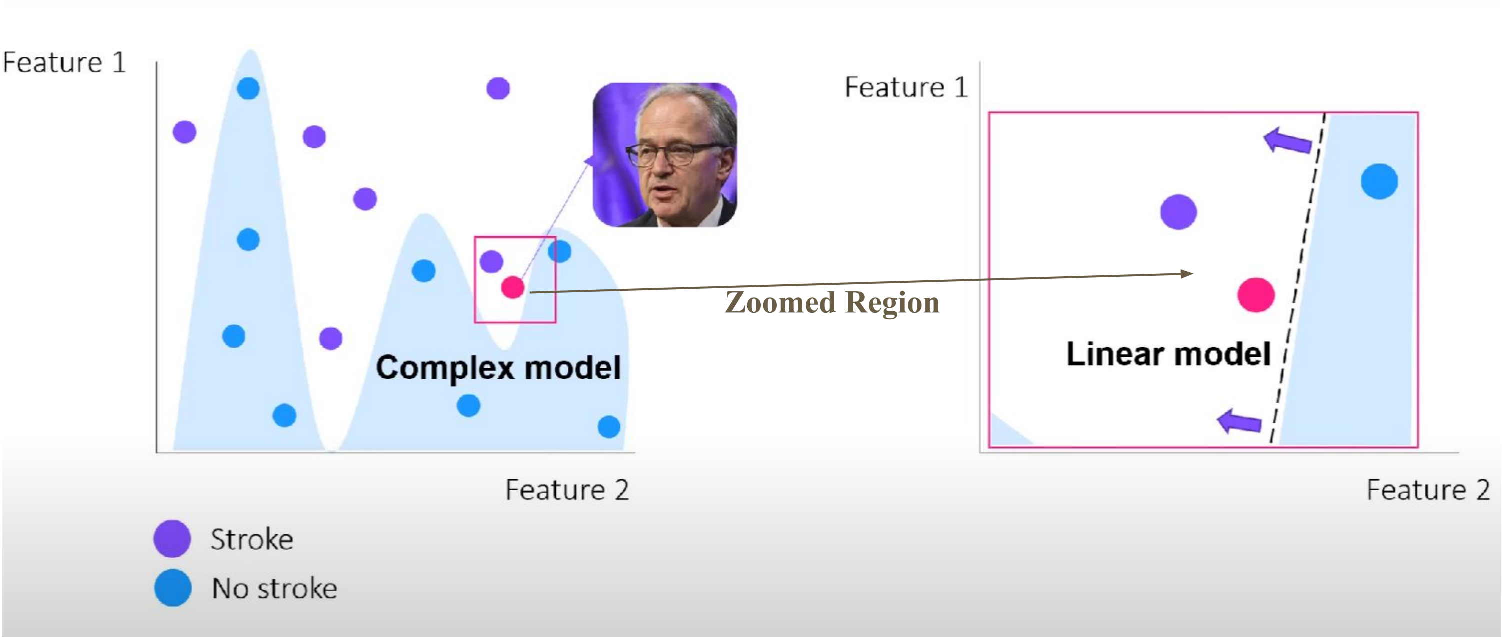
Source: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. [3]

LIME

- LIME [3] was introduced in **ACM SIGKDD** in 2016 by **Ribeiro et al.** and has over **14944 citations** till date.
- Complex Models **lack transparency**, making it challenging for users to comprehend **how** and **why** specific decisions are reached.
- LIME tackles this challenge by generating explanations for **individual** predictions.
- LIME focuses on providing insights into individual predictions by creating simpler, more understandable **surrogate** models in the **vicinity** of those predictions.

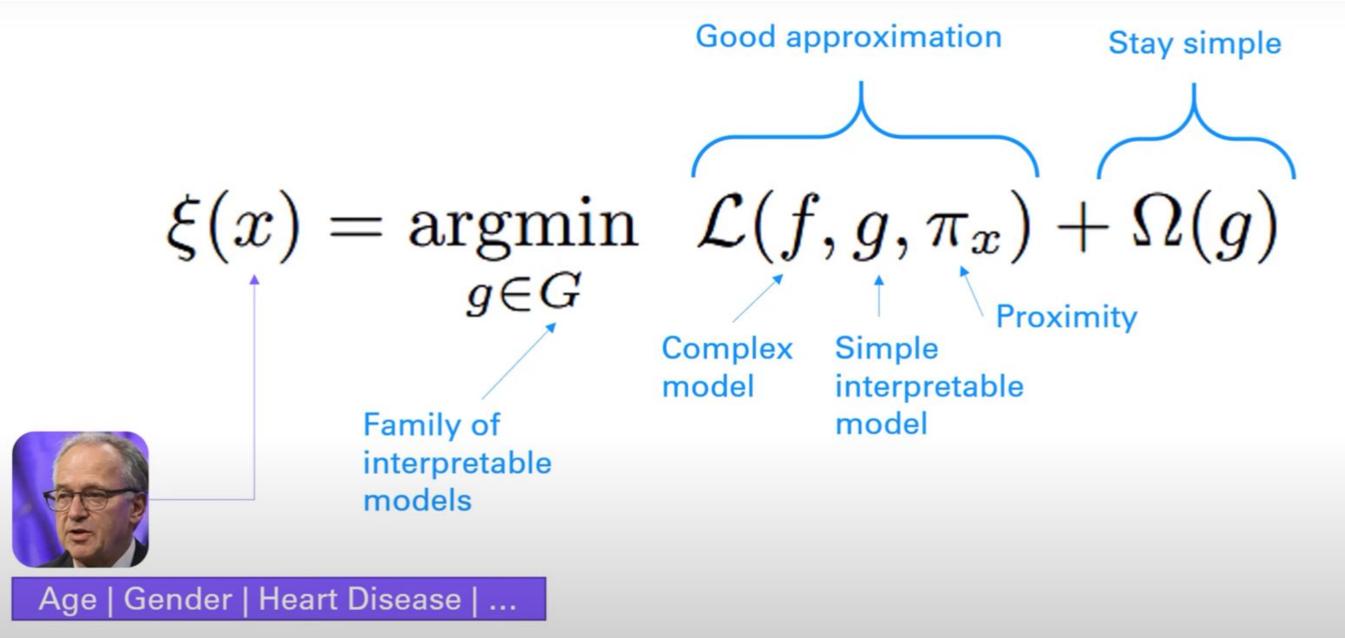
Source: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. [3]

LIME



Source: https://deepfindr.github.io/ [6]

Maths behind Local Interpretations



$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \overbrace{\mathcal{L}(f, g, \pi_x)}^{\text{Good approximation}} + \overbrace{\Omega(g)}^{\text{Stay simple}}$$

Family of interpretable models

Complex model

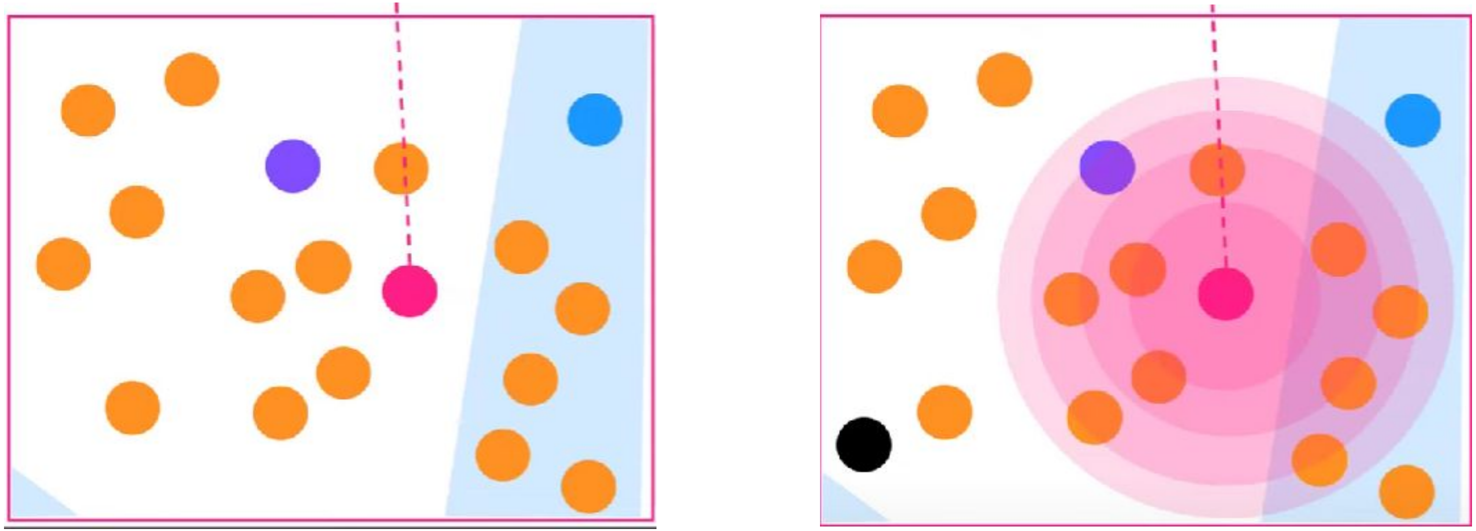
Simple interpretable model

Proximity

Age | Gender | Heart Disease | ...

Source: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016., [3] https://deepfindr.github.io/ [6]

Local Interpretations of LIME



Source: https://deepfindr.github.io/ [6]

Steps: LIME

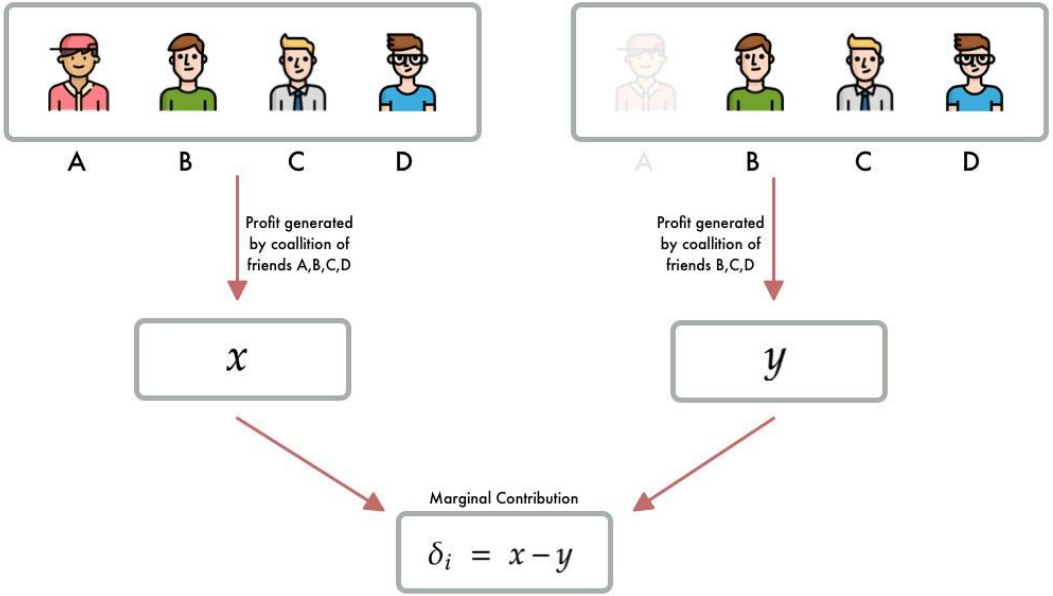
- The process involves **perturbing** the input data and observing how these perturbations influence the output of the model.
- LIME samples data points around the prediction of **interest**, generating a **dataset** that's more manageable for a simpler, interpretable model (often a linear model) to approximate the behavior of the complex model within that **local region**.
- These **local surrogate** models then provide explanations for the predictions, highlighting which features or attributes had the most significant impact on the model's decision-making process.

Source: Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. [3]

Shapely Values

The Shapley value, coined by **Shapley** (1953)63, is a method for assigning **payouts** to **players** depending on their **contribution** to the total payout.

Players cooperate in a **coalition** and receive a certain **profit** from this cooperation


A
B
C
D
A
B
C
D
Profit generated by coallition of friends A,B,C,D
Profit generated by coallition of friends B,C,D
x
y
Marginal Contribution
$\delta_i = x - y$

Source: https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3 [7]

Players? Game? Payout?

- What is the connection to machine learning predictions and interpretability?
- The **Game** is the **prediction task** for a single instance of the dataset.
- The **Gain** is the **actual prediction** for this instance **minus** the **average prediction** for all instances.
- The **Players** are the **feature** values of the instance that collaborate to receive the gain

Source: https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3 [7]

SHAP

SHAP [4] values provide insights into how the features in a model contribute to **individual** predictions.

They quantify the impact of **each feature** on a particular **prediction** by assessing how the prediction changes when the feature value changes.

Essentially, they break down the model prediction for a specific instance and attribute parts of that prediction to different features.

Source: Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017). [4]

GradCAM [1]

- Assigns each neuron a **relevance score** for the decision of interest.

$$L^{c}_{Grad-CAM} \in \mathbb{R}^{u\times v} = \underbrace{ReLU}_{\text{Pick positive values}} \left(\sum_{k} \alpha^{c}_{k} A^{k}\right)$$

- Our goal is to find the **localization map**, here u is the width, v the height of the explanation and c the class of interest.

- Forward-propagate the input image through the convolutional neural network.

- **Obtain** the **raw score for the class of interest**, meaning the activation of the neuron before the softmax layer.

- Set all other class activations to zero.

- Back-propagate the gradient of the class of interest to the last convolutional layer before the fully connected layers: $\frac{\delta y^c}{\delta A^k}$

Source: Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017. [1]

GradCAM [1]

- Weight each feature map “pixel” by the gradient for the class. Indices i and j refer to the width and height dimensions:

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \quad \underbrace{\frac{\delta y^c}{\delta A_{ij}^k}}_{\text{gradients via backprop}}$$

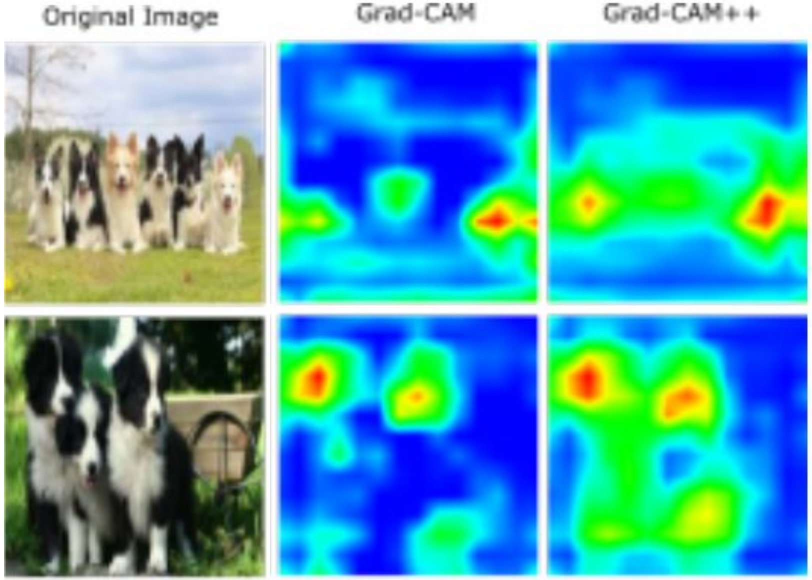
  This means that the gradients are globally pooled.

- Calculate an average of the feature maps, weighted per pixel by the gradient.
- Apply ReLU to the averaged feature map. (fancy way of saying that we set all negative values to zero).
- **For visualization**: Scale values to the interval between 0 and 1.
- Upscale the image and overlay it over the original image.

Source: Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017. [1]

Limitations of GradCAM [2]

- **Performance drops** when **localizing multiple occurrences** of the same class.
- For single object images, Grad-CAM heatmaps **often do not capture the entire object in completeness.**



Source: Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." 2018 IEEE winter conference on applications of computer vision (WACV). IEEE, 2018 [2]

GradCAM++ [2]

- Practically equivalent to a very simple variation of Grad-CAM in which gradients are **replaced with positive gradients.**
- **Grad-CAM** uses a **global average pooling of the partial derivatives** of the activation maps, while **Grad-CAM++** takes a **weighted combination of positive partial derivatives**.
- Considering higher-order derivatives, Grad-CAM++ captures more detailed information about the relationship between the activation maps and the class score.

$$\alpha_{ij}^{kc} = \frac{\frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2}}{2\frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2} + \sum_a \sum_b A_{ab}^k \{\frac{\partial^3 Y^c}{(\partial A_{ij}^k)^3}\}}$$

Y^c represents the class score for class c, A_{ij}^k represents the activation maps of the last convolutional layer.

Source: Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." 2018 IEEE winter conference on applications of computer vision (WACV). IEEE, 2018 [2]

GradCAM++ [2]

- Proposed solution is to replace the plain average of gradients at each feature map with a weighted average

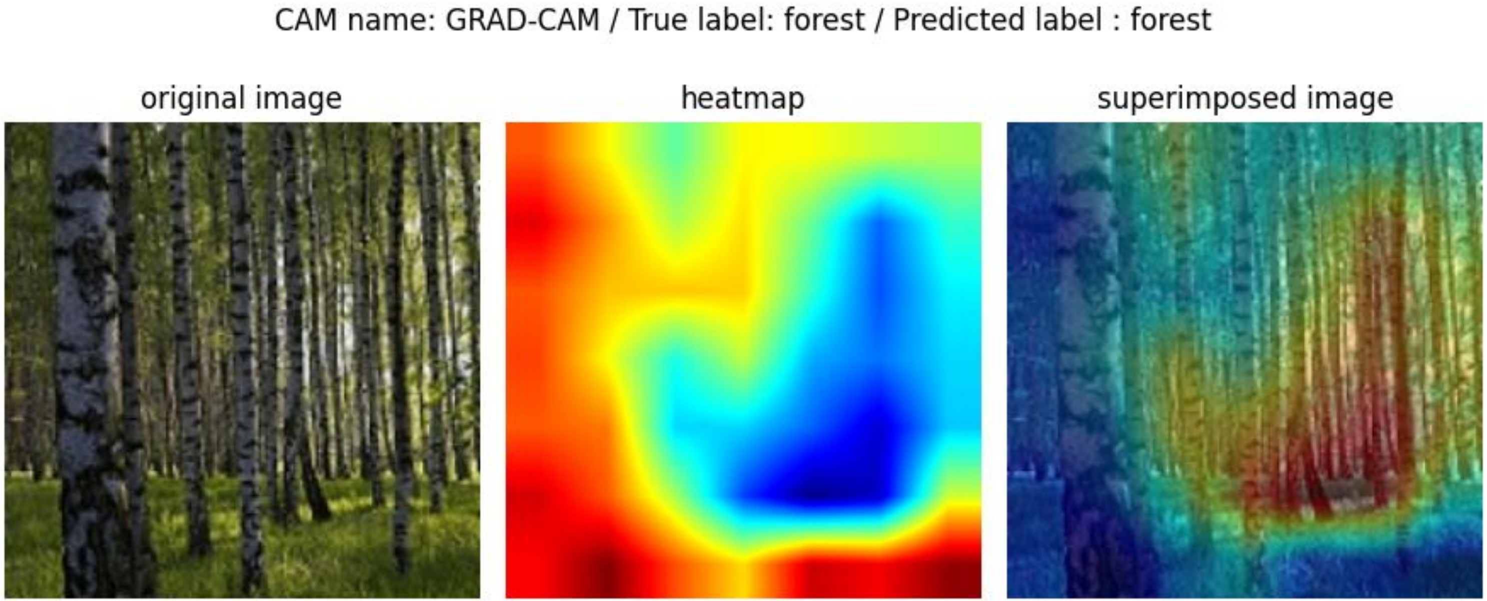
$$w_k^c = \frac{1}{Z}\sum_i\sum_j \mathrm{ReLU}\left(\frac{\partial y^c}{\partial A_{ij}^k}\right) \longrightarrow w_k^c = \sum_i\sum_j \alpha_{ij}^{kc} \cdot \mathrm{ReLU}\left(\frac{\partial y^c}{\partial A_{ij}^k}\right)$$

where the α_{ij}^{kc} measure the importance of each individual unit of a feature map

- α_{ij}^{kc} = 1 / Z , GradCAM++ reduces to the formulation for Grad-CAM

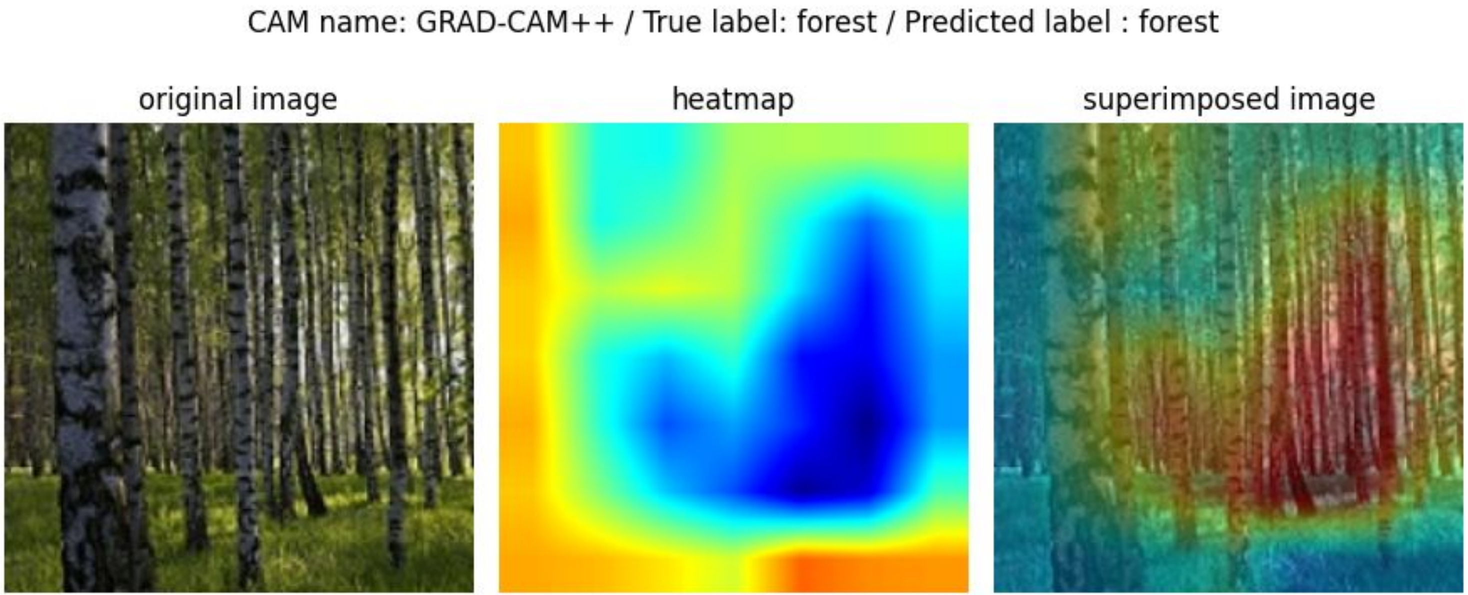
Source: Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." 2018 IEEE winter conference on applications of computer vision (WACV). IEEE, 2018 [2]

Interpretations on Intel Image Classification


CAM name: GRAD-CAM / True label: forest / Predicted label : forest
original image
heatmap
superimposed image

Source: https://www.kaggle.com/datasets/puneet6060/intel-image-classification/data [24]

Interpretations on Intel Image Classification



CAM name: GRAD-CAM++ / True label: forest / Predicted label : forest

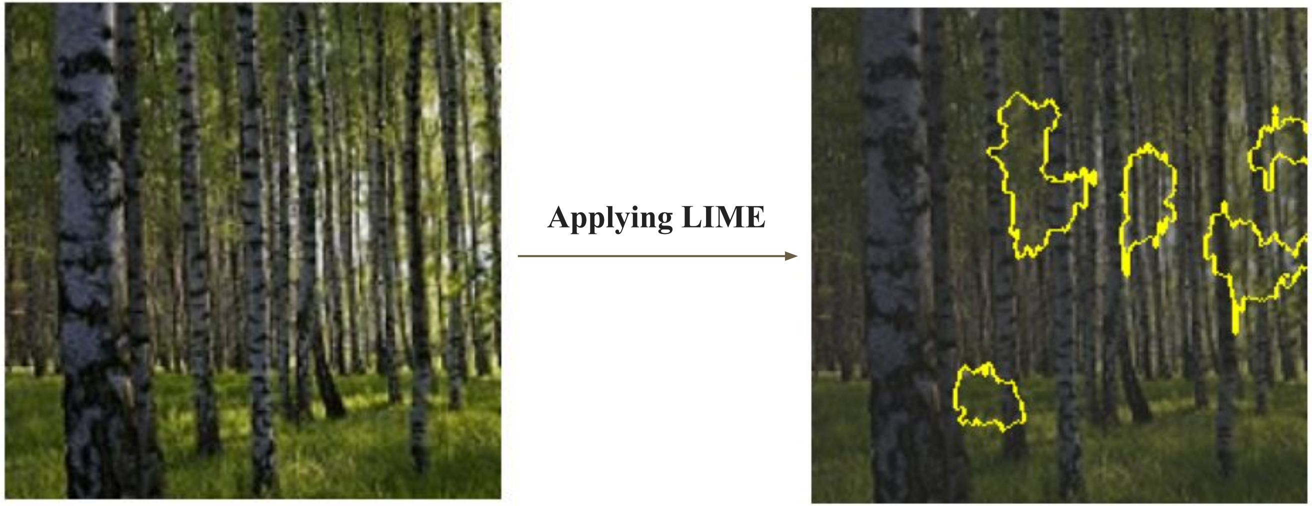
original image

heatmap

superimposed image

Source: https://www.kaggle.com/datasets/puneet6060/intel-image-classification/data [24]

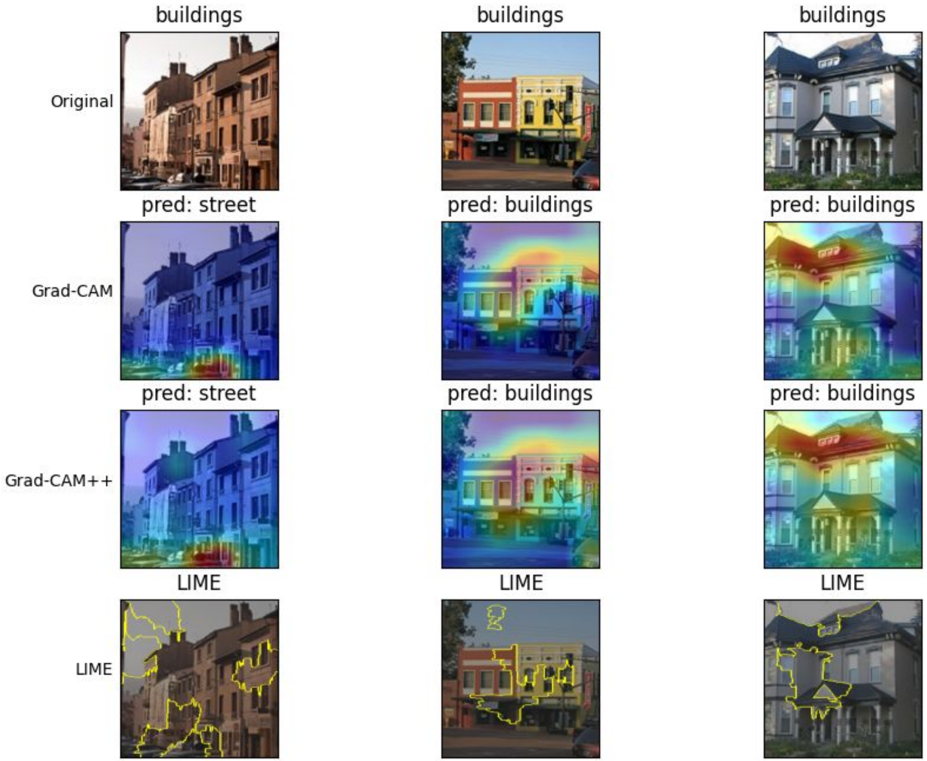
Interpretations on Intel Image Classification



Applying LIME

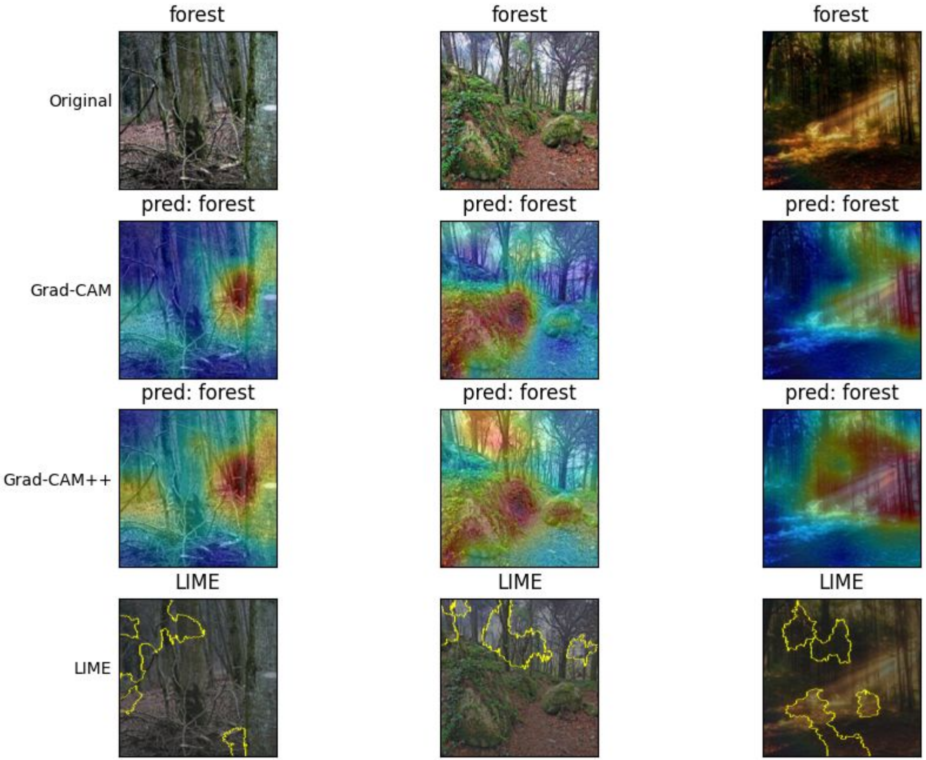
Source: https://www.kaggle.com/datasets/puneet6060/intel-image-classification/data [24]

Interpretations on Intel Image Classification



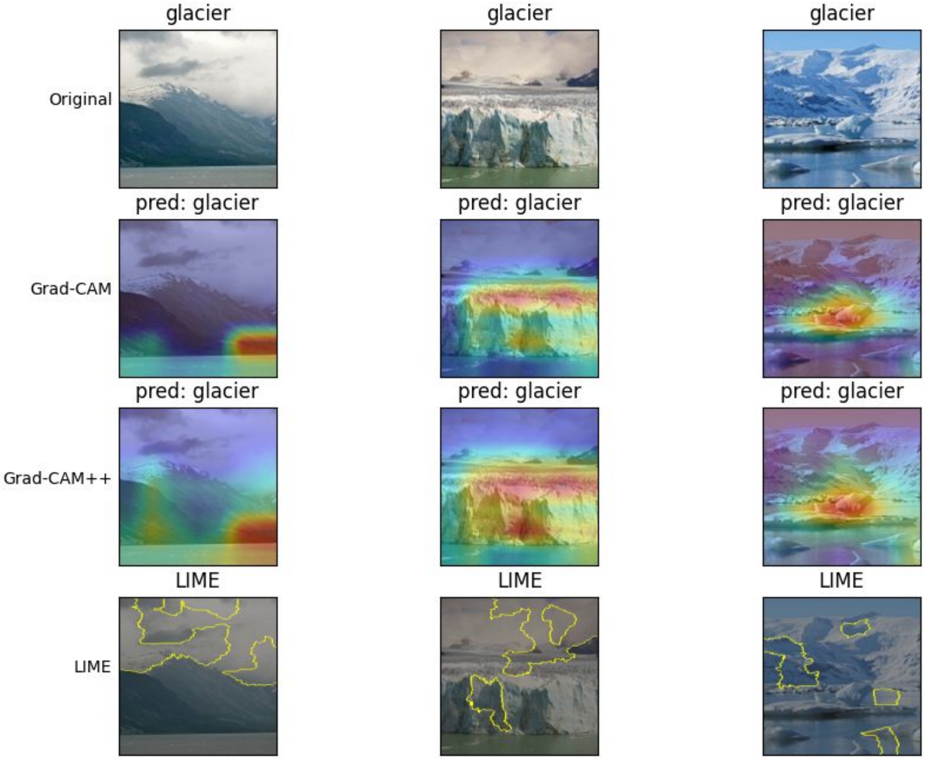
Source: https://www.kaggle.com/datasets/puneet6060/intel-image-classification/data [24]

Interpretations on Intel Image Classification



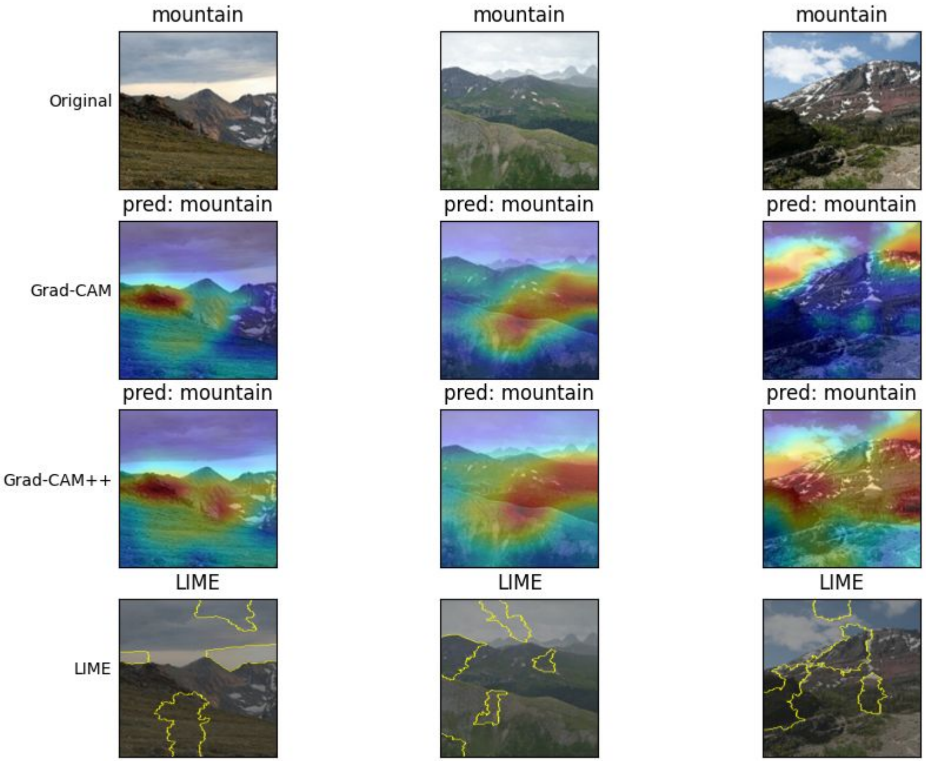
Source: https://www.kaggle.com/datasets/puneet6060/intel-image-classification/data [24]

Interpretations on Intel Image Classification



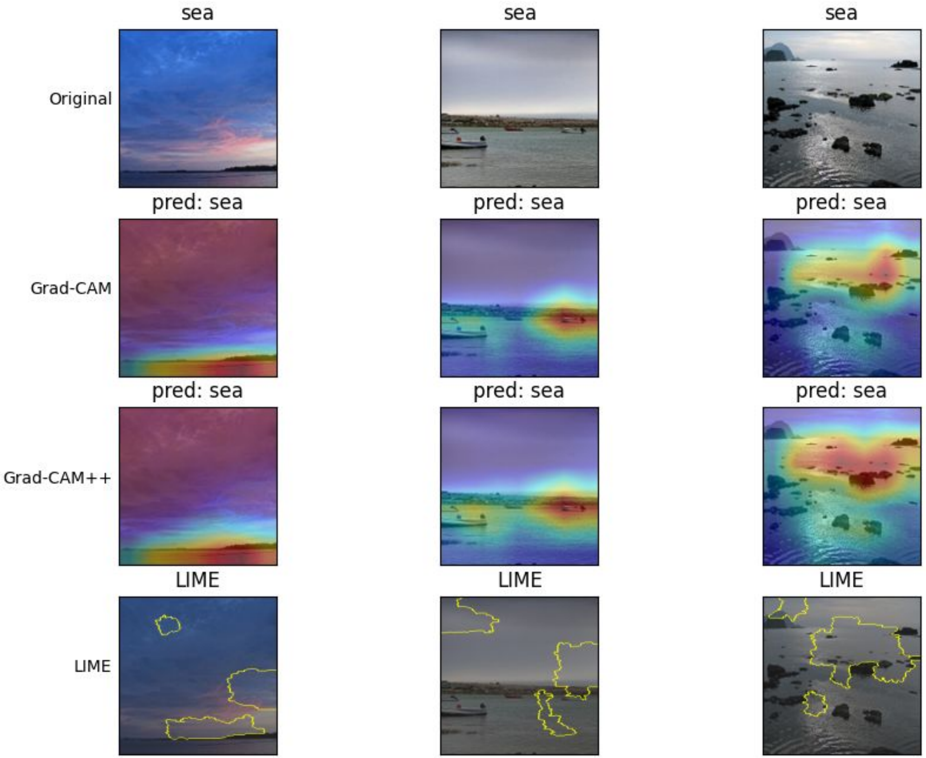
Source: https://www.kaggle.com/datasets/puneet6060/intel-image-classification/data [24]

Interpretations on Intel Image Classification



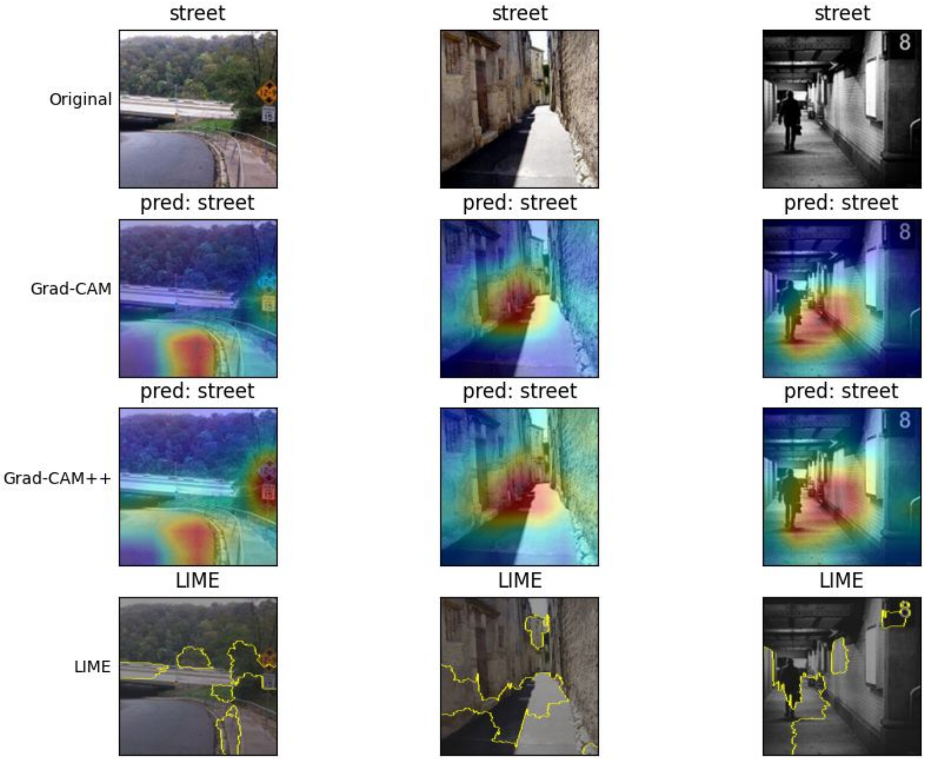
Source: https://www.kaggle.com/datasets/puneet6060/intel-image-classification/data [24]

Interpretations on Intel Image Classification



Source: https://www.kaggle.com/datasets/puneet6060/intel-image-classification/data [24]

Interpretations on Intel Image Classification



Source: https://www.kaggle.com/datasets/puneet6060/intel-image-classification/data [24]

References

[1] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.

[2] Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." 2018 IEEE winter conference on applications of computer vision (WACV). IEEE, 2018.

[3] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

[4] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).

[5] https://christophm.github.io/interpretable-ml-book/shapley.html

[6] https://deepfindr.github.io/

[7] https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3

[8] Abir, W. Hasan, et al. "Detecting Deepfake Images Using Deep Learning Techniques and Explainable AI Methods." *Intelligent Automation & Soft Computing.* (2023): 2151-2169.

[9] Chaddad, Ahmad, et al. "Survey of explainable AI techniques in healthcare." *Sensors* 23.2 (2023): 634.

References

[10] Singh, Bhuvanesh, and Dilip Kumar Sharma. "SiteForge: Detecting and localizing forged images on microblogging platforms using deep convolutional neural network." Computers & Industrial Engineering 162 (2021): 107733.

[11] Singh, Bhuvanesh, and Dilip Kumar Sharma. "Image forgery over social media platforms-A deep learning approach for its detection and localization." 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, 2021.

[12] Malolan, Badhrinarayan, Ankit Parekh, and Faruk Kazi. "Explainable deep-fake detection using visual interpretability methods." 2020 3rd International Conference on Information and Computer Technologies (ICICT). IEEE, 2020.

[13] Kuchumova, E., Martínez-Monterrubio, S. M., & Recio-Garcia, J. A. (2023). STEG-XAI: explainable steganalysis in images using neural networks. Multimedia Tools and Applications, 1-18

[14] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.

[15] Cogranne, R., Giboulot, Q., & Bas, P. (2020, December). ALASKA# 2: Challenging academic research on steganalysis with realistic images. In 2020 IEEE International Workshop on Information Forensics and Security (WIFS) (pp. 1-5). IEEE.

[16] Dwivedi, Rudresh, et al. "An Efficient Ensemble Explainable AI (XAI) Approach for Morphed Face Detection." arXiv preprint arXiv:2304.14509 (2023).

[17] Chollet, François. "Xception: Deep learning with depthwise separable convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

References

[18] Rössler, Andreas, et al. "FaceForensics++: Learning to Detect Manipulated Facial Images." ICCV. Vol. 1. No. 2. 2019.

[19] Szegedy, Christian, et al. "Inception-v4, inception-resnet and the impact of residual connections on learning." Proceedings of the AAAI conference on artificial intelligence. Vol. 31. No. 1. 2017.

[20] https://figshare.com/articles/dataset/Face_Research_Lab_London_Set/5047666

[21] George, Anjith, et al. "Biometric face presentation attack detection with multi-channel convolutional neural network." IEEE transactions on information forensics and security 15 (2019): 42-55.

[22] C. Chen, A. Dantcheva, T. Swearingen, A. Ross, "Spoofing Faces Using Makeup: An Investigative Study," Proc. of 3rd IEEE International Conference on Identity, Security and Behavior Analysis (ISBA 2017), (New Delhi, India), February 2017.

[23] https://www.shedloadofcode.com/blog/understanding-explainable-ai-for-classification-regression-and-clustering-with-python/

[24] https://www.kaggle.com/datasets/puneet6060/intel-image-classification/data

[25] https://censius.ai/blogs/global-local-cohort-explainability

[26] https://medium.com/@mohamedchetoui/grad-cam-gradient-weighted-class-activation-mapping-ffd72742243a

Thank You