TECHNO INDIA UNIVERSITY
WEST BENGAL

INTERNSHIP CERTIFICATE

This internship certificate has been awarded to

MAHEAK DAVE

Roll number: **211001001316**, **B. Tech** in **Computer Science Engineering** of **Techno India University, West Bengal** for successfully completing his internship of **One (1) Year (07.09.2024 to 06.09.2025)** for the DYSL-CT Project titled: **"Explaining the Cognitive Algorithms (XCA)"** and submitting his internship report on the topic of **"Iterative Constructive Perturbation and Firewall: Towards Robust and Explainable Artificial Intelligence"** under the guidance of **Prof. (Dr.) Debasis Chaudhuri**, Professor, DRDO Project Executive Lab, CSE Dept., TIU, West Bengal and **Mr. Manish Pratap Singh**, Director, DRDO Young Scientist's Lab-Cognitive Technologies, (DYSL-CT)-Chennai.

Place: Kolkata
Date: 8th Oct. 2025

Prof. (Dr.) Debasis Chaudhuri
Professor
DRDO Project Executive Lab, CSE Dept.
Techno India University, West Bengal

Prof. (Dr.) Debasis Chaudhuri
Principal Investigator
CSE Department
TIU, West Bengal

Shri Manish Pratap Singh
Director
DYSL-CT, Chennai
DRDO

मनीष प्रताप सिंह / MANISH PRATAP SINGH
निदेशक / Director
डीआरडीओ युवा वैज्ञानिक प्रयोगशाला - सीटी
DRDO Young Scientist Lab - CT
भारत सरकार, रक्षा मंत्रालय
Govt. of India, Min. of Defence
तरमणी, चेन्नै - 600 113.
Taramani, Chennai - 600 113.

TECHNO INDIA UNIVERSITY
WEST BENGAL

EM 4, Sector - V, Salt Lake, Kolkata – 700091, West Bengal, India
Phone: (91) 33-2357-6163/64/84/2658/1094, Fax: +91 33 2357 1097

To
Shri Maheak Dave
41 Elgin Road
Neelkamal Building, Flat – 8/A,
Kolkata – 700020

Subject: Temporary appointment for the paid DRDO project "XCA" internship.
Student ID: 211001001316

Dear Mr. Dave

I'm happy to inform you that you have been given the opportunity to work as a **Student of Internship** for the paid DRDO-sponsored project "***Explaining the Cognitive algorithms* (*XCA*)**" for a period of six months starting on dated 07.09.24. Without any further benefits, you will receive consolidated of **Rs. 7,500/- (Rs. Seven Thousand Five Hundred only**) every month. Please join the undersigned as soon as feasible.

With best wishes

2.9.24

Prof. (Dr.) Debasis Chaudhuri
CSE Dept
PI – project (XCA)
Techno India University
West Bengal

Prof. (Dr.) Debasis Chaudhuri
Principal Investigator
CSE Department
TIU, West Bengal

TECHNO INDIA UNIVERSITY
WEST BENGAL
EM 4, Sector - V, Salt Lake, Kolkata – 700091, West Bengal, India
Phone: (91) 33-2357-6163/64/84/2658/1094, Fax: +91 33 2357 1097

To
The Chief Finance Officer
Accounts Department
Techno India University
EM-4, EM Block, Sector – V
Kolkata - 700091

Subject: Extension of Temporary appointment for the paid DRDO project "XCA" internship.
Student Name: Shri Maheak Dave, Student ID: 211001001316

Dear Sir,

I'm happy to inform you that the DRDO-sponsored project "***Explaining the Cognitive algorithms* (*XCA*)**" (Contract No.: DYSL-CT/MMG/CARS/CS/23-24/01 dated 30/01/2024) has been sanctioned for a period of one and half years. Total cost of the project is **Rs. 31,88,360/-**. I am gladly informed you that Shri Maheak Dave, (Student ID: 211001001316) has been appointed as a **Student Internship** for the paid DRDO-sponsored project "***Explaining the Cognitive algorithms* (*XCA*)**" for a period of six months starting on dated 07.09.24. The director of DYSL-CT, Chennai, has requested an extension for his based on the satisfactory progress observed in the aforementioned project work. Therefore, the length of his paid student internship may be extended based on his performance. He will receive a total of **Rs. 7,500/-** (**Rs. Seven thousand five hundred only**) per month till the project is completed, with no further advantages. On 06.09.25, the project will be closed. Please take the necessary action to support his stipend.

With regards

12.2.25

Prof. (Dr.) Debasis Chaudhuri
Principal Investigator
CSE Department
TIU, West Bengal

Prof. (Dr.) Debasis Chaudhuri
CSE Dept
PI – project (XCA)
Techno India University
West Bengal

CC: HR, TIU, West Bengal, Kolkata

RESEARCH INTERNSHIP
REPORT

ON

Explaining the Cognitive Algorithms (XCA) project

AT

DRDO Project Executive Lab Techno India University

SUBMITTED BY

Maheak Dave

Research Intern

Under the Guidance of

Aryan Pareek (Research Scientist)

DECLARATION

I hereby declare that the work presented in this report entitled "Explaining the Cognitive Algorithms (XCA)" has been carried out by me, Maheak Dave, during my tenure as a research intern at the DRDO Project Executive Lab, Techno India University. The work described is original and is the result of my own investigations carried out under the supervision Dr. Debasis Chaudhuri at the DRDO Project Executive Lab. Wherever material from other sources has been used, due acknowledgement has been made in the text. I accept responsibility for the authenticity of the work and for any errors that may remain.

Date: 6.9.25

Signature: _____

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Debasis Chauduri, Project Lead, for his invaluable guidance, constant support, and insightful feedback throughout my research internship on the Explaining the Cognitive Algorithms (XCA) project. His direction greatly shaped the work presented in this report.

I am thankful to Aniket Kumar Singh (Ex-Research Scientist), for his technical assistance, patient explanations, and constructive discussions that helped resolve challenging problems during the project. I also extend my thanks to Aryan Pareek (Research Scientist), for his practical help in experiments, thoughtful suggestions, and collaborative spirit.

My thanks also go to the entire team at the DRDO Project Executive Lab, Techno India University, for providing a motivating research environment and access to necessary resources and facilities. Finally, I appreciate the encouragement and support of my peers and the administrative staff who made this internship a productive and rewarding experience.

Maheak Dave

Research Intern – DRDO Project Executive Lab, Techno India University

Explaining the Cognitive Algorithms (XCA) project.

Organization Profile

Defence Research & Development Organisation (DRDO) is the premier research and development body under the Department of Defence Research and Development, Ministry of Defence. DRDO is dedicated to the development of state-of-the-art defence systems, technologies and solutions to strengthen India's self-reliance in defence capabilities. Through an extensive network of laboratories, DRDO carries out basic and applied research, design and development of weapon systems, sensors, materials, electronics, computation and life-support technologies required by the armed forces.

DRDO's activities span a broad range of disciplines including aeronautics, armaments, combat vehicles, electronics, instrumentation, engineering systems, missiles, materials, naval systems, advanced computing, simulation and life sciences. The organisation works closely with academic institutions, industry partners and other government agencies to transition technologies into deployable products and systems that meet the operational requirements of the Defence Services.

Vision

To make India self-reliant by establishing world-class science and technology base and by equipping the Defence Services with internationally competitive systems and solutions. To establish the *DRDO Project Executive Lab at Techno India University* as a centre of excellence for cutting-edge research and innovation in cognitive algorithms and allied AI technologies; to cultivate a collaborative academic–industry ecosystem that develops secure, interpretable, and deployable AI solutions for defence and civilian applications; and to empower students and researchers through hands-on training, ethical practice, and strong translational partnerships that contribute to India's strategic technological self-reliance.

TABLE OF CONTENTS

INTRODUCTION

Neural networks have become the backbone of modern machine learning, enabling state-of-the-art results across tasks such as image recognition, language understanding, and autonomous control. However, while their predictive capabilities are impressive, neural networks often operate as opaque systems with limited interpretability. This lack of transparency not only hinders trust and understanding but also restricts our ability to optimize and improve models in a principled manner.

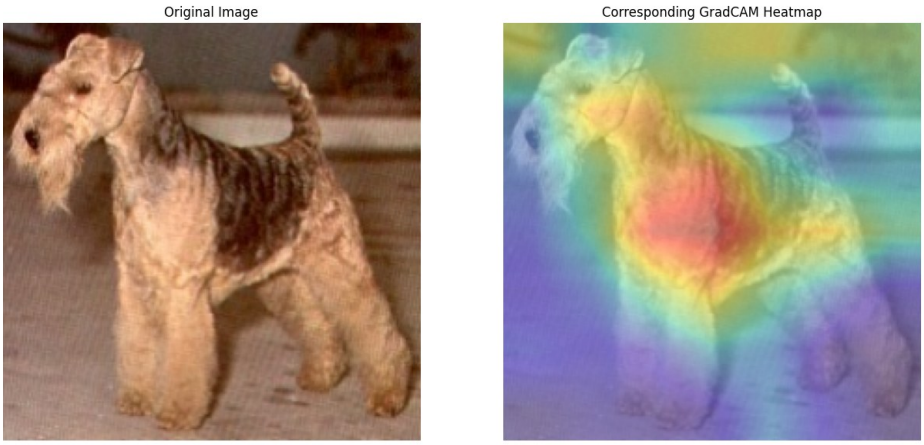


Figure 1. Grad-CAM visualization.

Explainable cognitive AI aims to bridge this gap by producing model outputs that are accompanied by intelligible, faithful information about *why* a decision was made. A large and practically useful family of explanation techniques for convolutional neural networks are gradient-based class localization methods. Class Activation Mapping (CAM) [1] and its gradient-aware extensions (for example, Grad-CAM [2] and Grad-CAM++ [3]) as depicted in figure 1. produce saliency or heat-maps that highlight image regions most influential for a particular class score. Conceptually, these methods use gradient information assign importance weights to spatial feature maps, and then combine those maps to produce a class-specific localization. Since they leverage the model's internal activations and gradients, CAM-style methods are inexpensive to compute and align naturally with the network's learned representation; consequently, they have become a de facto standard for visual explanation in many applied settings. However, relying on gradients also makes these methods sensitive to the same fragilities that affect the underlying network: noisy activations or subtle re-weightings of channels can change the saliency distribution, sometimes without large changes in the final prediction.

An alternative and complementary class of explanation techniques are *perturbation-based* methods. Instead of using analytic gradients, these approaches probe the model by systematically modifying the input and measuring how the prediction changes. Ablation-style CAMs (sometimes called AblationCAM [4] or related perturbation CAM variants) operate on this principle: they mask, occlude, or perturb spatial regions (or feature channels) and record the resulting effect on class scores to build an importance map. Perturbation methods tend to produce high-fidelity, model-agnostic explanations because they measure causal influence directly. In practice, hybrid approaches that combine gradient signals with localized perturbations seek to balance fidelity, efficiency, and stability.

The existence of perturbations is not limited to benign interpretability probes: adversarial perturbations are deliberately constructed, often imperceptible, input modifications that cause large changes in a model's output. Classic gradient-based attacks such as the Fast Gradient Sign Method (FGSM) [5] and its iterative variant, Projected Gradient Descent (PGD) [6] exploit the same gradient information used by explanation algorithms to shift predictions while remaining within a small norm ball around the original image. Importantly for explainability research, adversarial perturbations can alter saliency maps dramatically, even when the predicted class remains unchanged. This phenomenon undermines the usefulness of explanations: a saliency map that is easily perturbed cannot serve as a reliable explanation for human decision-making. Moreover, many explainability methods themselves rely on gradients, adversarial strategies that target gradients can simultaneously degrade both prediction and explanation.

While existing literature utilizes perturbations for adversarial attacks and explainability, it could also be used constructively to derive highly discriminative features and consequently use it to improve neural network performance, improve robustness against adversarial attacks and produce more refined explainations.

In computer-vision based deep learning, a *perturbation* is any deliberate modification to an input image ranging from small additive noise or pixel-level changes to localized occlusions, geometric transforms, or more semantic edits applied to probe or influence a model's behaviour. Formally, a perturbed image x' is often written as $x' = x + \delta$, where δ is constrained (e.g., $\| \delta \|_p \leq \epsilon$) when the goal is to keep the change imperceptible; other perturbations (occlusion patches, blurring, cropping) intentionally alter local content and need not be norm-bounded. Perturbations serve two main roles in vision research: as benign probes (perturbation-based attribution and robustness testing) that reveal which pixels or regions causally affect predictions, and as adversarial attacks that are optimized to induce misclassification or to manipulate saliency maps. Since perturbations change internal activations and gradients, they can both expose model vulnerabilities and help evaluate the fidelity and stability of explanations.

Gradient based adversarial perturbation techniques are a subclass of perturbation techniques in computer vision, which utilizes gradient of Deep Neural Network (DNN) to perturb the input sample in a way that increases the loss value for that input sample, consequently harming model performance. The most basic example of such a technique is FGSM. While other, more effective techniques such as JSMA (Jacobian Saliency Map Attack) [7], EOT-FGSM (Expectation Over Transformation + FGSM) [8], and many more exists, this report will mainly focus on FGSM and its variants.

FGSM

Given an input sample x and corresponding label y, a loss function $L(\theta, x, y)$, where θ represents model parameters. The input is then perturbed as:

$$x' = x + \epsilon \cdot sign(\nabla_x L(\theta, x, y)) \qquad (1)$$

Here ϵ is the perturbation constant, which in most cases is a very small value. Only sign of the gradients is taken, so as to constrain the perturbation, and make it imperceptible to the human eye.

FGSM constraints the gradients of the loss function w.r.t the input sample and uses to it to effectively perturb the input so as to fool the network. An example is depicted in figure 2.

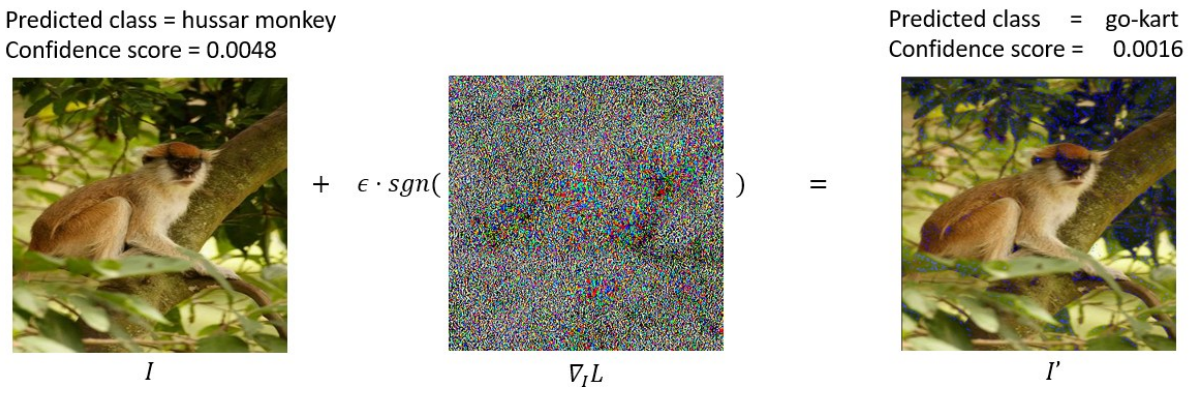


Figure 2.Example depicting effect of FGSM.

Defensive techniques against adversarial perturbations is a thoroughly researched topic and can be categorized as:

Empirical techniques:

These techniques exploit the nature of neural networks and perturbations techniques, to defend against such adversarial perturbations, there are three well distinct types into which an empirical technique can fall into, which are:

A. *Adversarial Training*: Involves training a neural network on both normal and perturbed samples so as to handle any perturbed sample during test time. However, using this type of technique, it is difficult to account for a large variety of perturbation techniques.
B. *Auxiliary Networks*: The main neural network works along with an auxiliary neural network which has been trained to detect or transform any perturbed input that has been passed through it. Training the auxiliary neural network increases the computational cost with no guarantee of defending against every type of adversarial sample.
C. *Gradient Obfuscation*: Since many adversarial techniques exploit gradient to perturb the sample, this type of techniques obfuscates or diminishes the gradient values during the backpropagation before the input sample stage. This type of technique can effectively weaken any gradient based adversarial techniques, but is infamous to negatively affect model performance [9,10].

Certified techniques:

Certified defences provide provable**,** mathematical guarantees that a model's prediction cannot be changed by any adversarial perturbation within a specified threat model. Unlike empirical defences (which may be broken by stronger or adaptive attacks), certified methods return either a certified radius (the input is provably robust up to that radius) or a proof of robustness/violation for a specific instance. The most well known example of such a technique is *Randomized Smoothing* [11].

Proposed Algorithm

During the tenure of the internship, a combination of auxiliary network and gradient obfuscation was explored. The algorithm involved training an autoencoder having only particular activation functions which introduces vanishing gradients, such as the Sigmoid function, or the TanH function. The autoencoder reproduces the input fed into it as its output and the output is then fed into the target model. In case of any gradient based attacks, the vanishing gradient effect introduced by the activation functions in the autoencoder weakens any such attack effectively enough to not cause harm to the main model, thus mitigating it. This type of algorithm as visualized in figure 3. enables a plug-and-play style of defence where the autoencoder could be used along with any such model assuming that both the models have been trained on the same dataset.

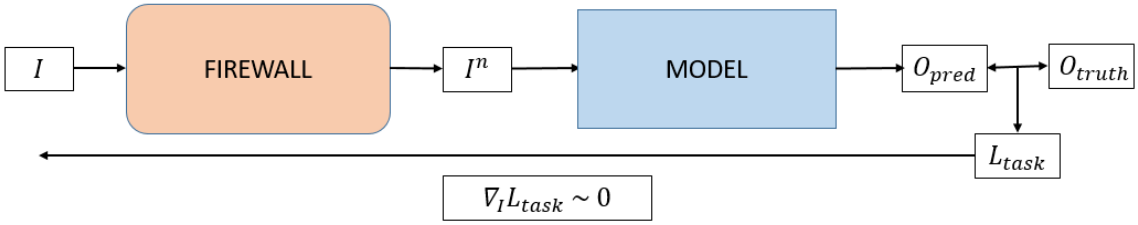


Figure 3. Flowchart of the firewall algorithm

Results

Initial testing of this algorithm was done on the MNIST dataset [12] using the LeNet pretrained classifier [13]. The result also showed no negative impact on the main model performance by the autoencoder.

Table 1. Initial testing result of the firewall algorithm.

	Without Firewall	**With Firewall**
No. of successful attacks.	37.94%	6.9%
No. of un-successful attacks.	62.06%	93.1%

While the initial testing on the small scale setup showed promising results, when tested on a larger scale using colored images, the autoencoder started to struggle in replicating the input samples, which then affected the performance of the main model. Various types of autoencoder were then tested to perfectly reproduce the input samples, however eventually all efforts were rendered futile, since no type of neural can give 100% positive results according to the "no-free lunch" theorem [14], and since it is guaranteed that few pixels in the output of the autoencoder will be perturbed, the autoencoder will almost always negatively impact the main model's performance, atleast when scaled up.

Self-distillation via Iterative Constructive Perturbations

The term 'constructive perturbations' refers to perturbation techniques which enhances a model's performance. During the internship, a concept was developed, called the *Iterative Constructive Perturbations,* inspired from FGSM, it uses the gradient information to perturb an input sample in a way that it minimizes the loss function value when the perturbed sample is fed into the model. Since, the technique uses parameters gradients of the model to enhance input sample, such perturbed inputs can then be interpreted as a more ideal input sample for the given model as compared to its original counterpart.

Iterative Constructive Perturbation

Iterative Constructive Perturbation (ICP) as expressed in Eq. 2. is a gradient-based input optimization technique designed to iteratively refine input representations to enhance model performance. Unlike single-step methods such as FGSM, ICP applies multiple iterations of gradient-based adjustments directly to the input data. This process systematically aligns the inputs with the model's learned feature space, reducing classification errors and enhancing class separability. Each iteration involves optimizing the input representation by utilizing full gradient information, ensuring precise and stable refinements.

$$x_t = x_{t-1} - (\epsilon \cdot \nabla_{x_{t-1}} L(\theta, x, y)) \qquad (2)$$

ICP addresses a fundamental challenge in adversarial settings, the overlap of data across decision boundaries. By iteratively modifying inputs, ICP tightens intra-class clustering while increasing inter-class separation as depicted in Fig. 5. This creates a more robust feature space, enabling the model to make confident predictions even in the presence of adversarial perturbations. The approach treats step sizes as dynamic learning rates, ensuring nuanced adjustments that align data points with their intended classes.

Proposed Methodology

Building upon ICP, a self-distillation [15] framework as illustrated in Fig. 4. enhances its efficacy by aligning feature representations between the original and ICP-modified inputs during training. Self-distillation enables the model to act as both teacher and student, refining its own predictions through intermediate-layer activations.

Iterative methods, such as ICP, offer a computationally efficient alternative to traditional adversarial defences strategies, which are often resource-intensive and computationally demanding. The efficiency of iterative methods lies in their ability to achieve robust results with fewer computational resources, making them highly suitable for deployment in environments with limited computational capabilities. This characteristic broadens the accessibility of advanced AI solutions, enabling robust defences to be applied across diverse scenarios, from resource-constrained edge devices to large-scale AI systems.

A key advantage of iterative methods is their ability to refine input representations for improved alignment with the model's learned feature space. By iteratively fine-tuning inputs, these methods ensure that the model's decision-making is based on data that is more representative, structured, and accurate. This enhanced alignment not only reduces the likelihood of errors in predictions but also leads

to more informed and reliable decision-making processes. As a result, the model benefits from improved generalization and better performance, particularly in scenarios involving noisy or adversarial data.

Moreover, iterative approaches address the inherent limitations of single-step perturbation methods by incorporating multiple refinements. Single-step techniques often fall short in countering complex adversarial attacks due to their simplistic nature. In contrast, iterative methods progressively enhance input robustness through a series of refinements, effectively neutralizing adversarial vulnerabilities. This progressive approach not only mitigates the impact of attacks but also enhances the model's overall interpretability by fostering a deeper understanding of how inputs are transformed and processed. Consequently, iterative techniques strengthen both the reliability and robustness of AI systems, making them a vital tool in the development of secure and trustworthy machine learning models.

The steps of the ICP driven self-distillation is as described below:

1. *Baseline Training Phase:* Moreover, iterative approaches address the inherent limitations of single-step perturbation methods by incorporating multiple refinements. Single-step techniques often fall short in countering complex adversarial attacks due to their simplistic nature. In contrast, iterative methods progressively enhance input robustness through a series of refinements, effectively neutralizing adversarial vulnerabilities. This progressive approach not only mitigates the impact of attacks but also enhances the model's overall interpretability by fostering a deeper understanding of how inputs are transformed and processed. Consequently, iterative techniques strengthen both the reliability and robustness of AI systems, making them a vital tool in the development of secure and trustworthy machine learning models.
2. *Alignment of intermediate feature representation:* The modified input I' is passed through the network again to obtain updated intermediate feature maps F_i' . Layer-wise distillation losses L_{dist}^{i} are computed by comparing feature maps F_i with F_i'. In this paper, the mean-squared error (MSE) loss has been utilised as the distillation loss metric such as expressed in Eq. 3.

$$L_{dist}^{i} = MSE(F_i, F'_i) \qquad (3)$$

3. *Combined training loss with cosine decay:* For the initial k baseline epochs, there is no distillation. Only after k baseline epochs, the self-distillation phase begins. The balance between task specific loss L_{task} and distillation losses L_{dist}^{i} is controlled by a parameter α_e , which evolves with each epoch e. The total loss is defined as:

$$L_{total} = \alpha_e \cdot L_{task} + (1 - \alpha_e) \cdot \sum_{i=1}^{n} L_{dist}^{i} \qquad (4)$$

With α_e calculated as:

$$\alpha_e = \begin{cases} 1, & e \leq k \\ \cos\left(\frac{\pi \cdot (e-k)}{2 \cdot (E-k)}\right), & e > k \end{cases} \qquad (5)$$

Here, E is the total number of epochs for training. For baseline epochs, $\alpha_e = 1$ results in no weightage for the self-distillation task. After k epochs, α_e follows cosine-decay scheduling method, inspired by principles in learning rate decay strategies. Thus, the use of the parameter α_e optimally balances between task performance and robust feature representation by adjusting the model's focus over time.

Here, E is the total number of epochs for training. For baseline epochs, $\alpha_e = 1$ results in no weightage for the self-distillation task. After k epochs, α_e follows cosine-decay scheduling method, inspired by principles in learning rate decay strategies. Thus, the use of the parameter α_e optimally balances between task performance and robust feature representation by adjusting the model's focus over time.

In summary, the framework processes both the original input I and the ICP-modified input I', aligning feature representations through layer-wise distillation losses. By leveraging ICP without a fixed perturbation limit, this approach effectively aligns input with learned features.

The cosine-decayed self-distillation approach further smoothens model convergence, enhancing performance and efficiency for real-world applications. One of the most significant advantages of certain defensive architectures, such as lightweight autoencoder-based models, is their ability to maintain the performance metrics of the original model without degradation

4. *ICP Variants:* Since Iterative Constructive Perturbation (ICP) is basically a process of gradient descent on input samples, it is possible to extend it based on contemporary optimization techniques. To see the influence of various modern optimization algorithms based ICP alternatives, we considered two alternatives: Adam-ICP and AdEMAMix-ICP. These alternatives take their name after the Adam and AdEMAMix [16] optimization techniques, respectively, and are motivated by the potential in improving the effectiveness of ICP.

Experimental Results

To rigorously evaluate the effectiveness of our proposed ICP-based self-distillation framework, we conducted extensive experiments across three distinct computer-vision tasks: image classification, image reconstruction, and semantic segmentation. This multi-task evaluation was chosen to validate our theoretical insights in varied settings classification tests the framework's ability to refine decision boundaries, reconstruction assesses its impact on generative fidelity, and segmentation examines how well input refinements translate to precise pixel-level predictions. In each case, we compared vanilla models against their ICP-refined, self-distilled counterparts to quantify gains in accuracy, F1 score, Structural Similarity Index Measure (SSIM) [17] as expressed in Eq. 12:

$$\boldsymbol{SSIM}(\boldsymbol{x},\boldsymbol{y}) = \frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)} \qquad (6)$$

Where $\boldsymbol{\mu_x}, \boldsymbol{\mu_y}$ are the mean of $\boldsymbol{x}, \boldsymbol{y}$ respectively, $\boldsymbol{\sigma_{xy}}$ is the sample covariance of $\boldsymbol{x}$ and $\boldsymbol{y}$, and $\boldsymbol{\sigma_x^2}, \boldsymbol{\sigma_y^2}$ are the sample variance of the two variables.

We also consider **F**réchet **I**nception **D**istance (FID) [18] for image generation tasks. To ensure a fair and consistent comparison, all models were trained for a fixed 100 epochs using identical training hyperparameters (learning rate schedules, batch sizes, optimizer settings) and data augmentation pipelines. We also controlled for model capacity by matching parameter counts between baseline and ICP-augmented variants, isolating the benefits of our optimization scheme from mere increases in architecture size. Furthermore, to assess robustness, we evaluated each model on both clean test sets and perturbed inputs introducing controlled noise and occlusions to measure performance degradation under challenging conditions. This rigorous, uniform experimental protocol underscores the broad applicability and practical advantages of integrating input refinement with self-distillation across diverse vision tasks.

1. *Hyperparameter Tuning using image classification:* We performed experiments for image classification task on the CIFAR-100 [19] dataset using a modified ResNet20. As shown in

Table 2, the optimal configuration for ICP-based self-distillation was found to be at $k = 25$ and $T = 5$, with weighted feature maps enabled for self-distillation. This configuration led to an improvement in accuracy for all ICP variants over the control baseline ($k = 100$). Specifically, the AdEMAMix-ICP variant achieved the highest recorded accuracy, which was 19.06% higher than the control baseline. Additionally, improvements in F1-score further validated the efficacy of our approach in refining feature representations. Consequently, we adopted the hyperparameters ($k = 25$, $T = 5$, *Weighted = True*) for all subsequent experiments to maintain consistency across tasks, while also comparing the performance of different ICP variants (SGD-ICP, Adam-ICP, and AdEMAMix-ICP).

Table 2.Ablation study on SGD-ICP, Adam-ICP and AdEMAMix-ICP with different k, T, and weighting schemes.

ICP variant	Baseline epochs	Iterations	Weighted Feats	Acc (%).	F1	Training time	ICP variant	Baseline epochs	Iterations	Weighted Feats	Acc(%).	F1	Training time
SGD-ICP	0	5	False	40.23	0.399	39.84	Adam-ICP	0	5	False	38.59	0.378	39.55
			True	39.97	0.393	39.88				True	37.77	0.373	39.51
		10	False	39.68	0.387	45.98			10	False	38.77	0.380	**46.31**
			True	39.08	0.385	**46.25**				True	37.70	0.370	46.06
	25	5	False	**40.90**	**0.402**	36.97		25	5	False	40.80	0.404	36.54
			True	39.64	0.385	36.96				True	41.31	**0.409**	37.48
		10	False	40.29	0.396	42.03			10	False	40.39	0.399	41.98
			True	40.60	0.400	42.03				True	**41.32**	0.405	41.50
	50	5	False	38.95	0.383	34.28		50	5	False	38.39	37.40	33.40
			True	38.83	0.377	34.38				True	36.96	0.356	34.23
		10	False	39.11	0.389	37.64			10	False	39.32	0.386	37.26
			True	38.71	0.385	37.64				True	39.38	0.390	36.98
	75	5	False	32.25	0.313	31.46		75	5	False	29.33	0.282	31.39
			True	31.19	0.304	31.69				True	29.43	0.285	31.00
		10	False	34.55	0.335	33.27			10	False	28.01	0.258	32.82
			True	34.40	0.336	33.09				True	28.88	0.272	32.97

ICP variant	Baseline epochs	Iterations	Weighted Feats	Acc(%).	F1	Training time
Ademamix-ICP	0	5	False	38.93	0.385	40.19
			True	40.51	0.394	39.87
		10	False	37.82	0.370	**47.90**
			True	36.30	0.352	46.98
	25	5	False	40.80	0.399	37.36
			True	**41.99**	**0.414**	37.37
		10	False	41.27	0.407	42.08
			True	41.43	0.405	42.32
	50	5	False	37.20	0.363	34.03
			True	37.80	0.373	33.81
		10	False	39.53	0.391	38.02
			True	39.61	0.391	37.84
	75	5	False	26.27	0.249	32.44
			True	26.15	0.246	31.66
		10	False	27.86	0.260	33.46
			True	28.26	0.268	34.05

2. *Image generation:* To broaden our investigation beyond classification, we applied the ICP-based self-distillation framework to the image generation task. In this experiment, we evaluated the framework using a Variational Autoencoder (VAE) trained on the CUB dataset. The VAE was optimized using the standard VAE loss, which is a combination of mean squared error (MSE) and KL divergence. The model was trained for 100 epochs using the previously determined optimal hyperparameters (k = 25, T = 5, Weighted = True). Evaluation metrics included the Structural Similarity Index (SSIM) and Fréchet Inception Distance (FID). The experiment was deliberately constrained, using a small dataset and a lightweight model with 128×128 image resolution to assess how well our method performs under challenging, resource-limited conditions.

Table 3.Quantitative evaluation on Image Generation on CUB Dataset.

Method	**SSIM ↑**	**FID ↓**	**Time (mins) ↓**
Control	0.2580	161.830	110.31
SGD-ICP	0.3365	159.905	232.52
Adam-ICP	0.3645	158.080	221.89
AdEMAMix-ICP	0.3893	157.604	232.80

results on CIFAR-100 indicate that aligning the intermediate representations of perturbed and unperturbed samples significantly enhances model performance. Figure 3 presents results from the three ICP-based methods alongside baseline training for image generation from the encoded representation of an image.

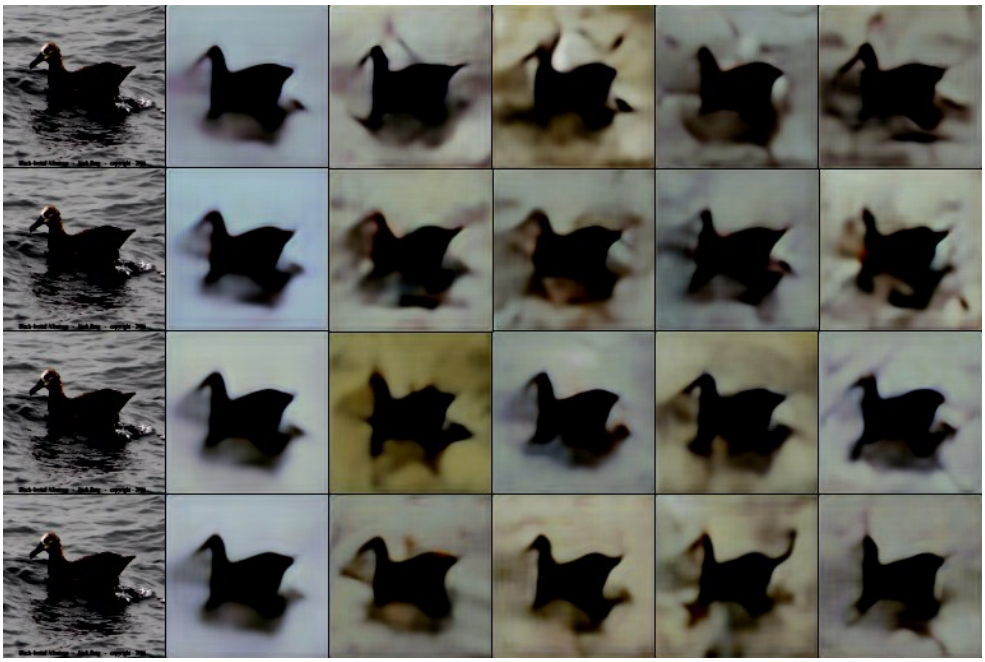


Figure 4.Left to right: Input image from CUB dataset, deterministic output of VAE(with not variance), outputs of VAE with 4 different noised latents with different seeds; ***Top to bottom:*** *Baseline control method, SGD-ICP,Adam-ICP, and AdEMAMix-ICP.*

Figure 4. shows results of the 3 methods alongside baseline training for image generation from the encoding of an image from CUB dataset. We used the latent encoding for the given image and generated 5 sample outputs, one with deterministic (i.e., without adding any noised variance) and 4 noised latents (using different noise samples for the reparameterization). Visually, there is not much difference to be seen, however the structure is more close to original in case of AdEMAMix-ICP for the deterministic outputs.

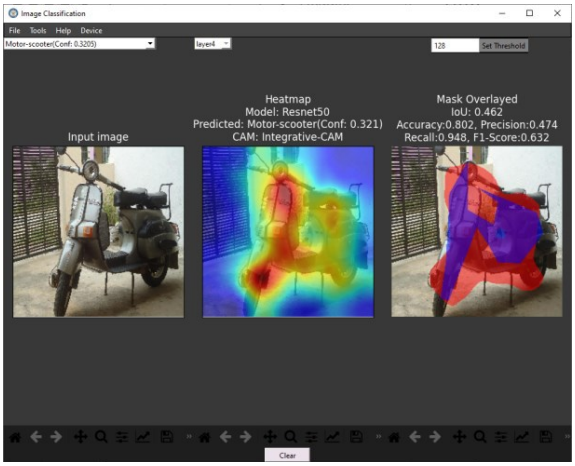
Table 3. show the SSIM and FID scores of VAE using the 4 methods. The results of image generation were expected to be poor due to the heavy constraints applied. However, even with such constraints, ICP still resulted in better scores as compared to the baseline. AdEMAMix-ICP performed best in terms of both SSIM and FID scores. Furthermore, even remaining 2 ICP variants still performed better than the baseline.

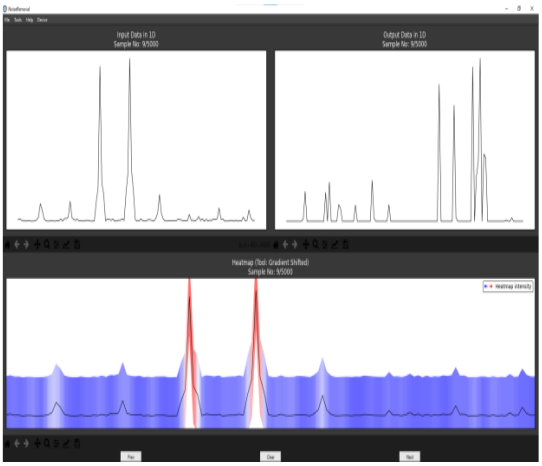
DRDO XCA DESKTOP APPLICATION

During the extension of the internship, I worked on developing the major portions of the *DRDO XCA DESKTOP* application, which was the final deliverable of the XCA project. The application was developed with the aim to explain few deep learning algorithms, such as the HRRP algorithm, and the CFAR algorithm, which were signal based deep learning algorithms given by the DYSL-CT Chennai lab. Other general algorithms used were based on image classification task and image fusion task. These features use famous explainability algorithms listed as below:

1. *HRRP algorithm*: This was signal classification task, for which a CNN trained on the given data was to be explained. Conventional algorithms such Grad-CAM, Grad-CAM++, Ablation-CAM, Pixel Ablation-CAM [20], LIME [21], SHAP [22].
2. *CFAR algorithm*: This was a signal based generation task, for which in-house developed explainability algorithms were used.
3. *Image Classification algorithm*: Family of ResNet [23], Swin-Transformers [24], and ConvNext models [25], along with the option uploading a pretrained CNN model by user were included. In addition to the aforementioned CAM algorithms, few extra algorithms were also included namely: Guided Backpropagation based algorithms (Normal, Guided-Grad-CAM, Guided-Grad-CAM++) [2], and a novel CAM algorithm called I-CAM [26]. A free form polygon based object annotator tool was also developed and included to compare and verify CAM algorithms against user based ground-truth masks.
4. *SVM Image Classification algorithm*: A pretrained SVM model was used, along with the option to upload a custom trained SVM model by the user, and to train an SVM on the user's desired dataset. Given the SVM model, the HiLite algorithm (an in-house explainability algorithm, which at the time of development of this report was still under review) was used to explain the output of the SVM model.
5. *Image fusion*: A CNN trained to fuse thermal and normal visual images, was used to explain the output using each modal inputs [27].

Few instances of the application have been depicted in figure 5.





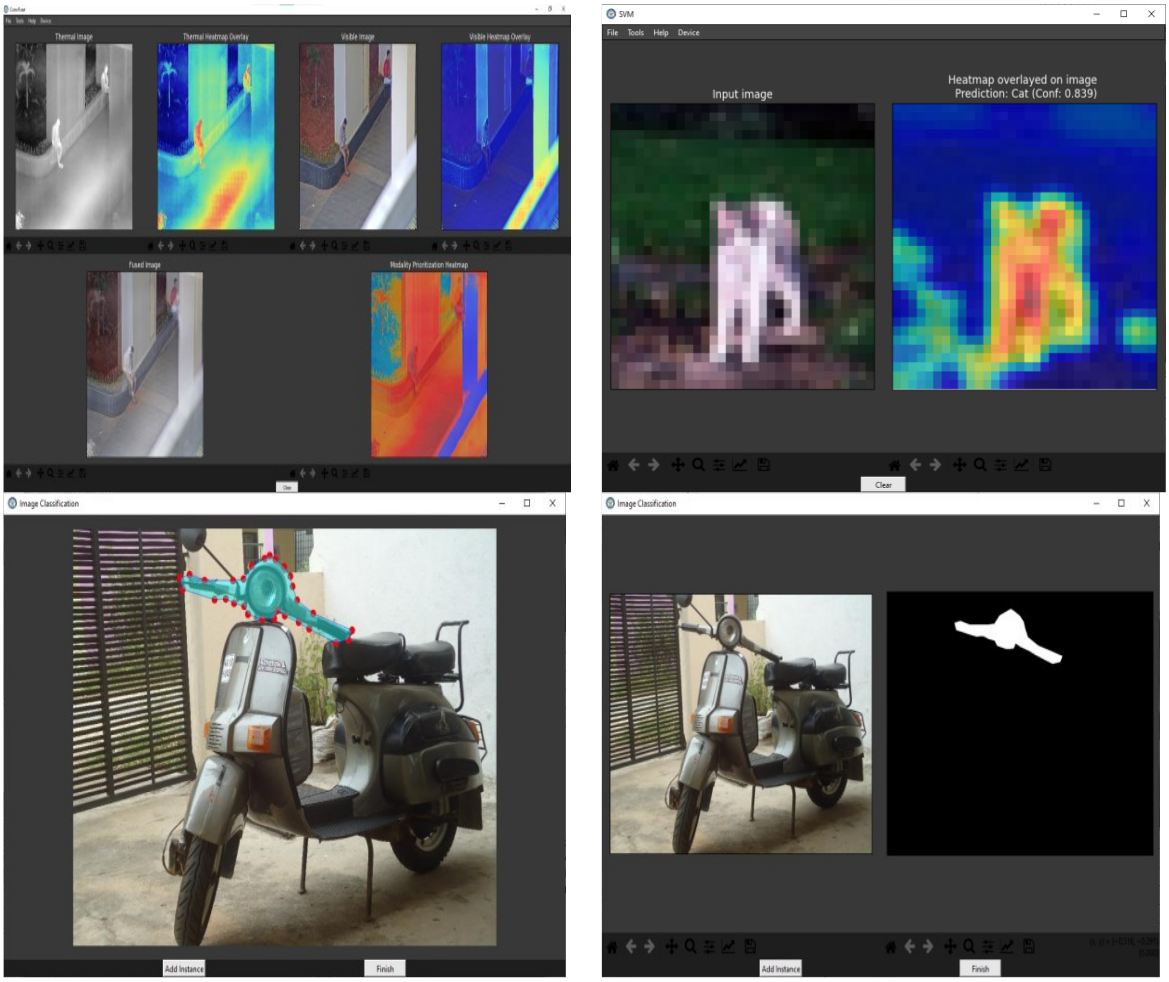


Figure 5. A collage of instances of the DRDO XCA desktop application.

CONCLUSION

This report summarizes the work completed during my research internship at the DRDO Project Executive Lab, Techno India University on the Explaining the Cognitive Algorithms (XCA) project. The internship delivered three core outcomes: (1) the design and evaluation of a "Firewall" auxiliary autoencoder intended to weaken gradient-based attacks while remaining plug-and-play with pretrained classifiers, (2) the development and thorough evaluation of Iterative Constructive Perturbation (ICP) and a self-distillation training framework that aligns perturbed and original feature representations, and (3) the implementation of the DRDO XCA desktop application that integrates a suite of explainability methods for both signal- and image-based models.

Experimental results showed promising directions and clear limitations. The Firewall idea yielded strong mitigation on a small-scale MNIST setup (successful attacks dropped from 37.94% to 6.9%), demonstrating that auxiliary modules which induce vanishing gradients can reduce the effectiveness of simple gradient attacks. ICP combined with layer-wise self-distillation produced robust gains across multiple vision tasks: in the CIFAR-100 classification ablation the AdEMAMix-ICP variant delivered the largest improvement (≈19.06% relative gain over the control baseline), and image-generation experiments exhibited consistent SSIM/FID improvements under constrained settings. These outcomes validate the central hypothesis that carefully constructed input refinements and iterative alignment can improve both performance and robustness.

At the same time, scaling challenges and practical trade-offs were identified. The Firewall's reliance on near-perfect reconstruction made it difficult to scale to colored, high-dimensional images without degrading the primary model's performance; the "no-free-lunch" limitations mean some reconstruction error is effectively unavoidable. ICP and its variants, while effective, add computational overhead and require careful hyperparameter tuning (k, T, weighting schedules) to balance baseline learning with distillation. The DRDO XCA desktop application centralizes many explainers and tools developed during the internship and provides a useful platform for further testing and user validation.

All code developed for these projects was written in Python using the PyTorch framework. Working end-to-end with PyTorch (model design, training loops, optimization schedules, mixed-precision experiments and deployment wrappers) allowed me to gain substantial hands-on experience in writing efficient, maintainable, and effective PyTorch code skills that materially improved experiment turnaround and reproducibility.

Future work should focus on making these defenses and refinement methods more practical for real deployments: improve reconstruction fidelity or find alternative gradient-diminishing transforms with lower utility loss; reduce the runtime cost of ICP (for example via adaptive iteration schedules or mixed-precision implementations); extend evaluations to larger, more realistic datasets and signal modalities used by DRDO partners; and conduct user studies with domain experts using the XCA application to validate explanation utility in operational workflows. Additionally, the ICP-driven self-distillation work described here is available as a preprint on arXiv for those who wish to review the methods and results in more detail [28]. Overall, the internship produced novel methods and a working application that together point toward a promising research direction: leveraging constructive perturbations and self-distillation to make AI systems both more interpretable and more robust.

REFERENCES

[1] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2921-2929).

[2] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

[3] Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018, March). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)* (pp. 839-847). IEEE.

[4] Ramaswamy, H. G. (2020). Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 983-991).

[5] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

[6] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

[7] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016, March). The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)* (pp. 372-387). IEEE.

[8] Liu, X., Li, Y., Wu, C., & Hsieh, C. J. (2018). Adv-bnn: Improved adversarial defense through robust bayesian neural network. *arXiv preprint arXiv:1810.01279*.

[9] Athalye, A., Carlini, N., & Wagner, D. (2018, July). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning* (pp. 274-283). PMLR.

[10] Yue, K., Jin, R., Wong, C. W., Baron, D., & Dai, H. (2023). Gradient obfuscation gives a false sense of security in federated learning. In *32nd USENIX security symposium (USENIX Security 23)* (pp. 6381-6398).

[11] Cohen, J., Rosenfeld, E., & Kolter, Z. (2019, May). Certified adversarial robustness via randomized smoothing. In *international conference on machine learning* (pp. 1310-1320). PMLR.

[12] L. Deng, "The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]," in *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141-142, Nov. 2012, doi: 10.1109/MSP.2012.2211477. keywords: {Machine learning}.

[13] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

[14] Wolpert, D. H., & Macready, W. G. (2002). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, *1*(1), 67-82.

[15] Zhang, L., Bao, C., & Ma, K. (2021). Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(8), 4388-4403.

[16] Pagliardini, M., Ablin, P., & Grangier, D. (2024). The ademamix optimizer: Better, faster, older. *arXiv preprint arXiv:2409.03137*.

[17] Nilsson, J., & Akenine-Möller, T. (2020). Understanding ssim. *arXiv preprint arXiv:2006.13846*.

[18] Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., & Kumar, S. (2024). Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9307-9315).

[19] Krizhevsky, A., & Hinton, G. (2009, September). *Learning multiple layers of features from tiny images.(2009)*.

[20] Chaudhuri, D., Samanta, A., Singh, A. K., & Singh, M. P. (2025). Pixel Ablation-CAM: A New Paradigm in CNN Interpretability for Feature Map Visual Explanations. *Defence Science Journal*, *75*(2), 188.

[21] Garreau, D., & Luxburg, U. (2020, June). Explaining the explainer: A first theoretical analysis of LIME. In *International conference on artificial intelligence and statistics* (pp. 1287-1296). PMLR.

[22] Oveis, A. H., Giusti, E., Meucci, G., Ghio, S., & Martorella, M. (2023, October). Explainability in hyperspectral image classification: a study of XAI through the Shap algorithm. In *2023 13th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)* (pp. 1-5). IEEE.

[23] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[24] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012-10022).

[25] Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11976-11986).

[26] Singh, A. K., Chaudhuri, D., Singh, M. P., & Chattopadhyay, S. (2024). Integrative CAM: Adaptive Layer Fusion for Comprehensive Interpretation of CNNs. *arXiv preprint arXiv:2412.01354*.

[27] Singh, M. P., Singh, A. K., & Chaudhuri, D. (2025, March). Towards Explainable Image Fusion: Gradient-Based Heatmaps for Modal Contributions. In *2025 4th International Conference on Range Technology (ICORT)* (pp. 1-6). IEEE.

[28] Dave, M., Singh, A. K., Pareek, A., Jha, H., Chaudhuri, D., & Singh, M. P. (2025). Self Distillation via Iterative Constructive Perturbations. *arXiv preprint arXiv:2505.14751*.