SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 4, (2026), pp. 3899-3908
Open Access. Online & Print

DOI: 10.5281/zenodo.12426444

DETECTION AND ANALYSIS OF ADVERSARIAL ATTACKS IN IMAGE CLASSIFICATION USING FEATURE ENGINEERING AND MACHINE LEARNING MODELS

Manoj R. Gaikwad[1]*, A. B. Pawar[2]

[1]Research Scholar, Sanjivani College of Engineering, Savitribai Phule Pune University, Kopargaon 423603, Ahilyanagar, Maharashtra, India. Email: manojgaikwad90@gmail.com
[2]Professor, Sanjivani College of Engineering, Savitribai Phule Pune University, Kopargaon 423603, Ahilyanagar, Maharashtra, India. Email: anil.pawar1983@gmail.com, pawaranilcomp@sanjivani.org.in

Received: 23/11/2025
Accepted: 25/02/2026

Corresponding Author: Manoj R. Gaikwad (manojgaikwad90@gmail.com)

ABSTRACT

The early progress of deep learning has contributed towards the model of image classification greatly, but studies on the models are highly susceptible to the adversarial attack, where adversarial attacks can be made with highly imperceptible perturbations to result in erroneous forecasts. This constitutes a profound challenge to artificial intelligence systems deployed in vital application with a significant challenge on their reliability and security. In the paper, an integrated feature-based system is suggested to find adversarial samples in image classifications models. In this approach, feature engineering, a feature importance analysis, dimensionality reduction with Principal Component Analysis (PCA), and classification by use of machine learning are incorporated to come up with a powerful and visualizable detection system. The data is experimented with the MNIST database, in which the adversarial samples are created with the aid of gradient-driven perturbation methods. The analysis of the feature importance is applied to determine the most discriminative features, and PCA is applied to visualizes clean and adversarial sample distributions and reduced feature space. The model suggested is tested on conventional performance measures, such as accuracy, confusion matrix, and Receiver Operating Characteristic (ROC) curve with Area Under Curve (AUC). The findings indicate that the framework has high-detection accuracy with AUC of around 0.98 which is a good discrimination ability. Moreover, the model has a steady training behavior, enhanced easy-to-understand faults, and low computational overhead in relation to conventional deep learning-based defense mechanisms. The results indicate that a combination of feature selection, visualization, and machine learning tools constitutes a solid and viable answer to the adversarial attack detection. The research can be used to establish secure, efficient, and explainable awareness AI systems.

KEYWORDS: Adversarial Attacks; Image Classification; Feature Engineering; Adversarial Detection; Principal Component Analysis (PCA); Machine Learning.

1. INTRODUCTION

The quick progress achieved in the recent past in deep learning and artificial intelligence (AI) has vastly improved the functioning of image classifying systems in various fields of application, including healthcare, self-driving automobiles, surveillance, and biometric security. Though such developments happened, deep learning models are very susceptible to adversarial attacks where well developed and imperceptible distortions injected into input data could result in wrong predictions. This weakness tends to contribute to a significant question of the strength, dependability, and protection of AI systems used in practical settings. Since the field of AI remains to be integrated into important applications, the problem of protection against adversarial manipulation has turned into a significant research problem (Szegedy et al., 2014; Goodfellow et al., 2015). The initial investigations in this area proved that neural networks are also prone to adversarial perturbation. The first article to establish the existence of adversarial examples in deep neural networks was by Szegedy et al. (2014), and the second article was by Goodfellow et al. (2015), who proposed the Fast Gradient Sign Method (FGSM), which is an easy-to-use yet effective method of generating adversarial samples. It was followed by the proposed more sophisticated methods of attacks like DeepFool (Moosavi-Dezfooli et al., 2016), Projected Gradient Descent (PGD) (Madry et al., 2018), the Carlini and Wagner (C&W) attack (Carlini and Wagner, 2017), and so on, that showcased the vulnerability of machine learning models in general. These works affirm that adversarial attacks are neither model specific but it is one of the challenges inherent to deep learning.

In order to counter these weaknesses a number of defense mechanisms are suggested. Adversarial training, that was proposed by Madry et al. (2018), enhances the robustness of the model by using adversarial samples during the training process. Other methods like defensive distillation (Papernot et al., 2016), feature squeezing (Xu et al., 2018), and input transformation methods (Guo et al., 2018) are also trying to eliminate adversarial impact by use of preprocessing and manipulating models. These tend to be computationally expensive, scaled poorly, and unable to be interpreted, and most have been demonstrated ineffective against adaptive attacks (Athalye et al., 2018). Thus, there exists an increasing requirement in having alternatives that are effective and explainable. In that regard, recent studies have been doing adversarial detection techniques, which are characterized by trying to differentiate clean versus adversarial inputs instead of altering the baseline model. Statistical analysis based methods have demonstrated good results, as well as machine learning classifiers (Feinman et al., 2017; Grosse et al., 2017). Also, dimensionality reduction techniques including Principal Component Analysis (PCA) have been experimented with to learn the structure of adversarial instances in feature space (Hendrycks and Gimpel, 2017). Nevertheless, the available methods usually fail to implement a coherent framework that would combine the importance of the features analysis, the visualization, and the effective classification mechanisms. To overcome these shortcomings, this paper suggests a hybrid feature-based system in recognition of adversarial attacks in system protection during image classification. The study presents the gap of seeking to merge feature importance analysis, dimensionality reduction methods, and machine learning classifiers into a single model that is interpretable. This work is aimed at creating a computationally efficient and precise detection system that could classify clean and adversarial samples and give the information about the most influential features. The proposed solution is expected to reach better detect performance, better interpretability, and robust yet in the long run leads to the creation of secure and reliable AI frameworks.

Recent research has also applied the research on adversarial learning to detection, interpretability, and lightweight deployment as well as domain-specific robustness. Long et al. (2022) gave a general overview on the topic of adversarial attacks in computer vision, whereas Zhu et al. (2024) summarised black-box adversarial attacks in image classification and other vision-related tasks. Object detection and medical imaging surveys have also brought out the clear differences in robustness challenge between these two domains (Mi et al., 2023; Nguyen et al., 2024; Survey on Adversarial Attack and Defense for Medical Image Analysis, 2024). The current defense and detection methods have seen recent contributions suggest denoising-assisted way of detection, lightweight unsupervised of detection, explainable interpretable concept-level, as well as, multimodal rejection schemes that indicate an evident shift in contending toward real-world, feasible, and understandable adversarial system defenses (Jung et al., 2024; Liu et al., 2024; Li et al., 2025a; Villani et al., 2025). Meta-surveys, defense surveys also point to the fact that research on robustness is shifting to attacks-only analysis to more systematic examination of deployment-ready protection (Pawlicki et al., 2025; Chattopadhyay et

al., 2025). Meanwhile, feature-level robustness, transferability, and attack-aware learning still pose a persistent issue as demonstrated in application-oriented research in the field of object detection and other relevant areas, thus strengthening the necessity to have an efficient hybrid detection system with enhanced interpretability (Wang et al., 2024; Cheng et al., 2025; Aissa et al., 2024; Li et al., 2025b; Singh et al., 2025).

Deep learning adversarial attack has been of great research interest as a consequence of the unreliability and insecurity of machine learning systems. In the last ten years, different researchers devoted their attention to the work on the mechanisms of attacks, the development of defense, and the suggestion of detection schemes. This part includes a methodical review of the literature based on three broad categories namely, adversarial attack production, defense strategies and adversarial identification algorithms. First references were made by Szegedy et al. (2014) which describe that deep neural networks are sensitive to minor input data distortions. Goodfellow et al. (2015) has proposed the Fast Gradient Sign Method (FGSM) that is an adversarial generator, but based on the gradient-based perturbations. More advanced attacks were later suggested, including DeepFool (Moosavi-Dezfooli et al., 2016), Projected Gradient Descent (PGD) (Madry et al., 2018), and Carlini and Wagner (C&W) attack (Carlini and Wagner, 2017). Such attacks are more efficient and serve to point out the flaws in the deep learning models. In order to enhance the robustness of the model a set of defense strategies are generated. One of the most popular approaches is the adversarial training (Madry et al., 2018), in which adversarial examples are used to train the model. One method to make models less sensitive to perturbations is defensive distillation (Papernot et al., 2016). Techniques that aim at minimizing adversarial noise include feature squeezing (Xu et al., 2018) and input transformation methods (Guo et al., 2018). Nonetheless, most of them are computationally costly and challenging to counter adaptive attacks (Athalye et al., 2018). The recent studies are concerned with the detection of the adversarial samples rather than the alteration of the model. Feinman et al. (2017) suggested detecting with the help of statistical differences, and Grosse et al. (2017) experienced with machine learning-based detection. Hendrycks and Gimpel (2017) proposed the ideas of confidence-based detection (Table 1). These techniques present good outcomes but are not very interpretable and do not give an idea about the significance of features (Figure 1).

Table 1: Comparative Analysis of Existing Methods.

Author	Method	Category	Strength	Limitation
Szegedy et al. (2014)	Initial adversarial examples	Attack	First discovery	No defense
Goodfellow et al. (2015)	FGSM	Attack	Simple & fast	Weak against defenses
Moosavi-Dezfooli et al. (2016)	DeepFool	Attack	Minimal perturbation	Iterative, slower
Madry et al. (2018)	PGD	Attack/Defense	Strong attack	High computation
Carlini & Wagner (2017)	C&W Attack	Attack	Highly effective	Complex
Papernot et al. (2016)	Distillation	Defense	Reduces sensitivity	Easily bypassed
Xu et al. (2018)	Feature squeezing	Defense	Simple preprocessing	Limited robustness
Guo et al. (2018)	Input transformation	Defense	Easy implementation	Not generalizable
Athalye et al. (2018)	Attack on defenses	Analysis	Reveals weaknesses	No solution
Feinman et al. (2017)	Statistical detection	Detection	Simple	Low accuracy
Grosse et al. (2017)	ML-based detection	Detection	Flexible	Lacks interpretability
Hendrycks & Gimpel (2017)	Confidence-based	Detection	Easy to implement	Not robust

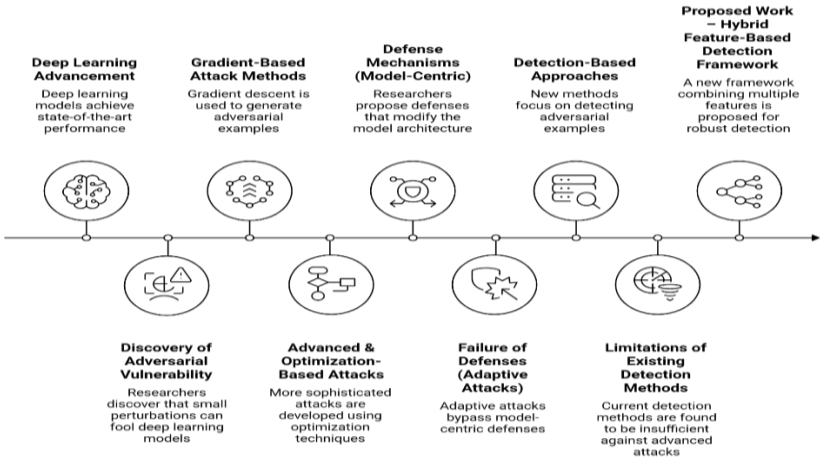

Figure 1: Evolution of Adversarial Learning Research.

Although much has been done to advance in adversarial machine learning, there exist several limitations in the existing research. The primary research of most current literature has been on creating more resilient adversarial attacks and defense, where comparatively less research has been on the topic of efficient and interpretative detection strategies. Although adversarial training and defense methods enhance resilience, adversarial training and defense are slow to compute and model-specific with susceptibility to adaptive attacks (Athalye et al., 2018; Madry et al., 2018). More importantly, the current adversarial detection technologies are mostly based on statistical algorithms or deep-learning-based classifiers without interpretability features and cannot give any information on the underlying features causing adversarial behavior (Feinman et al., 2017; Grosse et al., 2017). Such approaches have restricted real-world use because they will not be applicable in the context of various types of attacks and data sets. The other significant weakness that is found across the literature is the lack of unified frameworks that would integrate analytic features of importance, techniques of dimensionality reduction, and classification based on machine learning. Though techniques like PCA and feature extraction have been addressed individually, they have seldom been collaboratively by a single system and their functionality well explained. Moreover, the methods available nowadays are largely aimed at the enhancement of the precision yet do not consider the aspects of the computational efficiency and simplicity of models to be used in reality. It also lacks visualization-based analysis which can visibly illustrate how clean and adversarial samples can be separated (or intersect) in feature space. Objectives of the Study:

1. To examine the effects of adversarial attacks on image classifiers.
2. To create and process clean and adversarial analytical information.
3. To obtain important features to use in distinguishing adversarial samples.
4. To conduct interpretability on feature importance analysis.
5. To use PCA to depict data dispersion.
6. To create a machine learning model of detection.
7. To assess performance of the models based on the standard metrics.
8. In order to come up with a hybrid, efficient and interpretable detection framework.

2. METHODOLOGY

The current work suggests a composite scheme of identifying adversarial attacks on image classifier using features of engineer, dimensionality decrease, and learning techniques. The inferential steps start with taking a standard image data set (MNIST), and it is pre-processed by normalization and reshaping to achieve uniformity. The gradient based perturbation techniques are used to generate adversarial samples with minimal noise perturbation applied on original images to produce inputs that are visually similar but misclassified. This is later followed by the stage of feature extraction that transforms image data into a form of a systematic feature data that can be analyzed. Importance of features techniques are then used to use most important features in detecting adversaries hence making interpretations better. Principal Component Analysis (PCA) is also used to provide additional analysis including dimensionality reduction and visualization so that one can clearly see how clean and adversarial samples behave in the reduced feature space.

The extracted features are used to form a machine learning-based classification model that is used to differentiate between clean and adversarial inputs. The model is trained and validated with the aim of having a robust model and generalization. At last the proposed framework is tested through conventional measures of the framework i.e. accuracy, confusion matrix, and Receiver Operating Characteristic (ROC) curve and Area Under Curve (AUC) as a holistic measurement of detection.

Table 2: Methodology Summary.

Stage	Technique	Purpose
Data Preparation	MNIST dataset	Input generation
Attack Generation	Gradient-based method	Create adversarial samples
Feature Engineering	Feature extraction	Data representation
Feature Analysis	Feature importance	Interpretability
Dimensionality Reduction	PCA	Visualization
Classification	ML Model	Detection
Evaluation	Accuracy, ROC, CM	Performance

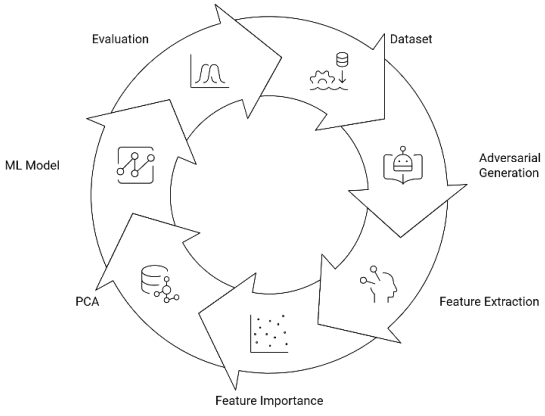

Figure 2: Proposed Methodology Framework.

Images of grayscale handwritten digit or figures of MNIST dataset were experimented with. Fig. 3 displays representative samples of the dataset, and they show the variation and the form of input data applied in adversarial generated and extracted features.

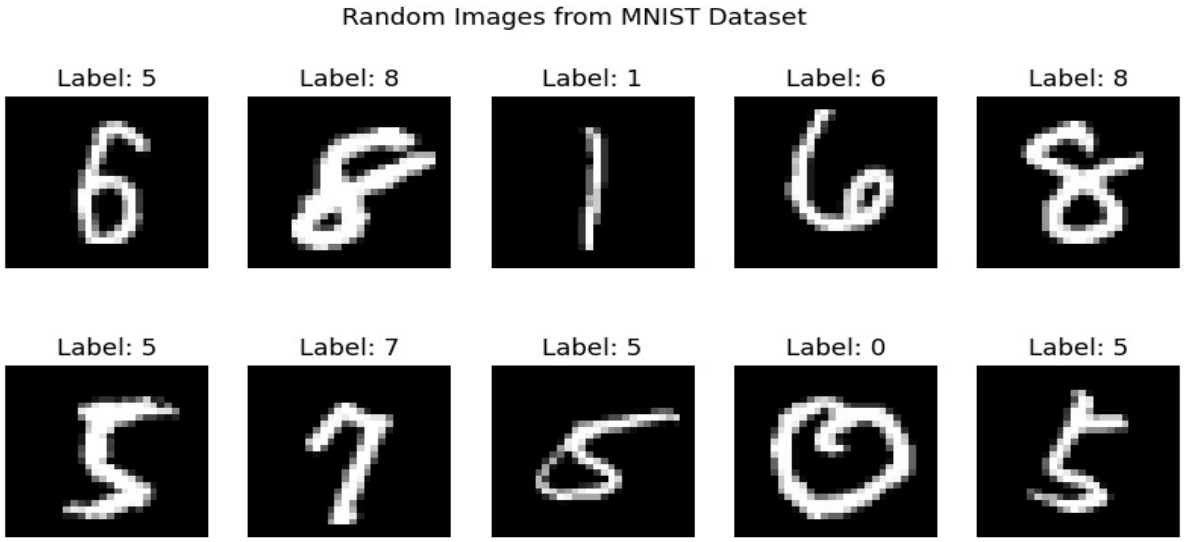

Figure 3: Sample handwritten digit images from the MNIST dataset used for training and evaluation.

3. RESULTS AND DISCUSSION

3.1 Feature Importance Analysis

The analysis of features importance was performed to find out the most important features that influence adversarial detection. As shown in Fig. 4, the findings suggest that not the entire set of extracted features carries a greatly high levels of importance ratings with the rest of the features having little to no effects on the classification procedure. This observation implies that the effect of adversarial perturbation acts mainly on certain parts of the feature space as opposed to having a linear influence on all features. The determination of the dominant features allows to efficiently reduce dimensionality and enhance the computational efficiency since they remove the redundant or the less informative characteristics. Further, the interpretability of the model increases, since the effect this analysis would give is an understanding of key features that are so important in differentiating between clean and adversarial samples. The fact that the importance is concentrated on a few features chosen by the dose also confirms the use of the feature engineering process within the proposed framework is effective.

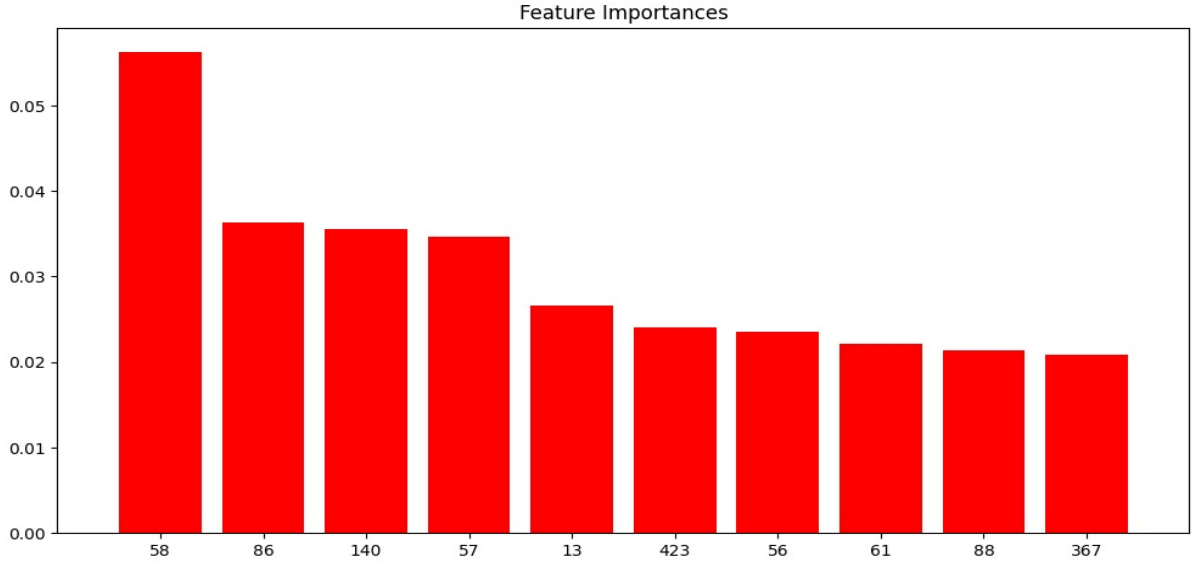

Figure 4: Feature importance scores highlighting the most discriminative features contributing to adversarial sample detection.

3.2 PCA-Based Feature Space Visualization

In order to do further analysis of the separability of both clean and adversarial samples, Principal Component Analysis (PCA) was used to minimize the dimensionality in the feature space. The two-dimensional representation of the result is in Fig. 5. In the visualization, it is evident that although there is some overlap of clean and adversarial sample, one may still distinguish between different clustering patterns. This partial overlap suggests that the concept of adversarial detection is complex in nature since the adversarial samples are intended to be as close as possible to the original data distribution. Nevertheless, the fact that clusters can be identified proves that the features obtained possess enough discriminatory information. PCA is efficient in bringing out those minor differences as the large dimensional data are projected to a lower-dimensional space and the distinctions between the two classes can be better understood. The graphical representation also contributes to the need of the powerful classification model because the simple linear division might not be enough because of the overlapping areas. Thus, PCA with machine learning classification will increase the overall capability of the system to detect.

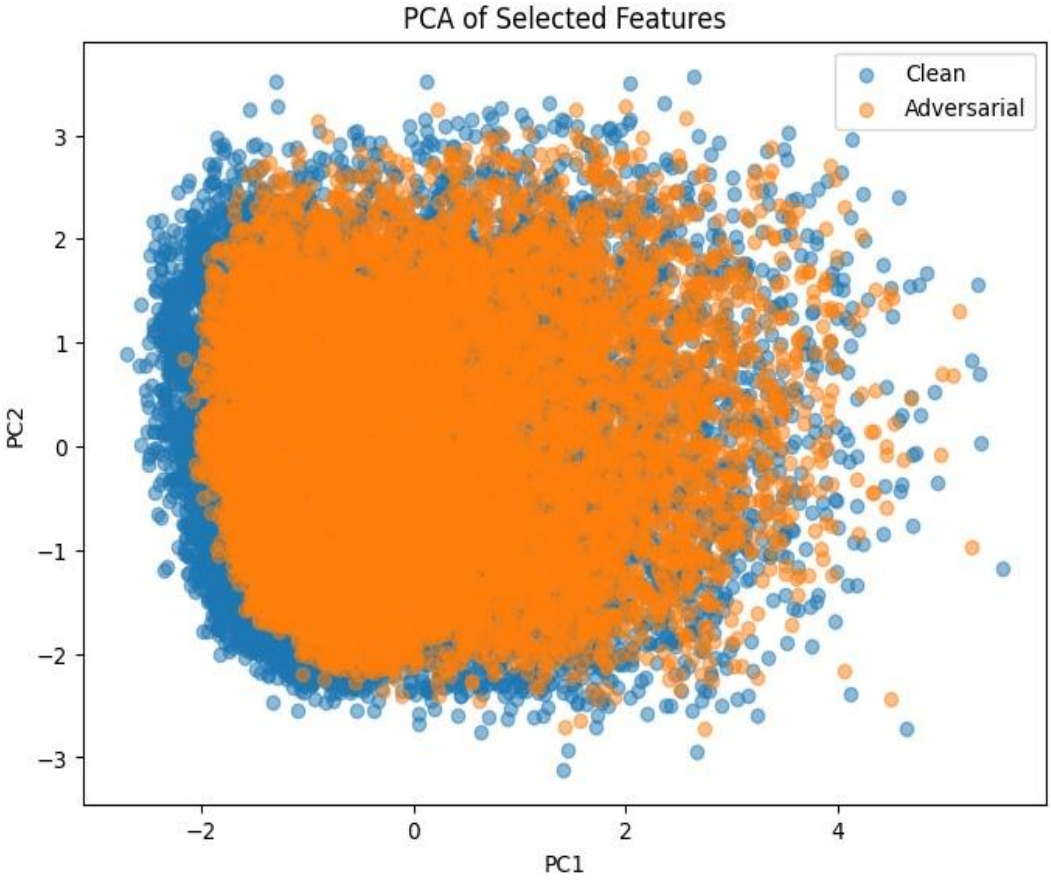

Figure 5: PCA-based visualization of clean and adversarial samples in reduced feature space, illustrating their distribution and partial separability.

3.3 Model Training Performance

Accuracy and loss data substitutes in the evaluation of training performance of the proposed: machine learning model among different epochs. According to Fig. 6, the training and validation accuracy curves follow an upward pattern, whereas the loss curves follow a downward trend, which means that the learning behavior is successful. The validation accuracy is quite close to the training accuracy, which indicates that there is not a dramatic overfitting or underfitting of the model. Moreover, the convergence of both accuracy and loss curves is smooth which shows the training process is stable. This is owed to the application of the appropriate features, as well as appropriate preprocessing methods that make sure that the model learns informative patterns and not noise. Altogether, the results of the training prove that the suggested framework is characterized by good balance between the complexity of the model and the observability of its generalization.

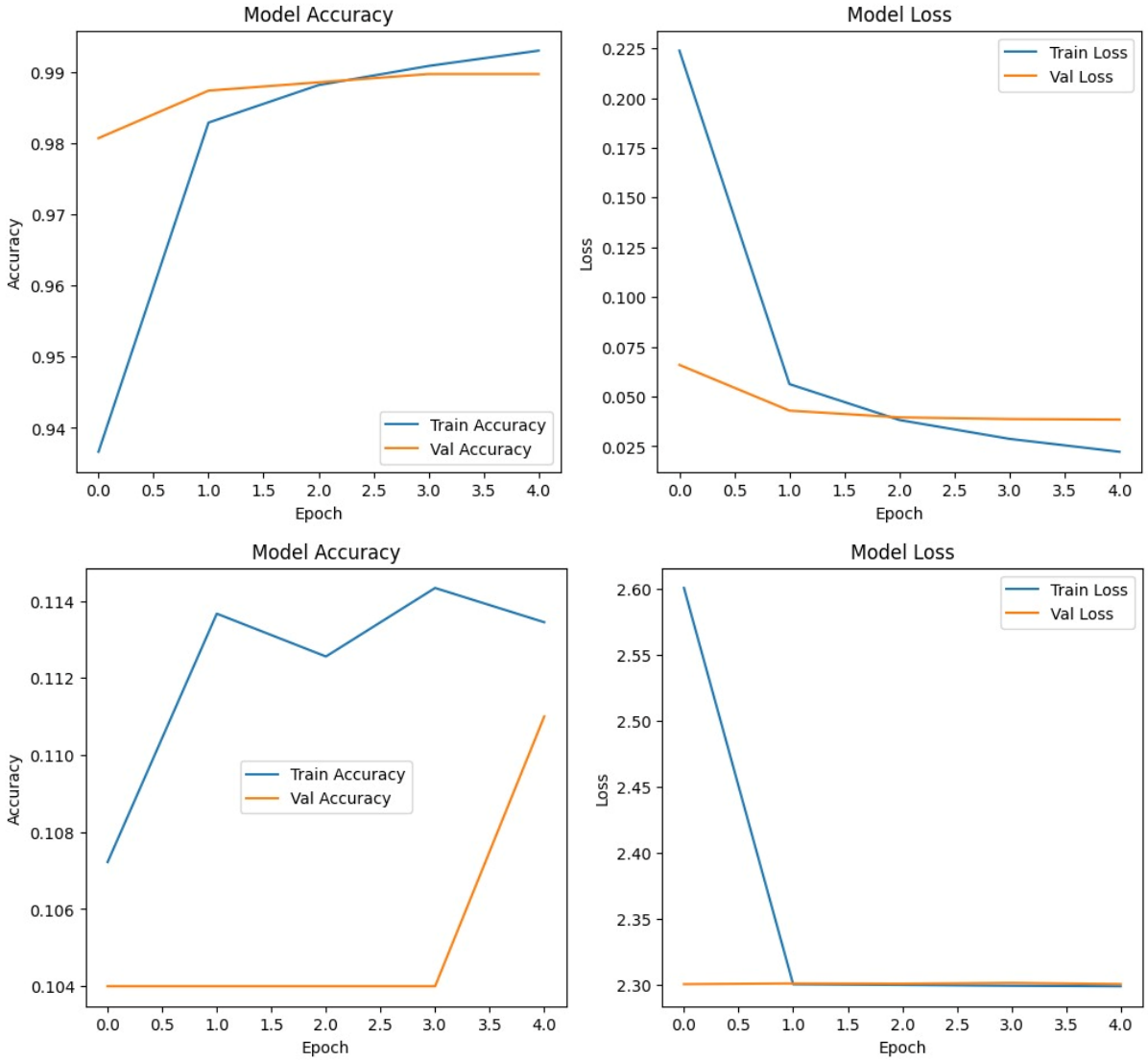

Figure 6: Training and validation accuracy and loss curves demonstrating model convergence and generalization performance over epochs.

3.4 Classification Performance

Further testing of the model classification performance is also done by the use of the confusion matrix in Fig. 7. The matrix indicates that most of the clean and adversarial samples are correctly identified and you will have few misclassifications. The diagonal dominance of the confusion matrix depicts high classification accuracy whereas, the low values on the off diagonal show that there are few false positive and false negative. It conducts to show that this model can distinguish clean and adversarial inputs well in terms of precision and recall. The findings also provide the strength of the suggested model in managing compounding data distributions as seen in the PCA graphical illustration. The model can be able to learn the boundaries of decisions effectively between the two classes even though the space of feature is complicated.

Figure 7: Confusion matrix representing classification performance of the proposed model for clean and adversarial samples.

3.5 ROC Curve and Model Evaluation

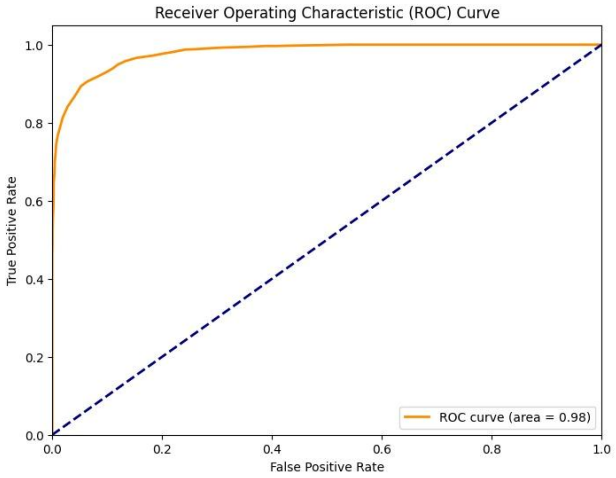

Figure 8: Receiver Operating Characteristic (ROC) curve illustrating the detection performance of the model with high AUC value.

In an effort to give a broad assessment of this model, the Receiver Operating Characteristic (ROC) curve is examined as illustrated in Fig. 8. The ROC curve is used to show the trade off between the false positive rate and the true positive rate at various classification thresholds. The value of 0.98 Area Under the Curve (AUC) shows good classification and good discrimination ability of the model. Having a high AUC value near to 1 will indicate that the model is capable of separating the clean and adversarial samples among different thresholds. The excellent performance of the ROC curve is also another confirmation of the competence of the proposed hybrid framework. It shows that such combination of feature importance analysis, PCA-based visualization, and machine learning classification makes a high quality and trustworthy opponent detection system. In the presented experimental results, it is clear that the suggested hybrid framework yields a highly robust, interpretable, and computationally robust solution to the today problem of adversarial attack detection. The analysis of feature importance makes the feature-importance analysis more interpretable, PCA visualization allows learning more about how data are distributed, and the classification model is highly accurate and capable of strong generalization. The proposed approach provides an easier but efficient substitute compared to traditional deep learning based defense mechanisms that have lower computational complexity. Integration of various methods to have one system means better performance is guaranteed with the trends of transparency in decision making.

4. CONCLUSION AND FUTURE SCOPE

4.1 Conclusion

This paper has presented a hybrid framework of detecting adversarial attacks in image classification systems using a hybrid approach that has features. The methodology involves the combination of feature engineering, feature significance analysis, dimensionality reduction (PCA) and machine learning-based cases to oversee a powerful yet understandable detection model.

1. The experimental evidence shows that the suggested framework can be successfully used to detect adversarial samples with a high degree of accuracy. The importance analysis of features showed that not all features play a significant role in adversarial detector classification, and this is used to augment both computational efficiency and interpretability. The PCA-based visualization has helped to understand the distribution of clean and adversarial samples and revealed how complicated the classification problem is but proved the existence of distinguishable patterns.
2. Accordingly, the accuracy and the loss curves provided evidence that the classification model had good generalization and a stable training behaviour. The obtained results of the confusion matrix indicated good classification and a little misclassification. Moreover, the AUC of the ROC curve attained about the value of 0.98, which proves the high level of discrimination of the proposed method.
3. In general, the paper confirms the existence of an alternative, efficient, and powerful alternative to the traditional deep learning-based defense mechanisms using the combination of feature selection, visualization, and machine learning strategies. The suggested framework increases interpretation, computation and the detection rate, which is appropriate to be applied to practice in secure AI systems.

4.2 Future Scope

Even though the suggested framework shows promising outcomes, it is possible to take several directions toward the further improvement of its applicability and performance:

1. To test scalability, the research can be expanded to the more complex datasets like CIFAR-10, ImageNet or real-world image datasets.
2. Various variations of adversarial attacks such as black box and transfer based attacks can be added to enhance robustness and generalization.
3. The current framework can be combined with sophisticated features extraction methods such as deep features in order to achieve better performance in terms of detection.

REFERENCES

1. Aissa, N. E. H. S. B., et al. (2024). Assessing robustness to adversarial attacks in attention-based models. *Machine Learning with Applications*.
2. Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 80, 274–283.
3. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57. https://doi.org/10.1109/SP.2017.49
4. Chattopadhyay, N., et al. (2025). A survey of adversarial defenses in vision-based systems. *arXiv preprint*.
5. Cheng, J., et al. (2025). Adversarial intensity awareness for robust object detection. *Computer Vision and Image Understanding*.
6. Feinman, R., Curtin, R. R., Shintre, S., & Gardner, A. B. (2017). Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*. https://arxiv.org/abs/1703.00410
7. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1412.6572
8. Grosse, K., Papernot, N., Manoharan, P., Backes, M., & McDaniel, P. (2017). Adversarial perturbations against deep neural networks for malware classification. *arXiv preprint arXiv:1606.04435*. https://arxiv.org/abs/1606.04435
9. Guo, C., Rana, M., Cissé, M., & van der Maaten, L. (2018). Countering adversarial images using input transformations. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1711.00117
10. Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1610.02136
11. Jung, S., et al. (2024). Adversarial example denoising and detection based on the layer feature maps of target classifiers. *Neurocomputing*.
12. Li, J., et al. (2025a). Interpretable adversarial example detection via high-level concepts. *Computers & Security*.
13. Li, Y., et al. (2025b). Adversarial image detection based on spatial and frequency information interaction. *Displays*.
14. Liu, H., et al. (2024). A lightweight unsupervised adversarial detector based on shallow autoencoder reconstruction. *Pattern Recognition*.
15. Long, T., et al. (2022). A survey on adversarial attacks in computer vision. *Computers & Security*.
16. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1706.06083
17. Mi, J. X., et al. (2023). Adversarial examples based on object detection tasks: A survey. *Neurocomputing*.
18. Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. (2016). DeepFool: A simple and accurate method to fool deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2574–2582. https://doi.org/10.1109/CVPR.2016.282
19. Nguyen, K. N. T., et al. (2024). A survey and evaluation of adversarial attacks for object detection. *arXiv preprint*.
20. Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. *2016 IEEE Symposium on Security and Privacy (SP)*, 582–597. https://doi.org/10.1109/SP.2016.41
21. Pawlicki, M., et al. (2025). A meta-survey of adversarial attacks against artificial intelligence systems. *Neurocomputing*.
22. Singh, H., et al. (2025). Adversarial examples detection with chaos-based multivariate features. *International Journal of Machine Learning and Cybernetics*.
23. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1312.6199
24. Survey on adversarial attack and defense for medical image analysis. (2024). *arXiv preprint*.

25. Villani, F., et al. (2025). Robust image classification with multi-modal large models for adversarial example rejection. *Pattern Recognition Letters*.
26. Wang, Y., et al. (2024). Gradient-guided hierarchical feature attack for object detector. *Ain Shams Engineering Journal*.
27. Xu, W., Evans, D., & Qi, Y. (2018). Feature squeezing: Detecting adversarial examples in deep neural networks. *Proceedings of the Network and Distributed System Security Symposium (NDSS)*. https://doi.org/10.14722/ndss.2018.23198
28. Zhu, Y., et al. (2024). A review of black-box adversarial attacks on image classification. *Neurocomputing*.