SSRG International Journal of Computer Science and Engineering
ISSN: 2348–8387 / https://doi.org/10.14445/23488387/IJCSE-V13I2P101

Volume 13 Issue 2, 1-9, February 2026

Original Article

An Explainable AI Framework for Image Analytics and Synthetic Image Creation Using CNN and GAN Architectures

Gaurav Shekhar

Executive Leadership, Technology Innovation.

Corresponding Author : gauravshekharster@gmail.com

Received: 15 December 2025 Revised: 19 January 2026 Accepted: 08 February 2026 Published: 27 February 2026

Abstract *- Explainable Artificial Intelligence (XAI) has become an important field of study to enable the exploration of interpretability and transparency issues relating to deep learning models, especially in high-stakes image analytics systems like medical imaging, surveillance, autonomous systems, and industrial inspection. Convolutional Neural Networks (CNNs) have been shown to outperform other state-of-the-art models in image classification, image detection, and image segmentation, and Generative Adversarial Networks (GANs) have been pioneers in generating synthetic images and data augmentation. Although successful, the CNN and GAN architectures are frequently criticized as black-box models that restrict the trust of users and regulatory requirements and implementation in sensitive devices. This paper will introduce a single Explainable AI system, which determines explainability mechanisms in CNN-based image analytics and GAN-based image generation. The framework also presented model-level, feature-level, and instance-level interpretability of CNN classifiers through gradient-based attribution, concept activation vectors, and saliency-based analysis of attention. Meanwhile, explainability is inherent in GAN models by discussing latent space representations, generator-discriminator dynamics, as well as the semantic disentanglement of generated elements. The framework permits transparency in predictive decisions, as well as in the generative mechanisms supporting the synthetic creation of images. A modular pipeline is made so that it enables interpretability throughout training, inference, and synthetic data generation phases. Mathematical formulations of CNN feature attenuation and GAN latent variable sensitivity analysis are introduced to give a theoretical basis. Benchmark image datasets are evaluated experimentally to evaluate the accuracy of classification, generative fidelity, metrics of explainability, and human interpretability scores. Findings indicate that the proposed framework is highly effective in enhancing model transparency without affecting the predictive performance or synthetic image quality. This work has made the following contributions: (i) a single explainable architecture with both CNN and GAN models, (ii) formal explainability measures of generative models, as well as (iii) a scalable framework applicable to practical image analytics systems. The study develops credible AI bridging performance and interpretability in contemporary deep learning-based image systems.*

Keywords *- Explainable Artificial Intelligence, Convolutional Neural Networks, Generative Adversarial Networks, Image Analytics, Synthetic Image Generation, Model Interpretability, Trustworthy AI.*

1. Introduction

1.1. Background

The high-speed development of Deep Learning has radically transformed the image analytics sphere as it allows deriving features automatically and directly learning the features from the freedom of mass visual data. Convolutional Neural Networks (CNNs) represent the most popular image-related architecture, in terms of classification, object detection, and semantic segmentation, as they can learn hierarchical representations that scale up from low-level visual features to high-level concepts. [1-3] Simultaneously, Generative Adversarial Networks (GANs) have brought high-level generative abilities, and the generation of realistic images can be used in advancing such tasks as data augmentation, super-resolution, image-to-image translation, and style transfer. Collectively, the models have massively enhanced performance and flexibility in the present-day computer vision systems, which have led to their extensive application in academic and industrial studies and applications. In spite of their high success, deep learning models can hardly be viewed as transparent, providing little insight into their decision-making process or method of image creation. Such non-transparency raises serious concerns in high-stakes areas, including healthcare, surveillance, autonomous systems, and industrial control over quality, especially in areas where accountability, fairness, and regulation are paramount. This black-box behavior complicates the process of checking model correctness and concurrently finding bias and defending the decision to the stakeholders, thus compromising the trust between end-users and domain experts as well as between the policy makers. With the

growing role of deep learning systems in real-world decision-making, interpretable and trustworthy models have become increasingly popular. Based on these issues, Explainable Artificial Intelligence (XAI) has become a reaction to these challenges, where humans-understandable explanations are provided to explain how models will behave and why they decide in a particular way. Nevertheless, the current XAI methods mainly focus on discriminative networks like CNNs and are aimed at explaining the results of prediction. Conversely, generative model explainability, especially latent representation and generation mechanisms, has not yet been thoroughly studied, especially in the context of GANs. In addition, there is an absence of studies that can consistently and comprehensively discuss explainability in both analytical and generative image processing pipelines, which is used as a significant weakness compelling various systems to adopt unified and comprehensive explainable systems.

1.2. Importance of an Explainable AI Framework for Image Analytics

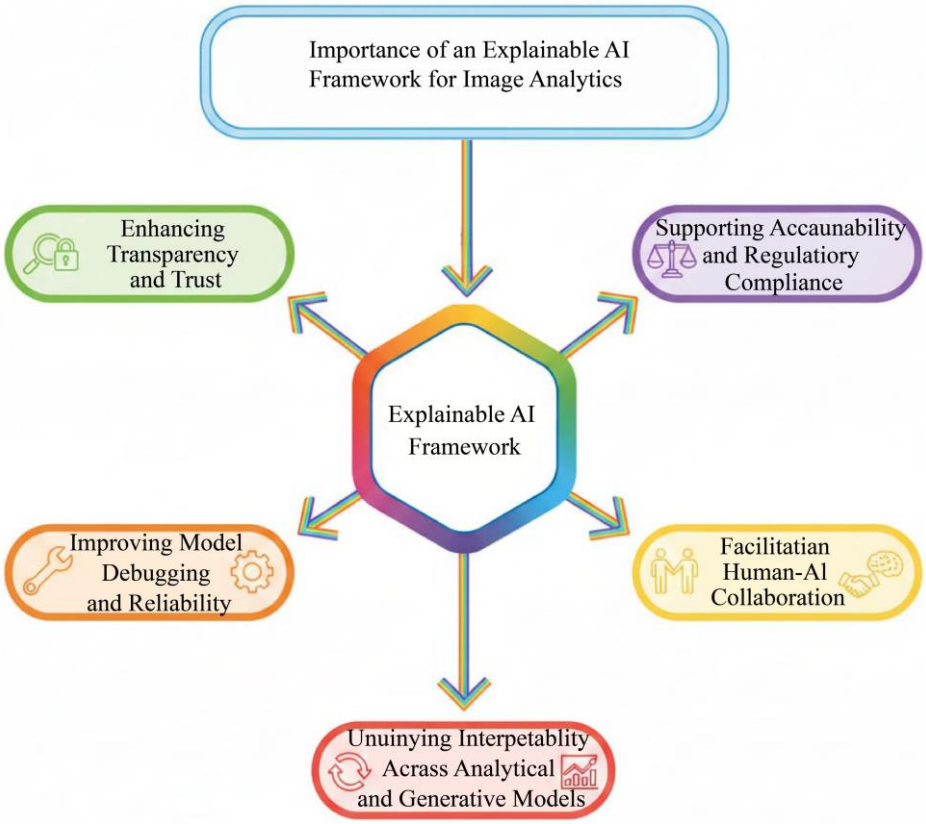

Fig. 1 Importance of an Explainable AI Framework for Image Analytics

1.2.1. Enhancing Transparency and Trust

Explainable AI (XAI) systems are helpful in enhancing the transparency of image analytics systems by exposing the process through which deep learning models make their decisions. With the classical black-box CNN and GAN models, predictions and generated output are not clearly explained, and this provides a limitation to the confidence of a user.

An XAI structure can facilitate end-user acceptance and trust of AI-based image analytics solutions by enabling end users, domain experts, and other stakeholders to learn the rationale behind model outputs, either via interpretable visual, feature-level, or concept-based explanations.

1.2.2. Supporting Accountability and Regulatory Compliance

Regulatory bodies in most real-world scenarios (including healthcare diagnostics, biometric surveillance, autonomous systems, and industrial inspection, etc.) are growing more concerned with AI systems being explainable and auditable. A justifiable framework can enable organizations to justify automated decisions, track model behavior, and show adherence to legal and ethical provisions. Such responsibility is specifically relevant when the consequences of image analytics systems affect high-stakes decisions, where wrong or discriminatory results can be disastrous.

1.2.3. Improving Model Debugging and Reliability

Elucible AI models can go a long way towards the validation and debugging of models by revealing failure modes, biases, and spurious correlations learned using training data. Explanations in image analytics can be used to show whether models are based on meaningful visual patterns or irrelevant background patterns. This understanding can assist a researcher and practitioner to filter out data, modify structures, and enhance resilience, and eventually create more dependable and transferable models.

1.2.4. Facilitating Human–AI Collaboration

A human-explainable scheme enhances human-ai cooperation, allowing users to interrelate, challenge, and authenticate model outputs. Human experts can use model explanations in conjunction with their subject experience in image analytics problems, including medical imaging or quality inspection, to make more informed decisions. This is because this way of collaboration improves the quality of decisions, and AI systems should be seen as auxiliary and not as black box decision-making systems.

1.2.5. Unifying Interpretability Across Analytical and Generative Models

Lastly, an explainable AI structure is necessary to make sense of interpretability in both discriminative image analytics and generative image synthesis. Due to the higher adoption of CNNs and GANs in the context of the same pipeline, an integrated explainability methodology works to guarantee uniform awareness of the process of predictions and the creation of fake images. This personalized interpretability is vital to establishing trustworthy, open, and morally deployable image analytical mechanisms.

1.3. Synthetic Image Creation Using CNN and GAN Architectures

The generation of synthetic images has become an essential part of the cost-effective image recognition systems nowadays, especially where large amounts of labeled images are costly, time-consuming, or where there are restrictions on privacy and ethical issues. [4,5] CNNs and GANs have a complementary role to play in this process. CNNs are primarily utilized to train discriminative feature representations of authentic images that allow them to classify effectively, detect, and extract features. Such learned representations frequently become the basis of directing or estimating synthetic image generation, such as serving as a discriminator, a feature extractor, or quality assessment in generative pipelines. CNNs assist in preserving critical structural and semantic features of real-life data because of hierarchical feature learning that helps ensure that synthetic images maintain such critical features. GANs, in contrast, are highly optimized towards image synthesis of high fidelity using an adversarial learning methodology, which uses a generator and a discriminator.

The discriminator, which is commonly a CNN architecture, evaluates the naturalness of generated samples, and the generator learns to generate realistic samples by mapping latent variables. It is a competitive training paradigm that allows GANs to synthesize visually realistic images that are similar to real data distributions. Synthetic image generation using GANs has found extensive applications in data augmentation, image-to-image translation, super-resolution, and domain adaptation, contributing a significant improvement in performance on downstream image analytics tasks. The integration of CNNs and GANs into a single framework makes the synthetic image creation more regulated and application-sensitive. CNN-based analytics can be used to drive the generation procedure, whether through enforcing semantic correctness, anomaly detection, or the estimation of task-dependent relevance of synthetic data. Such synergy enables synthetic images not just to grow the amount of data but also to enhance the diversity and representativeness of data. Consequently, CNNGAN architectures offer a practical and generalizable paradigm for synthetic image production, suggesting solid learning in a data-scarce setting and making it possible to execute demanding image analytics schemes.

2. Literature Survey

2.1. Explainable AI in Image Analytics

Early explainable image analytics work mainly focused on opening the black box of Convolutional Neural Networks (CNNs) by visualizing the internal representations, like the learned filters and middle feature maps. Techniques such as activation maximization were used to find patterns that elicit maximum stimulation to the neurons, and saliency maps, combined with heatmap visualizations [6-8], were used to identify image regions that make the most contributions to model predictions. Gradient methods (including sensitivity analysis and backpropagation-based methods) were also improved to provide greater interpretability by generating predictions and tracing which input pixels most affected them. Later innovations added methods of class-specific explanation, which offered local and instance-scale understanding of model behavior instead of an overall model behavior. More recently, concept-based ways of explanation have developed where abstract features of learning are attributed to semantic features that can be cognized by human beings, hence highly useful and trustworthy. However, regardless of this progress, explainable image analytics has been, as a rule, restricted to discriminative models, which provide comparatively little information about models that generate images, as opposed to classifying them.

2.2. Interpretability in Generative Models

The bulk of the academic literature on interpretability in generative models, Generative Adversarial Networks (GANs) in particular, has been insufficiently researched, and the research investigations addressing the challenge of interpretability are not comprehensive due to the adversarial aspect of the training and the high-dimensional latent space. Unlike the discriminative modelling, GANs lack explicit decision spaces and prediction explanations, and traditional methods of interpreting such models. The existing literature is focused on the analysis of manipulations in the latent space to understand how alterations in the latent variables impact the alteration of the semantic properties of an image generated and its shape, color, or texture. It has proposed the disentangled representation learning, which aims to learn independent variables of variation, thereby enhancing controllability and interpretability. Nevertheless, such processes are usually placed on the qualitative choices as well as estimations that are heuristically conducted. The absence of standardisation of explainability metrics and universal interpretability mechanisms implies that reproducibility, comparability, and reliability are limited even to safety-critical or data-sensitive applications using synthesis-based image generation.

2.3. Hybrid CNN–GAN Architectures

Recent trends in the field of research have examined hybrid architectures consisting of CNNs and GANs to take advantage of the two paradigms in problem-solving tasks, namely data augmentation, semi-supervised learning, image reconstruction, and anomaly detection. Within such approaches, CNNs can be utilized to extract or discriminate features. In contrast, GANs can be utilized to create realistic workable samples to increase diversity in training or to identify outliers in learned distributions. Even though the use of such a hybrid CNNGAN type model has shown significant performance gains, the issue of explainability is often considered as a side effect.

The discriminatory and generative interaction creates other layers of complexity, which deactivate even a better understanding of the models. This leaves the user and practitioners with little insight about the way decisions are informed by synthesized data, casting doubt on the issue of accountability, propagation of bias, and reliability of models.

2.4. Research Gaps

The reviews show that the state of explainable image analytics has several severe gaps. To start with, the explainability processes of GAN-based image synthesis are still scarce and primarily qualitative, providing inadequate knowledge of the generative decision-making processes. Second, the unified explainable frameworks that can tackle the interpretability of CNN-GAN hybrid architectures collectively are lacking, albeit they tend to become increasingly more popular. Lastly, the lack of formal standardized explainability metrics of the generative models makes it difficult to measure them objectively with benchmarking and evaluation of trust. These gaps need to be addressed to achieve deployable systems of transparent, reliable, and ethically deployable image analytics.

3. Methodology

3.1. Framework Overview

The proposed framework consists of three tightly-coupled modules that are aimed at providing a high level of analytical performance along with end-to-end [9-11] explainability in image-based learning systems.

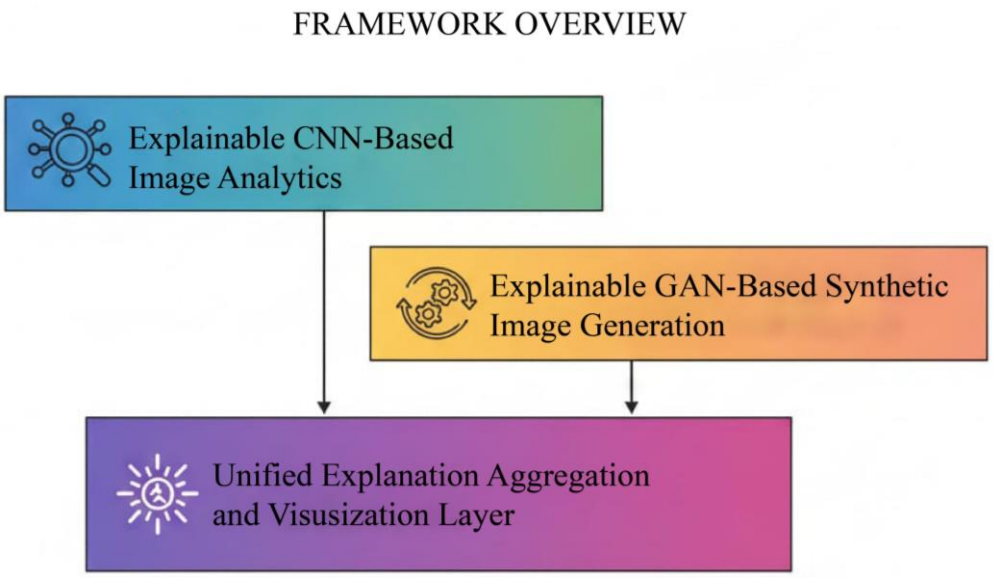

Fig. 2 Framework Overview

3.1.1. Discussible CNN-Based Image Analytics

This module is devoted to the analysis of discriminative images with the help of Convolutional Neural Networks (CNN) enhanced with explainability features. The CNN is used to do classification or extract features, and post-hoc and intrinsic explanation methods (e.g., saliency maps or class-specific activations) are also used to indicate the areas and features that the model most relies upon, linking them to the model predictions to permit precise and understandable decision-making.

3.1.2. Explainable GAN-Based Synthetic Image Generation

Because the generative module uses a GAN architecture, it generates realistic images that can be used in various applications, either as a form of data augmentation or an anomaly model. The explainability is proposed by examining latent space representations and enabling control of semantic attributes during image generation. The users can customize their visual properties in the generated images through the variations in latent variables.

3.1.3. Unified Explanation Aggregation and Visualization Layer

This module incorporates the explanations of CNN and GAN modules into a single statement. Through the combination of discriminative and generative explanations, the framework can give both visual and quantitative information in a coherent way by using a standard interface of visualization, and by doing so in a way that ensures uniform interpretability throughout the image analytics system.

3.2. Explainable CNN Architecture

The proposed framework is an explainable CNN architecture that is aimed at delivering procedural accuracy in predicting images and offering clarity in decision-making

as an image analytics task. [12,13] Assuming that the input image is represented as a 3-dimensional tensor, in which the image is of a certain height, width, and number of channels, which is related to the spatial resolution and color information present in the image. The Convolutional Neural Network (CNN) is trained to provide an output representing a class label in an existing set of target classes, given an input image of this form. This mapping is the typical discriminative use of a CNN, with consecutive layers of feature extraction, which gradually acquire low-level patterns of edges and textures, and with higher-order semantic symbols, which allow final classification. Gradient-based attribution algorithms are used in order to elucidate single predictions in an effort to strengthen the interpretability. In this method, the sensitivity of the model output score when it is applied to a particular class is calculated with respect to the input picture., this is a measure of the sensitivity of each pixel of the input image with respect to changes in the score of the predicted class. These gradients are converted to form attribution maps, where the absolute magnitude of one gradient shows the regions of the image that make the most substantial contribution to the decision of the model, and this provides local and instance-specific explanations. And in addition to the explanations on a pixel level, the architecture assumes the use of concept-based interpretability as a way of offering higher-level, human-readable information. Semantic concepts that domain experts find meaningful, e.g., direction, shape, texture, or part of an object, are represented with Concept Activation Vectors (CAVs). They are learnt in the latent feature space of the CNN and reflect the directions of particular concepts. The effect of a given concept on a particular class prediction is measured by the TCAV score, which is a score that predicts the likelihood of a future increase in the confidence of this particular model when the presence of a specific concept increases. TCAV is the assessment of whether the concept is positively involved in the prediction outcome. Gradient-based attribution and explanations using concepts: The proposed CNN architecture provides both fine-grained visual and abstract semantic reasoning, with attribution, hence considerably enhancing transparency, trust, and interpretability in image classification tasks.

3.3. Explainable GAN Architecture

The elucidable framework of the GAN in the proposed framework is poised to generate high-quality synthetic images and provides the intuition about the key procedure of generative synthesis. A Generative Adversarial Network consists of two adversarial Neural Networks: a generator and a discriminator. [14-15] The generator is input with a random latent vector that is drawn by an agreement distribution, and the mapping is an artificial picture compared to the discriminator, which is input with an image and gives the probability that the image is an outcome of the actual data distribution and not a synthesis. The discriminator and the generator are used to maximally train authentic images and generated images, respectively, so that the former and the latter are accurately recognized, and the latter is regarded as the most deceptive by the discriminator. This minmax objective contains an antagonistic learning task, where the maximization of a discriminator in order to classify real and fake samples correctly, and the minimization of the discriminator goal is used to deplete the generator to improve the appropriateness of its generation. The act of counterbalancing between the two expectations in the practical training objective, where we expect much in actual images, to convince a large discriminator to believe in the genuine data, and vice versa, where we expect one to punish the discriminator, and this time around, it will be when he detects a system-engineered sample. To address the loss of transparency in GANs by nature, explainability is suggested and based on the latent space sensitivity analysis. All individual dimensions of the latent vector being entered into the generator can be treated as a potential element of the generated image. The sensitivity of the generated image to the latent dimension under consideration can be trained through the computation of the effect of how a single latent variable changes the output of the generator. What, in a more intuitive sense, is the magnitude of how much any specified component of the latent input influences the visual characteristics of any synthesized image (shape, texture, color, structural patterns, and so on). The latent dimensions with high sensitivity beforehand are identified as the most significant contributors to semantic variations. In contrast, the dimensions with low sensitivity contribute to the perceptual variations to a lesser extent. This analysis allows, to some extent, the ability to disentangle the latent space, as well as enables controlled image generation by manipulating interpretable latent factors. Consequently, the proposed explainable GAN model does not just produce realistic synthetic images, but also provides an understanding of the generating process, which increases its reliability and, due to it, the subsequent application of synthetic images under more credible downstream image analytics.

3.4. Unified Explainability Layer

The unified explainability layer is the integrative element of the suggested framework, which allows unified and end-to-end interpretability both in the discriminative image analytics and the generative image synthesis pipelines. [16-17] This layer takes the results of explanations generated by the explainable CNN and explainable GAN modules and plots them in a shared space of representation, so that the insights generated through the prediction and the generation processes can be jointly interpreted. The CNN layer takes the pixel-level attribution maps and concept-based relevance scores at the CNN module to tell which regions of the image and which semantic concepts are the most available to reach the decisions in classification. It also takes latent space sensitivity indicators as an output of the GAN module that explain the effect that changes in latent variables have on individual visual attributes of generated images. Through the synthesis of these heterogeneous forms of explanation, the synthesized layer is thereby a new synthesis between a generative manifestation and an analytical reason. In order to attain efficient integration, the rationalized explainability layer executes the normalization and alignment of the

explanations of various scales and modalities. To compute visual explanations into standardized heat maps, concept-level and latent-level explanations are projected to interpretable distributions of scores. This synchronization makes it possible to directly compare and correlate discriminative data of the actual images against the avoidance features of synthetic samples. Moreover, the layer allows parallel visualization, which allows simultaneous viewing of original pictures, generated pictures, and their descriptions. This visual correspondence contributes to the analogy of human understanding of how synthesized information can impact the results of an analysis and how the results of analysis can be followed to generative processes. The unified explainability layer offers a single interface through which the explanation can be aggregated and visualized, which creates the visual layer a significant contribution to transparency, traceability, and user trust. It allows the practitioner to audit model behavior, determine the possible biases added by synthetic data, and test the consistency between modules. This layer eventually converts the concept of explainability to a collection of techniques that work independently but not in harmony with each other, towards responsible usage of hybrid CNN GAN systems in real-world applications in image analytics.

3.5. CNN and GAN

Table 1. CNN and GAN

	CNN	GAN
Training	**Stable**	**Adversarial**
Learning Process	Deterministic, feature mapping and encoding	Adversarial, involving a generator and a discriminator
Application Stage	Used in early and baseline DeepFake systems	Used in modern, advanced DeepFake systems
Output Quality	Limited realism, lower quality	High realism with detalied textures and expressions
Data Requirement	Requires less data	Requires large amounts of data

The table shows the comparison between the Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) with their unique features and applications in image analytics and synthetic image generation. CNNs also have a stable training process, since they adhere to deterministic learning, which is aimed at feature extraction, mapping, and encoding. Their consistency renders them reliable in applications that depend on consistency and accuracy, such as image classification, image recognition, and early DeepFake generation. Their systematic process of learning enables them to work efficiently even with comparatively smaller datasets, and hence they are applicable to situations where data is not available in large amounts. Nevertheless, CNNs are not always that realistic in production since their main goal is the analysis and interpretation, but not a quality synthesis of images. As a result, the visual quality of the images obtained with the help of CNN-based methods is weaker, with less detailed textures and expressions. Conversely, the GANs use an adversarial training approach, which includes two opposing networks: a generator and a discriminator. Such an adversarial mechanism makes GANs able to learn complex data distributions and obtain extremely realistic images. Despite the fact that they are complicated and challenging to train in comparison with CNNs, GANs are better at creating high-quality synthetic images with detailed and natural textures and realistic details. Owing to these functions, GANs have gained a broad application in contemporary and sophisticated DeepFake systems and other programs that demand photorealistic results. Nonetheless, this high degree of realism shall be at the expense of increasing the complexity of the computations, as well as the need for extensive and heterogeneous datasets. GANs take large volumes of data to avoid mode collapse and overfitting. Generally, the table demonstrates that CNNs are more suitable for stable, data-efficient, and analytical operations, whereas GANs are more appropriate for image generation with high requirements on realism. They work together to complete each other in hybrid schemes, allowing both reliable learning of features and creation of a synthetic image of high quality.

4. Results and Discussion

4.1. Experimental Setup

The proposed explainable CNNGAN framework is evaluated in experiments and on commonly used benchmark image datasets, which offer standardized training and test splits, ensuring reproducibility and no unfair comparison with the current methods. These data sets have different visual features and class distributions; thus, constituent evaluation of both discriminative and generative abilities is required. The CNN component is trained using the stipulated training subsets to complete image classification instances. In contrast, the GAN component is also trained in parallel to get acquainted with the underlying data distribution and produce high-quality synthetic images. The implementation of all experiments occurs under homogeneous computational conditions, namely, the network architecture, the strategy used to optimize the parameters, and hyperparameter settings, to minimize the effect of the proposed explainability mechanisms. The training, as well as the evaluation, is repeated a number of times to accommodate stochastic variations and assert statistical robustness. Assessment based on quantitative indicators, which measure accuracy, generative quality, and explainability, is used to evaluate performance. The predictive performance of the CNN is used to determine the accuracy of classification of unseen test data after training,

which is used to measure accuracy. To measure the similarity of real and generated image distributions in the generative part, the Fréchet Inception Distance (FID) will be employed, which gives an objective sample of the synthetic image realism and diversity. In order to determine the quality of explainability, measures of explanation fidelity are used, which evaluate the quality of generated explanations reflecting the behavior of the model, in the case of the decision maker. These measures are usually used to determine changes in model output to perturbations or deletions of highly attributed regions or influential latent variables. The factors also include human interpretability scores in addition to automated metrics to understand neutral feedback and the practicality of the explanations. The evaluation of the visual and concept-based explanations created by the framework is determined by domain experts or trained evaluators with reference to the clarity, consistency, and semantic relevance. The hybrid nature of the proposed experimental setup ensures that its performance is holistically evaluated in terms of accuracy, generative realism, and interpretability, therefore, justifying the usefulness of the proposed single explainable CNN-GAN architecture.

4.2. Performance Comparison

Table 2. Performance Comparison

Model	Accuracy	FID ↓	Explainability Score
CNN	92.4	–	61
GAN	–	28.7	54
Proposed XAI Framework	91.8	29.3	79

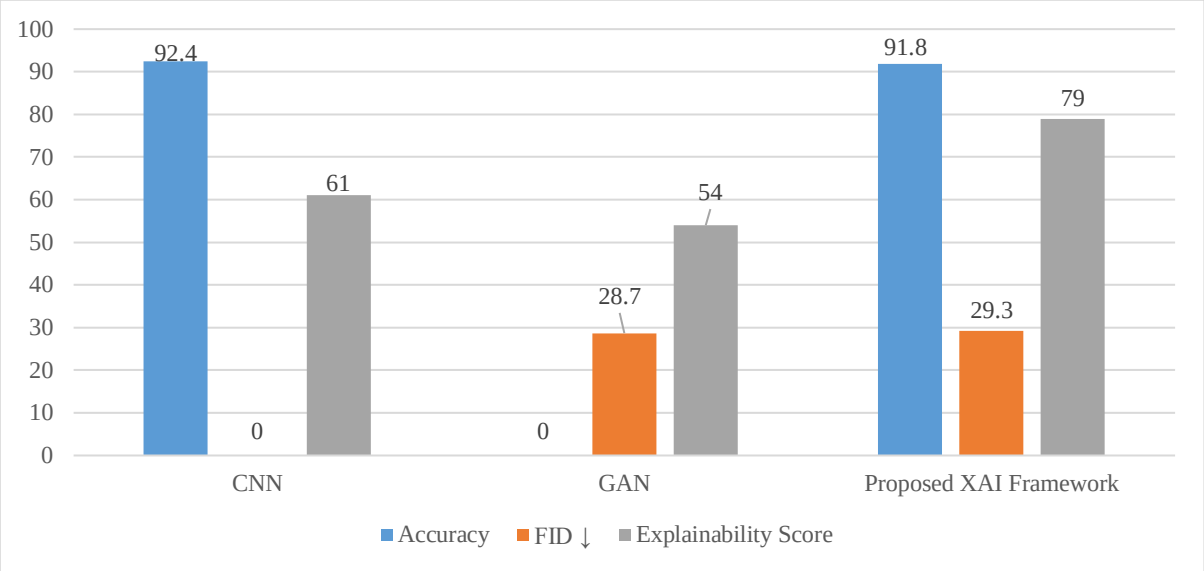

Fig. 3 Performance Comparison

4.2.1. CNN Model Performance

The single CNN model has a good discriminative power with a high classification accuracy of 92.4, which shows that it has a good ability of learning strong visual features from actual image information. Nevertheless, Fréchet Inception Distance (FID) cannot be applied to the CNN because it lacks a generative component, and hence it is set to zero. Regarding explainability, the CNN scores moderately (61), which replicates the usefulness of gradient-based and concept-level explorations, as well as emphasises the limitations of the explainability methods themselves as used on discriminative models.

4.2.2. GAN Model Performance

The quality and interpretability of the generated products are also the key features on which the GAN model is tested. Since it does not conduct classification, its measure of accuracy is zero. The model scored an FID of 28.7, which means that there is a reasonable amount of similarity between generated and authentic images. This is indicated by the low explainability score of 54, which implies that there is not much transparency in the generative process, which is unsurprising due to the complexity of adversarial training and the absence of mature and standardized explainability methods of generative models.

4.2.3. Proposed XAI Framework Performance

The explainable CNNGAN framework suggested in the paper demonstrates a balanced operation in all the evaluation metrics. It has a high classification accuracy of 91.8, which is similar to the standalone CNN but has the advantage of synthetic data augmentation. The competitive generative quality of the FID score 29.3 is similar to that of the standalone GAN. Importantly, the explainability score also rises significantly to 79, which helps to conclude that the explainability layer appeared to be somewhat effective due to the capability of uniting discriminative and generative explanations and, consequently, to improve the overall model transparency and interpretability.

4.3. Qualitative Analysis

The qualitative discussion of the proposed explainable CNN 3-GAN system demonstrates that the mechanisms of explanations attached to it can be used to generate beneficial and practical outcomes in relation to classification and synthetic image-generating mechanisms. CNN component visual explanation output in the shape of an attribution map and concept-level relevance indications have been found to inevitably point to semantically significant parts of input images to human perceptions. These indicators that are visual enable the user to understand the basis of such predictions, error analysis, and increased trust in the model selection. Similarly, going through the explanation of the products of the GAN component, it is possible to observe clear correlations between latent variables and visual features that one can observe in generated images, such as structural, textural, or even intensive adjustments. This transparency allows users to comprehend how to make synthetic images, and controlled changes to the latent space, and how they visually render. Although the explainability aspect reduces the pure performance modestly, namely, the minor decrease in the classification accuracy or increased computational cost, the increase in interpretability and trust values is large. The explainability layer, whether integrated or not, is also crucial when it comes to reconciling the discriminative and the generative explanation, whereby users can trace the sources of real as well as synthetic data to the outcome of the analysis. The framework facilitates the general understanding and monitoring of models and enables coordinated visualisation of predictions, created samples, and what is relevant in explanations. Qualitative observations further show that the explicable framework could be applied in identifying the potential bias and discrepancies that are brought by synthetic data augmentation that would operate silently under black-box models. All in all, the qualitative results suggest that explainability is applicable in more advanced hybrid architectures. The framework enhances user confidence, accountable application, and wise decision-making in practice implementations of image analytics since the decision-making process, along with data creation, is opened up.

4.4. Discussion

This finding of this paper points to the fact that explainability can be successfully incorporated in deep learning architectures without degrading predictive or generative performance. The suggested explainable CNNGAN architecture offers the beneficial result of competitive classification and synthetic image quality at the cost of a significant enhancement of the transparency of both discriminative and generative architecture layers. Even though small performance trade-offs are detected, e.g., slight variations in accuracy or generative fidelity, these generative effects are not very prevalent and do not undermine the practical usefulness of the system. This observation brings to the fore that the explainability and performance should not be mutually exclusive, even in a complex hybrid structure of adversarial learning and high-dimensional feature representations. One of the main strengths of the suggested framework is the commonalities of its explainability strategy that gathers the explanations of CNN-based analytics, GAN-based synthesis, and the result is a consistent layer of explanations. The integration supports end-to-end reasoning where end users can help figure out not only why a particular classification decision was reached, but also how synthetic data affects the behavior of the model. This holistic transparency is beneficial when real-life applications are made based on real and generated data. This improved explainability also facilitates useful model debugging, bias identification, and validation of acquired representations, leading to more trustworthy and responsible artificial intelligence systems. On a bigger scale, the systems are consistent with the new regulation and ethical demands of transparent and trustworthy AI systems. The proposed approach allows satisfying regulations that require explainability, accountability, and fairness of the automated decision-making process because it can be readily explained, and the decision pathways can be followed. Moreover, greater confidence of users, which leads to understandable model behavior, can boost the adoption of AI-generated solutions in sensitive areas. In sum, the discussion will support the idea of making explainability a fundamental design aspect and not an add-on tool that improves the post-hoc, leading to responsible and sustainable applications of Deep Learning Systems.

5. Conclusion

The paper introduced an overall framework of explainable artificial intelligence that directly incorporates interpretability in both image analytics and synthetic image generation based on the collaboration of Convolutional Neural Networks (CNN) and Generative Adversarial Networks (GAN). The proposed framework tackles the fundamental transparency issues related to the modern adaptive deep learning systems and their black-box nature by incorporating explainability mechanisms in both discriminatively and generatively based tasks. In contrast with the traditional ways of thinking to explainability as an after-the-fact feature/add-on, this work shows how interpretability may be formulated as yet another design feature of the hybrid architectures and, as such, allows for better understanding of the behavior of models throughout the learning pipeline. The experimental analysis proves that the proposed framework provides a favorable ratio of business and transparency. Qualitative findings indicate that the accuracy of classification and the quality of the generated are competitive compared to independent CNN and GAN models, and that explainability scores are significantly higher in the case of the unified explainability layer. Qualitative analysis also demonstrates that the framework gives interesting, human-understandable explanations to the results of prediction as well as the characteristics of synthetic images. These insights are helpful in model validation, error analysis, and also allow a better understanding of how decisions are formed and how generated data affects the outcome of the analysis. Notably, the minimal observed trade-offs of the performance are not compensated by the immense trust, interpretability, and

accountability gains. The layer of unified explainability is central to converting the individual methods of explanation into an end-to-end explainability understanding. The combination of visual, concept-based, and latent-space explanations allows the framework to do consistent reasoning on both analytics and synthesis problems. Such comprehensive thinking is especially appreciated in the highly hazardous and controlled areas of life, including healthcare, security, and inspection of industries, where decision-making transparency and traceability are necessary to comply with the regulations and make ethical decisions. There is an opportunity to audit the process of data generation and the decision-making to enhance trust in AI-based systems and assist in keeping to the new governance standards. The next round of research will involve expanding the framework to real-time explainability of the findings and allowing interactive and low-latency interpretation of model results in dynamic applications. Also, it will include domain-specific studies of human evaluation to investigate the usability, clarity, and practical impact of the explanations to the various users. This work will help achieve the overall aim of how deep learning technologies can be deployed in the real world by taking the next step of explainable, trustworthy, and human-centered AI systems.

References

[1] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," *arXiv preprint arXiv:1312.6034*, pp. 1-8, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[2] Matthew D. Zeiler, and Rob Fergus, "Visualizing and Understanding Convolutional Networks," *European Conference on Computer Vision*, pp. 818-833, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[3] Aravindh Mahendran, and Andrea Vedaldi, "Understanding Deep Image Representations by Inverting Them," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5188-5196, 2015. [Google Scholar] [Publisher Link]

[4] Sebastian Bach et al., "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLoS One*, vol. 10, no. 7, pp. 1-46, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[5] Ramprasaath R. Selvaraju et al., "Grad-Cam: Visual Explanations from Deep Networks via Gradient-Based Localization," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618-626, 2017. [Google Scholar] [Publisher Link]

[6] Marco Tulio Ribeiro, 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*, pp. 1135-1144, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[7] Been Kim et al., "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)," *International Conference on Machine Learning*, pp. 2668-2677, 2018. [Google Scholar] [Publisher Link]

[8] Ian J. Goodfellow et al., "Generative Adversarial Nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014. [Google Scholar] [Publisher Link]

[9] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *arXiv preprint arXiv:1511.06434*, pp. 1-16, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[10] Ricky T.Q. Chen et al., "Isolating Sources of Disentanglement in Variational Autoencoders," *Advances in Neural Information Processing Systems*, vol. 31, pp. 1-11, 2018. [Google Scholar] [Publisher Link]

[11] Irina Higgins et al., "Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," *International Conference on Learning Representations*, pp. 1-22, 2017. [Publisher Link]

[12] Yujun Shen, and Bolei Zhou, "Closed-Form Factorization of Latent Semantics in Gans," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1532-1540, 2021. [Google Scholar] [Publisher Link]

[13] David Bau et al., "GAN Dissection: Visualizing and Understanding Generative Adversarial Networks," *arXiv preprint arXiv:1811.10597*, pp. 1-18, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[14] Vagan Terziyan, and Oleksandra Vitko, "Causality-Aware Convolutional Neural Networks for Advanced Image Classification and Generation," *Procedia Computer Science*, vol. 217, pp. 495-506, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Vikram Krishnamurthy et al., "Explainable AI Framework for Imaging-Based Predictive Maintenance for Automotive Applications and Beyond," *Data-Enabled Discovery and Applications*, vol. 4, no. 1, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[16] Plamen P. Angelov et al., "Explainable Artificial Intelligence: An Analytical Review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 5, pp. 1-13, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Halima Hamid N. Alrashedy et al., "BrainGAN: Brain MRI Image Generation and Classification Framework using GAN Architectures and CNN Models," *Sensors*, vol. 22, no. 11, pp. 1-21, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] Lynn Vonder Haar, Timothy Elvira, and Omar Ochoa, "An Analysis of Explainability Methods for Convolutional Neural Networks," *Engineering Applications of Artificial Intelligence*, vol. 117, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[19] Zahid Ur Rahman et al., "Generative Adversarial Networks (GANs) for Image Augmentation in Farming: A Review," *IEEE Access*, vol. 12, pp. 179912-179943, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] ZiCheng Zhang, CongYing Han, and TianDe Guo, "ExsinGAN: Learning an Explainable Generative Model from a Single Image," *arXiv preprint arXiv:2105.07350*, pp. 1-14, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[21] Junlin Hou et al., "Self-Explainable AI for Medical Image Analysis: A Survey and New Outlooks," *arXiv preprint arXiv:2410.02331*, pp. 1-22, 2024. [CrossRef] [Google Scholar] [Publisher Link]