Proceedings of International Conference on "Recent Innovations in Engineering, Technology & Science for Sustainable Development (RIETSSD-2026)"
Career Point International Journal of Research (CPIJR)
©2022 CPIJR | Special Issue | ISSN : 2583-1895
Conference Paper | DOI: https://doi.org/10.5281/zenodo.20470289

Interpretable and Trustworthy Deepfake Detection Framework: Leveraging Transfer-Learned CNNs with Grad-CAM and SHAP for Robust Media Forensics

MD Adil Muzaffar[1], Dr Ihtiram Raza Khan[2]

[1]Research Scholar, Department of Computer Science & Engineering,
SEST, Jamia Hamdard, University, Delhi, India
adilmuzaffar96@gmail.com

[2]Associate Professor, Department of Computer Science & Engineering,
SEST, Jamia Hamdard, University, Delhi, India

Abstract: The rapid evolution of Deep-fake technologies has enabled AI based techniques like Generative Adversarial Networks (GANs) to create incredibly realistic yet totally fabricated video and image content. These developments are very exciting; however, they have serious implications to many areas including; financial fraud, misinformation, Identity Theft and Erosion of Public Trust. A significant weakness of most existing detection mechanisms is the lack of transparency- Many operate as "Black Boxes" which will identify fake media but provide no explanation for why this was done; Therefore, Most are untrustworthy. The purpose of this research paper is to develop an advanced Deepfake Detection Framework that is capable of identifying manipulated media at a very high confidence level; In addition, Provide the user with a clear and understandable justification for every prediction. The framework uses transfer learning from pre-trained CNN models: Xception and ResNet50. It is trained on diverse, publicly available datasets and follows a structured preprocessing pipeline consisting of face detection, alignment, resizing, and data augmentation in order to improve real-world robustness. This is demonstrated by embedding Explainable AI (XAI) techniques, Grad-CAM and SHAP, that highlight the particular facial regions responsible for the model's prediction. For instance, the heatmap could convey that the eye or mouth region looks unnatural, which might hint that this is where the system bases its decision on whether something is fake. Combining compelling classification with both visual and numerical explanations will help the system to build users' confidence in its potential applications to real-world digital forensics, content moderation, and media verification tasks.

Keywords: DeepFake Detection, Explainable AI (XAI), Transfer Learning, Grad-CAM, SHAP, Generative Adversarial Networks (GANs), Digital Trust

INTRODUCTION

In the modern digital era, the difference between real and fake media is very hard to detect by the naked eye because of the rapid advancement of deepfake technology, which poses a significant threat to the authenticity and integrity of digital media [1][2]. In the era of false news, people and society as a whole fear that they will be unable to believe anything they see online [2]. The word "Deepfake" was coined from the words "deep learning" and "fake". Deepfakes are artificially generated media through the use of sophisticated AI methods to manipulate images, videos, and audio, producing hyper-realistic but completely fabricated content [3]. Such manipulations are usually generated by Generative Adversarial Networks(GANs), comprising a pair of neural networks that compete against each other. While one is used as a generator, the other one operates as the discriminator, and competition between them yields images that at times appear hyper-realistic and adequately natural to the human eye.

The fast pace at which this technology is developing poses a serious threat to the credibility and authenticity of digital media. While deepfakes can be used in an amusing context, lately they have been increasingly applied for malicious intentions [4]. These technologies threaten both privacy and national security since they have the potential for use in cyberattacks, financial fraud, influencing elections, and the spreading of fake news. Moreover, deepfakes can be applied for defamation, cyberbullying, blackmailing, and to harm reputations and their psychological consequences on individuals due to non-consensual pornography [5][6]. Therefore, there is a dire need for serious detection methods that are widely applicable to counter these threats.

In 2017 a Reddit user showed how deep learning could be used to swap faces in videos, and since then fake media has grown very fast [9][10]. For instance, deepfakes were once used in one of the "CEO Fraud" cases in which criminals, using AI to impersonate a CEO's voice, had managed to swipe $243,000 from a UK-based energy company [7]. According to the Sensity survey, a company that monitors deepfake films online claims that, since 2018 the number has doubled every six months reaching 85,047 videos at the end of December 2020 [11]. Tools like “DeepFaceLab” and “FakeApp” made it easy for anyone to create realistic face-swaps, so harmful uses became more common. Deepfakes can be extremely harmful

while being used to create fake news. Politicians, including US presidents Barack Obama and Donald Trump, have also fallen victim to deepfakes in the past [12].

To fight this, researchers have developed various detection methods. This is very promising, since most deep learning-based detection schemes, including CNNs and LSTM networks, automatically learn and extract features from data [4][7][8]. However, one problem remains: most of the deepfake detection models are "black boxes." While they reach high accuracy on a certain dataset they have been trained on, even experts cannot interpret the concrete reasons behind their decisions. This lack of transparency results in a gap between model performance and user trust, particularly in high-stakes domains where decision accountability is essential. If a system flags a video as fake but cannot explain why, it is hard to use that finding in a court of law or for journalistic verification.

In this study, the proposed framework works as a robust and explainable deepfake image detection system. It employs powerful CNN-based transfer learning models capable of classifying images into real or fake, such as "ResNet50" and "Xception". To make the image detection process more transparent and trustworthy, the framework will make use of explainable AI models such as Grad-CAM and SHAP, which will help to depict exactly which parts of the image influenced the model's decision in this regard. This will improve accuracy and, most importantly, will help build trust in the results.The discussed framework represents a more reliable and transparent solution for preserving digital media authenticity by taking advantage of state-of-the-art methods. Future work in this area should be oriented towards setting better standards for testing these models on more diverse, real-world datasets to ensure they can generalize well against new attack types.

The paper is structured to achieve a comprehensive understanding of the proposed approach. Section 2 outlines the problem statement, highlighting how most of the proposed detection methods have very limited applicability to real-world problems in terms of accuracy and explainability. Section 3 provides a detailed literature review. Section 4 describes the proposed solution, including its data strategy, preprocessing pipeline, detection models, training workflow, and explainability layer. Table I gives an overview of key findings from reviewed literature, summarized for clarity and comparison with methodologies, datasets, metrics, strengths, and limitations. Similarly, Table II summarizes some of the important datasets used in the field of deepfakes, including types, size, availability, links, supporting reproducibility, and dataset diversity during both training and evaluation.

PROBLEM STATEMENT

While it has become a lot easier to create very realistic deepfake images, it has also become quite hard to tell whether the media is real or fake [9][13]. This rising challenge threatens to erode public confidence in digital content, where people cannot tell anymore whether what they see is real or manipulated. Misinformation spreads fast in such an environment, leading to devastating social, political, or financial outcomes.

There have been many detection-based deep learning models, and most of them work as black boxes [3][8]. Although they have high accuracy in predictions, they never tell us how they reached their decisions. Such black-box systems, although quite accurate, provide no transparency or justification behind their classifications. This becomes a critical limitation when such systems are expected to be used in serious areas like forensic analysis or content verification.

They are not very trustworthy, and so it is hard to adopt them in areas like digital forensics, journalism, and legal verification [5][15]. For instance, legal proceedings would require more than the output of a model, with necessary evidence or clear reasoning for support. Without transparency, such systems cannot be used with full confidence as trusted tools in sensitive sectors.

The majority of the models trained on a particular dataset do not generalize to another source or even real-world image conditions, i.e., compression, noise, or blurring. In fact, this largely reduces their performance outside the environment they were exposed to during training. From their findings, there is an urgent need for a detection system that is not only accurate and efficient but also explainable and generalizable.

This paper proposes deepfake detection using a CNN-based transfer learning model, using XAI techniques such as Grad-CAM and SHAP together with it. Finally, their integrated model shall improve the accuracy and explainability of the model in order to bridge the gap between the performance of deep learning and human trust.

LITERATURE REVIEW

Deepfakes are synthetic images, audio, or videos made with advanced AI so that they appear real but are not. They are typically created by deep learning techniques like Generative Adversarial Networks, autoencoders, and modern face-swap or lip-sync pipelines. As these tools continue to improve, it is becoming increasingly challenging to distinguish real from forged media with the naked eye, while false content can also lead to significant harm in terms of privacy, reputations, elections, and security. Many detection methods have been proposed: image-level CNNs and transfer learning models that identify artifacts in pixels and

textures, temporal models that detect abnormal motion in videos, forensic approaches analyzing frequency traces or blending, and more recently explainable AI methods that provide information on model decisions. This review of the literature surveys those families of methods, compares their strong points and weaknesses, and points to why robust detection should be combined with clear explanations if it is to be useful in real-world settings.

3.1 General Landscape: Evolution, Datasets & High-Level Findings

Garg and Gill [9] exploratory review explains how deepfakes grew from a 2017 Reddit face-swap example into widely available tools such as “DeepFaceLab” and “FakeApp”. The paper enumerates the common datasets used by the researchers-FaceForensics++, Celeb-DF, DFDC, UADFV-and explains why these datasets matter. They provide manipulated and real videos under different qualities and compression levels; these datasets are reused by researchers to test and compare the detectors. The main takeaway from the review seems to be that while these generation tools became easy to use, the detection methods still struggle to generalize across different datasets and video qualities. It asks for "universal" detectors working under many conditions and for clear evaluation standards, such as cross-dataset tests and robustness to compression.

The broader review papers add that the detection research is fragmented, with some works targeting image artifacts, some targeting motion inconsistencies, while others focus on traces in frequency space. The reviews consistently flag the same problems: dataset bias (models overfit to the dataset they were trained on), poor cross-dataset generalization, and a lack of interpretability for many high-accuracy systems. These high-level findings motivate to emphasize both accuracy and explainability in the proposed method.

3.2 Image-level detectors and transfer learning (CNN-based approaches)

Kaushal et al [16]. conducted a head-to-head comparison among different pretrained CNN backbones, such as ResNet50, Xception, and EfficientNet. Their approach consists of fine-tuning each pretrained network on standard deepfake datasets, using the same preprocessing, which includes face cropping and resizing, and then comparing the performance. According to them, Xception often performs best for spotting subtle texture and blending artifacts because of its depthwise separable convolutions, while ResNet50 gives stable results across datasets and is easier to train. The paper emphasizes the practical benefit of transfer learning: given the limited number of labeled deepfake images, starting from ImageNet weights

accelerates training and improves accuracy. They note that the main limitation is sensitivity to the datasets used for testing.

Kumar et al [7]. (Deepfake Detection Using AI and Machine Learning Algorithms) tested pipelines that combine traditional feature extractors with deep CNN classifiers. They underline the importance of careful preprocessing before passing images to a CNN, including face detection with MTCNN, alignment, and normalization. The experiments they conducted on a medium-sized dataset show very high accuracy, roughly 93-94%, but the paper warns that small or custom datasets can give optimistic results that may not hold in the wild. The limitation consists of the size of the dataset used and the need for more test data with more variations.

These reviews and comparative studies together give a practical rule of thumb: for image-level detection, use a strong pre-trained CNN backbone, good face preprocessing, and augmentation. These form strong baselines that are easy to reproduce. However, image-only models miss temporal cues in video and thus best suit still-image scenarios or serve as frame-level building blocks for video methods.

3.3 Temporal and geometry-based methods

Xiong et al [8]. propose a different angle: instead of raw pixels, they track facial landmarks (68 points) across frames, normalize coordinates, and feed the resulting time series into an LSTM. Their goal was to capture geometric and motion inconsistencies-for example, unnatural eye blinking, mouth motion, or head jitter that a face-swap algorithm may not reproduce consistently. Evaluated on FaceForensics++ and UADFV, this landmark+LSTM approach attained very high accuracy (reported ~93.5% on FaceForensics++), and it was robust to video compression and low-quality frames. The drawback is that the method depends on reliable landmark detection; if faces are occluded or at extreme angles, landmark extraction can fail, and the detector's performance drops.

Other approaches in this class combine the CNNs for the per-frame feature extraction and sequence models - LSTM/GRU or 3D-CNN - for temporal modeling. The pipeline commonly used is: detect and crop faces per frame → extract CNN features per frame → feed sequence to LSTM → classify. These papers show temporal modeling improves detection on video tasks because many manipulations are inconsistent over time even if single frames appear real. Temporal modeling increases complexity and requires more time to train than image-only models.

3.4 Hybrid, fusion and ensemble models

Hussein et al [3]. proposed a fusion architecture that leverages the image representation capabilities of ResNet50 for local texture and Vision Transformer for global context. Their key contribution is a cross-attention-based fusion block: instead of simply performing concatenation, cross-attention allows the network to weigh which CNN features with which transformer features should be emphasized together. This provides the model with both fine detail and an understanding of whether different parts of the face "fit" together, which helps it spot sophisticated fakes. Results on FF++ and Celeb-DF demonstrate improved accuracy compared to single backbones. On the flip side, it has higher computational and memory costs due to running two large models on every image.

Kumar et al [7]. presented an ensemble that combined several CNNs-VGG and Xception, among others-a custom-designed ConvNet, where outputs were combined with a meta-learner. They also used LIME for explaining the ensembled decisions. A larger array of manipulation styles was handled by this model, and the detection stability was improved. Strong accuracy, ~94.5%, and good AUC were reported in this study. Limitations include a much more complex model and slow inference, which may be problematic to apply in real-time. It is also difficult to interpret decisions in ensembles, unless XAI tools are carefully integrated.

According to the literature, hybrid/fusion models are powerful when accuracy and robustness are the priorities, but they cost more to train and run. They would be suitable for forensic tools or research prototypes rather than mobile or real-time systems unless pruned or optimized.

3.5 Explainable AI (XAI): Grad-CAM, LIME, SHAP and practical trade-offs

Several papers focus on making detectors explainable. For example, Sugiantoro(2024)[5] applied Grad-CAM to ResNet variants and demonstrated that heatmaps reliably highlight regions which drive a “fake” decision. Such a scheme provides evidence to users and examiners that the model has looked at plausible (eyes, mouth, hairline seams) artifacts. Due to speed reasons (one backward pass) and the visual intuitiveness of such a method, this might be useful for demonstrations or manual checks. The limitation is that heatmaps are coarse and do not provide numeric importance scores.

Roshinta & Gábor [17] compared LIME and SHAP: LIME is faster in explaining single images, by perturbing superpixels and fitting a simple surrogate model, while SHAP yields Shapley-value-based attributions that are more theoretically solid and consistent across

examples, but it requires many model evaluations and is computationally heavy. Their recommendation — also repeated in survey papers — is practical: use Grad-CAM for fast visual checks, LIME for quick local explanations, and SHAP for in-depth auditing of important or ambiguous cases.

The survey "Revealing the Unseen" synthesizes these findings and argues that explainability is needed for adoption in journalism, forensics, and court settings. The key message of this is that a detector has to do more than simply label a content as fake or real; it needs to provide evidence readable by humans that can also be verified. Explainability also helps in finding model weaknesses and biases during development.

3.6 Broader Reviews, Advanced Methods, and Future Directions

Recent review papers, such as Kumari and Singh [12] and Garg and Gill [9], discussed many modern techniques for the detection of deepfakes. These vary from specific models that focus on certain manipulated facial parts to transformer models that take into consideration both small and large features of an image and methods that check if the audio and video are synchronous. Some papers also explore ideas such as models that learn without needing a lot of labeled data, the combination of sound and video for better accuracy, and blockchain to prove whether media is real.

TABLE I. COMPARISON TABLE

Author(s) & Year	**Methodology**	**Dataset(s)**	**Evaluation Metrics**	**Key Findings**	**Limitations**
Garg & Gill (2023)[9]	Exploratory survey on deepfake tools and datasets	FaceForensics++, Celeb-DF, DFDC	Qualitative analysis	Emphasizes need for cross-dataset testing and better benchmarks	Weak generalization of models across datasets
Kaushal et al. (2022)[16]	CNN-based transfer learning	FaceForensics++, Celeb-DF	Accuracy, Recall	Xception best for texture; transfer learning boosts	Dataset-specific tuning, limited

	(ResNet, Xception, etc.)			performance	generalization
Kumar et al. (2024)[7]	CNN pipeline with preprocessing	Custom deepfake dataset	Accuracy (~94%)	Strong results with proper preprocessing	Small dataset, possible overfitting
Xiong et al. (2025)[8]	LSTM on facial landmark time-series	UADFV, FaceForensics++	Accuracy (~93.5%)	Captures temporal and geometric inconsistencies	Sensitive to occlusion, landmark errors
Hussein et al. (2025)[3]	ResNet + ViT hybrid with cross-attention	FF++, Celeb-DF	Accuracy, Precision	Combines local and global features effectively	High computation and memory usage
Kumar et al. (2024)[7]	Ensemble CNNs + LIME-based explanation	Custom Dataset	Accuracy (94.5%), AUC	Combines multiple detectors for stability and explanation	Complex, harder to deploy in real time
Sugiantoro (2024)[5]	CNN with Grad-CAM heatmaps	Deepfake Face Datasets	Visual maps	Highlights decision regions for explanation	Coarse outputs, not numerically informative
Roshinta & Gábor (2024)[17]	LIME vs. SHAP for XAI	Deepfake Images	XAI Consistency, Speed	SHAP is more accurate; LIME is faster	SHAP is computationally expensive
Kumari & Singh (2024)[12]	Review of advanced methods (transformers, blockchain)	Multi-modal datasets	Broad synthesis	Push for multimodal, XAI, and self-supervised approaches	Most models are resource-intensive and lack real-world validation

However, most of these models require very powerful computers and take immense time to train. Some perform well on very specific datasets and then completely fail if tested with new or real examples. Kumari and Singh [12], and Garg and

Gill [9] agree that an ideal approach would be to formulate robust models that are interpretable-using explainable AI-and test them on a wide range of data types for their potential applications within realistic scenarios.

To conclude this literature review, this paper demonstrates real progress but no single perfect solution. Many powerful detectors exist-especially CNNs with transfer learning-there are useful video methods like LSTM on landmarks, and powerful hybrid or ensemble systems. However, each has trade-offs in speed, cost, or how well it works on new data. Explainable tools like Grad-CAM, LIME, and SHAP are becoming essential so people can see why a model made a decision. Table 1 summarizes the key works by listing the method each paper used, which datasets they tested on, how they measured results, their main findings, and limitations. From this review, at least one practical next step is clear: build a solid image-level detector using transfer learning, add simple XAI outputs for human checks, and evaluate the system across several datasets-and later on video-so that it is both accurate and trustworthy in real situations.

TABLE II. DEEPFAKE DATASETS TABLE

S.No.	**Dataset Name**	**Type**	**Real:Fake Ratio**	**Total Samples**	**Publicly Available**	**Link**
1	UADFV	Video	1:1	98	No	-
2	FaceForensics++	Video	1:5	~6,000	No	-
3	DFDC	Video	1:5	~128,000	Yes	https://ai.facebook.com/datasets/dfdc/
4	Celeb-DF	Video	1:9.5	~6,229	Yes (on request)	https://github.com/yuezunli/celeb-deepfakeforensics
5	ForgeryNet	Image/ Video	1:1	2.9M / 221K+	Yes	https://yinanhe.github.io/projects/forgerynet.html#download
6	DFFD	Image	1:5	~299,000	Yes	http://cvlab.cse.msu.edu/category/downloads.html
7	FakeAVCeleb	Audio/ Video	1:19	~20,000	Yes	https://fakeavceleb.github.io/

8	FFHQ	Image	-	70,000	Yes	https://github.com/NVlabs/ffhq-dataset
9	iFakeFaceDB	Image	-	87,000	Yes	https://github.com/socialabubi/iFakeFaceDB
10	VGGFace2	Image	-	3.31 million	Yes	https://github.com/ox-vgg/vgg_face2

PROPOSED SOLUTION

This paper proposes a hybrid with an explainable deepfake detection system that will not only classify an image as real or fake but also explain why such a decision has been made. The system will use strong, proven CNN backbones under transfer learning and attach explainability layers Grad-CAM for fast visuals and SHAP for detailed attribution. In real life practice, high accuracy is not enough: real users need reliable and general results with clear evidence they can inspect. All in all, The aim is to build a detector that will provide good performance, can handle different data sources, and provide human-readable explanations for its outputs.

4.1 Data strategy

The model will be trained and tested on several public datasets-a mixture of image and frame extracts from video datasets-so that it sees many kinds of manipulations. Using such diverse datasets helps the model to learn general patterns, not just quirks of one source. In practice, this combines large image datasets with frames sampled from other sets for realistic, high-quality fakes. Detectors trained on one dataset usually fail when applied to another; mixing sources gives stronger, more realistic performance.

4.2 Data preprocessing pipeline

Each image or every frame from videos will be prepared very carefully in the pre-processing step to help the model learn better. First comes the face detection step, where the face of the person is located within the frame. The face area is cropped immediately after that, where unnecessary background is cut, focusing only on its important features. Then comes the alignment of the face so that the key points in each face, like the eyes, nose, and mouth, would take roughly the same positions in all the images, thereby reducing any confusion and keeping learning as consistent as possible.

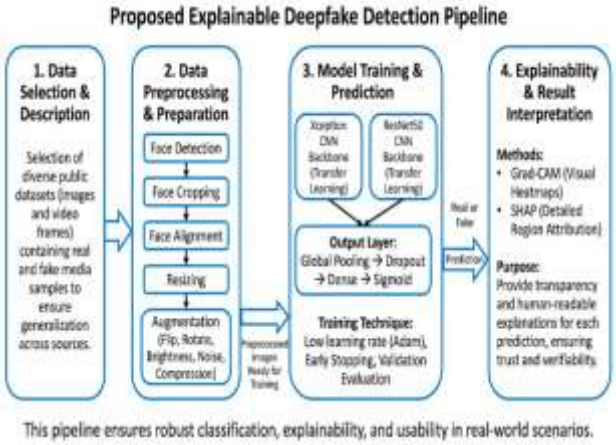
Next, the image is resized to fit the model's required input size. Several small changes are also made to each image in order to mimic the kinds of imperfections seen in real-world data. These include flipping the image, slightly rotating it, adjusting brightness, compression, and adding mild noise. These steps are called augmentation, and they train the model to recognize deepfakes even if the input is blurry, low quality, or altered. Together, these preprocessing steps ensure that the model learns to focus on meaningful facial features and stays strong across different visual conditions.

4.3 Detection models

For deepfake detection, powerful CNN models will be used that have already been trained on large image datasets; this is referred to as transfer learning. The two base models used are "Xception" and "ResNet50". Xception is the first choice since it's really good at finding small texture changes and blending marks common in fake images. ResNet50 is also used as a backup model because it gives stable and reliable results. In both models, The original top layers will be removed and add our own simple classifier. This includes steps like pooling of features, adding dropout to prevent overfitting, and finishing with a dense layer that outputs the final prediction — real or fake.

Transfer learning reduces the necessity of large collections of labeled deepfake images, as pre-trained models already provide strong and transferable feature representations. Multiple models further rise confidence in the detection outcome-if multiple architectures nod toward one single conclusion, the prediction is likely to be reliable. If their outputs differ, explainable AI techniques such as Grad-CAM or SHAP can be consulted to better understand the underlying reasoning behind each model's decision.

4.4 Training and prediction workflow



This pipeline ensures robust classification, explainability, and usability in real-world scenarios.

The model will carefully train in order to avoid overfitting. will use the Adam optimizer with a small learning rate, so the model may learn slowly but correctly. early-stop the training if performance on the validation set stops improving, train each model separately, and finally test its performance on data it has never seen, even from different datasets. Check some important results such as accuracy, AUC, precision, and recall, but also analyze the wrong predictions to understand where the model performs badly. That way, The model will not only good on the training data but can also work correctly in a real-world scenario.

4.5 Explainability layer

It will explain the decision of the model for its prediction using two tools. Firstly, "Grad-CAM", which generates a heat map that illustrates which part of the image the model looked at the most. This gives a fast understanding of where the model is focusing on. Then, apply "SHAP" on important or confusing cases, where it segments the image into pieces and informs us about the contribution rate of each piece to the decision of the model about whether the image is real or fake. By using both Grad-CAM and SHAP, we get a clear and useful explanation of how the model thinks, thus making the results more trustworthy for people.

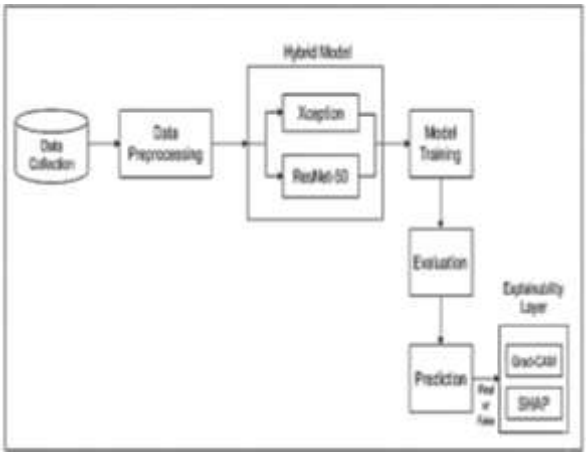


FIG I. FLOW CHART

To conclude, this solution couples deep learning models with explainability tools in order to build up a reliable deepfake detection system. The system, by training on diverse datasets, doing careful preprocessing, and using two performing CNN backbones, has as an end target high accuracy that generalizes across real-world cases. Furthermore, the addition of Grad-

CAM and SHAP will make sure that the predictions are not only accurate but also interpretable; this gives the user clear evidence to trust and verify each result.

FIG 2. PROPOSED SOLUTION PIPELINE

Summary

This framework addresses the increasing danger of realistic manipulated media by developing a deepfake detection system that is both accurate and explainable. This shows that simple accuracy is not enough: people, courts, and journalists need reasons they can check. this need will be fulfilled by reviewing recent detection methods, identifying common weaknesses, and proposing a practical detection pipeline that pairs strong CNN-based classifiers with explanation tools.

The problem which the paper defines is as follows: given the significant advance in generative models, fake images and videos are becoming hard to spot, and many detectors behave like "black boxes" or fail when tested on new data. The related work is organized around image-level CNN approaches, temporal/geometry-based approaches, hybrid and ensemble systems, and explainable-AI methods. Some of the biggest gaps that arise include poor cross-dataset generalization, dataset bias, and a lack of transparent decision evidence-a set of problems that motivated the design choices in my method.

The provided solution would include the following steps: (A) collect a number of diverse datasets so the model sees many manipulation styles; (B) sample preprocessing and augmentation would involve face detection, alignment, resizing, and realistic corruption, in order to make the model robust; (C) train and test transfer-learning CNNs, such as Xception and ResNet50, compare their results, perform cross-dataset and corruption robustness checks; and (D) provide human-friendly explanations by using Grad-CAM for fast visual checks and SHAP for detailed attributions on the most important cases. This also be thoroughly documenting experiments along with keeping the held-out dataset for a cross-sources evaluation-a step directly tackling the generalization problem discussed in the literature.

To evaluate, The standard metrics: accuracy, AUC, precision, recall, F1. Then provide confusion matrices, cross-dataset results, and degradation under common corruptions like JPEG, noise, resizing. For representative cases, example explanations (heatmaps and SHAP summaries) will show the readers both what the model decided and why. Deliverables will include the trained model weights, preprocessing and training scripts, evaluation tables, and a set of explanation artifacts. In sum, this work will yield a reproducible, easy-to-follow

detection system that balances the performance-interpretability trade-off. By bringing together multiple public datasets, careful preprocessing, transfer-learning backbones, and complementary XAI tools, this approach will be practical, auditable, and more trustworthy than purely black-box detectors. It positions the work as a concrete step toward detection tools that could be used and inspected in real-world settings.

References

1. M. Alrashoud, “Deepfake video detection methods, approaches, and challenges,” Alexandria Engineering Journal, vol. 125, pp. 265–277, 2025.
2. D. Pan, L. Sun, R. Wang, X. Zhang and R. O. Sinnott, "Deepfake Detection through Deep Learning," 2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT), Leicester, UK, 2020, pp. 134-143, doi: 10.1109/BDCAT50828.2020.00001.
3. Anitha, L. M. Saju, and B. Kamaraj, "Deepfake Detection using XAI based Deep Fusion Models," in *Proc. 2025 2nd Int. Conf. on Computational Intelligence, Communication Technology and Networking (CICTN)*, 2025, pp. 1–8, doi: 10.1109/CICTN64563.2025.10932587.
4. G. Chandel, A. Kumar, K. Malik, K. Gurani, K. Gahlawat, and S. K. Saini, "Deepfake Detection Using AI And Machine Learning Algorithms," in Proc. 2025 IEEE Int. Conf. on Computer, Electronics, Electrical Engineering & their Applications (IC2E3), 2025, pp. 1–7, doi: 10.1109/IC2E365635.2025.11167331.
5. B. Sugiantoro, "Deepfake face images: Explainable detection using deep neural networks and class activation mapping," in Proc. 2024 IEEE Int. Symp. on Consumer Technology (ISCT), 2024, pp. 86–89, doi: 10.1109/ISCT62336.2024.10791156.
6. Asad Malik, Minoru Kuribayashi, Sani M Abdullahi, and Ahmad Neyaz Khan. Deepfake detection for human face images and videos: A survey. Ieee Access, 10:18757–18775, 2022.
7. A. Kumar, D. P. S. J., M. P., M. Dheeraj, and A. R. Aarthi, "XAI-Empowered Ensemble Deep Learning for Deepfake Detection," in Proc. 2024 15th Int. Conf. on Computing Communication and Networking Technologies (ICCCNT), 2024, pp. 1–7, doi: 10.1109/ICCCNT61001.2024.10726125.
8. D. Xiong, Z. Wen, C. Zhang, P. Xiao, and M. Wu, "Deepfake Detection based on LSTM Networks with Facial Geometric Features," in Proc. 2025 4th Int. Conf. on Electronics,

Integrated Circuits and Communication Technology (EICCT), 2025, pp. 627–630, doi: 10.1109/EICCT65471.2025.11099876.

9. D. Garg and R. Gill, "Deepfake Generation and Detection – An Exploratory Study," in Proc. 2023 10th IEEE Uttar Pradesh Section Int. Conf. on Electrical, Electronics and Computer Engineering (UPCON), 2023, pp. 888–891, doi: 10.1109/UPCON59197.2023.10434896.
10. C. O¨ hman, “Introducing the pervert’s dilemma: a contribution to the critique of deepfake pornography,” Ethics and Information Technology, vol. 22, no. 2, pp. 133–140, 2020.
11. (2021) How deepfakes are a problem for us all and why the law needs to change. https://informationmatters.net/ deepfakesproblem- why-law-needs-to-change/. [Online; accessed 19-July-2023].
12. H. F. Shahzad, F. Rustam, E. S. Flores, J. Lu´ıs Vidal Mazo´n, I. de la Torre Diez, and I. Ashraf, “A review of image processing techniques for deepfakes,” Sensors, vol. 22, no. 12, p. 4556, 2022.
13. D. L. R. and B. B. Sujitha, "Advancements in Deepfake Detection: A Comprehensive Review of AI-Driven Approaches," in Proc. 2025 Int. Conf. on Machine Learning and Autonomous Systems (ICMLAS), 2025, pp. 1007–1011, doi: 10.1109/ICMLAS64557.2025.10967658.
14. H. Chotaliya, M. A. Khatri, S. Kanojiya, and M. Bivalkar, "Review: DeepFake Detection Techniques using Deep Neural Networks (DNN)," in Proc. 2023 6th Int. Conf. on Advances in Science and Technology (ICAST), 2023, pp. 480–484, doi: 10.1109/ICAST59062.2023.10454938.