IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Journal Européen des Systèmes Automatisés
Vol. 57, No. 5, October, 2024, pp. 1471-1480
Journal homepage: http://iieta.org/journals/jesa

Automated System for Credibility Diagnosis of Forged Facial Images Using Ensemble Deep Learning and Explainable AI

Pooja Kamat[1*], Shruti Patil[1,2], Kulveen Kaur[2], Manish Rathod[2], Preksha Pareek[3]

[1] Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 412115, India
[2] Symbiosis Centre for Applied Artificial Intelligence, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 412115, India
[3] Computer Engineering Department, Thakur College of Engineering and Technology, Mumbai 40010, India

Corresponding Author Email: pooja.kamat@sitpune.edu.in

https://doi.org/10.18280/jesa.570522

Received: 3 September 2024
Revised: 30 September 2024
Accepted: 14 October 2024
Available online: 28 October 2024

Keywords:
automated system diagnosis, credibility assessment, deep learning ensemble, explainable AI, forged image detection, facial health sector, responsible AI

ABSTRACT

In the digital age, the internet plays a pivotal role in shaping consumer perceptions, with vast amounts of data, particularly images and videos, being utilized by organizations for beneficial and deceptive purposes. In the facial health sector, forged facial images are increasingly used in advertising, misleading consumers into purchasing ineffective beauty products. To address this issue, we propose an automated system for the credibility diagnosis of facial images in online advertisements. Our approach employs an ensemble of state-of-the-art deep learning models—ResNet50V2, MobileNetV2, and InceptionV3—which achieved an ensemble score of 79%, outperforming individual model accuracies of 80% (ResNet50V2), 76% (MobileNetV2), and 75% (InceptionV3). Furthermore, stacked ensembling yielded 79% accuracy, showing a marked improvement over individual models. We also integrate Explainable AI techniques, where Score-CAM demonstrated the best performance with a 32.9% average drop and a 31.4% increase in confidence, providing interpretable visual explanations for the detected forgeries, thereby enhancing transparency and trustworthiness. By training our models on a curated dataset of real and fake facial images, we achieve robust detection and provide users with an interpretable analysis of non-credible images. This novel system not only advances the field of automated image credibility assessment but also contributes to the development of responsible AI systems that protect consumers from deceptive practices.

1. INTRODUCTION

We surely live in a time when we are experiencing a lot of visual imagery data from online sources. While we may have had faith in the purity of this imagery in the past; modern digital technology has begun to weaken that faith. Our mind has a tendency to trust what we're persuaded of. Altered visuals are emerging more often and with greater intricacy in sensationalist magazines, the fashion sector, prominent news outlets, scholarly periodicals, political endeavors, legal proceedings, and deceptive images arriving via email. The primary contributors to the dissemination of bogus photographs are social media, blogs, websites and hence these distortion helps brands to promote their product easy and convincible way specially in the field of health information. These falsified photographs undermine the legitimacy of online information in the eyes of the general public, prompting consumers to utilize items in the hopes of getting results similar to those shown in the Figure 1.

1.1 Image credibility analysis

Because of technological advancement and globalization, electronic devices are available widely and affordably available. As an outcome, cameras have increased recognition. There are numerous cameras near us, and apply it to take a significant number of images. Various documents that must be filed online require soft copies of their visuals [1], and every day huge images are shared on social media. Illiterate people take glance of photographs and take information from them, which is incredible. Hence the way individuals currently receive news has changed. To quickly learn more, they primarily use social media sites to search for a condensed version of the news [2]. In order to determine credibility, there are many aspects of it as shown in Figure 2. Hence, delving into the reasons behind the identification of an image as fraudulent proves valuable, as it can introduce fresh insights and information that might have been previously undiscovered, even among experts.

Figure 1. Fake vs real images

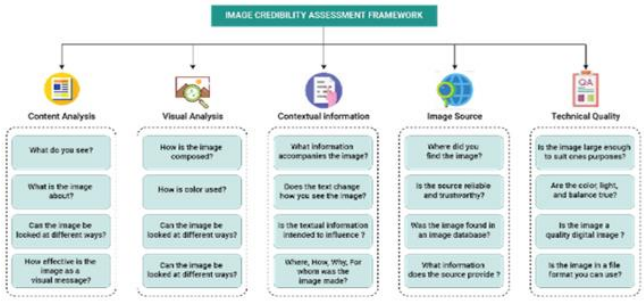

Figure 2. Image credibility assessment framework

1.2 Importance of visualization in image credibility analysis

In the context of AI ethics and performance, explainability refers to methods that clarify how an AI system reaches its conclusions [3], while interpretability focuses on how easily humans can understand the model's logic. For image forgery detection, especially in advertising, it is not enough to merely identify manipulated images; understanding why the model flagged an image as fake is crucial for building trust [4]. In Figure 3, each term invisible to the human eye, are explained. By integrating Explainable AI (XAI), this study aims to combat deceptive practices in facial health advertising, restore consumer confidence, and foster responsible AI use.The study contribution is mentioned in section D of the Introduction.

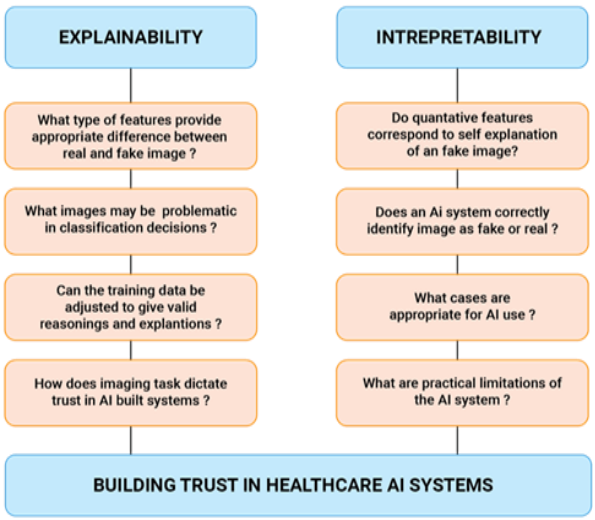

Figure 3. Impact of credibility

This transparency helps auditors grasp the key factors behind the AI's decisions, ensuring that even subtle manipulations, invisible to the human eye, are explained. By integrating Explainable AI (XAI), this study aims to combat deceptive practices in facial health advertising, restore consumer confidence, and foster responsible AI use. The study contribution is mentioned in section D of the Introduction.

1.3 Contribution of paper

Following is the contribution of the paper

- In literature, there is a lack of real and forged images which can't be differentiated by the naked eyes, authors have proposed their dataset consisting of 300+ images that are real as well as fake for implantation purposes
- After surveying Authors found that as per their knowledge very few works were done in the field of explaining the prediction in image credibility detection hence authors proposed a method in a user understandable format.
- Authors have visualized the credibility of image authors using 3 different explainable AI tools and did comparative analysis on 3 CNN models as well as their ensembled model.

Authors have used 3 different CNN for classification followed by their ensemble hence making the model more stable.

2. LITERATURE REVIEW

The literature has suggested several strategies to combat picture counterfeiting [5] define fake images as images in posts that do not accurately represent the events that they refer to. But they didn't explain why the images were not credible. First, we will describe various techniques used in image forgery and then move to explainable ai used in nearby areas. Wagle et al. [6] developed the NLP model to detect an image as fake or real. Jin et al. [7] suggested a domain-transferred CNN model that might utilize information from the auxiliary set and progressively apply it to the intended job. Mayer and Stamm [8] provided a statistical model that reflects the discrepancy between global and local LCA estimations. Then, they applied this model to formulate the problem of forgery detection as a hypothesis testing one and arrived at a detection statistic, which they demonstrated is the best under specific circumstances. Dua et al. [9] illustrated an approach employing JPEG compression. An image block, divided into discrete 8×8 pixel blocks, undergoes individual assessment of discrete DCT coefficients.

The statistical attributes of the AC components of these block DCT coefficients alter upon the failure of a JPEG compressed image. Using SVM, the recovered feature vector distinguishes authentic from counterfeit images. In the study [10], forged images were categorized into 'active,' termed non-blind, and 'passive,' termed blind. In the study [11], methods like SIFT (Scale Invariant Feature Transform), SURF (Speed Up Robust Features), GLOH, ORB (Oriented FAST), and others were applied to enhance outcomes. For detecting copy-move forgeries [12], BusterNet was proposed. A fusion module resides within a dual-branch architecture. Both branches employ visual cues to detect potential manipulation areas and visual similarities to identify copy-move regions. Wu et al. [13] employed a CNN to extract block-like features from an image, compute inter-block self-correlations, identify matching points via a point-wise feature extractor, and reconstruct forgery masks using a deconvolutional network. Wu et al. [14] introduced the fully convolutional network ManTra-Net, adaptable to various image sizes and forgery types such as copy-move, augmentation, removal, splicing, and unknown forms. These techniques were also extended to the medical domain for image forgery detection.

In 2020 cloud environment was used and a big data analytics engine, Ali et al. [15] proposed a novel healthcare monitoring framework to precisely analyze and store data of healthcare, and to enhance classification accuracy.

Ontologies, data mining techniques, and bidirectional long

short-term memory (Bi-LSTM) are used in the proposed big data analytics engine. Also, using ensemble deep learning and feature fusion, Ali et al. [16] proposed a smart healthcare system for predicting disease in heart. The proposed system is evaluated and compared to standard classifiers based on feature fusion, feature selection, and weighting approaches using heart disease data. The suggested approach outperforms existing systems with accuracy of 98.5%. The study [17] suggested a unique two-tier structure in which the first tier differentiates between normal and tumour MRI, and the second tire localizes tumor regions. According to their experimental results, the proposed framework achieved 97% accuracy for classification tasks using GoogleNet and 83% accuracy for localization tasks using pretrained YOLO v3 models after fine-tuning. Dash et al. [18] have used a fast and a matched filter vessel extraction for measuring the performance. They did an extension of the matched filter methodology by integrating a filter having fast-guide along with a matched filter to enhance fundus images. Also, they proposed combining hysteresis thresholding, mean-C thresholding, and Otsu thresholding for extraction of vessel, and is evaluated on DRV and CDB data sets. The high-volume dataset's attribute set was reduced in this research by a new tri-stage feature selection method that involves choosing a subset, crucial features [19]. It used four filter methods (MI, CS, RFF, and XV) at Phase 1 along with three classification algorithms (KNN, SVM, and NB) in order to select every feature that is most accurate regardless of the filter method or classification algorithm used. Srinivasu et al. [20] performed automatic segmentation of CT scan images to detect anomalies in the human liver using a effective in computation AW-HARIS method. In contrast to supervisory methods that demand substantial computational resources for training, the proposed approach achieves superior issue detection accuracy without the need for training. The study [21] employed Long Short-Term Memory (LSTM) combined with a multimodal multitasking Deep Learning (DL) approach, utilizing data from 47 patients to anticipate Length of Stay (LOS) and readmission. Within this multimodal DL model, the patient's readmission status is precisely classified, yielding a mean square error of 0.025 and root mean square error of 0.077, with an impressive accuracy of 94.84%. Similarly, the study [22] introduced an ensemble learning framework that integrates heterogeneous base learners into a unified model via the stacking technique. Leveraging multimodal time-series data, a 4-class ensemble classifier is constructed to forecast the progression of Alzheimer's disease 2.5 years into the future.

However, all these techniques didn't deliver user a self-explanatory reason of the claim of forged images. These works showed different methods to detect image credibility however no work was shown to explain why the image is not credible. While some work of XAI in medical imaging analysis was done by models discussed on single image modality, which conforms with image explanation settings [23] numerate numerous requirements as proxies for actual results to direct the development and assessment of XAI algorithms, including correctness and robustness. Singh et al. [24] studied 13 XAI algorithms for classifying eye disorders were studied. Utilizing retinal scans, they asked 14 medical professionals to assess the heatmaps with relation to their clinical utility (plausibility). The study [25] assessed five gradient-based XAI algorithms for early cancer classification from endoscopic images. They evaluated the consistency between heatmaps and the actual annotations of localized lesions using computational measures (plausibility). The gradient approach performed better than the other four algorithms that best match the ground-truth annotations made by clinicians. Table 1 shows strengths and weakness of XAI models used in medical domain. As no work is done is field of facial credibility, we propose our XAI system which also cover all the limitations of methods used here.

Table 1. Previous work done in association with XAI AND CNN

Reference	XAI Model Used	Dataset Used	Strengths	Limitations
[23]	Lime	Freddie Mac dataset	A sole dataset is altered by adjusting feature values, and the subsequent influence on the output is observed	suffers from labels and data shifts, explanations dependent on the choice of hyperparameters
[24]	Shapley value sampling	3D MNIST and an MRI dataset	It calculates rewards for every feature, accounting for potential variations in individual contributions among players	It uses feature dependencies, leading to confusion when interpreting SHAP plots.
[25]	Lime and Shapley	Own dataset consisting of bleeding and non-bleeding images	These both determined feature contributing using one feature, which is a friendly way	They lead to self-discrepancies and many times different results

3. PROPOSED METHODOLOGY

From the extensive literature review it was very clear that the Explainable AI was not used in case of forged images specially in facial health section field in which authors have identified a potential gap and have proposed a methodology, which was never created in such fashion. Authors proposes, to create a dataset and then use ensembling of unique combination of CNN [26] and among which the best performed combinations results will be explained by explainable ai. To cover in brief methodology is divided into further sub sections as follows as shown in Figure 4.

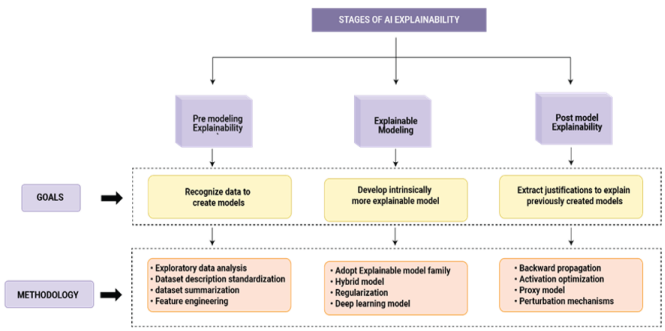

Figure 4. Proposed methodology architecture

3.1 Dataset preparation

In the dataset preparation process, a total of 600 images were curated, comprising 300 real and 300 fake images. As the authors conducted a survey, they found that 50-60% of health web blogs advertising products like “Get Glow in 7 days” or “Remove acne in 30 days” used before-after images to gain users' trust. These images, often used to showcase product results, attract consumers as the forgery is imperceptible to the naked eye. To address this problem, the authors scraped 300+ freely available images online portraying skincare issues such as acne, pimples, and dark spots. Using a face-editing application, the images were photoshopped to create fake versions where the forgery was not visible. This balanced dataset was then split into 70% for training, 15% for validation, and 15% for testing, ensuring effective training and robust model evaluation.

3.2 Image processing

Authors after creating the dataset, applied image processing techniques on it. Since every image was of different size it was necessary to make sure that image size was same so that amount of visual data and pixels are same and will not hamper the model. Image was resized into 224 × 224 using OpenCV and then the image was normalized so that the pixel intensity value is reduced such that the computation required gets reduced.

3.3 Deep learning and ensemble

Ganaie et al. [27] have decided to use deep learning model and ensemble learning technique in the proposed methodology. In order to employ deep learning model pretrained ImageNet TensorFlow model was taken. A specific list of CNN was chosen which amounted to total 6 CNN namely ResNet50V2, MobileNetV2 DenseNet121, VGG19, EfficientNetB0 and InceptionV3. ResNet50V2 and DenseNet121 would be more efficient, with the ability to learn a complex feature hierarchy that captures fine details in forged images. Then, MobileNetV2 and EfficientNetB0 provide computational efficiency, so the ensemble would be suitable for real applications. Finally, VGG19 and InceptionV3 ensure consistent performance while detecting multi-scale features for robust analysis of images. The varying lightweight and deep models together improve the ensemble's generalization capacity, thus making it a more accurate and efficient detector for subtle forgeries of facial images in diverse scenarios. These models were trained by dataset created by the authors and have been tested on unseen images. After training these models, their weights were saved and was used to employ stacked ensembling method by using three deep learning model CNN. ResNet50V2, MobileNetV2, and InceptionV3 models are trained separately, and then fed their prediction for meta-learner, learned to optimally combine the predictions from these base models. The stacking method enables the ensemble to leverage the strengths of each individual model, which ultimately yields better output than either simple or weighted averaging. This approach, in particular, is useful for retaining complementary information from different CNN architectures. Since there was no study which has applied explainable ai with combination of stacked ensemble technique and the fact that using stacked method that accepts sub-model outputs as input and attempts to learn how to best combine the input predictions to get a superior output prediction. Hence it was decided that among the trained and tested model of CNN has been used a good combination of CNN has to be employed to gain highest accuracy or ensemble score. In order to gain a good ensemble, score a combination of 3 CNN at a time was taken such that the resulting ensemble score is highest, the results associated with this is shown in Table 2.

Table 2. Ensemble score of unique combination of CNN

Experiment No.	CNN Combination	Ensemble Score
1	ResNet50V2 MobileNetV2 InceptionV3	79
2	MobileNetV2 InceptionV3 DenseNet121	78
3	ResNet50V2 InceptionV3 DenseNet121	77
4	ResNet50V2 DenseNet121 VGG19	76
5	MobileNetV2 DenseNet121 VGG19	75
6	InceptionV3 DenseNet121 VGG19	74
7	VGG19 MobileNetV2 InceptionV3	73
8	DenseNet121 EfficientNetB0 ResNet50V2	70
9	EfficientNetB0 ResNet50V2 MobileNetV2	70
11	EfficientNetB0 InceptionV3 DenseNet121	68
12	EfficientNetB0 MobileNetV2 InceptionV3	67
13	DenseNet121 VGG19 EfficientNetB0	66
14	VGG19 EfficientNetB0 ResNet50V2	66
15	ResNet50V2 VGG19 EfficientNetB0	66
16	MobileNetV2 VGG19 EfficientNetB0	65
17	InceptionV3 VGG19 EfficientNetB0	64

In order to come up with such a combination it was decided that two same CNN model will not be repeated in a combination of 3 Ensemble CNN model and every CNN used in a combination of 3 will be unique in nature. Using this logic, a stacked ensemble technique was employed where every CNN is stacked onto one another. Since the models are stacked onto another generalization ensemble happens which can utilize the collection of predictions as a context and conditionally decide how to weight the input predictions,

perhaps leading to higher performance.

The time complexity [28] of these ensemble learning methods is evaluated under both pruned and unpruned conditions. Assuming this model is employed to train the foundational algorithms (mbase), the time complexity for making predictions on new, unseen data (unknown instances) is represented as O (m x train), where train signifies the average time taken to train the model using one method, and test indicates the average time needed to test the model using one technique. The overall training time complexity for the ensemble learning model, excluding the pruning process, stands at O(m x test). The aggregate space required by an algorithm to function across diverse input sizes is termed its space complexity, essentially reflecting the amount of space needed for its execution. In this specific scenario, the space complexity amounts to n (n = number of photos).

The ensemble of CNN models in this study was constructed using a stacked ensemble approach, leveraging the strengths of multiple Convolutional Neural Networks (CNNs) to improve overall prediction accuracy. We employed six pre-trained CNN architectures—ResNet50V2, MobileNetV2, DenseNet121, VGG19, EfficientNetB0, and InceptionV3—each fine-tuned on our custom dataset of forged and real facial images. These architectures were chosen for their varying depth, parameterization, and suitability for image classification tasks.

The models were trained individually using the Adam optimizer with an initial learning rate of 0.001, categorical cross-entropy loss, and a batch size of 32. Each model was trained for 50 epochs, with early stopping criteria based on validation loss to prevent overfitting.

After individual training, the ensemble was constructed by stacking the top three performing CNNs based on validation accuracy. The stacked ensemble takes the predictions of the individual models as input and combines them using a meta-learner, which learns how to best combine these predictions to produce a superior output. The final layer of the stacked ensemble was a fully connected layer with a softmax activation function to output the classification probabilities. To ensure diversity in the ensemble, the same CNN model was not repeated across multiple combinations, and each ensemble contained unique CNN architectures.

Hyperparameters for each model were fine-tuned using grid search, varying the learning rates (0.0001, 0.001, 0.01), batch sizes (16, 32, 64), and dropout rates (0.2, 0.5). The best ensemble configuration was selected based on its ability to generalize on unseen data, and its performance was evaluated using metrics such as accuracy, precision, recall, and F1-score. This stacked ensemble approach ensured that the strengths of individual CNN models were captured, leading to a more robust detection system for identifying forged facial images.

3.4 Explainable AI

An understanding of why a model made a particular choice is critical to the process of detecting Image credibility. A fact-checker is provided with the information deemed most relevant to the conclusion that the content was fraudulent by the disclosure of the reasons it was deemed fraudulent. Here, Explainable AI provides users with the reasoning behind the model's decisions, which helps build trust in the model. The three stages of Explainable AI (XAI) from the study [29] are illustrated in Figure 5. Using these stages we implemented 3 methods GradCam, GradCam++ and ScoreCam on different CNN models and generated heat maps to visualize which areas are forged or photoshopped contributing to models decision as fake.

These 3 methods of XAI have been explained accordingly, which is also illustrated in Figure 6, where the image, when fed into an XAI system, goes through various convolutional layers where then the final layer is back propagated till the last convolutional layer for each of the XAI systems i.e. GradCam, GradCam++, and ScoreCam which then gives the final image representing areas of forgery.

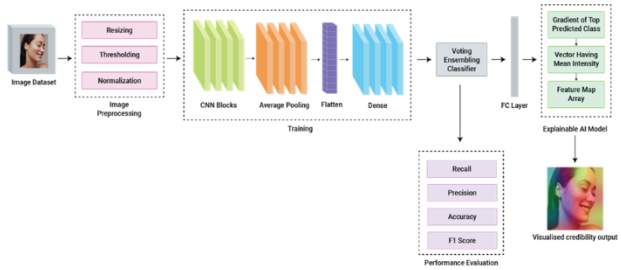

Figure 5. Working of explainable AI

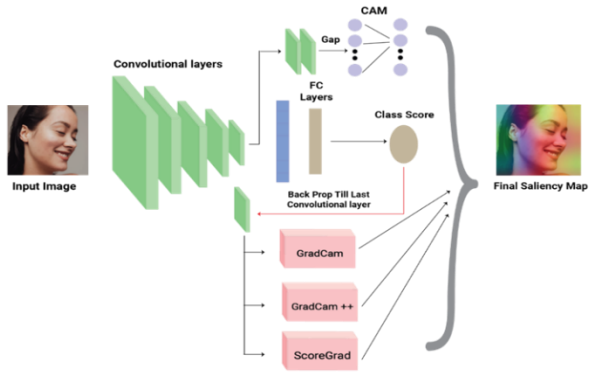

Figure 6. Explainable AI architecture

GradCam: Grad-CAM [30] employs gradient information from CNN's final convolutional layer to assign priority values to each neuron for a specific option. Obtain the class-discriminative localization map for any class C of width u and height v.

$$L^{C}_{Grad-CAM} \in R^{u\,X\,v} \tag{1}$$

The authors of this study [30] first calculate the gradient of the score for class c, yc of a convolutional layer's feature Map Activation Ak (before the softmax).

$$\frac{\partial y^{c}}{\partial A^{k}} \tag{2}$$

The neuron significance weights (indexed by i and j, respectively) are calculated using the global average of the gradients across the width and height dimensions.

$$a^{c}_{k} = \frac{1}{z} \sum_{i} \sum_{j} \frac{\partial y^{c}}{\partial A^{k}_{ij}} \tag{3}$$

The real amount of computation to successive weight matrix products and gradient to activation functions matrix products up until the final convolution layer to which the gradients are transmitted while backpropagating gradients to activations and computing ack. This weight ack thus captures the 'importance'

of feature map k for a target class c by partially linearizing the deep network downstream from A.

Authors perform a weighted combination of forwarding activation maps and follow it by a ReLu to obtain,

$$L^{C}_{Grad-CAM} = ReLu\left(\sum_k a^c_k A^k\right) \quad (4)$$

Because they are only concerned in parts that have a bigger impact on the image's emotional content, such as pixels whose intensity should be increased, the authors use a ReLu to boost yc. Brighter pixels are more likely to be visible in an image with fewer affecting or contributing factors. The final results using GradCam are comparable to one done in a recent study [30]. As one might expect, localization of maps without this ReLu occasionally shows darker regions than emotion recognition features.

GradCam++: Grad-CAM cannot correctly locate the objects when there are several instances of a particular class in a single image. This is a significant issue since there are frequently multiple instances of an object in the real world. The localization may not always be needed for the complete item but rather for portions of it, which is another effect of using an average of unweighted using partial derivatives. Grad-notion CAMs of deepening a CNN's transparency may be compromised. Since each pixel in a CNN's feature map directly affects the outcome, the authors of this study [31] employed the Gradcam plus model, which was developed in this manner. The authors, in particular, have transform Eq. 1 by explicitly coding the composition of the weights w_kc.

$$w^c_k = \sum_i \sum_j a^{kc}_{ij} . relu\left(\frac{\partial Y^c}{\partial A^k_{ij}}\right) \quad (5)$$

where, relu is the activation function for the Rectified Linear Unit. The idea is that w^c_k symbolizes the importance of certain maps of activation A^k. Positive gradients are critical in constructing saliency maps for a given convolutional layer in previous attempts at pixel-space visualization, such as Deconvolution and Guided Backpropagation. An activation map with a positive gradient at (i, j).

A^k signifies that increasing the pixel intensity (i,j) has a positive effect on the class score Y^c. Thus, a linear combination of the positive partial derivatives with respect to each pixel in an activation map is obtained, and the relevance of that map for class c is revealed. Because of this structure, the weights w^c_k are a weighted average of the gradients rather than a global average.

ScoreCam: Variations of GradCAM, like GradCAM++, exhibit distinctions primarily in the combinations of gradients employed to characterize a_k^c. Their objective is to make models without global pooling layers more universally applicable. The overarching contribution, functioning as a bridge between perturbation-based and CAM-based techniques for the relevant input features, encodes the activation map values instead of the local sensitivity measurement, also known as gradient information. This representation in Score-CAM [32] offers a more intuitive interpretation of the weight of activation maps. Unlike earlier approaches, the authors incorporate the increased confidence value within the gradient information that contributes to the final convolutional layer, emphasizing the importance of each activation map's relevance.

$$L^{C}_{Score-CAM} = ReLu\left(\sum_k a^c_k A^k_l\right) \quad (6)$$

4. RESULTS AND DISCUSSION

4.1 Deep learning

As discussed in the proposed methodology section C, multiple unique combinations of CNN were taken in a group of 3 using a specified logic by not repeating CNN in a particular combination and not repeating the same CNN in a combination. This enabled the creation of a set of huge results, among which the best was taken and studied further in detail. This result was mapped in Table 2, which shows how Experiment No. 1, which had a combination of ResNet50V2, MobileNetV2, and InceptionV3, performed the highest. In contrast, Experiment No. 17 performed lowest using InceptionV3, VGG19, and EfficientNetB0.

Among the best-performed Experiment No. 1 in Table 2 was taken and then was further studied in detail as to how this group of CNN performed individually by recording metrics such as Precision, Recall, F1Score and Accuracy using Keras AI library mapped in Table 3.

Table 3 shows how the individual CNN has performed on validated data, in which three pre-trained ImageNet CNN were used, namely ResNet50 V2, MobileNet V2, and Inception V3. The dataset was split into 80:20 rations in two sections: Testing and Validation. In such dataset distribution, the model used was trained, and then later stacked ensembling was done on it, which can be seen in Table 4, where many CNN indicates the CNN stacked onto one another and their associative results. These stacked CNNs are the same CNNs that, as a group, have achieved the highest ensemble score in Table 2.

Table 3. Model wise individual metrics

Model used	Accuracy %	Precision	Recall	F1 Score
ResNet50 V2	0.80	0.81	0.81	0.81
MobileNet V2	0.76	0.78	0.77	0.76
Inception V3	0.75	0.77	0.76	0.75

Table 4. Stacked ensembling accuracy

Number of CNN	Stacked Ensemble Accuracy %
1	0.81
2	0.78
3	0.79

From Table 4, we can see how stacked ensembling accuracy has performed; from this table, it is observed how an increasing the number of CNNs used for ensembling affects the accuracy. By using three CNN stacked ensembles, the accuracy recorded was 79%, which is greater than the average accuracy of all three CNN. The reason why ensembling accuracy is more than the average accuracy is that the predictive results accuracy increases as the stacking of one CNN onto another broadens the scope of the learned parameters. At the same time, training since every CNN has its own unique architectural designs, which lead to varied performance metrics. However, no matter how CNN architecture differs, authors have used a generalized approach to use multiple CNNs rather than making custom changes on every CNN. This was because to reach the perfect metrics are obtained using AI libraries offered by Keras. Using these combinations of specific CNN, which can give maximum accuracy, several combinations of different numbers of CNN models were tried, among which models used in Table 2 Experiment No. 1 have recorded the highest accuracy. To give

fair parametric learning across all models, the same epoch and batch were used, which can be observed in Table 5

Table 5 shows that the epochs used for all three CNN are 20, and the batch size used is 32. The batch size 32 was selected because, in terms of computation power used fits perfectly to the resources available at the time of study by the authors also another reason why batch size apart from 32 was not taken is that if we took less batch size, it would mean each step in the gradient might lead to less accuracy as only small portion of the dataset which might be lead to a local minimum rather than an overall lowest minima in the gradient graph. A similar vice versa reason can be applied that with the greater batch size, the computation power would also increase and may also lead to skipping of overall lowest minima in terms of gradient descent graph while training. Since the 32 batch size is neither small nor too big, it makes a good parametric value, leading to high hopes of getting good accuracy. Now, although the parameters used among all three CNNs were the same, the size of the computation power would also increase. It may lead to skipping of the overall lowest minima in terms of gradient descent graph while training. Since the 32 batch size is neither small nor too big, it makes a good parametric value, leading to high hopes of getting good accuracy. Now, although the parameters used among all three CNN were the same, however, the CNN used are still different and have their own individual plus points as compared to other CNN because of their architectural designs. Although the architecture for every CNN may differ, a generalized approach was taken, which can be observed in Table 6.

Table 5. Models used parameters

CNN Used	Epoch Used in Dataset	Batch Size	Pretrained ImageNet Weights Used	Additional Weights Used
ResNet50 V2	20	32	✓	×
MobileNet V2	20	32	✓	×
Inception V3	20	32	✓	×

Table 6. CNN model layer distribution

Layers	Detail Layers	Filters	Units	Kernel Size	Stride	Activation
1	Input Data	-	-	-	-	-
2	Pretrained ImageNet CNN	-	-	-	-	-
3	Dense_1	-	128	-	-	ReLu
4	Dense_2	-	128	-	-	ReLu
5	Dense_3	-	2	-	-	Sigmoid

4.2 Explainable AI

In this part, we run experiments to see how effective the proposed explanation approach is. Explainable AI is employed on this individual stacked CNN used for ensemble learning to determine how they perceive and label the credibility of an image. To achieve this, the authors qualitatively evaluate three explainable models—Grad-CAM, Grad-CAM++, and Score-CAM—by presenting heatmaps on three different CNN models and assessing their results. Figure 7 illustrates the increasing order of feature-contributing colors, where red indicates the least contribution and yellow indicates the highest contribution. Figure 8 displays the saliency map for sample unseen images. The heatmap images use colors to indicate the extent of contribution from each area, helping to identify which areas are most photoshopped and which are least or not altered, thereby determining the reasons for the image's lack of credibility.

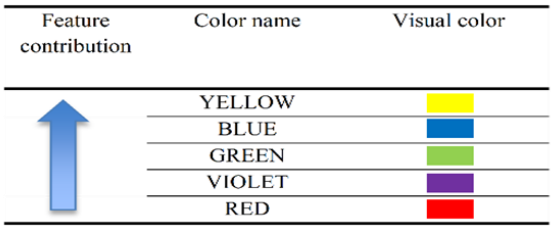

Figure 7. Explanation AI color range

In Figure 8, colors in the heatmap are more visible, implying the highest accuracy in the scorecard following GradCam ++ and GradCam. In this, the yellow colors are highlighted on the facial area, where the dark spots or pimples are removed using photoshopping or editing, followed by red, which shows that little or no editing has been done on that area. Figure 7 and Figure 8 together show which area is most edited, contributing to the forging of that image. In this way, users get a clear explanation of why the model predicts the image as fake / non-credible.

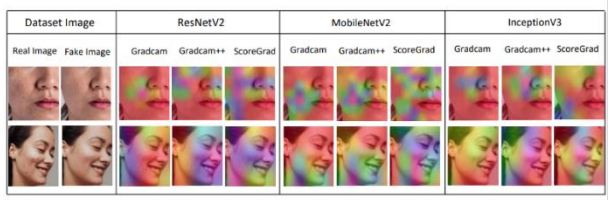

Figure 8. Heatmap visualizations by XAI techniques

4.3 Evaluation of explainable AI models

We evaluate the faithfulness of the explanations generated by all the explainable Ai models namely Grad-CAM, Gradcam ++ and ScoreGrad we studied the performance with two different metrics: (1) Average drop %; (2) % increase in confidence which is described below.

1) Average Drop The regions that are most important for making decisions should be highlighted on an explanation map for a class. As opposed to when the entire image is provided as input, it is anticipated that deleting portions of an image will decrease the model's confidence in its judgement. Using this, we investigate the effectiveness of the explanation maps produced by Score Grad, Grad-CAM++, and Grad-

CAM. This suggests that the visual explanation of GradCAM++ includes more of what is relevant (be it the object or the context) for a correct decision. Average Drop is expressed in equation 7 where Y^c_i is the predicated score for class c on image i and O^{C}_i is the predicated score for class c with the explanation map region as input. The average of this value throughout the entire dataset is calculated for each image.

$$\sum_{i=1}^{N} = \frac{max(0, Y_i^c - O_i^c)}{Y_i^c} \text{x } 100 \quad (7)$$

2) Increase % in Confidence: Furthermore, to the prior metric, it is expected that there will be cases were using only the explanation map region as input (rather than the complete image) increases the prediction's level of confidence (particularly when the context is distracting). This metric counts the number of times the model's confidence increased when only the explanation map regions were provided as input. Formally, the higher % Confidence metric is defined as in equation 8 where 1_Y is an indicator function that returns 1 when the argument is true. All other notations are as defined for the previous metric.

$$\sum_{i=1}^{N} = \left(\frac{1_{Y_i^c < O_i^c}}{N}\right) 100 \quad (8)$$

According to Table 7, Score-CAM beats alternative perturbation- and CAM-based approaches on a wide scale, achieving an average reduction and increase of 32.9% and 31.4%, respectively. A strong performance shows that Score-CAM, rather than only discovering what humans think is essential, can successfully identify the target object's most easily identifiable location. Compared to earlier methods, task results show that Score-CAM could more faithfully disclose the original CNN model's decision-making process.

Table 7. Explainable AI metrics

Method	GradCam	GradCam++	ScoreCam
Average Drop %	57.9	46.3	32.9
Average Increase %	22.5	19.2	31.4

5. CHALLENGES

This method of visualizing image credibility predictions of health blogs in a user-friendly way delivers essential information. According to an extensive literature analysis, most effort in picture credibility is done solely to identify these faked images. Even though the initial phase of identifying ways to help consumers understand the explanation of image credibility was consuming time, the findings were given as significant. The key applicability of the work is the visual description of the blog's image believability. This paper offers software-based reasons for the predictions made. According to the authors, very little equivalent work is done in healthcare. Authors encountered the following limits and obstacles:

1. A lack of studies on explainable AI algorithms in picture credibility hampered the project's initial research effort.

2. Due to a shortage of publicly available datasets, this project's dataset must be produced from scratch utilizing face editing tools.

3. The strategies used in this experiment were exclusively applied to facial photos from health blogs that were hosted online.

6. FUTURE SCOPE

Although this study is devised to address the challenge of image forgery detection in health-related advertisements for facial care products, there are several limitations prevailing. The method heavily relies on the availability of static image datasets, thereby eliminating the chance of its applicability toward more complex forgeries such as video-based manipulations. Further extension of this method toward video forgery detection has matured into an extremely relevant online marketing concept. Combining other XAI methods, like SHAP or LIME, can help generate even deeper understanding in the decisions of the model. A more challenging and diverse dataset should be used to benchmark the framework, along with extending the framework for other related medical applications, such as hair treatment or weight reduction. Reinforcement learning and one-shot learning will enhance the adaptability of the model toward changing data. In addition, federated learning integration will ensure privacy on sensitive user data, and the system will be more secure and scalable for real-world applications.

7. CONCLUSION

From the extensive literature review, we came to know the need for image credibility visualization and potential gaps, which were covered by our proposed methodology. Also, the available datasets were not useful as they were fake to the naked eye. Hence a new dataset was made where user with the naked cannot recognize whether the image is fake or not. Authors use an ensemble of state-of-the-art CNN models and have achieved an accuracy of 79% by training our model from the dataset created from these trained model 3, models of explainable ai were successfully implemented on them, which determines the sub-sections of the images, which are forged by color visualization offered by explainable ai techniques such as GradCam, GradCam++ and SoftGrad. This explainable approach, along with ensembling, will make the online consumer audience more alert from such products, which are marketed on forged images, enabling better accountability from the facial health/beauty product claiming companies. This approach will help the users to identify whether the claim made by the companies using before-after images in their advertisement to influence people to use their product is actual or not; hence, they will be able to know why the claim is not true. Also, with this real advertisements will gain trust of people will be able to reach more people without their marketing. This approach can also be used in the future, not only in facial but also in the full body, which will indicate more awareness of the treatments provided on the internet, like reducing weight in 10 days showing before and after photos, which can indicate areas of photoshopping. Hence, with this approach, a user will be able to decide whether the treatment or medicine is mentioning the claims true or not.

FUNDING

This work was supported by the Research Support Fund (RSF) of Symbiosis International (Deemed University), Pune, India.

DATA AVAILABILITY STATEMENT

The data used in this study is available at the following link: https://github.com/manishr0404/Credibility-Research-Paper-/blob/main/README.md. The dataset is created by scrapping images from online freely available resources and websites.

REFERENCES

[1] Ali, S.S., Ganapathi, I.I., Vu, N.S., Ali, S.D., Saxena, N., Werghi, N. (2022). Image forgery detection using deep learning by recompressing images. Electronics, 11(3): 403. https://doi.org/10.3390/electronics11030403

[2] Peterson, G. (2005). Forensic analysis of digital image tampering. In Advances in Digital Forensics: IFIP International Conference on Digital Forensics, National Center for Forensic Science, Orlando, Florida, pp. 259-270. https://doi.org/10.1007/0-387-31163-7_21

[3] Ambiata. (2021). Explainable AI (XAI) and interpretable machine learning (IML) models. https://www.ambiata.com/blog/2021-04-12-xai-part-1/.

[4] Kadam, K., Ahirrao, S., Kotecha, K. (2021). AHP validated literature review of forgery type dependent passive image forgery detection with explainable AI. International Journal of Electrical and Computer Engineering, 11(5): 4489-4501. https://doi.org/10.11591/ijece.v11i5.pp4489-4501

[5] Boididou, C., Papadopoulos, S., Zampoglou, M., Apostolidis, L., Papadopoulou, O., Kompatsiaris, Y. (2018). Detection and visualization of misleading content on Twitter. International Journal of Multimedia Information Retrieval, 7(1): 71-86. https://doi.org/10.1007/s13735-017-0143-x

[6] Wagle, V., Kaur, K., Kamat, P., Patil, S., Kotecha, K. (2021). Explainable AI for multimodal credibility analysis: Case study of online beauty health (mis)-information. IEEE Access, 9: 127985-128022. https://doi.org/10.1109/ACCESS.2021.3111527

[7] Jin, Z., Cao, J., Luo, J., Zhang, Y. (2016). Image credibility analysis with effective domain transferred deep networks. arXiv preprint arXiv:1611.05328. http://arxiv.org/abs/1611.05328

[8] Mayer, O., Stamm, M.C. (2018). Accurate and efficient image forgery detection using lateral chromatic aberration. IEEE Transactions on Information Forensics and Security, 13(7): 1762-1777. https://doi.org/10.1109/TIFS.2018.2799421

[9] Dua, S., Singh, J., Parthasarathy, H. (2020). Image forgery detection based on statistical features of block DCT coefficients. Procedia Computer Science, 171: 369-378. https://doi.org/10.1016/j.procs.2020.04.038

[10] Meena, K.B., Tyagi, V. (2019). Image forgery detection: Survey and future directions. Data, Engineering and Applications, 2: 163-194. https://doi.org/10.1007/978-981-13-6351-1_14

[11] Shyry, S.P., Meka, S., Moganti, M. (2019). Digital image forgery detection. International Journal of Recent Technology and Engineering, 8(2S3): 658-661. https://doi.org/10.35940/ijrte.B1121.0782S319

[12] Wu, Y., Abd-Almageed, W., Natarajan, P. (2018). Busternet: Detecting copy-move image forgery with source/target localization. In Proceedings of the European Conference on Computer Vision (ECCV), 168-184.

[13] Wu, Y., Abd-Almageed, W., Natarajan, P. (2018). Image copy-move forgery detection via an end-to-end deep neural network. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 1907-1915. https://doi.org/10.1109/WACV.2018.00211

[14] Wu, Y., AbdAlmageed, W., Natarajan, P. (2019). ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9535-9544. https://doi.org/10.1109/CVPR.2019.00977

[15] Ali, F., El-Sappagh, S., Islam, S.R., Ali, A., Attique, M., Imran, M., Kwak, K.S. (2021). An intelligent healthcare monitoring framework using wearable sensors and social networking data. Future Generation Computer Systems, 114: 23-43. https://doi.org/10.1016/j.future.2020.07.047

[16] Ali, F., El-Sappagh, S., Islam, S.R., Kwak, D., Ali, A., Imran, M., Kwak, K.S. (2020). A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Information Fusion, 63: 208-222. https://doi.org/10.1016/j.inffus.2020.06.008

[17] Ali, F., Khan, S., Abbas, A.W., Shah, B., Hussain, T., Song, D., EI-Sappagh, S., Singh, J. (2022). A two-tier framework based on GoogLeNet and YOLOv3 models for tumor detection in MRI. Computers, Materials and Continua, 72: 73. https://doi.org/10.32604/cmc.2022.024103

[18] Dash, S., Verma, S., Kavita, Bevinakoppa, S., Wozniak, M., Shafi, J., Ijaz, M.F. (2022). Guidance image-based enhanced matched filter with modified thresholding for blood vessel extraction. Symmetry, 14(2): 194. https://doi.org/10.3390/sym14020194

[19] Mandal, M., Singh, P.K., Ijaz, M.F., Shafi, J., Sarkar, R. (2021). A tri-stage wrapper-filter feature selection framework for disease classification. Sensors, 21(16): 5571. https://doi.org/10.3390/s21165571

[20] Srinivasu, P.N., Ahmed, S., Alhumam, A., Kumar, A.B., Ijaz, M.F. (2021). An AW-HARIS based automated segmentation of human liver using CT images. Computers, Materials and Continua, 69(3): 3303-3319. https://doi.org/10.32604/cmc.2021.018472

[21] Ali, S., El-Sappagh, S., Ali, F., Imran, M., Abuhmed, T. (2022). Multitask deep learning for cost-effective prediction of patient's length of stay and readmission state using multimodal physical activity sensory data. IEEE Journal of Biomedical and Health Informatics, 26(12): 5793-5804. https://doi.org/10.1109/JBHI.2022.3202178

[22] El-Sappagh, S., Ali, F., Abuhmed, T., Singh, J., Alonso, J.M. (2022). Automatic detection of Alzheimer's disease progression: An efficient information fusion approach with heterogeneous ensemble classifiers. Neurocomputing, 512: 203-224. https://doi.org/10.1016/j.neucom.2022.09.009

[23] Dutta, P., Muppalaneni, N.B., Patgiri, R. (2022). A

survey on explainability in artificial intelligence. In Handbook of Research on Advances in Data Analytics and Complex Communication Networks, IGI global, pp. 55-75. https://doi.org/10.4018/978-1-7998-7685-4.ch004

[24] Singh, S., Sengupta, S., Lakshminarayanan, V. (2020). Explainable deep learning models in medical image analysis. Journal of Imaging, 6(6): 52. https://doi.org/10.3390/jimaging6060052

[25] Knapič, S., Malhi, A., Saluja, R., Främling, K. (2021). Explainable artificial intelligence for human decision support system in the medical domain. Machine Learning and Knowledge Extraction, 3(3): 740-770. https://doi.org/10.3390/make3030037

[26] de Souza Jr, L.A., Mendel, R., Strasser, S., et al. (2021). Convolutional neural networks for the evaluation of cancer in Barrett's esophagus: Explainable AI to lighten up the black-box. Computers in Biology and Medicine, 135: 104578. https://doi.org/10.1016/j.compbiomed.2021.104578

[27] Ganaie, M.A., Hu, M., Malik, A.K., Tanveer, M., Suganthan, P.N. (2022). Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence, 115: 105151. https://doi.org/10.1016/j.engappai.2022.105151

[28] Afzal, S. (2022). Research on computational complexity of machine learning. https://www.researchgate.net/publication/359441560_Research_on_Computational_Complexity_of_Machine_Learning.

[29] Islam, M. R., Ahmed, M. U., Barua, S., Begum, S. (2022). A systematic review of explainable artificial intelligence in terms of different application domains and tasks. Applied Sciences, 12(3): 1353. https://doi.org/10.3390/app12031353

[30] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision, 128(2): 336-359. https://doi.org/10.1007/S11263-019-01228-7

[31] Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N. (2017). Grad-CAM++: Improved visual explanations for deep convolutional networks. In Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, pp. 839-847. https://doi.org/10.1109/WACV.2018.00097

[32] Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X. (2020). Score-CAM: Score-weighted visual explanations for convolutional neural networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 24-25. https://doi.org/10.48550/arxiv.1910.01279