AHP validated literature review of forgery type dependent passive image forgery detection with explainable AI

Kalyani Kadam[1], Swati Ahirrao[2], Ketan Kotecha[3]
[1,2]Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India
[3]Head SCAAI, Symbiosis International (Deemed University), India

Article Info

Article history:

Received Aug 26, 2020
Revised Dec 28, 2020
Accepted Jan 19, 2021

Keywords:

Deep learning
Explainable AI
Image forgery detection
Image splicing

ABSTRACT

Nowadays, a lot of significance is given to what we read today: newspapers, magazines, news channels, and internet media, such as leading social networking sites like Facebook, Instagram, and Twitter. These are the primary wellsprings of phony news and are frequently utilized in malignant manners, for example, for horde incitement. In the recent decade, a tremendous increase in image information generation is happening due to the massive use of social networking services. Various image editing software like Skylum Luminar, Corel PaintShop Pro, Adobe Photoshop, and many others are used to create, modify the images and videos, are significant concerns. A lot of earlier work of forgery detection was focused on traditional methods to solve the forgery detection. Recently, Deep learning algorithms have accomplished high-performance accuracies in the image processing domain, such as image classification and face recognition. Experts have applied deep learning techniques to detect a forgery in the image too. However, there is a real need to explain why the image is categorized under forged to understand the algorithm's validity; this explanation helps in mission-critical applications like forensic. Explainable AI (XAI) algorithms have been used to interpret a black box's decision in various cases. This paper contributes a survey on image forgery detection with deep learning approaches. It also focuses on the survey of explainable AI for images.

Corresponding Author:

Kalyani Dhananjay Kadam
Symbiosis Institute of Technology
Symbiosis International (Deemed University), Pune, India
Email: kalyanik@sitpune.edu.in

1. INTRODUCTION

Images are used in almost every field, such as medical systems, glamor, courts, military, and industries, and social networking platforms such as Instagram, Facebook, and so on over the internet [1]. Regardless of whether it is the space shuttle exploding during launch, a man strolling on the moon, or officers raising a banner on Iwo Jima during World War II, such ground-breaking images impact the society [1]. The advancement in digital imaging software and photo-realistic graphics allows people to make images more realistic or spread alternative meanings. Images can be fused, graphically improved, and created by computers, then detecting these controlled images can be troublesome. The realness of digital images becomes an important study area for research and development. It finds it difficult for forensic experts to identify genuine and forged images. Identifying such manipulated parts performed by the faker's manipulation operation is a significant work. Detecting the forgery in digital images is one of the challenges in the digital era. Nowadays, deep learning is gaining more attention due to its significant results. Deep

Learning algorithms have achieved high accuracy, but offering drawbacks, as the important features cannot be interpreted with the numbers which are given by deep learning model. The semantics is also not added. These networks are giving a higher performance without the understanding of its inside working. Black box issues are nothing but investigating the inner working of a deep learning model and interpreting why it delivered a given yield. These problems become more critical when these networks are used in real-life applications such as medical diagnosis or forensic. Explainable AI explains the decisions of a network. In computer vision, visualization networks provide fascinating ways to visualize the image's essential features, interpreting the image. In this case, the input data is the image, and image interpretation is performed using heat maps. Heat maps give the most significant input data areas for the interpretation. It allows the user to understand which image pixels are associated with the expected class, and it also checks whether the deep learning model focuses on the image's reasonable area. Figure 1 shows the mind map of the article. The article is mainly divided into digital image forgery detection, keyword analysis using various tools, deep learning for handling forgery type-dependent forgeries, AHP model for image forgery detection, and explainable AI.

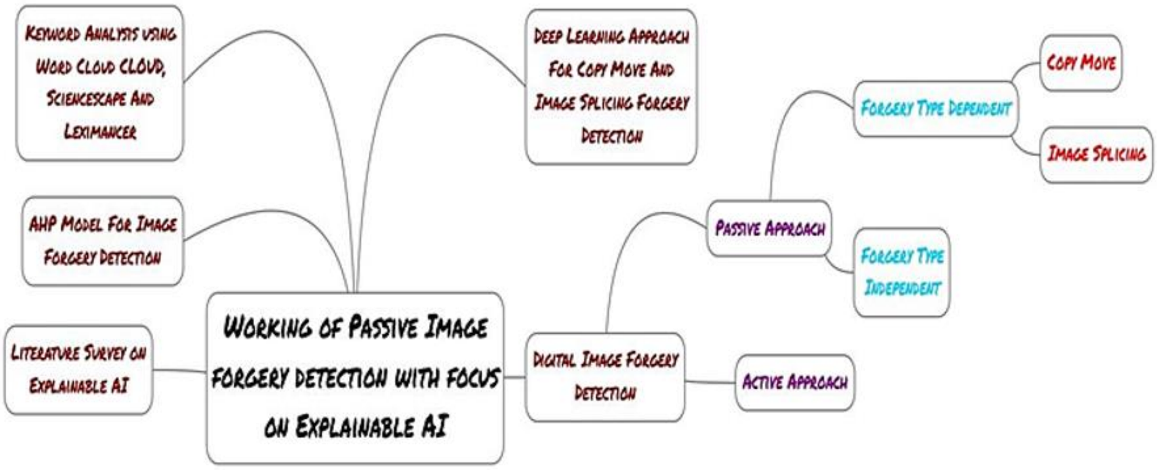

Keyword Analysis using Word Cloud CLOUD, Sciencescape And Leximancer
Deep Learning Approach For Copy Move And Image Splicing Forgery Detection
Copy Move
Forgery Type Dependent
Image Splicing
AHP Model For Image Forgery Detection
Passive Approach
Forgery Type Independent
Working of Passive Image forgery detection with focus on Explainable AI
Literature Survey on Explainable AI
Digital Image Forgery Detection
Active Approach

Figure 1. Workflow of passive image forgery detection with a focus on explainable AI

2. DIGITAL IMAGE FORGERY DETECTION

Image forgery detection is classified into two approaches [1]: passive or blind and active approach. Various types of image forgeries are shown in Figure 2.

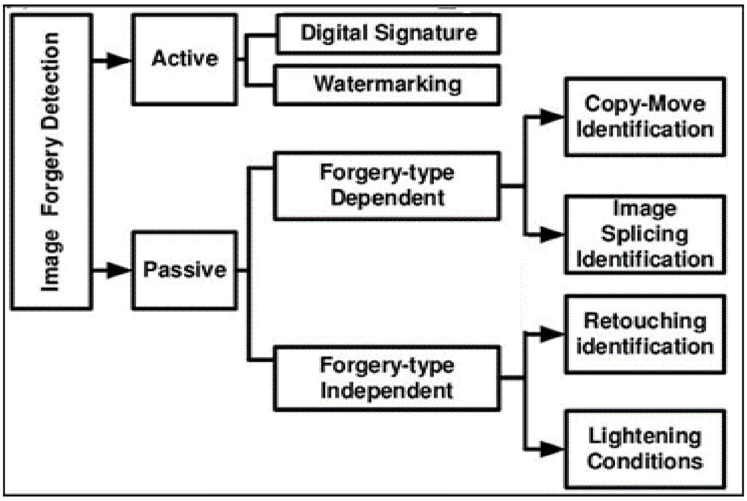


Figure 2. Digital image forgery detection

2.1. Active approach

In this approach, a computerized signature or watermark is injected in an image, and that is inserted either by an individual or by the acquisition device. Digital watermarking inserts particular information known as a message digest inside the image while capturing it. This information is then taken out from the image at later phases to get its authenticity. This extricated information is checked to see if it varies or not; if it changes, it shows that the image was altered after the image capturing procedure. It is a two-phase process, in the first phase, the message-digest is embedded in the image. In the second phase, i.e., after reaching the image to its s destination, the message digest is retrieved and compared with the obtained watermark [1]. Digital watermarking is an effective method to secure the image trustworthiness; however, different difficulties make its utilization unrealistic. Few cameras and gadgets have the property to insert watermark in the image at the time of creation. Some types of equipment, e.g., Canon EOS-1D, Nikon D2Xs, have embedding features but are overpriced. The disadvantage is that this method cannot differentiate legitimate and invalid operations in the image. Legitimate operations are performed for improving image quality, such as sharpening, contrast improvement, and so on. In the case of watermarking special software or hardware is needed to insert message digest in an image. Image forgery is identified with statistical information. This is done by dividing the image into similar region characteristics; after this, statistical measures such as mean, mode, median, and range of pixel values are inserted in the image with an encryption method [1]. Then this information is verified to check its authenticity.

2.2. Passive approach

Passive methodologies [1] uses statistical information of the image to detect forgeries in an image. These methods work without earlier information about the image, for example, digital watermarks or signatures. This method uses the existing image, and the image's manipulation operation changes its properties. It will make the image inconsistent, and thus the statistical image information is used for forgery identification. It is further classified into forgery type independent and dependent.

2.2.1. Forgery type independent

These techniques are used to identify the forgeries based on artifact clues left during retouching and lighting conditions [1]. This type of forgery uses statistical data of the image, and such invisible information is misused in different image handling activities, e.g., jpeg confining, image filtering, contrast improvements, and resampling. The resampling process changes the image by performing upsampling and downsampling operations and tools handling such forgeries works on either pixel or frequency domain. The median filtering tool is mostly used for noise elimination and image improvement. These methods give excellent performance, while the images are uncompressed and large. Commonly, the images are compressed with the JPEG compression strategy. The main goal of JPEG-based forgery identification techniques is to find out the locations of images with different JPEG compression locations.

– Forgery type dependent

This type of forgery focuses only on specific kinds of forgeries such as copy move and image splicing; these are gaining more attention as they are commonly performed manipulations on the image.

a. Copy move

This type of forgery copies part of the image and pasted into the same image at another region, which entirely changes its meaning. This is usually done to make an object "disappear" from the image by hiding it with a segment copied from another part of the image. This kind of forgery is very easy to perform, but it is very hard to detect as the copied section comes from the same image. Its different properties, such as shade palette, noise component, and other characteristics, are suitable to the rest of the image, which will not be recognized via the same methods that find inconsistencies in statistical measures in another portion of the image.

Figures 3 and 4 show that numerous papers are published in this exploration area from various countries. Data analysis is performed by using queries such as 'copy move forgery' and 'copy move forgery detection using deep learning' on IEEE (ieeexplore.org), Scopus (scopus.com), WOS (web of science), and ACM databases. The duration from 2001 to 2020 is considered for getting the result. Three different steps are performed for copy move forgery detection (CMFD): separating the image into blocks (overlying/non-overlying), executing the image property extraction technique on each block, comparison of features. In CMFD methods [1], the principal objective is to compare the portions inside an image and recognize the replicated and the original region.

b. Image splicing

Image compositing or splicing joins more than two images to create a forged image. It can be done by various software tools such as Adobe Photoshop, and Coral PaintShop. The copy move detection methods cannot be used in splicing as it works on multiple images; the manipulated part of the image has distinct properties compared with the rest of the image. This method uses numerous properties such as DCT and

DWT coefficients, Bi-coherence properties, camera response operation, and invariant image moments, The splicing methods consider the gadgets inborn properties instead of capturing the gadget's actual attributes. Figures 5 and 6 show that numerous papers are published in this exploration area from various countries. Data analysis is performed by using queries such as 'image splicing' and 'image splicing detection using deep learning' on IEEE (ieeexplore.org), Scopus(scopus.com), WOS(Web of Science), and ACM. The duration from 2001 to 2020 is considered for getting the result.

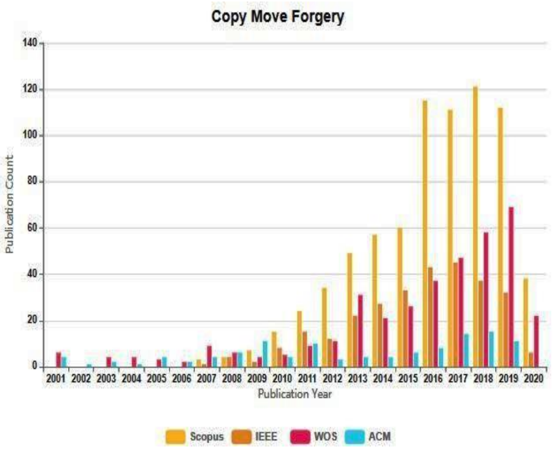


Figure 3. Last 19 years of publication data for 'copy move forgery' in scopus, IEEE, WOS (web of science), and ACM

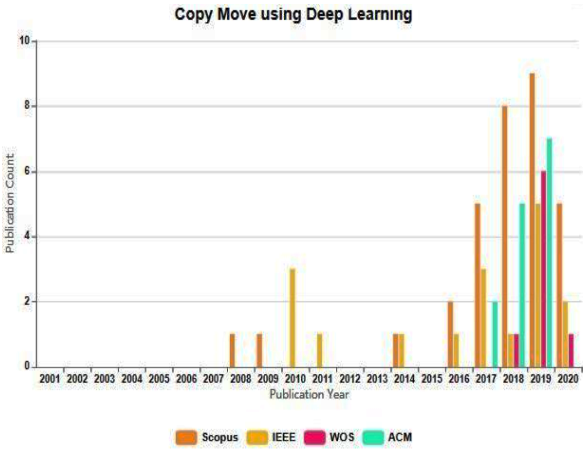


Figure 4. Past 19 years publication count for 'copy move forgery detection using deep learning' in Scopus, IEEE, WOS (web of science), and ACM

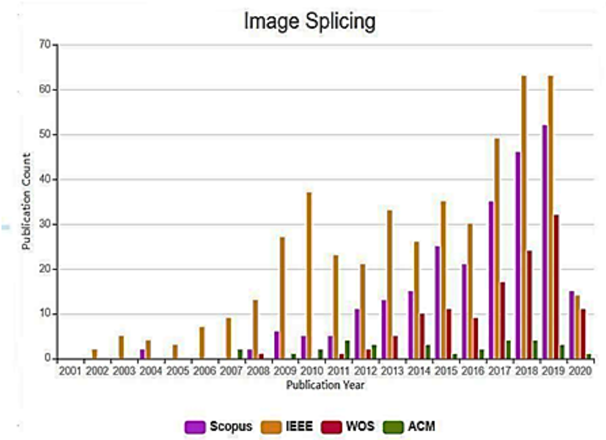


Figure 5. Past 19 years of the publication information for 'image splicing' in Scopus, IEEE, WOS (web of Science), and ACM

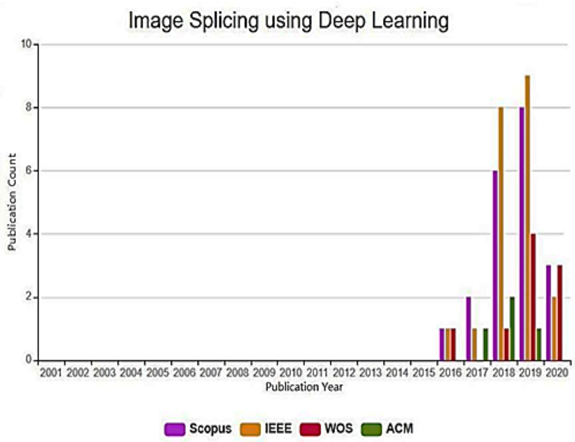


Figure 6. Last 19 years of publication information on 'image splicing detection using deep learning' in Scopus, IEEE, WOS (web of science), and ACM

3. KEYWORD ANALYSIS USING WORD CLOUD, SCIENSCAPE AND LEXIMANCER TOOL

Figures 7-9 indicate the top 3 paper's word clouds in a deep learning approach for image forgery detection and explainable AI [2]. Word cloud highlights important words in this research field. The words with smaller fonts are an indication of future research. All the word clouds are drawn from www.wordclouds.com. Important words are copy move, splicing, deep learning, forgery detection, localization of forgery, segmentation, classification, visualization, explainable, and XAI. Less important words are image tampering detection, tampered, manipulation, blocks, and. Figures 10 and 11 are drawn with the ScienScape tool, which shows the top keywords per year from Scopus and the WOS database. In Figure 12, the thematic graph drawn using the Leximancer tool uses a machine learning technique that finds the relation between different author keywords.

Figure 7. Word cloud for "a deep learning approach to detection of splicing and copy-move forgeries in images

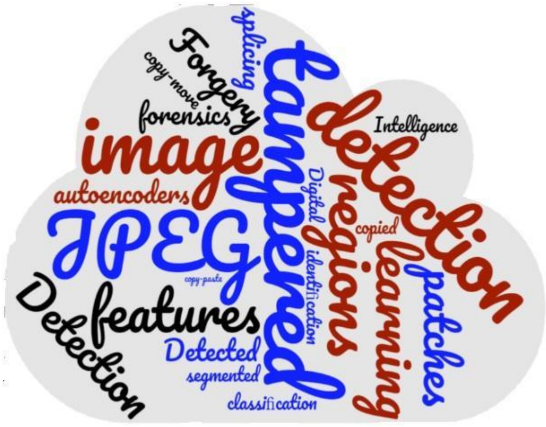


Figure 8. Word cloud for "image region forgery detection: A deep learning approach

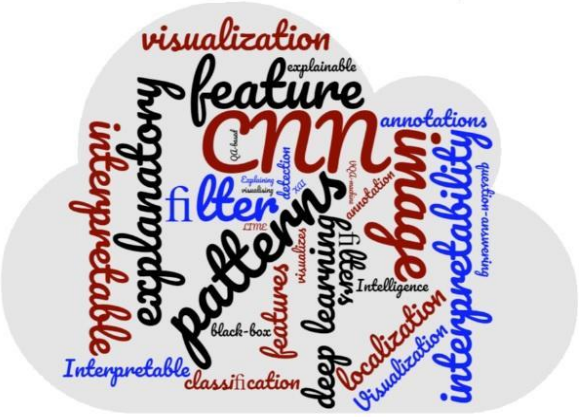
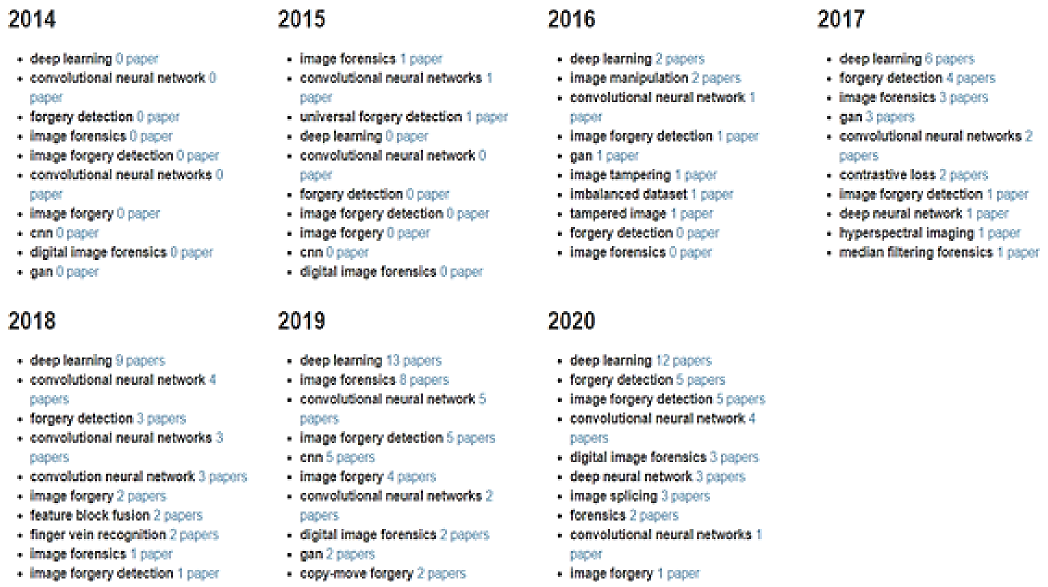


Figure 9. Word cloud for "visual interpretability for deep learning: a survey*"



2014

- deep learning 0 paper
- convolutional neural network 0 paper
- forgery detection 0 paper
- image forensics 0 paper
- image forgery detection 0 paper
- convolutional neural networks 0 paper
- image forgery 0 paper
- cnn 0 paper
- digital image forensics 0 paper
- gan 0 paper

2015

- image forensics 1 paper
- convolutional neural networks 1 paper
- universal forgery detection 1 paper
- deep learning 0 paper
- convolutional neural network 0 paper
- forgery detection 0 paper
- image forgery detection 0 paper
- image forgery 0 paper
- cnn 0 paper
- digital image forensics 0 paper

2016

- deep learning 2 papers
- image manipulation 2 papers
- convolutional neural network 1 paper
- image forgery detection 1 paper
- gan 1 paper
- image tampering 1 paper
- imbalanced dataset 1 paper
- tampered image 1 paper
- forgery detection 0 paper
- image forensics 0 paper

2017

- deep learning 6 papers
- forgery detection 4 papers
- image forensics 3 papers
- gan 3 papers
- convolutional neural networks 2 papers
- contrastive loss 2 papers
- image forgery detection 1 paper
- deep neural network 1 paper
- hyperspectral imaging 1 paper
- median filtering forensics 1 paper

2018

- deep learning 9 papers
- convolutional neural network 4 papers
- forgery detection 3 papers
- convolutional neural networks 3 papers
- convolution neural network 3 papers
- image forgery 2 papers
- feature block fusion 2 papers
- finger vein recognition 2 papers
- image forensics 1 paper
- image forgery detection 1 paper

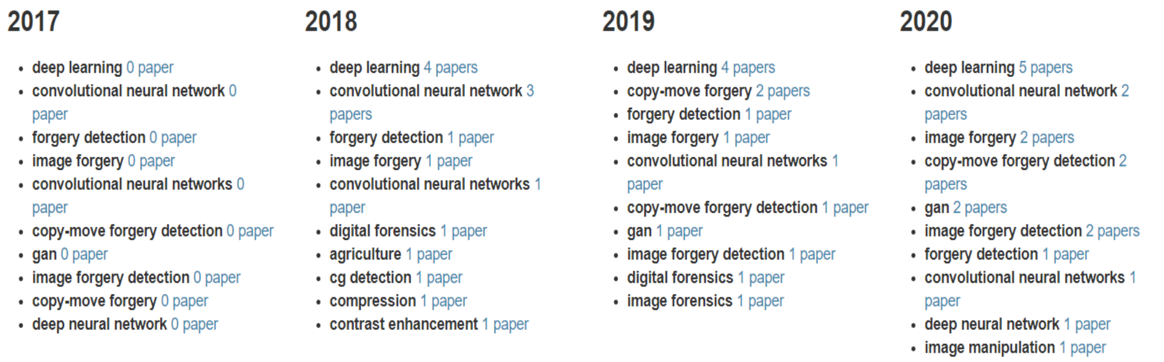
2019

- deep learning 13 papers
- image forensics 8 papers
- convolutional neural network 5 papers
- image forgery detection 5 papers
- cnn 5 papers
- image forgery 4 papers
- convolutional neural networks 2 papers
- digital image forensics 2 papers
- gan 2 papers
- copy-move forgery 2 papers

2020

- deep learning 12 papers
- forgery detection 5 papers
- image forgery detection 5 papers
- convolutional neural network 4 papers
- digital image forensics 3 papers
- deep neural network 3 papers
- image splicing 3 papers
- forensics 2 papers
- convolutional neural networks 1 paper
- image forgery 1 paper

Figure 10. Keywords per year (top keywords per year) for scopus database



2017

- deep learning 0 paper
- convolutional neural network 0 paper
- forgery detection 0 paper
- image forgery 0 paper
- convolutional neural networks 0 paper
- copy-move forgery detection 0 paper
- gan 0 paper
- image forgery detection 0 paper
- copy-move forgery 0 paper
- deep neural network 0 paper

2018

- deep learning 4 papers
- convolutional neural network 3 papers
- forgery detection 1 paper
- image forgery 1 paper
- convolutional neural networks 1 paper
- digital forensics 1 paper
- agriculture 1 paper
- cg detection 1 paper
- compression 1 paper
- contrast enhancement 1 paper

2019

- deep learning 4 papers
- copy-move forgery 2 papers
- forgery detection 1 paper
- image forgery 1 paper
- convolutional neural networks 1 paper
- copy-move forgery detection 1 paper
- gan 1 paper
- image forgery detection 1 paper
- digital forensics 1 paper
- image forensics 1 paper

2020

- deep learning 5 papers
- convolutional neural network 2 papers
- image forgery 2 papers
- copy-move forgery detection 2 papers
- gan 2 papers
- image forgery detection 2 papers
- forgery detection 1 paper
- convolutional neural networks 1 paper
- deep neural network 1 paper
- image manipulation 1 paper

Figure 11. Keywords per year (top keywords per year) for WOS database

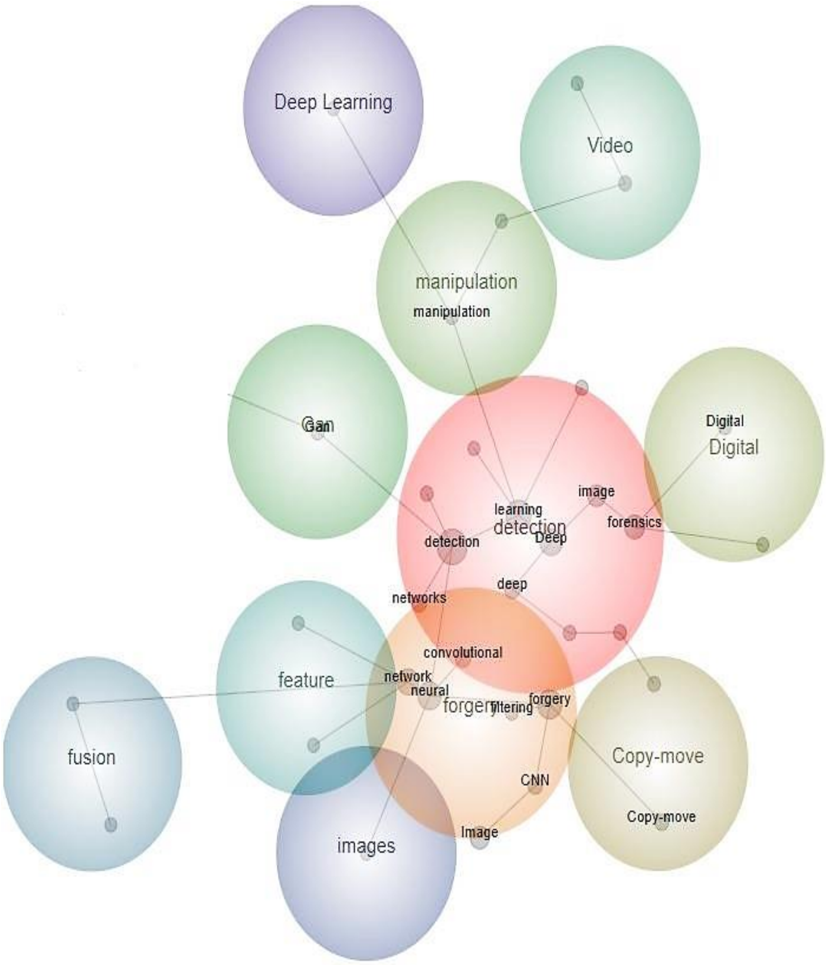


Figure 12. Thematic diagram for keywords

4. DEEP LEARNING APPROACH FOR COPY MOVE AND IMAGE SPLICING FORGERY DETECTION

ConvNets or convolutional neural networks (CNNs) [3] is one of the main categories for image recognition and classifications in neural networks. Face recognition, and detection of objects in an image are some of the fields where CNNs are extensively used. CNN is useful for retrieving meaningful features for image classification. This model [3] uses CNN for identifying copy move and image splicing forgery in an image. This network is pretrained with the help of labeled images from the training image dataset. The same network is then used for extracting the features for the patches, and these features are combined to train the SVM model. In [4], a two-phase deep learning approach is employed to learn the features for detecting image forgery for an image that comes in various formats. In the first stage, Stack AUTOENCODER is used to learn complex features from each patch. In the second stage, the image is broken into patches of 32x32. From each patch, the contextual data is used for finding tampering. In future work, deep belief networks can be used for feature learning. This work [5] presents a technique for tampering detection and recognizing the manipulated regions for the images delivered through the combination of images taken from various camera models. The image is broken into color patches of 64x64. In this, CNN is used to extract features from every patch, and iterative algorithms are used on these features to group them into pristine clusters and forged clusters. Future work focuses on the usage of CNN to learn more features such as blurring, rotations, and resizing. This work [6] uses two different methods for tampering detection and localization. The first method calculates resampling features over overlying areas. A deep learning network and Gaussian conditional random field are utilized for creating the heatmap. In the next method, the resampling features are then passed to the LSTM model to classify the regions as either forged or not. In the future, CNN and LSTM fusion can be used for forgery detection.

Convolutional neural network (CNN) [7] identifies the image tampering with automatic feature learning. This model has five convolutional layers. The network is made up of two fully connected layers and a softmax classifier. In this [8], the CNN network is employed for determining patch-based inpainting operation. CNN is utilized for learning the features of the image. CNN model is trained with a class label matrix of all the pixels of the image. In this weighted cross entropy is used to balance the inpainted and uninpainted pixels. CNN encoder-decoder network is employed for predicting the inpainting probability of each pixel in an image. In this work [9], illuminant maps and CNN are used for the splicing identification. The deep and transfer learning techniques are utilized for extracting features for forgery identification. The classifier is trained with these features. After identifying the image, whether it is fake or not, the next step is to locate the manipulated region. The future work considers localization maps for the identification of forged regions in an image. In [10], a fully convolutional network is used for detecting the image splicing. The Single task FCN is trained with a surface label that identifies every pixel in the image as real or not. SFCN generates a coarse localization output. MFCN is trained on the superficial label and boundary, which shows that the pixel is related to the spliced region’s border. The edge enhanced MFCN uses a surface label and edge probability map, which is better than SFCN and the MFCN approach. In this work [11], CNN based methods are used for handling copy move forgery detection. CNN with CFA (color filter array) features are used for finding and localizing the manipulated region. The CFA interpolation technique develops interrelationship and consistency between the pixels. Therefore, inconsistency in manipulated areas can be used for identifying the forged regions.

R-CNN [12] is employed for finding manipulated regions in forged images. This network provides two streams; one is RGB in which features are extracted from RGB image to detect tampering such as unnatural altered boundaries, and strong contrast difference. The second is the noise stream, which takes advantage of obtaining noise features and identify the inconsistency with real and manipulated regions. The bilinear pooling layer is then used for fusing these features retrieved from two streams. In this [13], deep neural network and conditional random field are used to identify the image’s spliced region, which does not need any prior data. Three unique variations of FCN were utilized to improve the accuracy. Discovery of small manipulated objects is difficult as down-sampling action reduces the real image measurements and makes smaller objects considerably harder to identify. Another issue is the overfitting issue, which will be addressed in future work. Another future work includes the optimization of the network for splicing detection. This framework [14] is used for the detection of forged images. In this network, the image is split into patches. Then resampling features are extracted from these patches. The Hilbert curve finds out the sequencing of patches which are supplied to LSTM. The sampling features detect the correlation between patches. The encoder is used for finding out the spatial location of a manipulated region. Spatial features from the encoder and out features from LSTM are combined for detecting forgery in an image. The decoder model gives the finer representation of the spatial map, which offers the altered region in an image. An improved mask R-CNN model [15] (regional convolutional neural network) is proposed with a Sobel filter to recognize the altered and unaltered region’s distinctive features. This network handles two types of forgeries, such as copy move and splicing. Pixel wise information is used for training the model, and Sobel filters use the edge’s information to identify the manipulated boundaries. Fully convolutional network model (FCNN) [16] is used for image splicing detection. It distinguishes the altering of an image as well as recognizes the forgery of spliced regions. The FCN gives misclassification when there are numerous people are there in the non-spliced region of an image. The improved FCN is used to capture the different features of spliced and non-spliced regions. The improved CNN network is trained with authentic and altered images.

5. ANALYTICAL HIERARCHICAL PROCESSING (AHP) MODEL FOR IMAGE FORGERY DETECTION

Analytical Hierarchical Process is a multi-criteria decision-making mathematical technique. It helps to make associations among qualitative and quantitative attributes and is almost used in every area. In this paper, the AHP calculation is done by using an online tool (https://bpmsg.com/ahp/). This process follows five steps, which are given below: i) Hierarchy modeling, ii) Priorities establishment; iii) Decision related to overall priorities of hierarchy; iv) Checking consistency; and v) Judgment preparation.

In AHP, numerical ranking [17] ranging from 1 to 9 compares attributes at various levels in the hierarchy by considering alternatives. Figure 2 decides different levels of decisions for the analytical hierarchy process of digital image forgery detection. Literature shows that passive image forgery detection is more significant than the active approach; therefore, in level 1 passive approach is used to make a decision. After looking at hierarchy modeling, Tables 1-3 indicates prioritizing attributes for performing AHP on image forgery. The color code in columns P and Q of Table 1-3 shows the number of comparisons that must be made between P and Q. By establishing relevance with scale (1-9), the yellow color indicates significance among the compared attributes. Table 4-6 show consistency checks for considered levels of a first and second set of important image forgery detection attributes. These tables show the priority and rank of a group of important attributes, with the highest priority indicating the most important attribute.

In the second level, the forgery type-dependent technique is considered. Literature shows that these forgeries emphases only on particular types of forgeries, and these are obtaining more recognition as they are frequently performed manipulations in the image. Figures 13-15 show rank levels 1, 2 and 3, respectively. Figure 16 summarizes the AHP implementation result for the undertaken research problem, which gives significant attributes at every level.

Table 1. Formulation of priorities for level 1(types of forgery detection such as active and passive approach)-first set of significant attributes for image forgery detection (1 comparison)

	First set of attributes		Equally important	How much important?							
	P	Q	1	2	3	4	5	6	7	8	9
1	Passive Approach	Active Approach									

Table 2. Consistency check for level 1-the first set of significant attributes for image forgery detection

Consistency check for level 1-the first set of significant attributes			+	-	1	2
	Priority	Rank				
Passive Approach	83.3%	1	0.0%	0.0%	1	5.00
Active Approach	16.7%	2	0.0%	0.0%	0.20	1
Comparisons	1			Principal eigenvalue	2.000	
Consistency check	0.0%		Eigen vector solution	1	iterations, delta	7.7E-34

Table 3. Formulation of priorities for level 2 (types of passive forgery detection such as forgery type-dependent and forgery type independent)-the second set of significant attributes for image forgery detection (1 comparison)

	Second set of attributes		Equally important	How much important?							
	P	Q	1	2	3	4	5	6	7	8	9
1	Forgery type dependent	Forgery type independent									

Table 4. Consistency checking for level 2-the second set of significant attributes for image forgery detection (1 comparison)

Consistency check for level 2 – the second set of significant attributes			+	-	1	2
	Priority	Rank				
Forgery type dependent	87.5%	1	0.0%	0.0%	1	0.142857
Forgery type independent	12.5%	2	0.0%	0.0%	7	1
Comparisons	1			Principal eigenvalue	2.000	
Consistency check	0.0%		Eigen vector solution	1	iterations, delta	0.0E+0

In the third level, equal priorities are given to both copy move and splicing

Table 5. Formulation of priorities for level 3 (types of forgery type-dependent such as copy move and splicing)-the third set of significant attributes for image forgery detection (1 comparison)

	Third set of attributes		Equally important	How much important?							
	P	Q	1	2	3	4	5	6	7	8	9
1	copy move	splicing									

Table 6. Consistency checking for level 3-the third set of significant attributes for image forgery detection (1 comparison)

Consistency check for level 3-the third set of significant attributes			+	-	1	2
	Priority	Rank				
Copy move	50.0%	1	0.0%	0.0%	1	1.00
Splicing	50.0%	1	0.0%	0.0%	1.00	1
Comparisons	1			Principal eigenvalue	2.000	
Consistency check	0.0%		Eigen vector solution	1	iterations, delta	0.0E+0

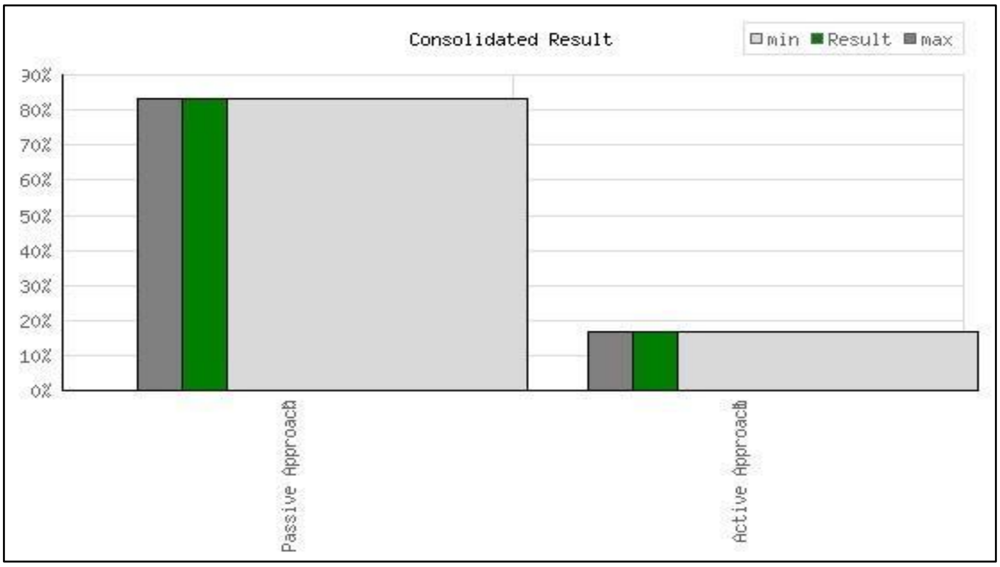


Figure 13. Rank level 1-the first set of significant attributes for image forgery detection

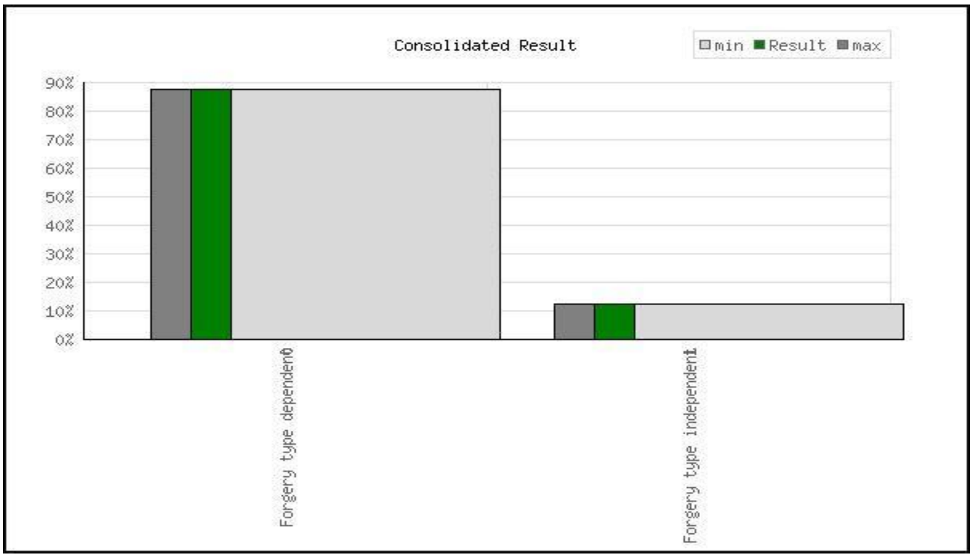


Figure 14. Rank level 2-the second set of significant attributes for image forgery detection

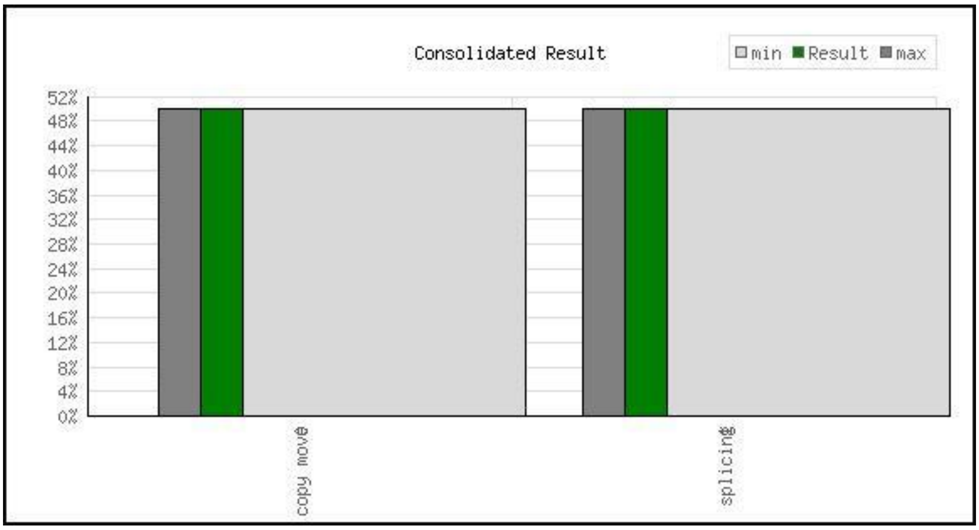


Figure 15. Rank level 3-the third set of significant attributes for image forgery detection

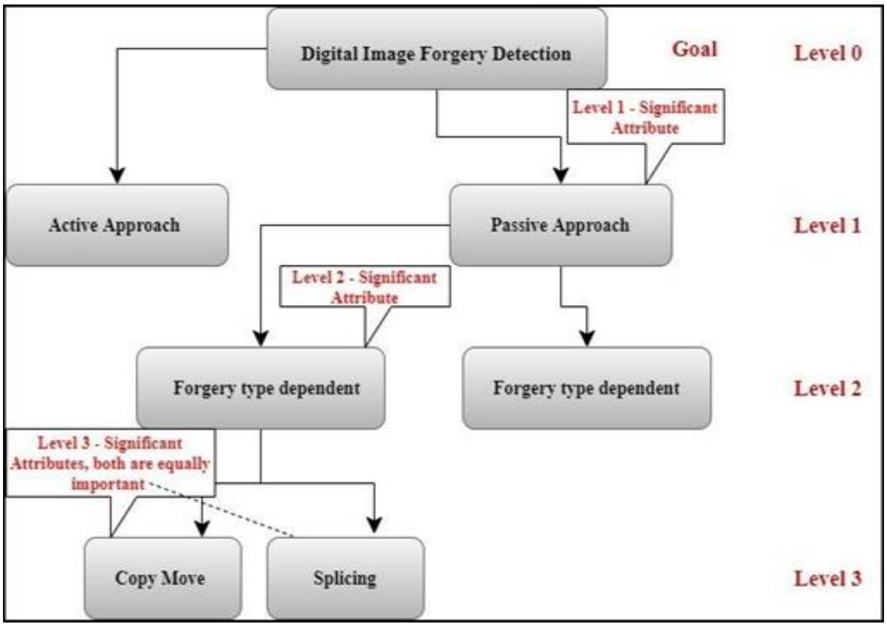

Digital Image Forgery Detection
Goal
Level 0
Level 1 - Significant Attribute
Active Approach
Passive Approach
Level 1
Level 2 - Significant Attribute
Forgery type dependent
Forgery type dependent
Level 2
Level 3 - Significant Attributes, both are equally important
Copy Move
Splicing
Level 3

Figure 16. Summarization of AHP implementation result for the undertaken research problem

6. LITERATURE SURVEY ON EXPLAINABLE AI

Machine Learning is frequently used in research and development and, mainly through Deep Learning and neural network achievement. Nowadays, machine learning networks are overgrowing for making predictions in critical contexts; different AI stakeholders require transparency in predictions [18]. This is achieved by utilizing interpretable networks or growing new strategies and using a local approximation for interpretation. XAI helps the clients to understand and trust a network using various interpretable networks.

AI exhibited to be important in discovering new uses for existing drugs, detecting cancer in tissues, distinguishing heart arrhythmia, and anticipating hypoglycemia in people with diabetes at an earlier time than that of the clinical industry. Machine Learning, Artificial Intelligence, Neural Networks, and Deep Learning models have recently shown their medical field performance, but justifications assisting the outputs of a network remain significant, e.g., in the precision medicine field. Deep Learning is also used in other domains, e.g., self-directed vehicles in transport systems, safety, and economics. Humans using these networks must be trustworthy, interpretable, and tractable. Numerous research and development papers show that XAI has become an essential field in machine learning, which proposed various procedures and systems that specify understandability in various fields.

6.1. XAI in deep learning

Numerous work is presently done in the explainable AI area, but several difficulties exist in explaining deep learning networks. The difficulties such as different challenges in feature selection, feature significance, and visualization methods. There is no consistency in heatmaps, saliency maps and masks, neuron stimulations, and other techniques similar to visualization methods.

This network [19] is practically a simple network based on the convolutional decomposition of images under a sparsity check, and this is an unsupervised approach. This method learns rich feature sets by developing a hierarchy of such decompositions. Applying this model to the natural image generates different sets of filters that capture the edge primitives such as parallel edges, edge connections, parallel lines, curves, and basic geometric shapes and high-order image structures. This does not require settings of different parameters of CNN models such as max-pooling, local contrast normalization. This work [20] presents two visualization methods for deep classification ConvNets. The first visualization method produces an image, which is illustrative of a class of interest. The second method calculates a saliency map, specifically for a given image. It also generates a class, highlighting the areas of the given image. Such a saliency map can be utilized to initialize GraphCut based object segmentation without the need to train devoted segmentation or detection networks. In the future, this method plans to incorporate the image’s explicit saliency maps into the learning process in a more principled way. This method [21] generates sharper visualizations of colorful image regions than the earlier methods specified in literature and can be utilized even in a lack of maxima in max-pooling regions. This work [22] combines bottom-up and two top-down attention. In the object level attention, various patches are given to the system, which selects significant patches to a particular object. The part level attention focuses mainly on local discriminate patterns. This method combines these attentions to train domain-specific deep networks. This technique [23] explored large convolutional neural networks trained for image classification using different ways. This network gives a way to visualize the internal representation of a network. This helps to identify the problems within the network. In this paper, a series of occlusion experiments are conducted for image classification problems, which shows the local structure is more sensitive than its broad scene. In [24], the authors proposed a method that shows single pixels’ visualization using a heatmap in result prediction. In this, a layer-wise relevance propagation (LRP) method is used, which depends on a Taylor series to estimate the point rather than partial derivatives to estimate that point. This paper [25] performs a sensitivity study of one layer CNN that determines different network components affecting its performance. It aims to differentiate between essential and reasonably inconsequential design decisions for sentence classification.

The technique used in this paper [26] can invert images represented in HOG format. This technique shows that few layers in CNNs hold photographically the same data of the image through geometrical and photometric invariance steps. This technique shows that continuously increasingly invariant and abstract information of an image is visible in the model. This method [27] utilized the CNN-LSTM model for the generation of image captions. In this technique for capturing multiple objects inside an image, features are extracted from the lower convolutional layers. Thus, multiple features represent a single image at different localities. The LSTM is trained with features that are extracted from images from various localities. LSTM generates word, and this process is repeated for k-times for creating K-words for image caption.

In this paper, layer-wise relevance propagation (LRP) [28] technique is compared with deep convolutional neural networks and fisher vector classifiers. LPP technique is used for calculating results for a single part of a given image, signifying its impact on the estimation of the classifier for one particular assessment point. Gradient-weighted class activation mapping (Grad-CAM) offers a new way to recognize any CNN-based model. Deep neural networks (DNN) show various pattern recognition tasks, particularly in vision classification problems. DNNs [29] performance is improved by understanding the brain’s internal working. One such method for enhancing DNN is an activation maximization, which works on the principle of neuron’s activation for recognizing any image. Activation maximization is improved by using deep generator network (DGN). This algorithm produces synthetic images that look almost real. Secondly, it discloses the features learned by every neuron in an interpretable way. Thirdly it generalizes to new datasets. And lastly, it tends to be considered as a high-quality generative technique. The proposed network [30] emphasizes the object’s selective features and predicts a class label suitable for that particular image. This method provides a loss function, which generates a collection of meaningful words with the help of global sentence property such as class specificity.

This paper [31] proposed a method that stacks various attention modules. The advantage of this method is that distinct attention modules hold various types of attention for feature learning. Future work focuses on detection and segmentation areas. This paper [32] shows two ways to clarify deep learning networks. The first way focuses on the sensitivity estimation based on the variations in the input. The second way expressively divides the decision based on the given input. The future work will study the hypothetical establishments of interpretability. It also focuses on the relationship between generalizability, compactness, and interpretability. This method [33] involves three steps; in the first step global-and-local attention (GALA) module is utilized for learning combinations of local saliency and global relevance variations in feedforward neural networks. GALA adjusts input feature maps with an attention mask of a similar dimension as the input. Secondly, a vast scale online analysis was defined to enhance ImageNet. Lastly, identification accuracy is improved by humans. Humans can administer attention and only concentrate on the visual features that people favor object identification. This paper uses a technique that endeavors to highlight conditions between successive layers in a CNN to get discriminative pixel areas that control its prediction. In this technique, CNNs are trained for different computer vision undertakings (for example, image captioning and image identification) can reliably confine the objects in an image with minimal extra calculations compared to gradient-based techniques. This method can be used across various types of networks and different data or information modalities.

7. CONCLUSION

In this paper, the survey on image forgeries with the deep learning approach is presented. It also focuses on the survey of XAI for images. Literature shows deep learning frameworks have higher performance accuracy, but the results are not interpretable because of which there is a need for techniques that interpret decisions either in visual format or in wordings. The explainable artificial intelligence survey focuses on various types of XAI techniques in deep learning frameworks that help to interpret the decision. For image forgery detection, three different databases, namely Scopus, Web of Science, and ACM Digital Library, have been used, which shows the number of papers published in this area from various countries. Keyword analysis for image forgery detection is done with the help of various tools such as word cloud, Sciencescape. The thematic graph using the Leximancer tool studies relationship between different author keywords. The words with smaller fonts indicate future research in forgery detection and explainable AI in word cloud figures. The figures drawn with the Sciencescape tool's help provide top keywords per year in forgery detection. The AHP model is carried out on digital image forgery detection at multiple levels. The digital image forgery detection AHP model is based on existing literature. This survey can be used by various research scholars, technologists, and experts working in multiple fields where image and interpretation have more significance. Till date, no research paper considered the combined scenario of image forgery with explainable AI; this is the key aspect of this research paper.

REFERENCES

[1] S. Walia and K. Kumar, "Digital image forgery detection: a systematic scrutiny," *Australian Journal of Forensic Sciences*, vol. 51, no. 5, pp. 488-526, 2019, doi: 10.1080/00450618.2018.1424241.

[2] K. D. Kadam, S. A. Ahirrao, and K. Kotecha," Bibliometric Analysis of Passive Image Forgery Detection and Explainable AI," DigitalCommons @ University of Nebraska-Lincoln, 2020

[3] Y. Rao and J. Ni, "A deep learning approach to detection of splicing and copy-move forgeries in images," *2016 IEEE International Workshop on Information Forensics and Security (WIFS)*, Abu Dhabi, United Arab Emirates, 2016, pp. 1-6, doi: 10.1109/WIFS.2016.7823911.

[4] Y. Zhang, J. Goh, L. L. Win, and V. Thing, "Image region forgery detection: A deep learning approach," *Proceedings of the Singapore Cyber-Security Conference (SG-CRC) 2016*, vol. 14, 2016, pp. 1-11.

[5] L. Bondi, S. Lameri, D. Guera, P. Bestagini, E. J. Delp, and S. Tubaro, "Tampering Detection and Localization Through Clustering of Camera-Based CNN Features," *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA, 2017, pp. 1855-1864, doi: 10.1109/CVPRW.2017.232.

[6] J. Bunk, J. H. Bappy, T. M. Mohammed, L. Nataraj, A. Flenner, B.S. Manjunath *et al.*, "Detection and Localization of Image Forgeries Using Resampling Features and Deep Learning," *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA, 2017, pp. 1881-1889, doi: 10.1109/CVPRW.2017.235.

[7] N. Huang, J. He, and N. Zhu, "A Novel Method for Detecting Image Forgery Based on Convolutional Neural Network," *2018 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, New York, NY, USA, 2018, pp. 1702-1705, doi: 10.1109/TrustCom/BigDataSE.2018.00255.

[8] X. Zhu, Y. Qian, X. Zhao, B. Sun, and Y. Sun, "A deep learning approach to patch-based image inpainting forensics," *Signal Processing: Image Communication*, vol. 67, pp. 90-99, 2018, doi: 10.1016/j.image.2018.05.015.

[9] T. Pomari, G. Ruppert, E. Rezende, A. Rocha and T. Carvalho, "Image Splicing Detection Through Illumination Inconsistencies and Deep Learning," *2018 25th IEEE International Conference on Image Processing (ICIP)*, Athens, Greece, 2018, pp. 3788-3792, doi: 10.1109/ICIP.2018.8451227.

[10] R. Salloum, Y. Ren, and C. C. Jay Kuo, "Image Splicing Localization using a Multi-task Fully Convolutional Network (MFCN)," *Journal of Visual Communication and Image Representation*, vol. 51, pp. 201-209, 2018, doi: 10.1016/j.jvcir.2018.01.010.

[11] L. Liu, Y. Zhao, R. Ni, and Q. Tian, "Copy-move forgery localization using convolutional neural networks and CFA features," *International Journal of Digital Crime and Forensics*, vol. 10, no. 4, pp. 140-155, 2018, doi: 10.4018/IJDCF.2018100110

[12] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning Rich Features for Image Manipulation Detection," 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 1053-1061, doi: 10.1109/CVPR.2018.00116.

[13] B. Liu and C. M. Pun, "Locating splicing forgery by fully convolutional networks and conditional random field," *Signal Processing: Image Communication*, vol. 66, pp. 103-112, 2018, doi: 10.1016/j.image.2018.04.011

[14] J. H. Bappy, C. Simons, L. Nataraj, B. S. Manjunath, and A. K. Roy-Chowdhury, "Hybrid LSTM and Encoder-Decoder Architecture for Detection of Image Forgeries," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3286-3300, 2019, doi: 10.1109/TIP.2019.2895466

[15] X. Wang, H. Wang, S. Niu, and J. Zhang, "Detection and localization of image forgeries using improved mask regional convolutional neural network," *Mathematical Biosciences and Engineering*, vol. 16, no. 5, pp. 4581-4593, 2019 doi: 10.3934/mbe.2019229.

[16] J. Zhang, Y. Li, S. Niu, Z. Cao, and X. Wang, "Improved Fully Convolutional Network for Digital Image Region Forgery Detection," *Computers, Materials and Continua*, vol. 58, no. 2, pp. 287-303, 2019, doi: 10.32604/cmc.2019.05353.

[17] T. L. Saaty, "What Is The Analytic Hierarchy Process?," *Mathematical Models for Decision Support,* vol. 48, pp. 109-121, 1988.

[18] A. Preece, D. Harborne, D. Braines, and R. Tomsett, "Stakeholders in Explainable AI," *arXiv:1810.00184v1*, pp. 1-6, 2016.

[19] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional Networks," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, 2010, pp. 2528-2535, doi: 10.1109/CVPR.2010.5539957.

[20] K. Simonyan, "Deep Inside Convolutional Networks : Visualising Image Classification Models and Saliency Maps," *arXiv :1312.6034v2,* pp. 1-8, 2013.

[21] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving For Simplicity : The All convolutional Net," *arXiv:1412.6806*, pp. 1-14, 2015.

[22] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The Application of Two-level Attention Models in Deep Convolutional Neural Network for Fine-grained Image Classification," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Boston, MA, USA, 2015, pp. 842-850, doi: 10.1109/CVPR.2015.7298685.

[23] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," *European Conference on Computer Vision-ECCV 2014*, vol. 8689, 2014, pp. 818-833.

[24] S. Bach, A. Binder, G. Montavon, and F. Klauschen, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *Plos One*, vol. 10, no. 7, 2015, Art. No. e0130140, doi: 10.1371/journal.pone.0130140.

[25] B. C. Wallace, Ye Zhang, "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification," *arXiv:1510.03820,* 2014.

[26] A. Mahendran and A. Vedaldi, "Understanding Deep Image Representations by Inverting Them," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Boston, MA, USA, 2015, pp. 5188-5196, doi: 10.1109/CVPR.2015.7299155.

[27] K. Xu, A. Courville, R. S. Zemel, and Y. Bengio, "Show, Attend and Tell : Neural Image Caption Generation with Visual Attention," *ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning,* vol. 37, 2014, pp. 2048-2057.

[28] F. Heinrich, "Controlling Explanatory Heatmap Resolution And Semantics Via," *arXiv:1603.06463*, 2016.

[29] J. Clune, J. Yosinski, and T. Brox, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," *NIPS 2016,* pp. 1-29, 2016.

[30] L. A. Hendricks and J. Donahue, "Generating Visual Explanations," *Conference: European Conference of Computer VisionA*t: Amsterdam, NL, 2016, doi: 10.1007/978-3-319-46493-0_1.

[31] F. Wang Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, Xiaoou Tang, "Residual Attention Network for Image Classification," *CVF,* no. 1, pp. 3156-3164, 2017.

[32] W. Samek, T. Wiegand, K. -R. Muller, "Explainable Artificial Intelligence: Understanding, visualizing and interpreting deep learning models," *ITU Journal: ICT Discoveries,* no. 1, pp. 39-48, 2017.

[33] D. Linsley, D. Shiebler, S. Eberhardt, and T. Serre, "Global-and-local attention networks for visual recognition," *A Conference Paper at ICLR 2019,* 2019, pp. 1-21.