symmetry

MDPI

Article

Using XAI for Deep Learning-Based Image Manipulation Detection with Shapley Additive Explanation

Savita Walia [1], Krishan Kumar [1], Saurabh Agarwal [2,*] and Hyunsung Kim [3,*]

[1] University Institute of Engineering and Technology, Panjab University, Chandigarh 160014, India
[2] Amity School of Engineering & Technology, Amity University Uttar Pradesh, Noida 201301, India
[3] School of Computer Science, Kyungil University, Kyungbuk 38428, Korea
* Correspondence: saurabhnsit2510@gmail.com (S.A.); kim@kiu.ac.kr (H.K.)

Abstract: In the arena of image forensics, detecting manipulations in an image is extremely significant because of the use of images in different fields. Various detection techniques have been suggested in the literature that are based on digging out the features from images to unveil the traces left by manipulation operations. In this paper, a deep learning-based approach is proposed in which a residual network is used to learn deep, complex features from preprocessed images for classification into authentic and forged images. There is statistical symmetry in similar types of images and asymmetry in different types of images. The proposed scheme can highlight the statistical asymmetry between authentic and forged images. In the proposed scheme, firstly, an RGB image is analyzed for different JPEG compression levels. The obtained difference between the error levels is used to extract enhanced LBP code. Then, the scale- and direction-invariant LBP (SD-LBP) code is transformed into SD-LBP feature maps to feed to a deep residual network. Next, the concept of explainable artificial intelligence (XAI) is used to help provide explanations and interpret the output, thereby raising the credibility of the proposed approach. The unique feature selection approach employed is the kernel SHAP method, which is focused on the Shapley values. This technique is used to pinpoint the specific characteristics that are responsible for the aberrant behavior of the forged images dataset. Later, the deep learning-based model is trained and validated using these feature sets. A pre-activation version of ResNet-50 architecture is used that achieved an accuracy of 99.31%, 99.52%, 98.05%, and 99.10% on CASIA v1, CASIA v2, IMD 2020, and DVMM datasets, respectively. The capability of the pretrained residual network and rich textural features, which are scale- and direction-invariant, helps to expand the detection accuracy of the proposed approach. The results confirmed that the method either produced competitive results or outperformed existing methods.

Keywords: explainable AI; image forensics; image splicing; deep learning technique; image manipulation detection

Citation: Walia, S.; Kumar, K.; Agarwal, S.; Kim, H. Using XAI for Deep Learning-Based Image Manipulation Detection with Shapley Additive Explanation. *Symmetry* **2022**, *14*, 1611. https://doi.org/10.3390/sym14081611

Academic Editor: Mihai Postolache

Received: 14 July 2022
Accepted: 2 August 2022
Published: 5 August 2022

1. Introduction

Due to powerful advancements in technology, the internet, and software, a plethora of applications are available that can be used to forge an image to fulfill some fraudulent purpose. The image editing software being developed can be used in a number of ways to forge images that cannot be distinguished by human judgment. Images are used all over the place, on the fronts of magazines, in papers, in courts, and so forth. Therefore, it is of utmost importance to develop an effective method for forgery detection. To detect manipulations, image-specific attributes are used to discriminate between forged and authentic images. In traditional passive methods of forgery detection [1,2], the learning workflow comprises preprocessing the images, drawing out features using hand-crafted algorithms, selecting features from the separated list of features, and then feeding the selected features to a learning model. The process of feature engineering [3] is a complex and time-consuming process in which an image undergoes various statistical, mathematical,

and image processing operations to generate a particular set of features. Taking into account the complexity of the feature extraction process in machine learning algorithms, deep learning frameworks have come into play in the recent couple of years for various image processing tasks and applications [4].

Multiple methods for the detection of image forgery using hand-crafted features are available in the literature to uncover the statistical asymmetry of the forged images. In this segment, we discuss a few inactive methods of forgery detection related to the proposed approach. A forgery detection strategy based on curvelet transform was proposed by Al-hammadi and Muhammad [5]. The chrominance element of the image was utilized to transform it into several curvelet wedges. Then, LBP histograms were extracted from wedges of varied scales and directions, which were further fed to the support vector machine (SVM) classifier as features. The method achieved 91.74% accuracy on CASIA v1 and 97.0% accuracy on CASIA v2. In Muhammad et al.'s [6] scheme, LBP was combined with steerable pyramid transform (SPT) for the detection of manipulations in images. Again, the authors utilized the YCbCr color space and applied SPT on chrominance components yielding sub-bands. Then, LBP histograms were used to represent texture in SPT sub-bands. The histograms from every sub-band were combined to make a feature vector. Classification was performed using LIBSVM, which obtained an accuracy of 97.33% on CASIA v2 and 94.89% on CASIA v1. Hussain et al. [7] proposed a Weber's law-based neighborhood descriptor to extricate features from an image, and the features were further reduced on the basis of local learning-based (LLB) selection. The reduced feature set was fed as input to SVM for categorization and achieved an accuracy of 94.19% with CASIA v1, 96.61% with CASIA v2, and 94.17% with the Columbia dataset. In [8], the chroma component of the input image was separated into blocks; then, the LBP pattern of every block was converted to the discrete cosine transform (DCT) domain. The standard deviation of DCT coefficients of every block altogether were treated as a feature to be used for categorization using SVM. The technique attained an accuracy of 97%, 97.7%, and 97.5% on the CASIA v1, DVMM, and CASIA v2 datasets, respectively. Another method proposed by Vidyadharan and Thampi [9] extracted multitexture features (LBP, BSIF, LPQ, and BGP) from various SPT sub-bands. All the features from all sub-bands were extracted and integrated to build a large dimensional feature vector. The best discriminating features were picked out using the relief method. The final feature set is then provided to the random forest classifier as input for classification. In [10], the handcrafted features were combined with deep features extracted from a pretrained network in order to get more abstract representations of the images. The major drawback of this method is the increase in computational complexity with the feature size.

In recent times, machine learning-based methods were not considered very well suited for forgery detection because, to extract appropriate image representations, very complex and mathematical algorithms are required to hand-craft the features as per the requirement of the problem. To overcome this problem, deep learning methods are being used in various digital image processing problems. Various data-driven approaches such as deep learning with the help of convolutional neural networks have demonstrated remarkable results in most image classification problems [11–15]. These networks can grasp rich feature portrayals straightforwardly from the images. The activation layers of pretrained CNN architectures can be adapted for extricating features for numerous applications in computer graphics and image processing.

In Rao and Ni's [16] scheme, a convolutional neural network was applied to acquire the image features for classification. Rather than using a random strategy for weight initialization, a set of high-pass filters was used for calculating residual maps in the spatial rich model. It also functioned as a regularizer that suppressed the contents of the image and captured the delicate traces established by manipulations. Two different kinds of input were fed to the CNN architecture i.e., spatial- and frequency-domain-based features, in Shi et al.'s [17] scheme. The architecture was called dual-domain CNN as it utilized features from both domains. In a recent approach [18], image patches were used as input to

the ResNet50v2 architecture. The network used the weights of the YOLO convolutional neural model for initialization. A very lightweight CNN model was proposed in [19] with minimum parameters that used the difference in the recompressed image and the original image to be fed to the proposed CNN architecture. Image manipulations were detected using a convolutional neural network with blockchain (CNNB), which combined blockchain technology with deep learning image processing methods. Recently, various approaches proposed by researchers were based on block or patched inputs to convolutional neural networks [16,17,20]. When patch-based data are fed to CNNs, it leads to a loss of evidence for manipulation detection. Furthermore, to train a CNN architecture afresh is very expensive and complex. Hence, a pretrained architecture, i.e., ResNet50 was commonly used for classifying the images into forged and authentic categories according to the textural statistics of the image. It has been observed in the literature [9,21,22] that LBP is an adequate and computationally efficient texture descriptor. These studies inspired us to integrate LBP along with error level analysis, which helps in recognizing the areas that belong to different compression levels to formulate an apt representation for the images.

The issue with deep learning and machine learning interpretability is not brand new. It has been around since the 1970s when scientists were attempting to understand how expert systems produced their results [23]. However, Van Lent [24] coined the term explainable AI (XAI) in 2004 for use in modeling, simulations, and game applications. Initially, this matter was taken seriously; however, as time went on, attention switched to honing the models' accuracy and creating new algorithms. Researchers and practitioners have recently shown a renewed interest in explainable AI [25], particularly for deep learning-based architectures such as autoencoders, pre-trained architectures, and other sophisticated models. One of the XAI strategies utilized to clarify and enhance the findings of the proposed approach is the SHAP framework [26] based on Shapley values. Shapley values show the real contribution of each variable to model prediction, which is helpful in identifying and explaining asymmetry [27]. Substantial Shapley values on variables result in large reconstruction errors, making them more crucial than other variables for such anomaly identification models [27]. However, the use of Shapley values to explain the performance of the pretrained models in image manipulation detection has not been implemented in any relevant research studies.

In this work, a unique technique is proposed for digital image manipulation detection using error level analysis (ELA) and scale- and direction-invariant local binary pattern (SD-LBP). Error level analysis helps in finding the areas within an image that have different compression levels. For JPEG compressed images, it is assumed that the entire image is approximately of the same magnitude. If a certain segment of an image is at a considerably different error level, then it is most likely that the region was manipulated. An extension of LBP is used in this work, which is scale- and direction-invariant. Furthermore, these LBP patterns are converted into feature maps to be fed to our deep neural network architecture.

A pretrained ResNet-50 architecture is used for the classification of images into two classes, i.e., forged and authentic. ResNet-50 is a state-of-art model trained to give high accuracy on image classification tasks. It extracts high-level hierarchical features that are very specific to image classification tasks. Therefore, to use this architecture specifically for image forgery detection, certain preprocessing is required. In our case, we used textural features to be fed to the selected neural network. The major contributions of the proposed work can be summarized as follows:

- A unique feature selection strategy is delivered in this paper that is based on Shapley additive explanation (SHAP), which is an XAI approach. The explanation framework for the real deep learning-based model is constructed using the kernelExplainer approach.
- A pretrained ResNet50 architecture is used by initializing the network with 'imagenet' weights to retrain the network with forgery detection datasets.
- The proposed approach takes advantage of the difference in compression levels of the image and identifies the possible manipulated regions to be fed to the network.

- The proposed approach uses a local binary pattern that helps in analyzing the manipulations based on the texture of the images.

The remainder of the paper is arranged as follows: Section 2 explains the details of the proposed approach. The experimental arrangement and observations are mentioned in Section 3. Conclusions are drawn in Section 4.

2. Proposed Scheme

Image manipulation is easy to perform with the help of the latest applications. These fake images can be the reason for big social and political conflicts. In this paper, a robust scheme is proposed to address the issue of fake images. The outline of the proposed research method for image manipulation detection is shown in Figure 1. The proposed image manipulation detection method works by analyzing different levels of compression and taking into consideration the texture information of the image. This is achieved by performing error level analysis on images and then finding the scale- and direction-invariant local binary pattern (SD-LBP) from the obtained images. Then, SD-LBP is converted into feature maps which are further fed to deep neural network architecture. The details of the various steps of the approach are mentioned below.

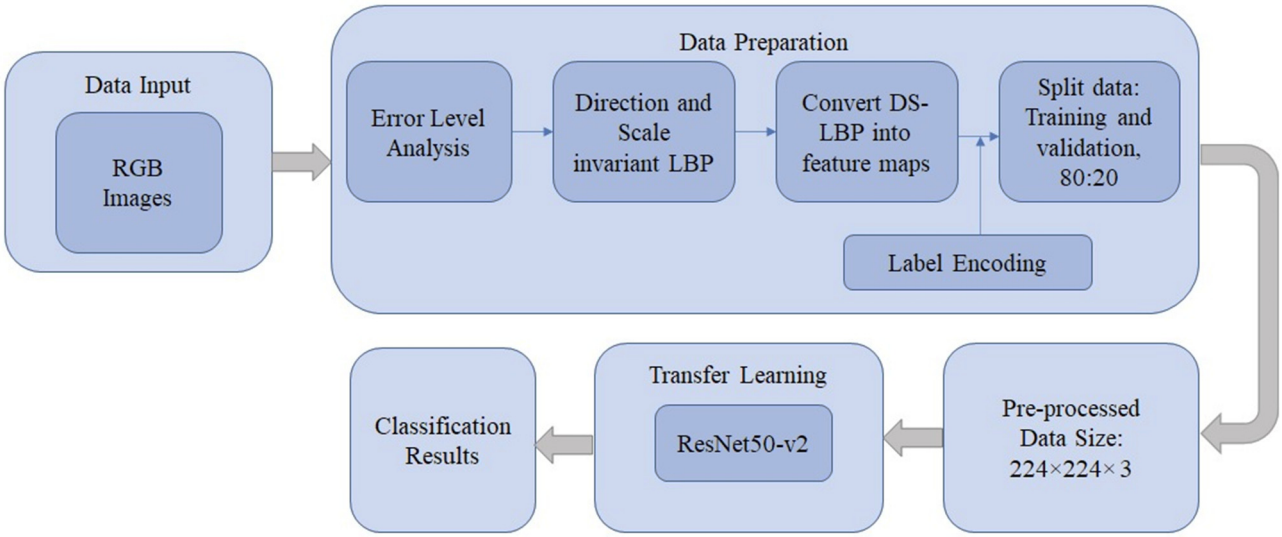


Figure 1. Overview of the proposed approach.

2.1. Error Level Analysis

Most of the methods in the literature discussed above converted the RGB image into grayscale to discover the textural features of the image. While the conversion process takes place, substantial useful information is discarded. The idea behind using the ELA technique for image manipulation detection is that the manipulated image will always be resaved after manipulation. Therefore, the number of resaves for the manipulated image will always be higher as compared to an authentic image. Hence, the compression levels of the authentic and forged image will be different. ELA [28] helps in detecting the difference in the compression levels of the image. It works by retaining the image at a known compression rate, usually at 95%, and then the change between the resaved image and the original image is calculated. This calculated difference is termed the error. On the off chance that there is just about no change, the cell has arrived at its local minimum for error at that specific quality level. Nonetheless, on the off chance that there is a curiously large amount of change, the pixels are not at their local minimum and viably unique. The resultant ELA pictures had their brightness upgraded to additionally separate the variations. An example of the obtained ELA image from an RGB image is shown in Figure 2. The image in Figure 2 was taken from the CASIA v2 dataset, which is a benchmark dataset in the field of image forensics.

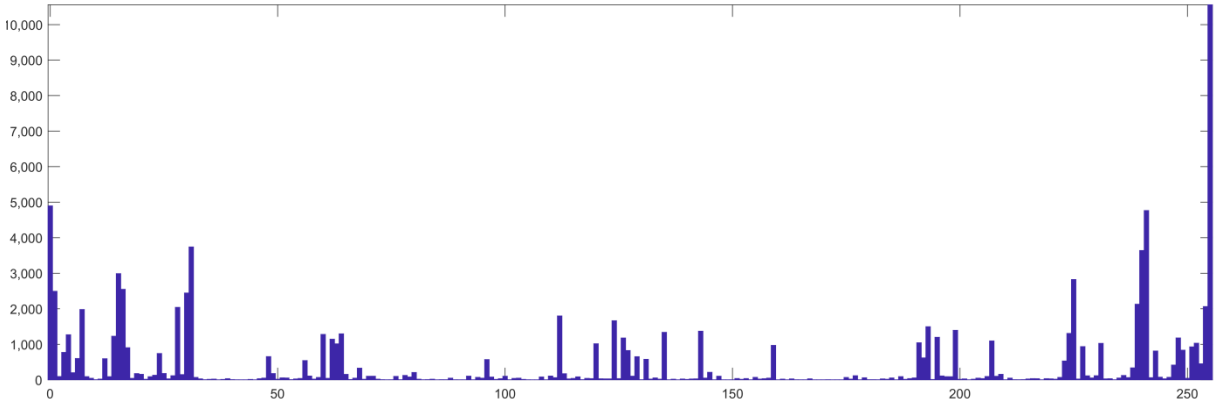
Figure 2. RGB image and corresponding obtained ELA image.

2.2. Direction- and Scale-Invariant Local Binary Patterns

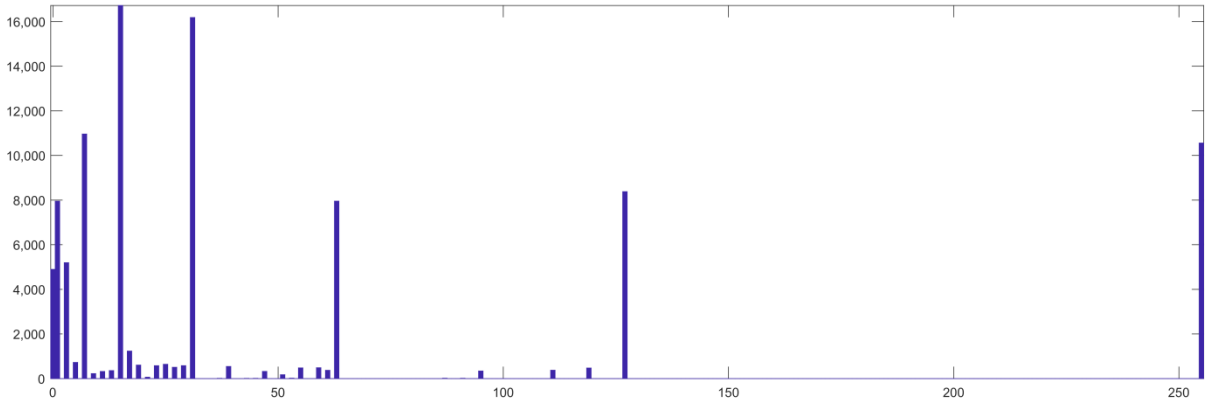
Local binary patterns are highly discriminative in nature; the only disadvantage of LBP is that it is exceptionally delicate to relative transformations. To compensate for image rotations, a comprehensively used turn-invariant encoding of LBP representation is applied, which requires either a verifiable or express plan of representations that will be finished at the encoding level. Different variations of LBP have been presented that have perhaps lessened the discriminative intensity of the component portrayals by extending the limit of describing the center microstructures at various scales.

Additionally, the LBP's rotation-invariant replacement is unable to account for image scaling. Therefore, we used scale- and direction-resistant LBP on the aforementioned ELA images that were collected. To do this, scale-invariant features and pivot-invariant features were processed separately, and then both variable representations were combined to increase the discriminatory strength of the LBP. By modifying the variables at the extraction level while using a powerful universal estimator, resistance to direction was achieved. On the basis of the transmission of scale-standardized Laplacian reactions in a scale-space representation, scale-adjusted features were determined in relation to the evaluated size of the image. The two properties mentioned above were combined to create the final SD-LBP using a multiscale representation. A histogram was created from the SD-LBP code occurrences in the image. Figure 3 illustrates the number of occurrences of regular LBP codes, a sparse histogram of SD-LBP codes, and a tight histogram of SD-LBP codes of the image shown in Figure 2.

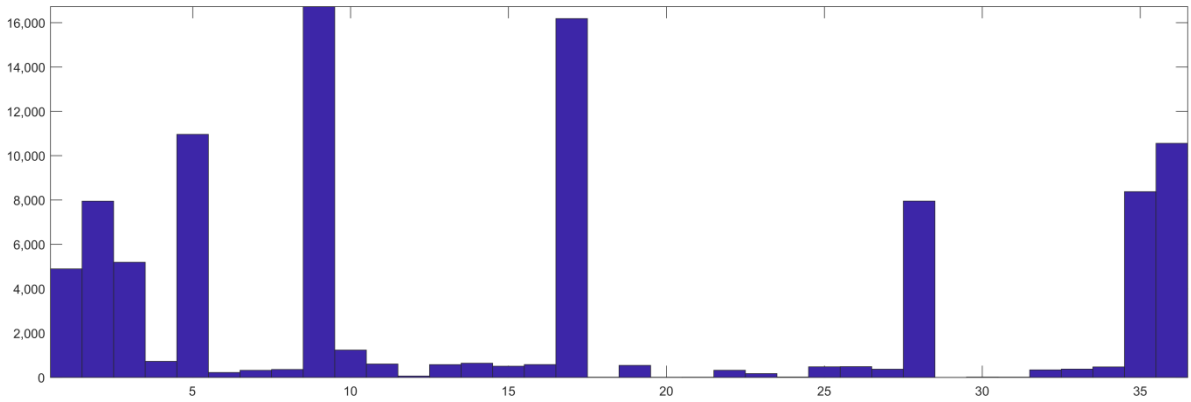
The neural network design cannot receive the acquired SD-LBP codes in their original form. Therefore, we must translate the SD-LBP into feature maps in order to feed them to the neural architecture. To transform SD-LBP variables into LBP maps, multidimensional scaling is used by converting the pattern values into points in a metric space. The altered points can be averaged using convolutional techniques, but their distances are essentially the same as the original code-to-code distances. Distance represents the fundamental similarity of the pixel intensity configurations that each LBP code pattern is made up of. The distances between every possible code value are shown on a comprehensive disparity grid. MDS searches for a mapping of the code patterns to a low-dimensional measurement space for a specific disparity grid. Furthermore, we used their earth mover's distance (EMD) in place of Hamming distance to adjust to variations in spatial areas of pixel code patterns. It is applied as a percentage of the contrast between the two LBP codes.



(**a**) Regular LBP histogram



(**b**) SD-LBP sparse histogram



(**c**) SD-LBP tight histogram

Figure 3. Histograms of regular LBP codes and SD-LBP codes.

2.3. ResNet50-v2 Architecture

The obtained SD-LBP maps are further used for the classification of images into authentic and forged classes. ResNet-50 version 2 is a deep neural network that has already been trained on numerous images. Networks with convolution, pooling, activation, and fully associated layers stacked on top of one another are comparable to residual networks [29]. The only difference is that the levels are connected by an identity. With these identity links, the network seeks out the output functions directly and without assistance. We used ResNet-50 V2, which is based on using the pre-activation of weight layers as a substitute for post-activation. Version 1 of ResNet-50 architecture uses the post-activation.

The pre-activation of ResNet-50 is shown in Figure 4. ResNet-50 V2 eliminates the last nonlinearity, subsequently making way for the input to produce output in the form of an identity association. In ResNet-50 V2, batch normalization and ReLu activation to input are applied before the convolution operation. This property of ResNet V2 helps in protecting the network from the vanishing gradient problem. The parameters used for the ResNet50 architecture are shown in Table 1. The strategy behind this network is to let the network fit the residual mapping rather than have layers learn the underlying mapping. Thus, the network is fitted instead of using, say, the initial mapping of H(x) as shown in Equation (1).

$$F(x) = H(x) - x \text{ which gives } H(x) = F(x) + x(1). \quad (1)$$

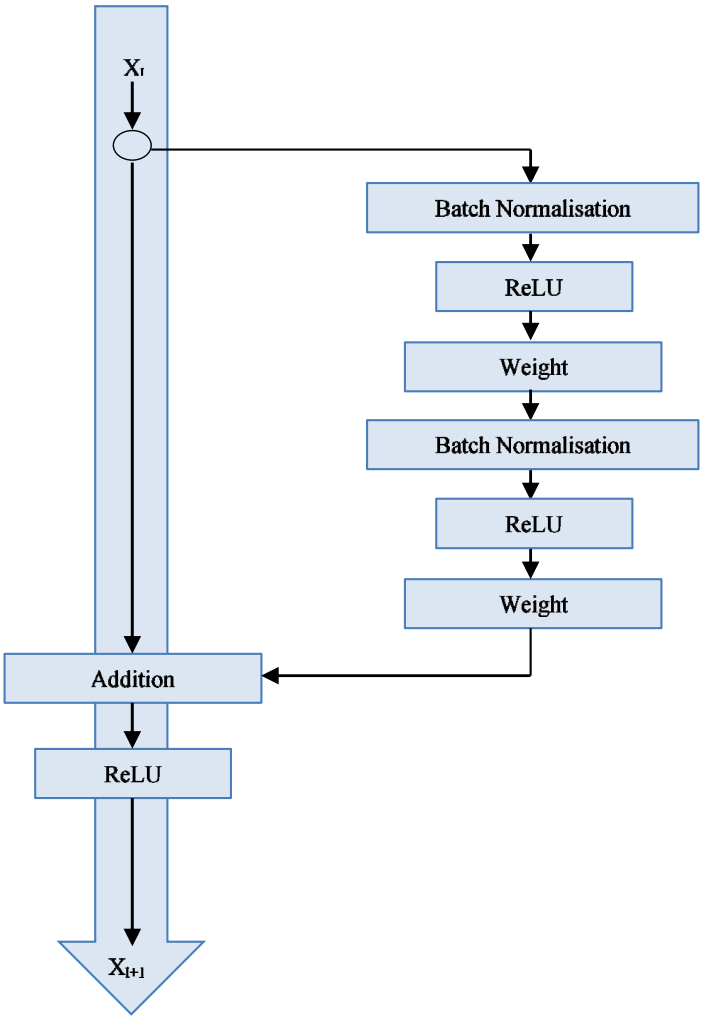


Figure 4. Pre-activation in ResNet-50 v2.

Table 1. Parameters used for ResNet50.

Number_of_Layers	50
Batch_size	32
Weights	'imagenet'
Input_size	224, 224, 3
ResNet_50_pooling	'max'
Dense_layer_activator	'sigmoid'
Objective_function	'binary_crossentropy'
Loss_metrics	['acc']
Learning_rate	0.00005
Epochs	50

2.4. Shapley Additive Explanation (SHAP)

The black-box nature of AI-based systems produces outstanding outcomes but without any justification; thus, people tend to lose faith in them when they are used for making important decisions. Similar to this, the unbounded nature and absence of supervisory character of anomalies in a network can occasionally become a significant barrier to the adoption of deep learning-based models, and it can be challenging to identify the anomaly's root cause in an unsupervised learning framework. Adadi et al. [30] provided illustrations of several XAI techniques, and the SHAP architecture was one of the techniques covered in this section. The SHAP architecture was also utilized in text analysis to predict tweet popularity [31].

Shapley additive explanation (SHAP) [26] is an integrated model put forth for explaining and addressing the interpretability of intricate models such as ensemble combinations and deep convolutional neural networks. The SHAP framework integrates previously proposed approaches from the additive feature acknowledgment class, which includes DeepLIFT [31] and LIME [32], whereby methods from this class contain explanatory models with a linear function of binary variables.

3. Experimental Setup and Results

The forged image issue is crucial and needs to be solved with better schemes. Four accessible benchmark datasets for image fakery detection were used for conducting the assessment of the proposed approach. The datasets utilized in this study were Columbia DVMM, CASIA TIDE v1 [33], CASIA TIDE v2 [33], and IMD 2020 [34]. The Columbia DVMM dataset was originally provided by the Digital Video Multimedia (DVMM) Lab; it contains images with no color information, and the images are very unrealistic. Details of the datasets are delivered in Table 2. IMD 2020 is a new dataset with which researchers have not experimented using their methods. On the other hand, the DVMM dataset contains very unrealistic image forgeries; thus, it is not very well suited for evaluating the performance of the forgery detection methods. For testing our method rigorously on different types of images based on the type of forgery, size of images, etc., we included the abovementioned four datasets for evaluation.

Table 2. Particulars of the datasets used for the assessment.

Dataset	No. of Original Images	No. of Forged Images	Image Resolution	Image Format
Columbia DVMM	933	912	128 × 128	BMP
CASIA TIDE v1	800	925	384 × 256	JPEG
CASIA TIDE v2	7491	5123	240 × 160 to 900 × 600	JPEG, TIFF, BMP
IMD 2020	35,000	35,000	384 × 256 to 1200 × 1051	JPEG

3.1. Effectiveness of the Approach

Various experiments were executed to learn the efficiency of the proposed approach. Firstly, the ResNet architecture was implemented on all four datasets. The effects of different input images were analyzed for the detection rate of the proposed approach. Simple RGB images were fed to the ResNet architecture without any preprocessing. In another experiment, LBP feature maps were used to feed the ResNet architecture. Then, ELA images were fed to the ResNet architecture. The detection rates of each of these cases were compared with the proposed method, and the results on all four datasets are shown in Figure 5. The results clearly show that a higher detection rate was achieved when local binary patterns were calculated from error level analysis of the image and were fed to the ResNet-50 architecture for classification into authentic and forged images. Figure 6 shows the ROC curves obtained during the training phase, validation phase, and testing phase, as well as combined curve for the overall performance of the approach, on the CASIA v2 dataset. Figure 7 shows the performance plot for the proposed model, which shows the cross-entropy errors at each epoch. The model achieved the best validation performance at epoch 40.

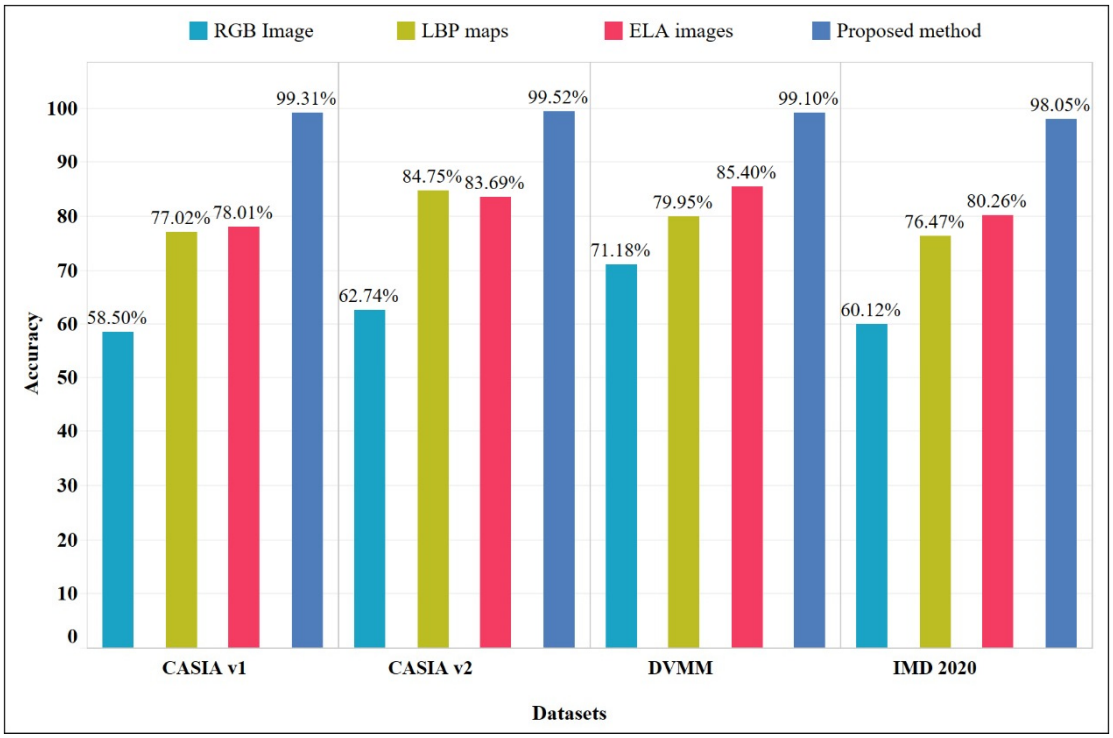


Figure 5. Detection accuracy on various datasets by considering RGB images, LBP maps, ELA images, and the proposed approach.

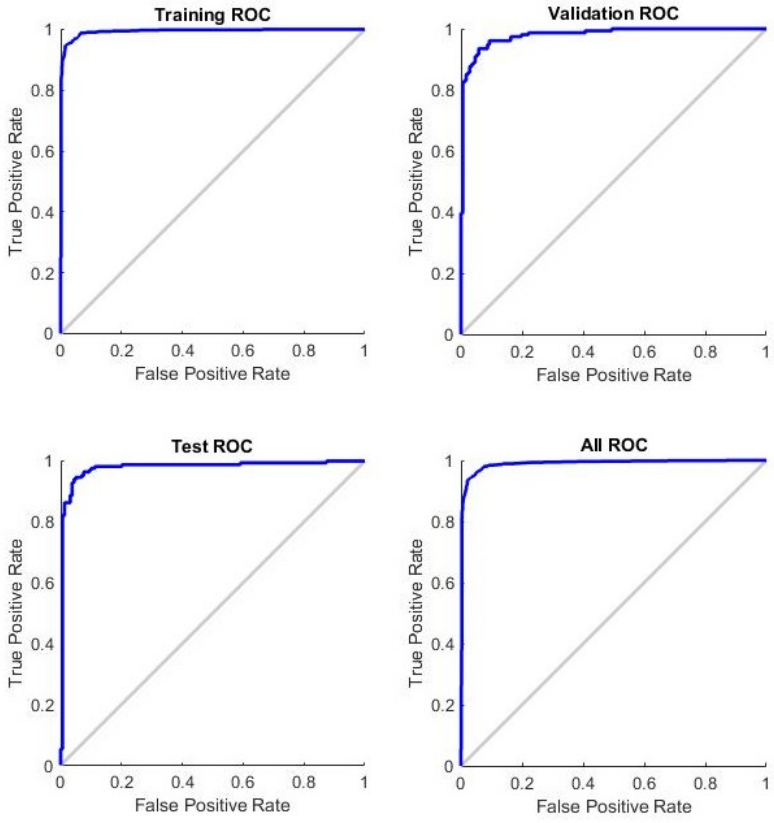


Figure 6. Receiver operating characteristic curve in the training phase, validation phase, and test phase, as well as combined curve, for the CASIA v2 dataset.

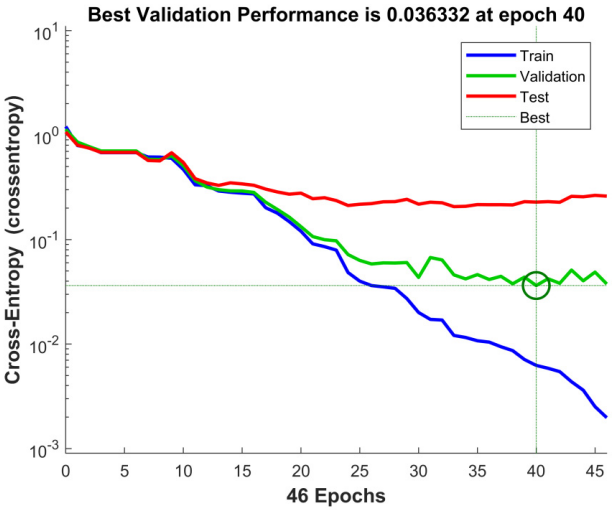


Figure 7. Performance plot showing the cross-entropy error at each epoch.

3.2. Results Explaining the Output of the Proposed Approach

SHAP decision graphs demonstrate how sophisticated models make their decisions (i.e., how models make predictions). Shapley values display each feature's actual contribution to the model's output. The SHAP methodology offers a variety of plots, including summary plots, dependency plots, force plots, and bee swarm plots, to help users visualize and analyze how model predictions based on input variables affect real-world phenomena. The summary plot shown in Figure 8a,b explains an instance of a normal image with a reconstruction error of 0.07. Figure 8c shows the decision plot. The *x*-axis is where the plot is centered using the explainer expected_value. Every SHAP value is relative to the estimated value of the model similarly to the way in which the linear model's effects are relative to the intercept. The features of the model are recorded on the *y*-axis. The attributes are by default listed in decreasing order of relevance. The significance is determined by the plotted explanations. The decision graph enables user-defined attribute ordering in addition to hierarchical cluster feature ordering and feature significance ordering.

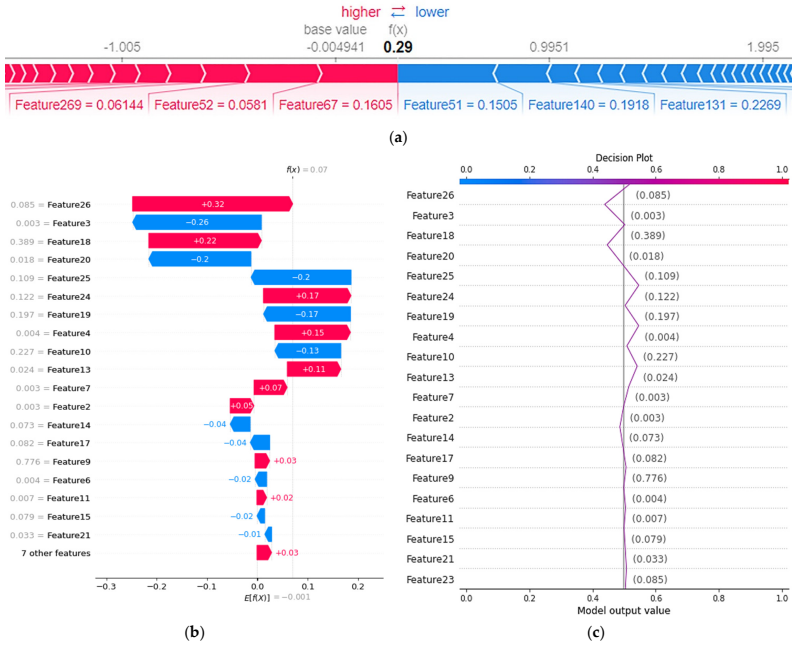


Figure 8. Interpreting a normal image with reconstruction error = 0.07 (**a**,**b**); the decision plot for the proposed approach (**c**).

3.3. Comparison with Recent Deep-Learning Methods

The performance of the proposed approach was compared with a few state-of-the-art methods published recently, and the results are shown in Table 3. In [35], LBP feature maps were extracted from ResNet's higher activation layer and fed to a common SVM classifier. In [36], a pretrained design called AlexNet was optimized to get satisfactory accuracy. In [20], a blocking method was proposed in which features from blocks of the image were extracted using rCNN and fed to an SVM classifier. In [16], a pretrained CNN was implemented to extract features from a high-pass filter set for calculating residual maps in the spatial rich model (SRM) used to collect the subtle artefacts of image manipulation operations. The method achieved an accuracy of 97.83% and 98.04% on CASIA v2 and CASIA v1 datasets, respectively. In [17], spatial- and frequency-domain features were extracted separately using CNN and combined to formulate the final feature vector to locate the forged region. The method reached an accuracy of 99.25% on CASIA v2. Among all these techniques, the proposed method achieved the highest accuracy of 99.31% and 99.52% on the CASIA v1 and CASIA v2 datasets, respectively. The proposed method also excelled in terms of training time and was faster as compared to the state-of-the-art methods. As most of these state-of-the-art methods train the entire network on forgery detection datasets such as rCNN, these methods took several days to train the network, whereas the proposed method implemented pretrained weights for training and took around 2 to 3 h for 12,000 images on an NVIDIA DGX GPU workstation.

Table 3. Comparison of the proposed approach with recent deep learning methods.

Recent Methods and Proposed Method	CASIA v1	CASIA v2
Deep textural RI-LBP maps [33]	99.10	99.30
rCNN [20]	98.04	98.02
CNN rely on feature fusion [16]	98.04	97.83
Dual-D-CNN [17]	-	99.25
Proposed method	**99.31**	**99.52**

3.4. Comparison with Traditional Machine Learning Methods

The proposed method was also compared with conventional machine learning methods based on manually crafted features, and the results were compared in terms of accuracy, as shown in Figure 9. The machine learning-based methods extract the features by image processing operations. With simple image processing operations, forged and authentic images cannot be classified because there is no visual distinction between the two categories. If we rely on image processing operations only, there has to be thresholds for every operation to give desirable results on every image. This is not possible because each image is different with respect to lighting, contrast, noise features, and variable image acquisition conditions. The automated deep learning-based approach proposed in this paper is helpful in such scenarios to extract deep and apt representations for the forgeries in the images. Furthermore, these representations were enhanced by the error level analysis of images, which captures the differences in JPEG compression levels of forged regions; the local binary pattern helps in identifying textural changes in neighboring pixels in the image region. In [8], the chroma channel of the image was used for extracting the local binary pattern of the image. Then, DCT was applied to the LBP code of individual blocks of the image. Furthermore, the standard deviation of all DCT coefficients belonging to each block was calculated and treated as features to be fed to the SVM classifier. In [6,7,9], the textural based features were used, and maximum accuracy was achieved by [6], i.e., 97.33% on CASIA v2. Figure 9 shows a graphical representation of the results obtained. The traditional methods had the disadvantage that they were computationally expensive, whereas deep learning-based techniques saved execution time.

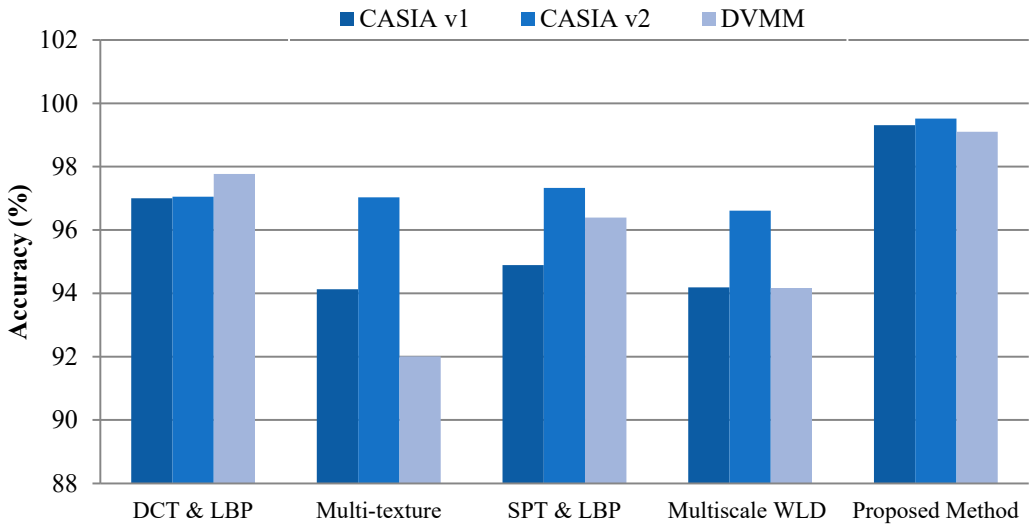


Figure 9. Comparison of the proposed approach with machine learning methods.

4. Conclusions

The approach for image manipulation detection suggested in the paper utilized the JPEG compression-based analysis of the image and textural-based features, and it was explained using the XAI approach. Both features can be combined in such a way as to use their benefits. In this approach, the RGB image is analyzed for JPEG compression levels by compressing the image at a certain level, and then the difference is calculated to find out the error. Furthermore, the obtained image difference is used to calculate the local binary code. The local binary code is further converted into 3D LBP feature maps using multidimensional scaling (MDS). The feature maps are served to the ResNet-50 v2 architecture. Here, the major reason for providing LBP feature maps to the neural network is that these networks are trained for specific image classification tasks; hence, they learn features that are general to image classification tasks. To make them more suitable for the detection of forgeries, LBP-based textural features are used. Residual networks are useful because support is provided by the addition of identity mapping to the residual output. The ResNet-50 v2 architecture is preferred in place of ResNet-50 v1 as version 2 uses pre-activation. This seems like a minor change, but it makes a big difference in the performance of the network. In post-activation, a single scale and bias are applied to the final output, whereas, in pre-activation, scaling and bias are applied to the input features by batch normalization. Batch normalization operation is applied to each layer, which then applies a unique scale and bias to earlier features.

The model-neutral kernel SHAP technique was chosen to interpret the results of the proposed technique. By estimating the Shapley values through kernelExplainer (explanation model) for each variable as a function of the anticipated reconstruction error of the proposed pretrained model, these variables can be identified. Instead of utilizing the raw error for each variable, Shapley values deliver the real influence of each variable, resulting in a significant reconstruction error. The proposed technique achieved an accuracy of 99.31%, 99.52%, 99.10%, and 98.05% on the CASIA v1, CASIA v2, DVMM, and IMD 2020 datasets, respectively. The capability of the pretrained CNN and rich textural features, which are scale- and direction-invariant, helps to escalate the detection accuracy of the proposed approach. In the future, the work can be extended to localize the forgery by providing batched input to the neural network.

Author Contributions: Conceptualization, S.W.; methodology, S.W. and K.K.; software, S.W.; validation, S.A. and H.K.; formal analysis, S.A.; investigation, S.W.; resources, H.K.; writing—original draft preparation, S.W. and S.A.; writing—review and editing, S.W., K.K., S.A. and H.K.; visualization, S.W. and S.A.; supervision, K.K. and H.K.; project administration, H.K.; funding acquisition, H.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A1B04032598).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this paper are publicly available, and their links are provided in the reference section.

Acknowledgments: We thank the anonymous reviewers for their valuable suggestions that improved the quality of this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Asghar, K.; Hussain, M. Copy-Move and Splicing Image Forgery Detection and Localization Techniques: A Review. *Aust. J. Forensic Sci.* **2017**, *49*, 281–307. [CrossRef]
2. Walia, S.; Kumar, K. Digital Image Forgery Detection: A Systematic Scrutiny. *Aust. J. Forensic Sci.* **2019**, *51*, 488–526. [CrossRef]
3. Nixon, M.S.; Aguado, A.S. *Feature Extraction and Image Processing for Computer Vision*, 4th ed.; Academic Press, Inc.: Cambridge, MA, USA, 2020; ISBN 978-0-12-814976-8.
4. Hemanth, D.J.; Vieira Estrela, V. *Deep Learning for Image Processing Applications*, 31st ed.; Advances in Parallel Computing; IOS Press: Amsterdam, The Netherlands, 2017.
5. Al-hammadi, M.H.; Muhammad, G. Curvelet Transform and Local Texture Based Image Forgery Detection. In *Advances in Visual Computing, ISVC 2013*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2013; pp. 503–512.
6. Muhammad, G.; Al-Hammadi, M.H.; Hussain, M.; Bebis, G. Image Forgery Detection Using Steerable Pyramid Transform and Local Binary Pattern. *Mach. Vis. Appl.* **2014**, *25*, 985–995. [CrossRef]
7. Hussain, M.; Qasem, S.; Bebis, G.; Muhammad, G.; Aboalsamh, H.; Mathkour, H. Evaluation of Image Forgery Detection Using Multi-Scale Weber Local Descriptors. *Int. J. Artif. Intell. Tools* **2015**, *24*, 416–424. [CrossRef]
8. Alahmadi, A.; Hussain, M.; Aboalsamh, H.; Muhammad, G.; Bebis, G.; Mathkour, H. Passive Detection of Image Forgery Using DCT and Local Binary Pattern. *Signal Image Video Process.* **2017**, *11*, 81–88. [CrossRef]
9. Vidyadharan, D.S.; Thampi, S.M. Digital Image Forgery Detection Using Compact Multi-Texture Representation. *J. Intell. Fuzzy Syst.* **2017**, *32*, 3177–3188. [CrossRef]
10. Walia, S.; Kumar, K.; Kumar, M.; Gao, X.Z. Fusion of Handcrafted and Deep Features for Forgery Detection in Digital Images. *IEEE Access* **2021**, *9*, 99742–99755. [CrossRef]
11. Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C.; Xu, W. CNN-RNN: A Unified Framework for Multi-Label Image Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2285–2294.
12. Lee, H.; Kwon, H. Going Deeper With Contextual CNN for Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2017**, *26*, 4843–4855. [CrossRef] [PubMed]
13. Han, D.; Liu, Q.; Fan, W. A New Image Classification Method Using CNN Transfer Learning and Web Data Augmentation. *Expert Syst. Appl.* **2018**, *95*, 43–56. [CrossRef]
14. Sun, Y.; Xue, B.; Zhang, M.; Yen, G.G.; Lv, J. Automatically Designing CNN Architectures Using the Genetic Algorithm for Image Classification. *IEEE Trans. Cybern.* **2020**, *50*, 3840–3854. [CrossRef] [PubMed]
15. Parashar, A.; Upadhyay, A.K.; Gupta, K. A Novel Machine Learning Approach for Forgery Detection and Verification in Digital Image. *ECS Trans.* **2022**, *107*, 11791–11798. [CrossRef]
16. Rao, Y.; Ni, J. A Deep Learning Approach to Detection of Splicing and Copy-Move Forgeries in Images. In Proceedings of the 8th IEEE International Workshop on Information Forensics and Security, WIFS 2016, Abu Dhabi, United Arab Emirates, 4–7 December 2016.
17. Shi, Z.; Shen, X.; Kang, H.; Lv, Y. Image Manipulation Detection and Localization Based on the Dual-Domain Convolutional Neural Networks. *IEEE Access* **2018**, *6*, 76437–76453. [CrossRef]
18. Qazi, E.U.H.; Zia, T.; Almorjan, A. Deep Learning-Based Digital Image Forgery Detection System. *Appl. Sci.* **2022**, *12*, 2851. [CrossRef]
19. Ali, S.S.; Ganapathi, I.I.; Vu, N.S.; Ali, S.D.; Saxena, N.; Werghi, N. Image Forgery Detection Using Deep Learning by Recompressing Images. *Electronics* **2022**, *11*, 403. [CrossRef]
20. Zhou, J.; Ni, J.; Rao, Y. Block-Based Convolutional Neural Network for Image Forgery Detection. In *Digital Forensics and Watermarking IWDW2017*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2017; pp. 65–76.
21. Farooq, S.; Yousaf, M.H.; Hussain, F. A Generic Passive Image Forgery Detection Scheme Using Local Binary Pattern with Rich Models. *Comput. Electr. Eng.* **2017**, *62*, 459–472. [CrossRef]

22. Muhammad, G.; Al-hammadi, M.H.; Hussain, M.; Mirza, A.M.; Bebis, G. Copy Move Image Forgery Detection Method Using Steerable Pyramid Transform and Texture Descriptor. In Proceedings of the Eurocon 2013, Zagreb, Croatia, 1–4 July 2013; pp. 1586–1592.
23. Moore, J.D.; Swartout, W.R. *Explanation in Expert Systems: A Survey*; Technical Report; University of Southern California, Information Sciences Institute: Marina del Rey, CA, USA, 1988.
24. Van Lent, M.; Fisher, W.; Mancuso, M. An Explainable Artificial Intelligence System for Small-Unit Tactical Behavior. In Proceedings of the 16th Conference on Innovative Applications of Artificial Intelligence (IAAI'04), San Jose, CA, USA, 27–29 July 2004; pp. 900–907.
25. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* **2018**, *51*, 1–42. [CrossRef]
26. Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Long Beach, CA, USA, 4–9 December 2017; pp. 4768–4777.
27. Takeishi, N.; Lee, S. Shapley Values of Reconstruction Errors of PCA for Explaining Anomaly Detection. In Proceedings of the International Conference on Data Mining Workshops (ICDMW), Beijing, China, 8 September 2019.
28. Sudiatmika, I.B.K.; Rahman, F.; Trisno, T.; Suyoto, S. Image Forgery Detection Using Error Level Analysis and Deep Learning. *TELKOMNIKA Telecommun. Comput. Electron. Control* **2019**, *17*, 653. [CrossRef]
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
30. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
31. Shrikumar, A.; Greenside, P.; Shcherbina, A.; Kundaje, A. Not Just a Black Box: Learning Important Features through Propagating Activation Differences. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017; Volume 70, pp. 3145–3153.
32. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should i Trust You?" Explaining the Predictions of Any Classifier. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
33. Dong, J.; Wang, W.; Tan, T. CASIA image tampering detection evaluation database. In Proceedings of the 2013 IEEE China Summit and International Conference on Signal and Information Processing, Beijing, China, 6–10 July 2013; pp. 422–426.
34. Novozámský, A.; Mahdian, B.; Saic, S. IMD2020: A Large-Scale Annotated Dataset Tailored for Detecting Manipulated Images. In Proceedings of the 2020 IEEE Winter Applications of Computer Vision Workshops (WACVW), Snowmass, CO, USA, 1–5 March 2020; pp. 71–80. [CrossRef]
35. Remya Revi, K.; Wilscy, M. Image Forgery Detection Using Deep Textural Features from Local Binary Pattern Map. *J. Intell. Fuzzy Syst.* **2020**, *38*, 6391–6401. [CrossRef]
36. Samir, S.; Emary, E.; El-Sayed, K.; Onsi, H. Optimization of a Pre-Trained AlexNet Model for Detecting and Localizing Image Forgeries. *Information* **2020**, *11*, 275. [CrossRef]