RESEARCH ARTICLE

SPECIAL ISSUE

International Research Journal on Advanced Science Hub
2582-4376
Vol. 05, Issue 05S May
www.rspsciencehub.com

RSP Science Hub

http://dx.doi.org/10.47392/irjash.2023.S003

International Conference on intelligent COMPUting TEchnologies and Research (i-COMPUTER) 2023

Analysis of Supervised and Unsupervised Deep Learning Approaches for Identifying and Localizing Image Forgeries

Krishnan S [1], Pranay Varma[1], Aravind J [1], Indra Gandhi K[2]

[1]Department of Information Science and Technology, College of Engineering, Guindy, Anna University, Chennai, Tamil Nadu, India

[2]Assistant Professor (Sl.Gr.), Department of Information Science and Technology, College of Engineering, Guindy, Anna University, Chennai, Tamil Nadu, India

Email: skrishnan2001@gmail.com

Article History

Received: 16 February 2023
Accepted: 8 March 2023

Keywords:

Deep Learning;
Image Processing;
CNN;
Unsupervised Self-Consistency Learning;
Forgery Detection;
Forgery Localization;
Splicing;
Copy-Move

Abstract

The field of image forensics has become important in recent years as the use of digital images continues to grow. With the rise of sophisticated image editing software, it has become increasingly difficult to detect whether an image has been tampered with or not. Moreover, social media platforms have made the distribution of forged images to the general public a simple task. It is hence very important to develop automated methods that can detect such forgeries. In this study, we detect and localize splicing and copy-move image forgeries in images by using two different deep-learning techniques - Convolutional Neural Networks (CNN), which is a supervised approach and Self-Consistency Learning, an unsupervised approach. By comparing and contrasting the performance of these approaches, the research aims to gain a better understanding of how to effectively detect and locate image forgeries using deep learning. Ultimately, this research will contribute to the development of more reliable and accurate image forensic techniques, which will be of great benefit in various fields such as criminal investigations, digital media, and photojournalism.

1. Introduction

Malicious image manipulation, previously restricted to dictators and spy agencies, is now available to legions of common Internet trolls and Facebook conmen. It is now possible to create realistic image composites, fill in large image regions, generate plausible video from speech, and so on with only basic editing skills. One might have expected that these new methods for creating synthetic visual content would be accompanied by equally powerful techniques for detecting fakes, but this has not been the case thus far. Thus detecting such forgeries becomes very important to stop the spread of false information.

Three of the most common image manipulation techniques are

- Splicing: In splicing a region from an authentic image is copied into a different image.
- Image Inpainting: In image inpainting, an image region is removed and the removed part is then filled in to complete the image.
- Copy-Move: When a particular section of an image is duplicated and then pasted onto a different location within the same image.

The goal of this research is to detect and localize Splicing and Copy-Move forgeries in images using both supervised and unsupervised deep learning techniques. To achieve this, two deep-learning approaches, CNN and unsupervised self-consistency learning have been implemented on various image forensics datasets like CASIA2 (Goldbloom and Ben), Dresden (Dresden image dataset) and in the In-the-Wild Image Splice Dataset dataset (website) and the performance of image forgery detection for each approach is analysed based on the test sample difficulty.

CNN Approach: Various computer vision and deep-learning approaches have been suggested to detect image forgeries to date. Specifically, a few CNN-based architectures have managed to predict images with an accuracy of close to 98%. However, the tampering done in these images can also be easily recognized by humans. In this research, we have developed a CNN network that attempts to detect forgeries on more difficult samples and analyse its performance on such examples. This is a supervised approach. It is robust in detecting both copy-move and splice forgeries.

Unsupervised Self-Consistency Learning: Standard supervised learning approaches, which have proven extremely effective for a wide range of detection problems, are unsuitable for image forensics. This is because the space of manipulated images is vast and diverse, making it unlikely that we will ever have enough manipulated training data for a supervised method to fully succeed. To address this, we are employing an unsupervised methodology known as self-consistency learning, which uses EXIF (Exchangeable Image File Format) metadata to determine whether or not an image has been tampered with. EXIF tags are camera specifications that are digitally engraved into an image file during the capture process. Thus, given two photographs, we can deduce from their EXIF metadata that there are several differences in the two imaging pipelines. This approach is well suited to identifying spliced images as the metadata for patches from different images is very likely to be different. However, copy-move forgeries cannot be identified as the imaging pipeline for the portion that has been copied and reproduced would not be different. When compared to CNN, more complex forms of splicing forgeries can be identified.

The objectives of this research are to

- Detect and localize splicing and copy-move image tamperings.
- Approximately recover the original image that was tampered with.
- Analyse the performance of different deep-learning approaches on different datasets.
- Develop a web application using Streamlit, which allows the users to upload test images and find the region of tampering if any.

2. Materials and Methodology

This section deals with the existing work that has been carried out to identify copy-move and spliced image tampering and the methods used to implement them. It gives an overview of the various supervised and unsupervised approaches. Under the supervised approaches, different variations of CNN are seen.

In (Hosny et al.), Khalid M. Hosny et al. present a model with a three-stage approach for detecting copy-move forgeries and is efficient in terms of computation time and memory usage, detecting copy-move forgeries with high accuracy. However, challenging test samples bring about mixed results. (Mallick et al.) focuses on detecting copy move and splicing image forgery using a CNN with three different models, and the steps involved are preprocessing, error level analysis and the CNN. Challenging test data again leads to inaccurate results, and the approach isn't generalizable

In (Rao and Ni), CNN automatically learns hierarchical representations from input RGB colour images. The approach outperforms many state-of-the-art models, in terms of speed and accuracy, however, the performance of the model deteriorates for more challenging image forgery datasets. The main advantage of (Takeda et al.) was that the tampered region was found accurate to a great extent, and the F-measure of the method was approximately 2.3 times that of the MDBD(Multiple Detection using Block noise and Double JPEG) methods. (Abdalmageed, Wu, and Natarajan) presents a method called "Mantra-Net" for detecting and localizing image forgeries. The method is based on a neural network that is trained to identify anomalous features in images that are indicative of tampering. The approach is able to detect and localize multiple types of manipulation in images.

The main contribution of (Qian et al.) is the development of a GAN-based approach to generate images that are resistant to tampering detection, which can help in improving the security of digital images. Qichao Ying et.al produces immunized images that are similar to the original images but are able to evade tampering detection methods. Mustafa Kaya, et al. use a pre-trained model called 'ResNet50' in (Sani, Kaya, and Karakus) to detect copy-move forgeries in digital images using CNN. The dataset used for analysis was CoMoFOD which contains both original and forged images. Post-processing attacks like the addition of noise, rotation, scaling, etc were used to implement the model. For the CoMOFoD dataset, the proposed solution achieved an accuracy of 100.00%, precision of 100.00%, recall of 100.00%, and f1-score of 100.00%.

In (Moon, W Park, and Eom), Y. H. Moon et al. use Singular Value Decomposition(SVD) for the purpose of estimation of prediction residue between acquired and interpolated images. The approach is based on colour filter array (CFA) pattern artifacts. The prediction residue of the proposed algorithm was more efficient for localizing forged regions.

The supervised approaches all had a common denominator of struggling with challenging datasets, which is why unsupervised approaches were looked at. In (Liu et al.) by Minyoung Huh et al., the algorithm employs the automatically recorded photo EXIF metadata as a supervisory signal for training a model to determine whether an image is self-consistent, or whether its content could have been produced by a single imaging pipeline. The concept explored is that patches from different images would have differing exif metadata and this could be used to identify splicing. Bhavsar A. Kumar et al. (Bhavsar, Kumar, and Verma) leverage the principle of domain adaptation, wherein the model is trained on a source dataset that is similar to a different target dataset in terms of features. A technique called Maximum Mean Discrepancy (MMD) is used to align the source and target domains, which helps the model to better generalize to the target domain. M. Baviskar et al. (Baviskar, Rathod, and Lohokare) conducted a comparative analysis of image forgery methods and proposed a new eight-layer CNN-based model trained on an ELA dataset for identifying forged images. S. I. Lee et al. (Lee, Park, and Eom) proposed a rotation-invariant feature based on wavelet coefficients and used a VGG16 network with a correlation module and simplified mask decoder module to detect copy-move forgery, outperforming existing methods on four test datasets.

The proposed system's high-level architecture is depicted in Figure 1. The user interface has been developed with Streamlit, which is used to accept a test image from the user as well as the deep learning model of choice. The sequence of actions will be followed depending on the model selected, as discussed further in sections 3.1 and 3.2 for the CNN-based approach and Unsupervised Self Consistency respectively.

Once the model has been executed, the web application will display the original image uploaded by the user as well as the parts of the image that have been tampered with (tampering localization). Two different datasets of varying difficulty have been used to test both these approaches. The first dataset used is the 'CASIA 2 Image Forensics dataset, which has 12,622 images, where the ratio of authentic to tampered images is 60:40. The tampering in this dataset is less challenging and can be recognized by humans. The second dataset used is the 'Labels in the Wild' dataset which contains 201 tampered images and the masks of each tampered image. This dataset is relatively much more challenging than the CASIA 2 dataset and the tampering cannot be easily recognized by humans. Both approaches' classification accuracy is evaluated using these two datasets.

2.1. CNN Approach

The overall architecture of the CNN approach is depicted in Figure 2. The images in the CASIA dataset's authentic and tampered classes are first preprocessed using techniques such as augmentation. Patches of 128 x 128 x 3 pixels are extracted from each image in both image classes.

These patches are then provided as input to the CNN model, producing an N-featured representation vector. Mean/Max fusion is then used to fuse this vector to a single feature vector. This fused vector is then fed into the SVM classifier, which determines whether or not the given image has been tampered with. An image manipulation trace feature extractor and a local anomaly detection network have been used to determine the region of tampering, which is further discussed in section 2.1.2.

2.1.1. CNN Model Architecture

The CNN model consists of 9 convolution and 2 pooling layers. Depending on the layer involved,

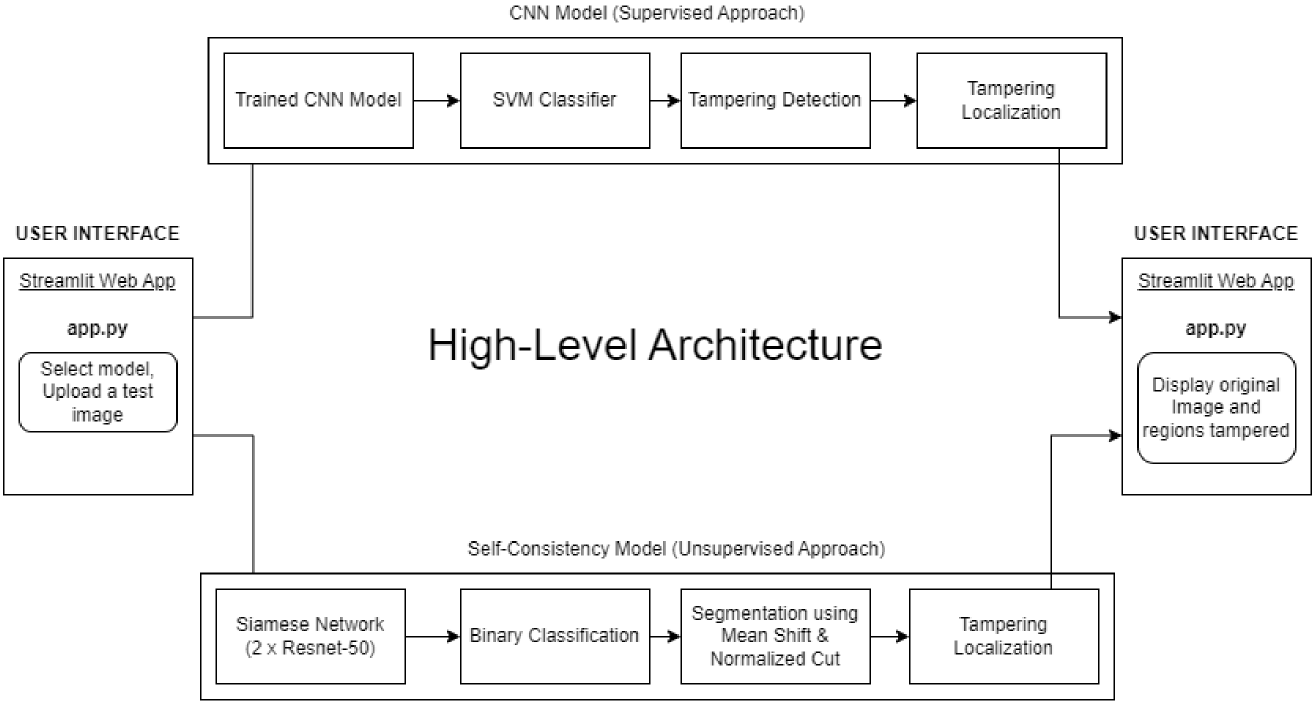

FIGURE 1. **High-Level Architecture**

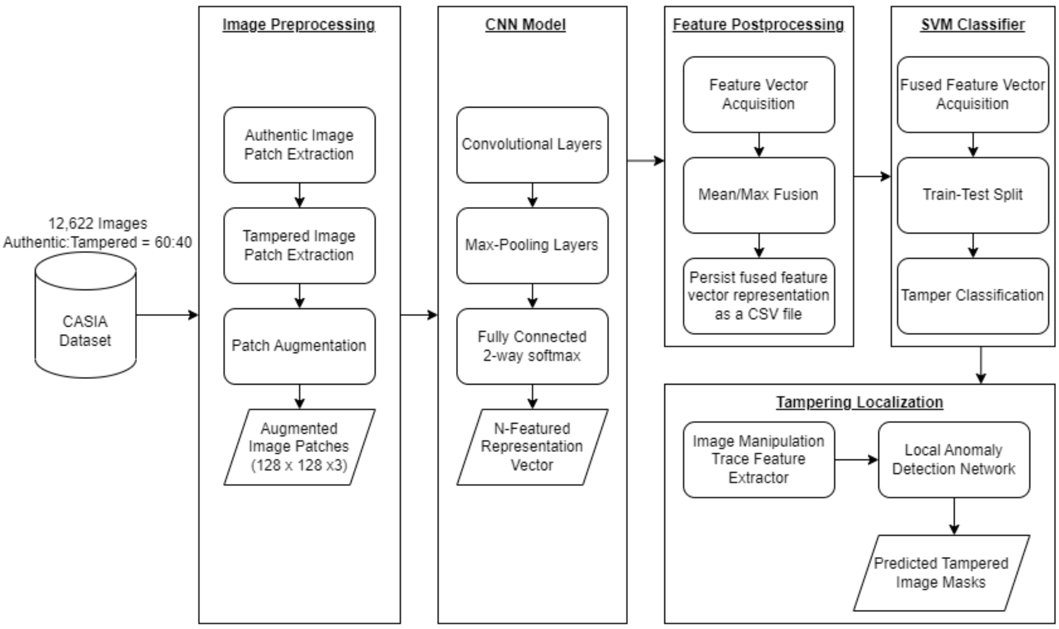

FIGURE 2. **CNN Approach Architecture**

the kernel size is fixed as either 3x3, 2x2 or 5x5 with a stride of 1 or 2. The convolution layers extract features from the input matrices, while the pooling layers perform down-sampling or dimensionality reduction of the features. The ReLu activation function is used by each of the convolution layers. The model is trained for 250 epochs. Figure 3 depicts the CNN architecture.

2.1.2. Mantra-Net Architecture for Forgery Localization

MantraNet is a machine learning-based image forgery detection method that uses deep learning techniques, specifically a CNN, to analyze images and identify whether they have been manipulated or altered in any way. This approach of image forgery

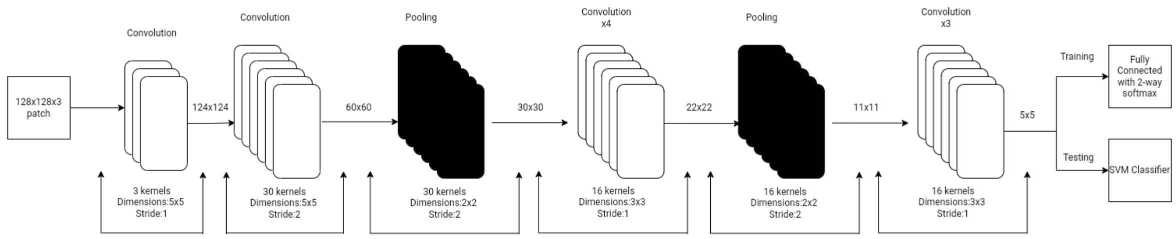

FIGURE 3. **CNN Architecture**

localization has been discussed in (Abdalmageed, Wu, and Natarajan) by W. AbdAlmageed, et al. The paper describes how the network can be trained on a large dataset of images, and then used to detect forgeries in new images by analyzing the output of various layers of the network. The authors of the paper also demonstrate how their method outperforms other state-of-the-art methods for localizing image forgeries, in terms of both accuracy and computational efficiency.

A testing image is used as the input, and a pre-trained “ManTraNet” model is used to predict a pixel-level forgery likelihood map as the output. It is made up of two smaller networks:

• The Image Manipulation Trace Feature Extractor is a feature extraction network for the purpose of classifying images that have been altered, and it encodes the altered image in a patch into a feature vector with a fixed dimension.

• The Local Anomaly Detection Network is a network that was created with the understanding that in order to effectively detect various types of forgeries, we must evaluate our extracted characteristics more and more locally.

2.1.3. Algorithms

Algorithm 1. Image Patch Extractor

Input: input_path, output_path, patches_per_image, no_of_rotations, stride

Output: Rotated image patches

1. START
2. FOR each image in Tampered Images and Authentic Images LOOP
3. Apply patch-sized sliding window of stride 128
4. IF image belongs to Tampered Images
5. Determine tampered patches where num_zeros <= 0 99 * (num_zeros + num_ones
6. END IF
7. Augment the patches by rotating them by 0, 90, 180 & 270 degrees
8. GOTO 2
9. END LOOP
10. Store the extracted patches in separate directories for authentic and tampered classes
11. STOP

Algorithm 2. Feature Extraction and Forgery Classification

Input: 128x128x3 image patches

Output: 1 or 0 (Binary Classification)

1. START
2. The patches are fed into the CNN model, which extracts a 400-D feature representation for each patch.
3. The (n x 400-D) feature representations for an image must be fused into a single feature vector.
4. These features are then passed to a fully-connected layer with a 2-way softmax classifier in the training phase and the SVM model in the testing phase.
5. The SVM model returns 1 if the image is tampered, and 0 otherwise
6. STOP

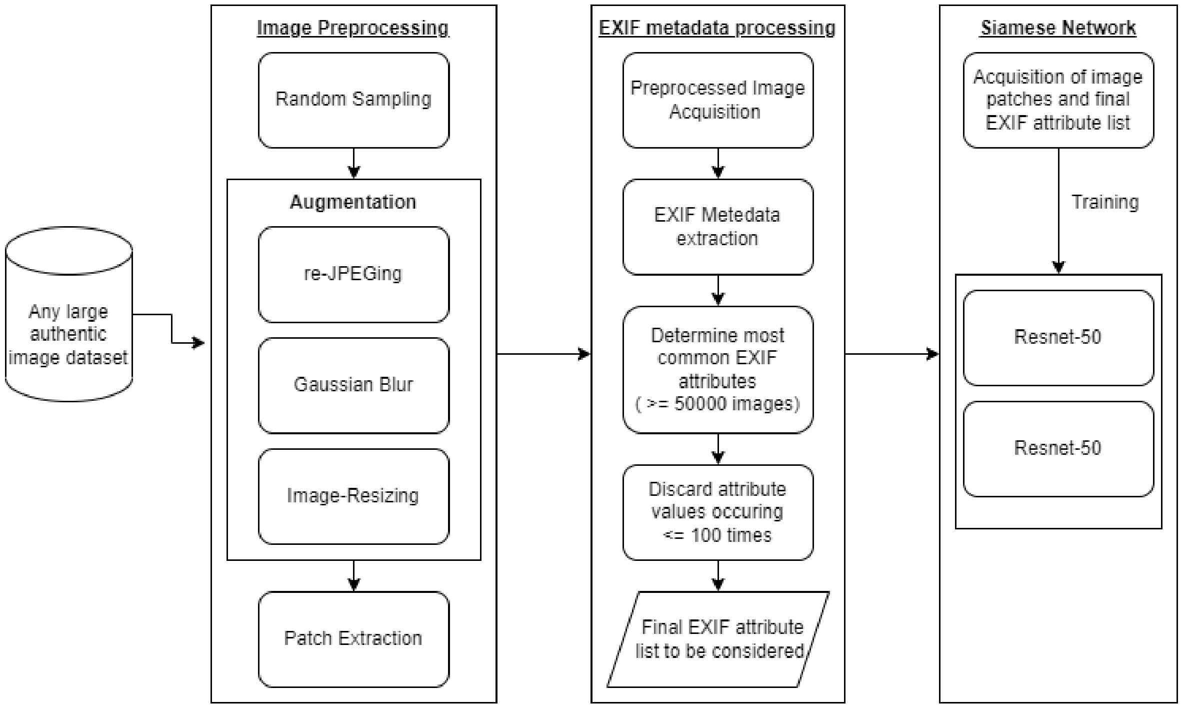

FIGURE 4. **Siamese Network training architecture**

2.2. Unsupervised Self-Consistency Learning

A Siamese network utilizing two Resnet-50 models has been employed to estimate the likelihood that two 128x128 image patches have the same EXIF metadata attribute values. Architecture for training the Siamese Network in the unsupervised self-consistency approach is depicted in Figure 4.

Any large authentic image dataset can be used to train the network. The Flickr dataset, which contains over 400000 authentic images, was used in this study. The input dataset is first preprocessed using random sampling and image augmentation techniques to produce a subset of well-distributed, augmented images. The EXIF metadata from these images is extracted, and the most frequently occurring EXIF attributes are determined. The Resnet-50 model is trained using the authentic image patches and the final EXIF attribute list. The model calculates the percentage of consistency between two image patches based on their EXIF values. Figure 5 depicts the architecture for determining the tampered region in an image using self-consistency learning.

The input image is first preprocessed and patches of size 128 x 128 are extracted. These extracted image patches as passed as input to the pre-trained Siamese Network, which returns a consistency map list after performing a pairwise consistency check of all the patches of the image. The relative consistency percentage values to the first patch are contained in each element of the consistency map list. Finally, using the obtained consistency map list, segmentation methods such as Mean Shift and Normalized Cut are applied to the input image to determine the exact region of tampering.

2.2.1. Siamese Network

Internally, the Siamese Network employs the traditional ResNet 50 neural network. It is a PyTorch predefined model that can be trained on the input dataset to predict the results. The Siamese network is used to predict the likelihood that two EXIF fields in a 128 x 128 image patch have the same value attribute. It makes use of pre-trained shared ResNet50 sub-networks. Each sub-network generates a 4096-dimension feature vector. These vectors are concatenated and then passed through a four-layer perceptron with 4096, 2048, and 1024 units, followed by the final output layer. The network forecasts the likelihood of the images having the same value for each of the n metadata attributes.

2.2.2. Algorithms

Algorithm 3. Input preprocessing and consistency map extraction

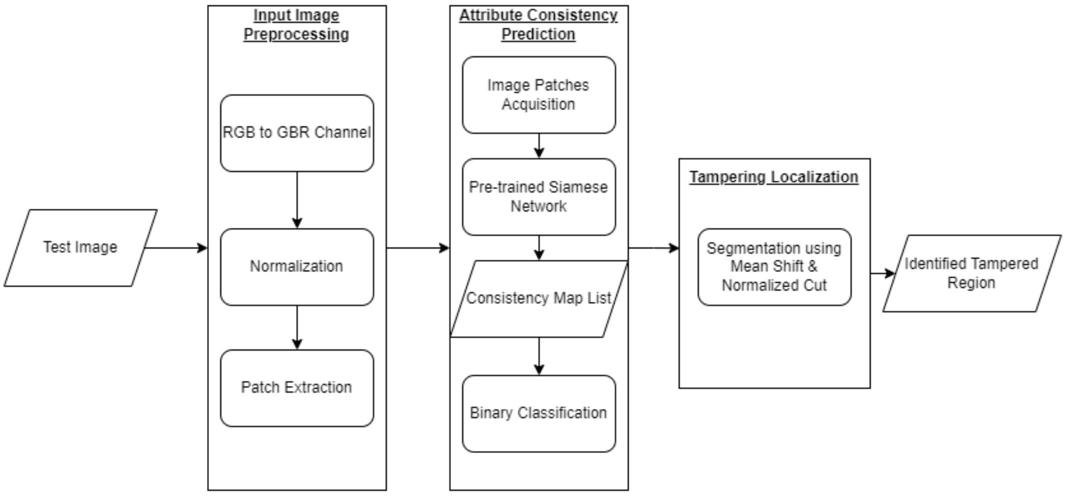

FIGURE 5. **Self-Consistency Learning Localization Architecture**

Input: Test Images
Output: Consistency Map List

1. START
2. Load the images to be tested
3. Convert RGB to GBR colour scheme
4. Unsqueeze the image's dimensions from (w, h, 3 to (1, 3, w, h
5. Calculate stride size based on the image dimensions
6. Apply a patch-sized sliding window to extract patches of size 128 x 128
7. Compare the obtained patches pairwise and get the probability score of consistency
8. Obtain the consistency map list
9. STOP

Algorithm 4. Image Segmentation
Input: Image, Image_Patches
Output: Segmented Images using Mean Shift and Normalized Cut

1. START
2. Compute the consistency map of a patch with respect to other patches considering each metadata attribute independently.
3. The resultant consistency map is used to plot the mean shift, taking the top 10 percentile of the nearest points into consideration for a given point.
4. The normalized cut is obtained from the consistency maps using the spectral clustering method
5. If most of the image is a high probability, flip it
6. The resultant images for mean shift and normalized cut are resized, showing the segments clearly.
7. STOP

3. RESULTS AND DISCUSSION

This section contains the results obtained and the performance analysis of this research: "Analysis of Supervised and Unsupervised Deep Learning Approaches for Identifying and Localizing Image Forgeries".

3.1. CNN Training Accuracy

The training accuracy vs epoch graph is depicted in Figure 6 and the training loss vs epoch graph is depicted in Figure 7. These results were obtained when the CNN is trained with the CASIA2 dataset.

From the graph, we can infer that as the number of epochs increases, the training accuracy also increases and reaches saturation after which the training accuracy doesn't change much. So the number of epochs is stopped at 250 to prevent overfitting.

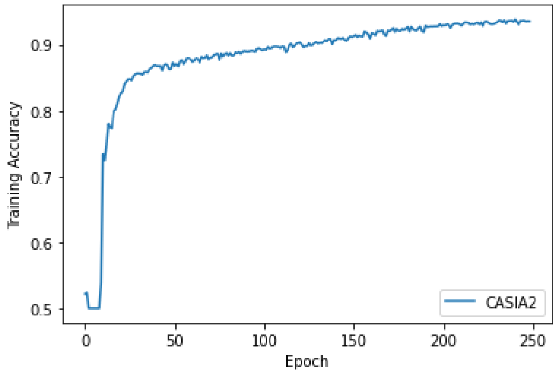

FIGURE 6. **Training Accuracy vs Epoch**

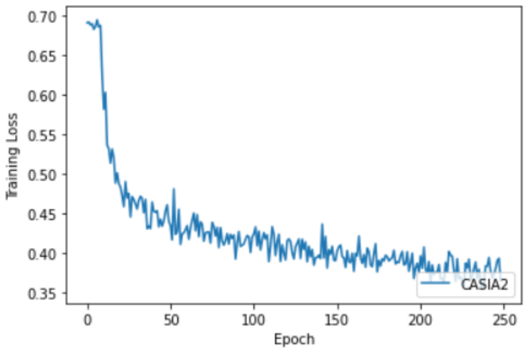

FIGURE 7. **Training Loss vs Epoch**

3.2. SVM Performance

A confusion matrix is a table that provides a concise representation of how well a classifier is performing in terms of correct and incorrect predictions. It serves as a useful tool for assessing the performance of a classification model, allowing for the calculation of performance metrics such as accuracy, precision, recall, and F1-score. Figure 8 shows the confusion matrix when tested with the CASIA dataset.

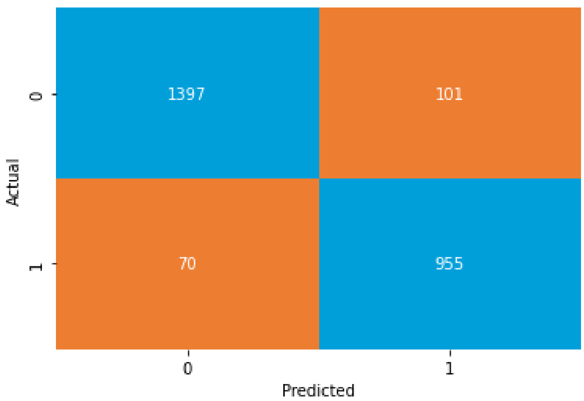

FIGURE 8. **Confusion Matrix**

TN: 1397, FP: 101, FN: 70, TP: 955

TP (True Positive) - Image is tampered and predicted as tampered

FP (False Positive) - Image is authentic but predicted as tampered

TN (True Negative) - Image is authentic and predicted as authentic

FN (False Negative) - Image is tampered with but predicted as authentic

SVM's performance with respect to the various metrics like F1 Score, Specificity, Recall and Precision has been summarized in Figure 9 below.

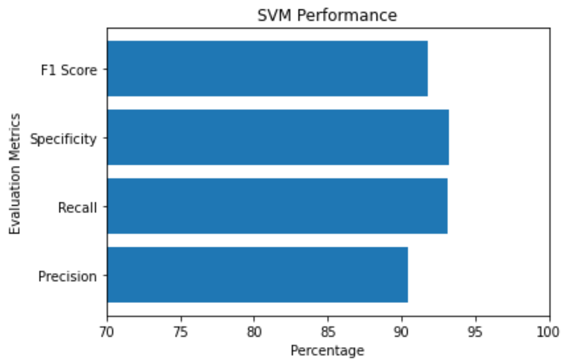

FIGURE 9. **SVM Performance Metrics**

It performs well, with a score greater than 90% for each of the performance metrics considered. The results are thus found to be satisfactory. Table 1 shows the Precision, Recall, Specificity and F1 Score of the SVM classifier.

TABLE 1. **Performance analysis of SVM classifier**

Precision	Recall	Specificity	F1 Score
90.43 %	93.1 %	93.25 %	91.74 %

3.3. CNN Outputs

Figure 10 shows the outputs of the CNN approach when applied to copy-move tampered images. The images seen below have been taken from the CASIA dataset.

The original image has been followed by its predicted forgery mask and the suspicious region. The final image shows the ground truth of the tampered image. Figure 11 shows the outputs of the CNN approach when applied to spliced images. The images below for testing of spliced images are taken from CASIA and 'Labels in the Wild' datasets.

The original image has been followed by its predicted forgery mask and the suspicious region. The final image shows the ground truth of the tampered image.

FIGURE 10. **CNN outputs for localization (copy-move images)**

FIGURE 11. **CNN outputs for localization (spliced images)**

3.4. Self-Consistency Learning Outputs

Figure 12 shows the outputs of the Self-consistency learning approach. The below images are from Labels in the Wild which contains forgeries which are more difficult to predict compared to CASIA.

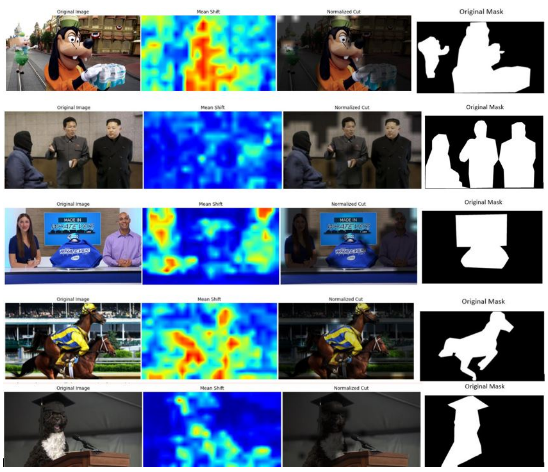

FIGURE 12. **Self-consistency learning outputs**

The original image has been followed by its corresponding Mean shift and Normalized cut-based segmentation. The final image shows the ground truth of the tampered image.

3.5. User Interface

The UI of this work is made using Streamlit, an open-source library available in python. Figure 13 is a screenshot of the web application developed. The web app prompts the user to upload an image for testing. On image upload, the web app runs the deep learning model in the background to localize the exact region of tampering, if any in the uploaded image.

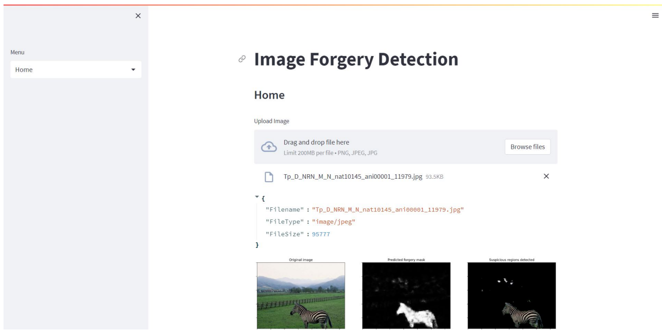

FIGURE 13. **Tampering Localization in UI**

3.6. Performance Analysis

The CNN model performs well with obvious instances of tampering, in terms of recognizing whether an image has been tampered with or not. Instances where the tampering is less obvious produce mixed results. The localization of the tampered region (using MantraNet) produces satisfactory results when the training and testing data are from the same source. As it is a supervised CNN-based approach, the accuracy suffers when the domains for testing and training differ. Self-consistency is more effective than CNN when it comes to identifying more subtle forms of tampering, but carries the downside of being unable to detect copy-move forgeries as the EXIF attributes of forged regions would match those of the original image. Factors like light exposure and the size of the region of splicing were found to impact the performance of the tamper localization. Over and under-exposed images often produced varying results. On testing with authentic images, the model sometimes flags tiny portions of the image that appear to have a different level of exposure as tampered and hence these portions appear as separate segments from the rest of the image.

4. Conclusion

This research work uses two deep learning approaches to detect and localise splicing and copy-move forgeries in images: the CNN approach and Unsupervised Self-Consistency Learning. It was developed using Python and its libraries like PyTorch, Pandas, and Matplotlib. A user interface was also developed using Streamlit, which allows a user to upload a test image and get the exact region of tampering in the image. The CNN approach achieved a precision of 90.43%, Recall of 93.1%, Specificity of 93.25% and F1 Score of 91.74% for the classification of forged or authentic images based on the CASIA dataset. The CNN approach was found to be more robust in detecting both splicing and copy-move image forgeries, whereas the self-consistency learning approach could only detect splicing image forgeries. However, when tested with the 'Label in the Wild' dataset, it was discovered that the CNN approach did not perform well when the tamperings in the images were much more complex and difficult to be identified by the human eye. The Self-Consistency Learning approach, on the other hand, could detect much more complex splicing tamperings in images, but the total time taken to localise the region of forgery is significantly longer when compared to the CNN approach.

The future work includes making the self-consistency approach more efficient so that it takes lesser time to localise the region of tampering. Future work will also include expanding the system to be able to reconstruct/recover the original image given a tampered image.

References

Abdalmageed, W, Y Wu, and P Natarajan. "Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features". *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019): 9535–9544.

Baviskar, Mohit, Sheetal Rathod, and Jay Lohokare. "A Comparative Analysis of Image Forgery Detection Techniques". *2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*. IEEE, 2022. 1–6.

Bhavsar, A, A Kumar, and R Verma. "Forgery classification via unsupervised domain adaptation". *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops* (2020): 63–70.

Hosny, Khalid M, et al. "An Efficient CNN Model to Detect Copy-Move Image Forgery". *IEEE Access* 10 (2022): 48622–48632. 10.1109/ACCESS.2022.3172273.

Lee, Sang In, Jun Young Park, and Il Kyu Eom. "CNN-Based Copy-Move Forgery Detection Using Rotation-Invariant Wavelet Feature". *IEEE Access* 10 (2022): 106217–106229. 10.1109/ACCESS.2022.3212069.

Liu, A, et al. "Fighting fake news: Image splice detection via learned self-consistency". *Proceedings of the European conference on computer vision (ECCV)* (2018): 101–117.

Mallick, Devjani, et al. "Copy Move and Splicing Image Forgery Detection using CNN". *ITM Web of Conferences* 44 (2022): 03052–03052. 10.1051/itmconf/20224403052.

Moon, Y H, C W Park, and I K Eom. "Image tampering localization using demosaicing patterns and singular value based prediction residue". *Image tampering localization using demosaicing patterns and singular value based prediction residue* (2021): 91921–91933. 10.1109/ACCESS.2021.3091161.

Qian, Z, et al. "From image to imuge: Immunized image generation". *Proceedings of the 29th ACM international conference on Multimedia* (2021): 3565–3573. 10.1145/3474085.3475520.

Rao, Yuan and Jiangqun 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)* (2016): 1–6. 10.1109/WIFS.2016.7823911.

Sani, K J, M Kaya, and S Karakuş. "Copy-move forgery detection in digital forensic images using cnn". *7th International Conference on Computer Science and Engineering (UBMK)* (2022): 239–245.

Takeda, N, et al. "Detecting tampered regions in jpeg images via cnn". *18th IEEE International New Circuits and Systems Conference* (2020): 202–205. 10 . 1109 / NEWCAS49341 . 2020 . 9159761.

Embargo period: The article has no embargo period.

To cite this Article: , Krishnan S, Pranay Varma, Aravind J , and Indra Gandhi K. "Analysis of Supervised and Unsupervised Deep Learning Approaches for Identifying and Localizing Image Forgeries." International Research Journal on Advanced Science Hub 05.05S May (2023): 15–25. http://dx.doi.org/10.47392/irjash.2023.S003