Application of Convolutional Neural Network for Image Forgery Detection and Localization

Neelanjan Manna, School of Computer Engineering, KIIT University, neelanjanmanna@gmail.com
Sahil Kumar, School of Computer Engineering, KIIT University, rkakar2000@gmail.com
Rohan Kakar, School of Computer Engineering, KIIT University, io.sahilkr@gmail.com
Sankalp Nayak, School of Computer Engineering, KIIT University, sankalp.nayakfcs@kiit.ac.in

Abstract- With the quick adoption of internet, social media, and widespread availability of unsophisticated image manipulation tools, the adverse consequences of image forgery are taking a toll on society. More than often visual inspection is ineffective to determine a forgery in an image, making it potentially perilous for all the illicit means of its application. To overcome this problem, we analyze the effectiveness of a deep convolutional neural network for the detection as well as localization of the area of manipulation in forged images, bearing forgeries of simple as well as complex nature. Further along, we interface the trained model with a web application for users to interact with the model in a simple and effective manner, and finally, we also develop a chatbot for further easing the process of interaction with the model and effectively tackling the problem of fake news forwards in popular internet messaging platforms such as WhatsApp.

Index Terms- Chatbot, Convolutional Neural Network, Digital Forensics, Image Forgery, Supervised Learning, Web Application

I. INTRODUCTION

The advent of the digital era and social media has incepted several unforeseen problems, out of which arguably the most severe is that of image forgery. Image Forgery refers to the act of manipulating the digital images in order to mask some meaningful data in an image. It has been a widespread predicament for several years now, contributing to the growth of illicit domains such as fake news, false insurance claims, blackmailing, election voter manipulation, communal violence, and tampered academic publications. Out of the 20, 621 academic papers that had published in 40 biomedical research journals between 1995 to 2014, at least 3.8% of them were found to have objectionable images, with tampered results, with at least a staggering half of them suggestive of deliberate manipulation [2]. With extensive adoption of image manipulation tools like Adobe Photoshop, GIMP, and Corel Image Draw, as well as the ever-soaring number of unattributed, free images found on the internet, potentially anyone can tamper and manipulate digital images. This has also raised concerns over the admissibility of digital images as evidence in legal proceedings [1]. Notably, the dissemination of fake news in India has relied largely on the social media platforms such as Facebook and WhatsApp, with 269 million and 250-300 million active users respectively as of October 2020. In 2018, an incident took place in Maharashtra, an India state, where five people belonging to a nomadic group were subjected to lynching by a mob on the basis of a rumour that was incited through WhatsApp about a gang responsible for multiple child-abduction being operative in that area. This incident, in which a total of 30 people succumbed to death, would later be known as the "WhatsApp Killings" [10]. Thus a reliable tool for automatic detection of image tampering is the need of the hour.

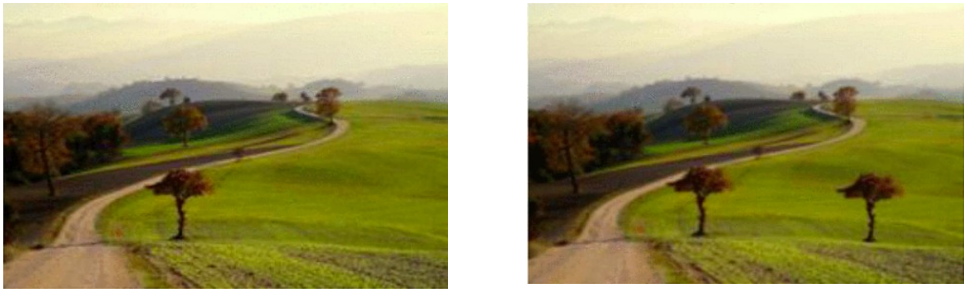
II. BASIC CONCEPTS/ TECHNOLOGY USED

Image forgery is a conglomeration of distinct methods for image tampering, most common of which includes image enhancement, splicing, copy-move, and removal.

Figure 1. Image Enhancement (a) Untampered Image (b) Contrast Enhanced Image [13]

(a) (b) (c)

Figure 2. Image Splicing (a), (b) Untampered Image (c) Tampered Image [13]



(a) (b)

Figure 3. Copy-Move (a) Untampered Image (b) Tampered Image [14]

(a) (b)

Figure 4. Removal (a) Untampered Image (b) Tampered Image [15]

Image enhancement refers to the wide array of local manipulation techniques such as sharpness adjustment, contrast adjustment, brightness adjustment, blurring, etc. Copy-move manipulation, also sometimes referred to as cloning, is formed by copy-and-paste of a portion of the image within the same image. Image splicing is the practice of combining the fragments of images sourcing from the same image or from different images, without the application of any kind of post-processing. Finally, removal which is better known as image inpainting is the act of erasing a certain region of the image and filling the eroded space with pixel values matching the background of the image. However, real-life image forgeries are a sequence of more than one type of forgery, which are carefully articulated in an attempt to mask the forgery, with the use of extremely sophisticated tools such as DNN aided face-swapping (Fig. 5), image inpainting and neural style-transfer.

Image forgery detection methods can be broadly classified in two groups: active image forgery and passive image forgery. Active image forgery techniques depend upon a priorly affixed authentication code, which is added into the image. The implementation of this technique includes digital signatures or digital watermarking. Thus in order to detect a potential forgery, both the original image as well as the forged image is necessary.

On the other hand, detection of passive image forgery, which is otherwise called blind image forgery, relies only upon the forged image to ascertain the presence of forgery without any kind of embedded authenticity code. It is based on the fact that even though a digital image manipulation can't be ascertained by visual inspection, they are very likely to introduce digital artifacts in the form of a number of inconsistencies, resulting from the disturbance of the underlying statistical properties of the original image.

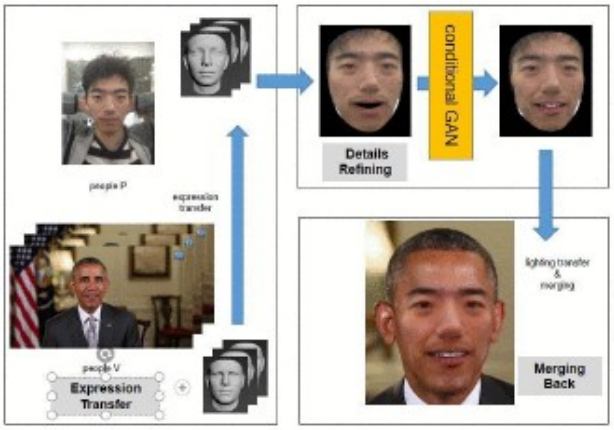


Figure 5. Face-Swapping

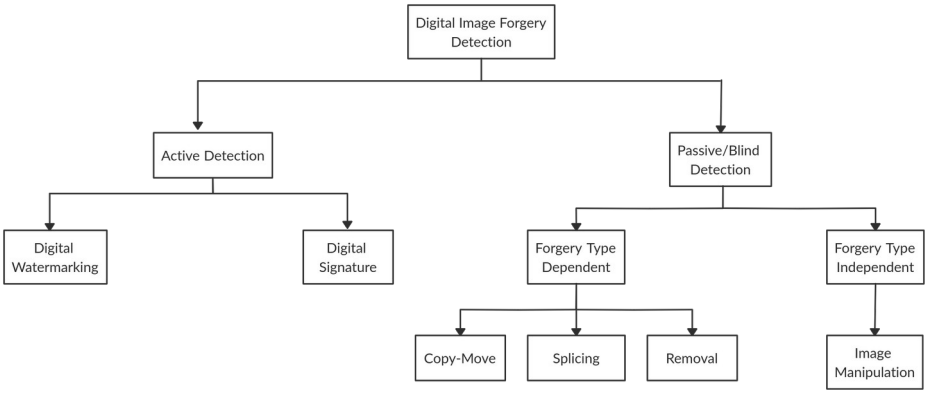


Figure 6. Classification of Digital Image Forgery Detection

III. Study of similar projects or technology / literature review

A CNN oriented approach involving the notion of a new convolutional layer was put forward by Bayar et al. [4], which learns the features attributing to image manipulation by suppressing the content of the image. This novel layer doesn't consider the image content, instead it only uses the local structural relationship between the pixels, as the tampering causes the alteration of some local relationships only. This approach enabled multiple forgeries present in the same image to be detected, which was a major problem for the pre-existing techniques. The method was found to be able to successfully classify several of the image forgeries with an average accuracy of 99.10%.

An approach for localization of image splicing with the help of Multi-task Fully Convolutional Network (MFCN) was proposed by Salloum et al. [5]. The model architecture was inspired by FCN VGG-16 neural network. It is established that MFCN gives better results as compared to FCN as it is able to overcome the problem of coarse localization of outputs. This particular approach splits two output branches, out of which the first branch learns the surface label features and the second branch learns the features along the manipulation region boundary. Although the approach mildly incapacitates in the case of post-processing, yet it offers a comparatively good result, with an average F1 score of 0.5410 on the CASIA dataset [12] and an average MCC (Matthews Correlation Coefficient) score of 0.5201.

Islam et al. [7] put forward a GAN architecture incorporated with a dual-attention network for the classification as well as localization of image forgeries. A first-order attention map is used by the generator part of the network for learning the features along the copy-move region while the second-

order attention learns the features attributing to the patch co-occurrence. Following this, both of the attention maps are obtained to aid with the fusing of location-dependent and co-occurrence features for the detection and localization of forgery. To obtain more accurate results for forgery localization the discriminator network is used. This model is currently state-of-the-art, which produced an F1 score of 69.53 for forgery detection and an F1 score of 41.44 for forgery localization on the CASIA dataset and CMFD dataset.

IV. Proposed Model / Tool

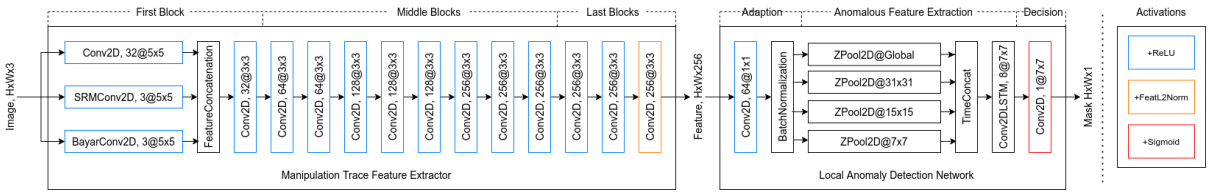


Figure 7. CNN Model

As shown in Fig. 7, the base model for the deep CNN used in this project has been derived from Wu et. al.'s ManTra-Net [6], with parametric modifications, such as altering the number of convolutional filters to 32 in the first block, to improve the model performance.

The model is a fusion of two separate neural networks, the first network being a feature extractor for detection of image forgery and the second network being a local anomaly detection network (Fig. 7). The first neural network extracts those features from an image which are suggestive of its manipulation to create a unified feature rendering for both the networks, whereas the second neural network localizes the forged regions in the image. The manipulation trace feature extractor is used for of classifying 385 types of image forgeries, which includes the basic manipulations as well as more complex manipulations such as the DNN based manipulations and sequential manipulations, which is a definite advantage over the previous methods [8]. For the manipulation trace feature extractor, VGG network architecture was chosen as the base network architecture with the default ReLU activation, as defined in Eq. (1).

$$\sigma(x) = \begin{cases} max(0, x) & , x >= 0 \\ 0 & , x < 0 \end{cases} \tag{1}$$

The base network was modified to be wider using more number of convolutional filters as well as deeper using more number of convolutional blocks, along with the use of SRMConv2D [11], BayarConv2D [3] and classic Conv2D in the first block. Softmax activation defined in Eq. (2) was used to account for multi-class classification of the forgeries, where σ refers to the softmax function, $\vec{z}$ refers to the input vector, z_i refers to the i^{th} element of the input vector and K refers to the number of classes.

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{2}$$

To improve upon the generalizability of the model, the image forgery and localization detection is defined to be a local abnormality detection task, unlike the previous methods that defined it as a semantic segmentation task. For this purpose, a decision function is learned using the dissimilarity between the local feature of the forgery and the forgery label its referenced to. The local anomaly detection network is divided into three stages namely "adaption", which processes the features extracted for the manipulation trace detection so as to be used for anomaly detection, "anomalous feature extraction", which extracts the anomalous features, and lastly "decision" which classifies whether a pixel is forged or not. The most important stage out of these is the anomalous feature extraction step, where the model seeks to first identify the most dominant feature and then any feature which is sufficiently distinct from the dominant feature is considered as an anomalous feature. This intuition is represented with the help of two novel neural network architectures, the ZPool2D layer, which standardizes the aforementioned dissimilarity like the Z-score and also the far-to-near observation using the Conv2DLSTM layer for the processing of the ZPool2D layer outputs stacked together from different resolutions [6].

V. IMPLEMENTATION AND RESULTS

The model was defined in Keras API with Tensorflow as the back-end. The training was commenced with at 800 steps per training epoch, featuring a batch-size of 32. Adam optimizer was used to enhance the training, which was initialized with a non-decaying learning rate of 1e-2. The learning rate was set to become half of its previous value during the training if the validation loss failed to update after a patience of 10 epochs. Finally, the model was fit with a categorical cross-entropy loss function and the trained model was saved as an HDF5 file. For evaluation, the original ManTra-Net model was tested against the classical unsupervised models like ELA [17], NOI1 [18] and CFA1 [19] and DNN solutions like MFCN [5] and J-LSTM [16], with standard benchmark datasets like NIST 2016 [20], CASIA [12], COVERAGE [11], and Columbia dataset [9]. The comparison of AUC scores is tabulated in Table 1.

Model Name	NIST	Columbia	COVERAGE	CASIA
Forgery Types	splicing, copy-move, enhancement	splicing	copy-move	splicing, copy-move, removal
ELA	42.9%	58.1%	58.3%	61.3%
EOI1	48.7%	54.6%	58.7%	61.2%
CFA1	50.1%	72.0%	48.5%	52.2%
J-LSTM	76.4%	N/A	61.4%	N/A
RGB-N	93.7%	85.8%	81.7%	79.5%
ManTra-Net	**79.5%**	**82.4%**	**81.9%**	**81.7%**

Table 1. Comparison of the AUC score of the model with the SOTA models using the benchmark datasets

Figure 8. Website Front-end User Interface

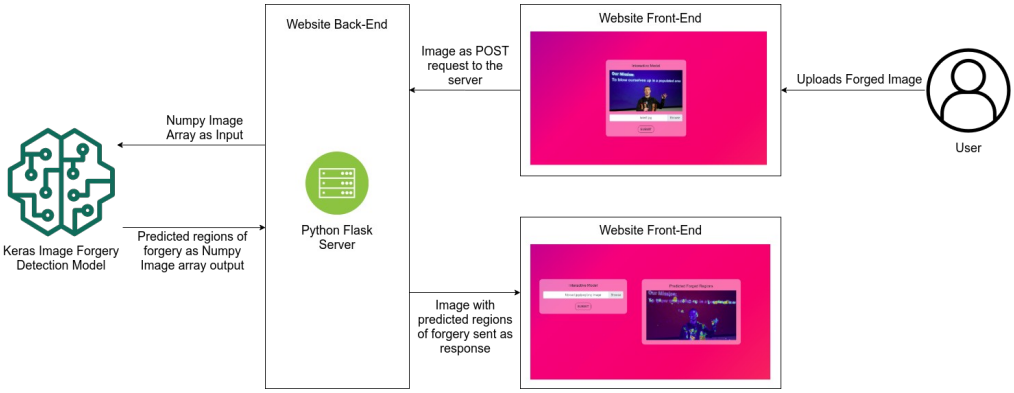


Figure 9. Workflow diagram of the web application for interacting with the model

A web interface was developed for interacting with the trained model, which utilizes a Python back-end API endpoint server developed using the Flask framework. The server listens for any POST request made on the API endpoint with an image. The POST request can be made using the front-end web interface developed using HTML, CSS, and JavaScript where users can upload the images to be checked for manipulation (Fig. 8). Upon receiving the uploaded image as a POST request by the server, the deployed Keras model with Tensorflow backend accepts the image as a Numpy array and outputs a Numpy image array with localized regions of forgery highlighted. The output image is then sent back to the front-end interface as a response by the server (Fig. 9).

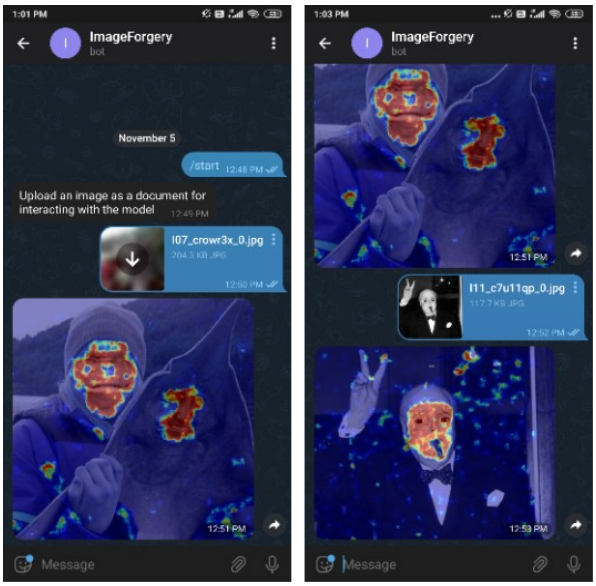


Figure 10. User Interface of the Telegram Chatbot

We also developed a chatbot for the Telegram messenger to further ease the process of interaction with the model, so that the manipulated images can be directly forwarded to the chatbot for the purpose of classification by the model. The chatbot uses a second independent Flask server with another API endpoint which listens for any image sent through the chatbot. The front-end of the chatbot is integrated within the Telegram messenger application itself, through which users can send the manipulated image with greater ease (Fig. 10). Though notably, the users are required to send the image "as a document" in the interface to prevent any further JPEG compression while processing the image. Upon receiving the image, the chatbot sends the image to the server as a POST request. After the successful reception of the POST request by the server, as before, the image is input to the Keras model, which outputs an image array with the predicted regions of forgery, which is then sent back to the chatbot interface as a response by the server (Fig. 11).

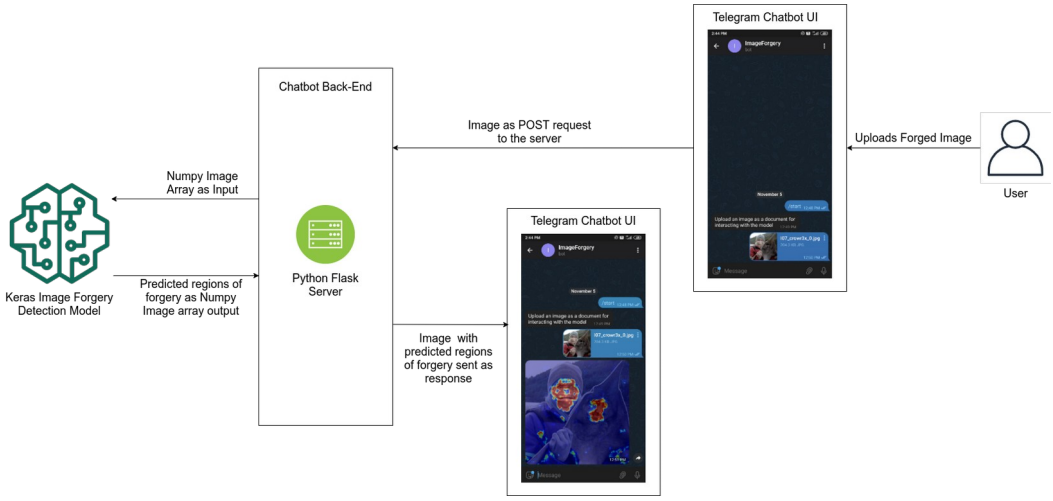


Figure 11. Work Flow Diagram of the Chatbot for interacting with the model

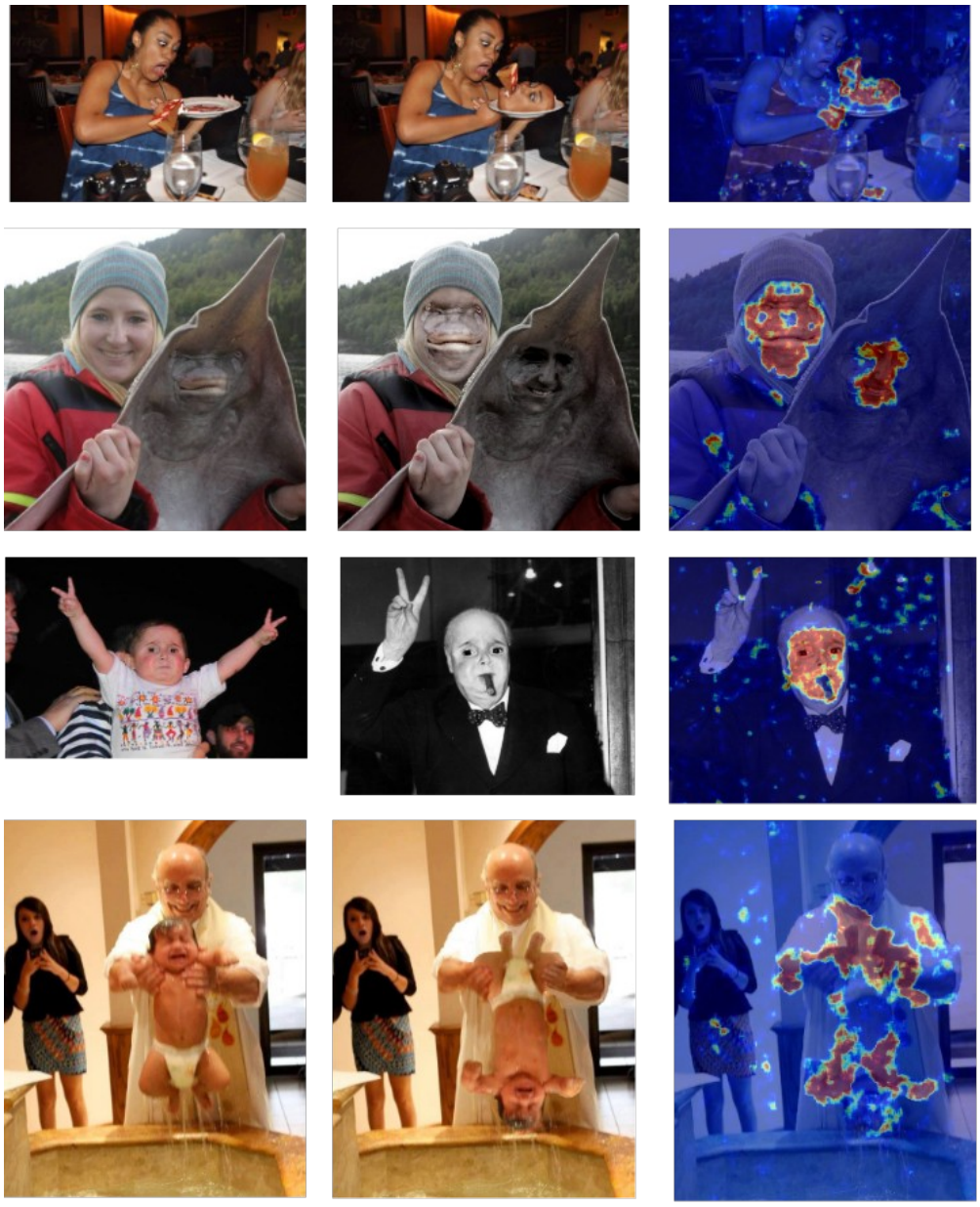


Figure 12. Sample Results from the model (a) Untampered Image (b) Tampered Image (c) Predicted regions of forgery

VI. CONCLUSION

In this project, we analyzed the application of a CNN for image forgery detection as well as anomaly region localization. A total of 385 different types of image forgeries can be detected using this model, including forgeries of complex nature such as DNN aided image forgeries. Furthermore, we developed a web application interface for interacting with the model in an easy and effective manner, and finally, we developed a chatbot interface for the purpose of further simplifying the process of interaction with the model while effectively tackling the plight of fake news forward through internet messaging platforms. One way to improve the practical usability of the model will be to reduce the amount of memory consumed by the model during execution so that it can be executed on resource-constrained hardware as well. Additionally, an NLP aided text classifier can be integrated with the chatbot to make the classification of fake news more robust.

REFERENCES

[1] Nagosky P. D., *Admissibility of Digital Photographs in Criminal Cases,* In: FBI Law Enforcement Bulletin, Volume:74, Issue:12, Dated: December 2005, pages 1-8

[2] E. M. Bik, A. Casadevall, and F. C. Fang. *The prevalence of inappropriate image duplication in biomedical research publications*. In: MBio, 7(3):e00809–16, 2016, pages 1-8

[3] B. Bayar and M. C. Stamm, *Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection*, In: IEEE Transactions on Information Forensics and Security, vol. 13, no. 11, Nov 2018, pages 2691-2706

[4] B. Bayar and M. C. Stamm, *A deep learning approach to universal image manipulation detection using a new convolutional layer*, In: Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security (2016), pages 5-10

[5] R. Salloum, Y. Ren and C.-C. J. Kuo, *Image splicing localization using a multi-task fully convolutional network (mfcn)*, In: Journal of Visual Communication and Image Representation, vol. 51, 2018, pages 201-209

[6] Y. Wu, W. AbdAlmageed and P. Natarajan, *ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image Forgeries With Anomalous Features* (2019), In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pages 9535-9544

[7] A. Islam, C. Long, A. Basharat and A. Hoogs, *DOA-GAN: Dual-Order Attentive Generative Adversarial Network for Image Copy-Move Forgery Detection and Localization* (2020), In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pages 4675-4684

[8] I. Amerini, T. Uricchio, L. Ballan and R. Caldelli, *Localization of jpeg double compression through multi-domain con-volutional neural networks* (2017), In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, 2017, pages 1865-1871

[9] Image splicing detection evaluation dataset, 2004, [online] Available: http://www.ee.columbia.edu/ln/dvmm/downloads/AuthSplicedDataSet/AuthSplicedDataSet.htm

[10] Billy Perigo, How Volunteers for India's Ruling Party Are Using WhatsApp to Fuel Fake News Ahead of Elections, [online] Available: https://time.com/5512032/whatsapp-india-election-2019/

[11] P. Zhou, X. Han, V. I. Morariu and L. S. Davis, *Learning rich features for image manipulation detection*, In: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pages 1907-1915

[12] J. Dong, W. Wang and T. Tan, *CASIA Image Tampering Detection Evaluation Database* (2013), In: IEEE China Summit and International Conference on Signal and Information Processing, Beijing, 2013, pages 422-426

[13] Kaur, C. deep, & Kanwal, N. (2019), *An Analysis of Image Forgery Detection Techniques,* In: Statistics, Optimization & Information Computing, *7*(2), pages 486-500

[14] Arun Anoop M, *Image forgery and its detection: A survey*, 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore, 2015, pages 1-9

[15] P. B. Shailaja Rani and A. Kumar, *Digital Image Forgery Detection Techniques: A Comprehensive Review* (2019), In: 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2019, pages 959-963

[16] J. H. Bappy, A. K. Roy-Chowdhury, J. Bunk, L. Nataraj and B. Manjunath, Exploiting spatial structure for localizing manipulated image regions (2017), In: IEEE International Conference on Computer Vision, pages 4970-4979

[17] N. Krawetz and H. F. Solutions, A pictures worth (2007), In: Hacker Factor Solutions, vol. 6, 2007

[18] B. Mahdian and S. Saic, Using noise inconsistencies for blind image forensics (2009), In: Image and Vision Computing, vol. 27, no. 10, pages 1497-1503

[19] P. Ferrara, T. Bianchi, A. De Rosa and A. Piva, Image forgery localization via fine-grained analysis of cfa artifacts (2012), In: IEEE Transactions on Information Forensics and Security, vol. 7, no. 5, pages 1566-1577

[20] NIST manipulation evaluation dataset, 2016, [online] Available: https://www.nist.gov/itl/iad/mignimble-challenge-2017-evaluation