Document Forgery Detection

Nandini N, Keerthi Joshi K, Devprakash, Madhura C, Vandana M Ladwani

Abstract: ***Document forgery is an increasing problem for both private companies and public administrations. It can be said to represent the loss of time and resources. There are many classical solutions to these problems such as the detection of an integrated security pattern. In such cases, it is important that we resort to forensic techniques for the detection. The idea behind using these forensic techniques can also be implemented using artificial intelligence/machine learning which can be of lower cost and can provide the same or better results. The experimental result shows that multiple models have strong detection capability to detect multiple forgeries. In this paper, we have developed a different approach to detecting forgery in a document. The forgery we detect can be classified as hand-written signature forgery and copy-move forgery of any photo, text, or signature. We have developed a novel approach using capsule layers to detect a forgery in handwritten signatures. We also use ELA (Error Level Analysis) to detect any error in the compression levels of the image.***

Keywords: Document, Forgery Detection, Capsule Neural Networks, CNN, Copy-Move Forgery, Signature Forgery

I. INTRODUCTION

We now live in a digital age where technology has become a key aspect in our day-to-day lives, from creating, processing, and storing information to displaying user needs and desires. Knowledge representation is multi-dimensional and stored as bits and bytes in various formats such as texts, photos, and videos. Technology can be an issue. With recent advances in technology, people are misusing technology in ways that can harm society. Changing data without visible traces of intervention has never been easier. With this, there are several opportunities as well as risks, and the gravity of the threats cannot be understated.

Manuscript received on 10 May 2023 | Revised Manuscript received on 22 May 2023 | Manuscript Accepted on 15 June 2023 | Manuscript published on 30 June 2023.

*Correspondence Author(s)
Nandini N*, Department of Computer Science and Engineering, PES University, Bangalore (Karnataka), India. E-mail: nandinins236@gmail.com, ORCID ID: https://orcid.org/0009-0002-5371-4667
Madhura C, Department of Computer Science and Engineering, PES University, Bangalore (Karnataka), India. E-mail: cmadhura122001@gmail.com, ORCID ID: https://orcid.org/0009-0009-4394-112X
Keerthi Joshi K, Department of Computer Science and Engineering, PES University, Bangalore (Karnataka), India. E-mail: keerthi.joshi2822@gmail.com, ORCID ID: https://orcid.org/0009-0008-8709-9539
Devprakash B, Department of Computer Science and Engineering, PES University, Bangalore (Karnataka), India. E-mail: devprakashbesoi@gmail.com, ORCID ID: https://orcid.org/0009-0002-6176-578X
Vandana M Ladwani, Department of Computer Science and Engineering, PES University, Bangalore (Karnataka), India. E-mail: vandanamd@pes.edu, ORCID ID: https://orcid.org/0000-0002-4099-6936

Impersonating or altering artifacts, such as significant papers, pictures, news articles, works of art, etc., is called forgery. It is always coupled with other fraudulent activities including check fraud, insurance fraud, identity theft, and smuggling. False social media profiles and modified email messages are examples of technological forgeries that are not always tangible. We have noticed a significant increase in demand for online document authentication in e-commerce and e-government applications due to the pandemic's continuous issue. Documents are placed on internet platforms for a variety of reasons. However, certain editing software or other technology may alter the content of the document. It is important to identify the changes done to document photocopies. Because images can serve as small pieces of forensic evidence or can be used in court. Proving the authenticity of documents is very important. Forgery process of imitating or changing objects which can include important documents, images, news, works of art, etc. It is always accompanied by other fraudulent behaviors like identity takeover, smuggling, insurance fraud, check fraud, and many more. Forgeries do not always have to be physical, they can also be electronics such as fake social media pages or adaption of email correspondence. Since the concern of our topic is documented forgery, we will show the limelight on different kinds of document forgeries. Genuine passports without any personal data or without any stamp, or any signed memorandum letters without content can be considered real documents as they are blank documents. The forger inserts the data to perform the fraudulent activity. The forgeries in these kinds of documents are difficult to detect. The detection of any manipulation in document images is important as an image can be used as legal evidence, in any forensics investigation, and in many other fields. Hence, there is a dire need to prove the authentication of a document.

II. RELATED WORK

Digital forgery detection has been an ongoing topic in the research field for many years. Numerous solutions for the same have been proposed addressing different kinds of forgery detection. The existing document forgery detection methods can be broadly classified into main categories i.e active and passive methods. The paper [1] proposes a system architecture based on the inspection of probed documents with the analysis of ink. The paper produced a new method to find any mismatch of ink color in the HSD images. The approach is based on an NMF model with orthogonal as well as graph regularization. The assumption made here is that under some attainment protocols, some of the latent actors present in the HSD images can be forced to be orthogonal. The author of this paper also has proposed an efficient algorithm that is multiplier based to incorporate in the method.

The paper [2] proposes an efficient method to detect signature forgery which is based on the siamese neural network. The method uses CNN for data preprocessing and evaluation is done using a siamese network working with the CNN model. Unique features of the implementation include a contrastive loss function. A high recall was achieved and loss was minimized to 0.43. The paper [3] proposes a robust system to detect digital forgery using CNN architecture for compressed images. The CNN architecture consists of multiple layers such as a pooling layer, a convolutional layer, and fully connected layers. The authors in [4] offer a shallow convolutional neural network(CNN) that can identify manufactured region boundaries from original edges in low-resolution images. SCNN was created to make use of chroma and saturation data. Two techniques based on SCNN, term sliding windows detection(SWD) and fast SCNN have been developed to detect and identify image forgery regions.

The paper[5] provides a new deep learning-based image fraud detection system for automatically learning hierarchical representation from input RGB color photographs. Image splicing and copy-move forgeries can be detected using the suggested CNN, The fundamental high-pass filter set employed in the spatial rich model(SRM) is utilized to establish the weights at the first layer of our network, which serves as a regularizer to efficiently suppress the effect of picture contents and capture the subtle artifacts created by the tampering operations. The pre-trained CNN is used as a patch descriptor to retrieve dense features from the test images, and the final discriminative features for SVM classification are obtained via a feature fusion technique.

III. DATASET

Finding an appropriate dataset for training and testing the model is one of the challenging tasks while creating an ML model. Since training of the ML model almost depends on the dataset that has been used. Forged document images dataset are not available easily. We planned to create the dataset by collecting real document images from the web and forging the images to create the forged image dataset. The dataset consists of two types of data, one is for training copy move forgery detection model and the other is for training the signature forgery detection model. Since the dataset is created manually, it requires preprocessing. The data preprocessing is done using the openCV technique. Since the noise in the image has a great effect on the result of the models. Preprocessing includes slant correction and denoising. Since We have approached the forgery detection problem by creating different models for different types of intrinsic features, extraction of that particular feature is also needed. The signature forgery detection model accepts the extracted signature from the uploaded image.

IV. PROPOSED WORK

The implemented system consists of preprocessing, capsule network, error level analysis, and CNN.

a) Data preprocessing: preprocessing is done to reduce the noise in the uploaded image, the presence of which might result in false positives or lesser efficiency of the detection model. Preprocessing phase plays a vital role in the product implementation since the errors or unwanted data present in this phase will be carried out till the final stages of the product development. Cleaning the data in the earlier stages will improve the system. The pre-processing stage consists of slant correction, cropping of the unwanted region from the input document image, and image denoising. The slant correction and denoising are done using the OpenCV technique.

b) Capsule neural network: To detect the forgery of signature, a capsule neural network is used instead of CNN. Despite the benefits of CNN, there are always

challenges in using them. CNNs do not have the ability to model the changes in new fields and dimensions and cannot distinguish between similar components that are placed in different locations of an image. The reason to use CapsNet is that it can reduce the number of layers and parameters in the network architecture and decrease complexity. CapsNet is capable of detecting the details of angular and spatial changes in components. Not only is a capsule network able to represent the existence of particular visual features, but also it can detect the transformations that might have occurred in the features. In capsule networks, spatial details are completely differentiated.

c) Error level analysis and CNN: To convert an image to ELA, the preprocessed images must be reframed at a certain level of quality. The image is whitened or brightened as a result of this technique. In order to reframe the images, forged and real images are considered that have been processed in preprocessed. Finally, the preprocessed image and reframed image are compared to see the difference between them. The tampered parts of the image in the forged ELA framed image are brighter than the corresponding original segments of the image. Having ELA analysis before the processing of the neural networks has a huge advantage as the ELA reframed image contains only non redundant information and nearby pixels are compared based on similar intensity. The reframed image needs to be resized to make the RGB values lie between 0 and 1 so each cell value can be normalized by dividing the values by 255. The resizing helps the neural network converge faster than usual. CNN model consists of 2 layers of a convolutional network. The first layer of the CNN is a convolutional layer with 32 filters. Dropout is used as a regularization technique for reducing overfitting in neural networks, preventing complex co-adaptations on training data. The CNN consists of 2 convolution layers, max pooling, and 2 dropout layers

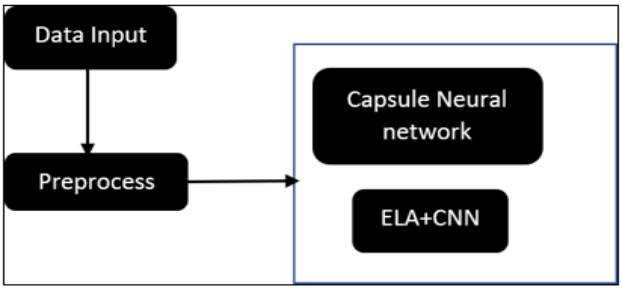

Fig 1: System architecture

The result of the forgery detection is produced considering the result of individual models. The individual models are enclosed within an ensemble model. Since the fakeness of each feature is detected by making use of different models, the results from the respective models need to be combined to get the overall fakeness result. All the models are combined together with an ensemble model and the overall fakeness of the document is produced as the output.

V. RESULTS AND DISCUSSIONS

The training images for the copy-move model, CASIA-2 and MICC dataset have been considered. The dataset consists of images from these two datasets which are split into training and testing datasets and then passed onto our model to classify them into two classes i.e authenticate and forged.

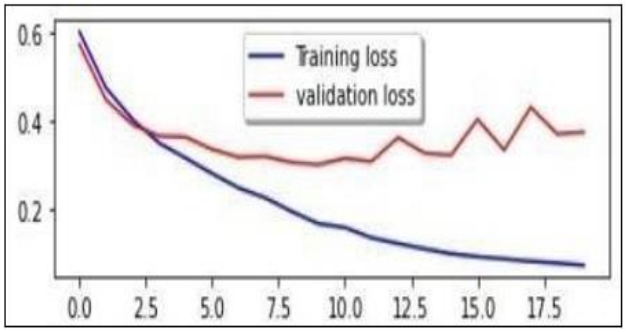

Fig 2: Training loss and Validation Loss of Copy Move Forgery Detection Model

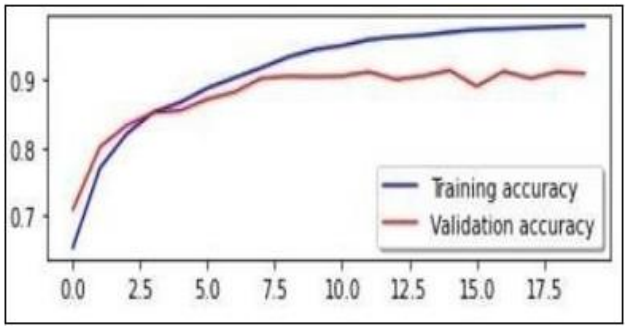

Fig 3: Training Accuracy and Validation Accuracy of Copy Move Forgery Detection Model

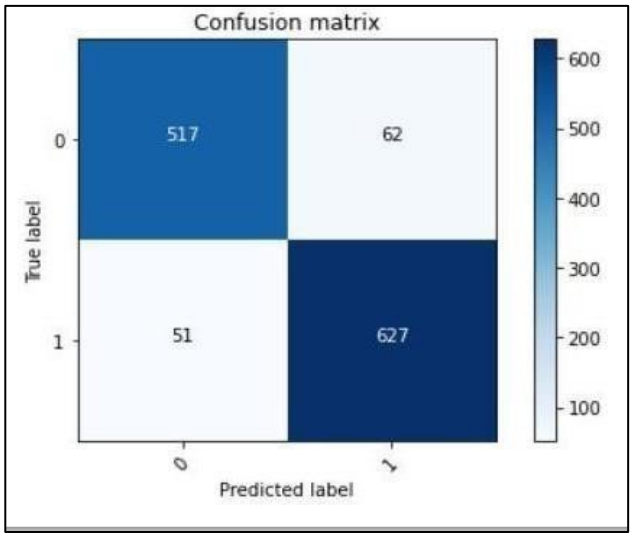

Fig 4: Confusion Matrix of Copy Move Forgery Detection Model

VI. CONCLUSION

In this paper, a method to detect a digital document forgery using a neural network has been presented. The proposed solution uses a capsule neural network for signature forgery detection and a combined model of ELA and CNN is used for copy-paste forgery detection. The proposed method uses features from two datasets of varying difficulty. The experiment results validate that the classification performance decreases when the samples are more challenging. However, the implemented architecture does not easily generalize to datasets with different underlying distributions. The capsule neural network was found to be efficient in detecting the forge-ness of the signatures in the document images. ELP was found to be efficient with detecting copy-move forgery detection. The overall positive and negative forgery detection rate of the designed system was very promising. It is encouraged that future work is required to investigate completely on a wider range of algorithms with even higher efficiency and also to detect the fraudulent use of video [MP4] format files. Overall, the project work has given us an opportunity to dig deeper into forgery detection methods. Image forgery detection is not only an emerging topic in research but an important topic in which faster and more accurate work is needed. The accurate and efficient methods to detect forgery are increasing day by day. Nevertheless, there is surely a lot of work still to be done in the image forgery detection domain and neural networks will be able to detect tampered images regardless of their difficulty. In the future, there is scope to improvise the training model by increasing the variance in the dataset and also it has scope to detect the forgery in video files.

DECLARATION

Funding/ Grants/ Financial Support	No, I did not receive.
Conflicts of Interest/ Competing Interests	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material/ Data Access Statement	Not relevant.
Authors Contributions	All authors have equal participation in this article

REFERENCES

1. Rahiche, Abderrahmane; Cheriet, Mohamed (2020). [IEEE 2020 IEEE/Cvf Conference On Computer Vision And Pattern Recognition Workshops (Cvprw) - Seattle, Wa, USA (2020.6.14-2020.6.19)] 2020 IEEE/Cvf Conference On Computer Vision And Pattern Recognition Workshops(Cvprw) - Forgery Detection in Hyperspectral Document Images Using Graph Orthogonal Nonnegativematrixfactorization.,(),2823-2831.Doi:10.1109/Cvprw50498.2020.00339 [CrossRef]

2. S. Jain, M. Khanna, and A. Singh, “Comparison among different CNN architectures for signature forgery detection using siamese neural network”; 2021 international conference on computing, communication, and intelligent systems (ICCCIS), 2021, pp. 481-486, DOI:10.1109/ICCCIS51004.2021.9397114. [CrossRef]
3. Robust forgery detection for compressed images using CNN supervision Boubacar Diallo *, Thierry Urruty, Pascal Bourdon, Christine Fernandez- Maloigne Universite de Poitiers, cnrs, xlim, umr 7252, f-86000 poitiers, france,volume 2, december 2020, 100112. [CrossRef]
4. IJCSNS International Journal of Computer Science and Network Security, vol.20no.12 december 2020 :https://doi.org/10.22937/ijcsns.2020.20.12.12 [CrossRef]
5. Zhang, Zhongping & Zhang, Yixuan & Zhou, Zheng & Luo, Boundary-based image forgery detection by fast shallow CNN, conference: computer vision and pattern recognition Doi: 10.11.09/ICPR (2009).

AUTHOR PROFILE

Ms. Nandini N is a passionate researcher and writer with an interest in machine learning and artificial intelligence. Ms Nandini is currently pursuing her Bachelor's of Technology degree with a specialization in computer science and engineering at PES University. Ms Nandini has got a deep understanding on image processing and machine learning concepts through deep research work and academics. Through the academic journey, Ms Nandini has consistently worked on machine learning projects to enhance her knowledge about the machine learning domain. Ms Nandini also has knowledge about image processing through the course work and also has worked on the team project on the same domain. In conclusion Ms Nandini is a passionate and dedicated researcher to work beyond the domain boundaries, With the well-equipped to tackle the challenges and contribute towards the research domain.

Ms. Keerthi Joshi K is a dedicated researcher and writer with a passion for Data Science and Machine Learning. Currently pursuing Computer Science Engineering in B.Tech at PES University, Bangalore, Ms Joshi has developed a deep understanding of Data science and Machine Learning through extensive research and coursework. Throughout her academic journey, Ms Joshi has consistently demonstrated exceptional analytical skills and a keen eye for detail. She and her team have conducted thorough investigations, analyzing complex data sets and examining diverse perspectives to arrive at well-rounded conclusions. Their commitment to academic excellence is evident in their consistent high grades and their active participation in class discussions and research projects. In conclusion, Ms. Joshi is an accomplished and driven researcher, dedicated to expanding the boundaries of knowledge in not only Data Science but also in Machine Learning. With their academic achievements, research experience, and commitment to excellence, she is well-equipped to tackle the challenges and plan to make valuable contributions to the scholarly community.

Devprakash is a dedicated and ambitious student at PES University Bangalore, currently pursuing their academic journey with a focus on data analytics and machine learning. As a student by status and a continuous learner by mindset, Devprakash is committed to acquiring knowledge and honing their skills in these rapidly evolving fields. With a deep passion for data analytics and machine learning, Devprakash enthusiastically explores the vast potential of these subjects. They actively engage in coursework, projects, and extracurricular activities that contribute to their understanding and expertise in the field. Devprakash's favorite subject, data analytics and machine learning, serves as a driving force behind their academic pursuits. Devprakash's proactive nature has led them to actively participate in hackathons, where they have consistently demonstrated their ability to solve complex problems and create innovative solutions. By collaborating with diverse teams and leveraging their analytical skills, Devprakash has successfully showcased their talent, earning recognition and respect from peers and mentors. Looking ahead, Devprakash envisions utilizing their skills in data analytics and machine learning to make a significant impact. They have a clear vision of leveraging these technologies to drive positive change in various domains such as healthcare, finance, and social issues. Devprakash firmly believes that the power of data, when combined with advanced analytical techniques, can revolutionize industries and improve lives on a global scale.

Madhura C is a bright and driven student currently pursuing her B.Tech graduation from PES University. She is currently working as a student intern at Textron Company, where she is showcasing her impressive problem-solving skills and ability to make strong decisions. Her exceptional technical expertise in areas such as Python, Data Science, Cyber Security, and Machine Learning are highly valued by her peers and colleagues. Madhura's passion for data science is evident in her impressive projects and academic achievements. Her work showcases her exceptional ability to utilize her technical skills to solve complex problems while keeping an eye on the bigger picture. Madhura's decision-making skills and attention to detail are assets that have helped her excel in her academic and professional career. In conclusion, Madhura is an exceptional student with strong technical skills and a passion for data science and machine learning. Her dedication, problem-solving skills, and ability to make sound decisions make her an asset to any team. We are confident that she will continue to excel in her academic and professional career, and we look forward to seeing her make a significant impact in her field.

Vandana M Ladwani is an assistant professor with 14 years of teaching experience. She has Bachelor of Technology and Master of Technology degree in computer science from RTM Nagpur University. She is working in the area of signal processing and machine learning. She has around more than 15 publications in international conferences and journals. She had also contributed to book chapters.