

A constrained convolutional neural network with attention mechanism for image manipulation detection

Kamagate Beman Hamidja¹, Fatoumata Wongbé Rosalie Tokpa², Vincent Monsan²,
Souleymane Oumtanaga³

¹Laboratoire des Sciences et Technologies de l'Information et de la Communication (LASTIC), Ecole Supérieure Africaine des Technologies de l'Information et de la Communication (ESATIC), Abidjan, Côte d'Ivoire

²Laboratoire de Mécanique et d'Informatique (LAMI), Université Félix Houphouët-Boigny (UFHB), Abidjan, Côte d'Ivoire

³Laboratoire de Recherche en Informatique et Télécommunication (LARIT), Institut National Polytechnique Félix Houphouët-Boigny (INPHB), Yamoussoukro, Côte d'Ivoire

Article Info

Article history:

Received May 15, 2024

Revised Oct 5, 2024

Accepted Oct 23, 2024

Keywords:

Attention mechanism
Constrained convolution
Convolutional neural network
Image manipulation
Transfer learning

ABSTRACT

The information disseminated by online media is often presented in the form of images, in order to quickly captivate readers and increase audience ratings. However, these images can be manipulated for malicious purposes, such as influencing public opinion, undermining media credibility, disrupting democratic processes or creating conflict within society. Various approaches, whether relying on manually developed features or deep learning, have been devised to detect falsified images. However, they frequently prove less effective when confronted with widespread and multiple manipulations. To address this challenge, in our study, we have designed a model comprising a constrained convolution layer combined with an attention mechanism and a transfer learning ResNet50 network. These components are intended to automatically learn image manipulation features in the initial layer and extract spatial features, respectively. It makes possible to detect various falsifications with much more accuracy and precision. The proposed model has been trained and tested on real datasets sourced from the literature, which include MediaEval and Casia. The obtained results indicate that our proposal surpasses other models documented in the literature. Specifically, we achieve an accuracy of 87% and a precision of 93% on the MediaEval dataset. In comparison, the performance of methods from the literature on the same dataset does not exceed 84% for accuracy and 90% for precision.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Kamagate Beman Hamidja

Laboratoire des Sciences et Technologies de l'Information et de la Communication (LASTIC) de l'Ecole Supérieure Africaine des Technologies de l'Information et de la Communication (ESATIC)

Bd. de Marseille face à Bernabe - Km 4. Treichville - 18 BP 1501 Abidjan 18. Abidjan - Côte d'Ivoire

E-mail: beman.kamagate@esatic.edu.ci

1. INTRODUCTION

In an age of information technology advances and rapid development of social networks, information can be provided by the media in the form of images or video to quickly captivate readers and increase audience numbers. Visual content, which includes images or videos, is a dynamic means of expression that is more striking than simple text, stimulating the dissemination of information. Visual content is also frequently used as evidence, reinforcing the credibility of information. However, individuals or interest groups disseminate false visual content in order to increase the visibility and achieve their objective, which is to manipulate the opinion of a mass of people, destabilize a government, increase their financial

income or influence the results democratic process [1] These objectives have negative impacts, such as undermining the credibility of social media, weakening democratic processes and generating conflict between members of a society.

Image falsification is a set of digital techniques designed to convey misleading information through modified images [2]. Copy and paste consists in replacing the content within a region of an image with content from another image [3]. Erase-fill or inpainting consists in filling a region of an image with a composition of elementary parts of the original image or another image [4]. Also, with the advent of modern artificial intelligence technology such as generative adversarial networks (GANs), images can be generated or falsified to resemble authentic images [5].

Most research on fake news detection has predominantly concentrated on textual content, *i.e.*, news that relies solely on text. In this context, Tokpa *et al.* [6] introduces a method that combines two neural network architectures, convolutional neural networks (CNNs) and bidirectional long short-term memory (BiLSTM), to improve fake news detection accuracy across various text-based datasets. Similarly, Ajao *et al.* [7] explores the use of a hybrid of CNNs and recurrent neural networks (RNNs) to identify and classify fake news messages on Twitter. Meanwhile, Popat *et al.* [8] present a neural network model that integrates signals from external evidence articles, considers the language used, and assesses the credibility of sources. Their model also generates useful features that provide clear explanations for users, ensuring transparency in the predictions. Finally, Rani *et al.* [9] propose a hybrid approach combining CNN and BiLSTM with global vectors for word representation (GloVe) embeddings to classify tweets as rumors or non-rumors. However, these studies overlook the issue of fake news involving falsified images [10]. Research in [11] indicates that news posts containing images receive more interaction compared to those with only text. As a result, there is growing interest in detecting falsified images on social media.

The falsified image can be detected using manually developed feature-based approaches. It is assumed that manipulated or fake images exhibit visual and statistical distribution patterns distinct from those of genuine images. Jin *et al.* [12] propose several visual and statistical features to detect fake images which are visual clarity, consistency score, visual similarity distribution histogram, visual diversity, clustering score, number of images in a post, image size and image popularity. These features are used as inputs to classifiers algorithms such as decision trees, support vector machines (SVMs), logistic regression and other classifiers to classify an image as forged or not. For these types of methods, the robustness of the feature vectors obtained is not sufficient, as knowledge of the falsification traces in an image is lacking. As a result, it is difficult to use these features to detect false images with acceptable accuracy.

Cao *et al.* [13] provides an in-depth analysis of visual content, highlighting fundamental concepts, important visual features, effective detection methods, and the challenges encountered in this field. Berthet [14] focuses on artificial intelligence-based compression tools to detect forged images. In study [15], principal component analysis (PCA) is used to detect manipulated artifacts in JPEG format images. Vijayalakshmi *et al.* [16] introduces an autoencoder-based method for identifying copy-paste forgeries in digital images. This approach includes image normalization, rescaling, and error level analysis (ELA) to enhance accuracy and reduce overfitting in the network model. To further improve performance, image augmentation is applied to increase the dataset size. Ultimately, the proposed autoencoder-based technique effectively classifies forged images. Study [17] addresses the detection of cut-paste manipulations in images using texture analysis of spliced images. Specifically, it extracts features based on the local entropy of the median filter residual (MFR) of the manipulated image, which helps reduce noise while preserving edges. These features are then used to create the ground truth mask. The goal of study [18] is to create a photo forensics algorithm capable of detecting all types of photo manipulation. To enhance the error level analysis, the study employs vertical and horizontal histograms of the ELA image to accurately identify the location of modifications. Through the previously cited works in this section, we observe that different types of residuals, such as MFR and ELA, are used. These methods achieve high accuracy for single manipulations but show lower accuracy for multiple manipulations. Consequently, their performance when applied to social media images is insufficient, as these images often undergo multiple manipulations [19].

In recent years, approaches based on deep neural networks, in particular convolution neural networks, have emerged [14]. Xue *et al.* [10] proposes a multimodal neural network composed of several modules, including semantic feature extraction and visual falsification. In the first module, features are extracted from the pre-trained ResNet50 neural network, and these features are passed on to the neural network (BiGRU) for semantic feature extraction. The features of the falsified image are obtained by applying ResNet50 to the ELA transformation of the image. In studies [20] and [21], to detect manipulations in images, the authors propose a deep learning model. Images generated by ELA are used as inputs for the neural network model EfficientNetB0. This model is trained and tested on MediaEval [22]. Singh and Sharma [20] achieve an accuracy of 79.47% on MediaEval and in [21] they achieve an accuracy of 81.07% on MediaEval. Some widely used manipulated images may have undergone several types of processing, thus increasing the difficulty of capturing manipulation traces. Indeed, convolutional neural networks such as

ResNet and EfficientNet have demonstrated some capability to detect various types of manipulations. However, they do so with less efficiency and are more effective at learning features that represent the content of the image rather than features that indicate the presence of manipulation. Their performance, which hovers around 80% accuracy, needs improvement. This is why we propose incorporating a constrained convolution layer to automatically learn prediction error filters in the initial layer. This approach enables the isolation of multiple falsification artifacts introduced into an image by eliminating irrelevant information, thereby effectively detecting multiple manipulations. The structure of this paper is as follows: Section 2 outlines the methodology employed and elucidates the proposed algorithms. Section 3 delineates the conducted experiments, provides analysis, and interprets the obtained results. Finally, the conclusion offers a summary of our study and outlines future directions.

2. METHOD

This section is subdivided into two subsections, the first of which is the problem formulation and features extraction methodology, followed by the experimentation method.

2.1. Problem formulation and features extraction methodology

Falsified image detection can be modeled as a binary classification problem that indicates whether an image is genuine or falsified. Consider $I = \{I_1, \dots, I_m\} \subset \mathbb{R}^{L \times l \times c}$ the input features, $Y = \{y_1, \dots, y_m\} \subset \mathbb{R}$ the corresponding labels. The problem is to find a function \mathcal{F} that automatically learns to recognize the characteristics of an image I_i and to predict its truthfulness, *i.e.*

$$\mathcal{F}(I_i) = \begin{cases} 0 & \text{if } I_i \text{ is falsified} \\ 1 & \text{if } I_i \text{ is genuine} \end{cases} \quad (1)$$

The feature extraction model consists of two modules as shown in Figure 1. One for the extraction of features from weathering traces and the second for the extraction of spatial features from the image.

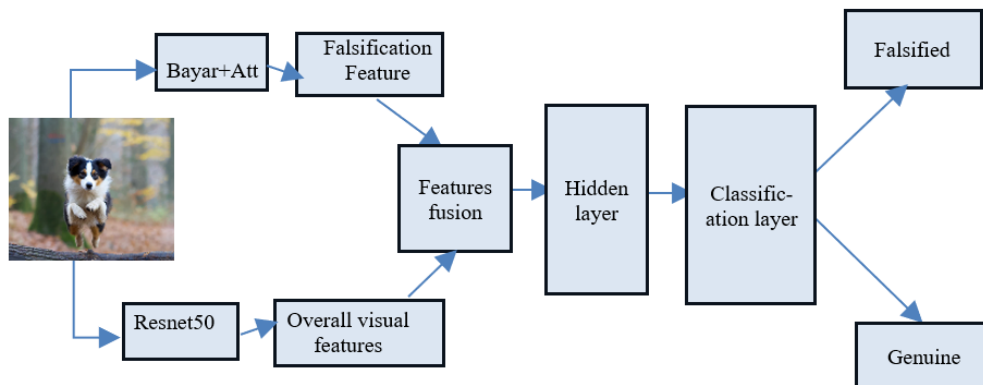


Figure 1. Falsification and spatial visual features extraction by Bayar_Att

The feature extraction module for weathering traces as shown in Figure 2 is composed with the combination of constrained convolution proposed by Bayar and Stamm [23] and attention mechanism for relevant features extraction based on attention mechanism applied in [24]. It is made up of several neural network layers: a constrained convolution layer for extracting low-level tampering features, an attention module for extracting more important features, and a convolution layer followed by a pooling layer to improve the generalization capability of the proposed model. A Convolution layer is a set of filters or matrices applied by convolution operation on another matrix. Filters are feature extractors and the result of the convolution is called a feature map. Let I be the input image of dimension $L \times l \times c$ where L is the length of the image, l the width and c the number of channels. This image convolves with a filter f_i of dimension $n \times n \times c$ with a step of 1 produces the characteristic map $C_i = f_{act}(I * f_i) + b$ of dimension $(L - n + 1) \times (l - n + 1)$ where $*$ is the convolution operation, f_{act} an activation function and b the bias. For k filters applied on the image we obtain k features maps at the output of the convolution layer *i.e.* the output of the convolution layer is of dimension $(L - n + 1) \times (l - n + 1) \times k$. In our proposal, in the Bayar constrained

convolution layer, we propose to use 7 filters of size 5×5 on an input of $150 \times 150 \times 3$. This layer produces a feature vector f_{Bayar} .

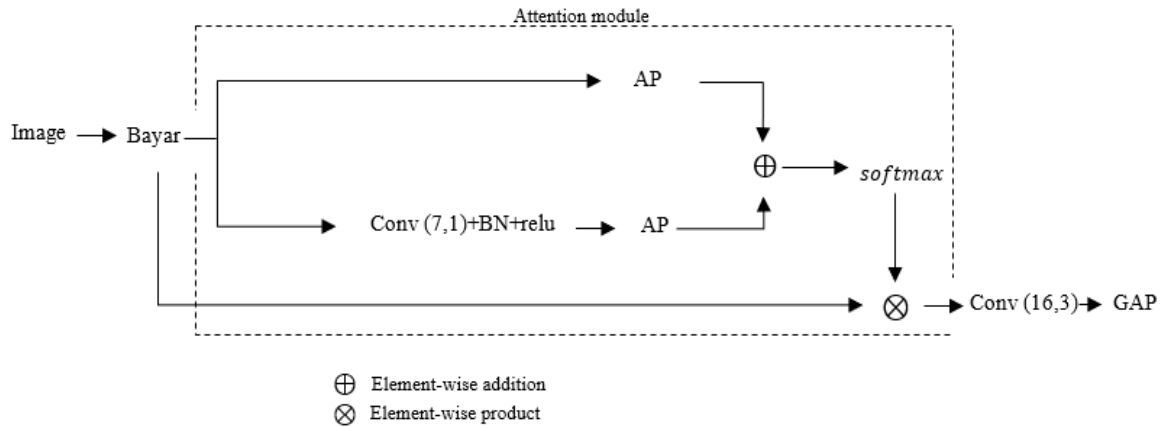


Figure 2. Weathering traces extraction module

Inspired by the attention mechanism of study [25] which allows important features to be selected and adaptively, we add the attention mechanism module in Figure 2. This attention module is first made up of an average pooling (AP) layer with which we obtain the feature vector $f_{AP} = AP(f_{Bayar})$. Secondly it is made up of a spatial feature extraction layer f_{ConvB} followed by an AP with which we obtain the feature vector $f_{sp} = AP(f_{ConvB} \circ f_{Bayar})$. A sum of f_{AP} and f_{sp} is produced in order to capture the two types of characteristics, then a $softmax(x)$ activation function is applied to the sum of f_{sp} and f_{AP} to allow the network to be more attentive to the most important regions. The obtained features are propagated on the Bayar layer in order to obtain the attention features k in (2).

$$f_{Att} = softmax(f_{AP} \oplus f_{sp}) \otimes f_{Bayar} \quad (2)$$

To enhance the model's ability to generalize and refine the feature selection, we employ a convolutional operation on the preceding layer. This operation utilizes 16 filters, each with a size of 3×3 . Subsequently, we apply a global average pooling (GAP) layer. The resulting feature vector is denoted by (3):

$$F_{Man} = GAP(conv \circ f_{Att}) \quad (3)$$

This module, responsible for extracting features from alteration traces, is detailed in Algorithm 2, which is called Algorithm 1. Algorithm 2 stands for artifact features extraction and Algorithm 1 is for the contained convolution layer.

Algorithm 1. Constrained convolution layer

```

Initialize randomly the weights  $w_k$ 
 $i = 1$ 
While ( $i \leq \max\_it$ ) {
  perform a feedforward pass
  Update the filter weights using stochastic gradient
  descent and backpropagate the errors
  For each  $k$  filters
    Define  $w_k(0,0)^{(1)} = 0$ 
    Normalize  $w_k$  so that
       $\sum_{l,m \neq 0} w_k(l,m)^{(1)} = 1$ 
    Define  $w_k(0,0)^{(1)} = -1$ 
  End for
   $i = i + 1$ 
  If the training accuracy converges, then Exit
}

```

Algorithm 2. Artifact featuresInput: $I \in \mathbb{R}^{L \times l \times c}$: an imageOutput: F_{man} : Manipulation traces features**Begin**Use Algorithm 1 to obtain f_{bayar}

Feature selection by attention mechanism

followed by Average Pooling: $f_{AP} = AP(f_{bayar})$ Select by Convolution and by Average Pooling": $f_{sp} = AP(f_{convB} \circ f_{bayar})$ Element-wise summation: $f_{att} = f_{AP} \oplus f_{sp}$ Normalization: $f_{att} = softmax(f_{att})$ Diffusion on f_{bayar} : $f_{att} = f_{att} \otimes f_{bayar}$ Refine the selection to obtain artifact features F_{man} $F_{man} = GAP(conv \circ f_{att})$ **End**

As for the spatial feature extraction as shown in Figure 1, the first layers learn low-level features such as edges, colors, and as the number of these layers increases, the feature learning becomes more accurate [26]. Setting up such a network is costly in terms of computing power and the size of the training data [27]. Another alternative is to apply transfer learning. Transfer learning is a machine learning technique that transfers knowledge acquired in one or more source tasks in order to use it to improve learning in a related target task [28]. Therefore, we use the pre-trained ResNet50 [27] model to obtain the spatial feature vector. The ResNet50 model is a CNN model composed of 50 layers. The architecture of the ResNet50 in Figure 3 model used is that of [27], except that we removed the last layer consisting of the average pooling (AP), the fully connected layer and the classification layer by the global average pooling (GAP) layer. This architecture is described as follows:

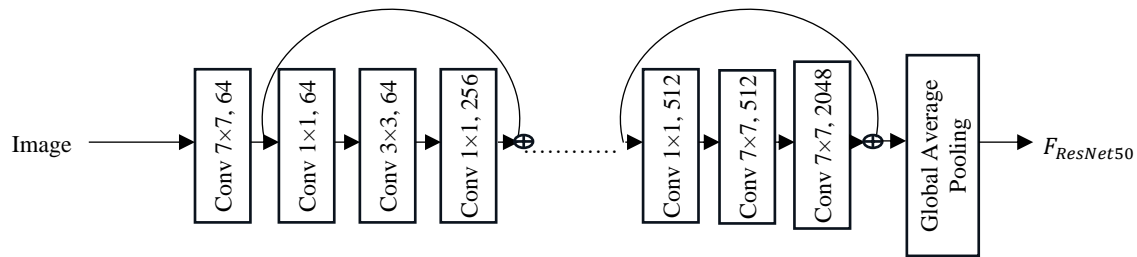


Figure 3. Transfer learn ResNet50 architecture

Let $F_{ResNet50} = ResNet50(I)$ be the feature vector obtained after application of ResNet50. Before the last layer which is the classification layer, we first pool the previous characteristics F_{Man} and $F_{ResNet50}$. Then a dense layer is added in order to learn the shared features. We finally obtain the following feature vector: $F_{Image} = \varphi(w(F_{Man} \oplus F_{ResNet50}))$ where φ is the rectified linear unit (ReLU) activation function and w the weights of the dense layer. As our problem is a binary classification, we use the sigmoid function for the distribution of predictions. At the output of our architecture, we obtain the prediction function (4):

$$\mathcal{F} = \sigma(F_{Image}) = \frac{1}{1 + e^{-F_{Image}}} \quad (4)$$

To allow our model to learn F_{Image} and allow it to improve in the prediction of \mathcal{F} , an error function ξ is calculated and minimized according to parameters θ . In this study, we adopt the cross-entropy-based error function. It is a function that measures the difference between the model's probability distribution and the predicted distribution. It is described as follows by (5):

$$\xi(\mathcal{F}; \theta) = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\mathcal{F}(I_i)) + (1 - y_i) \log(1 - \mathcal{F}(I_i))] \quad (5)$$

$y_i \in \{0,1\}$, N the number of observations, I_i the input characteristics of the i^{th} image and θ the classification parameters. The parameters θ are optimized by minimization of the error function ξ which gives by (6):

$$\hat{\theta} = \min_{\theta} \xi(\mathcal{F}; \theta) \quad (6)$$

2.2. Experimentation method

For implementing our models, we used an HP Core i7 computer with 16 GB of memory and a 64-bit operating system. We employed Python 3 and the following libraries: Pandas for transforming the dataset into a DataFrame, NumPy for matrix calculations, Matplotlib for data visualization, OpenCV for preprocessing raw images, and Keras with TensorFlow for designing and training deep learning models.

Regarding the dataset, we used royalty-free datasets commonly used in the literature for evaluating image manipulation models. The first is MediaEval [22], a dataset collected from Twitter as part of the automatic detection of the manipulation and misuse of multimedia content on the web. These manipulations include assembling, deleting, adding, and out-of-context images. Each entry in this dataset is accompanied by textual content, an image or video, and social context information. The dimensions of this dataset range from 100×100 pixels to 2709×3400 pixels. The second dataset is CASIA [29], which contains 7,491 genuine images and 5,123 tampered images. The falsified images in this dataset are real images manipulated first by preprocessing techniques such as cropping, distortion, and rotation, then by stitching operations and post-processing operations like blurring on edges or altered regions. Image dimensions in this dataset range from 320×240 pixels to 800×600 pixels.

In the data preprocessing, we reduced RGB images to 150×150 pixels. The OpenCV, Pillow, and NumPy libraries were used to read and digitize the images. We also removed duplicate images. All images were resized to a width of 150 pixels and a height of 150 pixels. The experiment was conducted by randomly dividing the datasets with 80% allocated for training data and 20% for test data. The training data were used for hyperparameter search and model selection. Performance-optimizing hyperparameters, such as the number and size of filters in convolution layers, dropout rate, batch size, and stride in pooling layers, were obtained using GridSearch from the Keras library on training data from [23]. We varied the batch size in the set {5, 15, 30, 32, 40, 50}, the number of filters in the set {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 16, 32}, the filter size in the set {3, 5, 7}, the dropout rate in the set {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}, the pooling size in the set {2, 3, 4, 5}, and the pooling step in the set {1, 2, 3}. The training process was stopped when the loss function converged, and the Adam optimizer was used. The parameters obtained were: a batch size of 32, a constrained convolution layer with 7 filters of size 5, pooling of size 3 with a step size of 2, and a second convolution layer with 16 filters of size 3. The final results were obtained by selecting the model that gave the best AUC in cross-validation with 5 folds.

To conduct a comparative evaluation of the performance of our proposed model, named BayarResnet, which utilizes a constrained convolution layer along with an attention mechanism and a transfer learning ResNet50 network, two additional models were trained and tested on the previously mentioned datasets, using the same parameters established during the data preprocessing step. The first is the SinghZamil model, as outlined in references [21], [22], which integrates ELA with the pre-trained EfficientB0 model to identify manipulated images. The second is BayarEff, which refers to the Bayar EfficientNet model, likely integrating the methodologies or enhancements proposed by Bayar and Stamm [23] within the EfficientNet architecture.

3. RESULTS AND DISCUSSION

3.1. Results

The results show that the proposed BayarResnet method is better than the others in terms of accuracy, precision, recall and specificity on both cassia and Medieval dataset in Tables 1 and 2. On the MediaEval dataset, the BayarResnet proposal gives better results for accuracy, precision, recall and F1-score and even on specificity. SingZamil model outperforms BayarResnet only on specificity when using CASIA Dataset.

Table 1. Performance of various approaches using MediaEval dataset

Models	Accuracy	Precision	Recall	Specificity	F1-score
SingZamil	0.716	0.702	0.722	0.71	0.712
BayarEff	0.851	0.903	0.777	0.921	0.835
BayarResnet	0.878	0.9354	0.805	0.947	0.865

Table 2. Performance of various approaches using CASIA dataset

Models	Accuracy	Precision	Recall	Specificity	F1-score
SingZamil	0.665	0.607	0.494	0.781	0.545
BayarEff	0.684	0.604	0.65	0.708	0.626
BayarResnet	0.705	0.623	0.696	0.711	0.657

The MediaEval dataset is trained with a proportion of 80% of the data over 30 epochs. Already at the 10th epoch during training, the BayarResnet and BayarEff models achieve over 90% higher accuracy than the other models. The training and validation accuracies of the BayarResnet and BayarEff models are better and increase with each epoch. But finally, BayarResnet performs well than BayarEff. This situation is illustrated by Figure 4.

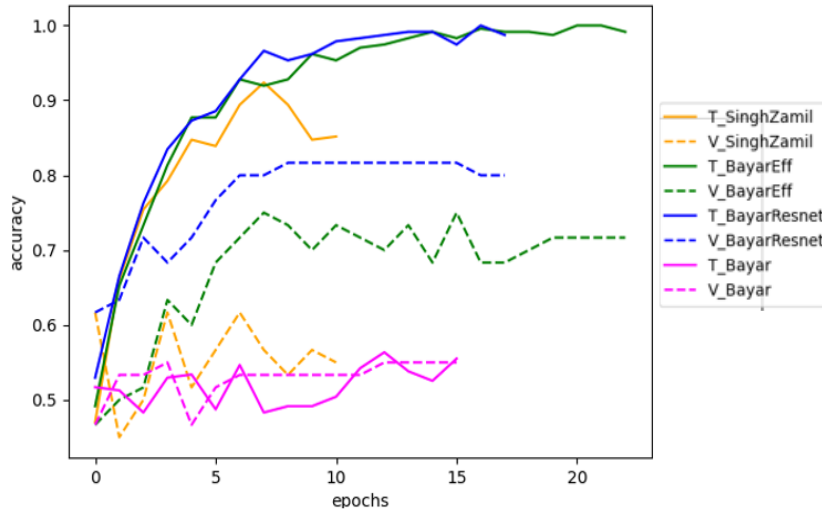


Figure 4. Training and validation accuracy

Concerning the loss values in Figure 5, overlearning is observed for the Bayar model from the 10th epoch and for the SinghZamil model from the 5th epoch. In addition, the validation and training loss values of the other two models (BayarResnet, BayaEff) decrease progressively towards zero and stabilize from the 15th epoch. The smallest loss values are observed with the BayarResnet proposal.

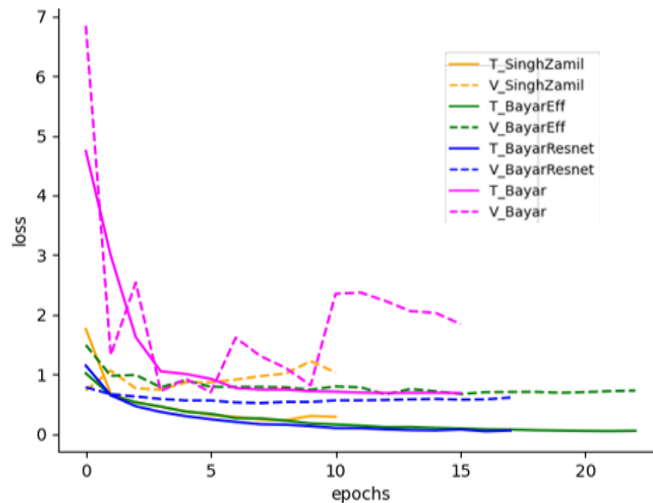


Figure 5. Loss curve during training and validation

Figure 6 displays the ROC curve achieved on the MediEval dataset, while Figure 7 depicts the curve for the CASIA dataset. The BayarResnet model exhibits the highest area under the curve, approximately 87% on the MediEval dataset and 70% on the CASIA dataset. In contrast, the SinghZamil model has the lowest respective areas under the curve, approximately 71% for MediEval and 63% for CASIA.

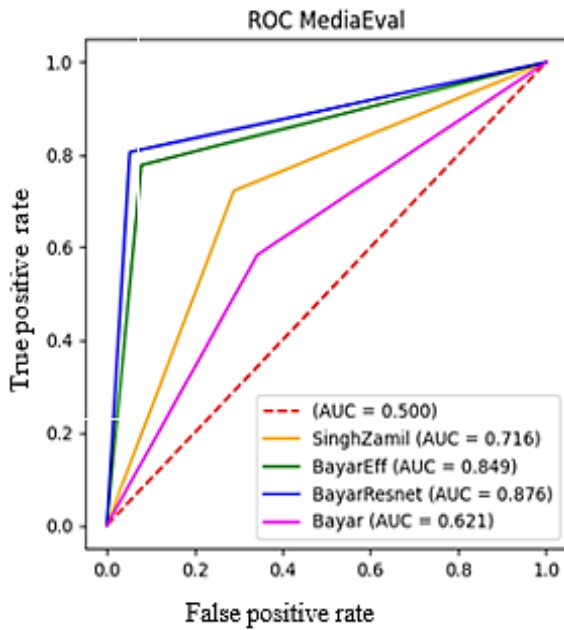


Figure 6. ROC curve on MediEVAL

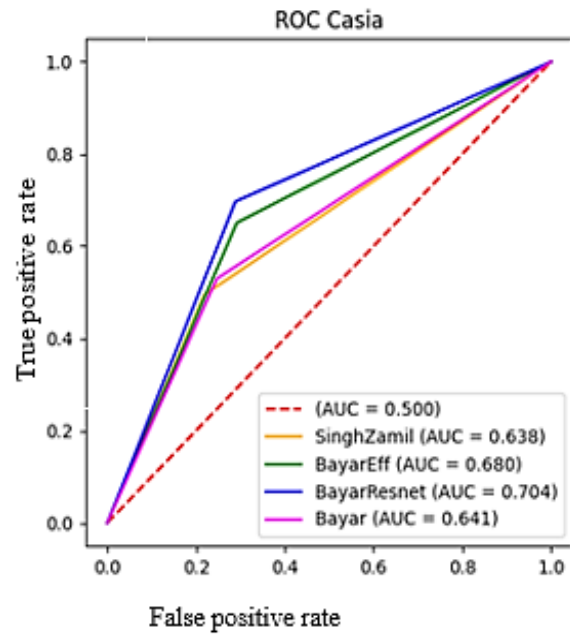


Figure 7. ROC curve on CASIA

3.2. Discussion

In this study, we conducted experiments to compare the performance of our proposed approach, based on the BayarResnet model, with recent approaches from the existing literature, such as BayarEfficient [23] and SinghZamil [21], [22]. As indicated in the results section in Tables 1 and 2, our proposed BayarResnet model demonstrated strong performance across two distinct datasets (MediEVAL and CASIA). We achieved higher accuracy and precision on both datasets compared to the alternative proposals. Thus, we can assert that our approach is well-suited for detecting forged images and exhibits greater potential for generalization. Analyzing the loss curve in Figure 5, we noted a consistent decrease in loss with the BayarResnet model, indicating its robust learning capacity when compared to BayarEfficient [23] and SinghZamil [21], [22]. Moreover, the loss curve of BayarResnet displayed stability with fewer fluctuations, suggesting an optimal learning rate. Conversely, the loss curve of the Bayar model, lacking the constrained layer found in BayarResnet, exhibited higher fluctuations in Figure 5. This suggests that the addition of the constrained layer in our proposal enhances its efficiency in identifying falsification artifacts. Upon examining the receiver operating characteristic (ROC) curve on the MediEVAL dataset in Figure 6 and the CASIA dataset in Figure 7, it became apparent that the area under the ROC curve of BayarResnet surpassed that of Bayar, BayarEfficient [23] and SinghZamil [21], [22]. This implies that BayarResnet yields fewer false positives compared to the other models. This improvement is predominantly attributed to the inclusion of the constraining layer and attention mechanism for detecting image falsification features.




4. CONCLUSION

To tackle the challenge of designing and training a deep neural network capable of autonomously learning features from various manipulations while minimizing false alarms, the BayarResnet model was introduced. This model integrates both noise residual features and global image features to identify falsified images. Its residue extraction module includes a convolutional constrained layer paired with an attention mechanism, enabling autonomous learning of falsification patterns. Meanwhile, the feature extraction module focuses on capturing spatial features across the entire image, utilizing convolutional neural network ResNet50 to effectively extract global features. The proposed model is evaluated through training and testing on two established datasets, demonstrating superior performance compared to recent models in the literature. Future research directions may explore new possibilities by incorporating multimodal information, such as text, video, and imagery, into fake news vectors, and investigating the adaptability of the proposed model in such scenarios.




REFERENCES

- [1] M. Cantarella, N. Fraccaroli, and R. Volpe, "Does fake news affect voting behaviour?," *Research Policy*, vol. 52, no. 1, Jan. 2023, doi: 10.1016/j.respol.2022.104628.
- [2] L. Zheng, Y. Zhang, and V. L. L. Thing, "A survey on image tampering and its detection in real-world photos," *Journal of Visual Communication and Image Representation*, vol. 58, pp. 380–399, Jan. 2019, doi: 10.1016/j.jvcir.2018.12.022.
- [3] A. Dixit and S. Bag, "A fast technique to detect copy-move image forgery with reflection and non-affine transformation attacks," *Expert Systems with Applications*, vol. 182, Nov. 2021, doi: 10.1016/j.eswa.2021.115282.
- [4] Y. Rao and J. Ni, "A deep learning approach to detection of splicing and copy-move forgeries in images," in *2016 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, Dec. 2016, pp. 1–6, doi: 10.1109/WIFS.2016.7823911.
- [5] D. K. Sharma, B. Singh, S. Agarwal, L. Garg, C. Kim, and K.-H. Jung, "A survey of detection and mitigation for fake images on social media platforms," *Applied Sciences*, vol. 13, no. 19, Oct. 2023, doi: 10.3390/app131910980.
- [6] F. W. R. Tokpa, B. H. Kamagaté, V. Monsan, and S. Oumtanaga, "Fake news detection in social media: Hybrid deep learning approaches," *Journal of Advances in Information Technology*, vol. 14, no. 3, pp. 606–615, 2023, doi: 10.12720/jait.14.3.606-615.
- [7] O. Ajao, D. Bhowmik, and S. Zargari, "Fake news identification on Twitter with hybrid CNN and RNN models," in *Proceedings of the 9th International Conference on Social Media and Society*, in SMSociety '18. ACM, Jul. 2018, doi: 10.1145/3217804.3217917.
- [8] K. Papat, S. Mukherjee, A. Yates, and G. Weikum, "DeClarE: Debunking fake news and false claims using evidence-aware deep learning," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2018, doi: 10.18653/v1/d18-1003.
- [9] N. Rani, P. Das, and A. K. Bhardwaj, "A hybrid deep learning model based on CNN-BiLSTM for rumor detection," in *2021 6th International Conference on Communication and Electronics Systems (ICCES)*, IEEE, Jul. 2021, pp. 1423–1427, doi: 10.1109/iccce51350.2021.9489214.
- [10] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, and L. Wei, "Detecting fake news by exploring the consistency of multimodal data," *Information Processing and Management*, vol. 58, Sep. 2021, doi: 10.1016/j.ipm.2021.102610.
- [11] Y. Li and Y. Xie, "Is a picture worth a thousand words? An empirical study of image content and social media engagement," *Journal of Marketing Research*, vol. 57, no. 1, pp. 1–19, Nov. 2019, doi: 10.1177/0022243719881113.
- [12] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 598–608, Mar. 2017, doi: 10.1109/tmm.2016.2617078.
- [13] J. Cao, P. Qi, Q. Sheng, T. Yang, J. Guo, and J. Li, "Exploring the role of visual content in fake news detection," in *Disinformation, Misinformation, and Fake News in Social Media*, Springer International Publishing, 2020, pp. 141–161, doi: 10.1007/978-3-030-42699-6_8.
- [14] A. Berthet, "Deep learning methods and advancements in digital image forensics," Ph.D. dissertation, Department of Computer Science, Telecommunications and Electronics, Sorbonne University, 2022.
- [15] N. Krawetz, "A picture's worth," *Hacker Factor Solutions*, 2007. Accessed: May 15, 2024. [Online]. Available: <https://www.blackhat.com/presentations/bh-usa-07/Krawetz/Whitepaper/bh-usa-07-krawetz-WP.pdf>
- [16] N. V. S. K. Vijayalakshmi K, J. Sasikala, and C. Shanmuganathan, "Copy-paste forgery detection using deep learning with error level analysis," *Multimedia Tools and Applications*, vol. 83, no. 2, pp. 3425–3449, May 2023, doi: 10.1007/s11042-023-15594-5.
- [17] K. H. Rhee, "Detection of spliced image forensics using texture analysis of median filter residual," *IEEE Access*, vol. 8, pp. 103374–103384, 2020, doi: 10.1109/access.2020.2999308.
- [18] T. S. Gunawan, S. A. M. Hanafiah, M. Kartiwi, N. Ismail, N. F. Za'bah, and A. N. Nordin, "Development of photo forensics algorithm by detecting photoshop manipulation using error level analysis," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 7, no. 1, Jul. 2017, doi: 10.11591/ijeecs.v7.i1.pp131-137.
- [19] X. Lin *et al.*, "Image manipulation detection by multiple tampering traces and edge artifact enhancement," *Pattern Recognition*, vol. 133, Jan. 2023, doi: 10.1016/j.patcog.2022.109026.
- [20] B. Singh and D. K. Sharma, "Predicting image credibility in fake news over social media using multi-modal approach," *Neural Computing and Applications*, vol. 34, no. 24, pp. 21503–21517, May 2021, doi: 10.1007/s00521-021-06086-4.
- [21] Y. K. Zamil and N. M. Charkari, "Combating fake news on social media: A fusion approach for improved detection and interpretability," *IEEE Access*, vol. 12, pp. 2074–2085, 2024, doi: 10.1109/access.2023.3342843.
- [22] C. Boididou, S. Papadopoulou, M. Zampoglou, L. Apostolidis, O. Papadopoulou, and Y. Kompatsiaris, "Detection and visualization of misleading content on Twitter," *International Journal of Multimedia Information Retrieval*, vol. 7, no. 1, pp. 71–86, Dec. 2017, doi: 10.1007/s13735-017-0143-x.
- [23] B. Bayar and M. C. Stamm, "Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2691–2706, Nov. 2018, doi: 10.1109/tifs.2018.2825953.
- [24] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11211, Springer International Publishing, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.
- [25] Q. Xu *et al.*, "Enhancing adaptive history reserving by spiking convolutional block attention module in recurrent neural networks," in *The Conference on Neural Information Processing Systems*, 2023.
- [26] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *Computer Vision – ECCV 2014*, Springer International Publishing, 2014, pp. 818–833, doi: 10.1007/978-3-319-10590-1_53.
- [27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
- [28] Y. Ma, S. Chen, S. Ermon, and D. B. Lobell, "Transfer learning in environmental remote sensing," *Remote Sensing of Environment*, vol. 301, Feb. 2024, doi: 10.1016/j.rse.2023.113924.
- [29] J. Gu, Y. Xu, J. Sun, and W. Liu, "Image tampering detection based on feature consistency attention," *International Journal of Information and Computer Security*, vol. 23, no. 1, pp. 1–15, 2024, doi: 10.1504/ijics.2024.136704.




BIOGRAPHIES OF AUTHORS

Kamagate Beman Hamidja    received his master degree in computer science 2014 at Unité de Formation et de Recherche en Sciences Fondamentales et Appliqués (UFR -SFA), Université Nangui Abrogoua, Ivory Coast. He received PhD degree in the same field at Institut National Polytechnique Houphouët-Boigny de Yamoussoukro (INPHB) in 2018. From that date to now he assistant Professor at Ecole Supérieure Africain des TIC (ESATIC) Abidjan, Treichville. He fields of research include network optimization, IoT, cybersecurity and applied artificial intelligent. He is a member of team Informatique et Sécurité des systèmes Numérique (I2SN) of Laboratoire des Sciences et Technologies de l'Information et de la Communication (LASTIC) of ESATIC. He can be contacted at email: beman.kamagate@esatic.edu.ci.






Fatoumata Wongbé Rosalie Tokpa    received his master degree in database and software engineering at Unité de Formation et de Recherche Mathématiques et Informatique (UFR-MI) at Université Félix Houphouët-Boigny (UFHB) Côte d'Ivoire, in 2019. Currently, she is a Ph.D. student in computer science at (UFHB). She is a member of the Mathematics and Computer Science Laboratory (LAMI) at this University, also a member of the Artificial Intelligence and Database Modelling team of Laboratoire de Recherche en Informatique et Télécommunication (LARIT) of Institut National Polytechnique Félix Houphouët-Boigny (INP-HB), Yamoussoukro. His research focuses on mathematical modelling machine learning and data veracity. She can be contacted at email: tokpafatou@gmail.com.



Vincent Monsan    obtained his Ph.D. in common sciences and techniques from the University of Rouen in France in 1994. During his doctorate, he worked on spectral estimation in the periodically correlated processes. He is currently a lecturer in statistic at Unité de Formation et de Recherche Mathématiques et Informatique (UFR-MI) at Université Félix Houphouët-Boigny (UFHB), Côte d'Ivoire. His research interests include Fourier coefficient, Periodic correlation, nonparametric estimation, and Mathematical statistics. He currently holds the position of vice-president at Université Felix Houphouët-Boigny, Abidjan, Ivory Coast. He can be contacted at email: vmonsan@yahoo.fr.



Souleymane Oumtanaga    is a current director of Unité d Mixte de Recherche en Mathématiques et Sciences du Numériques (UMRI MSN) at Institut National Polytechnique Houphouët-Boigny, Yamoussoukro (INPHB), Ivory Coast. He is also a member of the WACREN board and member of Scientifi Council of FONTSI (Fund for Science, Techonology and Innovation). He played a key role in the establishment of Internet domain of Ivory Cost. He received his Ph.D. in computer science at Université Paul Sabatier (Toulouse, France) in 1995. He is full professor in computer science at INP-HB since 2007. His research interests include network, IoT, cybersecurity and applied artificial intelligent. He can be contacted at email: oumtana@gmail.com.