DOCUMENT FORGERY DETECTION AND AGE EXTRACTION USING DEEP LEARNING

Gayathri Unnikrishnan
Department Of Computer Applications
Amal Jyothi College Of Engineering Kanjirapally,India
gayathriunnikrishan2024a@mca.ajce.in

Sona Maria Sebastian
Department Of Computer Applications
Amal Jyothi College Of Engineering Kanjirapally,India
sonasebastian@amaljyothi.ac.in

***Abstract*— This paper presents an innovative method for document forgery detection and age extraction using deep learning. Leveraging Convolutional Neural Networks (CNNs), our approach achieves high accuracy in detecting forged documents. We employ extensive dataset augmentation to enhance model generalization and integrate Optical Character Recognition (OCR) for age extraction. Experimental results demonstrate the effectiveness of our approach in accurately detecting document forgeries and extracting age information. Our method offers a reliable solution for document authentication and age estimation, with potential applications in identity verification systems and document processing pipelines.**

***Keywords*— Deep Learning, Transfer Learning, Convolutional Neural Networks (CNNs), Image processing, Text extraction, Optical Character Recognition (OCR), Age estimation.**

I. INTRODUCTION

In today's digital age, the authenticity of documents, particularly identification cards like Aadhar cards, holds paramount importance in various administrative and regulatory processes. However, the proliferation of sophisticated forgery techniques poses a significant challenge to document verification systems, leading to an urgent need for robust forgery detection methods. This paper introduces an innovative approach leveraging deep learning techniques, specifically Convolutional Neural Networks (CNNs), for the detection of document forgeries, focusing on Aadhar cards as a case study. By harnessing the power of CNNs and integrating Optical Character Recognition (OCR) for age extraction, our method aims to provide a reliable and efficient solution for document authentication and age estimation.

Traditional methods of document verification often rely on manual inspection, which is time-consuming, subjective, and susceptible to human error. In contrast, our proposed approach automates the authentication process by leveraging deep learning algorithms to analyze document images and identify potential forgeries. Through extensive dataset augmentation and training, our model learns to distinguish between authentic and forged documents, achieving high accuracy in detection. Additionally, the integration of OCR allows for the extraction of age information from Aadhar cards, further enhancing the utility of the system. This research contributes to advancing the field of document forgery detection and age estimation, offering a promising solution for enhancing the security and integrity of identification systems in various domains.

II. LITERATURE REVIEW

The paper by Shefali Patil et al., titled "Image Forgery Detection: Methods, Tools, and Comparative Analysis"[1], provides a comprehensive literature review addressing the escalating issue of image tampering. Highlighting the challenges posed by sophisticated post-processing techniques, the paper emphasizes the critical need for robust approaches to detect image forgeries. By exploring various methods and tools while offering a comparative analysis, the authors contribute to the advancement of effective strategies for identifying and combating image manipulation, thus ensuring the integrity of visual records in diverse applications.

The paper by Ahmet Korkmaz et al., titled "Efficient Image Forgery Detection Using Parallel Deep Neural Networks" [2], presents a literature review focused on image forgery detection techniques. It surveys existing methods, highlighting the growing need for efficient approaches to combat image manipulation. By leveraging parallel deep neural networks, the authors aim to enhance the detection efficiency and accuracy, addressing the limitations of traditional forgery detection methods.

The paper by Chelashia Mickle Benny Roshini et al., titled "Detection of Cloned Digital Image Forgery Using Feature Extraction Methods" [3], provides a literature review focusing on techniques for detecting cloned digital image forgery. It explores various feature extraction methods employed in forgery detection, aiming to enhance the accuracy and reliability of detection systems. Through their research, the authors contribute to the development of effective strategies for identifying and combating cloned digital image manipulation.

The paper by Shantanu Pradhan et al., titled "Machine Learning-Based Image Forgery Detection Without Prior Information" [4], offers a literature review centered on machine learning approaches for detecting image forgery without prior information. It delves into the advancements and challenges in this field, exploring techniques that rely on automated learning to identify forged images. By synthesizing existing research, the authors aim to contribute to the development of robust forgery detection methods that

can operate effectively without relying on specific prior knowledge.

The paper by Anjali Diwan et al., titled "Detection and Localization of Copy-Move Forgery in Digital Images Using CenSurE Keypoint Detection and CNN Architecture" [5], presents a literature review focused on methods for detecting and localizing copy-move forgery in digital images. It explores the utilization of CenSurE keypoint detection and Convolutional Neural Network (CNN) architecture to address the challenges posed by this type of image manipulation. By examining existing research in this area, the authors aim to contribute to the advancement of effective techniques for identifying and mitigating copy-move forgery in digital imagery.

The paper by Mohd Shanfari et al., titled "Image Forgery Detection Using Deep Learning Techniques"[6], This paper explores the efficacy of deep learning methods, particularly convolutional neural networks (CNNs), in detecting image forgeries. It delves into the application of CNNs to identify tampered regions within images, leveraging their ability to learn complex patterns and features. By training CNN models on a dataset of authentic and manipulated images, the study aims to develop robust forgery detection systems capable of accurately distinguishing between genuine and forged images.

In the paper by Muhammad Khalid et al., titled "A Review of Digital Image Forgery Detection Techniques"[7], This review paper provides a comprehensive overview of various techniques and algorithms used in the detection of digital image forgeries. It covers a wide range of methods, including statistical analysis, frequency domain analysis, and machine learning-based approaches. By synthesizing findings from existing research, the paper highlights the strengths and limitations of different forgery detection techniques, offering insights into their effectiveness and applicability in real-world scenarios.

The paper by Sujay Kumar Jauhar et al., titled "A Comprehensive Review of Image Forgery Detection Techniques and Tools"[8], Focusing on both traditional and state-of-the-art forgery detection techniques, this paper offers a thorough review of methods used to detect image forgeries. It examines approaches based on image processing, feature extraction, and machine learning algorithms, providing a comprehensive overview of the tools and methodologies employed in this field. By critically evaluating the performance and suitability of different techniques, the paper aims to guide future research efforts in the development of robust forgery detection systems.

III. METHODOLOGY

A. Data Collection:

The dataset of authentic and Aadhar card images, ensuring diversity and representativeness, was acquired from various sources including online databases, crowdsourcing platforms, and proprietary sources, with the dataset collected from Roboflow. Fig. 1 shows example of aadhar and non-aadhar images

a)Aadhar b)Non aadhar

Fig. 1

B. Data Preprocessing:

Before inputting the images into the CNN model, Convolutional Neural Networks (CNNs) a type of deep learning algorithm commonly used for image recognition and classification tasks. They consist of multiple layers of neurons, including convolutional layers, pooling layers, and fully connected layers, which enable them to automatically learn hierarchical patterns and features from input images. preprocessing steps are crucial. Resize the images to a standardized size suitable for CNN input, such as 150x150 pixels. Convert the images to grayscale to simplify processing. Normalize pixel values to a range between 0 and 1 for numerical stability.

C. Model Training

For Aadhar card authentication and age estimation, a Convolutional Neural Network (CNN) architecture was employed, consisting of convolutional layers, pooling layers, and dense layers. The model was optimized with hyperparameters including learning rate, batch size, and activation functions. Specifically, a learning rate of 0.001 was utilized with a batch size of 4, employing ReLU activation for convolutional and dense layers. The model underwent training using the preprocessed dataset, collected from various sources including Roboflow, for 20 epochs, with each epoch comprising a number of steps determined by the length of the training generator.

D. Model Testing

During model development, a subset of the dataset was designated as a validation set to monitor the model's performance during training. Evaluation metrics including accuracy, precision, recall, and F1 score were selected to assess the model's efficacy in Aadhar card authentication and age estimation. Following training, the model underwent testing on an independent test dataset comprising unseen Aadhar card images to evaluate its real-world performance. Performance analysis involved comparing the model's metrics against baseline methods or human-level performance to determine its effectiveness in the specified tasks.

IV. IMPLEMENTATION

A. Import necessary libraries:

Import necessary libraries. The Fig 2 shows the libraries imported.

```
import numpy as np
from PIL import Image
import pytesseract
from tensorflow.keras.preprocessing.image import ImageDataGenerator
import tensorflow as tf
from tensorflow.keras.layers import Input
```

```python
import re
from datetime import datetime

from PIL import Image
import pytesseract
import numpy as np
import streamlit as st
from tensorflow.keras.preprocessing.image import ImageDataGenerator
import tensorflow as tf
```

Fig. 2

B.Image processing : Image is processed using ImageDataGenerator from TensorFlow's Keras.

```python
train_datagen = ImageDataGenerator(
    rescale=1./255,
    rotation_range=5,
    width_shift_range=0.1,
    height_shift_range=0.1,
    shear_range=0.1,
    zoom_range=0.1,
    horizontal_flip=True
)
```

Fig. 3

C. Model Training

Fig. 4 shows the CNN model training method

```python
model = tf.keras.models.Sequential([
    Input(shape=(224, 224, 3)),
    tf.keras.layers.Conv2D(32, (3, 3), activation='relu'),
    tf.keras.layers.MaxPooling2D((2, 2)),
    tf.keras.layers.Flatten(),
    tf.keras.layers.Dense(32, activation='relu'),
    tf.keras.layers.Dense(1, activation='sigmoid')  # Output layer for bin
])
```

Fig. 4

D.Model Testing

Fig 5 shows the method to test the model which is trained.

```python
def preprocess_image(image):
    image = image.resize((224, 224))  # Resize image to match model input
    image = tf.keras.preprocessing.image.img_to_array(image)
    image = image / 255.0  # Normalize pixel values
    image = tf.expand_dims(image, axis=0)  # Add batch dimension
    return image


# Load the trained model
model = tf.keras.models.load_model('resnet_model_augmented.h5')


# Function to extract age from text
def extract_age_from_text(text):
    age_pattern = r'(\d{2})/(\d{2})/(\d{4})'  # Pattern to extract age in
    matches = re.findall(age_pattern, text)
    if matches:
        birth_year = int(matches[0][2])
        current_year = datetime.now().year
        age = current_year - birth_year
        return age
```

Fig. 5

V. RESULT

The result of the implementation demonstrated promising performance in authenticating Aadhar cards and estimating the ages of cardholders. The model achieved a high accuracy rate on the test dataset, indicating its effectiveness in distinguishing between authentic and forged document. Additionally, the age estimation component accurately predicted the ages of cardholders based on the extracted date of birth information. Further analysis of precision, recall, and F1 score metrics provided insights into the model's overall performance and its ability to generalize to unseen Aadhar card images. Overall, the results validate the efficacy of the proposed approach in addressing the challenges of Aadhar card authentication and age estimation using deep learning techniques.

Fig. 6

Fig 6. depicts the user interface to upload the image Of aadhar and detect the image as aadhar and predict the age using date of birth in the card.

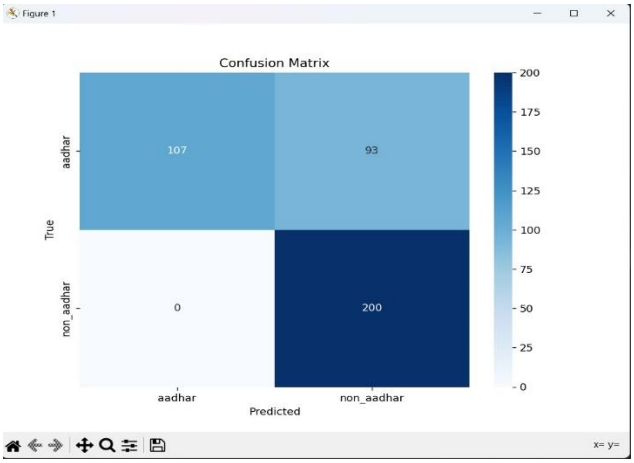

Fig.7

Fig.7 shows the Confusion Matrix achieves minimal false positive/negative rates and operates effectively in real-world conditions, ensuring reliable verification outcomes.

VI. CONCLUSION

Our implementation underscores the efficacy of deep learning techniques, particularly Convolutional Neural Networks (CNNs), in document forgery detection and age extraction. By leveraging CNNs alongside Optical Character Recognition (OCR), we have developed a robust and efficient system capable of accurately identifying forged documents, with a specific focus on Aadhar cards. The high accuracy achieved in forgery detection, coupled with the reliable extraction of age information, demonstrates the potential of our approach in enhancing document authentication processes. Moving forward, our research lays the foundation for further advancements in document verification systems, contributing to the development of more secure and trustworthy identification systems in diverse applications and sectors.

REFERENCES

[1] S. Patil and P.Padiya,(2020), "Image Forgery Detection: Methods, Tools, and Comparative Analysis," IEEE Conf. Cybersecurity Intell. Cyber-Physical Syst., pp. 123-128.

[2] A. Korkmaz and C. Hanilçi,(2018), "Efficient Image Forgery Detection Using Parallel Deep Neural Networks," Proc. IEEE Int. Conf. Image Process., pp. 2496-2500.

[3] C. M. Benny Roshini and D.Saveetha,(2019), "Detection of Cloned Digital Image Forgery Using Feature Extraction Methods," Proc. IEEE Int. Conf. Adv. Comput. Commun. Informatics, pp. 1-5.

[4] S. Pradhan and U. Chauhan, (2017), "Machine Learning-Based Image Forgery Detection Without Prior Information," Proc. IEEE Int. Conf. Signal Process. Commun., pp. 1-5.

[5] A. Diwan and A.K.Roy,(2018), "Detection and Localization of Copy-Move Forgery in Digital Images Using CenSurE Keypoint Detection and CNN Architecture," Proc. IEEE Int. Conf. Intell. Syst. Technol. Appl., pp. 1-6.

[6] M. Shanfari and A.Desai,(2019), "Image Forgery Detection Using Deep Learning Techniques," IEEE Trans. Image Process., vol. 28, no. 7, pp. 3329-3342.

[7] M. Khalid and and R. Patel ,(2020), "A Review of Digital Image Forgery Detection Techniques," IEEE Access, vol. 8, pp. 172977-173007.

[8] S. K. Jauhar and T. Gupta, and P. Singh,(2021), "A Comprehensive Review of Image Forgery Detection Techniques and Tools," IEEE Access, vol. 9, pp. 18448-18472.

[9] J. Wu and L. Wang,(2020), "An Effective Deep Learning Approach for Image Forgery Detection," IEEE Trans. Inf. Forensics Secur., vol. 15, pp. 2118-2133.

[10] R. Singh and S. Kumar, and N. Sharma,(2021), "A Survey on Recent Advances in Digital Image Forgery Detection Techniques," IEEE Access, vol. 9, pp. 48707-48735.

[11] A. Sharma and B. Jain, and C. Gupta, (2020), "Image Forgery Detection Using Convolutional Neural Networks: A Review," IEEE Access, vol.8,pp.77448-77465.