Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 2 (2024)

Machine Learning for Identifying and Validating Document Authenticity

Dr. Jhankar Moolchandani[1], Mrs. Rinki Pakshwa[2], Dr. Kulvinder Singh[3]

[1]*Assistant Professor, Department of CSE ITM University, Gwalior,*
[2]*Assistant Professor, Department of CSE ITM University, Gwalior,*
[3]*Assistant Professor, Department of computer science, Tantia University Sri Ganganagar,*

Abstract:- The aim of the study is to identify and validate information authenticity using machine learning (ML). The primary objective is to develop a robust model capable of accurately distinguishing between authentic and fraudulent documents. To achieve this, the study employs advanced techniques such as Res-Net 50, a deep learning architecture renowned for its image classification prowess, and SHA-256 cryptography, a secure hashing algorithm. The impressive outcomes of the model are showcased by its remarkable achievements, boasting a 99.26% accuracy level. This accomplishment is clearly reflected in both the confusion matrix and the classification report. This study underscores the potential of combining Res-Net 50 and SHA-256 cryptography in crafting a potent solution for identifying and validating document authenticity, with far-reaching implications for fraud detection and document verification processes.

Keywords*:* Document Authentication, Machine Learning, Res-Net 50, Cryptography, Fraud Detection.

1. Introduction

Document authenticity and integrity have become more important due to the widespread use of electronic communication and the growing digitalization of documents. A person's identity (name, age, residence, Identity Document (ID) number, etc.) could be established or confirmed using an ID. While some nations only accept informal forms of ID for identity verification, others offer only formal documentation [1-2]. Many kinds of identification documents include passports, national ID cards, residence cards, birth certificates, death certificates, driver's licenses, military IDs, and so on [3-4]. A person's identity could be verified using the picture on their ID. ID fraud is built on the foundation of a fictitious identity, often created with a combination of real data and fabricated information. For instance, the fraudster could (tomorrow) utilize a stolen Social Security Number (SSN), name, and address to construct a new identity. The criminal could then use this identity to seek credit, make large purchases, or engage in other actions that establish the identity as having a legitimate financial history. ID forging has become widespread in recent years because of the comparably inexpensive cost of high-quality computers, printers, and scanners. Illegal immigrants, pharmaceutical traffickers, human traffickers, and terrorists are some groups that benefit from the widespread use of counterfeit picture identification [5].

Nowadays, ID theft and fraud are major issues. It is on the rise as technology improves, and it is easier than ever to make a convincing fake ID with the help of cheap but high-quality printers, scanners, and computers. Traveling with a photo ID that could be used to verify identity and get entry to restricted places or installations is crucial. Fake picture IDs, like the number of illegal immigrants entering Europe, have a vibrant industry and widespread usage. They see migrant boats daily, carrying people who may or may not make it to Europe. Fake picture IDs facilitate drug trafficking, human trafficking, and acts of terrorism. For example, someone who is not an airport official can gain access to a restricted area with fake photo IDs. It follows that the Russian flight that went down half an hour after takeoff over the Sinai Peninsula was brought down like the Russian airliner had gone down: a terrorist gained access to a taxed aircraft and put a bomb. Every year, fraudulent IDs cause a negative financial crisis of around $750 million [6], affecting almost half a million individuals in the United States. The American Government Accountability Office discovered many fraudulent licenses in three states in 2012. The licensing authorities in those jurisdictions failed to notice that the applicants' birthdates did not match those on the licenses.

John F. Kennedy International Airport also uncovered about 4,585 phony passports. Canada spent millions of dollars securing its border with the USA after the September 11 attacks on the American Trade Centre, with counterfeit passports being its primary security issue [7-8].

The problem of Machine Learning (ML) for Identifying and Validating Document Authenticity revolves around developing effective and efficient algorithms and models that can automatically discern the authenticity of various documents, such as certificates, identification cards, financial records, and legal contracts. This involves designing a system that can accurately differentiate between genuine and forged documents, considering factors such as handwriting analysis, image processing, text extraction, and potential digital manipulation. The aim is to enhance security measures, reduce fraudulent activities, and provide a reliable solution for organizations and individuals to verify the legitimacy of important documents in a technologically advanced and increasingly digitized world [9]. The following are the research objectives are as follows:

- To create robust ML algorithms that can effectively analyze various document features such as text, images, signatures, and watermarks to accurately differentiate between authentic and forged documents.
- Design and implement novel ML algorithms, including deep learning architectures, ensemble methods, and anomaly detection techniques, tailored specifically for document authenticity assessment, considering both structured and unstructured data.
- Investigate the vulnerabilities of the proposed models to adversarial attacks, both in the digital and physical domains and devise strategies to enhance the robustness of the models against various manipulation attempts.
- Define appropriate evaluation metrics and benchmarks for assessing the performance of the developed models, considering factors like accuracy, precision, recall, F1-score, and handling imbalanced datasets.

The remainder of this paper is organized as follows. Section II reviews the related literature on ML-based identifying and validating document authenticity. Section III describes the technique which we have used in this paper. Section IV introduces the proposed work of identity document verification using the ML method.

A. Identity Theft and Identity Verification

Protecting the security, privacy, and accessibility of private information has become more difficult due to technological progress and increasing reliance on digital information exchange [10]. Identity theft, in which criminals assume other people's identities to do illegal acts, is a growing problem. Identity theft, defined as "the fraudulent use of an individual's personally identifiable information" [11], entails two distinct but related acts: (a) the illegal acquisition of personally identifiable information [12] and (b) the use of that information to establish a new, fictitious identity. With the rise in identity theft, the problem of authenticating and verifying identification paperwork has risen to the forefront [13]. Counterfeit identity cards have become more popular. When people provide their names, birthdates, places of birth, addresses, education levels, and occupations to verify their identities, they make a claim based on a wide range of credentials. However, these assertions alone are insufficient to authenticate identity; further proof is necessary to confirm that the identification document and the information inside are genuine and that the identity of the person being verified is confirmed [14]. Identity cards must have theft-resistant authentication techniques built in so that the actual and legitimate identity may be protected from those who would use a fake one to impersonate the holder.

Businesses use a wide variety of security and verification elements in ID cards, including tamper-proof laminates, holograms, and even ultraviolet ink and microprint [15-18]. The card's security measures can only confirm the card's legitimacy; they cannot confirm the cardholder's identity. If the card is to authenticate the cardholder's identification, it must connect in real time to a central repository that confirms the cardholder's right to possess the card [19].

2. Review of Literature

This section defines the previous studies of several authors built on ML for identifying and validating document authenticity.

Wu G. et al., (2018) [20] offered an innovative approach to providing a continuous authentication system by combining data from mobility and physiological sensors to form a multi-sensor synthesis. The authors used three lightweight algorithms for user motion recognition. Then, they used three distinct one-class algorithms in two distinct authentication contexts. The comprehensive tests to test the practicality and usefulness of the suggested authentication mechanism. Extensive studies were carried out to test the efficacy of the suggested method, which ended up with an F1 score of 86.67% and an average accuracy of 98.5%. The suggested technique seems to be a viable and workable authentication method based on the findings.

Ghanmi N. et al., (2018) [21] provided a more useful visual description for contrasting patterns. In contrast to most already used descriptors, the suggested descriptor, Grid-3CD, uses color and spatial information. This identifier is based on the quantified image's color-connected components (CC). Two pattern comparison methods demonstrate this descriptor's usefulness for checking identification documents. One is unsupervised and uses a distance measure as its basis, while the other is supervised and uses a Support Vector Machine (SVM) with a single class. The novel descriptor outperforms state-of-the-art descriptors in experiments conducted on four datasets, including 3,250 identification documents, demonstrating an average accuracy of roughly 90%.

Chinapas A. et al., (2019) [22] introduced face detection and face comparison to replace the need for a separate picture of the person's face to be included in the ID verification process. Face recognition software is built using Dlib, Facenet, and Arc-Face, all freely available online. The empirical evaluation demonstrates that the Arc-Face-based system achieves the maximum accuracy, with a detection rate of 99.06% and a comparison rate of 96.09%. Arc-Face beats competing approaches because it employs MTCNN and, in addition, align the facial picture in a straight line and fixes the eyebrow, eye, nose, and mouth locations such that all images have consistent frames of reference.

Castelblanco A. et al., (2020) [23] provided an ML-based pipeline to analyze images of documents in such cases, using several analytical modules and visual elements to confirm the document's kind and authenticity. Signing up for services, including banking, is becoming more common on mobile devices. In these procedures, individuals are often prompted to upload a photo of a government-issued ID to verify who they are. They test the methodology on official identification papers from the Republic of Colombia. The results showed that the ML background identification approach had a 98.4% success rate, while the authenticity classifier had a 97.7% success rate and an F1-score of 0.974.

Nasyrov N. et al., (2020) [24] described locating text document components with common design characteristics. The service's client-server interaction framework is described in detail, along with its simulated execution. Gradient boosting on decision trees is one example of an ML method. A multi-classification technique called Cat-Boost was selected. This method could isolate the elements of docx files who's formatting the classifier misunderstood. The classifier's results could be modified occasionally to improve the precision with which docx file components' formatting is checked.

Konlenko M. et al., (2021) [25] provided a novel model architecture for IDs using a CNN and a semantic segmentation technique. Simulation outputs in the form of numbers are used to quantify quality measures. The findings intersect with the union (IoU) value's threshold versus accuracy. For an IoU threshold value of 0.8, the study estimates an accuracy of more than 0.75. In addition, they determined the model's file size and demonstrated that it could be executed on a single-board microcomputer or smartphone from a commercial manufacturer.

Zhao L. et al., (2021) [26] developed a document forgeries method that cheaply uses current deep learning-based technologies to modify real-world document photos. Quantitative comparisons between the suggested technique and the existing procedure have shown the benefits of the design, including a reduction in the approximately 2/3 reconstruction error assessed in MSE, an improvement in reconstruction quality measured in PSNR of 4 dB, and a reduction in SSIM of 0.21. Qualitative testing has shown that the suggested method's reconstruction outcomes are visually superior to the current technique in complex characters and texture.

Ghadi M. et al., (2022) [27] suggested a method for authenticating identification documents that relies on detecting guilloche forgeries. The recommended technique consists of feature extraction and similarity measures between

a pair of ID feature vectors. The feature extraction process first learns their similarity using a Convolutional Neural Network (CNN) architecture to extract highly discriminative characteristics between identification papers. The findings demonstrate the effectiveness of the suggested method in extracting features from the processed identification papers that can be used to model the guilloche patterns and enable accurate discrimination.

Table 1 below summarizes the summary of the Review of Literature and the authors' process used in their studies.

Table 1. Comparison of the Related Work

Authors	**Technique Used**	**Outcomes**
Wu G. et al., (2018) [20]	ML	They concluded by suggesting an authentication system, which was subsequently put to the test with ten participants. Overall, the findings indicated an F1-score of 81.67% and an accuracy of 98.5%.
Ghanmi N. et al., (2018) [21]	SVM	However, the achieved outcomes fall short of expectations when applied to the actual world.
Chinapas A. et al., (2019) [22]	Dlib, Facenet, and Arc-Face.	The experiments demonstrate that Arc-Face offers the best overall answer, with an accuracy of 96%, since its straight face is better able to compare important features of the face than the other approaches.
Castelblanco A. et al., (2020) [23]	ML	The findings of this case study illustrate the validity of the approaches for use in their entirety during the enrolling procedure.
Nasyrov N. et al., (2020) [24]	ML	The accuracy of the elements' categorization was calculated to be 94.43% after using the stated method.
Konlenko M. et al., (2021) [25]	Semantic Segmentation and CNN	The primary result indicates that the suggested CNN produces respectable results. The computational architectures of CNN and deep neural layers (DNN layers) are simple to implement on contemporary hardware platforms like smartphones, microcontrollers, and industrial one-board microcomputers.
Zhao L. et al., (2021) [26]	DL	The experimental findings confirm that various post-processing techniques successfully preserve the coherence between distinct picture portions inside a text.
Ghadi M. et al., (2022) [27]	CNN	The findings validated the suggested method's effectiveness in accurately extracting the necessary features from the processed pair of identification documents to model the guilloche pattern.

3. Technique Used

In this section, we have used two techniques, i.e., SHA-256 and ML for Identifying and Validating Document Authenticity.

A. SHA-256

Document authenticity is verified using SHA-256, an advanced cryptographic technique from the Secure Hash Algorithm (SHA) family. SHA-256 creates a fixed-size, 256-bit hash result that uniquely reflects document content. To calculate this hash value, a sophisticated method comprises initializing a buffer with values, bitwise operations, modular additions, and logical functions on the incoming data. Importantly, the technique has the "avalanche effect," where even tiny document content changes change the hash result. SHA-256 gives each document a unique hash value to verify authenticity. This hash value is a digital fingerprint that condenses the document's information. If the document is unmodified, its hash value is consistent. Any change, no matter how tiny, changes the hash value. SHA-256 effectively detects even little document content manipulation due to this characteristic. The following Fig 1 below illustrates the working of SHA-256 for document authentication.

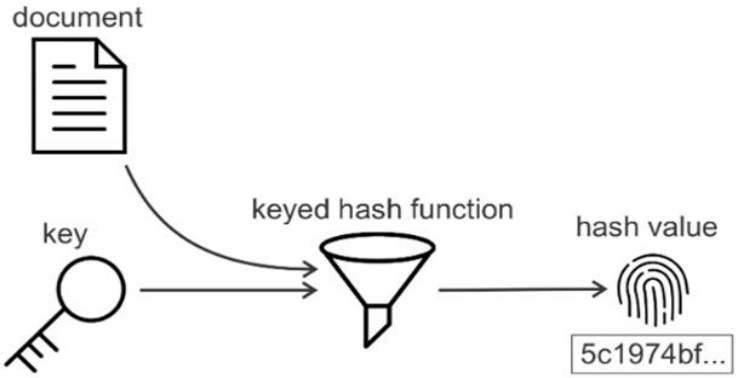


Fig 1. SHA-256 for document authentication [28].

The SHA-256 hash value of the original document is compared to the delivered document to verify its validity. If the two hash values match, the document has not changed since the initial hash was created. This comprehensive method verifies the document's integrity, guaranteeing no illegal changes.

B. Machine learning (ML)

ML has emerged as a valuable tool for identifying and validating document authenticity. ML techniques leverage patterns and features in document data to distinguish genuine documents from fraudulent or tampered ones. By analyzing various attributes such as text content, visual elements, and metadata, ML models can learn to make accurate decisions about a document's legitimacy. ML algorithms can be trained on a labeled dataset of authentic and fake documents, enabling them to generalize and detect subtle signs of manipulation or forgery. This technology finds applications in fraud detection, legal document validation, and secure data transmission.

- ResNet-50: ResNet-50, "Residual Network with 50 layers," is a well-known deep convolutional neural network design adept at challenging image identification problems. When it comes to determining whether a document is genuine, ResNet-50 is a crucial piece of the puzzle. ResNet-50 starts processing once a broad dataset of genuine and possibly fraudulent document pictures has been collected and pre-processed. A training set and a test set have been created from this dataset. Each document in the dataset is given its unique hash value by applying a secure hash algorithm to the dataset. These hash values are used as digital signatures to guarantee the authenticity of the associated documents. The following Fig 2 illustrates the architecture of ResNet-50 [29].

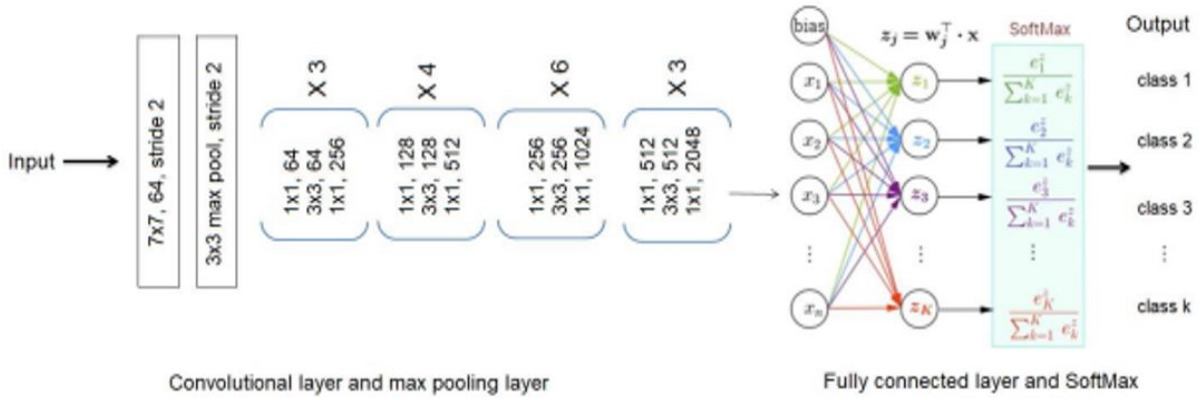


Fig 2: Architecture of ResNet-50

ResNet-50's deep design, which uses residual blocks, allows training very deep neural networks while avoiding vanishing gradient concerns. The ResNet-50 model is first pre-trained on a big-picture dataset to pick up on broad categories of features and regularities. The pre-trained ResNet-50 is then fine-tuned using the training dataset, where it learns to distinguish between real and counterfeit documents by analyzing their characteristic characteristics.

After the ResNet-50 model has been trained, its efficacy is measured. By comparing each document's predicted authenticity score to a threshold value, the model may make a binary prediction about whether the picture represents a genuine document. The performance of a model can be measured through various performance metrics that can evaluate how well a model can distinguish real documents from fake ones. In addition, the pre-processed hash values are crucial for document validation. To validate a hash, it must be compared to its original value. Any inconsistency suggests that the substance of the document has been altered.

4. Proposed Methodology

The proposed methodology for identifying and validating document authenticity follows a systematic approach involving several key steps. Initially, a diverse dataset containing authentic and fraudulent document images is collected. These images are then pre-processed, including resizing, normalization, and data augmentation to ensure consistency and increase dataset variability. The dataset is divided into training and test sets for model training and assessment. The methodology's core involves building and training a ResNet-50 deep learning model, initially pre-trained on a large dataset. The final classification layer of the ResNet-50 model is adapted to distinguish between authentic and fraudulent documents. During model training, techniques like mini-batch gradient descent and backpropagation are used, coupled with dropout and batch normalization strategies, to prevent overfitting.

The trained model is rigorously evaluated on the training dataset by assessing key metrics to determine its performance. Hyperparameter tuning is then conducted to optimize the model's parameters, including learning rate, batch size, and regularization, ensuring its effectiveness.

In the final stages, the ResNet-50 model is tested on the test dataset to simulate real-world scenarios. Upon successful testing, the model is deployed in a production environment. This involves setting up APIs to verify document authenticity enhancing user experience. Security measures are implemented to safeguard against adversarial attacks and unauthorized access, including encryption, secure communication protocols, and strong authentication mechanisms.

This methodology integrates cutting-edge deep learning techniques with robust security measures to create a comprehensive solution for identifying and validating document authenticity. It encompasses data preprocessing, model building, training, evaluation, tuning, and deployment, all while prioritizing data privacy, model robustness, and system security. The following Fig 3, given below, illustrates the flowchart of the proposed methodology.

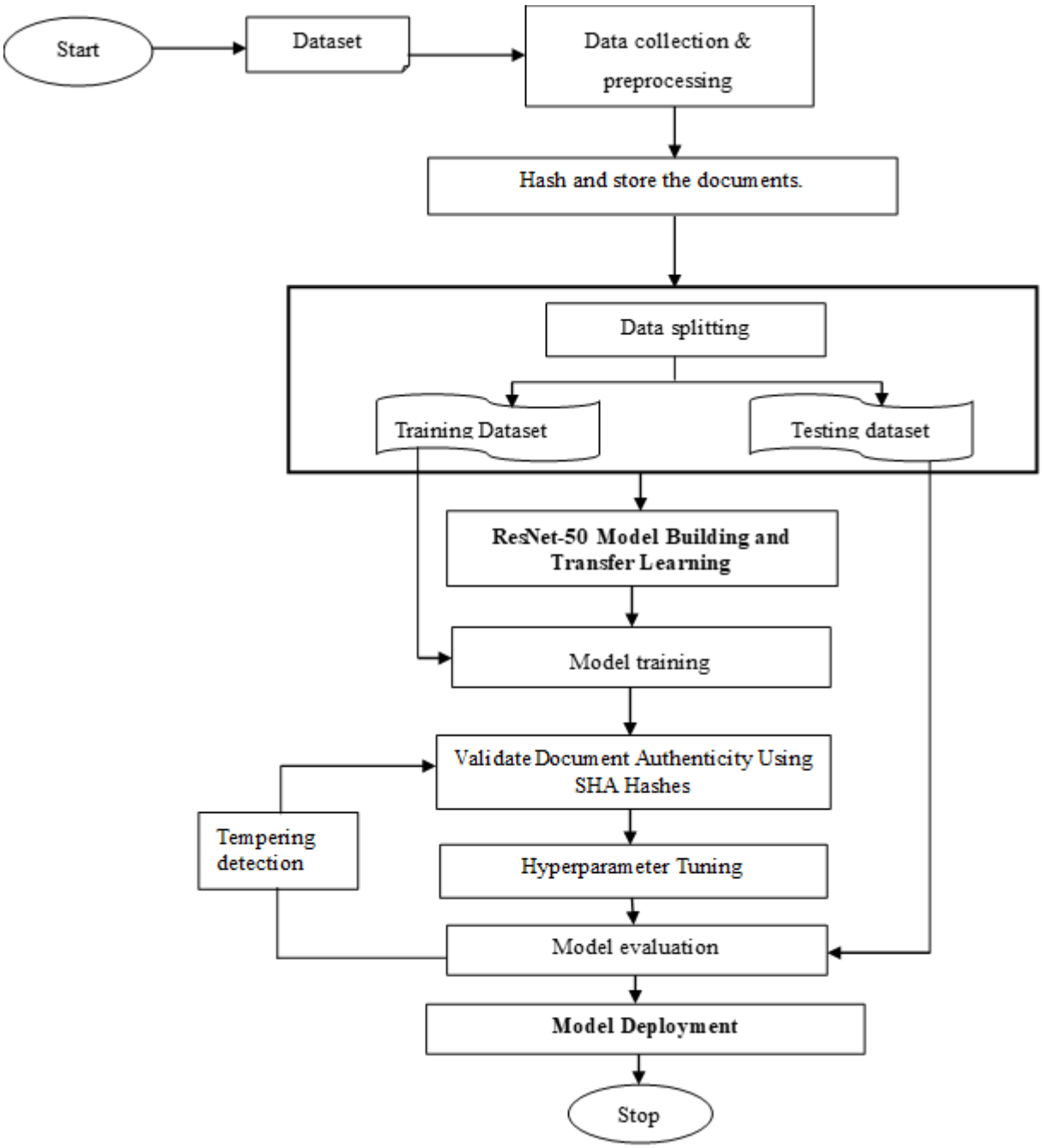


Fig 3. Proposed methodology

The following steps given below explain the above flowchart in detail:

1. Data Collection and Preprocessing:

- Collect a diverse dataset of authentic and fraudulent document images.
- Normalize pixel values, perform data augmentation (e.g., rotation, flipping), and resize photos to a standard size (224x224) to improve dataset diversity.

2. Hash and Store the Documents

- Initialize an empty list of hashed_documents to store tuples of documents and their hash values.
- Loop through each document in the dataset.

- Apply a secure hash algorithm (SHA) to compute the document's hash value.
- Append the document along with its hash value to the hashed_documents list.

3. Dataset Splitting:

- Split the dataset into training, validation, and testing sets, e.g., 70% training and 30% testing.

4. ResNet-50 Model Building and Transfer Learning:

- Load the pre-trained ResNet-50 model with weights trained on a large dataset.
- Set include_top to False to exclude the original classification layer from the model.

5. Model Training:

- Train the modified ResNet-50 model on the training dataset.
- Use techniques like mini-batch gradient descent and backpropagation.
- Apply techniques to prevent overfitting, such as dropout and batch normalization.

6. Validate Document Authenticity Using SHA Hashes

7. Hyperparameter Tuning:

- Perform hyperparameter tuning to optimize the model's performance.
- Tune learning rate, batch size, and regularization parameters.

8. Model Evaluation:

- Evaluate the trained ResNet-50 model on the training dataset.
- Calculate accuracy, precision, recall, F1-score, and ROC-AUC.

9. ReValidate Document Authenticity Using SHA Hashes

- Like Step 6, iterate through each document in the test set and its hash value, recalculating and comparing hash values to check for tampering.

10. Deploy the Model in a Production Environment

- Deploy the model in a production environment.
- Set up APIs for document authenticity verification.
- Implement security measures to protect the model from adversarial attacks

A. Proposed algorithm

ALGORITHM: Document Authenticity

Start

Phase I: Dataset Acquisition and Preprocessing

Step 1: dataset = load_and_preprocess_data()

Phase II: Hash and store the documents.

Step 2: hashed_documents = []

for the document in the dataset:

hash_value = secure_hash_algorithm(document)

hashed_documents.append((document, hash_value))

Phase III: Dataset Splitting

Step 3: X_trn, X_test, y_trn, y_test = split_dataset(dataset)

Phase IV: ResNet-50 Model Building

Step 4: base_model = ResNet50(weights=' ', include_top=False)

Step 5: Train the modified ResNet-50 model on the training dataset.

trn_model(X_trn, y_trn, model)

Phase V: Model evaluation

a. y_trn_pred = odel.predict(X_trn)

Step 6: y_trn_pred_binary = y_trn_pred > .5

Step 7: Calculate:

accuracy_trn = Accuracy_value (y_trn, y_trn_pred_binary)

precision_trn = precision_ value (y_trn, y_trn_pred_binary)

recall_trn = recall_ value (y_trn, y_trn_pred_binary)

f1_trn = f1_value (y_trn, y_trn_pred_binary)

roc_auc_trn = roc_auc_ value (y_trn, y_trn_pred)

Phase VI: Validate document authenticity using SHA hashes.

Step 8: for i, (document, hash_value) in enumerate (zip (X_trn, hashed_documents)):

if not secure_hash_algorithm(document) == hash_value:

printf ("Warning: Document {i} in the trainninhg set has been tampered with.")

Phase VII: Perform hyperparameter tuning.

Step 9: best_params = hyperparameter_tuning (X_trn, y_trn)

Phase VII: Evaluate the final ResNet-50 model on the testing dataset.

Step 10: y_test_pred = model.predict(X_test)

Phase VIII: Revalidate document authenticity using SHA hashes.

Phase IX: Deploy the model in a production environment.

Step 11: deploy_model(model)

Step 12: Implement security measures to protect the model from adversarial attacks.

implement_security_measures ()

End

5. Result and Discussion

A. Confusion matrix

The confusion matrix is a tabular representation that assists in comprehending the degree to which the model's predictions coincide with the actual labels of the documents. The confusion matrix that has been given in Fig 4 is a representation of the performance of a machine learning model that has been trained to recognize and confirm the validity of documents. "Authentic" and "Fraudulent" are the two distinct buckets that the model intends to place documents into when applied to this setting.

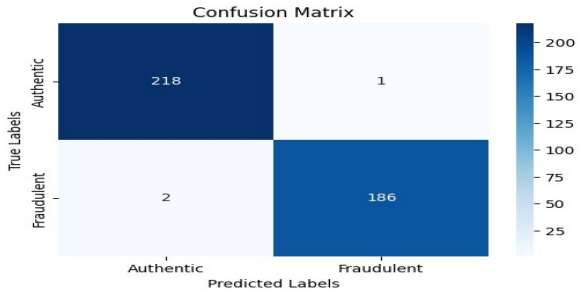


Fig 4. Confusion matrix.

• True Positives (TP): The total number of "Authentic" documents that were correctly classified. In the present instance, the model has successfully verified the authenticity of 218 papers.

• False Positives (FP): Incorrectly identified "Authentic" documents that are really "Fraudulent" documents. Only 1 instance of a fake document being wrongly identified as genuine exists in this matrix.

• True Negatives (TN): The total amount of "Fraudulent" documents that were properly labelled. A total of 186 fake papers have been recognized by the model.

• False Negatives (FN): Incorrectly labelled "Fraudulent" papers as "Authentic" ones. 2 authentic papers were incorrectly identified as fakes in this matrix.

The confusion matrix demonstrates that the model has attained an impressive accuracy of 0.99, which is quite impressive. 218 of the papers that were recognized as legitimate were properly identified, whereas just one of the documents that were misclassified as fake was identified correctly. In a similar manner, 186 of the counterfeit papers were accurately detected, while only two of them were incorrectly classified as legitimate.

B. Classification report

Based on the above confusion matrix, the various performance metrics for the machine learning model are calculated as follows:

• Accuracy: Accuracy represents the overall correctness of the model's predictions.

$$Accuracy = \frac{TP+TN}{Total} = \frac{218+186}{218+186+1+2} = 0.992628992629 \approx 99.26\%$$

• Precision: Precision measures the accuracy of the positive predictions made by the model.

$$Precision = \frac{TP}{TP+FP} = \frac{218}{218+1} = 0.995433789954 \approx 99.54\%$$

• Recall: Recall indicates the model's ability to correctly identify positive cases.

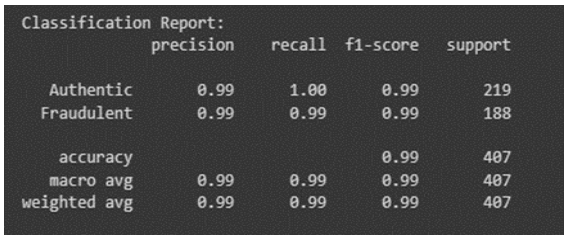
$$Recall = \frac{TP}{TP+FN} = \frac{218}{218+2} = 0.990909090909 \approx 99.09\%$$

• F1-Score: F1-score is the harmonic mean of precision and recall, providing a balanced performance metric.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} = 2 \times \frac{0.9954 \times 0.9909}{0.9954+0.9909} = 0.993144903 \approx 99.32\%$$

In order to evaluate the efficacy of the model, the value of accuracy, recall, and F1-score are calculated are shown in the classification report given in Fig 5. For genuine documents, the model achieves a remarkable 0.99 in accuracy and recall, showing that it is both very accurate in detecting real documents and adept at reducing false negatives. Legitimate papers also have a high F1-score, which considers both accuracy and recall. The model can efficiently identify counterfeit papers without producing an excessive number of false positives or false negatives, as seen by the accuracy, recall, and F1-score for these cases all being around 99%. The model's balanced performance in both legitimate and fraudulent categories is shown by a macro average (mean of accuracy, recall, and F1-score over both classes) that is continuously close to 99.3%. The model's overall performance is further

confirmed by the fact that the weighted average, which adjusts for class differences, likewise retains a high score of around 0.99.



Classification Report:

	precision	recall	f1-score	support
Authentic	0.99	1.00	0.99	219
Fraudulent	0.99	0.99	0.99	188
accuracy			0.99	407
macro avg	0.99	0.99	0.99	407
weighted avg	0.99	0.99	0.99	407

Fig 5. Classification report.

C. Accuracy

The proposed model demonstrates an accuracy of 0.992628992629 which is around 99%, indicating the model's strong ability to accurately identify document authenticity. Fig. 6 indicates the accuracy of the proposed model.

Fig 6. Accuracy of the proposed model.

Overall, the model seems to perform remarkably well based on the confusion matrix that was supplied, exhibiting excellent levels of accuracy, precision, recall, and F1-score. It is essential to indicate that the assessment of the model must consider the particular prerequisites and objectives of the document authenticity identification activity. The confusion matrix and classification report that was obtained demonstrate that the results obtained by the machine learning model designed for the purpose of determining and certifying document authenticity are quite encouraging.

6. Conclusion and Future Scope

In conclusion, the application of machine learning for identifying and validating document authenticity has shown remarkable success, as evidenced by the highly accurate results obtained from the confusion matrix and classification report. The model demonstrated exceptional precision, recall, and F1-score values for both authentic and fraudulent documents. With an accuracy rate of 99.26%, the model's ability to effectively differentiate between authentic and fraudulent documents is evident. This achievement holds significant implications for enhancing document verification processes, bolstering security measures, and mitigating instances of fraud.

The success of the proposed machine learning approach opens promising avenues for future research and development in the realm of document authenticity. One notable future scope lies in refining the model to handle more complex and diverse types of documents. Document formats, languages, and styles can vary widely, and expanding the model's capabilities to accommodate this variability would be beneficial. Moreover, the model's performance could be further improved by incorporating more advanced techniques such as deep learning and neural networks, which could capture intricate patterns and features in documents that might elude traditional machine learning approaches. In essence, while the current results are impressive, there is a rich landscape of opportunities to explore in the future. Advancements in machine learning techniques, coupled with a deep understanding of document verification challenges, can pave the way for more sophisticated, efficient, and robust systems for identifying and validating document authenticity.

Refrences

[1] "National Document Fraud Unit (UK Home Office), Guidance on examining identity documents, London," *Britain*, 2016.
[2] "National Document Fraud Unit (UK Home Office), Guidance on examining identity documents, London," *Britain*, 2015.
[3] Cifas, "ID documents report" Fraud Prevention. London: Britain, 2014.
[4] quifax and E. F. X., *The New Reality Synthetic ID Fraud*. Atlanta, Georgia, 2015.

[5] A. B. Hassan and Y. A. Fadlalla, "A survey on techniques of detecting identity documents forgery" in Sudan Conference on Computer Science and Information Technology (SCCSIT). IEEE, *2017*, pp. 1-5 [doi:10.1109/SCCSIT.2017.8293052].
[6] B. Perry et al., "Digital watermarks as a security feature for identity documents" in *Proc. SPIE Int. Soc. Opt. Eng.*, vol. 3973, pp. 80-87, 2000.
[7] Fake identity: Brits warned that their lives are in danger, The Independent, 2010. [Online]. Available: https://www.independent.co.uk/news/world/middle-east/fake-identity-brits-warned-that-their-lives-are-in-danger-1905971.html. [Accessed: March 1, 2023].
[8] A. Bashir and Y. A. Fadlalla, *'Techniques of detecting forgery in identity documents.' no. February*, 2018.
[9] Castelblanco, Alejandra, Jesus Solano, Christian Lopez, Esteban Rivera, Lizzy Tengana, and Martín Ochoa. "Machine learning techniques for identity document verification in uncontrolled environments: A case study." In *Pattern Recognition: 12th Mexican Conference, MCPR 2020, Morelia, Mexico, June 24–27, 2020, Proceedings 12*, pp. 271-281. Springer International Publishing, 2020.
[10] A. Asrani et al., "Review of network steganography techniques," *Imperial J. Interdiscip. Res. (IJIR)*. Mumbai: Thadomal Shahani Engineering College, vol. 2, no. 12, ISSN: 2454-1362, 2016.
[11] A. R. Gonzales and D. P. Majoras, "Combating identity theft: A strategic plan," *Of Fice President S Départ. Justice*, 2007.
[12] G. R. Gordon et al., "Identity Fraud: A critical national and global threat," *J. Econ. Crime Manag.*, pp. 1-48, 2004.
[13] A. Klenk et al., "Preventing identity theft with electronic identity cards and the trusted platform module" in *ACM, EUROSEC '09, Proc. Second European Workshop on System Security*. Nuremberg, Germany, Mar. 2009, pp. 44-51 [doi:10.1145/1519144.1519151].
[14] Zafarullah, J., R. Ashwanth, and S. A. Gowtham. "FACE COUNTERFEIT DETECTION IN IDENTITY CARDS USING IMAGE STEGANOGRAPHY."
[15] V. Martínez et al., "A comparative study of three Spanish eGoverment smart cards," *Log J. IGPL*, vol. 25, no. 1, pp. 42-53, August 2016, 2017.
[16] P. D. Di Lazzaro et al., "Invisible marking system by extreme ultraviolet radiation: The new frontier for anti-counterfeiting tags," *J. Instrum.* IOPScience Publishing Ltd/Sissa Medialab SRL, vol. 11, no. 7, 4th Intern Confer Frontiers in Diagnostics fix Technologies (ICFDT4), C07002-C07002, Jul. 2016 [doi:10.1088/1748-0221/11/07/C07002].
[17] D. Lushnikov et al., "Experimental study of the method of recording color volume security holograms on different photosensitive materials on the base of the diffuser with a narrow indicatrix of laser radiation," *Proc. SPIE* 10022, Holography, Diffractive Optics, and Applications VII. Beijing, China: October, p. 100221S, 2016.
[18] The Council of the European Union, "'PRADO Glossary (013) Technical terms related to security features and to security documents in general', Directorate-General Justice/ Home Affairs, Visas and Borders (DGD 1A) Brussels, Belgium," *Europe*, 2015.
[19] J. Blue et al., "Identity document authentication using steganographic techniques: The challenges of noise" in 28th Irish Signals and Systems Conference (ISSC). IEEE, *2017*, pp. 1-6 [doi:10.1109/ISSC.2017.7983646].
[20] G. Wu et al., "A continuous identity authentication scheme based on physiological and behavioral characteristics," *Sensors (Basel)*, vol. 18, no. 1, p. 179, 2018 [doi:10.3390/s18010179].
[21] N. Ghanmi and A. M. Awal, "A new descriptor for pattern matching: Application to identity document verification" in 13th IAPR international workshop on document analysis systems (DAS). IEEE, *2018*, pp. 375-380 [doi:10.1109/DAS.2018.74].
[22] A. Chinapas et al., "Personal verification system using ID card and face photo," *Int. J. Mach. Learn. Comput.*, vol. 9, no. 4, pp. 407-412, 2019 [doi:10.18178/ijmlc.2019.9.4.818].
[23] A. Castelblanco et al., "Machine learning techniques for identity document verification in uncontrolled environments: A case study" in *Pattern Recognit.*: 12th Mexican Conference, MCPR 2020, Morelia, Mexico, June 24-27, 2020, *Proceedings*, vol. 12. Springer International Publishing, 2020, pp. 271-281.
[24] N. Nasyrov et al., "Automated formatting verification technique of paperwork based on the gradient boosting on decision trees," *Procedia Comput. Sci.*, vol. 178, pp. 365-374, 2020 [doi:10.1016/j.procs.2020.11.038].
[25] M. Kozlenko et al., *Identity Documents Recognition and Detection Using Semantic Segmentation with Convolutional Neural Network*, 2021.
[26] L. Zhao et al., "Deep learning-based forgery attack on document images," *IEEE Trans. Image Process.*, vol. 30, pp. 7964-7979, 2021 [doi:10.1109/TIP.2021.3112048].
[27] M. Al-Ghadi et al., *'Identity Documents Authentication based on Forgery Detection of Guilloche Pattern.' arXiv Preprint ArXiv:2206.10989*, 2022.

[28] Available at: https://freecontent.manning.com/data-authentication-with-keyed-hashing/.
[29] R. S. Khudeyer and N. M. Almoosawi, "Combination of machine learning algorithms and Resnet50 for Arabic Handwritten Classification," *Informatica*, vol. 46, no. 9, 2023 [doi:10.31449/inf.v46i9.4375].