User Documentation Verification Portal

Vivek Sinha, Vaishnav Suguru, Vaibhav Prakash, Jyoti Malhotra

Computer Science & Engineering, MIT ADT University, Pune Maharashtra, India

Abstract - Any type of proof of identity must be shown to complete any legal formality or verification process. Regardless of the industry in which a client wants to participate, whether it is banking, insurance, healthcare, technology, travel, education, or any other online service, the consumer is required to show Identity proof to prove that they are who they claim to be. Verifying customer identities can help to lessen the risk of identity theft. While validating IDs sounds great, it will take money, effort, and human resources to adopt secure document verification solutions as a corporation. The overly demanding authentication process degrades the entire client experience. While most customers despise the lengthy verification procedure, not having one diminishes brand trust and can lead to businesses losing more clients throughout the onboarding process. This project will help the different organizations in detecting whether the Id provided to them by their employees or customers or anyone is original or not as it is a verification portal.

***Key Words*: Computer Vision, Document Verification portal, documents verification, Tampered, analyzer, docs verification.**

1. INTRODUCTION

We have recently witnessed the tremendous rate of globalization and its consequences for humanity. New technologies, such as the Internet and other networks, have both advantages and disadvantages. Speed, efficiency, better time management, access to an ocean of information, mobility, agility, automation, connectivity, remote sharing, better resource management, and the list goes on are just a few of the benefits of emerging technology. In a country like India, for example, Document Verification is a very hectic process. People need to stand in queue for the verification and the process is very slow. As also some people submit fake documents so the employees have to be very careful while checking Documents so it becomes very time-consuming, sometimes the process takes months and months. So to overcome this issue we are making our project.

Before we go into detail about how digital identity verification works, it's crucial to understand why organizations need to verify papers in the first place, whether using technology or by hand. Online document verification has two purposes: a: it protects against threats such as financial fraud and identity theft; b: it aids in industry-wide legal compliance.

According to research - 2020 Identity Fraud Study: Genesis of the Identity Fraud Crisis published on 07 April 2020 By Krista Tedder, and John Buzzard, the overall damage from identity fraud in 2020 is expected to be around $16.9 billion. Theft of identities causes more than simply financial losses; it also leads to sexual, racist, and gender-related comments on social media, which can bring a slew of problems.

The global epidemic of covid-19 has increased overall online transactions and digital banking, as well as the number of digital fraud and identity theft cases, causing more harm than good. Banks and financial institutions are investing heavily in AI-based identity verification systems to combat this form of fraud.

2. Literature Survey:

The section presents the existing work done for the verification system relating to multi-format document and a Generic Certificate Verification System

2.1 Multi-Format Document Verification System

In December 2020 a paper was published by Rajapakshe, Madura & Adnan, Muammar & Dissanayaka, Ashen & Guneratne, Dasith & Yapa Abeywardena, and Kavinga. (2020).

The author presents the widespread distribution of bogus documents purporting to be from official sources on social media has increased public distrust and doubt. Currently, there is no easily accessible technique of document verification that the general public may use. Using digital signatures and Blockchain, this study presents a mechanism for multi-format document verification. To sign document contents extracted using Optical Character Recognition (OCR) technologies, authors use digital signature algorithms and convert the signature into a 2D barcode format. This code can then be used to extract the digital signature of a shared document, and OCR can be used to validate the signature. Furthermore, the authors offer an alternate method of verification in the form of forgery detection systems. These signed documents are maintained in a blockchain-based decentralized storage solution, increasing the solution's overall reliability and security.

2.2 A Generic Certificate Verification System for Universities.

Obilikwu, Patrick, Usman, Karim, Dekera, Kenneth, and Kwaghtyo, Dekera published a paper in October 2019. (2019).

One of the most crucial documents for a graduate is the certificate granted by educational institutions. Certificate verification is required to ensure that the certificate holder is real and that the certificate is issued by a trusted source. On the other side, verifying certifications is demanding work for the verifier (the prospective employer who wants to verify the certificate). A Certificate Verification System for Institutions (CVSI) is being developed to address this issue, and it uses a Top-Down Design technique and an iterative paradigm. A NoSQL database (MongoDB) is used for certificate storage, while PHP is used for front-end design. Table 1 summarises the outline of the publications that were examined.

Understanding the importance of document verification for almost every official tasks; motivated us to choose this model to work with. While driving or traveling, we have to give our documents to authorized persons, and it takes a long time for them to verify them and return them to us. As a result, we decided to automate this task because we use it on a specific portal, but any organization can use it wherever they want, depending on their needs. Section 3 outlines the working architecture of the proposed model.

3. System architecture

This section highlights the working of the model "**doc**ument **V**erification **P**ortal" **- docVP** for user Document Verification Portal –uDVP. The uDVP accepts the documents and verifies their authenticity with the decision parameters (dp) as formulated in the equation (1) and equation (2).

$$\text{uDVP} \rightarrow \text{Doc}_{upload} + \text{Doc}_{verification} \qquad \text{Eq. 1)}$$

$$Verify(documents) = \begin{cases} valid, \ correct \ dp \\ invalid, \ otherwise \end{cases} \qquad \text{(Eq. 2)}$$

As outlined in figure 1, the model displays the User Interface that shows the front page of the website, which includes basic information about the website, a contact section, and a service section that offers a few document verification-related services.

The model works in three stages (refer to figure 2):
Stage I: Registration/Login
Stage II: Email verification and Document uploading
Stage III: Document verification against the valid document

For the new user, the work flow initiates with the registration process followed by the login process wherein authentic users can access the system with valid login credentials.

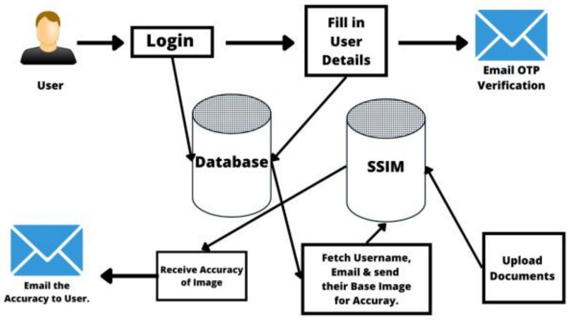

User
Login
Fill in User Details
Email OTP Verification
Database
SSIM
Email the Accuracy to User.
Receive Accuracy of Image
Fetch Username, Email & send their Base Image for Accuray.
Upload Documents

Figure 1: System architecture of docVP

Once a user has successfully logged in, the upload process of stage II starts as listed below –

- Basic details are entered into the system { First Name, Last Name, Contact No, Date of Birth }
- User enters the email id to get the OTP on their mail for mail verification.
- The user uploads the documents { PAN and/or Aadhar card } for verification

In Stage III, once the user uploads the document and submits it; the system performs the following tasks:

- Fetch the username and the verified email-id based on the input given
- The base image is pulled out and is compared with the uploaded image.
- The accuracy of the image is shared through email to the user.

4. Process of Document Verification

The method is based on notions of computer vision. Computer vision is an area of AI that teaches computers how to evaluate and comprehend images. Machines use digital images from cameras and webcams, as well as deep learning models, to reliably distinguish and classify objects, and then respond to what they "observe." Computer vision tasks include methods for getting, processing, analyzing, and comprehending digital images.

In image processing, computer vision is primarily concerned with enhancing or preparing raw input images for use in other applications. Computer vision is concerned with gathering data from input images or videos to comprehend them thoroughly and predicting visual input in the same way that the human brain does.

Is the greyscale conversion of the input and original photos required? Because many applications in image processing do not assist us in identifying the importance, edges of colored images and colored images are a bit complex to understand by machine because they have three channels while the grey

Table 1: Summary of the existing verification models

Author(s)	Method	Approach	Methodology
Obilikwu, Patrick & Usman, Karim & Dekera, Kenneth & Kwaghtyo, Dekera	A generic certificate verification system for Nigerian universities	Top-Down Design approach with iterative model	▪ Certificate verification is a difficult task for the verifier (the prospective employer who wants to verify the certificate). ▪ Certificate Verification System for Institutions (CVSI) is developed, which employs a Top-Down Design method and an iterative model. ▪ The system stores certificates in a NoSQL database (MongoDB) and uses PHP for the front-end design.
Samit Shivadekar	Document Validation and Verification System	Digital Signature Certificate (DSC)	▪ The system consists of a DigiVault [Digital Storage] website that may be linked to several government ministries. ▪ Documents generated are digitally signed and authenticated by the website's government authorities who are authorized to do so. ▪ Public Key Infrastructure is used to implement document digital signatures.
Rajapakshe, Madura & Adnan, Muammar & Dissanayaka, Ashen & Guneratne, Dasith & Yapa Abeywardena, Kavinga	Multi-Format Document Verification System	Digital signature algorithms	▪ Using digital signatures and blockchain, this study presents a mechanism for multi-format document verification. ▪ Digital signature algorithms are used to sign document, contents extracted using Optical Character Recognition (OCR) technologies ▪ The signatures are further converted to 20 barcode format
Shaimaa H shakir, Nour Zwyer	Forgery Detection Baped Image Processing i Techniques	(PCA) and (LDA)	▪ This work focuses on methods to detect digital forgeries based digital pixel properties in the grey level ▪ The max frequency intensity of plante is executed as the first method and the edge gradient as the second one to find some variance between the folgery and the original one then can be detected

scale has only one, In image processing inaccuracy, changing photos to greyscale is particularly advantageous.

Then we must compare the two photographs using the Structural Similarity Index (SSIM). As a result, the Structural Similarity Index (SSIM) is a perceptual metric for evaluating image quality degradation due to processing such as data compression or transmission losses. This metric is essentially a full reference that necessitates the use of two photos from the same shot, i.e. two graphically identical images to the naked eye. The second image is usually compressed or of a lower quality, which is the index's purpose.

Structural Similarity Index (SSIM) is most commonly associated with the video business, although it also has a large presence in photography. SSIM is used to determine the perceptual difference between two similar photos. It can't distinguish which of the two is better because it doesn't know which is the original and which has been processed further with compression or filters.

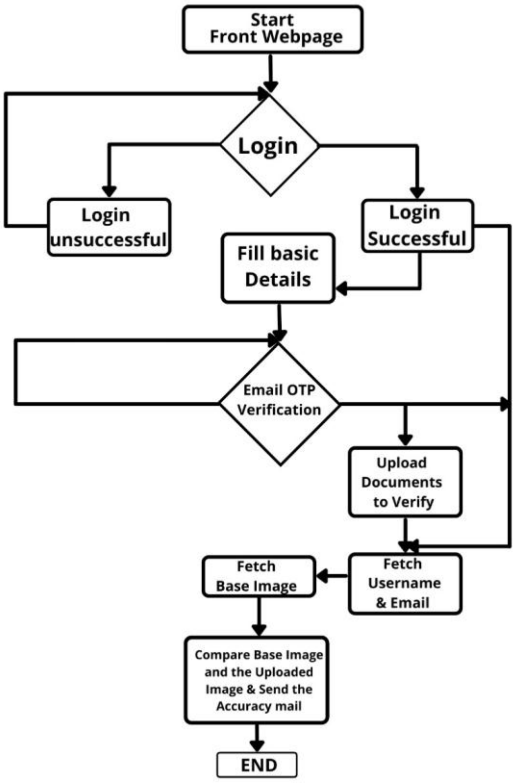


Figure -2: uDVP work-flow

5. METHODOLOGY:

We'll discuss what this model's requirements are in this part, its system architecture, and the required technologies to create it. There are several software used in the project, Python is the programming language used for programming. Front-end technologies and relevant frameworks were employed for the GUI (Graphical User Interface).

The model's working steps are as follows: (refer to figure 3)

Step 1 - Import necessary libraries
Step 2- Upload the original and modified documents from the website or local files on the computer.
Step 3- Reduce the size of the changed image by scaling down the source image.
Step 4 - Read original and tampered image
Step 5 - Converting an image into a grey scale image
Step 6 - To compare the two images, the Structural Similarity Index (SSIM) approach is used.
Step 7 - Calculate Threshold and contours
Step 8 - On images, users should see real-time contours and thresholds.

The SSIM index is calculated using several image windows. The distance between two N*N-sized windows x and y is:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

where,

- μ_x *the average of x*
- μ_y *the average of y*
- σ_x^2 *the variance of x*
- σ_y^2 *the variance of y*
- σ_{xy} *the covariance of x and y*
- $c_1 = (k_1L)^2, c_2 = (k_2L)^2$ *variables to stabilize the division*
- L the dynamic range of the pixel-values ($2^{\#bits/pixel}$ – 1)
- k_1=0.01 and k_2=0.03 (default)

The SSIM values range from 1 to +1, with 1 indicating that the two images are identical. A sliding window of size 11 11 and a circularly symmetric Gaussian weighting function are used in the usual approach for calculating SSIM. The maximal/minimal SSIM images synthesized along the equal-MSE hypersphere in the space of all images are shown in the following example. The MSE values of all images along the hypersphere are the same as the reference image, yet the perceived quality is substantially different.

Image quality assessment is critical in digital image processing applications. In this work, the metrics (MSE, PSNR, SSIM, and FSIM) are used to determine the optimal quality metric. To mimic trials with Gaussian noise, we employed the Gaussian filtering approach. The image quality obtained was assessed using the metrics listed above (refer to figure 4)

6. WORKING IMAGE OF THE MODEL

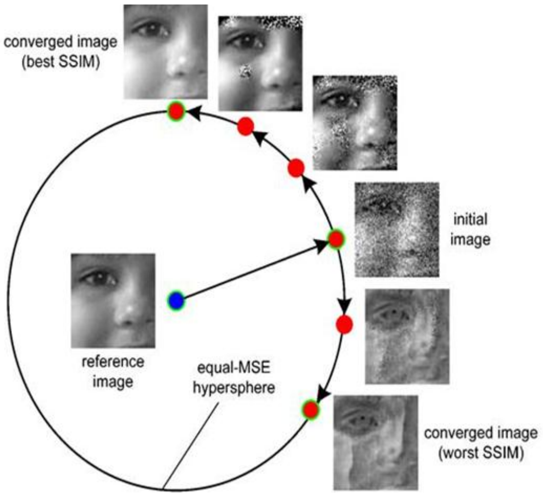


Figure 3: Working steps of SSIM

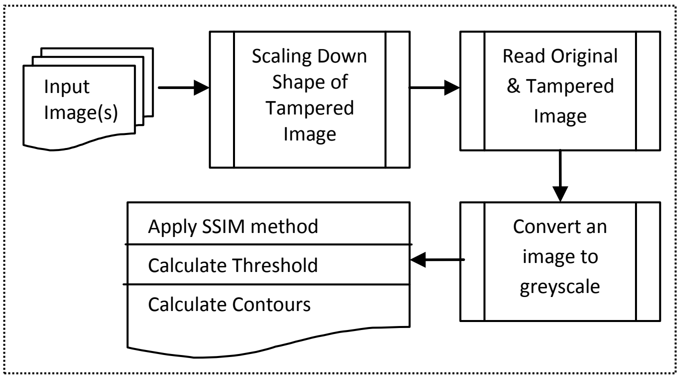
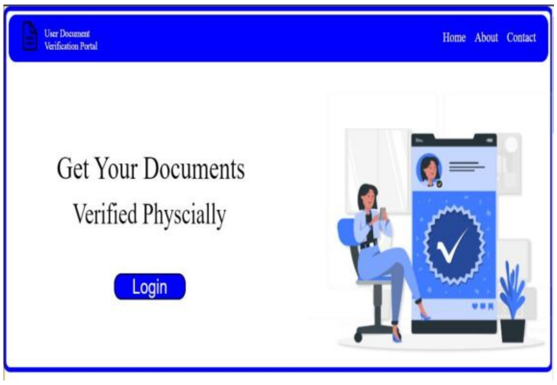

Input Image(s)
Scaling Down Shape of Tampered Image
Read Original & Tampered Image
Convert an image to greyscale
Apply SSIM method
Calculate Threshold
Calculate Contours

Figure 4: Working steps of the model

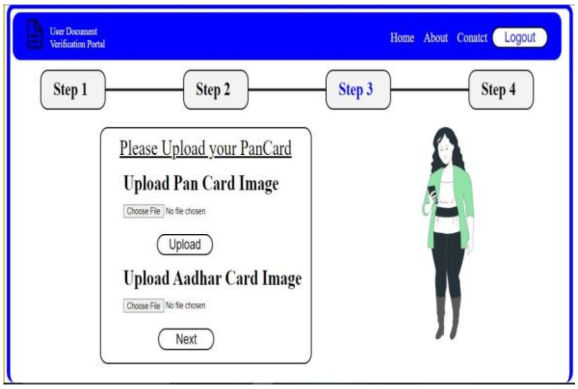
7. PERFORMANCE OUTCOME

The discussed system in the form of 3 activities is presented in the figure 5(a), 5(b), and 5(c)



Get Your Documents Verified Physcially

Figure 5(a) : Welcome screen

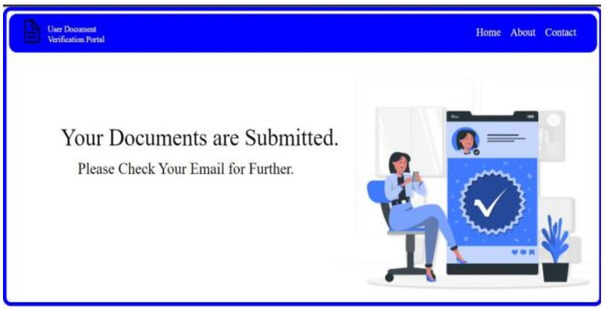

User Document Verification Portal
Home About Contact Logout
Step 1
Step 2
Step 3
Step 4
Choose File No file chosen
Upload
Choose File No file chosen
Next

Please Upload your PanCard

Upload Pan Card Image

Upload Aadhar Card Image

Figure 5(b) : Upload documents



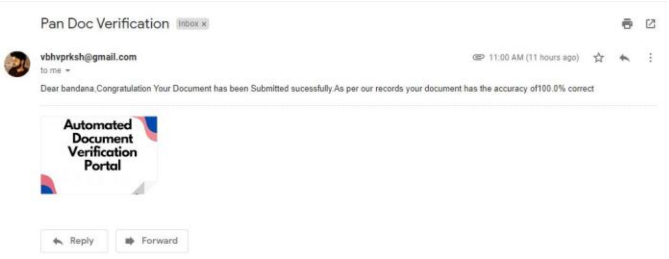
Your Documents are Submitted.

Please Check Your Email for Further.

Figure 5(c) : Upload confirmation

OUTCOME

By using the formulas of SSIM and the methodology as described above in the mathematical model and working image we got the output in percentage value as shown in the figure 6(a). The accuracy is generated through SSIM methodology and that value will be sent in the mail to the user.



Pan Doc Verification

vbhvprksh@gmail.com

to me

11:00 AM (11 hours ago)

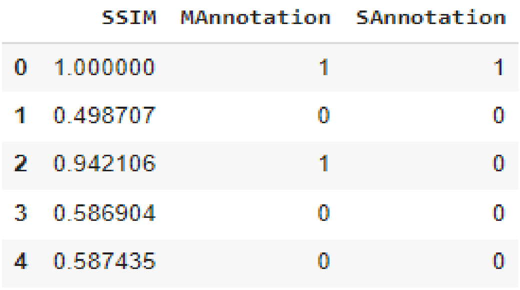
Dear bandana,Congratulation Your Document has been Submitted sucessfully.As per our records your document has the accuracy of100.0% correct

Automated Document Verification Portal

Figure 6(a) : Verification accuracy

To test the model, we created a manual dataset with two annotation values – **S Annotation** and **M Annotation.** (refer to figure 6(b))

S Annotation - Actual Value; **M Annotation** – Predicted Value



	SSIM	MAnnotation	SAnnotation
0	1.000000	1	1
1	0.498707	0	0
2	0.942106	1	0
3	0.586904	0	0
4	0.587435	0	0

Figure 6(b) : Annotations

By using these two annotation values we generated a confusion matrix (refer to figure 6(c)). The confusion matrix produced four results, each representing a distinct mix of Actual and Predicted values.

TP: True Positive: Values that were both truly positive and projected to be positive.

FP: False Positive: Values that were genuinely negative but were incorrectly expected to be positive. Type I Error is another name for it.

FN: False Negative: Positive readings that were incorrectly forecasted as negative. Type II Error is another name for it.

TN: True Negative: Values that were both genuinely negative and projected to be negative.

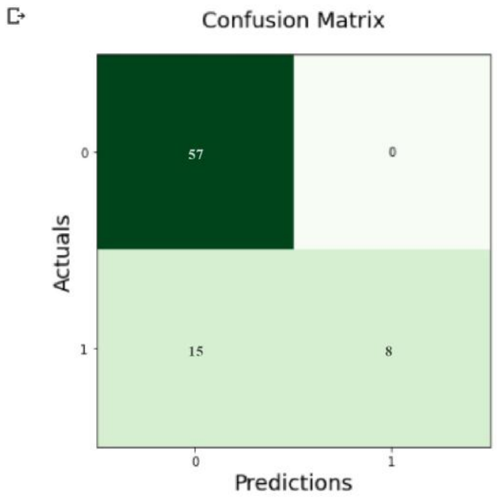


Figure 6(c): Confusion matrix

When the value in the confusion matrix is [1,1] it is said to be True Positives. Similarly, when the value is [0,0] it is True Negatives, when the value is [0,1] it is False Positives and when the value is [1,0] it is True Negatives. As it is observed from the figure 6(c), docVP model produces more true positives.

The discussed model can be enhanced further, as presently, we need to have a base image and SSIM needs the image structure, pixels, and luminous to be the same when we compare two images.

8. CONCLUSION

The working of this system is mainly based on Computer Vision and SSIM (Structural Similarity Index). The outcome of the document verification is generated according to the document uploaded by the user. The document is being compared to the original document submitted by the admin then the document uploaded by the user is being compared to the original picture to get the accuracy and if the document which is being uploaded by the user has accuracy greater than 70% then the document gets verified.

9. REFERENCES

[1] Mahdian, B., and Saic, S.: 'Blind methods for detecting image fakery',IEEE Aerospace and Electronic Systems Magazine, 2010, 25, (4), pp.

[2] Maintz, T.: 'Digital and medical image processing', Universiteit Utrecht,2005

[3] Farid, H.; Lyu, S. Higher-order Wavelet Statistics and their Application to Digital Forensics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Madison, WI, USA, 6–22 June 2003; p. 94.

[4] Thakur, R.; Rohilla, R. Recent advances in digital image manipulation detection techniques: A brief review. Forensic Sci. Int. 2020, 312, 110311. [CrossRef]

[5] Li, J.; Li, X.; Yang, B.; Sun, X. Segmentation-Based Image Copy-Move Forgery Detection Scheme. IEEE Trans.Inf. Forensics Secur. 2015, 10, 507–518.

[6] Y. Wu, X. Kong, and X. You, "Printer forensics based on documents geometric distortion," in Proc. 2009 16th IEEE Int. Conf. Image Processing (ICIP 2009), pp. 2909–2912, IEEE, Piscataway, New Jersey (2009).

[7] C. Schulze et al., "Using DCT features for printing technique and copy detection," Adv. Digital Forensics 306, 95–106 (2009).

[8] D. Lee and S. Lee, "A new methodology for grey-scale character segmentation and recognition," in Proc. Int. Doc.

[9]C. H. Lampert and T. M. Breuel, "Printer technique classification for document counterfeit detection," in Int. Conf. Comput. Intell. Secur.ICCIAS, pp. 639–644, IEEE, Piscataway, New Jersey (2006).

[10]R. Bertrand et al., "A system based on intrinsic features for fraudulent document detection," in Proc. Int. Conf. Document Analysis and Recognition, ICDAR, pp. 106–110,IEEE, Piscataway, New Jersey(2013).