Proposing a Blockchain-based Solution to Verify the Integrity of Hardcopy Documents

Sthembile Mthethwa, Nelisiwe Dlamini, Dr. Graham Barbour
Council for Scientific and Industrial Research (CSIR)
Modelling and Digital Science Unit, Information Security
Pretoria, South Africa
{smthethwa, ndlamini2, gbarbour}@csir.co.za

Abstract—**Even with the ability to produce documents digitally, the paperless environment has yet to become a reality in South Africa. Hardcopy documents are still printed daily which makes them susceptible to document fraud. In South Africa, a case was reported recently, where someone who was creating fake documents was exposed. This introduces the challenge when using hardcopy documents which is loss of integrity. Thus, it is vital to have systems in place to verify document integrity and be able to determine when a document has been tampered with. Various techniques have been used to secure documents, yet the challenge persists. The combination of 2D barcodes, digital signatures, Optical Character Recognition (OCR), cryptographic hashing has proved the potential to achieve good results when combined. Recently, blockchain has been added as one of the techniques to be employed for document verification. This paper presents a proposed solution that incorporates the combination of 2D barcodes, OCR, cryptographic hashing and blockchain. As this is still on-going work, experiments are still required to demonstrate the viability of the solution.**

Keywords—blockchain, 2D barcodes, cryptographic hashing, digital signatures, document generation and validation, document verification, integrity, optical character recognition (OCR), secure hash algorithm (SHA-256), tesseract

I. INTRODUCTION

Nowadays digital documents have become a large part of every sector whether public or private resulting from the transformation of modern technology. This not only allows the dissemination of information but also preserves these documents in digital form and promotes paperless environments. However, a variety of these documents are printed every day such as; academic certificates, wills, case files, birth and marriage certificates, national identity documents, insurance documents, passports and drivers' licenses, etc. and issued to different people because printed documents are the most prevalent form of trusted communication. A great challenge is therefore experienced along with the advancement in technology; which now make it possible to easily reproduce falsified documents as they are now susceptible to unauthorized alterations [1, 2].

A series of forgery cases have been reported in the past. In India, a woman who was applying for a passport presented a falsified birth certificate, and it took some time for the concerned agency to detect the forgery. The police reported this as a common occurrence when it comes to applicants [5]. Incidents of falsified academic documents have also been reported, for instance, in Singapore, where foreign nationals were convicted after being charged with forgery of academic documents [4]. Additionally, in South Africa, two people presented fraudulent asylum documents at their workplace which they claimed were received from a state's agency official [3]. Recently a man, who produced unauthorized pay slips and bank statement and sold them to people, was exposed [6]. The counterfeiting of these documents enabled people to get credit and the credit companies are unable to trace these people after that, which has cost them a lot of money [6]. Such cases indicate the seriousness of hardcopy document fraud and demand a need to augment the solutions that solve this problem by introducing numerous methods to secure the original document from being forged in any way.

Luckily, this downside in the security of hardcopy documents has attracted attention from many researchers, and research in this area is advancing, with the aim to alleviate unauthorized altering and compromised integrity of these documents and the use of falsified documents. This entails proving that a scanned/physical copy of a document is the same as its original document [1, 2]. In doing this, various methods are implemented to record information pertaining to the original document, called the original template; and encode this information in a barcode then insert the barcode into the document and print it. Upon presentation of a copy of the original document, recognition techniques attempt to extract this information from the copy, producing a copy template. The two templates are then compared. Two key issues arise; first is the problem of storing the original template content, and secondly, the problem of extracting the copy template.

While the original template can be stored in a database, an ideal solution would be to store the original template as visible information on the original document itself. To supplement this, there are methods that are used to add information to the original document to ensure that a copy is not tampered with in any way. Amongst these methods are watermarks, document signatures, barcodes, hashing, etc. However, information stored using these methods is limited, for instance all document content cannot be included in a barcode because of the space limitations. Rather than storing the original template on the document, a hash is stored instead [2], and the problem of extracting the stored template from the copy emerges as a hash maps the document content to a string that represents the content, it does not contain the actual content in the original document.

Despite this, the use of hash values in blockchain-based methods, is undoubtedly a reliable solution that is now applied in document verification systems [7]. The blockchain which is

978-1-5386-6477-3/18/$31.00 ©2018 IEEE

a distributed, replicated and synchronized public ledger has made it possible to implement solutions that validate the integrity of the documents issued. However; more work still remains, the transition from eliminating the use of hardcopy documents to using digital documents hasn't been successfully navigated as yet considering that hardcopy documents are still widely used. In this paper we use the standard Optical Character Recognition (OCR) technique and incorporate the use of blockchain technology to present an agnostic solution that focuses on the ability to verify the integrity of both digital and hardcopy documents.

This paper is organized as follows, in Section 2, we present the literature review. Then a discussion of the proposed solution is presented in Section 3 and Section 4, concludes the study.

II. Literature Review

Over the years, technology has made it very easy to produce digital documents which are easy to retrieve, access and store, and encourages the change to more paperless environments. However; printed documents are still predominant and used to serve the purpose of communicating relevant information to people even though these documents have been perceived as cumbersome and inefficient [8, 9]. With the advancement of technology, the demand to verify important hardcopy documents has escalated, as the issue of fraudulent documents continues to aggravate. Falsifying a hardcopy document requires less effort these days, because these documents are inherently insecure and most have no passwords or digital signatures unlike digital documents. A long list of techniques have been proposed to mitigate the problem of forged documents mostly digital documents. But research in the area of securing hardcopy documents is starting to gain a lot of traction, since the use of these documents has become undeniable [1].

Some of the prevalent techniques include the use of watermarking, which aims to preserve the integrity of a document. Watermarking can either be in a digital or printed format [1]. This technique is still vulnerable to attacks, which may not necessarily remove the watermark imprinted, but rather disable its readability, the success of watermarks also relies on high quality printers, which incurs cost [8]. Nevertheless, it remains an active research area and continues to be improved [10]. The use of OCR, to recognise text from an image file is also prevalent. OCR is the best tool with regards to character recognition, whereby it takes in an image and returns the recognised text. Tesseract is quiet popular as an open source OCR tool, and is identified as having better accuracy and precision than other OCR tools, e.g. Transym OCR and GOCR [11]. The main issue with using OCR independently, is that it is not sufficiently reliable to determine the accuracy and is not generally 100% especially when a document has text that is not solidly black and a noisy white background, nonetheless it can be trained to achieve the expected accuracy [11, 12, 13].

Cryptographic techniques such as cryptographic hashing, Public Key Infrastructure (PKI) and digital signatures, are also very common in document verification, matter of fact it has been used in many studies to secure documents. [10] presented a solution to prove the authenticity of a document and verify it, using digital signatures. They also considered incorporating blockchain technology but decided using PKI digital signatures was sufficient for their system. Blockchain technology is flourishing in this area, its properties such as immutability, transparency and authenticity of digital records; has attracted it to a number of private and public sectors which have welcomed its use to counter document fraud [14]. Civic is one of the companies that have successfully implemented secure identity verification using blockchain technology, for this system to work, cryptographic hashing which plays a major role in Blockchain-based solutions, is employed [14]. Academic institutions have also adopted Blockchain use, e.g. Massachusetts Institute of Technology (MIT) is one of many institutions that now uses the blockchain to register digital educational certificates and allows people to authenticate these certificates, also applying vast use of cryptographic hashing for verification [6, 15]. Another system, Stampery uses a combination of blockchains, to ensure the integrity, existence and the ascription of any file or document, even communication. Once these files have been anchored to the blockchain anyone, anywhere in the world can verify their integrity [16].

Several research studies have also explored the use of two – dimensional (2D) barcodes, whereby information about the document is stored in a barcode and used later for the process of verification [1]. 2D barcodes are commonly used for document verification as they can store more data than 1D barcodes. To strengthen the security of barcodes, various cryptographic techniques are used i.e. PKI, data compression, hash functions, digital signatures [1]. [1] proposed a system whereby, barcodes are used with the help of these cryptographic techniques. Thus, showing the importance of integrating different components to design a suitable solution for the problem of document forgery. A limitation that comes with the use of barcodes is size (the amount of data that could be stored in a barcode) and once a document has numerous barcodes, it starts using a lot of space that can be used for content. Most of the proposed solutions that utilize barcodes store the entire document content in the barcode [17, 18]. In [19] we eliminated this by only storing the information that should be validated in a document, but still encountered the challenge of having numerous barcodes. Hence, in continuing with this ongoing work, the solution proposed in this study aims to decrease the number of barcodes by transferring the information used to verify the hardcopy documents to a blockchain.

From all these studies, and present implementations it can be concluded that efficient techniques must not only be effective but affordable, and implemented well to ensure the security of a document in making sure that unauthorized alterations can be detected. Thus, this study aims to provide an effective, simple and fast method of document integrity verification through the usage of 2D barcodes, OCR, cryptographic techniques and blockchain.

III. Proposed Solution

This section describes the proposed solution for verifying a hardcopy document and a digital document which is an extension of the solution that was presented initially in [19, 20]. In [19] the solution consisted of 4 components, namely; cryptographic hashing, digital signatures, OCR and 2D barcodes. Experiments were conducted using 3 different fonts i.e. Times New Roman, OCRB and AnyOCR and the highest accuracy obtained for AnyOCR was 100% which presented an opportunity to improve the solution so that it can work with different fonts. The documents generated in [20] consisted of 7 barcodes positioned at the bottom of the document, which presents a challenge to those who might want to adopt the system. The number of barcodes is dependent on the information that needs to be stored as each barcode has storage space limitation.

This led to some research on finding ways we can make the solution easily adoptable without the issue of having more barcodes to deal with when documents are generated as this might not necessarily fit in with the company's objective. All of the components used in the previous solution will still be used for this solution except only one barcode that contains information used to verify will be placed on the document. The solution is designed in a way that, if an attacker tries to tamper with the document, the system can detect those changes. In this paper we won't discuss the other components used for this solution as this was done in [20]. Only the added component and the modifications introduced are discussed.

A. Components of the proposed solution

1) Barcodes: This component was used in our previous solution, however; the limitations of storage space led to the use of more barcodes in order to accommodate all the information required to validate a document. The maximum capacity of a version 40 barcode in byte mode is between 1273 – 2953 bytes. In [20] , we observed that this barcode capacity presented flaws as it possessed high resolution which yielded negative results after printing and scanning the document. The quality of the data stored in the barcode was poor. To improve this the metadata was divided into 7 portions and stored in 7 smaller barcodes making sure that the capacity used in the barcodes is distributed equally and doesn't reach a growing rate that presents poor quality when the barcodes are decoded and read. The possibility of the capacity growing in the barcode therefore presents a downfall. This challenge might prevent the adoption of this system as a company would not agree to change their structure in-order to accommodate for more barcodes. For the reason that existing companies already have an acceptable format and layout they use to create documents. We realised when we demonstrated the first developed prototype a lot of questions were centred on the multiple barcodes, and some companies were concerned about the space used by the barcodes and the aesthetics of the document design. This challenge led us to trying other means in order to eliminate the inclusion of numerous barcodes. As a solution, we decided to transfer all the information previously stored in the barcodes to a blockchain and only have one small barcode in a document that would contain less information which will be used to verify the integrity of the document. Having one barcode would make it easier for the solution to blend perfectly with the existing structure of documents without changing it massively.

2) Blockchain: The introduction of blockchain has sparked a lot of interest in the research field. Researchers are constantly looking for means where this technology can be applied. The inception of this technology presented a lot of opportunities in the field and not only is it sparking interest in the field of cryptocurrencies, but its being studied for other purposes as well. Blockchain introduced a decentralized method of storing information, whereby all the participants have a copy of the blockchain. Thus eliminating one single point of failure. There is a plethora of blockchains e.g. Bitcoin, Ethereum, Hyperledger, Ripple, etc., one can choose from depending on what they are trying to achieve. Just like any other technology, blockchain posesses limitations i.e. the size of information one can store in the blockchain. This solution focuses more on the components selection rather than the cost of using each of these components particularly the use of barcodes and blockchain by the companies. The cost of the selected components and affordability of the company will be established accurately during the implementation of the solution, certain measures will be used to determine which blockchain will be used and deployment of this system in different company environments.

B. Proposed Solution Design

The solution consists of 2 main processes; generation and validation process. These processes utilize all the components discussed in the previous solution [20] as well as the ones discussed in the previous sub-section.

1) The Generation Process: This process includes the definition of two types of text; normal and validation text, validation text is hashed to produce a single hash value that is then encrypted with a secret key to obtain a digital signature. This process is illustrated in fig. 1 and 2.

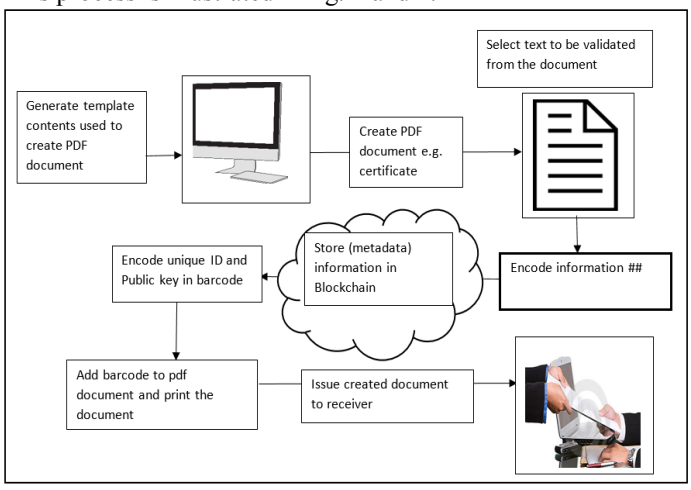

Fig. 1. Generation Process.

The digital signature and metadata are stored on the blockchain and a unique key is generated. This unique key and public key associated with the secret key used to create a digital signature are encoded to the barcode which is placed on a document. The metadata consists of; position, length, width and checksum values that are derived for each validation text, hash produced for all the validation text labels and timestamp of when the document was created. Finally, a digital copy is sent to the recipient and the pdf document is generated, printed and presented.

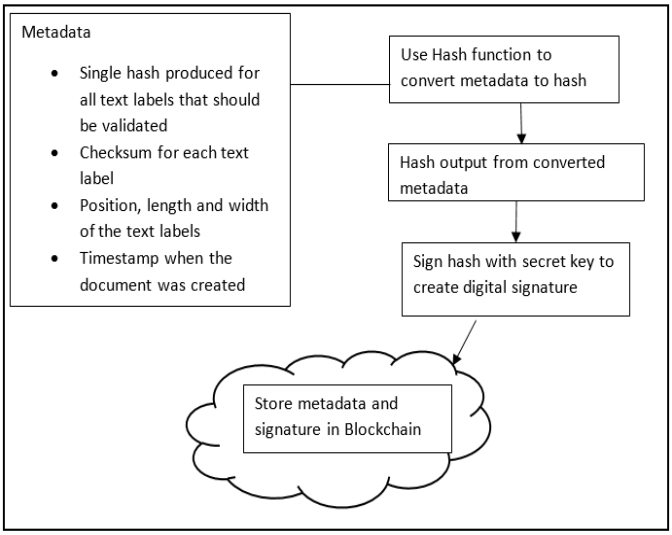

Fig. 2. Encode Information.

2) The Validation Process: Once the documents have been generated, printed and issued, the recipient can then use the documents in various cases, i.e. applying for jobs, applying for bank acounts etc. When these documents are submitted, they must be verified in order to determine whether their integrity has been maintained. Users can either submit a digital copy or a hardcopy document. If a hardcopy document is presented, the document must be scanned first in order to obtain a digital copy. To start the process of validation, the system reads in a scanned image and the barcode is identified and decoded. The barcode consists of a public key and unique key (which is used to fetch metadata related to the document from the blockchain). With the use of the public key extracted from the barcode, the digital signature is validated, if valid the metadata is extracted and used to locate the validated labels in the document. Thereafter, Tesseract OCR is used to validate the text labels. The hash (of all the validated text labels) is calculated and compared with the one from the original document (retrieved from the blockchain). If the comparison fails, it means the document has been altered. In addition to the hash that is included, a checksum for each text label is also calculated, this aids to point the exact text label that is not matching. Fig. 3, illustrates the process of document validation.

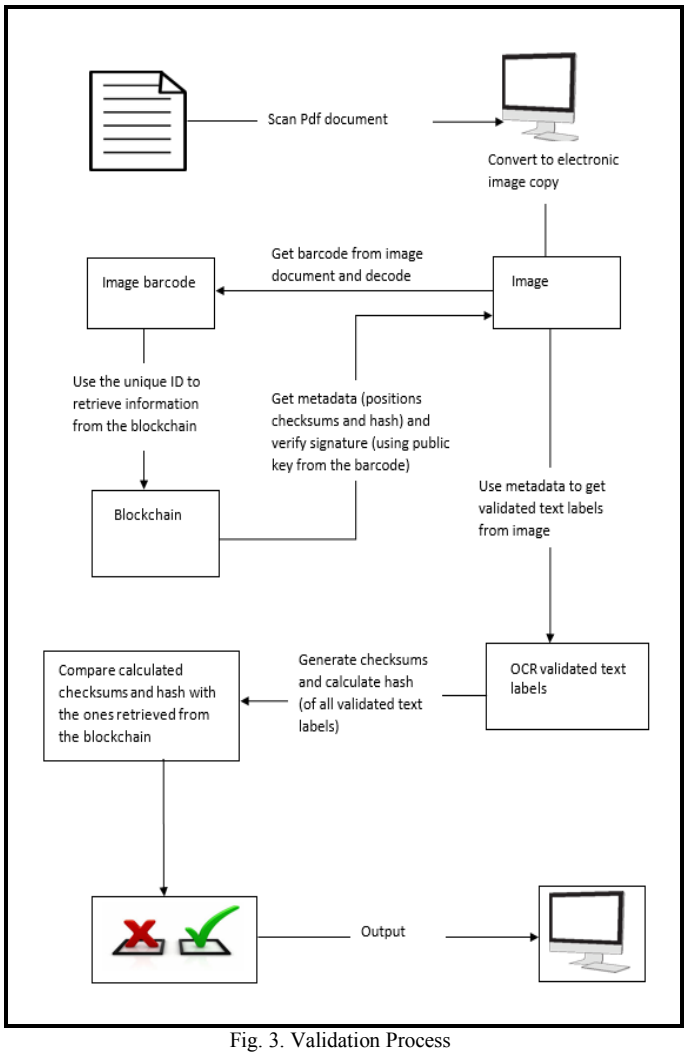

Fig. 3. Validation Process

The evaluation of the proposed solution will use a dataset generated using the generation process. The dataset would consist of 100 generated documents, using different fonts, i.e. Courier, Arial, Times New Roman etc., in order to determine the accuracy of OCR on these documents. The dataset will also be separated into different groups, on the size of text, which are; small, medium and large text. The validation process will be used to verify the integrity of the documents. To evaluate this, the information retrieved from the blockchain must be the same as the one presented when hashed.

Based on the studies conducted, the use of blockchain to verify documents is not something new and the implementations discussed in this paper have been a great success. Be that as it may most proposed and implemented solutions focus more on utilizing a single hash for verification purposes, whereby the document content is hashed and the calculated hash is saved on the blockchain [15]. To augment the existing solutions, our solution uses the blockchain to save more information about the document which is used during the process of validation, and also aims to exclude the irrelevant information in the document. Not only are we intending to identify a document that has been tampered with, but we also want to be able to show where the document has been changed.

This is made possible by including the exact position of the text that should be validated, (x and y coordinates) and the width of the text in the metadata. Before using OCR when the text is extracted from the document all white spaces are removed to minimize any additional white spaces introduced in the process. In most case a single hash value is used to represent the document's content, which cannot be obtained as a hash is designed to be a one way function, whereby we don't know the contents of the document, nor are we aware of the location of the altered text.

IV. Conclusion

This paper presented a proposed solution for the problem of document forgery, which is an extension to our previous proposed solution. The solution employed 4 techniques; OCR, cryptographic hashing, digital signatures and 2D barcodes. OCR was the first technique to be implemented, whereby documents were generated using a font known as AnyOCR (which is designed for OCR tools) and Tesseract was used to validate the documents. The experimental results yielded an accuracy of 100%, which is good. The second part of the experiment was to combine all the techniques, whereby new documents were generated and validation text was specified which was then added to the barcodes that are positioned at the bottom of the documents. Using our validation process, the system was able to detect when documents have been tampered with. This paper extends the previous solution by limiting the number of barcodes used to a single barcode and using blockchain to store the information that was previously stored in barcodes. This proposed solution will be implemented and tested for its practicability to detect forgery and ensure that the integrity and authenticity of a hardcopy document is maintained.

Acknowledgment

The authors would like to express their gratitude to the CSIR Modelling and Digital Science Unit for sponsoring this research.

References

[1] A. Husain, M. Bakhtiari, and A. Zainal, "Printed Document Integrity Verification Using Barcode," *Journal Teknologi (Sciences and Eng*, pp.99-106, 2014.

[2] M.H. Eldefrawy, K. Alghathbar, and M.K. Khan, "Hardcopy document authentication based on public key encryption and 2D barcodes," In *Biometrics and Security Technologies (ISBAST), 2012 International Symposium,* pp. 77-81, IEEE, March 2012.

[3] R. Jain, and D. Doermann, "Visualdiff: Document image verification and change detection," In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on* pp. 40-44, IEEE, August 2013.

[4] N. Ganesan, "*Three foreigners jailed in Singapore for submitting fake academic certificates _ Human Resources Online*, *Human Resources*," Available at: http://www.humanresourcesonline.net/three-foreigners-jailed-for-submitting-fake-academic-certificates/ (Accessed: 07 February 2018), 2017.

[5] S. Kalipa, "*Home Affairs official sold us fake papers _ IOL News*, *Crime and Courts IOL News*," Available at: https://www.iol.co.za/news/crime-courts/home-affairs-official-sold-us-fake-papers-1929048 (Accessed: 08 February 2018), 2015.

[6] C. Lewis, "SABC News exposes fake payslips, bank statements," 8 May 2018, 2018. [Online]. Available: http://www.sabcnews.com/sabcnews/sabc-news-exposes-fake-payslips-bank-statements/. [Accessed: 13-Sep-2018].

[7] B. Cresitello-Dittmar, "Application of the Blockchain For Authentication and Verification of Identity," Independent Paper, 2016.

[8] Y. S. Joshi, "The Future of Enterprise Printing: Securing Hardcopy Documents in the Digital Age: white paper," CIO Insight, no. July, pp. 1–12, 2014.

[9] C. Lakmal, S. Dangalla, C. Herath, C. Wickramarathna, G. Dias, and S. Fernando, "IDStack - The common protocol for document verification built on digital signatures," 2017 Natl. Inf. Technol. Conf. NITC 2017, vol. 2017–Septe, no. September, pp. 96–99, 2018.

[10] S. R. M. Oliveira, M. A. Nascimento, and O. R. Zaiane, "Digital Watermarking: Status, Limitations and Prospects," Technical Report TR 02-01, Department of Computing Science, Alberta University, Edmonton, Alberta, Canada, 2002.

[11] S. Dhiman, and A. Singh, 2013. "Tesseract vs gocr a comparative study," *International Journal of Recent Technology and Engineering*, *2*(4), pp.80, 2013.

[12] C. Patel, A. Patel, and D. Patel, 2012. Optical character recognition by open source OCR tool tesseract: A case study. *International Journal of Computer Applications*, *55*(10), 2012.

[13] V. S. Chandel, "Deep Learning based Text Recognition (OCR) using Tesseract and OpenCV," 06 June 2018, 2018. [Online]. Available: https://www.learnopencv.com/deep-learning-based-text-recognition-ocr-using-tesseract-and-opencv/. [Accessed: 13-Sep-2018].

[14] B. Karanjia, A. G. Karanth, S. Veerapaneni, S. Goswami, A. Sharma, and M. Boda, "Blockchain in the Public Sector – Transforming Government Services through Exponential Technologies," 2017.

[15] Universa, "Blockchain in Education," 23 May, 2018. [Online]. Available: https://medium.com/universablockchain/blockchain-in-education-49ad413b9e12. [Accessed: 04-Sep-2018].

[16] A. S. de P. Crespo and L. I. C. García, "Stampery Blockchain Timestamping Architecture (BTA) - Version 6," 2017, pp. 1–21.

[17] M. Salleh, and T.C. Yew, "Application of 2D Barcode in Hardcopy Document Verification System," In *ISA*, pp. 644-651, June 2009.

[18] C.M. Li, P. Hu, and W.C. Lau, "Authpaper: Protecting paper-based documents and credentials using authenticated 2D barcodes," In *Communications (ICC), 2015 IEEE International Conference,* pp. 7400-7406, IEEE, June 2015.

[19] S. Mthethwa and N. P. Dlamini, "Verifying the Integrity of Hardcopy Document Using OCR," in 2nd International Women in Science Without Borders (WiSWB)-Indaba, 2018.

[20] N. Dlamini, S. Mthethwa, and G. Barbour, "Mitigating the Challenge of Hardcopy Document Forgery," in International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), 2018, pp. 1–6.