DOCUMENT VERIFICATION SYSTEM FOR FRAUD DETECTION USING MACHINE LEARNING TECHNIQUE

By
Nwanze D. E.[1], Okechukwu O. P.[2], Nnaji C. H.[3]

1,3: Department of Computer Science, Novena University, Ogume, Delta State, Nigeria.
2: Department of Computer Science, Nnamdi Azikiwe University, Awka, Nigeria

Abstract:
This research paper discusses an intelligent document verification system for fraud detection using machine learning technique. This research seeks to address the problems encountered in the conventional verification system such as delay time, lack of intelligence and not being reliable by developing a machine learning based verification system and localizing it for the verification of certificate at the Nnamdi Azikiwe University (Unizik), Awka, Nigeria. To achieve this, the methods of data collection, data acquisition, data processing, feature extraction, artificial neural network training, and classification were used. Self-defining equations and modeling diagrams were used to develop the artificial neural network model and then train with 1180 authorized data collection of Nnamdi Azikiwe University, Awka certificates from 2016 to 2020, to generate the reference verification model which was used to develop the expert system for verification of documents. The system was implemented using image acquisition toolbox, image processing toolbox, statistical and feature extraction toolbox, neural network toolbox in Matlab and then tested for evaluation. The result recorded however, achieved a Mean Square Error (MSE) performance of 0.000100Mu and Regression value of R= 0.99373 which is very good, with implication that the proposed system is very reliable.

Keywords: Data acquisition, Intelligent System, Certificate verification, artificial neural network

I. INTRODUCTION

Public and private institutions in most of the countries rely on traditional verification processes that extensively use paper records and manual procedures to verify and authenticate presented academic certificates. Moreover, employers and other academic institutions have been experiencing a high alarming rate of fake university academic certificates.

According to Wandera (2015), the current system is a typical traditional verification process characterized by tedious, repetitive and boring search and locate procedures that accompany every verification request. A process that could require less than one hour duration sometimes takes not less than two weeks; one day's costs may become more than six days costs.

It should be noted that a graduation certificate is one of the most important documents issued by universities and other educational institutions. It is proof of a graduate's qualification. However, due to advances in printing and photocopying technologies, fake certificates can be created easily and the quality of a fake certificate nowadays can be as good as the original. The certificates of many prominent universities have been forged and these forgeries are very difficult to detect. Lack of a mechanism for instant verification of graduates' academic certificates has been and remains a major wish in higher academic institutions all over the world (Rhead, 2012).

Currently, employers, recruiters, and other organizations experience difficulties to authenticate academic documents instantly, since one has to come physically to the university or send an email, which takes time. Incidences of graduates losing opportunities due to the delay in the verification process have been common. In view of the public concerns raised in different forums, there is need to undertake a study to develop a prototype of an academic document verification system.

II. LITERATURE REVIEW

Shivram et al. (2018) researched on document analysis and data recognition system. The study was developed to provide a solution to the challenges encountered during forensic

investigation. In the research, image processing technique was used for the segmentation and analysis of documents, the machine learning was used for the recognition process. The limitation with the study is that only handwritten document can be verified.

Jamilah (2015) used image processing for the recognition and analysis of documents. In the study, the binarization and Otsu thresholding algorithm was developed and deployed as a document verification system. This system was able to differentiate between two different but physically similar documents when tested. However, despite the success of the algorithm, the performance can be improved using machine learning technique.

Subramanian et al. (2017) researched on the use of stroke detection algorithm for the detection, extraction and segmentation of handwritten feature vectors in documents. The algorithm developed was deployed as an expert system for the intelligence analysis of documents. However, despite the success the system cannot perfectly differential real and fake documents except using machine learning technique.

Oleka (2018) presented research on the impact of optical character recognition with putative analysis for legal document notarization in MATLAB for forensic analysis. In the study, optical character recognition technology was used for data collection. Then putative clustering technique was used for classification and recognition. The system was deployed on MATLAB and tested with various real and fake documents. The result showed high level of accuracy but the performance is limited to only legal documents.

Salvador et al (2017) presented a research on improving office document recognition via text extraction. In the study, features of handwritten text were extracted and used for the training of an intelligent algorithm for the verification of documents. The system, despite the success, was limited to only office documents and the accuracy achieved need to be improved.

Amariot et al. (2018) surveyed on optical character recognition applications. In the study, the various technologies used for the analysis of documents was discussed and presented. The review submitted that the use of artificial intelligence technique will provide reliable means to verify document originality.

Scott (2018) researched on handwriting document recognition system using multiple pattern recognition class models. In the study, the problem of handwritten document was identified as a pattern recognition type and then machine learning was used for the training and classification. The system when deployed and tested showed high recognition accuracy and will be considered in this research for certificate recognition.

Namukose (2018) developed a secured online document verification system. In the study a web-based app was developed and used for the verification of education certificate. This was achieved using certificates as dataset and then compared with the query certificate for verification. The study despite the success can be improved using machine learning algorithms.
Their work addresses a pixel-based binarization evaluation methodology for historical handwritten/machine-printed document images. In the proposed evaluation scheme, the recall and precision evaluation measures are properly modified using a weighting scheme that diminishes any potential evaluation bias.

III. MATERIALS AND METHOD

The methodologies used for the development of the proposed system are image processing, artificial intelligence and object-oriented analysis methodology. The various processes involved are data collection, data processing, optical image acquisition, binarization, segmentation, feature extraction, training and classification. These are to be explained in the following sections.
The materials used are the post graduate certificates, under graduate certificates, scanner, laptop, and microsoft certificate publisher software. The specifications for the hardware and software materials are presented below:

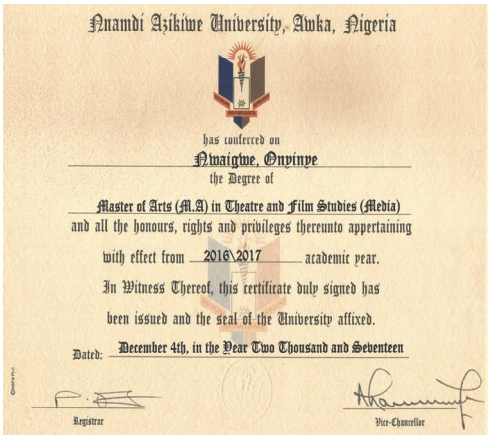
Software Specification
Language: MATLAB
Database: MySQL

Operating System: Mac, Linux, Windows NT/95/98/2000 and above.
RAM: 4GB-8GB

Hard ware Specification:
Processor: Core, Intel P-III based system
Processor Speed: 250 MHz to 833MHz
RAM: 64MB to 256MB
Hard Disk Space: 2GB and above
HD Scanners
Key Board: Standard or Enhanced Keyboard

3.1. Data collection

The primary source of data for this research is from the Unizik Exams and Records department which provided data of certificates and other non-sensitive official documents of the university from 2016-2020.
The secondary sources of data collection are from self-volunteered Alumni of the institutions from 2016- 2020 which provided data of their certificates with a total sample size of 40. Also, expert was consulted to replicate 20 varieties of Unizik certificates using micro soft publisher and was used to develop the testing dataset. The total data collected and used for the training dataset is 1160 and 20 for the testing dataset, making a total sample size of 1180 data collected. The data were acquired using optical character recognition (OCR) system and then returned back to the sources; while the scanned softcopy was used for the research. The data sample is shown in Figure 1.


Nnamdi Azikiwe University, Awka, Nigeria

has conferred on
Nwaigwe, Onyinye
the Degree of
Master of Arts (M.A) in Theatre and Film Studies (Media)
and all the honours, rights and privileges thereunto appertaining
with effect from 2016\2017 academic year.
In Witness Thereof, this certificate duly signed has
been issued and the seal of the University affixed.
Dated: December 4th, in the Year Two Thousand and Seventeen

Registrar Vice-Chancellor

Figure 1: The Data Sample

3.2. Data Processing

To format the data and achieve singularity in the properties and characteristics, the Matlab database toolbox was used to convert all the data irrespective of the size into Matlab.File (.m file). When this process was completed, the data was fed to the artificial intelligence system for training. The method of artificial intelligence and training will be discussed later.

3.3. Optical Image Acquisition

The quality of optical character image acquisition device is very important and a sure step of achieving accuracy in the document recognition process. While OCR technology is highly accurate when scanning straight or printed text, accuracy can be greatly decreased if the quality of the print scanned document is not good or if the document contains mixed columns, recursive words, charts, diagrams, or graphics. There are large number of image acquisition scanners available, but most existing OCR software do not work perfectly with all hardware models. One of the reasons is that the software designed idea was focused more on OCR image processing and recognition, thus contributes less to the image acquisition features and accuracy. However, the quality and accuracy of the scanned image plays a very key role in the recognition process. For this research work, high-definition image acquisition tools were employed to capture samples of provided document certificates which were used to create a testing database for the system evaluation.

3.4. Artificial intelligence

This is a machine learning algorithm with the capacity to learn feature vectors from training dataset and make accurate decision. This research will use the clustering technique to solve this problem. The neural network approach is an unsupervised machine learning algorithm which uses set of clusters feed forward via neurons and activation function to make decision.

3.5. Training and Classification

The training and classification are processes that work hand in hand in artificial intelligence system. The training is the process of learning the artificial intelligence technique with feature vectors extracted from the training dataset to develop a reference model for classification. In the classification parts when customer data in time series are feed forward, the ANN

compares the extracted features with that in the reference model and then make classifications to decide if the document is real or fake. The process flowchart is presented in Figure 2.

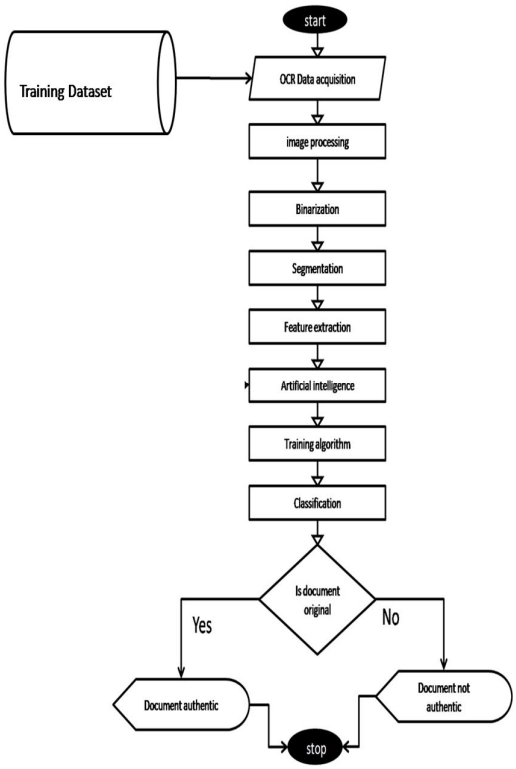


Figure 2: The Logical Dataflow Model of the Processes

The flowchart presents the logical data flow of how each method is interrelated with the other, starting with the training dataset. The training dataset was used to learn the artificial intelligence technique of the features for the data collected with the intelligence model which serves as base for future classification. When new data is uploaded for verification using the OCR device (scanner), the data is processed using binarization for bi-level conversion, then data processing for resizing. The features of the documents are revealed more for better extraction performance using segmentation process and then extracted using feature extraction technique which converts the image features into a compact feature vector and then feed to the artificial intelligence system for training and classification. The A.I is neural network which uses training algorithm to learn and classify features for accurate decisions.

3.2 To Develop a Machine Learning Algorithm for Classification of Certificate Documents

The machine learning algorithm used for this research is the Artificial Neural Network (ANN). This ANN was discussed in details in the literature review and will be designed and trained with the data collected in this section. The design of the ANN will be done using the block diagram as shown in Figure 3.

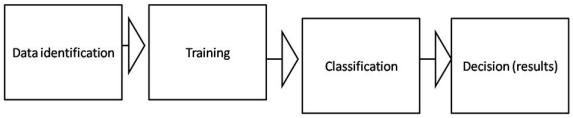


Figure 3: Block Diagram of the ANN Operation

The ANN operated in four basic steps as submitted using the block diagram in figure 3.3. The system uses neurons which has weights and bias to identify feature vectors from the dataset of original documents (training dataset) and then use the training algorithm to learn the data for classification and accurate decision. The activity model of the ANN is presented as shown in Figure 4.

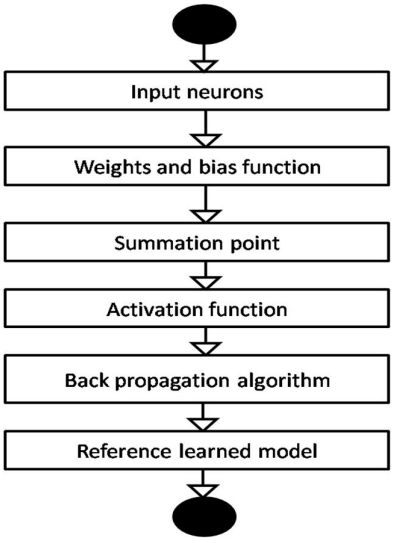


Figure 4: The Activity Model of the ANN

The model in the Figure 4 presents the logical data flow in the neural architecture as showing the interconnected neurons which forms the layers, the summation, activation function and

.
training algorithm which produced the reference model. The training parameters are presented in Table 1

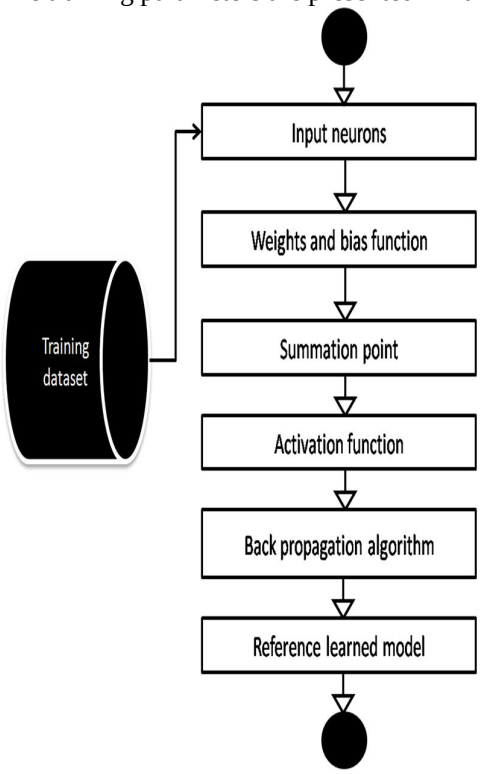


Figure 5: The Train ANN model

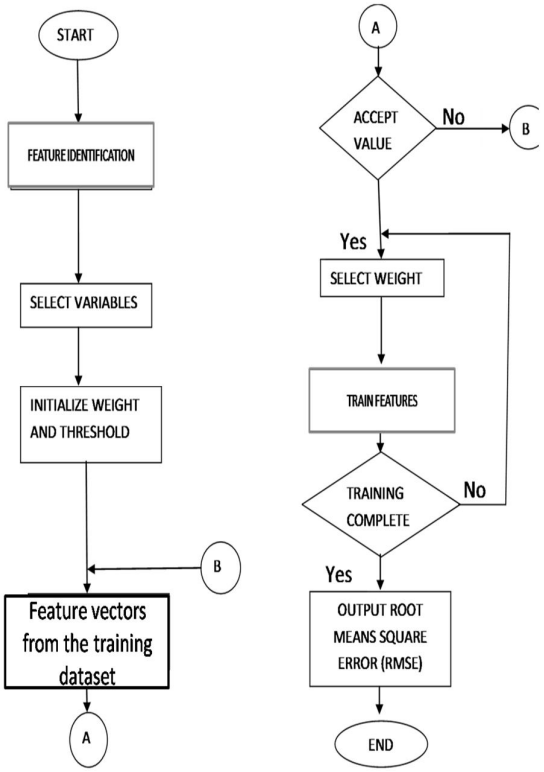


Figure 6: Training Algorithm

In the Figure 5, the ANN configured with multiple neurons which have weights and bias function was feed forward with the feature vectors from the training dataset. The summation of the feature points was activated using nonlinear activation function which converts the feature vectors into statistical values from 0 to 1 based on Tansig function which is a mathematical function used to introduce nonlinearity into feature vectors and eliminate negative values. The real values are then trained using the back propagation algorithm in figure 6 to learn the feature vectors and achieve a reference classification model. The Pseudo Code for the Machine learning algorithm developed is presented below.

IV. RESULTS AND DISCUSSIONS

4.1 Result of the Algorithm

The performance of the algorithm was generated using the neural training tool as identified in the implementation section. The tool used necessary elements which can measure how effective the learning process was done to evaluate the performance as shown in figure 7.

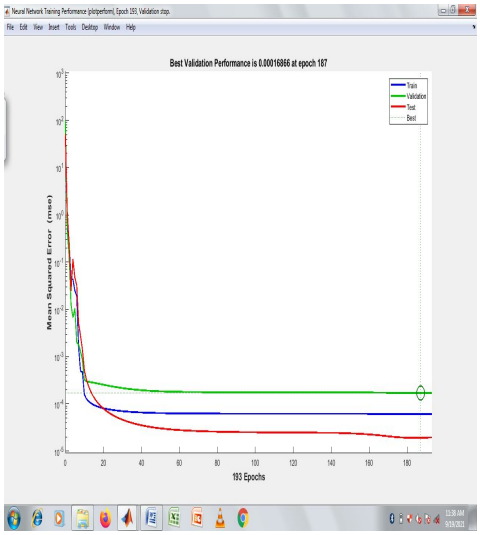


Figure 7: Mean Square Error (MSE) Performance

The aim of this tool is to measure the error margin recorded during the system training, with the hope of achieving a MSE value equal or approximately zero. From the result, it was observed that the MSE performance is 0.00100Mu which is good at epoch 187 and best validation value of 0.0016866. The implication of this result shows that the training error margin was almost zero which is very good and proved that the training was very good and the algorithm accurately learned the data fed to it for training. The next result presented the regression performance of the system as shown in Figure 8.

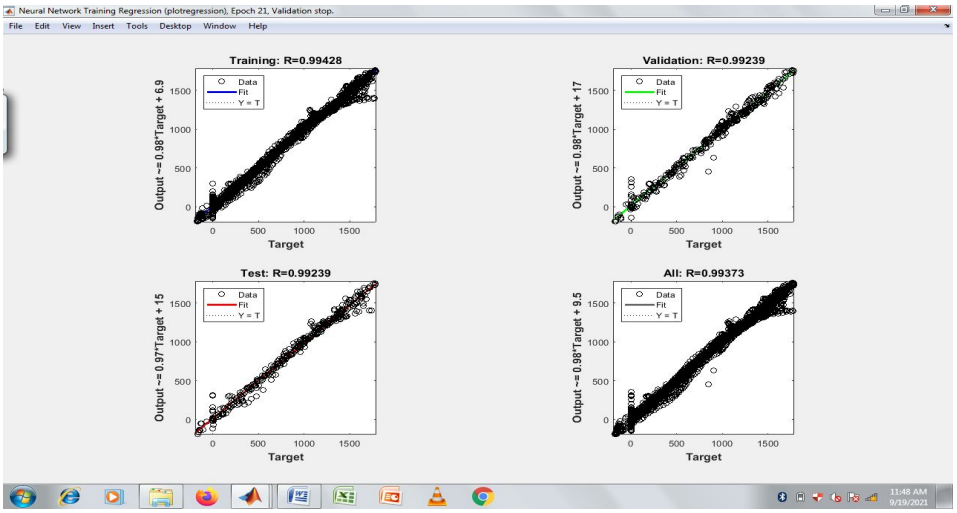
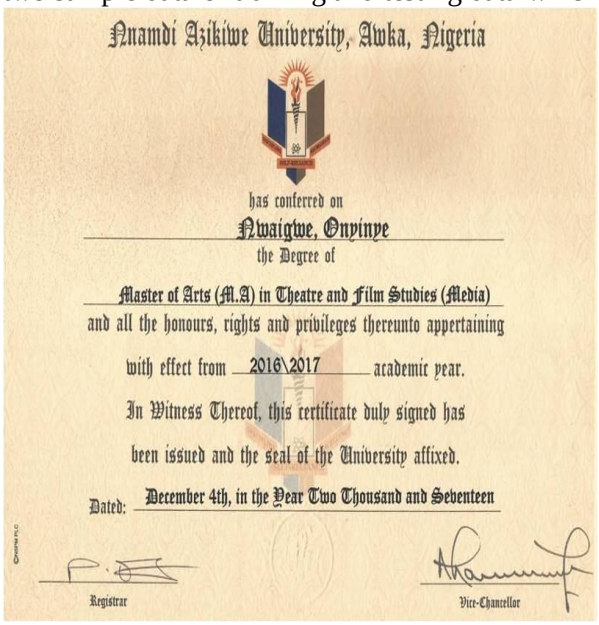


Figure 8: Regression Result

4.2 Result of the Verification Software

The next result presented the performance of the system when the algorithm was deployed as an expert system using Matlab. The figure 9 presented two sample data of training and testing data which was used to evaluate the performance of the document verification software developed. The figures 9(a) and (b) are the training document sample and the testing document sample for verification.


Nnamdi Azikiwe University, Awka, Nigeria

has conferred on

Nwaigwe, Onyinye

the Degree of

Master of Arts (M.A) in Theatre and Film Studies (Media)

and all the honours, rights and privileges thereunto appertaining

with effect from 2016\2017 academic year.

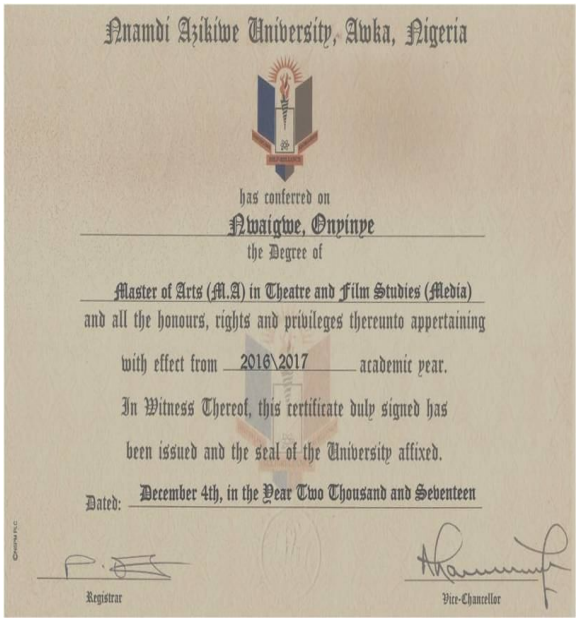
In Witness Thereof, this certificate duly signed has

been issued and the seal of the University affixed.

Dated: December 4th, in the Year Two Thousand and Seventeen

Registrar Vice-Chancellor

Figure 9 (a): Query Document Sample


Nnamdi Azikiwe University, Awka, Nigeria

has conferred on

Nwaigwe, Onyinye

the Degree of

Master of Arts (M.A) in Theatre and Film Studies (Media)

and all the honours, rights and privileges thereunto appertaining

with effect from 2016\2017 academic year.

In Witness Thereof, this certificate duly signed has

been issued and the seal of the University affixed.

Dated: December 4th, in the Year Two Thousand and Seventeen

Registrar Vice-Chancellor

Figure 9 (b): Testing Document Sample

Before the system verification, recall that the training sample was part of the training dataset which was used in training the system and generated the algorithm used to develop the verification system. Here the query certificate was uploaded into the system for verification.
The uploaded sample was identified by the software and the result is presented in figure 10.

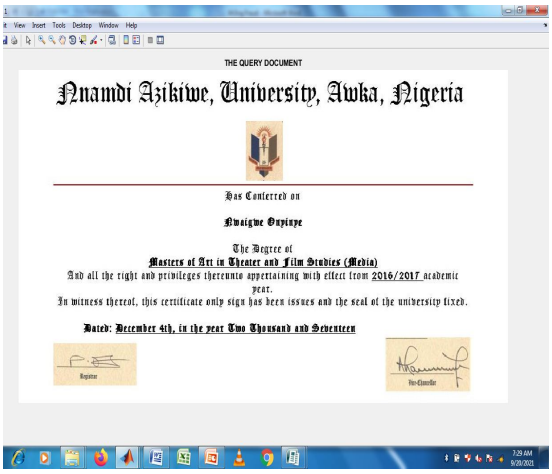


Figure 10: Sample of the Query Certificate

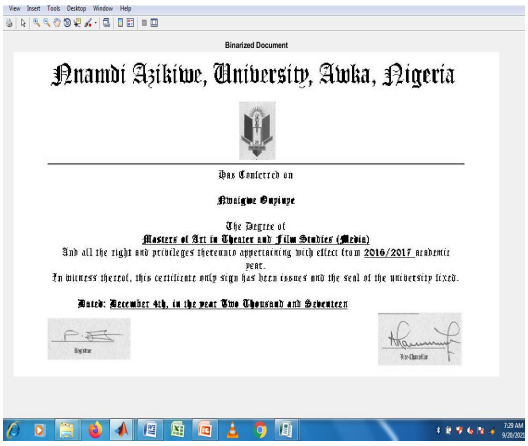


Figure 11: Sample of the Binarized Result

The figure 11 presents the sampled query image and the pre-processing result. The pre-processed output was segmented and then trained for classification as shown in figure 12 for segmentation, and the figure 13 for the verification result.

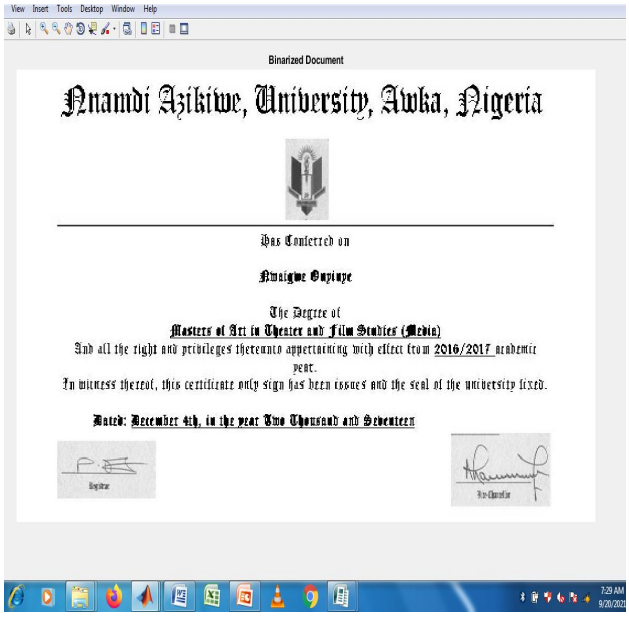


Figure 12: Segmentation Result

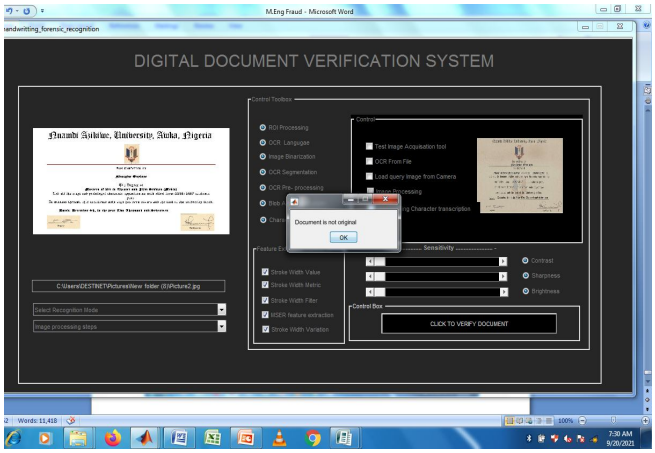


Figure 13: Verification Result

V. CONCLUSION

The designed document verification system for fraud detection using machine learning technique was presented. This work addressed the problems inherent in the conventional verification system such as delay time, cumbersome, high cost via the developed machine learning based verification system and localizing it for the verification of certificate at the Nnamdi Azikiwe University (Unizik), Awka. This was achieved using the methods of data collection, data acquisition, data processing, feature extraction, artificial neural network training, and classification. Self defining equations and modeling diagrams were used to develop the artificial neural network model and then train with authorized data collection of Unizik certificates from 2016

to 2020, to generate the reference verification model which was used to develop the expert system for verification of documents. The system was implemented using image acquisition toolbox, image processing toolbox, statistical and feature extraction toolbox, neural network toolbox in Matlab and then tested for evaluation. The result recorded however, achieved a Mean Square Error (MSE) performance of 0.000100Mu and Regression value of R= 0.99373 which is very good, with implication that the proposed system is very reliable.

REFERENCES

Aisltd. (2017). Document Verification System. Retrieved June 1, 2018, from www.aisltd.ie: https://www.aisltd.ie/latest-news/document-verification-systems-why-it%E2%80%99s-time-you-invested-in-one.1514.html.

Ajayi, O. B., Sodiya, A. S., & Ibrahim, S. A. (2015). The State of Information Security in South- Western Nigerian Educational Institutions. Department of Mathematical Sciences, 2004. . Retrieved from www.unaab.edu.ng: http://www.unaab.edu.ng/journal/index.php/COLNAS/article/download/174/172

Ayush Ruhu Purohit and Shardul Singh Chauhan Rusu (2015). A Literature Survey on Handwritten document Character Recognition, International Journal of Computer Science and Information Technologies, Vol. 7 (1), Rusu.

Ke Y and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors." in *Proc. CVPR*. Vol. 2, pp. 506–513, 2004.

Levine, S., Finn, C., Darrell, T. & Abbeel, P. (2016). "End-to-end training of deep visuomotor policies," J. Mach. Learning Res., vol. 17, no. 39, pp. 1–40.

Linus U. (2017) "Tackling the rise of fake qualifications in Nigeria"; Politics and society; this Africa Publisher; 2017.

Oleka Chioma (2018), "The Impact of Optical Character Recognition With Putative Analysis For Legal Document Notarization In Mathlab For Forensic Analysis, Enugu State University Of Science And Technology, Enugu, Nigeria, Department Of Computer Engineering.

Reena Bajaj, Lipika Dey, and S. Chaudhury (2002), "Devnagari numeral recognition by combining decision of multiple connectionist classifiers", Sadhana, Vol.27, part. 1, pp.-59-72.