WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Certificate Fraud Detection Using Artificial Intelligence Technique

Isizoh A.N.[1], Anyi D.O.[2], Onyeyili T.C.[3], Ebih U.J.[4], Ejimofor I.A.[5]

[1,2,3,4]Dept. of Electronic & Computer Engineering, Nnamdi Azikiwe University, Awka, Nigeria.
[5:] Dept. of Computer Engineering, Madonna University, Nigeria.

ABSTRACT

This paper presents the development of an intelligent certificate verification system for fraud detection using machine learning technique. The research was embarked upon after noticing the rate of document forgery in the Nigerian society. Thus an exhaustive review of literatures was made which identified the challenges public and private institutions encounter due to lack of automated means to verify any legal document. The flaws in the conventional verification system such as delay time, cumbersome, high cost, lack of intelligence and above all, not being reliable; have been exploited over the years by fraudsters to fabricate fake documents such as certificates mostly and commit fraud. This research seeks to address the problems via the development of 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 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, 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 new system is very reliable.

KEYWORDS: Certificates, fraud, machine learning, image processing, feature extraction, and modeling.

I. INTRODUCTION

Today, in both the public and private enterprises, the traditional means of document verification is via observation method where documents are observed and then automatically accepted for certain actions or proceedings based on signatures, stamps, seals, etc. However, these criteria used for authenticating documents are very easy to be reproduced and commit fraud, and as a result has become a major problem, as the daily media hardly make broadcast today all over the world without citing a scenario of fraud with fake documents.

One of the document types which is the current trend today is the fabrication of fake certificates. Certificates are official documents given by an organization to certain qualified persons as a proof that the person has completed a certain professional training, course, study or program. This document is accepted worldwide for employment, education, training purposes etc. However due to the importance of this certificate, and the advancement in certain photocopy, scanning and printing machines, the process of reproducing them are very easy [1].
In Nigeria for instance, certificates are really big deals as most people are actually more interested in getting the certificates than acquiring the requirements such as skill, knowledge, experienced among others needed for the award of the certificates. One of such certificates is the education certificate. According to [2], the disease of forging educational certificate has infected every sphere of public life in Nigeria, including politics. In fact, many politicians are allegedly victims of false documents fabrication. In Tanzania alone, their government once fired 10,000 civil servants over fake documents used to acquire work.

In the United States, a fraudster who allegedly defrauds millions from victims using fake documents was arrested and sentenced on February 16, 2021 by the Department of Justice. On 20 January 2018, the BBC news reported that 20 fraudsters were arrested for fake travelling documents in Nigeria. Another report by Nairaland Online News (April 15, 2018) submitted that a fake doctor was arrested with falsified certificate of the Nigerian Medical Association at the Lagos State Teaching Hospital. Even the Nigerian police that is supposed to fight fraud is also a victim as the premium times daily on March 27, 2018 reported that over 80000 ghost workers in the agency were uncovered with false documents representations. These issues among others, to mention but a few, are major problems worldwide and have affected the economy, credibility, integrity and standards of many countries in negative way.

The main cause of these problems is simply the lack of an intelligent document verification system. The fraudsters can easily provide details which convince the conventional means of document authentication "observation method" which authenticates documents by signature, dates, seals and stamps, however these information are not enough to validate document originality, and as a result these limitations are taken as advantage by fraudsters to replicate documents and commit fraud.

Overtime, many solutions have been proffered to solve this problem. In the academic sectors for instance, every 5 years, administrative signatures of various official documents are changed. Some private companies change their certificate design, seal signatures and add unique concepts to improve uniqueness of their documents. Some introduce private serial numbers; electronics stamps among other means to improve document standards and make false document fabrication difficult. However, despite the success, it is clear that these approaches only provide short term solution to a long-term problem of fraud.

To solve this problem completely, there is need for a digital document verification system. When an intelligent system which can verify document is developed and deployed to the public and private sectors, it will be impossible for criminals to commit fraud. Overtime, many researchers have proposed this using many techniques such as image processing, optical character recognition techniques, colour-based approaches among others. However despite the success, the techniques suffer from issues of false alarm, high cost, unreliability among others, however the performance can be improved using machine learning techniques.

Machine learning is thus a set of artificial intelligence algorithms which have capabilities to learn patterns in data and make correct predictions. This research proposes the use of this technique to solve this problem by training the algorithm with data of authentic documents to develop a reference document algorithm which will be deployed to build a document verification system. This when achieved will detect fake certificate documents with high level of accuracy for elimination of problems of fake documents, ghost worker, frauds, among others.

II. LITERATURE SURVEY

2.1 Machine Learning (ML)

Machine learning is a branch of artificial intelligence (A.I) that focuses on building applications that learn from data and improves its accuracy over time without being programmed to do so. This means a single program, once created, will be able to learn how to do some intelligent activities outside the notion of programming. This is in contrasts with purpose-built programs whose behavior is defined by hand-crafted heuristics that explicitly and statically define their behavior. So, Machine Learning is an approach to achieve Artificial Intelligence. Machine learning combines data with statistical tools to predict an output. In machine learning, algorithms are 'trained' to find patterns and features in massive amounts of data in order to make decisions and predictions based on new data. The better the algorithm, the more accurate the decisions and predictions will become as it processes more data [2].

In the views of [3], machine learning is a subfield of computer science often also referred to as predictive analytics, or predictive modeling. Its goal and usage are to build new and/or leverage on existing algorithms to learn from data, in order to build generalizes models that give accurate predictions, or to find patterns, particularly with new and unseen similar data.

According to [3], there are four basic steps for building a machine learning application (or model). These steps are listed thus:

Step 1: Select and prepare a training data set

Step 2: Choose an algorithm to run on the training data set

Step 3: Training the algorithm to create the model

Step 4: Using and improving the model

2.2 Artificial Neural Networks (ANN)

The field of neural networks is a subarea of machine learning. The human brain has about 100 billion nerve cells. We humans owe our intelligence and our ability to learn various motor and intellectual capabilities to the brain's complex relays and adaptivity. For many centuries, biologists, psychologists, and doctors have tried to understand how the brain functions. Around 1900 came the revolutionary realization that these tiny physical building blocks of the brain, the nerve cells and their connections are responsible for awareness, associations, thoughts, consciousness, and the ability to learn. Human brain is one of the best 'machines' we know for learning and solving problems. Within the machine learning fields, there is an area often referred to as brain-inspired computation. The brain-inspired technique is indeed inspired by how our human brain works. It is believed that the main computational element of our brain is neuron. The complex connected network of neurons forms the basis of all the decisions made based on the various information gathered. This is exactly what Artificial Neural Network technique does.

An artificial neural network is a system of hardware or software that is patterned after the workings of neurons in the human brain and nervous system. Artificial neural networks are a variety of deep learning technology which comes under the broad domain of Artificial Intelligence. However, the tools used for modeling, namely mathematics, programming languages, and digital computers have very little in common with the human brain. With artificial neural networks, the approach is different. Starting from knowledge about the function of natural neural networks, we attempt to model, simulate, and even reconstruct them in hardware.

However, the efficiency of a neural network is a function of how extremely adaptive it is to learning very quickly. Each node weighs the importance of the input it receives from the nodes before it. The inputs that contribute the most towards the right output are given the highest weight. There are many

different types of neural networks which function on the same principles as the nervous system in the human body. Howard Rheingold said, "The neural network is this kind of technology that is not an algorithm, it is a network that has weights on it, and you can adjust the weights so that it learns. You teach it through trials."

A neural network has a large number of processors. These processors operate parallel but are arranged as tiers. The first tier receives the raw input similar to how the optic nerve receives the raw information in human beings. Each successive tier then receives input from the tier before it and then passes on its output to the tier after it. The last tier processes the final output. Small nodes make up each tier. The nodes are highly interconnected with the nodes in the tier before and after. Each node in the neural network has its own sphere of knowledge, including rules that it was programmed with and rules it has learnt by itself .

2.3 Mathematical Presentation of a Simple Neural Network Structure

The Neural Network is an interconnected massive parallel computational models, units or nodes, whose functionality mimic the animal neural network in order to process information from the input to the output using the connection strength (weight) obtained by adaptation or learning from a set of training patterns. The mathematical description of the neural network process is shown in Figure 1.

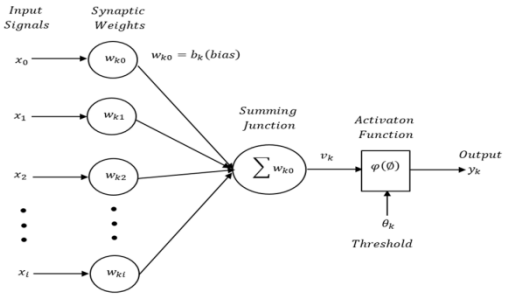


Figure 1: Mathematical Model of an Artificial Neuron

The neuron is a unit of computation that reads the inputs given, processes the input via its weights, sum the result collected and transform into statistical values using the activation function and gives the output in processed form. The weighted sum of the input's neurons is presented as:

$$v_k = \sum_{i=1}^{N} w_{ki} x_i \qquad (2.1)$$

Where;

x_i is the neuron's input from the training dataset.

w_{ki} is the corresponding weight to the input x_i.

The neuron's output is obtained by sending the weighted sum v_k as the activation function φinput that resolves the output of the specific neuron. $y_k = \varphi(v_k)$. A step function with threshold t can be used to express a simple activation as:

$$\varphi(x) = \begin{cases} 1\ if x \geq t \\ 0\ if x < t \end{cases} \qquad (2.2)$$

However, bias is most times used instead of a threshold in the network to learn optimal threshold by itself by adding $x_o = 1$ to every neuron in the network. The step activation function for the bias becomes:

$$\varphi(x) = \begin{cases} 1\ if\ x \geq 0 \\ 0\ if x < 0 \end{cases} \qquad (2.3)$$

More advanced learning or adaptation can be achieved by using a multilayer network of neurons formed by feeding the output of one neuron to the input of another neuron as shown in Figure 2. The layers between the input and output layers are termed hidden layers. Each layer of the multilayer network is made up of a bunch of neuron nodes. One circle represents one neuron in Figure 2. Feed forward multi layered networks are known to solve nonlinear problems when a nonlinear activation is used.

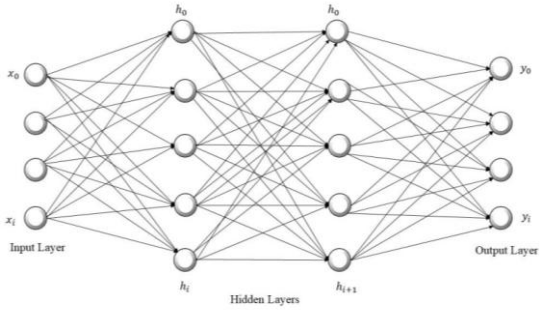


Figure 2: Basic Structure of a Neural Network

Any two neurons are connected by a link that has a weight which represents the connection strength between the two neurons. In Figure 2, the w_{ij}^{l} denotes the weight for a link between unit j in layer l and unit i in layer $l+1$. Also b_i^lrepresents the bias of the unit i in layer $l+1$. For any neural network, the associated parameters inside it are expressed as a function of the weight and the bias of the neurons as:

$(w, b) = (w^1 b^1, w^2 b^2, w^3 b^3, w^4 b^4 \ \ldots)$ (2.4)

The components of equation 2.4 can be written in the form of $w^1 \in \mathbb{R}^{3\times 3}$ and

$w^1 \in \mathbb{R}^{1\times 3}$.

Let the activation of unit i in layer $l$be represented by a_i^l, then the input for the layer labelled as L_1 we have $a_i^1 = x_i$ for the ith input of the whole network. Other layers are given by $a_i^l = f(z_i^l)$, where z_i^l is the total weighted sum of the inputs to unit i in layer l in addition to the bias term.

Each of the classes from the training set are feed to one input layer of the neurons with the activation function and bias for each neuron presented as the models below; the essence of the bias function is to shift the activation functions to either left or right when the training starts.

$a_1^2 = f(w_{11}^1 x_1 + w_{12}^1 x_2 + w_{13}^1 x_4 + b_1^1)$ (2.5)

$a_2^2 = f(w_{21}^1 x_1 + w_{22}^1 x_2 + w_{23}^1 x_4 + b_2^1)$ (2.6)

$a_3^2 = f(w_{31}^1 x_1 + w_{32}^1 x_2 + w_{33}^1 x_4 + b_3^1)$ (2.7)

$a_4^2 = f(w_{41}^1 x_1 + w_{42}^1 x_2 + w_{43}^1 x_4 + b_4^1)$ (2.8)

The four equations presented from (2.5) – (2.8) represented the input neuron feed with each class of the training data model (assuming the dataset has four classes). The summation of the neurons is presented using the model below:

$h_{w,b}(x) = a_1^4 = f(w_{11}^2 a_1^2 + w_{12}^2 a_2^2 + w_{13}^2 a_3^2 + w_{14}^2 a_4^2 + b_1^2)$ (2.9)

Where $h_{w,b}(x)$ is a real number representing the output of the ANN.

The activation function $f(\cdot)$ can be applied to vectors in element-wise as $f([z_1,\ z_2,\ z_3,\ z_4]) = [f(z_1), f(z_2), f(z_3), f(z_4)]$.

Therefore equation (2.9) can be written as (2.10).

$a^1 = x,$

$z^2 = w^1 a^1 + b^1,$

$a^2 = f(z^2)$ (2.10)

$z^3 = w^2 a^2 + b^2,$

$a^3 = f(z^3)$

$z^4 = w^3 a^3 + b^3$

$h_{w,b}(x) = a^3 = f(z^3)$ (2.11)

So, for any given layer l with activationa^l, the activation a^{l+1} of the next layer $l+1$ is obtained as:

$z^{l+1} = w^l a^l + b^l,$

$a^{l+1} = f(z^{l+1})$ (2.12)

When the computation of the signal moves from the input to the output of the forward network, it is called Forward Propagation. To make the network recurrent, the ANN could have a closed-loop back to itself from a neuron. Also, when an ANN has every neuron in each layer connected to the neurons in the next layer as in the developed structure in figure 2, it is called feed forward neural network (FFNN). A nonlinear activation function is used in FFNN networks as sigmoid functions which are like the logistic function as shown in Figure 3.

$\sigma(z) = \dfrac{1}{1+e^{-z}}$ (2.13)

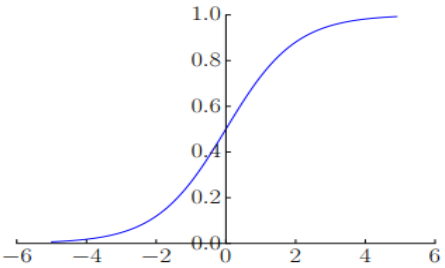


Figure 3: Non-linear Logistic Activation Function

From Figure 3, if z is large, then e^{-z} tends to zero (0), so $\sigma(z) = 1$. Conversely, if z is a small or very large or very large negative number, then e^{-z} tends to one (1), so $\sigma(z) = 0$. The essence of this is to ensure that only real statistical values between 0s and 1s are extracted and feed forward to the next hidden layer for training.

2.4 Review of Past Related Literatures

Some researchers have carried out many works on document verifications. [2] presented a study on the use of convolutional neural network for handwritten document recognition. In the study, the deep learning technique was used to extract handwritten features from documents and the use the vectors as a convolutional algorithm for the recognition of feature test document. The limitation of the study is that only hand written document was considered in the scope.

[3] 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.

[4] 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.

[5] presented research on online handwritten document recognition system for tail document. In the study, systematic review was done to evaluate the various techniques used in the past for document analysis and then the Tamil script was used as the study for the development of an improved document recognition system. The developed system when tested was able to recognize Tamil.

[6] presented a study on diagonal-based feature extraction technique for the recognition of handwritten document using artificial neural network. In the study, data of office documents were collected and used to train a neural network algorithm and then deployed for the classification of documents. This system was used for the notarization of documents accurate but despite the success is limited to only office documents.

[7] researched on new cursive approach for the verification of bank cheques. From the study, it was observed that signatures on bank cheques are easily forged and used to steal from customers. The research developed a system which extracts handwritten character on the cheques documents and then used the features for classification and training of the new system. The result when tested showed improved recognition performance but despite the success was limited to only detect bank cheques.

[8] 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.

[9] 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.

2.6 Summary of the Past Related Literatures/Research Gap

From the literatures reviewed, it is surprising that despite the huge problems of fraud due to falsification of certificates, receipts, among other issues of impersonation and theft, etc., it is seen that very little research has been done to improve authenticity of certificate verification in the public and private sector.
In fact, the researcher can boldly say that no reliable system is in the market for verification of certificates. Although, there are various methods for the verification of certificate documents as identified in the literatures but all have their weak-points which center on inefficiency and unreliability. The main challenge with the artificial intelligence method which presents a promising result compared to other methods is its difficulty in getting training data quality and algorithm. This research proposes to address this problem using image processing and neural network to improve on quality of data and classification decisions.

III. MATERIALS AND METHODS

3.1 Materials

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
Hardware 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

3.2 The Methodology

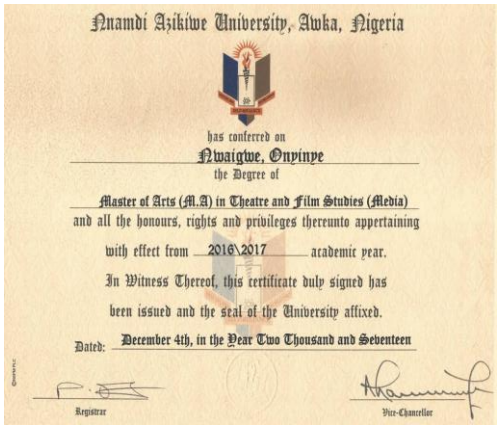
The Methodology used for the development of the proposed system are image processing, artificial intelligence and object oriented analysis methodology. The various processes used are data collection, data processing, optical image acquisition, binarization, segmentation, feature extraction, artificial intelligence technique, training and classification. These are to be explained here-under.

3.2.1 Data collection

One of the major problems with this research as identified in the previous chapter is the unavailability of dataset for this research. This is because most organizations and firms protect their data with strict legal ethics which makes it very difficult for access. However, to conquer this issue and provides data for the research, the researcher collected data from the Unizik Exams and Records department, based on the Letter of Permission written to the Registrar of the School (see the copy of the Letter in appendix A).

However, the primary source of data 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 sample size of data collected here is 1120 which consist of 970 first degree certificates and 150 postgraduate certificates. These data collected in the primary source was integrated together as the training dataset alongside 40 data collected from Alumni.

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 4.


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 4: The Data Sample

3.2.2 Data Processing

The data collected are of various properties such as size, content, quality, and colour, among other characteristics which are non-uniform. To format this 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.

3.2.3 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 does 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.

3.2.4 Image Binarization

Image binarization is found useful in many image processing applications due to its simplicity and effectiveness. This process was used for the mitigation of excess color frequency in the data into bi-level as in black and white. This way the quality of data being queried will be the same with that of the training dataset.

3.2.5 Segmentation

This operates iteratively by grouping together pixels which have similar values and splitting groups of pixels which are dissimilar in value. The aim is to highlight and identify the important characters on the data such as text, signature, stamps, handwritten and label for feature extraction.

3.2.6 Feature Extraction

Feature extraction is the process of extracting the pertinent features from data to build feature vectors which are then employed by classifiers to identify the input unit with objective output unit. This process simply extracts the important feature vectors with focus on the hand written text pattern and signatures from the data into a compact feature vector in statistical format.

3.2.7 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.

3.2.8 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 5.

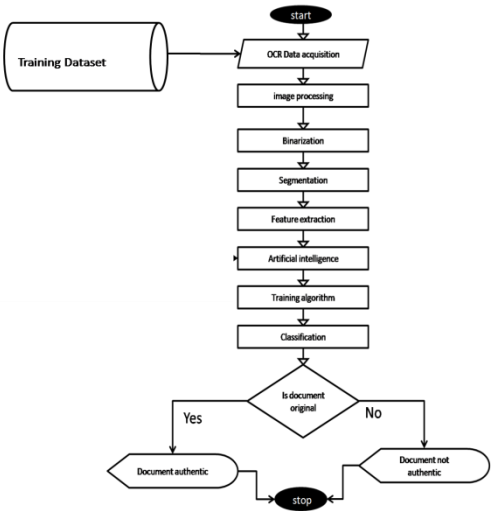


Figure 5: The Logical Dataflow Model of the Processes

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.

3.3 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 6.

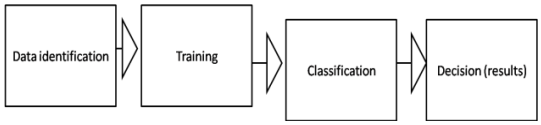


Figure 6: Block Diagram of the ANN Operation

The ANN operated in four basic steps as submitted using the block diagram in figure 6. 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 7.

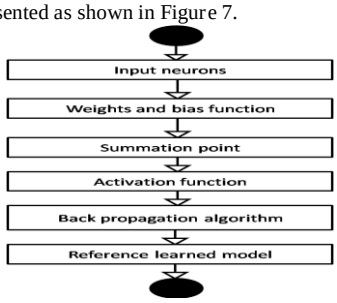


Figure 7: The Activity Model of the ANN

The model in the Figure 7 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.

Table 1: Neural Network Parameters

Parameters	Values
Train epoch values	16
The network hidden layers	10
Max epoch values	30
No. delayed reference input	2
Maximum feature output	3.1
Number of non-hidden layers	10
Maximum interval per sec	2
No. delayed output	2
No. delayed feature output	2
Minimum reference value	-0.7
Maximum reference value	0.7

To develop the model, there is need for document classifications for the neural network model in Figure 7. This is fed with the training dataset collected which consist of original certificates from the Unizik, and then trained as shown in Figure 8 to achieve the Reference Modelling diagram of Figure 9.

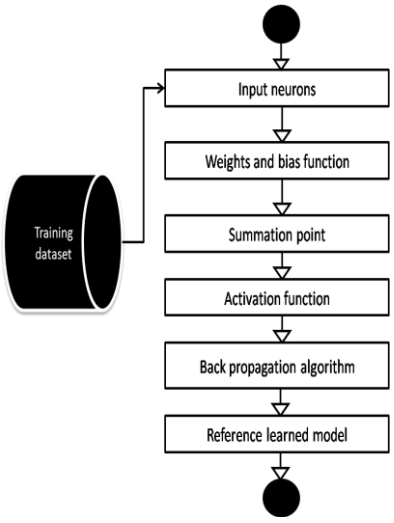


Figure 8: The Train ANN model

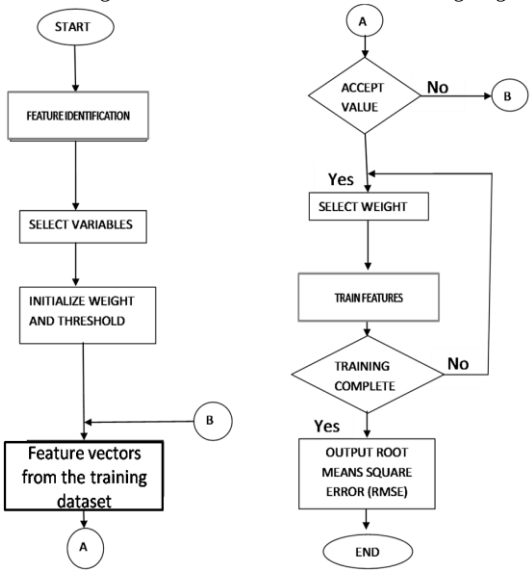


Figure 9: Training Algorithm

In the Figure 8, 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 9 to learn the feature vectors and achieve a reference classification model. The Pseudocode for the Machine learning algorithm developed is presented below.

The Pseudocode:

Start
% Initialize neural network parameter
Initialize random weight and bias function
Specify learning rate at 0.001
Initialize error rate at 0.001
% Load data of document extracted with features (Signature and handwritten text)
Input the converted m.file
System identification of the m.file from training dataset
Initialize activation function
Call training algorithm in figure 9
Split data into multiset of (70:15:15) and train
Generate reference model
Get output of training, regression and mean square error
Stop

3.4 System Modeling

In order to realize the model, Unified Modeling Language (UML) was employed. This modeling technique uses diagrams to document an object-based decomposition of systems revealing the interactions between these objects and their dynamics. UML aims to provide a common vocabulary of object-based terms and diagramming techniques that is rich enough to model any system development project from analysis to design. For our modeling, we make use of use case diagram, activity diagram and class diagram to design the new system.

3.4.1 Use Case and Domain Analysis

Use Case Universal Modeling Languages are deployed in research to explain in details the main components of the requirement definition. They buttress the activity through which the system will satisfy the aforementioned functional requirements and would then be deployed in constructing the process model which explains the operations (user action) in a more formal manner. The process uses diagrams to document an object-based decomposition of systems showing the interaction between these objects and their dynamics. The author's objective here is to provide a common vocabulary of object-based terms and diagramming techniques.

Use case diagrams give a user point of view of the new system with different users referred to as the Actors. Here in the User Case diagrams below the actors includes: new admin user, existing admin user, user while the supporting actors includes the suspect in possession of the query certificate.
The Use Case diagram for data collection is shown in Figure 10.

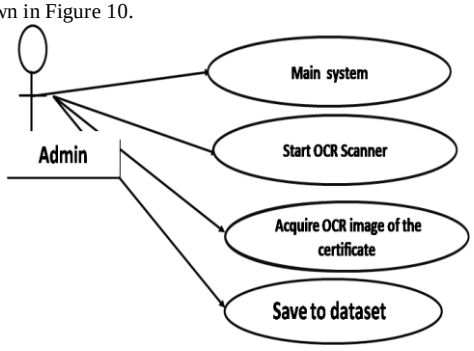


Figure 10: Use Case diagram for data collection

Primary actors: Admin
Brief description of event: The modeling diagram explains how the admin collected the data needed for the authentication. This was done using OCR system to collect the data in a software-based image format and then saved in the testing dataset for verification process.
Pre-conditions: It was assumed that the OCR device is connected with the monitoring PC installed with the image processing software.
Post-conditions: The saved image will be uploaded from the testing dataset for verification.
Main flow of events:

1. The admin gains access to the main system software

2. The image acquisition system was initiated for image scanning
3. The OCR Scanned the document into image format
4. The data collected is saved in the testing dataset.

The Use Case Diagram for Data Verification is shown in Figure 11.

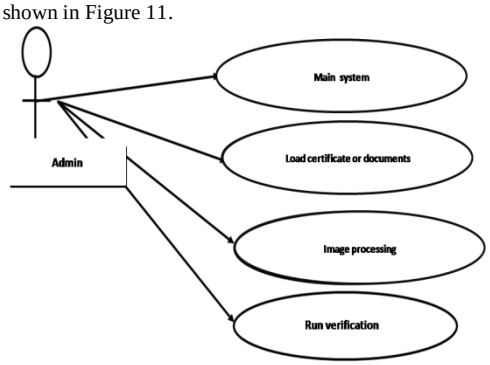


Figure 11: Use Case Diagram for Data Verification only

Primary actor: Admin

Brief description of event: From the modeling diagram in Figure 11, the admin loads the suspected certificate or document into the software; then the verification process was done to check if the document is real or fake.

Pre-conditions: It was assumed that the query data was already collected from the suspect and then the software was already trained with training dataset of authentic documents.

Post-conditions: The recognition process when completed presented a decision that the document is real or fake.

Main flow of events:

1. The admin gains access to the main system
2. Uploads the query data for verification
3. This is processed for verification.

The Use Case Diagram for Suspect Data Collection and Verification is shown in Figure 12.

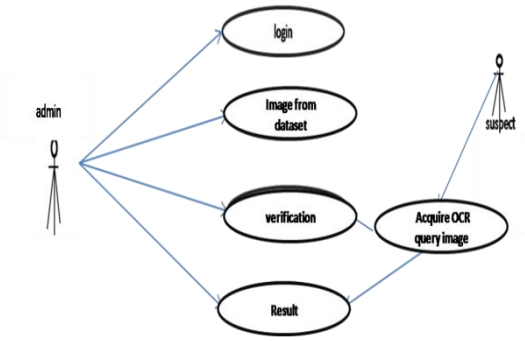


Figure 12: Use Case Diagram for Suspect Data Collection and Verification

Primary actor: The admin

Secondary actor: Suspect (person)

Brief description of event: From the modeling diagram above the admin loads the OCR data of the training set into the system for learning, then the query certificate is uploaded and then tested by the system for authenticity, while the result is popped out after the verification.

Pre-conditions: We assumed that the admin already has a confirmed and authentic document to be in the training dataset to learn the artificial intelligence system with the feature vectors before comparing with the query certificate.

Post-conditions: the result when verified presented a match document report of otherwise.

Main flow of events:

1) The admin gains access to the main system

2) Uploads the training and query certificate to the software for verification

3) Runs verification

4) Results

3.4.2 Activity Diagram for the Proposed System

The activity diagram describes the process flow among the multiple objects of a class during the activity processing. They are used in association with the UML modeling methods with a major objective of molding templates for workflow behind the system being developed. It is employed in describing a use case by giving details of a complex algorithm as a required action that is needed to take place and when they are to occur. The figure 13 shows the

possible activity diagram of the new system for verification of document.

To achieve this, the machine learning algorithm, which in this case is the artificial neural network was used to train the training dataset containing feature vectors of original data extracted and then stored as a reference model for classification. When query document or certificate from the testing data set is uploaded, the data is processed using the necessary methods such as binarization, segmentation, feature extraction and then fed to the ANN (A.I algorithm) for classification with the reference learned model already stored. The classification result then presents if the document is original owing to the similarity index with the reference model; or if not original, this is when there is no similarity with the feature vectors of the original documents. The activity diagram is presented in the figure 13.

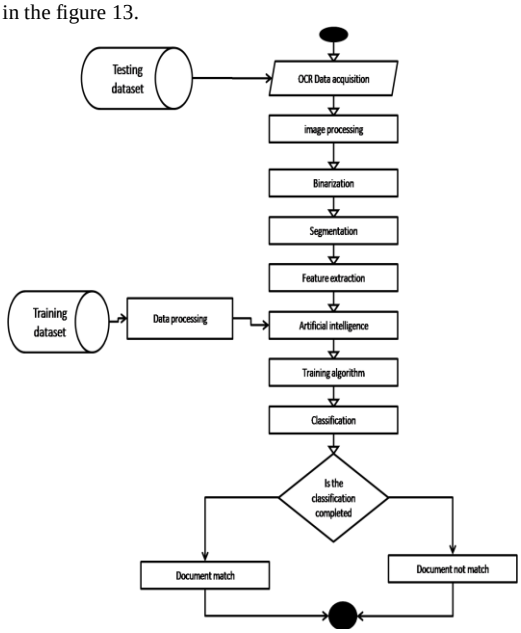


Figure 13: Activity Diagram of the Proposed System

3.4.3 The System Flowchart

The system flowchart shows the data flow of the entire system operation. In the flowchart, the neural network algorithm developed with the training dataset was used as the classification model to train and compare testing documents as shown in Figure 14.

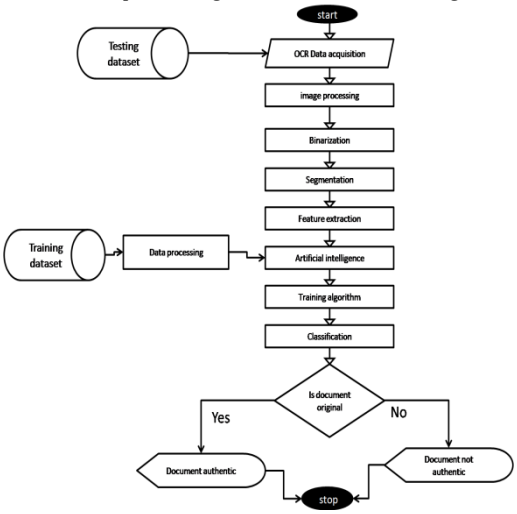


Figure 14: The System Flowchart

From Figure 14, data from the testing dataset are processed using binarization, segmentation, feature extraction and then feed forward to the system for training and classification using the already learned reference model. The classification model was used to make decisions whether the document is original or fake.

3.5 The System Implementation

To implement the system, the artificial neural network algorithm developed with the training dataset was implemented using neural network toolbox as shown in Figure 15.

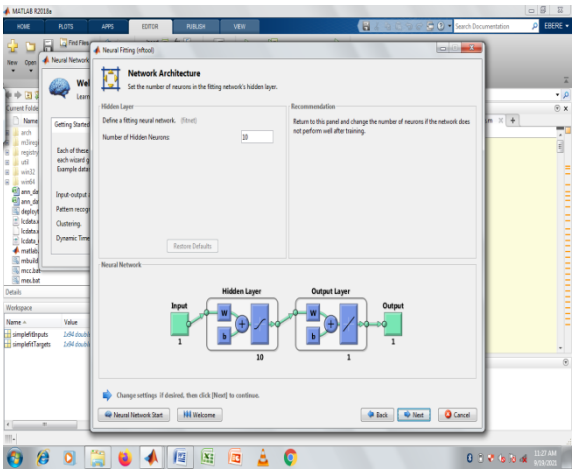


Figure 15: The ANN Tool

The tool is a neural network app in MATLAB used for the development of ANN algorithm. To achieve this, the ANN was loaded with the training dataset, and then the tool automatically configured the network architecture based on the data configuration of two input classes for documents and certificates. Then the number of hidden layers and training algorithm were introduced based on the training parameters in table 1 which enables the network to train and generate a reference model deployed in the MATLAB. The MATLAB environment was used to develop the digital document verification system as shown in Figure 16.

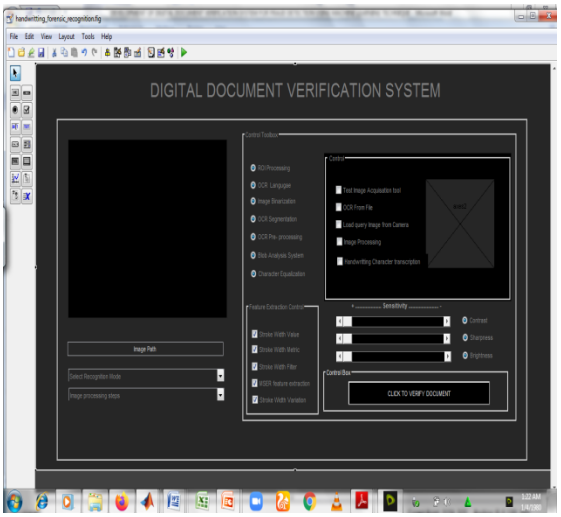


Figure 16: The MATLAB Programming Environment for the System Development

IV. RESULTS AND DISCUSSIONS

This section presents the performance of the trained algorithm developed using regression analysis; mean square error and confusion matrix. The algorithm when deployed as the intelligence document verification system was also tested and the result presented and discussed here.

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 17.

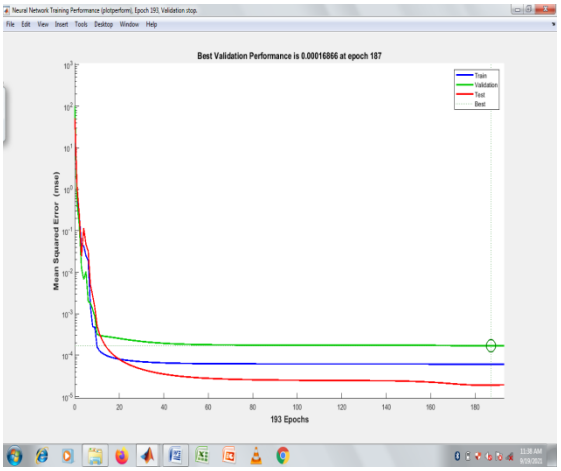


Figure 17: Mean Square Error (MSE) Performance

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 18.

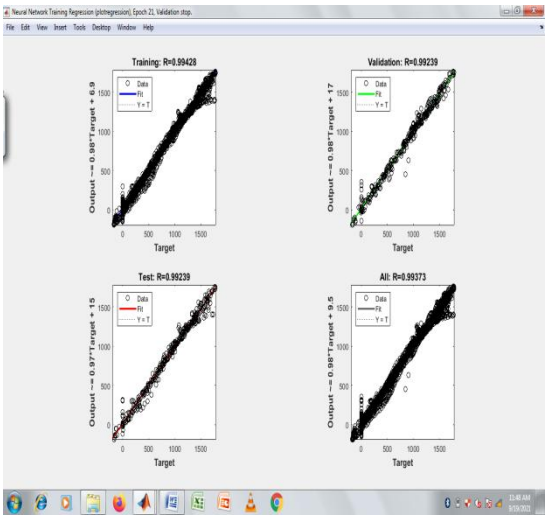


Figure 18: Regression Result

This regression was used to evaluate the performance of the training process. The aim is to achieve a regression value equal or approximately one. From the result it was observed that the mean regression value which was used to evaluate the overall performance of the multiset for the training, test and validation is R = 0.99373. The result implies that the system was reliable with high ability to detect fake or real document when deployed for verification.

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 19 presents two sample data of training and testing data which was used to evaluate the performance of the document verification software developed. The figures 19(a) and (b) are the training document sample and the testing document sample for verification.

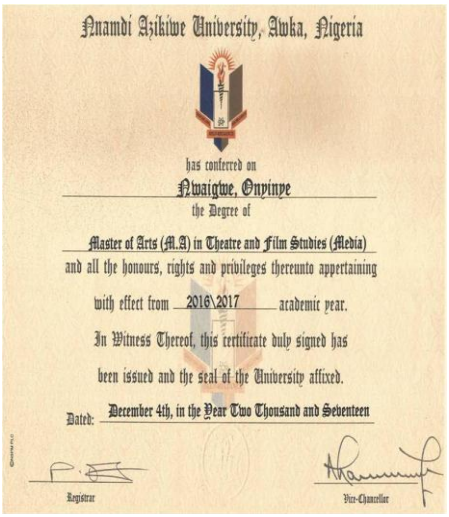


Figure 19(a): Query Document Sample

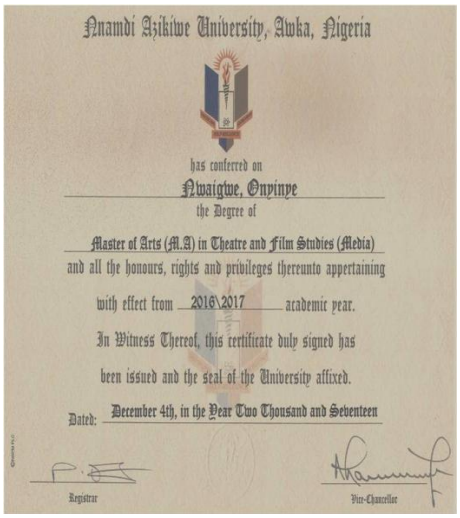


Figure 19(b): Testing Document Sample

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 as shown in figure 20.

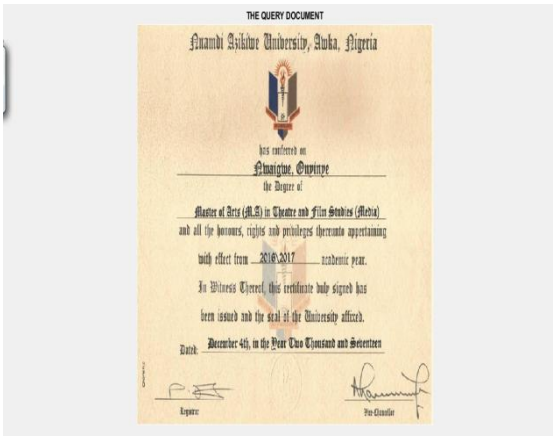


Figure 20: Testing Certificate

The Figure 20 presents a certificate from the case study institution (Unizik) uploaded into the software for verification. The system identified it as a query image and preprocessed using binarization which eliminated the various colors on the image which are unnecessary due to noise and then leaving a bi-color image as shown in figure 21.

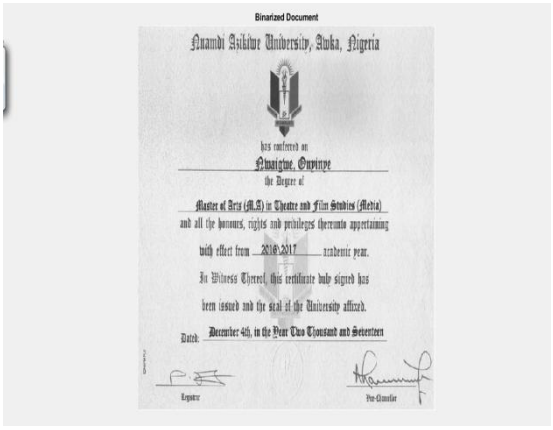


Figure 21: Binarization Result

The figure 21 presents the result of the binarization process which was the first pre-processing step to improve the query image quality for segmentation

which enhances the image features into a skeletal format. The result of the segmentation is presented in figure 22.

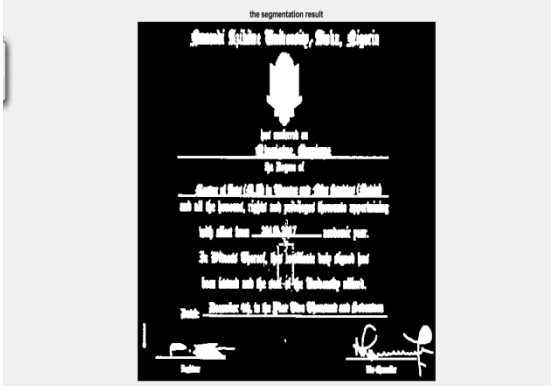


Figure 22: Segmentation Result

The aim of the segmentation process is to provide an image with enhanced feature for easy extraction by the neural network and classification. The classification compared the patterns in the input with the pattern of the training features as shown using the figure 23.

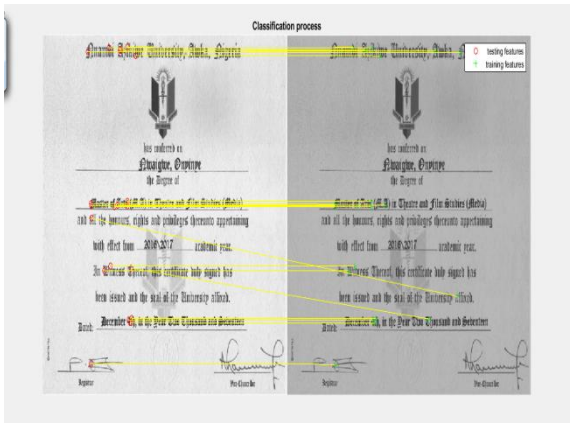


Figure 23: Classification Result

Here the feature of the training and query data were classified based on the pattern recognition and then decision was intelligently made showing that despite the degrade in the quality of the query image due to certain environmental factor, that the software was still able to recognize it as original as shown in figure 24.

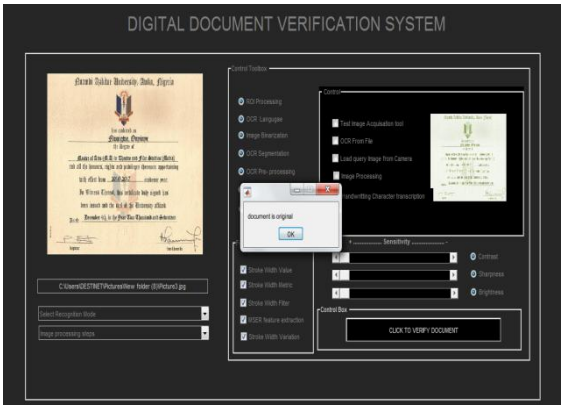


Figure 24: Result of the Verification Process.

The result showed the decision made by the intelligent document verification system developed which correctly detected the certificate as an original copy.

The next result presented a case where the certificate was fabricated as shown in the figure 25(a) and (b) alongside the original trained document.

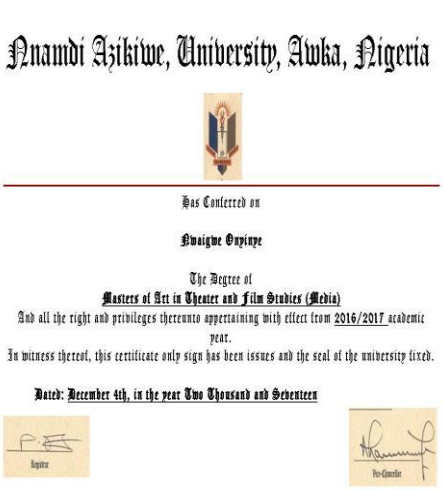


Figure 25(a): Fabricated Query Certificate

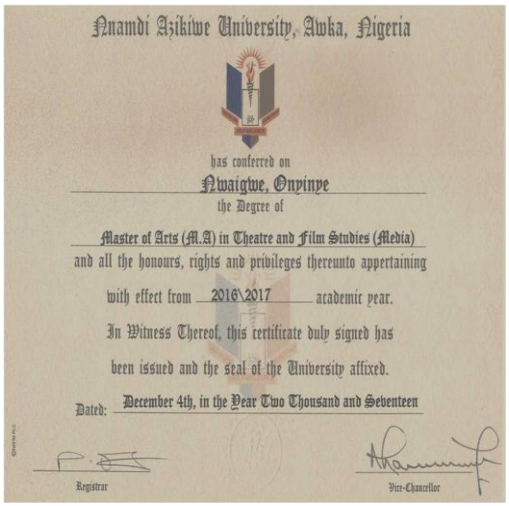


Figure 25(b): Trained Sample

The figure 25 presented the performance of the software when used to test a certificate which was fabricated by the researcher for demonstration sake using Microsoft certificate publisher. The uploaded sample was identified by the software and the result is presented in figure 26.

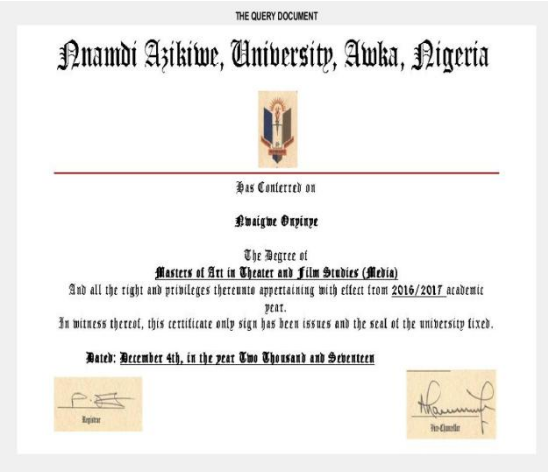


Figure 26: Sample of the Query Certificate

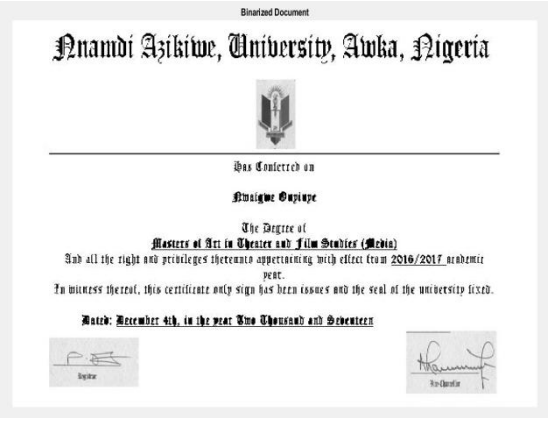


Figure 27: Sample of the Binarized Result

The figures 26 and 27 present the sampled query image and the pre-processing result. The pre-processed output was segmented and then trained for classification as shown in figu**re 28 for segmentation, and the figure 29 for the verification result.**

Figure 28: Segmentation Result

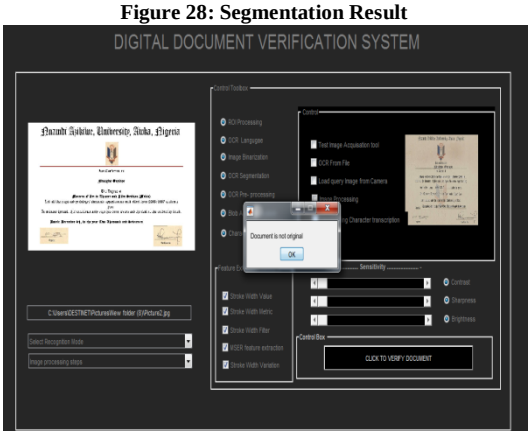


Figure 29: Verification Result

4.3 System Validation

To validate the result, tenfold validation technique was used to validate the training performance and regression of the document verification process. This was used by iterating the training and testing process ten times, using the neural network training toolbox.

Table 2: Validation Result

Iteration	Mean Square Error (MSE)	Regression
1	0.00101	0.98373
2	0.00103	0.99373
3	0.00105	0.99749
4	0.00098	0.99773
5	0.00103	0.99339
6	0.00101	0.99350
7	0.00095	0.99443
8	0.00097	0.99339
9	0.00099	0.99753
10	0.00107	0.99179
Average	0.00100	0.99373

From the result in table 2, it was observed that the average MSE is 0.00100Mu which is approximately 0Mu. The implication of the result is that the margin for error in the new system developed is almost none. The regression performance was also used to support the result when the average result was computed and the value for R = 0.99373 which implied that the system is reliable and very effective for the verification of documents.

V. SUMMARY, CONCLUSION AND RECOMMEDATION

5.1 Summary

The study has successfully presented an intelligent system for the verification of documents using artificial intelligence technique. This was done to combat the increased fraud and the rate at which documents are fabricated and manipulated all over the world today for many selfish reasons, in order to gain wealth, fame, employment, among others. This study collected data from the bursary department, Unizik; and then developed an intelligent document verification algorithm and deployed as an expert system using Matlab. The system was tested and the result showed high rate of verification regression and MSE performance. The implication is that when deployed, it was able to recognize and verify results accurately.

5.2 Conclusion

All over in Nigeria, there is no reliable system which can be used for verify documents in both public and private establishments. The conventional means of achieving this aim is via manual approach which takes lots of time, processes and still prone to all sorts of manipulations due to the increased rate of corruption within the country. Today, corruption has eaten deep into the helm of our affairs and things have fallen apart. As a result, the conventional means to verify documents despite the other limitations is not reliable as it lacks integrity. To solve this problem, this study has collected all these documents from the case study institution (Unizik) from 2016 till 2020 and used to train, and then neural network algorithm designed and then deployed an intelligent document verification expert system. This was tested and the result achieved showed that it is reliable with high regression value of R = 0.99373 and MSE = 0.000100.

REFERENCES

[1] Ayush Ruhu Purohit and Shardul Singh Chauhan Rusu, “A Literature Survey on Handwritten document Character Recognition”, International Journal of Computer Science and Information Technologies, Vol. 7 (1), 2015, Rusu 1-5.

[3] Gaurav K. and Bhatia P. K., “Analytical Review of Preprocessing Techniques for Offline Handwritten document Character Recognition”, 2nd International Conference on Emerging Trends in Engineering and Management, ICETEM, 2013.

[4] Gollin, G. D., “Verification of the Integrity and Legitimacy of Academic Credential Documents in an International Setting”, New York: AACRAO, 2014.

[5] Jordan, M.I and Mitchell, T.M., “Machine learning: trends, perspectives, and prospects. *Science* 349, 255–260. doi: 10.1126/science.aaa8415, 2015.

[6] Linus U., “Tackling the rise of fake qualifications in Nigeria”; Politics and society; this Africa Publisher, 2017.

[7] Namukose Mpongo Sarah, “Design Of A Secure Online Academic Document Verification System”, Nkumba University, 2018.

[8] Ntirogiannis, Konstantinos, Basilis Gatos, and Ioannis Pratikakis, "A Performance Evaluation Methodology for Historical Document Image Binarization." : 1-1, 2017.

[9] Russell, S. and Norvig, P., “Artificial Intelligence: A Modern Approach”, 3rd Edition. New Jersey, Prentice Hall, 2010.