

Cyber Security Experts Association of Nigeria (CSEAN)  
Society for Multidisciplinary & Advanced Research Techniques (SMART)  
Faculty of Computational Sciences & Informatics - Academic City University College, Accra, Ghana  
SMART Scientific Projects & Research Consortium (SMART SPaRC)  
Sekinah-Hope Foundation for Female STEM Education  
ICT University Foundations USA

---

---

Proceedings of the Cyber Secure Nigeria Conference – 2023

---

---

## Development of a Fake News Detection Model Using Decision Tree Algorithm

Adeola, Olajide O., Bello, Ganiyat R., Oluwasola, Babatunde S. and Lateef, Kafayat R.

Department of Computer Science  
Oyo State College of Agriculture and Technology  
Igboora, Oyo State, Nigeria

**Corresponding Email:** ooadeola@gmail.com

**Phone Nos:** +2347033261184 or +2348051242349

### ABSTRACT

Over 500,000 Americans are influenced each year by fake news, which has an impact on society. Researchers are examining data on fake news identification and looking into machine learning models for a quicker and more accurate categorization as part of their investigation of false news detection techniques. This study's main objective is to examine, contrast, and evaluate how well three distinct machine learning algorithms do at spotting fake news. Decision tree, Random forest, and logistic regression are the machine learning algorithms. There are 21418 datasets available for real news, compared to 23503 for fake news. The open-source Kaggle dataset, which was selected for the project, served as the source of both datasets. The results of the construction and comparison showed that the decision tree model did the best forecasting with a maximum accuracy of 99.64% and the random forest model performed well with an accuracy of 98.89%. However, the logistic regression model did well in this assignment with an accuracy of 98.88%. Using these models has the critical goal of improving the news verification system's accuracy and dependability.

**Keyword:** Fake news, Verification System, Machine Learning Model.

---

---

#### Proceedings Citation Format

Adeola, O.O., Bello, G.R. Oluwasola, B.S. & Lateef, K.R. (2023): Development of a Fake News Detection Model Using Decision Tree. Proceedings of the Cyber Secure Nigeria Conference. Nigerian Army Resource Centre (NARC) Abuja, Nigeria. 11-12<sup>th</sup> July, 2023. Pp 81-88. <https://www.csean.org/>. dx.doi.org/10.22624/AIMS/CSEAN-SMART2023P10

---

---

### 1.. INTRODUCTION

Fake news poses a significant challenge for forensic specialists in the news industry, potentially leading to extortion and character assassination. Forgers employ sophisticated techniques to produce fake news.

Forensic investigators can tell whether a similar method or material was previously used to fake a news item by consulting an up-to-date database of previously falsified news pictures. Therefore, it is feasible to learn more about the origin of fake news by describing how it operates. Multinational businesses in banking, public, and private sectors have adopted computer-based strategies, including digital news, which has increased consumption. IDC reports 36% of people used digital media in 2018.

However, bogus or falsified news is also growing rapidly. Every year, fake news has an impact on 500,000 Americans and results in a \$750 million financial catastrophe, (Khan, Shafait, and Mian, 2013). Press releases sometimes include consumer photos, seals, holograms, words, and stamps, making it difficult to identify bogus news. Checking security components is essential. The process of news verification is being outsourced to a third party Business Process Outsourcing providers (BPO) by numerous multinational corporations, private businesses, and governmental bodies. According to a recent analysis from Thomson Reuters, 32 billion US dollars will be spent on outsourced BPOs for "Know Your Customer" (KYC) verification in 2017. Automating news verification reduces costs and labor, enabling firms to focus on core capabilities and focus on core capabilities, resulting in a more efficient commercial solution.

To check all the security elements of the news, we developed an automatic news verification system based on the Decision Tree algorithm in this research project. The model enables users, customers, and workers to post digital news directly to the verification system, which contributes to the creation of a quicker and more precise counterfeit news detection system utilizing the Decision Tree algorithm. Digital news verification challenges include detecting counterfeit news using techniques like watermarks, seals, and holograms. However, scanning physical news using scanners or mobile phones reduces their quality. Verifying news authenticity is challenging, but two-dimensional barcodes and QR codes offer a visual, machine-readable format for verification. However, barcodes can be easily edited or replaced.

Pawel and Andrzej (2018) developed a method to detect fake documents and news using image processing and pattern recognition techniques. They used shape-based algorithms to retrieve rubber stamps from scanned news. Image processing techniques verify signature authenticity in scanned news. Muhammad et al. (2014) proposed a method using local features for signature stability analysis, achieving a 15% error rate. Hamadene et al. (2017) developed a one-class writer-independent system for classifying handwritten signatures using feature dissimilarity measures and contour-let transform-based DCCM for feature generation, without relying on machine learning classifiers. Yanzhi, Wang, Chen, and Yang (2020) developed a signature verification system using critical segment method for secure mobile transactions, capturing intrinsic signing behavior and geometric layout. Luiz, Robert, and Luiz (2019) used convolutional neural networks for signature classification, while Nabil and Awal (2018) proposed a method for verifying personal identity using pattern matching and recognition techniques.

Traditional forgery news detection methods focus on features appearances, ranging from 62.6% to 85.09%, without verifying manipulated regions. High need for forgery detection in scanned news; no machine learning-based automatic verification system detects all possibilities. CNN algorithm checks news for discrepancies and forgeries using OCR and LBP to retrieve crucial textual information, including names, identification numbers, and scores.

The OCR system converts physical news into machine-readable format by converting it into grayscale images and identifying dark and light areas as characters or letters. By partitioning photos into patches and stacking auto encoders, Zhang, Goh, Win, and Vrizlynn (2016) provide a two-step method for identifying faked images using machine learning approaches.

A machine neural network technique to identify faked images was developed by Hou, Jang, Park, and Lee (2018) using context learning and region-convolutional networks. The region-based object detector's inability to recognize the backdrop, however, causes the accuracy to be low. Hyperspectral imaging was employed by Khan, Shafait, and Mian (2013) to recognize actual handwritten notes. Ink color mismatches are discovered using hyperspectral imaging in news articles to confirm authenticity. If the letters match, the entire story is not falsified; but, if there are any mismatches, a particular section of the story is revealed to be forged.

Using the Copy Move Forgery identification algorithm for textual forgeries, Abramova, Svetlana, and Bohme (2016) created block-based near duplicate identification in scanned news, and Khan, Shafait, and Mian (2013) increased ink mismatch detection accuracy. In contrast, Maryam et al. (2013) offer a text fraud detection method utilizing CNN for efficient characterization of source printers using machine visual cues. Fuzzy C-Mean clustering separates ink pixels in handwritten notes into many clusters. Using SVM and other techniques, Tsai et al. created a system to identify fake text in printed news.

Visual artifacts, blockiness, and blur are used to identify forged sections in the methods for identifying image splicing proposed by Manu and Mehtre (2015) and Ambili and George (2016). LDA is used to distinguish between authentic and fabricated areas. To combat image splicing forgeries, Salloum, Ren, and Kuo (2018) designed a multi-task MFCN. Su, Ueng, and Chung (2019) employed SVM to distinguish between authentic and fabricated regions and the Hough transformation to detect imprint borders and edge differences.

Research focuses on the identification of textual and visual forgeries in scanned news, which includes all varieties of news from the public and commercial sectors. In order to address current issues, this work provides a quick, precise counterfeit news detection method employing CNN and sliding window techniques. We developed and tested a Fake News detection model using a decision tree approach that had a 99.4% accuracy rate.

## 2. METHODOLOGY

The suggested technique for forgery detection analyzes news forgeries more accurately and in less time. Kaggle's ISOT Fake News Dataset, an open-source dataset appropriate for machine learning models, is used for data collecting. The created model was preprocessed, cleaned, trained, tested, and evaluated using Python, Jupyter, and Number tools. Filling and thinning techniques were utilized to enhance character quality. The user interface system rejects low-resolution news, necessitating high-quality content. Preprocessing, extraction, and image format conversion are done on scanned news. The algorithms used are Logistic Regression, Random Forest and Decision Tree.

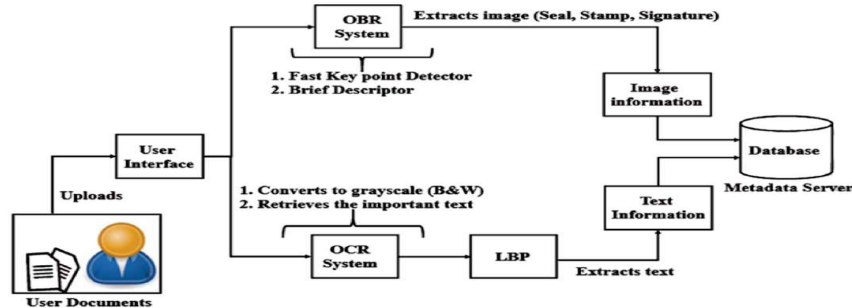


Fig. 2.1: Google Chrome Architecture of Feature Extraction and Detection.

Analyzing the local or nearer edges of the letters, the changes in the region can be identified. To find out the edges of the counterfeited letters, Local Binary Pattern (LBP) algorithm is used along with the OCR. Algorithms used in the feature extraction phase is shown in Fig. 3.1.

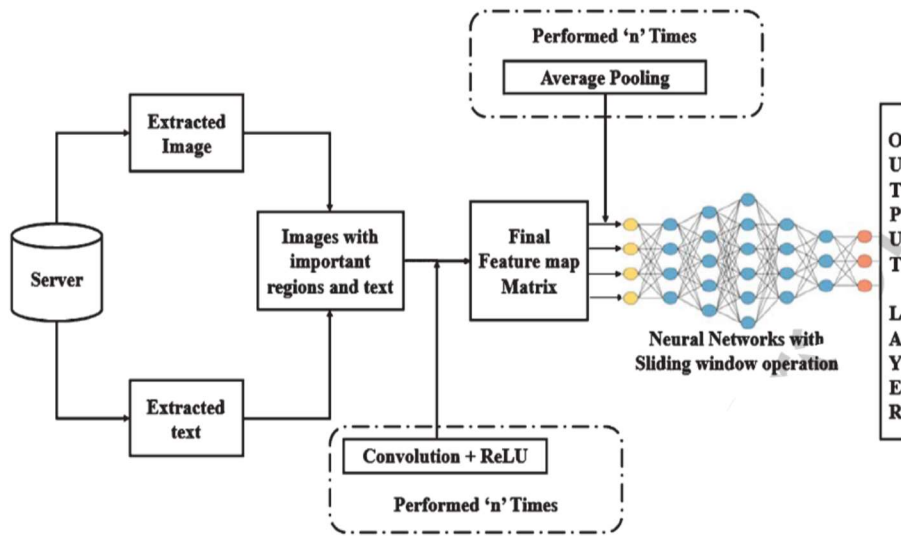


Fig. 2.2: Google Chrome Workflow of Sliding Window Based CNN Technique.

### 2.1 Logistic Regression:

Using a machine learning process called logistic regression, text can be classified into multiple or binary classes based on binary alternatives.

$$hg(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} \quad (1)$$

The cost function is minimized in order to achieve the optimal prospect and therefore the cost function is calculated as the cost function is minimized in order to achieve the optimal prospect and therefore the cost function is calculated as

$$Cost(h_g(x), y) = \begin{cases} \log(h_g(x)), y=1 \\ -\log(1-h_g(x)), y=0 \end{cases} \quad (2)$$

## 2.2 Decision Tree

A supervised learning technique called a decision tree starts at the root and provides class predictions for planned and retrospective challenges. By comparing the original attribute values to the real data sets, it chooses the subsequent node. Attribute Selection Methods (ASM) include the following:

i. Information Gain: Information gain establishes attribute order in decision tree nodes by measuring the availability of class informational features. Let information gain is denoted as IG and is calculated as given in the following expression.

$$IG = \frac{Entropy(S) \times (Weighted.Avg)}{\times Entropy(each\ future)} \quad (3)$$

Where

Entropy is the probability distribution of observations in the dataset belonging to one class or another. For example, in a two class ('Yes' and 'No') dataset, Entropy can be calculate

$$Entropy = \frac{(p(No) \times \log(p(No))) + (p(Yes) \times \log(p(Yes)))}{p(Yes) \times \log(p(Yes))} \quad (4)$$

ii. Gini Index: Gini Index measures the inequality among the values of a variable that is used in the construction of a decision tree in Classification and Regression Tree algorithm. Let Gini Index is denoted as GI and is calculated as given in the following expression.

$$G1 = 1 - \sum P^2 \quad (5)$$

## 2.3 Random Forest:

A supervised machine learning method called random forest uses the bagging principle to produce precise conclusions on huge data sets. With each decision tree operating independently, the outcome is determined by the class that received the majority of votes.

To summarize, the grid search used different numbers of scales to generate the most accurate model that could accurately predict the results.

$$G_{ind} = 1 - \sum_{i=0}^e (P_i)^2 \quad (6)$$

## 3. RESULTS

Model accuracy is improved by machine learning methods including decision trees, logistic regression, and random forests. Decision trees have the highest accuracy, at 99.64%.

### 3.1 Logistic Regression Result

A bogus value was predicted by an algorithm 4664 times, with 57 situations presuming it was untrue, according to the confusion matrix. The logistic model correctly predicted 35 occurrences of bogus values, yet the confusion matrix exposes them. However, 4224 real value cases were predicted, proving the model was successful in identifying true instances.

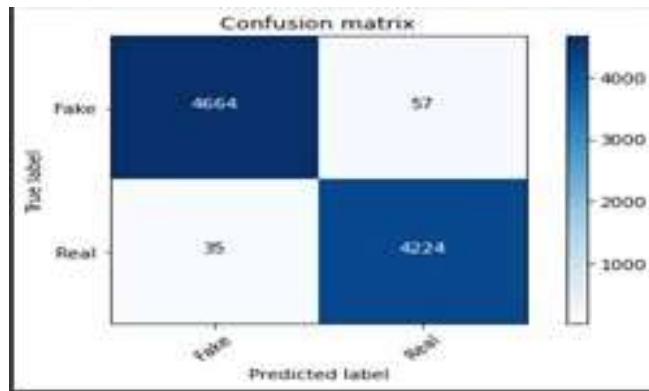


Fig 3.1: Logistic Confusion Matrix

### 3.2 Decision Tree Classifier

Out of 4254 instances, the Decision tree classifier model correctly identified 4242 instances as "true" while classifying 4709 instances as "fake," demonstrating its superior performance in identifying real occurrences utilizing a tree-like structure.

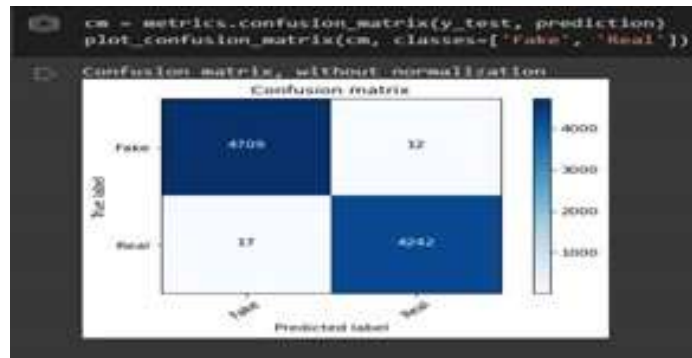


Fig 3.2: Decision Tree Classifier Result

### 3.3 Random Forest Classifier

The confusion matrix displays 54 occurrences of genuine values and 4667 instances of the algorithm anticipating false values. The Random Forest classifier model only properly predicted real values in 37 occasions. In 4222 examples, the Random Forest classifier model fared better at correctly predicting "real" values than decision trees, showing improved results in identifying actual "real" instances.

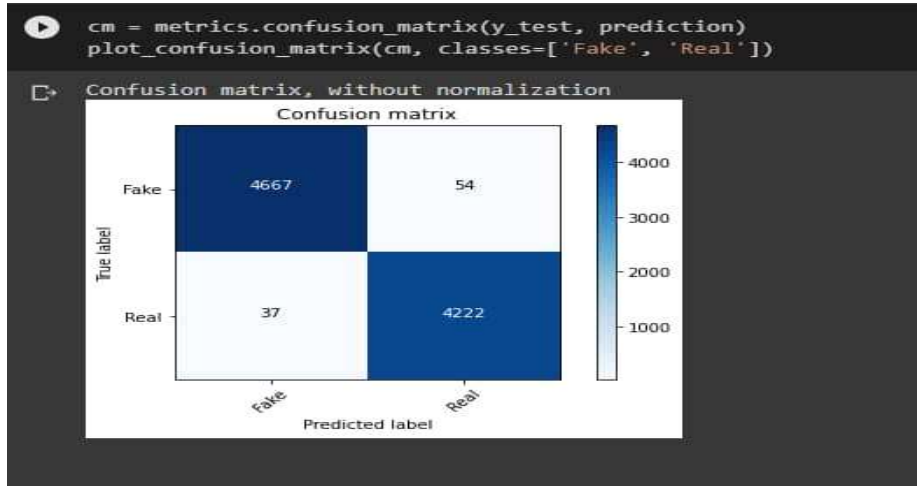


Fig 3.3: Random Forest Classifier Results

Table 3.1: Performance of the Algorithms

S/N	Algorithm	Accuracy
1	Logistic Regression	98.88%
2	Decision Tree	99.64%
3	Random Forest	98.89%

Comparison on accuracy of different algorithms on the same dataset is presented below

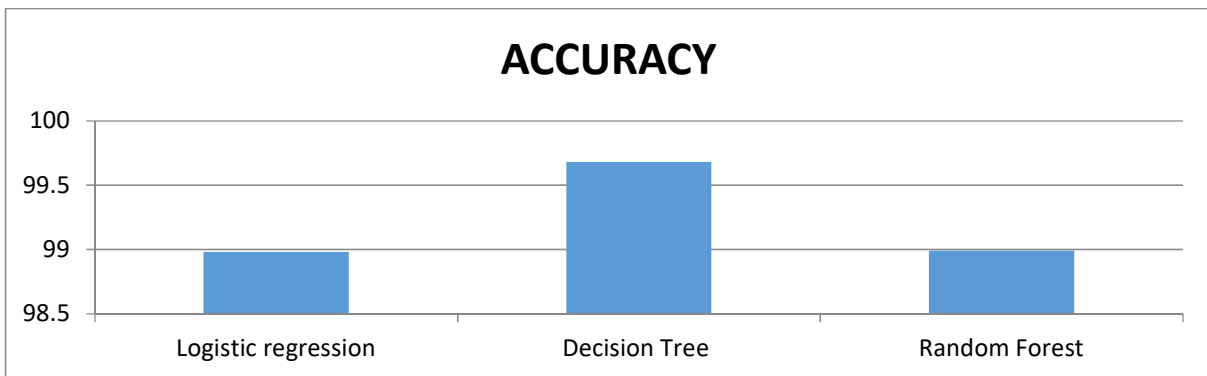


Fig 3.4: Comparison of Algorithms

#### 4. CONCLUSION

Financial crises can be impacted by new machine learning techniques that can identify fake news, reducing fake detection. Results for the decision tree, random forest, and logistic regression models demonstrate accuracy of 99.64%, 98.89%, and 98.88%, respectively. In a variety of industries, including banking, government, and law enforcement, machine learning models successfully identify bogus news, increasing reliability and accuracy.

## REFERENCES

1. Abramova, Svetlana and Bohme, R. (2016). Detecting Copy Move Forgeries in Scanned Text News, *Int. Conf. Media Watermarking, Security and Forensics*, (2016). pp. 1–9.790
2. Hamadene, A. and Chibani, A. (2016). One-Class Writer Independent Offline Signature Verification Using Feature Dissimilarity Thresholding, *IEEE Transaction on Information Forensics and Security* (2016), 1226–1238. 765
3. Khan, M.J, Yousaf, A., Khurshid, K., Abbas, A. and Shafait, A. (2018). Automated ForgeryDetection in Multispectral News Images Using Fuzzy Clustering, *IAPR Int. Work. News Analysis Systems*, (2018). pp. 393–398. 794 [doi:10.1109/DAS.2018.26].
4. Khan, Z., Shafait, F. and Mian, A. (2013) hyper spectral Imaging for Ink Mismatch Detection, *IAPR Int. Workshop on News Analysis and Recognition*, pp. 877–891. 783[doi:10.1109/ICDAR.2013.179]
5. Luiz, G.H, and Luiz, S.O (2019). Characterizing and Evaluating Adversarial Examples for Offline Handwritten Signature Verification, *IEEE Transaction on Information Forensics and Security*, 1–1. 773
6. Moffitt, K.C., Rozario, A.M and Vasarhelyi, M.A (2018). Robotic Process Automation for Auditing, *Journal of Emerging Technology in Accounting*.
7. Muhammed, I.M., Wick, L.I, Dengel, A., Uchida, S. and Frinken, V. (2014). Automatic SIGNATURE Stability Analysis and Verification Using Local Features, *Int. Conf. on Frontiers In Handwriting and Recognition*, pp. 621–626. 761
8. Nabil, G. and Awal, A.M (2018). A New Descriptor for Pattern Matching: Application to Identity News Verification, *IAPR Int. Workshop on News Analysis System*, pp. 375–380. 777
9. Partha, P.R, Umapada, P. and Joseph, L. (2010) Seal object detection in news images using GHT of local component shapes, *ACM Symposium on Applied Computing*, pp. 754 23–27.
10. Pawel, F. and Andrzej, M. (2018) Stamps Detection and Classification Using Simple Features Ensemble *Mathematical Problem in Engineering*, pp. 1–15. 757
11. Reinsel, D., Gantz J., and Rydning. J. (2018) the digitalization of world from the edge to core, White paper, IDC, 744.
12. Shang, S., Memon, N. and Kong, X. (2014) Detecting news forged by printing and copying, *EURASIP J Advanced Signal Processing*, 1–13. 786.
13. Teerakanok, S. and Uehara, T. (2019) Copy-move forgery detection: A state of the art technical Review and analysis, *IEEE Access*, 40550–40568.748
14. Yanzhi, R., Wang, Y.C, Chen, M.C., and Yang, J. (2020). Signature Verification Using Critical Segments for Securing Mobile Transactions, *IEEE Transactions in Mobile Computing*, 724-739. 769