Original Article
ISSN (Online): 3048-9245

ShodhAI: Journal of Artificial Intelligence
July-December 2024 1(1), 76–96

TRANSFORMING CLOUD COMPUTING DATA SECURITY WITH AN INNOVATIVE VIRTUALIZATION MODEL

Nseobong Archibong Michael [1], Friday E. Onuodu [2], E. E. Ogheneovo [2]

[1] Lecturer, Department of Mathematics and Computer Sciences, Ritman University, Mkpatak, Nigeria
[2] Lecturer, Department of Computer Science, University of Port Harcourt, Nigeria

Received 30 July 2024
Accepted 29 August 2024
Published 18 September 2024

Corresponding Author
Nseobong Archibong Michael,
nsemike@yahoo.com

DOI 10.29121/ShodhAI.v1.i1.2024.7

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

This study presented an optimized approach to Data Security Monitoring in Cloud Computing Infrastructure via an Improved Robust Virtualization Model. Malicious activities have continued to become an alarming issue in cloud computing. Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service-provider interaction. In this work, a new virtualization model for data security monitoring was developed. Structured System Analysis and Design Methodology was adopted and design was achieved with tools such as Dataflow Diagram, Use-Case Diagram, Unified Modelling Language (UML) Diagram, and Sequence Diagram. The study utilized Five Hundred (500) datasets from robust repositories, of which 30% was used for training, while 70% was used for testing. Parameters for analyzing and evaluating the results of both systems encompassed the number of adopted algorithms, the number of adopted technologies, the number of adopted design tools, and the number of tested records. From the performance evaluation, the new system showed better performance than the existing system as it achieved an accuracy rate of 1.07% as compared to the existing system which achieved an accuracy rate of 0.48%. The newly developed model was for fraudulent data detection in cloud computing infrastructure with a special emphasis on financial fraud. This is because; financial frauds are committed against property, involving the unlawful conversion of the ownership of the property to one's personal use and benefit. In addition, the new system was further optimized with a deep neural network and logistic regression technique. This study could be beneficial to anti-corruption agencies, corporate organizations, and researchers with keen interest in the study area.

Keywords: Logistic Regression Technique, Virtualization, Cloud Computing, Security Monitoring, Virtualization Architecture

1. INTRODUCTION

The worrying rate of data fraud in computing today has permeated even the most advanced cloud computing infrastructures. The absence of an improved virtualization model for data security is the root cause of this problem. Data fraud and other security issues are a persistent risk for people, groups, and businesses who use cloud computing and save their data online. Organizational data can now be accessed without the owner's knowledge or authorization, which increases the frequency of account breaches and hacking. The possibility of losing all of your data

saved on the cloud is nearly unavoidable due to the seriousness of the providers' data protection issues. Even at academic institutions, students occasionally break into school networks to change information, including adding their names to records without paying fees, which seriously disrupts the school's cloud records.

Researchers in a variety of academic disciplines, including computer science, have expressed serious concerns about the issue of data fraud. Many academics have attempted to give answers to this issue. In light of this, Omar and Jefffery (2015) conducted research on "challenges and issues within cloud computing technology" and were able to uncover several difficulties, including those about trust, security, and privacy. Their study looked at the difficulties and problems that can come up when businesses and government institutions adopt cloud computing. This study was unable to pinpoint a technique for identifying data fraud. The intrusion detection system was put on the virtual switch in Yogesh's 2009 paper "Securing Cloud from DDOS Attacks Using Intrusion Detection System in a Virtual Machine," which records all network traffic—both inbound and outbound—into a database for auditing purposes. The guidelines are established using well-known intrusion techniques. What about lesser-known trespassers? There are still limitations in the work, and data fraud is still an issue. A study on a cloud computing-based network monitoring threat detection system for critical infrastructures was conducted by another researcher in 2015. Zhejiang.

This paper describes a cloud computing-based threat detection and monitoring system for critical infrastructure systems. The three main parts of the system are an operations center, cloud infrastructure, and monitoring agents. Although cloud computing machine learning algorithms have demonstrated potential in identifying data fraud, these useful techniques were not used in this study. Similar to this, earlier work by Vishal (2018) investigated machine learning-related rule-based and game-theoretic methods for detecting online credit card fraud. To mimic their competing interests, two players—an intrusion detection system and an intruder—play a multi-stage game in which each player aims to maximize their reward. James et al. (2012) used this paradigm to suggest a two-tier system for credit card fraud detection. The first tier would comprise rule-based components, while the second tier would involve game-theoretic components. Nevertheless, the rule-based approach has drawbacks because it depends too heavily on rules and could miss subtle intrusion patterns. The goal of the current work, however, is to go beyond this by identifying even minute and obscure patterns of incursion.

1.1. BACKGROUND TO THE STUDY

An improved virtualization approach is presented in this paper to improve data security in cloud computing architecture. Notwithstanding its advantages, cloud computing poses a serious hazard since it is more open to malevolent activity. With no administrative labor or provider participation, cloud computing is a flexible and on-demand network access architecture that provides a shared pool of scalable resources, such as networks, servers, storage, applications, and services. However, criminals have persisted in using cloud computing infrastructure as a means of data fraud, underscoring the necessity of stronger security protocols Rouse (2018).

Basic, unprocessed facts like test results, student names, or records of everyday organizational transactions are examples of data. Data examples include pay slip details, hours worked, client names, student registration numbers, date and quantity of items invoiced, and numerical values utilized in mathematical computations Ndukwe (2009). When data is processed and organized into a more

meaningful format, it becomes information and conveys important insights and intelligence. In its raw state, data is meaningless.

There hasn't been a consistent and all-encompassing definition of fraud, which goes beyond simple wrongdoing or criminal activity. Deliberate deception, such as purposeful misrepresentation, distortion of the truth, or suppression of material facts in order to attain an unfair advantage, acquire anything of value, or violate the rights of another party, is what defines fraud. Therefore, fraud may result in the cancellation of a transaction or give rise to a claim for damages. Intentionality is a fundamental component of deception and can be expressed directly, subtly, or through the activities themselves Brenner (2001). By merging the phrases "data" and "fraud," one might expand on the knowledge of these concepts and investigate the idea of "data fraud." Data fraud is the use of dishonest methods to hide or falsify information in order to obtain sensitive information that is not supposed to be made public, gain an unfair edge over others, or acquire valuable assets. According to this definition, data fraud occurs when someone gains illegal access to personal information, regardless of where it is stored. Data fraud is essentially the act of gaining access to sensitive information without the owner's knowledge or approval. In the banking sector, for example, client information is private and secure; any attempt to gain unauthorized access to, or hack into, these accounts is blatant data fraud.

Fraud detection is a component of data quality assurance assignments in businesses and sectors. A variety of tactics are included in data fraud detection with the goal of preventing sensitive data or money from being obtained without authorization through dishonest ways. This idea is used in a variety of industries, such as banking and insurance, where fraudulent activity might take the form of credit card theft, check forging, or other dishonest practices. Furthermore, as noted by Knatterud et al. (1998), the potential of data fraud affects businesses that depend on cloud computing in addition to the banking sector.

These days, cloud computing is a common conversation subject. Even though the word "cloud" refers only to the internet, cloud computing has become a distinct entity. Although there are many definitions of cloud computing that have been put forth in academic and industrial circles, the US National Institute of Standards and Technology (NIST) offers a thorough description that covers all of the crucial components James et al. (2023), James et al. (2022), James et al. (2011), James et al. (2012), James et al. (2012), James et al. (2010). NIST defines cloud computing as a model that facilitates shared, instantly provisioned, and released scalable computing resources—such as servers, networks, storage, and application services—with the least amount of provider interaction or administrative work Mell and Grance (2009). Cloud computing is essentially a paradigm that moves the emphasis from local computing to internet-based computing by giving users access to a shared pool of computer resources through the internet.

Cloud computing, to put it simply, is the provision of different computing services via the Internet, such as networking, database management, server storage, analytics, and intelligence. It offers economies of scale, flexible resources, and accelerated innovation Chukwu et al. (2023), James et al. (2022), James et al. (2023). Cloud computing removes the need for capital expenditures on hardware, software, and on-site data centers, including server maintenance, power, cooling, and expert management. This is a major departure from traditional IT resource management. Cloud computing does, however, come with some drawbacks, security being one of the main ones. One of the main concerns in cloud computing today is the possibility

of security breaches, data fraud, and other security risks Chen et al. (2014), James et al. (2016).

According to several studies Chukwu et al. (2023), James et al. (2024), James et al. (2024), Ekong et al. (2024), machine learning has repeatedly shown its worth in solving difficulties in the actual world. These days, it is used in many other areas, such as banking, insurance, e-commerce, and medicine. In the past, manual review activities were carried out; but, due to advances in mathematical modeling and system processing capacity, machine learning has become widely used in various industries James et al. (2024). Notably, data fraud detection has been successfully achieved using machine learning algorithms. Neural networks used in deep learning, which are inspired by how the human brain functions, use a number of computing layers to do this. Using cognitive computing, this method creates machines that use algorithms for self-learning that combine data mining, pattern recognition, and natural language processing. The machine is trained to improve its accuracy by repeatedly being exposed to a dataset over several layers James et al. (2024). Through the application of cognitive computing, the computer gains the ability to discern between authentic and fraudulent transactions with increased precision by learning from authorized behavior patterns. As an alternative, data fraud detection can also be achieved through the use of logistic regression, a supervised learning method. Using categorical decision-making, this approach produces results that categorize transactions as either fraudulent or not Obinachi (2012).

To protect their vital assets, businesses worldwide are investing heavily in IT cybersecurity skills. Enterprises rely on three crucial components for incident detection and response: technology, established processes, and human knowledge to safeguard their interests, be it their brand, intellectual property, customer information, or vital infrastructure Chinagolum et al. (2020). A key component of this protection is data security, which includes the steps required to prevent data from being corrupted or accessed without authorization at any point in its lifespan. To guarantee that data is secure across all platforms and apps, this entails putting data encryption, hashing, tokenization, and key management procedures into effect Intellipat (2013).

Users are able to carry out manual tasks like signing in and out and adjusting permission levels thanks to a hardware device. In order to prevent unauthorized users from accessing or changing these settings, this device uses biometric technology C et al. (2020), James et al. (2023). In order to stop malicious access, the device's controllers keep an eye on the user's present condition and interact with peripherals like hard drives. The controllers prevent illicit access to the data by stopping efforts by unauthorized users or programs to access the system. Notably, operating system defenses are vulnerable to viruses and hacking attempts, while hardware-based access control provides far higher security. Hard disk data corruption can result from malicious access, although software manipulation of user privilege levels is prevented by hardware-based protection Onu et al. (2015). Unless the hardware is corrupted or has a backdoor, hardware protection makes sure that hackers or harmful programs cannot access confidential data. According to Bart et al. (2017), this security measure stops manipulation of the operating system image and file system rights. Hardware-based security essentially adds another level of protection, making it very difficult for unauthorized parties to access or alter sensitive data Ituma et al. (2020), James et al. (2012), Ekong et al. (2024), James et al. (2023).

1.2. RESEARCH FOCUS

The focus of this research is on malicious operations carried out by means of fictitious Uniform Resource Locators (URLs), an increasing worry for cloud users. Alarmingly high rates of malicious activity have been observed online, frequently enabled by malware and other malicious software such as trojans, spyware, and bots. According to Onu et al. (2015), malicious transactions carried out using malware and phony URLs can be defined as the introduction of information-stealing software that targets gullible individuals, leading to compromised security and possible data breaches. The proliferation of malware families continues, with new variants emerging from original authors or others exploiting leaked source code. Building on previous research, such as Jarrod et al. (2018) fuzzy-based model addressing financial fraud in cloud computing, this study aims to develop a system for detecting fraud in cloud computing using advanced machine learning techniques like deep neural networks and logistic regression. Specifically, the study focuses on designing a fake URL detection system tailored for cloud computing infrastructure in Nigerian financial institutions, enhancing security measures to prevent financial fraud.

1.3. THEORETICAL FRAMEWORK

The theoretical underpinning of this study will be based on machine learning theory, also known as computational learning theory. Arthur Samuel, a pioneer in artificial intelligence and computer gaming, first proposed this idea in 1959. Machine learning theory aims to precisely identify the necessary skills, knowledge, and algorithmic principles needed for task learning success by trying to understand the fundamentals of learning as a computational process. The theory aims to clarify fundamental learning characteristics and increase our knowledge of how computers learn from data and adjust through feedback. These goals will help design better-automated learning techniques.

Machine Learning Theory draws elements from both the Theory of Computation and Statistics and involves tasks such as:

Among the objectives of machine learning theory are:

1) Creating mathematical models that capture the fundamentals of machine learning so that different learning issues' intrinsic complexity or simplicity can be examined.
2) Creating theoretical guarantees for algorithms, defining success criteria, necessary data, and computational resources, and creating algorithms that can be proven to satisfy performance requirements.
3) Using mathematical analysis to look into basic topics like:

- A rationale for using Occam's Razor as a theoretical compass
- Calculating the level of confidence in forecasts based on scant information
- Assessing the benefits of active learning over passive observation.

The Occam's razor establishes that straightforward explanations are preferable to complex ones. ¿Désde el punto de vista del rendimiento, es posible argumentar matemáticamente en favor de la navaja de Occam, aunque existen diversas razones por favor de explicaciones simples, tales como su facilidad de comprensión? En especial, ¿deberían medir la simplicidad y los programas informáticos que aprenden de la experience usar algún concepto de la navaja de Occam? Al diseñar reglas de predicción, hay una razón para buscar explicaciones simples, según uno de los primeros resultados de la Teoría del Aprendizaje Computacional. Specifically, for simplicity measures like the length of description in bits, the Vapnik-Chervonenkis dimensión that measures the effective number of parameters, and more recent measures studied in the research. These theoretical results' underlying intuition can be summed up as follows: There are many more complex explanations than there are simple ones. Consequently, it's unlikely to be a coincidence if a straightforward explanation adequately explains your facts. In contrast, even with a huge dataset, it is challenging to rule out all of the possible complex explanations due to their vast quantity. Even yet, certain superfluous intricate explanations can continue to exist and trick your system. Mathematical assurances can be derived from this understanding, guiding machine learning algorithms to favor straightforward explanations and refrain from overfitting to superfluous complexity.

One of the major contributions of Computational Learning Theory is the development of algorithms that can learn effectively even when inundated with a large volume of irrelevant data. To produce predictions like document relevance or picture classification, machine learning algorithms often describe data through features, such as words in a text or attributes of an image. The hardest part of the process, though, is deciding which properties are the most useful because the algorithm creator has no idea which traits will be most useful in advance. In a sea of distracting data, this emphasizes the need for algorithms that can adaptively learn and concentrate on pertinent information.

The intention is to make it possible for designers to supply the learning algorithm with a large number of features so that it can quickly discover and rank the most pertinent ones. The development of algorithms that, in many circumstances, have a convergence rate that is only slightly impacted by distracting elements represents a major advancement in Computational Learning Theory. Accordingly, a twofold increase in data amount will, at most, cause a slight drop in performance. Consequently, designers can supply copious amounts of data without fear of overloading the system. Furthermore, in order to achieve even more flexibility and efficiency, current research has concentrated on creating learning algorithms that can leverage kernel functions—which can be learned from data—to dynamically adjust their input representation.

Cryptography and machine learning theory are closely related fields. Enabling secure communication and stopping eavesdroppers from intercepting private information is the main goal of cryptography. Fascinatingly, machine learning can be thought of as creating eavesdropper algorithms—that is, trying to decipher encrypted data in order to retrieve information. Significant progress has resulted from this relationship, such as the transformation of difficult learning issues into suggested cryptosystems and the conversion of secure cryptosystems into difficulties that cannot be learned. Additionally, the interdisciplinary nature of these fields is demonstrated by the strong technical connections that exist between key techniques in Machine Learning and Cryptography. For example, Boosting, a technique for improving learning algorithms, and techniques for amplifying cryptosystems in Cryptography share similarities.

The development of guaranteed-performance auctions and pricing mechanisms is made possible by the theoretical underpinnings of machine learning, which have profound effects on economics. Furthermore, the paradigm for comprehending how people adjust to changing surroundings is provided by adaptive machine learning algorithms. The design of fast-learning algorithms provides important new understandings of how to approximate equilibrium in complex systems with multiple options. Specifically, machine learning algorithms that affect their surroundings and how other entities behave in them bring up significant economic issues. The close connection between Machine Learning and Economics is highlighted by this mutual link, which also promotes a vibrant intellectual exchange between the two fields.

The relationship between electronic commerce and machine learning theory has strengthened dramatically in the last several years as both domains work to create cutting-edge instruments for simulating and promoting online trade. As a foundational framework, machine learning theory addresses fundamental issues and advances software development practically by providing mathematical foundations for creating state-of-the-art algorithms. The discipline is going through an exciting time right now, with new applications being explored that raise interesting problems that need to be answered and links to other fields starting to emerge. Beyond anything we could have imagined, machine learning and its theoretical foundations have enormous promise that will likely lead to ground-breaking discoveries in the years to come.

1.4. DATA FRAUD

One of the prevalent sins of our era is data fraud. Although the scale of data fraud is unknown, it is rare for a day to go by in the media without news of yet another purported or actual fraud.

Compared to severe crimes like murder or rape, data fraud is typically seen as a victimless crime and hence escapes political and public scrutiny Hapman and Smith (2011). But many experts now acknowledge that fraud poses serious hazards to an organization's operations, finances, legal standing, and strategy. In fact, some frauds can have even more dire effects than certain street crimes Rebovich and Kane (2012). As such, organizations should give data fraud careful consideration. Furthermore, data fraud encompasses issues such as fraudulent contracts and tort law, extending beyond criminal situations Podgor (2019). The definition of "data fraud" is not clear and consistent; it includes both intentional deception of victims by communicating inaccurate statements as well as fraudulent actions. This demonstrates the intricacy and breadth of data fraud, underscoring the need for clarity and action.

varied organizations and nations have varied definitions of what constitutes fraud. According to Australia's government's Fraud Control Policy Commonwealth of Australia (2010), fraud is defined as misleading people or organizations with the goal of avoiding responsibilities or obtaining financial gain via dishonest behavior, false statements, or omissions. The Government of Western Australia, on the other hand, defines data fraud more broadly (1999:9), including non-financial gains such stealing corporate time or resources in addition to financial ones. This definition highlights the various viewpoints on fraud across various jurisdictions by encompassing any dishonest or misleading activities intended to get benefits from the government.

1.5. CLOUD COMPUTING

Formal definitions of cloud computing have been proposed by academic and industry circles; however, the National Institute of Standards and Technology (NIST) of the United States Mell and Grance (2009) provides a definition that is particularly noteworthy for encapsulating the key features that are frequently linked to cloud computing environments. NIST defines cloud computing as a methodology that makes it simple and flexible to access a shared pool of computing resources, such as servers, networks, storage, apps, and services, over the internet whenever and wherever needed. The efficiency and flexibility of cloud computing are demonstrated by the simplicity with which these resources may be swiftly deployed and released, requiring little in the way of administrative work or communication with service providers.

Figure 1

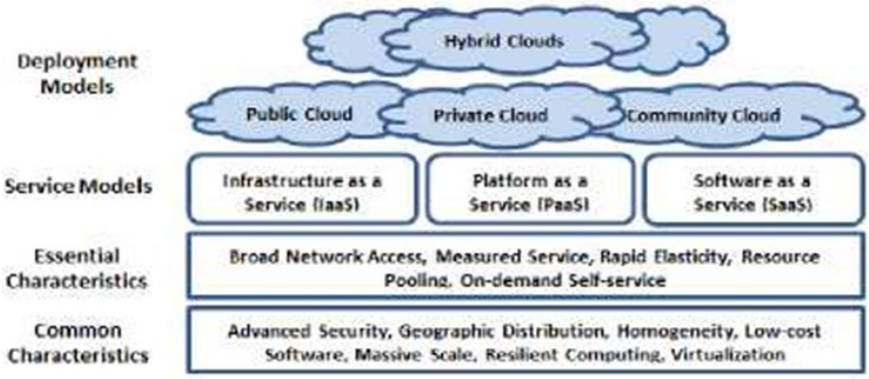

Figure 1 The NIST Cloud Definition Framework
Source Mell and Grance (2009)

1.6. OVERVIEW OF DEEP LEARNING

Artificial neural networks (ANNs) are the foundation for most modern deep learning models, with CNNs serving as the main architecture. Nevertheless, alternative models of deep learning, such deep generative models, might also include extra elements like latent variables or propositional formulae. Ciresan, Meier and Schmidhuber (2012) remark that these elements are frequently arranged layer-wise, much like the nodes in deep belief networks and deep Boltzmann machines. This demonstrates the variety of architectures utilized in contemporary deep learning.

In deep learning, the input data is transformed into progressively more abstract and sophisticated representations as each level builds on the one before it. When it comes to image recognition, for example, the first layer might take in a pixel matrix, which the second layer processes into edge arrangements, the third layer processes into facial features like noses and eyes, and the fourth layer finally produces face recognition. Interestingly, deep learning is capable of figuring out on its own what the best feature allocation is at every level. To attain the appropriate degree of abstraction, some manual fine-tuning is still required, such as modifying the quantity and size of layers, as noted by Krizhevsky et al. (2012).

The amount of layers that alter the data, suggesting a significant credit assignment path (CAP) depth, is what is meant to be understood as "deep" in deep learning. The series of transformations from input to output is represented by the CAP, which also describes possible causal links between them. The CAP depth in feedforward neural networks is equal to the sum of the hidden layers plus one; in recurrent neural networks, on the other hand, because of repeated signal propagation, it may be infinite. While experts cannot agree upon a precise cutoff, deep learning is generally understood to require a CAP depth larger than 2. Although any function may be universally approximated with a CAP depth of 2, further layers do not improve function approximation performance. Nevertheless, additional layers are helpful for feature learning because deeper models (CAP > 2) are superior at extracting features and learning them.

A greedy layer-by-layer method can be used to create deep learning models, adding layers one after the other in order to increase performance. Deep learning is made possible by this technique, which helps to decipher intricate abstractions and pinpoint the most important characteristics that lead to improved results Bengio, Y., Courville & Vincent (2013), Ekong et al. (2024), Okafor et al. (2023), James et al. (2017), James et al. (2024) . Deep learning can extract only the most informative features by detangling these abstractions, which improves accuracy and performance. Deep learning models can learn and get better over time because to this iterative, layer-by-layer structure, which makes use of the capability of many layers to produce better outcomes.

1.7. LOGISTIC REGRESSION

A statistical method for predicting the chance of an event or class occurring—such as pass/fail, win/lose, or healthy/ill—is called logistic regression, or logit model. This technique can be extended to categorize more than one event, such as recognizing different things in an image (such as a lion, dog, or cat). To ensure that the sum of the probabilities equals 1, each identified object is assigned a probability value ranging from 0 to 1. With increasingly sophisticated extensions available, logistic regression models binary outcomes using a logistic function. Essentially, it calculates the parameters of a logistic model in which the dependent variable (pass/fail, denoted by 0 and 1) has two possible values. It is possible to compute the log odds of the "1" result as a linear combination of binary or continuous independent variables. This makes it possible for Logistic Regression to forecast an event's probability using one or more predictor variables Because the log-odds scale is measured in units called logits, which are derived from the phrase "logistic unit," it is sometimes known by many names. Although the logistic model uses the logistic function, other models that are related to it, such the probit model, might use alternative sigmoid functions. The capacity of the logistic model, which has discrete parameters for each independent variable, to multiplicatively scale the probabilities of a given result at a constant rate when one of the independent variables is increased is what makes it unique. Tolles et al. (2016) point out that this property generalizes the odds ratio for binary dependent variables.

The dependant variable in binary logistic regression has two levels and is categorical. Multinomial logistic regression is used for outputs with more than two values, while ordinal logistic regression (e.g., proportional odds ordinal logistic model) is used if the categories are ordered. The logistic regression model does not do categorization by default; rather, it calculates the likelihood of an output given its inputs. Nevertheless, by establishing a cutoff value and classifying inputs with probabilities above or below the threshold, it can be utilized to build a classifier.

Logistic regression coefficients require iterative computation (see § Model fitting) and do not have a closed-form solution like linear least squares. According to Cramer (2002), Joseph Berkson was the one who first created and popularized the statistical model known as logistic regression.

1.8. VIRTUALIZATION/VIRTUALIZATION ARCHITECTURE

Making virtual copies of different entities, such as computer hardware, storage devices, software systems, and networks, is the process of virtualization Graziano (2013). Virtualization was first introduced in the 1960s with the intention of distributing mainframe resources logically among several applications. However, its use and breadth have since grown Graziano (2013). Numerous segregated virtual machines, operating systems, or numerous instances of a single OS can all operate concurrently on a single system thanks to virtualization. Even with its advantages, cloud virtualization security is still an issue. James et al. (2024) have conducted study on significant security issues around virtualization in cloud systems.

Virtualization, a key component of cloud computing, is defined by Sara Angeles (2014) as software that divides physical infrastructure into several specialized resources. According to Mike Adams, Director of Product Marketing at VMware, a top vendor of virtualization and cloud software, virtualization is essential to cloud infrastructure because it allows several operating systems and applications to run concurrently on a single server. In order to enable simultaneous use of physical resources across different systems, virtualization essentially entails building virtual replicas or models of those resources, which might include servers, storage devices, operating systems, or network resources.

Virtualization's main goal is to revolutionize existing computing techniques and make them more scalable, efficient, and economical in order to optimize workload management. Operating system virtualization, hardware-level virtualization, and server virtualization are just a few of the many uses for virtualization that enable adaptability and flexibility in the management of computer resources. Virtualization helps companies to become more agile, save money, and make better use of their resources by changing legacy computing.

By eliminating hardware dependence, virtualization technology is an affordable and energy-efficient option that is changing the computer environment Krishnatej et al. (2013). Virtualization primarily uses a virtual manager, also known as a virtual machine monitor (VMM), to separate computing environments from physical infrastructure. As stated by InfraNet vice president John Livesay, "Virtualization liberates servers, workstations, storage, and other systems from the constraints of physical hardware" Ekong et al. (2024). In order to achieve this, a hypervisor program is installed on the hardware layer. This software then hosts all other systems, hence abstracting the hardware layer itself Livesay (2013).

2. METHODOLOGY

The Structured System Analysis and Design Methodology is the methodology used for the proposed system design (SSADM). A systems approach to the analysis and design of information systems is the Structured Systems Analysis and Design Methodology James et al. (2022). The next stages systems provide the methodology by which SSADM was applied to the proposed system design, hence abstracting the actual hardware layer Livesay (2013):

1) **Stage One:** Examining Current strategies used to address the problem:

A thorough evaluation of the current system's effectiveness in providing users with trustworthy information was conducted James et al. (2024).

2) **Stage Two:** Analyzing Current System Performance and Finding Improvement Opportunities: This stage comprised evaluating the system's performance to find any irregularities, mistakes, or inefficiencies that need attention.

3) **Stage Three:** According to the findings of the comparison between the Existing and Proposed Systems, the most effective solution to address the stated issue would be adopted at this phase Ituma et al. (2020).

2.1. SYSTEM ARCHITECTURE

Figure 2

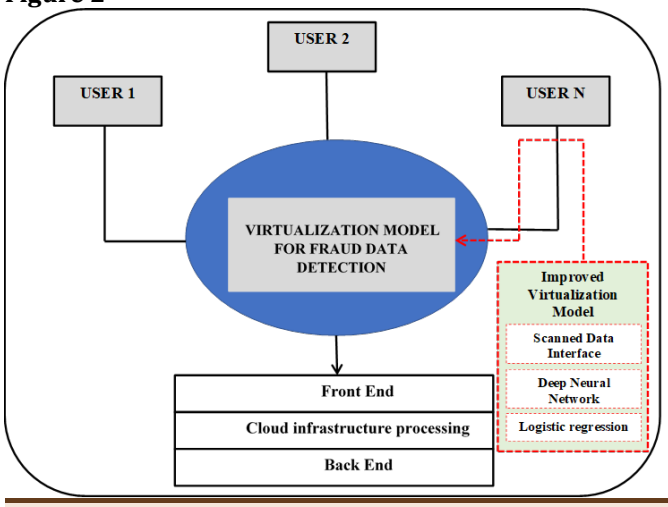

Figure 2 Architectural Design of the Proposed System

2.1.1. COMPONENTS OF THE PROPOSED SYSTEM

The following are components of the proposed system;

1) Scanned Data Interface

To provide a template of a set of attributes describing a specific entity, this component allows users to scan any data item belonging to a prospective criminal suspect. This template may then be used to design processes that read from or write to interfaces rather than directly to data sources or targets.

2) Deep Neural Network

This part demonstrates a portion of machine learning, which is basically a three-layer neural network. Although they fall well short of the human brain's capacity to "learn" from vast amounts of data, these neural networks make an effort to mimic its behavior.

3) Logistic Regression

This component uses a given dataset of independent factors to predict the likelihood of an event occurring, such as voting or not. Since probability determines the outcome, the dependent variable has a restricted range of 0 to 1.

2.1.2. ALGORITHM OF THE PROPOSED SYSTEM

Step 1: Start

Step 2: Initialize System
Step 3: Input set of classes in program being refactored
Step 4: Input set of datasets types (e.g. null-up method)
Step 5: Test System Cloud Server
Step 6: Output datasets and models accepted and initialized
Step 7: refactoring_count = 0
Step 8: repeat
Step 9: Activate Hybrid Model for Text Classification (HM)
Step 10: HM = X2 + PSO + k-NN
Step 11: Where
Step 12: X2 = Chi-Square
Step 13: PSO = Particle Swarm Optimization Technique
Step 14: k-NN = K-Nearest Neighbor
Step 15: Pre-process inputted text
Step 16: Classify and categorize processed texts
Step 17: Display categorized texts as result to user
Step 18 :classes = set of classes in program
Step 19: while !empty (classes) do
Step 20: class = classes.pick()
Step 21: scan inputted file for any trace of data fraud
Step 22: refactoring_count++
Step 23: Initialize Neural Network and Logistic Regression
Step 24: Text for Over-fitting (TFO) before pre-processing
Step 25: Increment TFO
Step 26: TFO = TFO + 1
Step 27: Pre-process inputted text
Step 28: Classify and categorize processed texts
Step 29: Display categorized texts as result to user
Step 30: Stop
Step 31: update system output
Step 32: else
Step 33: refactoring.undo()
Step 34: end.

2.1.3. MATHEMATICAL MODEL OF DEEP NEURAL NETWORK

Step 1: Start
Step 2: Initialize input and output layers of Neural Networks for training Models
Input = 0
Step 3: Increment Input
Input = Input + 1
Step 4: Access Learning Rate of Model Weight

Win+1 = win + n(yi – y1)xi

Step5: Test Learning Rate of Model Weight

Step 6: Transfer learned model to output layer of Neural Networks

Step 7: End

2.1.4. MATHEMATICAL MODEL OF LOGISTIC REGRESSION

Step 1: Start

Step 2: Initialize Regression Analysis

Step 3: Activate Analysis Modules

Step 4: plot the boundaries with original data

```
x_min, x_max = X[:, 0].min() - 1, X[:, 0].max() + 1
y_min, y_max = X[:, 1].min() - 1, X[:, 1].max() + 1
h = (x_max / x_min)/100
xx, yy = np.meshgrid(np.arange(x_min, x_max, h), np.arange(y_min, y_max, h))
X_plot = np.c_[xx.ravel(), yy.ravel()]
```

Step 5: Stop

3. EXPERIMENTAL RESULTS

The programming language Hypertext Pre-processor was used to implement the new system model. One high-level, general-purpose programming language is called Hypertext Pre-processor. The design philosophy of Hypertext Pre-processor prioritizes code readability through the substantial use of whitespace. Its object-oriented methodology and language elements are designed to assist programmers in writing logical, understandable code for both small and large-scale projects. Additionally, the Hypertext Pre-processor is compatible with Java and C++, among other computer languages. The Hypertext Pre-processor is garbage-collected and dynamically typed.

Guido van Rossum developed the Hypertext Pre-processor in the late 1980s to replace the ABC programming language. It was initially made available in 1991. The 2000 release of Hypertext Pre-processor 2.0 brought additional capabilities like list comprehensions and a garbage collection mechanism with reference counting. Version 2.7 of the program was terminated in 2020. 2008 saw the release of Hypertext Pre-processor 3.0, a significant language update that is not entirely backward compatible and requires modifications for many Hypertext Pre-processor 2 programs to function properly. Supported on most popular operating systems, Hypertext Pre-processor interpreters are also available for a few more (and previously supported by many more). The Hypertext Pre-processor Software Foundation, a non-profit, oversees and allocates funds for the development of the Hypertext Pre-processor and CHypertext Pre-processor. Netbeans IDE initialization is seen in Figure 4. The new system's welcome screen is displayed in Figure 5. The updated user registration page is displayed in Figure 6. The picture data upload page is displayed in Figure 7. Figure 8 displays the processed data result for handwriting-based malpractice identification together with the identified fraudster.

Figure 3

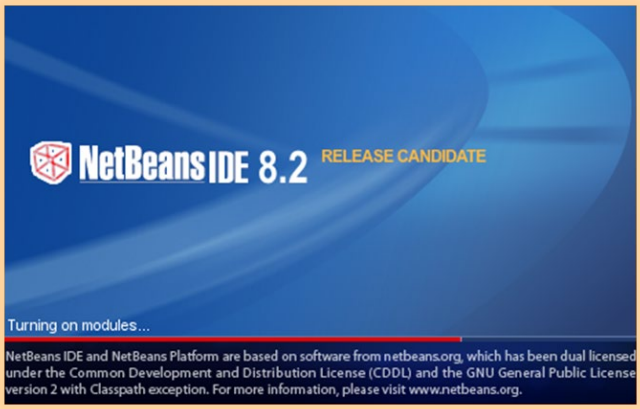

Figure 3 Initialization of Netbeans 8.2RC IDE

Figure 4

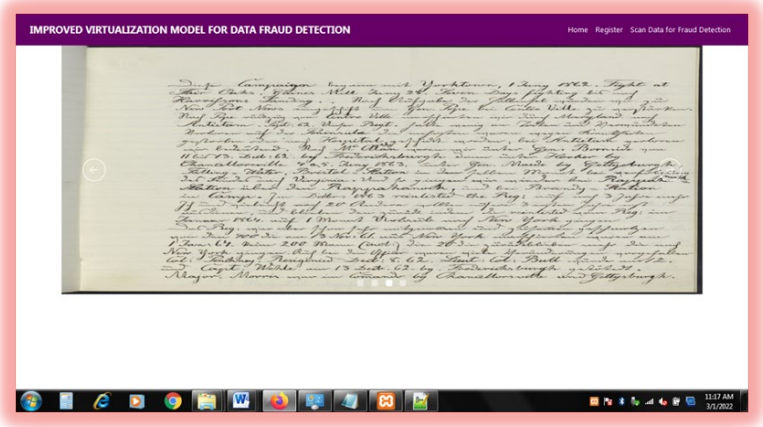

Figure 4 Welcome Page of the New System

Figure 5

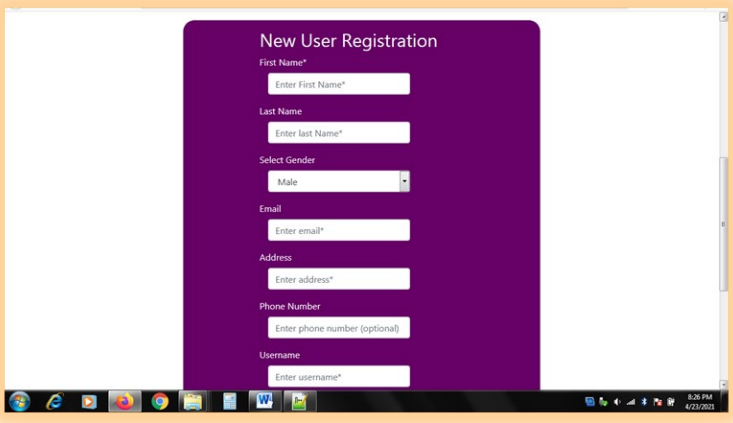

Figure 5 Registered Page

Figure 6

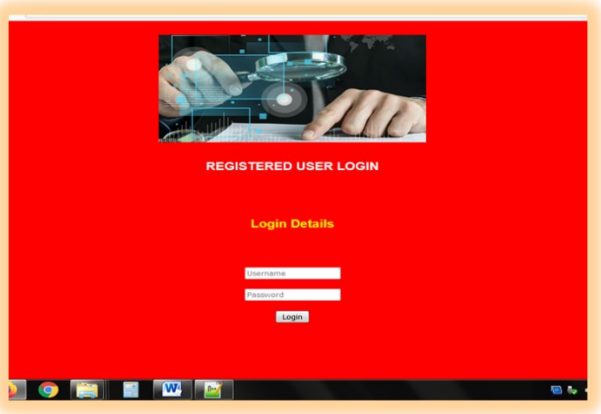

Figure 6 Registered User Login Page

Figure 7

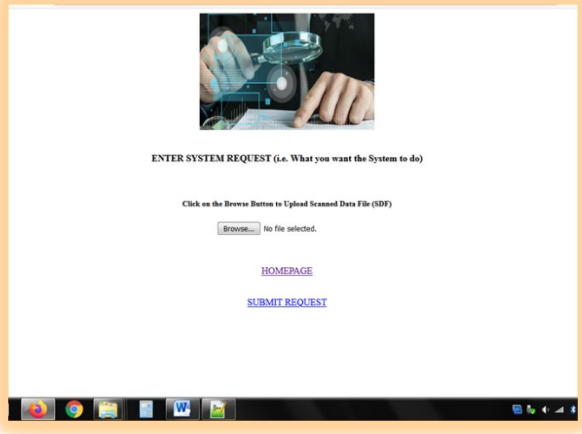

Figure 7 Scanned Data File Upload Page

Figure 8

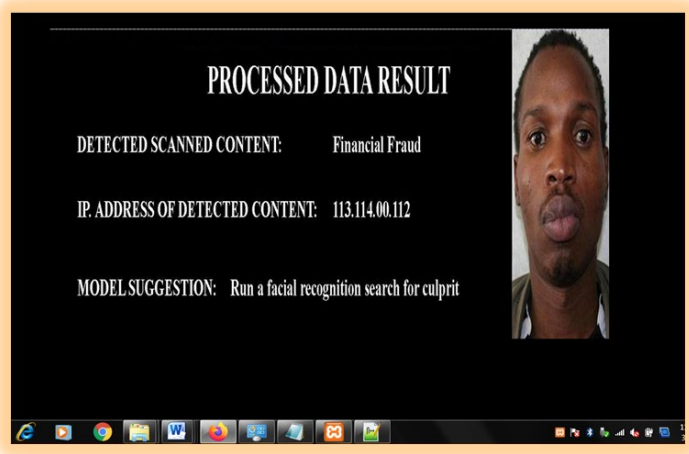

Figure 8 Processed Data Result

3.1. PERFORMANCE EVALUATION

Table 1

Table 1 Shows the Result Evaluation of the Existing System

S/N	Evaluated parameters	Values (V)
1	No. of adopted algorithm	1
2	No. of adopted technologies	1
3	No. of adopted design tools	1
4	No. of tested records	45

Accuracy level of the existing system is given by:

Summation of values x 1/100

= (1 + 1 + 1 + 45) x 1/100

= 48/100

= 0.48%

Figure 9

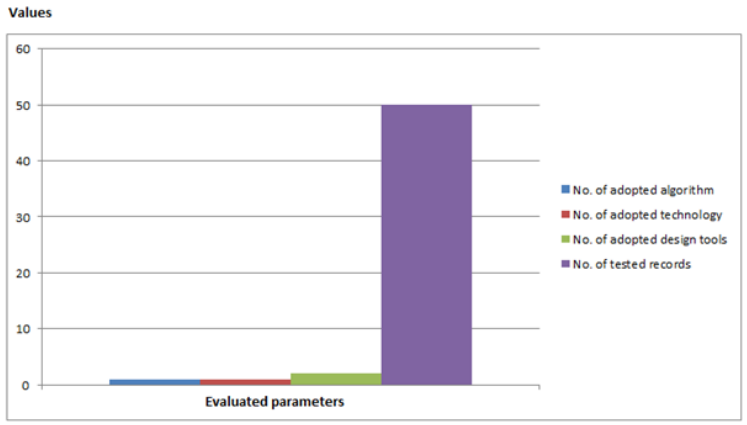

Figure 9 Performance Evaluation Chart of the Existing System

Table 2

Table 2 Result Evaluation of the New System

S/N	Evaluated parameters	Values (V)
1	No. of adopted algorithm	1
2	No. of adopted technologies	3
3	No. of adopted design tools	3
4	No. of tested records	100

Accuracy level of the new system is given by:

Summation of values x 1/100

= (1 + 3 + 3 + 100) x 1/100

= 107/100

= 1.07%

Figure 10

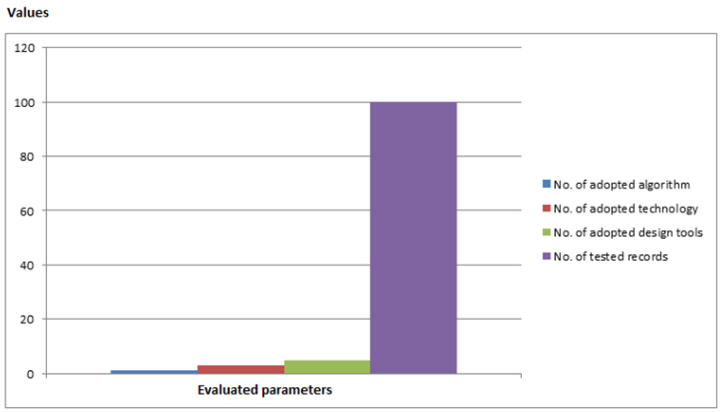

Figure 10 Performance Evaluation Chart of the New System

4. DISCUSSION OF RESULTS

The new system was put into place and checked for data fraud using Table 1. The outcomes of the new and current technologies for detecting data fraud are displayed in Table 2. The number of adopted technologies, the number of adopted design tools, the number of adopted algorithms, and the number of tested records were all used as parameters to analyze and assess the performance of both systems. According to the performance review, the new system performed better than the old one, achieving an accuracy rate of 1.07% as opposed to the old system's 0.48% accuracy rate.

The study's conclusions showed that fraud involving cell phones, insurance claims, tax return claims, credit card transactions, government procurement, and other areas pose serious issues for companies and governments, necessitating the use of specialized analysis techniques to find fraud involving them. These techniques are used in the fields of statistics, data mining, machine learning, and knowledge discovery in databases (KDD). They provide useful and effective solutions for various types of electronic fraud offenses. Since many internal control systems have significant flaws, combating fraud is typically the main goal of using data analytics approaches. For example, obtaining circumstantial evidence or accusations from whistleblowers is currently the predominant method used by many law enforcement organizations to identify businesses implicated in possible fraud cases. Consequently, a great deal of fraud cases go unreported and unpunished. Businesses entities and organizations rely on specialized data analytics techniques like data mining, data matching, and sounds like function, Regression analysis, Clustering analysis, and Gap in order to efficiently test, detect, validate, correct error, and monitor control systems against fraudulent activities. Artificial intelligence and statistical methods are the two main categories of fraud detection tools.

5. CONCLUSION

Umoh et al. (2012) provided an improved virtualization model for data fraud detection via fake uniform resource locators (URLs). This is due to the fact that malicious activity in cloud computing has remained a concerning problem. A shared pool of reconfigurable computing resources (networks, servers, storage, apps, and services) that can be quickly provided and released with little administration work

or service-provider interaction is made possible by the cloud computing concept. Nonetheless, the majority of guilty parties have persisted in using cloud computing infrastructure to spread data fraud. Data are fragments of fundamental raw facts, like a student's name or exam result (C. I. Ituma Iwok, Sunday Obot and James, G. G., 2020). It may take the shape of an organization's daily transaction records. For instance, the date, the amount of products on the invoice, the specifics of the pay stubs, the number of hours worked, the customer's name, the student registration number, and the numbers that are entered into a mathematical formula.

The statistics shown above lack meaning, but when they are transformed into a more interpretable and practical structure, they are referred to as information (data that transmit sufficient intelligence). The term "fraud" has never been defined in a consistent, comprehensive manner, and its definition is not restricted to issues of conduct, fraud, or criminal concepts. The definition of "fraud" is the deliberate use of deception, including purposeful fabrication of facts, concealing of material facts, or distorting the truth, to obtain an unfair advantage over another party in order to obtain valuables or deny another person their right. Fraud gives rise to the right to seek damages or to set aside a transaction at the option that it prejudiced.

A novel virtualization approach for cloud computing infrastructure was created in this work. The recently created model focused specifically on financial fraud and was intended to detect fake data in cloud computing infrastructure. This is due to the fact that financial frauds against property include the illegitimate transfer of property ownership for an individual's own use and advantage. Additionally, logistic regression and deep neural networks were used to further enhance the new system.

6. RECOMMENDATIONS

In order to ensure trust and confidentiality in financial management systems and cloud computing infrastructure, the study recommended the application of machine-learning oriented virtualization model for accurate and automated detection and prevention of fraudulent data.

7. CONTRIBUTIONS TO KNOWLEDGE

The study contributed the following to enhancing cloud computing infrastructures in Nigeria:

1) An improved machine-learning model for accurate and automated detection and prevention of fraudulent data
2) An enhanced situation awareness framework for accurately tracking malicious and inadvertent activities which could lead to data fraud in cloud computing infrastructures.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

All Glory returns to the Most High God for the gift of wisdom and life, which enabled us to succeed in this project. We use this medium to appreciate all who contributed positively to the actualization of this project.

REFERENCES

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798–1828.

Brenner, S.W. (2001). 'Is There Such a Thing as "Virtual Crime"?'in 4 California Criminal Law.

C, Ituma, G. G James, & F. U, Onu (2020). A Neuro-Fuzzy Based Document Tracking & Classification System. International Journal of Engineering Applied Sciences and Technology, 4(10), 414–423. https://doi.org/10.33564/IJEAST.2020.v04i10.075

Chukwu, E. G., James, G. G, Benson-Emenike, M. E., & Michael, N. A. (2023). Observed and Evaluated Service Quality on Patients Waiting Time of University of UYO Teaching Hospital using Queuing Models, 8(5), 2094–2098.

Ciresan, D., Meier, U., &Schmidhuber, J. (2012). Multi-Column Deep Neural Networks for Image Classification. 2012 IEEE Conference on Computer Vision and Pattern Recognition, 3642–3649.

Ekong, A. P., James, G. G., & Ohaeri, I. (2024). Oil and Gas Pipeline Leakage Detection using IoT and Deep Learning Algorithm. 6(1).

Ekong, A., James, G., Ekpe, G., Edet, A., & Dominic, E. (2024). A Model for the Classification of Bladder State Based on Bayesian Network, 5(2).

Ituma, C. I., Iwok, Sunday Obot & James, G. G. (2020). Implementation of an Optimized Packet Switching Parameters in Wireless Communication Networks. International Journal of Scientific & Engineering Research, 11(1).

Ituma, C., James, G. G., & Onu, F. U. (2020). A Neuro-Fuzzy Based Document Tracking & Classification System. International Journal of Engineering Applied Sciences and Technology, 4(10), 414–423. https://doi.org/10.33564/IJEAST.2020.v04i10.075

Ituma, C., James, G. G., & Onu, F. U. (2020). Implementation of Intelligent Document Retrieval Model Using Neuro-Fuzzy Technology. International Journal of Engineering Applied Sciences and Technology, 4(10), 65–74. https://doi.org/10.33564/IJEAST.2020.v04i10.013

James, G. G., Okafor P. C., Chukwu E. G., Michael N. A., Ebong O. A. (2024). Predictions of Criminal Tendency Through Facial Expression Using Convolutional Neural Network. Journal of Information Systems and Informatics, 6(1).

James, G. G., Asuquo, J. E., and Etim, E. O. (2023). Adaptive Predictive Model for Post Covid'19 Health-Care Assistive Medication Adherence System. In Contemporary Discourse on Nigeria's Economic Profile A FESTSCHRIFT in Honour of Prof. Nyaudoh Ukpabio Ndaeyo on his 62nd Birthday (Vol. 1, pp. 622–631). University of Uyo, Nigeria.

James, G. G., Chukwu, E. G. & Ekwe, P. O. (2023). Design of an Intelligent based System for the Diagnosis of Lung Cancer. International Journal of Innovative Science and Research Technology, 8(6), 791–796.

James, G. G., Ejaita, O. A., & Inam, I. A. (2016). Development of Water Billing System: A Case Study of Akwa Ibom State Water Company Limited, Eket Branch. The International Journal of Science & Technoledge, 4(7).

James, G. G., Ekanem, G. J., Okon, E. A. And Ben, O. M. (2012). The Design of e-Cash Transfer System for Modern Bank Using Generic Algorithm. International Journal of Science and Technology Research. International Journal of Science and Technology Research, 9(1).

James, G. G., Okpako, A. E., & Agwu, C. O. (2023). Tention to use IoT Technology on Agricultural Processes in Nigeria based on Modified UTAUT Model:

Perpectives of Nigerians' farmers. Scientia Africana, 21(3), 199–214. https://doi.org/10.4314/sa.v21i3.16

James, G. G., Okpako, A. E., Ituma, C., & Asuquo, J. E. (2022). Development of Hybrid Intelligent based Information Retreival Technique. International Journal of Computer Applications, 184(34), 1–13. https://doi.org/10.5120/ijca2022922401

James, G. G., U., U. A., Umoeka, Ini J., U., Edward N., & Umoh, A. A. (2010). Pattern Recognition System for the Diagnosis of Gonorrhea Disease. International Journal of Development in Medical Sciences, 3(1&2), 63–77.

James, G. G., Ufford, O. U., Ben, O. M., & Udoudo, J. J. (2011). Dynamic Path Planning Algorithm for Human Resource Planning. International Journal of Engineering and Technological Mathematics, 4(1 & 2), 44–53.

James, G. G., Umoh, U. A., Inyang, U. G. And Ben, O. M. (2012). File Allocation in a Distributed Processing Environment using Gabriel's Allocation Models. International Journal of Engineering and Technical Mathematics, 5(1&2).

James, G., Anietie, E., Abraham, E., Oduobuk, E., & Okafor, P. (2024). Analysis of Support Vector Machine and Random Forest Models for PrBroadband Networkedicting the Scalability of a. Journal of the Nigerian Society of Physical Sciences, 2093-2093.

James, G., Ekong, A., & Odikwa, H. (2024). Intelligent Model for the Early Detection of Breast Cancer Using Fine Needle Aspiration of Breast Mass. International Journal of Research and Innovation in Applied Science, IX(III), 348–359. https://doi.org/10.51584/IJRIAS.2024.90332

James, G., Umoren, I., Ekong, A., Inyang, S. & Aloysius, O. (2024). Analysis of Support Vector Machine and Random Forest Models for Classification of the Impact of Technostress in Covid and Post-Covid Era. Journal of the Nigerian Society of Physical Sciences, 2102-2102.

James, G.G., Archibong, M.N., Onuodu, F.E., Abraham, E.E., Okafor, P.C. (2024). Development of the Internet of Robotic Things for Smart and Sustainable Health Care. ShodhAI: Journal of Artificial Intelligence 1 (1), 9–27-9–27.

James, G.G., Okpako, A.E., & Ndunagu, J.N. (2017). Fuzzy Cluster Means Algorithm for the Diagnosis of Confusable Disease, 23(1).

James, Gregory G., & Ben, Oto-Abasi M. (2012). Fuzzy Diagnostic Support System for Asthma. International Journal of Engineering and Technological Mathematics, 5(1 & 2), 8–13.

James, V. O., G. G., Asuquo, J. E., & Etim, V. O. (2023). Combating Cybercrime in Nigeria: A Tool For Economic Development. In Contemporary Discourse on Nigeria's Economic Profile A FESTSCHRIFT in Honour of Prof. Nyaudoh U. Ndaeyo (Vol. 1, pp. 478–485). University of Uyo, Nigeria.

Jensen, M., Schwenk, J., Gruschka, N., & Iacono, L. L. (2010). Technical Security Issues in Cloud Computing, IEEE International Conference on Cloud Computing, 109.

Knatterud, G.L., Rockhold, F.W., & George, S.L. (1998). Guidelines for Quality Assurance Procedures for Multicenter Trials: A Position Paper. Controlled Clinical Trials, 19(5), 477–493.

Mell and Grance (2009). Effectively and Securely Using the Cloud Computing Paradigm (NIST Information Technology laboratory.

Ndukwe (2009). Data Structures and Algorithms.

Okafor, P. C., Ituma, C, & James, G. G. (2023). Implementation of a Radio Frequency Identification (RFID) Based Cashless Vending Machine. International Journal of Computer Applications Technology and Research, 12(8), 90–98. https://doi.org/10.7753/IJCATR1208.1013

Okafor, P. C., James, G. G., & Ituma, C. (2024). Design of an Intelligent Radio Frequency Identification (RFID) Based Cashless Vending Machine for Sales of Drinks. British Journal of Computer, Networking and Information Technology 7(3), 36-57.

Omar and Jefffery (2015). Challenges and Issues within Cloud Computing Technology.

Onu, F. U., Osisikankwu, P. U., Madubuike, C. E. & James, G. G. (2015). Impacts of Object Oriented Programming on Web Application Development. International Journal of Computer Applications Technology and Research, 4(9), 706–710.

Schmidhuber, J. (2015). Deep Learning in Neural Networks: An Overview". Neural Networks, 61, 85–117.

Tolles, J., Meurer, William J (2016). Logistic Regression Relating Patient Characteristics to Outcomes. JAMA, 316 (5), 533–4.

Umoh, U. A., Umoh, A. A., James, G. G., Oton, U. U. & Udoudo, J. J. (2012). Design of Pattern Recognition System for the Diagnosis of Gonorrhea Disease. International Journal of Scientific & Technology Research (IJSTR) 1 (5), 74-79.

Vishal (2018). Rule-Based and Game-Theoretic Approach to Online Credit Card Fraud Detection.

Yogesh (2009). Securing Cloud from DDOS Attacks Using Intrusion Detection System in Virtual Machine.

Zhijian (2015). A Cloud Computing Based Network Monitoring and Threat Detection System for Critical Infrastructures.