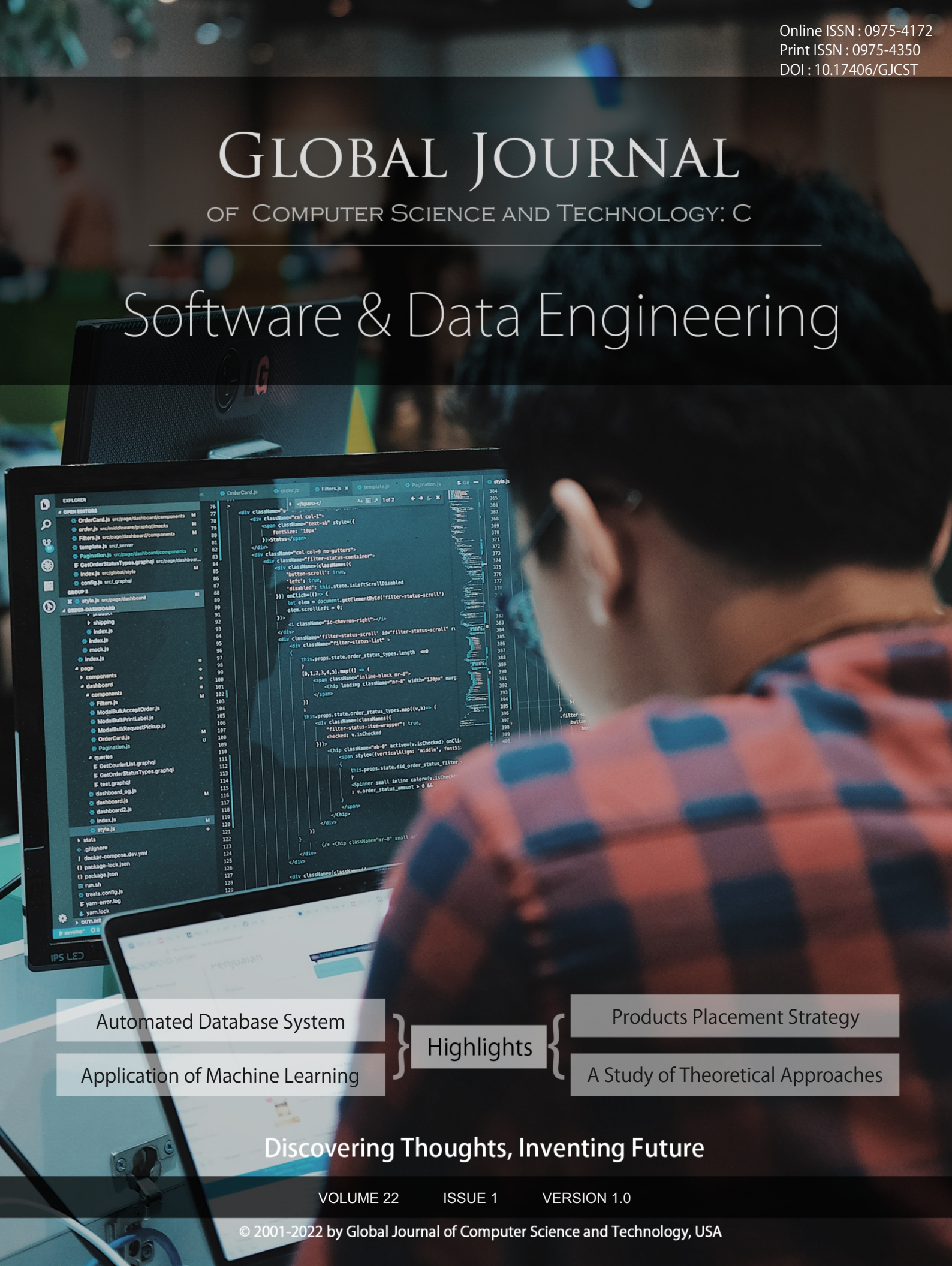
Online ISSN : 0975-4172
Print ISSN : 0975-4350
DOI : 10.17406/GJCST

Global Journal

of Computer Science and Technology: C

Software & Data Engineering



Automated Database System

Application of Machine Learning

Highlights

Products Placement Strategy

A Study of Theoretical Approaches

Discovering Thoughts, Inventing Future

VOLUME 22 ISSUE 1 VERSION 1.0

Global Journal of Computer Science and Technology: C
Software & Data Engineering

Global Journal of Computer Science and Technology: C
Software & Data Engineering

Volume 22 Issue 1 (Ver. 1.0)

Open Association of Research Society

Incorporation No.: 0423089
License No.: 42125/022010/1186
Registration No.: 430374
Import-Export Code: 1109007027
Employer Identification Number (EIN):
USA Tax ID: 98-0673427

Global Journals Inc.

*(A Delaware USA Incorporation with "Good Standing"; **Reg. Number: 0423089**)*

Sponsors: Open Association of Research Society
Open Scientific Standards

Publisher's Headquarters office

Global Journals® Headquarters
945th Concord Streets,
Framingham Massachusetts Pin: 01701,
United States of America
USA Toll Free: +001-888-839-7392
USA Toll Free Fax: +001-888-839-7392

Offset Typesetting

Global Journals Incorporated
2nd, Lansdowne, Lansdowne Rd., Croydon-Surrey,
Pin: CR9 2ER, United Kingdom

Packaging & Continental Dispatching

Global Journals Pvt Ltd
E-3130 Sudama Nagar, Near Gopur Square,
Indore, M.P., Pin:452009, India

Find a correspondence nodal officer near you

To find nodal officer of your country, please email us at *local@globaljournals.org*

eContacts

Press Inquiries: *press@globaljournals.org*
Investor Inquiries: *investors@globaljournals.org*
Technical Support: *technology@globaljournals.org*
Media & Releases: *media@globaljournals.org*

Pricing (Excluding Air Parcel Charges):

Yearly Subscription (Personal & Institutional)
250 USD (B/W) & 350 USD (Color)

Editorial Board

Global Journal of Computer Science and Technology

Dr. Stefano Berretti

Ph.D. in Computer Engineering and Telecommunications, University of Firenze Professor Department of Information Engineering, University of Firenze, Italy

Dr. Abdurrahman Arslanyilmaz

Computer Science & Information Systems Department Youngstown State University Ph.D., Texas A&M University University of Missouri, Columbia Gazi University, Turkey

Dr. Lamri Sayad

Ph.d in Computer science, University of BEJAIA, Algeria

Dr. Tauqeer Ahmad Usmani

Ph.D in Computer Science, Oman

Dr. Hazra Imran

Ph.D in Computer Science (Information Retrieval), Athabasca University, Canada

Dr. Magdy Shayboub Ali

Ph.D in Computer Sciences, MSc in Computer Sciences and Engineering, BSc in Electronic Engineering, Suez Canal University, Egypt

Dr. Nurul Akmar Binti Emran

Ph.D in Computer Science, MSc in Computer Science, Universiti Teknikal Malaysia Melaka, Malaysia

Dr. Asim Sinan Yuksel

Ph.D in Computer Engineering, M.Sc., B.Eng., Suleyman Demirel University, Turkey

Dr. Anis Bey

Dept. of Computer Science, Badji Mokhtar-Annaba University, Annaba, Algeria

Alessandra Lumini

Associate Researcher Department of Computer Science and Engineering University of Bologna Italy

Dr. Rajesh Kumar Rolen

Ph.D in Computer Science, MCA & BCA - IGNOU, MCTS & MCP - MIcrosoft, SCJP - Sun Microsystems, Singhania University, India

Dr. Rajneesh Kumar Gujral

Ph.D in Computer Science and Engineering, M.TECH in Information Technology, B. E. in Computer Science and Engineering, CCNA Certified Network Instructor, Diploma Course in Computer Servicing and Maintenance (DCS), Maharishi Markandeshwar University Mullana, India

Dr. Aziz M. Barbar

Ph.D. IEEE Senior Member Chairperson, Department of Computer Science AUST - American University of Science & Technology Alfred Naccash Avenue Ashrafieh, Lebanon

Dr. Federico Tramarin

Ph.D., Computer Engineering and Networks Group, Institute of Electronics, Italy Department of Information Engineering of the University of Padova, Italy

Dr. Chutisant Kerdvibulvech

Dept. of Inf. & Commun. Technol., Rangsit University Pathum Thani, Thailand Chulalongkorn University Ph.D. Thailand Keio University, Tokyo, Japan

Dr. Roheet Bhatnagar

Ph.D in Computer Science, B.Tech in Computer Science, M.Tech in Remote Sensing, Sikkim Manipal University, India

Contents of the Issue

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 22 Issue 1 Version 1.0 Year 2022
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Design of Machine Learning Framework for Products Placement Strategy in Grocery Store

By Olasehinde Olayemi
Federal Polytechnic

Abstract- The well-known and most used support-confidence framework for Association rule mining has some drawbacks when employ to generate strong rules, this weakness has led to its poor predictive performances. This framework predict customers buying behavior based on the assumption of the confidence value, which limits its competent at making good business decision. This work presents a better Association Rule Mining conceptualized framework for mining previous customers' transactions dataset of grocery store for the optimal prediction of products placement on the shelves, physical shelf arrangement and identification of products that needs promotion. Sampled transaction records were used to demonstrate the proposed framework. The proposed framework leverage on the ability of lift metric at improving the predictive performance of Association Rule Mining. The Lift discloses how much better an association rule is at predicting products to be placed together on the shelve rather than assuming. The proposed conceptualized framework will assist retailers and grocery stores owners to easily unlock the latent knowledge or patterns in their large day to day stored transaction dataset to make important business decision that will make them competitive and maximized their profit margin.

Keywords: *association rule mining, market basket analysis, frequent itemset, support, confidence, lift.*

GJCST-C Classification: *F.1.1*

Strictly as per the compliance and regulations of:

Design of Machine Learning Framework for Products Placement Strategy in Grocery Store

Olasehinde Olayemi

Abstract- The well-known and most used support-confidence framework for Association rule mining has some drawbacks when employ to generate strong rules, this weakness has led to its poor predictive performances. This framework predict customers buying behavior based on the assumption of the confidence value, which limits its competent at making good business decision. This work presents a better Association Rule Mining conceptualized framework for mining previous customers' transactions dataset of grocery store for the optimal prediction of products placement on the shelves, physical shelf arrangement and identification of products that needs promotion. Sampled transaction records were used to demonstrate the proposed framework. The proposed framework leverage on the ability of lift metric at improving the predictive performance of Association Rule Mining. The Lift discloses how much better an association rule is at predicting products to be placed together on the shelve rather than assuming. The proposed conceptualized framework will assist retailers and grocery stores owners to easily unlock the latent knowledge or patterns in their large day to day stored transaction dataset to make important business decision that will make them competitive and maximized their profit margin.

Keywords: association rule mining, market basket analysis, frequent itemset, support, confidence, lift.

I. Introduction

Grocery stores are stores that involves in the primary sales of general range of food products and daily needs. [1]Identified Cereals, Toothpaste, Beer, Butter, Cake Mix, Chips, Cookies, Facial Tissues, Laundry Detergent, Loaf Bread, Toilet paper and Coffee to be the twelve categories of products in a grocery store. These categories are selected for purchases based on certain parameter, which include price, always buy, satisfaction, recommendation, brand name, shelf space. Retailers regularly are faced with the challenges of allocating products to shelves due to shelf space being a scarce resource in retail stores and needs to increase the no of products to be included in the assortment [2].Product shelving allocates products in the shelves in an optimized way that will maximize sales and profit. According to [3].Products shelving tremendously affect consumer buying behaviors. Efficient allocation of product on shelves curtail the economic threats of unfilled product shelves, improves consumer satisfaction, healthier consumer relationship [4], and improve product sales [5].

Product shelving is a modern-day marketing strategy for products to get to end users without using overt traditional advertising. Product placement is becoming an increasingly important way for brands to reach their target audience in subtle ways. Businesses are exploiting product shelving to enhance brand awareness, increase sales and draw in customers without traditional marketing, Shelf shelving strategies are the various methods of arrangement of products on the shelves to induce impulse purchases and thereby increase sales and profit margin of the retailers. An ingenious display of product on shelve will increase customer's purchase decision, which habitually influenced in-store factors [6].The way customer's picks items to purchase on shelves are based on certain behavioral patterns and factors. Analytic of the past consumer purchasing behavior's record using Machine Learning (ML)algorithms will enhance the store's overall profitability[7].

ML is an aspect of artificial intelligence that learns with the aids of algorithm from data to obtain knowledge or pattern from it to make decision without human intervention. ML automate the process of data analytical for model building. ML's goal is to make an excellent guess useful to the predictive (classification) problem[8]. Supervised ML algorithms extract valuable knowledge from the mapping of supplied inputs and its desired output (class label) of the training dataset, then validates the testing dataset's obtained knowledge. Regression and classification are examples of supervised ML techniques. Unsupervised learning draws knowledge from a dataset consisting of input data without label responses. It partitions the dataset into clusters based on similarities that exists among the dataset. It validates by assigning a new test instance into the appropriate cluster; clustering analysis and association rule mining are examples of unsupervised learning methods.

Association Rule Mining (ARM) is rule-based ML algorithm for the discovering of interesting relationship among entities of a transactional dataset, ARM aim to identify patterns (combinations of events that occurred together) of entities in a transaction that frequently appear together among the whole transaction dataset. It generate rules that summarizes these patterns and use the generated rules to predicts presence of one or more

Author: Department of Computer Science, Federal Polytechnic, Ile Oluji, Nigeria. e-mail: olaolasehinde@fedpolel.edu.ng

products based on the occurrences of some products in a new transaction. Products that are capable to influence the presence of other products in transaction are predicted to be placed together on the shelved with the aim to create impulse purchases. Grocery store generates lot of data on daily basic from customer's transactions, this dataset contains hidden knowledge and patterns that can be used to make important business intelligent decision, unlocking this knowledge and patterns remains a mirage to several grocery stores, provision of a framework for discovering latent pattern or knowledge from transactional dataset will help grocery store's owners to make important business decision that will make them competitive and maximized their profit margin. This work presents an Association Rule Mining framework for mining previous transactions of consumers' buying patterns for the optimal prediction of products placement on the shelves, physical shelf arrangement and identification of products that needs promotion

II. Review of Related Literature

Several authors have applied Association Rule Mining algorithms to provide solution to different problems; Olasehinde et.al. (2018),applied ARM to mine customers buying behavior to improve customers relationship management, results from the research suggest products that should be shelved close to each other, products that needs promotion and products that promotion will not improve it sales [9].

[10] applied ARM to extract knowledge from the Market Basket Analysis (MBA) to predicts products that will be bought together and hence be placed close to each other on the shelve to induce and increase impulse buying. Serban et.al (11). proposed the application of relational ARM to predict the probability of certain diseases and predicts likely therapy [11]. Gupta et al. adopted ARM to determine the relationship among sequence of protein [12].The research in [13] applied ARM to the analysis of huge supermarket data exploiting the customer behavior to make market competitive decision. Liuet. al. (2007) applied ARM to generate important rules to extract strategic Business Intelligence (BI) from the mining of organization transaction. The experimenter results from the application of ARM to records of business transactions and customer's data analysis shows interesting patterns for customer's satisfaction and improvement of quality of service and profit [14]. [15] applied ARM to determine probability of purchases in online stores, result from this work shows that customers that have spent 10 to 25 minutes in an online book store and has opened thirty to seventy pages has a probability of 92% to confirm a purchase. The work in [16] applied ARM to the historic customer's transaction data from a grocery store to segment customers for targeted marketing.[17] conducted a research on Market Basket Analysis, Apriori Algorithm was used to discover frequent item sets among products stored in a large database, rules generated from this work were used to cluster customers based on their buying patterns and further subjected to selective marketing

III. Association Rule Mining

Association mining concerns the discernment of rules that cut across good percentage of dataset [18]. Given a set of transactions, T, the goal of ARM is to find all rules that predicts products to be placed closer to each other on the shelve and products that needs promotion.ARM involves two stages; in the first stage, frequent item set from the transaction dataset are generated that satisfied the predefined minimum support level. The second stage involves the generation of association rules that satisfied the minimum user's defined confidence rate among the frequent item-sets. Item-sets are one or more products in each record of the transaction dataset. A frequent itemset is a pattern that occurred frequently than a predefined threshold [19],frequent itemset is products combinations that satisfied the user's predefined minimum support. All subsets of a frequent itemsets are also frequent itemsets, while subsets of infrequent itemsets are infrequent item-sets. ARM is defined as follow:

$$let\ P = P_1, P_2 \ldots, P_n \quad (1)$$

Beset of n binary attributes called products.

$$let\ D = T_1, T_2 \ldots T_n \quad (2)$$

be set of all possible transactions D.
where each transaction T_i is a set of products such that $T_i \subseteq P$

Each transaction in D has a unique transaction ID and contains a subset of the products in P. A rule is defined as an implication of the form $X \Rightarrow Y$ interpreted as X implies Y.

$$\text{Where } X, Y \subseteq I \text{ and } X \cap Y = \emptyset \quad (3)$$

To select interesting rules for optimal products placement strategy, Support and confidence constrains are applied to all the generated rules from the transaction dataset.

The support often expressed as a percentage of total number of transactions in the dataset is basically the number of transactions that include all items in the antecedent and consequent parts of the rule [20]. The support of item-set containing products X and Y [21], written as $supp(X \Rightarrow Y)$ is the ratio of number of transactions that contains item-set X and Y to the total number of transaction in the dataset as shown in equation 4. Support of 0.75 for item-set X implies that 75% of the whole transactions in dataset contains item-set X. Itemsets that satisfied the minimum support threshold are considered to be frequent.

$$supp\,(X \Rightarrow Y) = \frac{No\ of\ trasactions\ \ that\ cointains\ \ itemset\ \ (X \cup Y)}{Total\ \ No\ of\ trasactions\ \ in\ the\ dataset} \quad (4)$$

The confidence of a frequent itemset (rule) is the percentage of all transactions that contain all products in both the consequent and the antecedent of the rule to the number of transactions that contain products in the antecedent [20]. The confidence of a frequent itemset (rule) X ⇒ Y is a conditional probability that Y will occurs whenever X occurred [22], it is the ratio of the support of X ∪ Y to support of X given in equation 5. The implication of the confidence of a rule X ⇒ Y to be 0.90 implies that, 90% of customers that buys X also buys Y.

$$conf\,(X \Rightarrow Y) = \frac{\text{supp}\ (X \cup Y)}{\text{supp}\ (X)} \quad (5)$$

IV. Proposed Framework for Association Rules Products Shelving Strategy

The proposed framework for the products placement (shelving) strategy is based on horizontal dataset layout, basically the framework consists of the following major components as shown in Figure 1

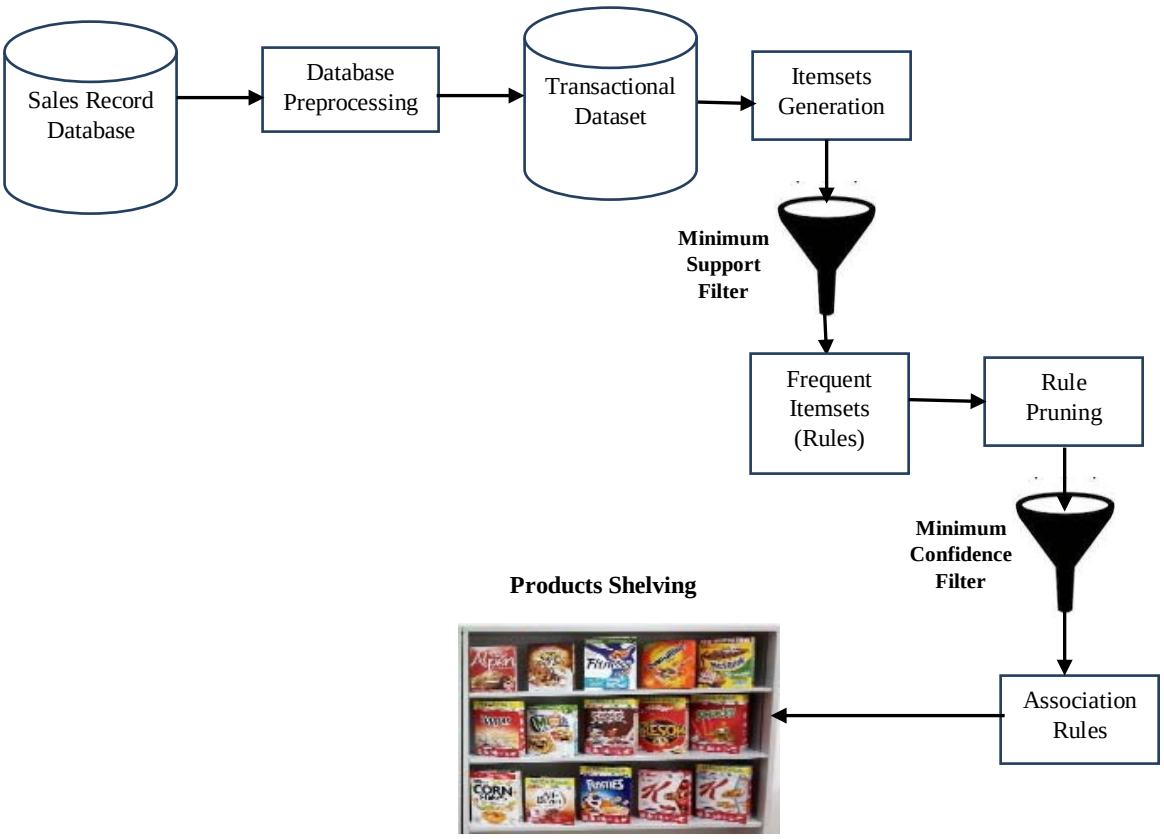


Figure 1: Proposed Framework for Association Rule Product Shelving Strategy.

V. Sales Record Dataset

Availability and quality of data is the bedrock of successful database modeling project, availability involves existence of relevant and suitable quantity of data while quality of data involves the fitness of data for the purpose to which it is intended for. High quality data are free of redundancy, defects and

possess desired features fit for the modeling purpose. Grocery stores generate and store large amount of data on daily basis, extraction of sensitive information implicitly contained in data will provide a lot direct benefits to the store. In order to obtain desire results of great benefits to the store from the store data, the data must be large enough to represent all the possible patterns of events in the store, a transaction data of 6to 24 months is recommended in order to provide good and effective decision that will benefit the grocery store [23]. Sales record contains many items such as transaction ID, customer's name, customer's ID, Product(s) bought, quantity bought, data and time of sales, unit price, product(s) code, Receipt number, product description.

VI. Sales Record Database Preprocessing

All the constitutes of the sales record are not relevant for the modeling of the products placement strategy, there is need to select the relevant constitute and represent them appropriately for ARM algorithms to be able to model them. Data preprocessing is critical to a successful data modeling process, presence of missing data, noise and irrelevant attributes will degrade the quality of the modeling results. For products placement strategy, transaction ID and the list of products in the transaction are the two relevant attributes, Table 2 shows a sampled preprocessed of a sales record containing five transactions depicted in Table 1, each row of the sales record represent a transaction, and each column represent a product (an item). Present of an item in the every transaction is represented with 1's while 0's implies absence of a product

Table 1: Horizontal Representation of Transaction Record

Transaction ID	Transaction Details
T1	{Bread, Egg}
T2	{Milk, Bread, Egg}
T3	{Bread, Butter, Egg}
T4	{Bread, Butter}
T5	{Milk, Bread, Butter, Egg}

Table 2: Sampled Preprocessed Five Transactions

Transaction ID	Milk	Bread	Butter	Egg
1	0	1	0	1
2	1	1	0	1
3	0	1	1	1
4	0	1	1	0
5	1	1	1	1

VII. Itemset Mining

Itemsets are one or more than one products bought together by customers and recorded in the transaction dataset, Itemset mining, the process of determining itemsets in the transaction dataset was first introduced by [24] in 1993, and it is nowadays called Frequent Itemset Mining (FIM). Frequent itemsets are pattern that transpired repeatedly than the predefined verge denoted as L_K, where K is the no of elements in the itemset. FIM mines group of items regularly bought together from the dataset. Any itemset X with its frequency of occurrences in the transaction dataset is more than the user predefined verge known as minimum support threshold (i.e. sup(X) minsup) is called frequent itemset. A transaction dataset with n distinct items (products), there will be 2^n-1 possible itemsets. The five transactions represented in Table 1 has four distinct items; {Milk, Bread, Butter,and Egg} with possible itemsets $=2^4$-1 = 15 itemsets. The fifteen itemsets from Table 1 with their support are ;{ Bread}: 1.0, it appeared in all the five transactions in the dataset, it support is 5/5 = 1.0,{Egg}: 0.8, {Milk}: 0.4, {Butter}: 0.6. {Milk and Bread}:0.4, {Milk and Butter}:0.2, {Milk and Egg}:0.4, {Butter and Egg}:0.4, {Bread and Egg}:0.8, {Bread and Butter}:0.6, {Milk, Bread and Butter}: 0.2, {Milk, Bread and Egg}:0.4, {Bread, Butter and Egg}:0.4, {Milk, Butter and Egg}:0.2, {Milk, Bread, Butter and Egg}: 0.2.

Given a user defined minimum support of 0.5, itemsets that has it support equals or greater than 0.5 will be filtered as the frequent itemsets, from Table 1, the itemsets that meet the minimum support threshold (0.5) set by the user are: {Bread}:1.0, {Egg}: 0.8, {Butter}: 0.6. {Bread and Egg}:0.8, {Bread and Butter}:0.6. Considering all conceivable itemsets, the mining of frequent itemsets is huge, naive, time consuming, expensive in terms of computer resources employed and not efficient particularly when the number of items under consideration are many. Efficient way to mine frequent itemsetsis via design of algorithms that circumvent exploring the search space of all

conceivable itemsets and analyses each itemset in the search space as efficient as possible.

The first algorithm used to mine frequent itemsets and association rules was Artificial Immune System (AIS) algorithm proposed by [25], improvement on AIS renamed as Apriori[24], other algorithms proposed for FIM include , Frequent pattern (FP) Growth algorithm [26], Equivalence Class Transformation (ECLATt) [27], Hyper-links Mine[28],Linear time Closed Mining (LCM) [29] and SET-oriented Mining (SETM) [30]. Apriori algorithm has been a predominantly implemented algorithm for mining frequent itemsets, but it is not efficient in its high overhead and consumption of the computer resources, an improvement to overcome it inefficiency was proposed in vertical representation of its dataset, Apriori TID [31] improve the efficiency of Apriori by avoiding multiple scan of the dataset during it valuation process. All these algorithms employ different strategies and data structures to discover frequent itemsets efficiently. According to [32], FIM algorithms differs in the following areas;

1. Mode of dataset representation, and how to compute minimum support
2. Search Space techniques, such as Depth-first or Breadth-first search and how it determine the next item sets to explore in the search universe

The two dataset representation formats used in FIM algorithms are Horizontal and vertical data format, horizontal format is presented in Table 1, it represents each transactions by its transaction ID, the vertical format is depleted in Table 3, it represents transactions with same items together, horizontal format can be easily converted to vertical format, the vertical format is more effective than horizontal format, it scan the dataset once to compute the support for each itemsets, it is faster than horizontal format in computing the support, but it also required more computer memory space to store the transactions ID. FIM algorithms employs Breadth-first and Depth-First search to mine frequent itemsets, Breadth-First search (BFS) explore all available nodes and select the shortest path between the starting node and other nodes, its memory consumption is higher than the Depth-First Search (DFS). in Breadth-first Search, the algorithm first evaluate single itemsets {Bread}, {Milk},{Butter}, {Egg}, then itemsets with two itemsets such as{{Milk and Bread}, {Milk and Butter}, {Milk and Egg}, {Butter and Egg}, {Bread and Egg}, {Bread and Butter}, follows by three elements, {Milk, Bread and Butter}, {Milk, Bread and Egg}, {Bread, Butter and Egg}, {Milk, Butter and Egg}and so on until all the number of items has been generated. On the other hand, depth-first search explore itemsets starting with single itemset and then recursively append items to the existing itemset to create another itemset, in the following order; {Milk}, {Milk, Bread}, {Milk, Bread, Egg}, {Milk, Bread, Butter}, {Milk, Bread, Butter, Egg}, {Milk, Butter}, {Milk, Butter, Egg}, {Milk, Egg}, {Bread}, {Bread, Egg}, {Bread, Butter}, {Bread, Butter, Egg}, {Butter},{Butter, Milk},{Butter, Egg}, {Egg}. Table 4 depletes the features of some FIM algorithms.

Table 3: Vertical Representation of Transections in Table 1

Itemsets	Transaction ID
Milk	T2,T5.
Bread	T1, T2, T3, T4, T5.
Butter	T3, T4, T5.
Egg	T1, T2, T3, T5.
Milk and Bread	T2, T5.
Milk and Butter	T5.
Milk and Egg	T2, T5.
Bread and Butter	T3, T4, T5.
Bread and Egg	T1, T2, T3, T5.
Butter and Egg	T3, T5.
Milk, Bread and Butter	T5.
Milk, Bread and Egg	T2, T5.
Bread, Butter and Egg	T3, T5.
Milk, Butter and Egg	T5.
Milk, Bread, Butter and Egg	T5.

Table 4: Features of Frequent Itemsets Mining Algorithms

Algorithms	Search Methods	Dataset Representation
AIS [25]	BFS (Candidate generation	Horizontal
Apriori [24]	BFS (Candidate generation	Horizontal
Apriori TID [31]	BFS (Candidate generation	Vertical (TID)
SETM [30]	BFS (Candidate generation	Horizontal (Sql)
ECLAT [27]	DFS (Candidate generation	Vertical (TID-List)
FP-GROWTH [26]	DFS (Pattern Growth	Horizontal (Prefix-tree)
H-MINE [28]	DFS (Pattern Growth	Horizontal (Hyperlink Structure)
LCM [29]	DFS (Pattern Growth	Horizontal (transaction merging)

VIII. Association Rules Generations

Association Rules (AR) generation in ARM involves two stages, in the first stage, frequent itemsets were generated, while the second stage has to do with creation of all possible rules from each of identified frequent itemsets that satisfied the minimum confidence threshold. AR are conditional probability that indicate the likelihood of a customers to buy a certain product provided if he or she had bought another product in the same purchase.AR is of the form {X ⇒ Y} has two part; the antecedent and the consequent, X is the antecedent (if) and Y (then is the consequent. Antecedent are items found within the data while consequent are items found in combination with the antecedent.AR are created from binary partitioning of each itemsets, the following binary rules will be generated from {Bread, Egg, Milk} frequent itemset;{Bread ⇒ Egg}, {Bread ⇒ Milk}, {Bread ⇒ Egg, Milk}, {Egg ⇒ Bread}, {Egg⇒ Milk}, {Egg ⇒ Bread, Milk}, {Milk ⇒ Egg}, {Milk ⇒ Bread}, {Milk ⇒ Bread, Egg}, {Bread, Egg ⇒ Milk}, {Bread, Milk ⇒ Egg}, {Egg, Milk ⇒ Bread},etc. The total number of possible binary rules R, generated from an itemset with d no of items is given in equation 6

$$R = 3^d - 2^{d+1} + 1 \quad (6)$$

AR generate a lot of rules, most these rules are not relevant and important, to prune the rules and obtain important rules, confidence of the each rule are computed using equation 5 based on the user defined minimum confidence threshold filter. AR that does not meet the minimum confidence threshold will be discarded. Note that the confidence of the rule {Bread ⇒ Egg} may not be same with the confidence of rule {Egg ⇒ Bread}.

From Table 1, the itemsets that meet the minimum support threshold (0.5) set by the user are: {Bread}:1.0, {Egg}: 0.8, {Butter}: 0.6. {Bread and Egg}:0.8, {Bread and Butter}:0.6. Given a user defined confidence of 60% (0.6).The number of AR with their support and confidence values are listed below;

Rule 1: {Bread ⇒ Egg}, support: 0.8, confidence: 0.8
Rule 2: {Egg ⇒ Bread}, support: 0.8, confidence: 1.0
Rule 3: {Bread ⇒ Butter}, support: 0.6, confidence: 0.6
Rule 4: {Butter ⇒ Bread}, support: 0.6, confidence: 1.0

The rules are interpreted as follows, in rule 1, 80% of customers that bought Bread also bought Eggs. In rule 2, all the customers that bought Egg also bought Bread. 60% of customers that bought bread in rule 3 also bought Butter, while all the customers that bought butter in rule 4, also bought Bread. Rules that satisfied the minimum support and confidence threshold are strong rules. Often, an AR with high confidence implies a strong rule, this can be misleading and deceptive when the antecedent and/or the consequent have a high support. Whenever the consequent of any AR is very frequent, its confidence will high. High confidence may be misleading at times, and does not always implies strong rules.

Lift ratio is a better metric to measure the strength of AR, it is the ratio of confidence of a rule to the expected confidence a rule. The expected confidence of a rule is probability of buying the consequent of the AR without any knowledge about antecedent. The lift ratio of AR $(X \Rightarrow Y)$ is given in equation 7.

$$Lift\,(X \Rightarrow Y) = \frac{conf\,(X \Rightarrow Y)}{conf\,(Y)} \quad (7)$$

A Lift value greater than one (1) implies positive association (correlation) between the antecedent and consequent of the AR, it implies that if a customer buy products in the antecedent there is great chances that products in the consequent will also be bought also. A lift value less than one (1) implies negative association between the antecedent and consequent of the AR, lift value of one (1) indicates no association between the antecedent and consequent of the AR. Applying Equation 7 to Table 1 gives the following lift values for the Rules 1, 2, 3and 4.

Rule 1: {Bread ⇒ Egg}, support: 0.8, confidence: 0.8, lift: 1.0
Rule 2: {Egg ⇒ Bread}, support: 0.8, confidence: 1.0, lift: 1.25
Rule 3: {Bread ⇒ Butter}, support: 0.6, confidence: 0.6, lift: 1.0

Rule 4: {Butter ⇒ Bread}, support: 0.6, confidence: 1.0, lift: 1.25

The values of the lift of the rules above shows that there is no association between the rules {Bread ⇒ Egg} and {Bread ⇒ Butter}, while there is a positive correlation between the antecedent and the consequent of rules {Egg ⇒ Bread} and {Butter ⇒ Bread}, with 25% more chances of buying the antecedent and the consequent products together. Considering the confidence of an AR alone will limit the competency of making good business decision, The Lift discloses how much better an AR is at predicting products to be placed together on the shelve rather than assuming, confidence assumes, *Lift* is a measure that assist store managers to determine the products to be placed together on shelve.

IX. Conclusion

The vast amount of transaction dataset being generated by grocery store remain useless unless the latent knowledge and patterns hidden in it is unlock and discovered. Discovered latent pattern or knowledge from transactional dataset will help grocery store's owners to make important business decision that will make them competitive and maximized their profit margin. The well-known and most used support-confidence framework for Association Rule Mining has some drawbacks when employ to generate strong rules, this weakness has led to it poor predictive performances. This framework predict customers buying behavior based on the assumption of the confidence value, which limits it competent at making good business decision. This work presents a better Association Rule Mining framework for mining data of previous transactions of consumers' buying patterns for the optimal prediction of products placement on the shelves, physical shelf arrangement and identification of products that needs promotion. The proposed framework leverage on the ability of lift metric at improving the predictive performance of association rule mining. The Lift discloses how much better an AR is at predicting products to be placed together on the shelve rather than assuming, confidence assumes. *Lift* is a measure that assist store managers to determine the products to be placed together on shelve. The proposed framework will assist retailers and grocery store's owners on products placement on the shelves, physical shelf arrangement and identification of products that needs promotion

References Références Referencias

1. Hoyer, W.D & Walgren, C.J.C. (1988). Consumer Decision Making Across Product Categories: The Influence of Task Environment. John Wiley & Sons Inc., 5(1) 45-69.
2. Hübner, A. H. and Kuhn, H. (2012). Retail category management: State-of-the-art review of quantitative research and software applications in assortment and shelf space management. Omega, 40(2):199–209.
3. Dreze, X., Hoch, S. J., & Purk, M. E. (1994). Shelf management and space elasticity. Journal of Retailing, 70, 301–326.
4. Fancher, L. A. (1991). Computerized space management: A strategic weapon. Discount Merchandiser, 31(3), 64-65.
5. Hwang, H., Choi, B., & Lee, M. J. (2005). A model for shelf space allocation and inventory control considering location and inventory level effects on demand. International Journal of Production Economics, 97(2), 185-195.
6. Hübner, A. H. and Schaal, K. (2017). An integrated assortment and shelf-space optimization model with demand substitution and space-elasticity effects. European Journal of Operational Research, 261(1):302–316.
7. Luís F. M. S. (2018).Optimizing Shelf Space Allocation under Merchandising Rules. A Master's Dissertation submitted University of Porto.
8. Olasehinde O. O. (2020).A Stacked Ensemble Intrusion Detection Approach for Security of Information System, International Journal for Information Security Research (IJISR),10(1).
9. Olasehinde, O.O., Williams, K.O. and Ajinaja, M.O. (2018): Application of Association Rule Learning in Customer Relationship Management. Proceedings of the 14th iSTEAMS International Multidisciplinary Conference, AlHikmah University, Ilorin, Nigeria, 14: 29-36.
10. Sherdiwala B., Khanna O. (2015). Association Rule Mining: An Overview", International Multidisciplinary Research Journal (RHIMRJ).
11. Serban G., Czibulal. G. and Campan A. (2016). A Programming Interface For Medical diagnosis Prediction", Studia Universitatis, "Babes-Bolyai", Informatica, LI(1), pages 21-30, 2006.
12. Gupta N., Mangal N, Tiwari K. and Mitra P. (200). Mining Quantitative Association Rules in Protein Sequences", In Proceedings of Australasian Conference on Knowledge Discovery and Data Mining –AUSDM, 273-281, 2000.
13. Raorane A.A., Kulkarni R.V. and Jitkar B.D. (2012). Association Rule – Extracting Knowledge Using Market Basket Analysis, Research Journal of Recent Sciences, 1(2): 19-27.
14. Liu, H., Su, B., & Zhang, B. (2007). The Application of Association Rules in Retail Marketing Mix. *2007 IEEE International Conference on Automation and Logistics*, 2514-2517.
15. Suchacka, G., Chodak, G. (2017) Using association rules to assess purchase probability in online

stores. *InfSyst E-Bus Manage* 15, 751–780. https://doi.org/10.1007/s10257-016-0329-4
16. March N., Reutterer T. (2008) Building an Association Rules Framework for Target Marketing. In: Preisach C., Burkhardt H., Schmidt-Thieme L., Decker R. (eds) Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-78246-9_52
17. Phani, P., and Murlidher, M. (2013). A Study on Market Basket Analysis Using a Data Mining Algorithm. International Journal of Emerging Technology and Advanced Engineering, 3(6): 361-363. Retrieved from www.ijetea.com
18. Zaki M.J. (2000). Scalable Algorithms for Association Mining. IEEE Transaction on Knowledge and Data Engineering, 12(3): 372-390.
19. P. Yazgan, A.O. Kusakci. (2016). A Literature Survey on Association Rule Mining Algorithms, Southeast Europe Journal of Soft Computing 5(1), Doi; 10.21533/scjour(nal.v5i1.102
20. Vidhate, D. (2014). To improve Association Rule Mining using New Technique: Multilevel Relationship Algorithm towards Cooperative Learning 241–246.
21. Cuzzocrea A., Leung C.K., MacKinnon R.K. (2015). Approximation to expected support of frequent itemsets in mining probabilistic sets of uncertain data, Procedia Comput. Sci. 613–622.
22. Naresh P. and Suguna R. (2019). "Association Rule Mining Algorithms on Large and Small Datasets: A Comparative Study," *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, 2019, pp. 587-592, doi: 10.1109/ICCS45141.2019.9065836.
23. Al-hagery, M.A. (2015). Knowledge Discovery in the Data Sets of Hepatitis Disease for Diagnosis and Prediction to Support and Serve Community. Int. J. Comput. Electron. Res. 4, 118–125.
24. Agrawal, R.,Srikant, R. (1994) Fast algorithms for mining association rules. In: Proc. 20th int. conf. very large data bases, VLDB 1994, Santiago de Chile, Chile, (487–499).
25. Aggarwal, C. C. Data mining: the textbook. Heidelberg: Springer; 2015.
26. Han, J, Pei, J, Ying, Y, Mao, R. Mining frequent patterns without candidate generation: a frequent-pattern tree approach. Data Min. Knowl. Discov., 2004, 8(1):53–87.
27. Zaki, M. J. (2000). Scalable algorithms for association mining. IEEE Transactions on Knowledge and Data Engineering. 12 (3), 371- 390. doi:10.1109/69.846291
28. Pei J., Han, J., Lu H., Nishio S, Tang S, Yang D. (2001) H-mine: Hyper-structure mining of frequent patterns in large databases. In: Proc. 2001 IEEE Intern. Conf. Data Mining, 441–448.
29. T. Uno, M. Kiyomi, and H. Arimura, (2004) "Lcm ver. 2: Efficient mining algorithms for frequent/closed/maximal itemsets," in Fimi, vol. 126, Proc. ICDM'04 Workshop on Frequent Itemset Mining Implementations, CEUR, 2004.
30. Houtsma M. and Swami A. (1995) Set-oriented mining of association rules". Elsevier journal Data & Knowledge Engineering 245-262 Technical Report RJ 9567
31. Gosain, A., & Bhugra, M. (2013). A comprehensive survey of association rules on quantitative data in data mining. *2013 IEEE Conference on Information and Communication Technologies*, 1003-1008.DOI: https://doi. org/10.1109/cict.2013.6558244.
32. Fournier-Viger, P., Lin, C., Vo, B., Truong, T.C., Zhang, J., & Le, H. (2017). A survey of itemset mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 7(4) https://doi.org/10.1002/widm.1207

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 22 Issue 1 Version 1.0 Year 2022
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Performance Analysis of D-Mosk Modulation in Mobile Diffusive-Drift Molecular Communication Relay System

By Jiaxing Wang , Dengchao Feng & Wanjun Li
North China Institute of Aerospace

Abstract- Molecular communication (MC) is a new wireless communication technology, which uses molecules as information carriers. Diffusion-based MC is one of the most common MC methods. With the increase of diffusion distance, the molecular signal attenuation is serious, so the traditional communication technology of relay is introduced into the MC system. In this work, a mobile diffusive-drift MC relay model is investigated, in which the depleted molecule shift keying (D-MoSK) modulation is used. The closed-form expression of symbole error rate (SER) and the channel capacity are derived, meanwhile the impacts of several crucial parameters on the performance are discussed comprehensively.

Index Terms: molecular communication, relay, diffusivedrift, symbol error rate, depleted molecule shift keying.

GJCST-C Classification: D.2.2

Strictly as per the compliance and regulations of:

Performance Analysis of D-Mosk Modulation in Mobile Diffusive-Drift Molecular Communication Relay System

Jiaxing Wang [α], Dengchao Feng [σ] & Wanjun Li [ρ]

Abstract- **Molecular communication (MC) is a new wireless communication technology, which uses molecules as information carriers. Diffusion-based MC is one of the most common MC methods. With the increase of diffusion distance, the molecular signal attenuation is serious, so the traditional communication technology of relay is introduced into the MC system. In this work, a mobile diffusive-drift MC relay model is investigated, in which the depleted molecule shift keying (D-MoSK) modulation is used. The closed-form expression of symbole error rate (SER) and the channel capacity are derived, meanwhile the impacts of several crucial parameters on the performance are discussed comprehensively.**

Index Terms: molecular communication, relay, diffusivedrift, symbol error rate, depleted molecule shift keying.

I. Introduction

MOLECULAR communication (MC) is a new type of communication using molecules as information carriers, which can be used to its advantage in scenarios where conventional wireless communication is not suitable, such as in confined pipes, seawater or body areas. In MC systems, the molecules usually undergo Brownian motion, and as the diffusion distance increases, the molecular signal attenuates severely, making the diffusion transmission distance very limited. At the same time, the large transmission delay of freely diffusing molecules causes severe inter-symbol interference (ISI), which is an important factor affecting system performance. In order to extend the transmission distance and improve the system performance, a suitable channel transmission model is needed to study the mechanism. Because their small size and the fact that they do not easily communicate using electrons or electromagnetic waves, MC offers a new mechanism for nanometers communication by transporting molecules to represent information [1], [2]. These nanomachines have computing, storage and drive functions [3]. Due to their own limitations, they cannot perform the corresponding tasks, so they are interconnected to overcome their limitations and form a nanonetwork with certain functions that work together in a collaborative area to accomplish specific tasks[4]–[6].

The idea that nanomachines achieve information exchange through the emission, transmission and reception of information molecules comes from the exchange of information between cells in nature [7]. Diffusion-based nanotechnology for MC has a wide range of promising applications, mainly in biomedicine. A biological system (e.g., nanomachines), each performing simple and specific operations such as the uptaking, processing and releasing of molecules, as well as cellular interactions to perform various functions of the body(e.g. , cell metabolism, molecular replication, etc.) [8].

The current demodulation algorithm applied to the received message in MC is mainly based on the detection of the number of molecules. In a fixed time slot, the transmitter sends a certain number of molecules to represent message “1”, while no molecules are sent to represnt message “0”. The receiver demodulates the message to bit “1” when and only when the number of molecules received exceeds the set threshold, otherwise the message is demodulated to bit “0”. The whole communication process is based on the time slot for message transmission. This demodulation algorithm based on the number of molecules is very simple to implement. However, due to the random diffusion of molecules and ISI caused by the accumulated molecules in the medium, the recognition rate of the signal during demodulation at the receiver side is reduced, resulting in a higher BER and the reliability of the communication is greatly affected.

In order to solve the above problems, research on diffusion based MC systems has attracted a lot of academic attention. By adding relay nodes, the transmission distance of diffusing molecules can be enhanced and the system performance can be improved. By introducing relay nodes, the transmission distance of diffusion MC can be extended. Meanwhile, the system performance can be improved [9], [10]. In conventional wireless communication systems, decode-and-forward (DF) is used to enhance system performance. A point-to-point relay model based on MC can significantly improve the transmission reliability, and related on MC relay has been studied in several

Author α σ ρ: Langfang Aerospace Testing Technology and Instrument Research and Develpment Center, North China Institute of Aerospace Engineering, Langfang, Hebei. e-mails: jx19882008@163.com, tyfdc001@163.com, jermeslee@163.com

literature. In [11], diffusion-based sensory relay transmission strategies for MC systems were investigated, and although bacteria were used as information carriers, the essence of the transmission was still diffusion. The DF relay transmission model for diffusion MC systems was proposed in [12], where the channel model considers the effects of noise and channel memory, while exploring the performance of the BER in the sysytem subject to channel fading versus the optimal relay location. Literature [12] and [13] explored the BER performance of the system in DF and amplify-and-forward (AF) transmission modes, respectively. A diffusion-based model of a reversible binding system for DF relay ligands and receptor for MC is presented in [14], where the time-varying spatial distribution of information is characterised based on the reversible biding and separation of ligands and receptors on the receiver surface. The literature [15] describes a diffusion based molecular theory system model based on the influence of molecular delivery sequences and obtains the information transmission rate of the relay channel under the influence of the sequence-based approach. In [16] the authors built an AF relay system and analysed the performance of the system under different detection schemes. In [17] the authors developed a DF relay model for MC systems based on drift diffusion, applied to the human vascular scenario, approximated the number of molecules received to a normal distribution and solved for a closed-form expression for the BER of the system. These authors after further research, proposed a DF relay system model based on energy detection in [18]. In [19], a cooprative diffusion-based MC network model is considered, which consisting of single source, single DF relay, and single destination.

However, mobile MC is needed in many envisioned applications. A static transmitter and a mobile bacterium-based receiver are considered in [20], meanwhile an adaptive ISI mitigation method and two adaptive detection schemes are proposed for the mobile scenario. In [21], authors consider a mobile MC system where the fluid medium has a fully developed homogeneous turbulence, and both the transmit and the receive nano-devices are mobile. The mobile multiuser diffusive MC system with drift which is composed of multiple mobile transmitter nanomachines and one mobile receiver nanomachine is built in [22], in which both the ISI and multiuser interference unavoidably exit in the same fluid medium. The closed-form expressions for the probabilities of detection and false alarm are derived at the cooperative and destination nanomachines considereing the multiple-source interference and the ISI are obtained in [23]. The authors propose an adaptive etection scheme for mobile MC with a low computational complexity by utilizing the local convex propoty of the channel impulse response in [24], in which the results show that the proposed scheme achieves good detection accuracy with low computational complexity. The mobile MC system is built in[25], in which tranmistter and receiver move randomly in a free diffusion manner. The closed-form expressions of the mean and variance of the received signal are derived by considering two kinds of randomness.

In this work, a DF relay for mobile MC system is presented to improve the system performance in the long-distance scene. DF relay can reduce the accumulation noise. Meanwhile, D-MoSK modulation is used in the system. The novelties of this work are summarized as follows:

1. A DF MC relay scheme for long-distance communication is concerned, in which the source and destination are mobile, and the D-MoSK modulation is used to deduce ISI and decoding complexity.
2. The corresponding SER and capacity are characterized, and the impacts of the key factors on the performance are evaluated, , such as the velcocity of fluid, the coefficient of molecules and so on. The obtained results are expected to provide guidance significance for the design of a practical mobile diffusive-drift MC relay system.

The remainder of this paper is organized as follows. The mobile diffusive-drift MC relay system model, including the mobile S and the mobile D, is introduced in Section II. In Section III, we will give the mathematical derivations with respect to the detection scheme. In Section IV, numerical results and performance discussions are presented. Section V concludes the paper.

II. System Model

In this work, a mobile MC relay system model is built, which considers a source node, a relay node and a destination node. They are in mobility, and the fluid medium has a certain velocity. The system model is shown in Fig. 1, where S, R and D represent the information source node, the relay node and the destination node, respectively. In this system model, the depleted molecules shift keying (D-MoSK) modulation method is utilized. It uses two different types of molecules to represent Quaternary information, that is, the emission of molecules “a” represents information “10”, the emission of molecules “b” represnts information “01”, molecules “a” and “b” simultaneously emission represents information “11”, and neither molecules “a” nor molecules “b” emission represents information “00”. When the information molecules drift to the R, the R detects the signal. The R decodes the information using threshold detection. When it detects that molecules “a” exceed the threshold while molecules “b” does not, it decodes the information as “10”; otherwise, it decodes the information as “01”. When both molecules “a” and “b” are detected to exceed the

threshold, the information is decoded as “11”. When neither molecules “a” nor molecules “b” exceeds the threshold, the information is decoded as “00”. The R adopts DF mode. In order to reduce the influence of ISI, the R uses different types of molecules to re-encode the decoded information. When the docoding information is “01”, the R releases a constant number of molecules “c”; when the decoding information is “10”, the R releases a constant number of molecules “d”; when the decoding information is “11”, the R releases a constant number of molecules “c” and “d”; when the decoding information is “00”, the R does not release any molecules.

Suppose that information molecules make Brownian motion in the fluid environment. That is, the information molecules have a velocity drift. The information molecules obeys the second Fick’s law of diffusion [26]. At first, we consider that the S, R and D are stationary. Then we consider the processing that information from S to R. The time slot is a random variable, which means that a molecule diffusion with drift from S to R, it defined as *t*, following

$$f(t) = \frac{d_0}{4\pi D\ t^3} e^{-\frac{(Vt-d_0)^2}{4D_p t}}, \qquad (1)$$

where *d*0 means the distance between S and R. Additionally, *V* stands for the drift velocity of the fluid medium, and *Dp* is the diffusion coefficient for the information molecules.

Next, we investigate a practical case, where both S and R are in mobility. Under this case, assuming the molecules are

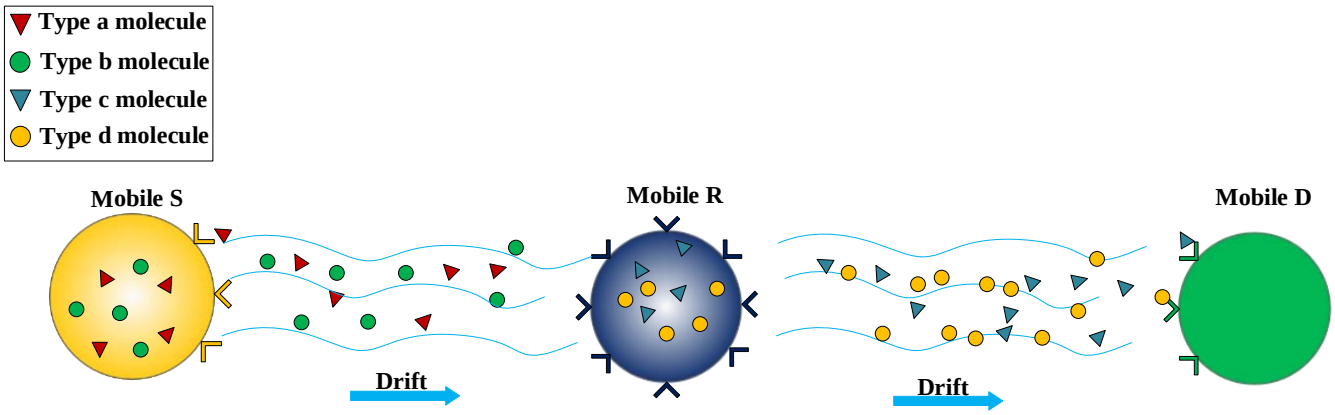


Fig. 1: The diffusive-drift molecular communication (MC) relay model with the mobile S, R and the mobile D. The red triangle represents the molecule of Type “a”; while the green circle stands for the molecule of Type “b”. The blue triangle represents the molecule of Type “c”; while the yellow circle stands for the molecule of Type “d”.

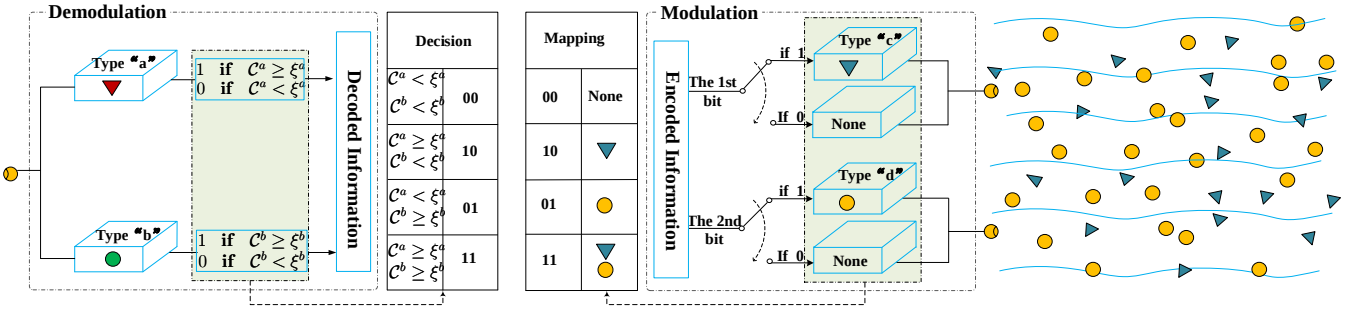


Fig. 2: Schematic diagram of the quarternary D-MoSK demodulation/modulation in a MC system at relay node. After receiving the information from the S, the R decodes it first. At the demoduator, the R captures the information molecule and decodes the data via a specially-designed decision device, in which the number of received molecules is the decision variable. *Ca* and *Cb* refer to the numbers of received molecules with Type “a” and Type “b”, respectively. Also, *_a* and *_b* denote the decision thresholds with respect to Type “a” and Type “b”. Then the R according to the D-MoSK modulation scheme to encode the information using molecules with Type “c” and Type “d”

transmitted by S at beginning of the *i*th time slot. According to the results in [27], the probability distribution function (PDF) for molecules of the first hitting time to reach R is shown in (2), where *erf*(·) denotes the standard error function. Also, *DS* and *DR* refer to diffusion coefficients of S and R, respectively. Besides, α and β are defined as $\alpha \triangleq DS + DR$ and $\beta \triangleq DR + Dp$, respectively.

It is supposed that a molecule transmitted by S at the beginning of the first time slot and reach R after *j* time slots, then the probability of the molecule being captured by the R satisfies,

$$F(t;j) = \int_0^{jT_s} f(t;i)\,\mathrm{d}t, \tag{3}$$

where *Ts* denotes the length of each time slot.

The D-MoSK modulation is used to modulate Quaternary information. We use two different types of molecules, namely Type "a" and Type "b". In this model, it is assumed that Type "a" and Type "b" have the same diffusion coefficients. The number of molecules captured by R at *n*th time slot can be expressed by

$$R^{a\text{ or }b}[n] = C^{a\text{ or }b}[n] + C_I^{a\text{ or }b}[n] + C_N[n], \tag{4}$$

in which *Ca* or *b*[*n*] stands for the number of information molecules that reach R at the *n*th time slot. *Ca* or *b I* [*n*], standing for the ISI, refers to the molecules released in the previous time slots but reach R at the *n*th time slot. Besides, *CN*[*n*] refers to the noise.

The molecules freely diffusive and keep independence with each other, then the number of the molecules captured by R, denoted by *Ca* or *b*[*n*], follows the binomial distribution, it follows

$$C^{a\text{ or }b}[n] \sim \mathcal{B}\left(Q^{a\text{ or }b}x[n],\ F(t;n)\right), \tag{5}$$

where *Q* denotes the released molecular number by S when the input bit is "1" as shown in Fig. 2, and *x*[*n*] represnts the transmitted bit by S at the *n*th time slot. Besides, the ISI term in (4), i.e., *Ca* or *I* (*n*), follows

$$C_I^{a\text{ or }b}[n] \sim \sum_{i=1} \mathcal{B}\left(Q^{a\text{ or }b}x[n-i], F(t;n) - F(t;n-1)\right), \tag{6}$$

where *I* denotes as the number of ISI. Generally, the influence of ISI gradually weak over time, and hence it is reasonable to assume the number of ISI is finite. Also, *x*[*n* − *i*] represents the transmitted bit by S at the (*n* − *i*)th time slot. The noise.

$$f(t;i) = \frac{\sqrt{iT_s\alpha\beta}}{\pi\sqrt{t}\,(iT_s\alpha+\beta t)} e^{-\frac{d_0^2}{4iT_s\alpha}} + f\left(t + iT_s\frac{\alpha}{\beta}\right)\cdot erf\quad \frac{d_0}{2}\sqrt{\frac{\beta t}{iT_s\alpha\,(iT_s\alpha+\beta t)}}\Bigg) \tag{2}$$

term in (4), i.e., *N*[*n*], follows Gaussian distribution, that is $C_N[n] \sim \mathcal{N}\ \ \mu_\omega,\ \sigma_\omega^2$.

On the basis of the central limit theorem, if *Ca* or *b* is sufficiently large, the binomial distribution in Eqns. (5) and (6) will approximates to the Gaussian distribution [28]. It is found in (4) that the terms keep statistically independent, and hence the receive signal at R follows Gaussian distribution, as shown in (7), in which *qi* and *F* are defined as *qi* ≜ *F*(*t*; *i*+1)−*F*(*t*; *i*) and *F* ≜ *F*(*t*; 1), respectively. The conditional probability of *Ra* or *b*[*n*] follows the Gaussian distribution as

$$\begin{cases} R^{a\text{ or }b}[n] \sim \mathcal{N}\left[\mu_0^{a\text{ or }b},\ \left(\sigma_0^{a\text{ or }b}\right)^2\right], & \text{if } x[n]=0; \\ R^{a\text{ or }b}[n] \sim \mathcal{N}\left[\mu_1^{a\text{ or }b},\ \left(\sigma_1^{a\text{ or }b}\right)^2\right], & \text{if } x[n]=1. \end{cases} \tag{8}$$

Define *p*0 and *p*1 as the probabilities with respect to transmitted bit to be "0" and "1", respectively. Furtherly, we can have

$$\begin{cases} \mu_0^{a\text{ or }b}[n] = & p_1 Q^{a\text{ or }b}\sum\limits_{i=1}^{I} q_i + \mu_\omega, \\ \mu_1^{a\text{ or }b}[n] = & Q^{a\text{ or }b}F + \mu_0^{a\text{ or }b}[n], \end{cases} \tag{9}$$

in which $\mu_0^{a\text{ or }b}$ and $\mu_1^{a\text{ or }b}$ denote the means of *Ra* or *b*[*n*] with the transmitted bit to be "0" and "1", respectively. Besides, the variations of *Ra* or *b*[*n*] can be calculated by

$$\begin{cases} \left(\sigma_0^{a\text{ or }b}[n]\right)^2 = & p_1 Q^{a\text{ or }b}\sum\limits_{i=1}^{I} q_i\,(1-q_i) \\ & +p_1 p_0\left(Q^{a\text{ or }b}\right)^2\sum\limits_{i=1}^{I} q_i^2 + \sigma_\omega^2, \\ \left(\sigma_1^{a\text{ or }b}[n]\right)^2 = & Q^{a\text{ or }b}F\,(1-F) + \left(\sigma_0^{a\text{ or }b}[n]\right)^2. \end{cases} \tag{10}$$

The R uses decode and forward scheme, so when the signal arrives R, it wiil be decoded, then re-encode the decoded information. The R uses different types of molecules to transmit information. It is assumed that R uses Type "c" and Type "d" to forward information.

The number of molecules captured by D at (*n* + 1)th time slot can be expressed by

$$D^{c\text{ or }d}[n+1] = C^{c\text{ or }d}[n+1] + C_I^{c\text{ or }d}[n+1] + N[n+1], \tag{11}$$

in which *Cc* or *d*[*n* + 1] stands for the number of information molecules that reach D at the (*n*+1)th time

slot. *Cc* or *d I* [*n*+1], standing for the ISI, refers to the molecules released in the previous time slots but reach D at the (*n* + 1)th time slot. Besides, *N*[*n* + 1] refers to the noise.

The molecules still diffusion and keep independence on each other, then the number of the molecules captured by D, denoted by *Cc* or *d*[*n*+1], follows the binomial distribution, it follows I

$$N^{c\text{ or }d}[n+1] \sim \mathcal{B}\left(Q^{c\text{ or }d}x[n+1],\ F(t;n+1)\right), \quad (12)$$

where *Qc* or *d* denotes the released molecular number by R when the input bit is "1", and *x*[*n*+1] represnts the transmitted bit by R at the (*n* + 1)th time slot. Besides, the ISI term in (11), i.e., *Cc* or *d I* (*n* + 1), follow

$$C_I^{c\text{ or }d}[n+1] \sim \sum_{i=1}^{I} \mathcal{B}\left(Q^{c\text{ or }d}x[n+1-i], F(t;n+1) - F(t;n)\right), \quad (13)$$

where *I* denotes as the number of ISI. Generally, the influence of ISI gradually weaks over time, and hence it is reasonable to assume the number of ISI is finite. Also, *x*[*n*+1 *i*] represents the transmitted bit by R at the (*n*+1 *i*)th time slot. The noise term in (11), i.e., *N*[*n*+1], follows Gaussian distribution, that is $N[n+1] \sim \mathcal{N}(\mu_\omega, \sigma_\omega^2)$.

Then the information diffusion follows same distribution from R to D, the mean and variance are calculated in the same way as from S to R. According to Eqns. (9) and (10), the corresponding parameters can be changed to calculate.

III. Signal Detection and Data Decoding

The maximum likelihood (ML) detection method is used in signal detection at R, and the likelihood ratio test (LRT) scheme satisfies

$$\begin{aligned} H_0:&\quad \xi = N_I^{a\text{ or }b}[n] + C_N[n], \\ H_1:&\quad \xi = R^{a\text{ or }b}[n], \end{aligned} \quad (14)$$

where *H*0 and *H*1 represent the ML conditions. So the detection threshold function follows

$$f(\xi) = \frac{p(\xi|H_1)}{p(\xi|H_0)} = \frac{f_\xi(\xi)}{f_\xi^{(0)}(\xi)} \underset{H_0}{\overset{1}{\gtrless}} \frac{p_0}{p_1} \quad (15)$$

in which $f(0)_\xi(\xi)$ and $f^{(1)}{}_\xi(\xi)$ represent the probability density function (PDF) of ξ in terms of *H*0 and *H*1, respectively, defined as

$$\begin{aligned} f_\xi^{(0)}(\xi) &\triangleq \frac{1}{\sqrt{2\pi\left(\sigma_0^{a\text{ or }b}\right)^2}} e^{-\frac{\left(\xi-\mu_0^{a\text{ or }b}\right)^2}{2\left(\sigma_0^{a\text{ or }b}\right)^2}}, \\ f_\xi^{(1)}(\xi) &\triangleq \frac{1}{\sqrt{2\pi\left(\sigma_1^{a\text{ or }b}\right)^2}} e^{-\frac{\left(\xi-\mu_1^{a\text{ or }b}\right)^2}{2(\sigma_1^{a\text{ or }b})^2}}. \end{aligned} \quad (16)$$

According to Eqns. (15) and (16), we can get the likelihood-ratio function,

$$\lambda(\xi) \triangleq \frac{f_\xi^{(0)}}{f_\xi^{(1)}} \quad (17)$$

It is assumed that the transmission of "0" and "1" are equal probability, let $\lambda(\xi) = 1$, and we can have,

$$\frac{1}{\sqrt{2\pi}\sigma_1^{a\text{ or }b}} e^{-\frac{\left(\xi-\mu_1^{a\text{ or }b}\right)^2}{2(\sigma_1^{a\text{ or }b})^2}} = \frac{1}{\sqrt{2\pi}\sigma_0^{a\text{ or }b}} e^{-\frac{\left(\xi-\mu_0^{a\text{ or }b}\right)^2}{2(\sigma_0^{a\text{ or }b})^2}} \quad (18)$$

$$R^{a\text{ or }b}[n] \sim \mathcal{N}\left(Q^{a\text{ or }b}F + \sum_{i=1}^{I} Q^{a\text{ or }b}x[n-i]q_i + \mu_\omega,\ Q^{a\text{ or }b}F(1-F) + \sum_{i=1}^{I} Q^{a\text{ or }b}x[n-i]q_i(1-q_i) + \sigma_\omega^2\right) \quad (7)$$

$$\begin{aligned} &\left[\left(\sigma_1^{a\text{ or }b}\right)^2 - \left(\sigma_0^{a\text{ or }b}\right)^2\right]\xi^2 - \left[2\mu_0^{a\text{ or }b}\left(\sigma_1^{a\text{ or }b}\right)^2 - 2\mu_1^{a\text{ or }b}\left(\sigma_0^{a\text{ or }b}\right)^2\right]\xi + \mu_0^{2a\text{ or }b}\left(\sigma_1^{a\text{ or }b}\right)^2 \\ &- \mu_1^{2a\text{ or }b}\left(\sigma_0^{a\text{ or }b}\right)^2 - 2\left(\sigma_1^{a\text{ or }b}\right)^2\left(\sigma_0^{a\text{ or }b}\right)^2 \ln\frac{\sigma_1^{a\text{ or }b}}{\sigma_0^{a\text{ or }b}} - 2\left(\sigma_1^{a\text{ or }b}\right)^2\left(\sigma_0^{a\text{ or }b}\right)^2 \ln\frac{p_0}{p_1} = 0. \end{aligned} \quad (19)$$

Then, according to calculate we can get the conclusion which is shown in (19). Furtherly, the detection threshold, i.e., _, can be calculated by

$$\xi = \text{round}\quad \frac{B + \sqrt{B^2 - AC}}{A}\quad , \quad (20)$$

where *A*, *B* and *C* are defined as

$$\begin{aligned} A &\triangleq \left(\sigma_1^{a\text{ or }b}\right)^2 - \left(\sigma_0^{a\text{ or }b}\right)^2, \\ B &\triangleq \mu_0^{a\text{ or }b}\left(\sigma_1^{a\text{ or }b}\right)^2 - \mu_1^{a\text{ or }b}\left(\sigma_0^{a\text{ or }b}\right)^2, \\ C &\triangleq \left(\mu_0^{a\text{ or }b}\sigma_1^{a\text{ or }b}\right)^2 - \left(\mu_1^{a\text{ or }b}\sigma_0^{a\text{ or }b}\right)^2 \\ &\quad - 2\left(\sigma_0^{a\text{ or }b}\sigma_1^{a\text{ or }b}\right)^2\left(\ln\frac{p_0}{p_1} - \ln\frac{\sigma_0^{a\text{ or }b}}{\sigma_1^{a\text{ or }b}}\right). \end{aligned}$$

Using the transfer probability, the symbol error rate (SER) of from S to R is derived by (21), in which $Q(\_)$ refers to the well-known Q-function.
Then, the SER of from R to D is derived by (22).
So the BER of from S to D can be calculated as

$$P_e = 1 - (1 - P_{e_{s,r}})(1 - P_{e_{r,d}}). \tag{23}$$

Based upon Shannon's information theory, the channel capacity is defined as the maximum of the mutual information, which is denoted by $I(X; Y)$, between the transmitted symbol X and the received symbol Y. Let Xn represent the signal transmitted by S in the nth time slot, and Yn represent the signal received by R at the nth time slot. Thus, the channel capacity of from S to R can be expressed by

$$C_{s,r} = \max I(X_n, Y_n) \text{ bit/slot.} \tag{24}$$

The mutual information can be calculated by (25), where $px = \Pr(Xn = x)$ and $\Pr(Yn = yjXn = x)$ refer to the *priori* probability and conditional probability, respectively.
while the channel capacity of from R to D can be expressed by

$$C_{r,d} = \max I(X_{n+1}, Y_{n+1}) \text{ bit/slot.} \tag{26}$$

in which $Xn+1$ represent the signal transmitted by R in the $(n+1)$th time slot, and $Yn+1$ represent the signal received by D at the $(n + 1)$th time slot.

So all the channel capacity of from S to D is expressed as

$$C = \min(C_{s,r}, C_{r,d}) \text{ bit/slot.} \tag{27}$$

IV. Numerical Results and Performance Analysis

In this section, the numerical results in (23) to evaluate the SER performance of D-MoSK modulation in a mobile diffusive-drift DF communication system are presented. The parameters used in the evaluations are summarized in Table I.

It can be found in Fig. 3 that D-MoSK exhibits a much better SER performance than MoSK in the decode-and-forward communication system. Here we need to point out that the D-MoSK modulation employs much fewer molecular types than MoSK modulation. Thus, the D-MoSK modulation is considered to have the capability to reduce the decoding complexity, as well as the hardware complexity. In this work, we use the number of molecular types that the R and D need to identify to evaluate the decoding complexity. Take the quaternary modulation as an example. For MoSK, the D needs to identify four types of molecules; while for D-MoSK, the D only needs to identify two types of molecules. For general comparisons, we investigate the ratio of decoding complexity between MoSK and D-MoSK, that is M=log2M, in which M stands for the modulation order. It can be accessible that along with the increase of modulation order, the advantage of D-MoSK modulation in terms of complexity performance will become more evident. Also, as mentioned above, four types of molecules are needed for quaternary MoSK modulation to form a symbol; while only two types are needed for quaternary D-MoSK modulation. Assume that S can release Q molecules within a bit time. For the MoSK modulation, the number of released molecules is $n\text{MoSK} = 4 \times 2 \times Q$. For the D-MoSK modulation, the number of released molecules is nD-MoSK = 2×2×Q. We can conclude that for the quaternary modulation, the number of molecules released in D-MoSK is half of that for MoSK.

In Fig. 4, we investigate the SER performances versus of the ISI length with different Q. From Fig. 4, we can find that SER curves go up along with the ISI length increases. The ISI refers to the information molecules transmitted by the previous time slot arrive in the current time slot. As can be seen from Fig.4, the closer the slot is to the current slot, the greater the impact of ISI. With the increase of time slot distance, the influence of ISI is smaller. When the length of ISI in more than 10, the influence on the system is basically unchanged. That is to say, the current time slot will be affected within 10 slots before the current time slot. In addition, the impact of ISI on the current time slot can be ignored. Therefore, the length of ISI is set to 10 in this paper.

$$\begin{aligned} P_{e_{s,r}} \triangleq\ & p(00)\left[p(01|00) + p(10|00) + p(11|00)\right] + p(01)\left[p(00|01) + p(10|01) + p(11|01)\right] \\ & + p(10)\left[p(00|10) + p(01|10) + p(11|10)\right] + p(11)\left[p(00|11) + p(01|11) + p(10|11)\right] \\ =& \frac{1}{4}\left\{4 - \left[1 - Q\left(\frac{\xi - \mu_0^a}{\sigma_0^a}\right)\right]\left[1 - Q\left(\frac{\xi - \mu_0^b}{\sigma_0^b}\right)\right] - \left[1 - Q\left(\frac{\mu_1^a - \xi}{\sigma_1^a}\right)\right]\left[1 - Q\left(\frac{\xi - \mu_0^b}{\sigma_0^b}\right)\right]\right. \\ & \left. - \left[1 - Q\left(\frac{\mu_1^b - \xi}{\sigma_1^b}\right)\right]\left[1 - Q\left(\frac{\xi - \mu_0^a}{\sigma_0^a}\right)\right] - \left[1 - Q\left(\frac{\mu_1^b - \xi}{\sigma_1^b}\right)\right]\left[1 - Q\left(\frac{\mu_1^a - \xi}{\sigma_1^a}\right)\right]\right\} \end{aligned} \tag{21}$$

$$
\begin{aligned}
P_{e_{r,d}} \triangleq\ & p(00)\left[p(01|00)+p(10|00)+p(11|00)\right]+p(01)\left[p(00|01)+p(10|01)+p(11|01)\right] \\
& +p(10)\left[p(00|10)+p(01|10)+p(11|10)\right]+p(11)\left[p(00|11)+p(01|11)+p(10|11)\right] \\
=& \frac{1}{4}\left\{4-\left[1-Q\left(\frac{\xi-\mu_0^c}{\sigma_0^c}\right)\right]\left[1-Q\left(\frac{\xi-\mu_0^d}{\sigma_0^d}\right)\right]-\left[1-Q\left(\frac{\mu_1^c-\xi}{\sigma_1^c}\right)\right]\left[1-Q\left(\frac{\xi-\mu_0^d}{\sigma_0^d}\right)\right]\right. \\
& \left.-\left[1-Q\left(\frac{\mu_1^d-\xi}{\sigma_1^d}\right)\right]\left[1-Q\left(\frac{\xi-\mu_0^c}{\sigma_0^c}\right)\right]-\left[1-Q\left(\frac{\mu_1^d-\xi}{\sigma_1^d}\right)\right]\left[1-Q\left(\frac{\mu_1^c-\xi}{\sigma_1^c}\right)\right]\right\}
\end{aligned}
\tag{22}
$$

$$
I(X_n;Y_n)=\sum_y\sum_x p_x \Pr(Y_n=y|X_n=x)\log_2\frac{\Pr(Y_n=y|X_n=x)}{\sum\limits_x p_x \Pr(Y_n=y|X_n=x)}
\tag{25}
$$

Table 1: Parameters Used in the Numerical Results

Definition	Symbol	Value
Diffusion coefficient of information molecules	D_a, D_b, D_c, D_d	$[1, 50] \times 10^{-10}\ m^2/s$
Diffusion coefficient of S	D_S	$[1, 100] \times 10^{-14}\ m^2/s$
Diffusion coefficient of R	D_R	$[1, 100] \times 10^{-13}\ m^2/s$
Initial distance between S and R	d_0	$10\ \mu m$
Symbol time	T_s	$[0,\ 0.2\ s]$
Number of molecules transmitted by S	Q_a, Q_b, Q_c, Q_d	$(0, 800]$
Velocity of fluid medium	V	$[0.1, 1] \times 10^{-3}$
Mean of noisy molecule	μ_ω	0
Variance of noisy molecule	σ_ω^2	300
Length of ISI	I	10

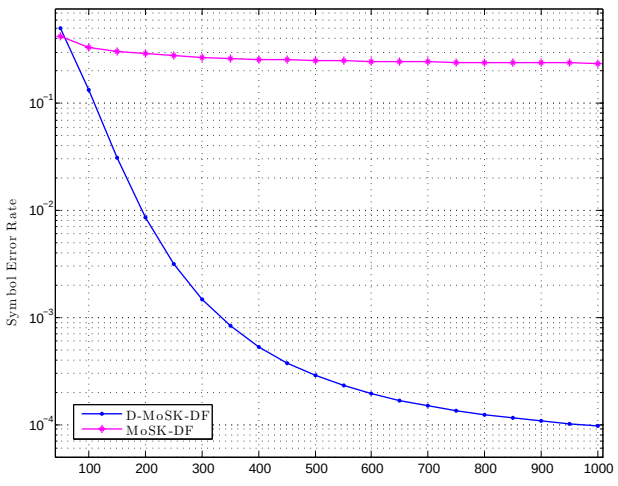


Fig. 3: Comparisons of SER performances between MoSK and D-MoSK.

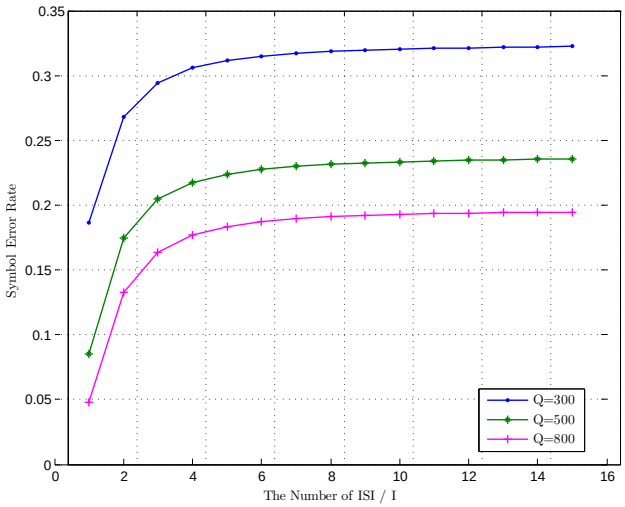


Fig. 4: SER performance versus length of ISI with different Q.

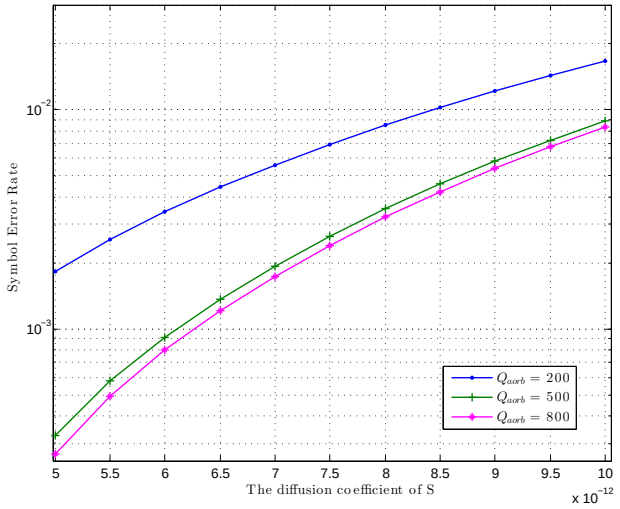


Fig. 5: SER performance versus diffusion coefficient of S. Here *Q* refers to the number of molecules transmitted by S.

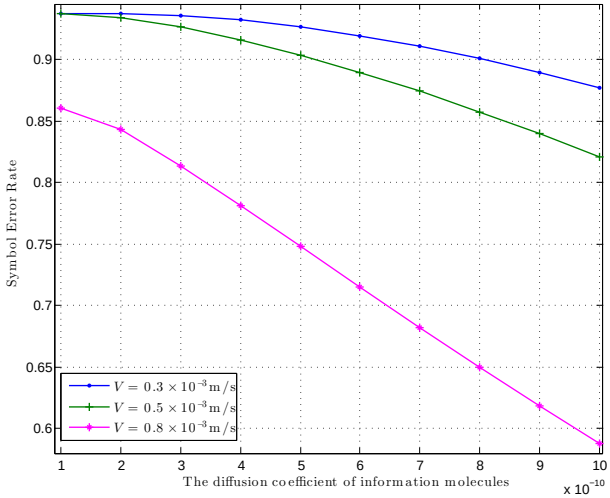


Fig. 7: SER versus different diffusion coefficient of information molecules with different *V* .

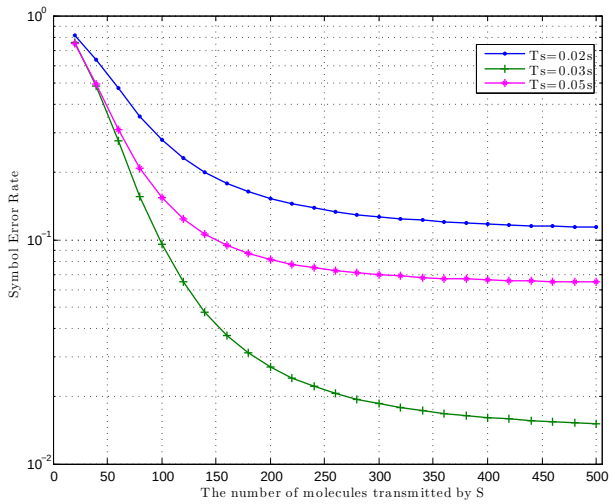


Fig. 6: SER performance versus number of information molecules, i.e., Q transmitted by S with different Ts.

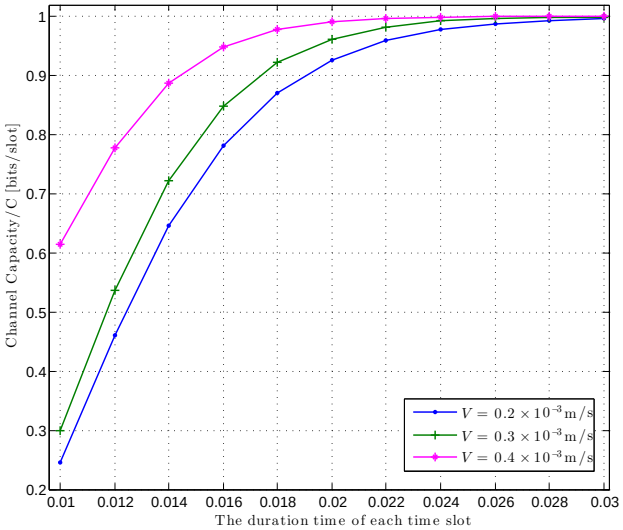


Fig. 8: Channel capacity versus duration time of each time slot with differentV . Here V denotes the driftvelocity.

From Fig. 5, it can be concluded that with the increase of the diffusion coefficient of S, the SER of system increases. The diffusion coefficient of S means the moving speed. The speed increases will bring much uncertainty, it causes channel fading. Then the molecules captured by R or D will decrease. So from Fig. 5, we can see that SER becomes decrease with larger diffusion coefficient. The increase of the number of molecules can make up for part of the channel fading, so the number of molecules increases and the SER decreases.

In Fig. 6 , we explore the SER performance versus the number of transmitted molecules with different *Ts*. It can be seen from Fig. 6 that SER decreases as the number of transmitted molecules increases, since more information molecules will be captured by R or D within a time slot. Additionally, we can find that with the same number of transmitted molecules, prolonging the time slot will decrease the SER.

In Fig. 7, we investigate the SER performance versus the Fig. 8. Channel capacity versus duration time of each time slot with different *V* . Here *V* denotes the drift velocity. diffusion coefficient of information molecules with different *V* . With the increase of diffusion coefficient of information molecules, the diffusion speed of information molecules in the channel is accelerated, which directly leads to the decreases of the time for molecules to arrive the destination, and the probability of being captured by the receiver in the same time increases, which makes the SER decrease. In addition, the increase of liquid velocity also accelerates the speed of information molecules. Therefore, the faster the liquid flow rate, the smaller the SER.

From Fig. 8, we can see that channel capacity along with the time duration of each time slot increase. With the increase of slot length, the number of information molecules captured at destination increases which can reduce the SER. Along with the increase of liquid velocity, the velocity of molecular diffusion is also

increased, and the number of captured information molecules increase, and the SER decrease.

From the previous discussion, we can draw the following conclusions: with the increase of liquid flow velocity, the molecular movement speed is accelerated, which makes the number of molecules captured in the per time slot increase, and the system performance is improved. It can be seen from the Fig. 9 that the simulation results are consistent with the previous conclusions. Meanwhile, with the increase of slot length, the number of molecules captured in the same time slot increases, which can also improve the system performance.

V. Conclusions

In this work, a diffusive-drift MC relay system model with mobile S and mobile D was investigated. The D-MoSK modulation is employed to this system model and the performance is analyzed. In order to reduce the decode complicated, the R uses DF scheme and different types of molecules. We introduce the ML criterion at R and D to decode the information. Meanwhile, the analytical results in terms of SER and capacity are derived. The numerical results show that D-MoSK exhibits better SER performances than the MoSK modulation.

References Références Referencias

1. J. Wang, B. Yin, and M. Peng, "Diffusion based molecular communication: Principle, key technologies, and challenges," *China Communications*, vol. 14, no. 2, pp. 1-18, Feb. 2017.
2. H. Sawai, "Biological functions for information and communication techmologies," *Springer*, 2011.
3. T. Nakano, A.W. Eckford, T. Haraguchi, "Molecular communication," *Cambridge University Press*, 2013.
4. N. Farsad, H. B. Yilmaz, A. Eckford, *et al:*, "A comprehensive survey of recent advancements in molecular communication," *IEEE Communications surveys & tutorials*, vol. 18, no. 3, pp. 1887-1919, Feb. 2016.
5. T. Nakano, M. J. Moore, F. Wei, *et al:*, "Molecular communication and networking: opportunities and challenges," *IEEE Transactions on Nanobioscience*, vol. 11, no.2, pp. 135-148, June. 2012.
6. U. Okonkwo, R. Malekian, B. T. Maharaj, *et al:*,"Molecular communication and nanonetwork for targeted drug delivery: A survey," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 3046-3096, May. 2017.
7. B. Atakan, "Molecular communications and nanonetworks," *Springer*, 2014.
8. D. Malak, O. B. Akan," Molecular communication nanonetworks inside human body", *Nano Communication Networks*, vol. 3, no. 1, pp. 19-35, Mar. 2012.
9. L. Felicetti, M. Femminella, G. Reali, *et al:*, "Applications of molecular communications to medicine: A survey," *Nano Commun. Netw.*, vol. 7, pp. 27-45, Mar. 2016.
10. T. Nakano, M. J. Moore, F. Wei, *et al:*, "Molecular communication and networking: Opportunities and challenges," *IEEE Trans. Nanobiosci.*, vol. 11, no. 2, pp. 135-148, Jun. 2012.
11. A. Einolghozati, M. Sardari, F. Fekri, "Relaying in diffusion-based molecular communication," *IEEE International Symposium on Information Theory (ISIT)*, Istanbul, Turkey, July, 2013.
12. X. Wang, M. Higgins, M. Leeson, "Relay analysis in molecular communications with time-dependent concentration", *IEEE Transactions on Molecular, Biological & Multi-Scale Communications*, vol. 19, no. 11, pp. 1977-1980, Sep. 2015.
13. A. Einolghozati, M. Sardari, F. Fekri, "Decode-and-forward relaying in diffusion-based molecular communication between two populations of biological agents", *IEEE International Conference on Communications (ICC)*, Sydney, NSW, Australia, June, 2014.
14. A. Ahmadzadeh, A. Noel, A. Burkovski,*et al:*, "Amplify-and-forward relaying in two-hop diffusion-based molecular communication networks", *IEEE Global Communications Conference(GLOBECOM)*, San Diego, CA, USA, Dec, 2015.
15. S. Yuan, J. Wang, M. Peng, "Performance analysis of reversible binding receptor based decode-and-forward relay in molecular communication systems", *IEEE Wireless Communications Letters*,vol. 7, no. 5, pp. 880- 883,Oct. 2018.
16. P. Manocha, G. Chandwani, S. Das, "Dielectrophoretic relay assisted molecular communication for in-sequence molecule delivery", *IEEE Transactions on Nanobioscience*, vol. 15, no. 7, pp. 781-791,Oct. 2016.
17. J. Wang, M. Peng, X. Liu, *et al:* , "Performance analysis of signal detection for amplify-and-forward relay in diffusion-based molecular communication systems ", *IEEE Internet of Things Journal*, vol. 7, no. 2, pp. 1401-1412,Feb. 2020.
18. N. Tavakkoli, P. Azmi, N. Mokari, "Performance evaluation and optimal detection of relay-assisted diffusion-based molecular communication with drift ", *IEEE Transactions on Nanobioscience*, vol. 16, no. 1, pp. 34-42, Jan. 2017.
19. N. Tavakkoli, P. Azmi, N. Mokari, "Optimal positioning of relay node in cooperative molecular communication networks ", *IEEE Transactions on Communications*, vol. 65, no. 12, pp. 5293-5304, Dec. 2017.
20. G. Chang, L. Lin, H. Yan, "Adaptive detection and ISI mitigation for mobile molecular communication",

IEEE Transactions on NanoBioScience, vol. 17, no. 1, pp.21-35, Dec. 2017.

21. N. Pandey, S. Hoshi, R. Mallik, *et al:*, "Channel characterization for devices in a turbulent diffusive environment: A mobile molecular communication approach", *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 5, no. 3, pp.222-232, Dec. 2019.
22. Z. Cheng, Y. Zhang, M. Xia, "Performance analysis of diffusive mobile multiuser molecular communication with drift", *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 4, no. 4, pp.237-247, Dec. 2018.
23. L. Chouhan, P. K. Sharma, N. Varshney, "Optimal transmitted molecules and decision threshold for drift-induced diffusive molecular channel with mobile nanomachines", *IEEE Transactions on NanoBioscience*, vol. 18, no. 4, pp.651-660, Oct. 2019.
24. X. Mu, H. Yan, B. Li, *et al:* , "Low-complexity adaptive signal detection for mobile molecular communication", *IEEE Transactions on NanoBioscience*, vol. 19, no. 2, pp. 237-248, Apr. 2020.
25. S. Huang, L. Lin, H. Yan, *et al:* , "Statistical analysis of received signal and error performance for mobile molecular communication", *IEEE Transactions on NanoBioscience*, vol. 18, no. 3, pp. 415-427, July. 2019.
26. R. Mosayebi, A. Gohari, M. Mirmohseni, *et al:* , "Type-based sign modulation and its application for ISI mitigation in molecular communication", *IEEE Transactions on Communications*, vol. 66, no. 1, pp. 180-193, Jan. 2018.
27. N. Varshney, W. Haselmayr, W. Guo , "On flow-induced diffusive mobile molecular communication: First hitting time and performance analysis", *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 4, no. 4, pp. 195-207, Dec. 2018.
28. T. Nakano, M. Moore, A. Enomote , "Biological Function for Information and Communication Technologies", *Heidelberg, Germany: Springer*, 2011.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 22 Issue 1 Version 1.0 Year 2022
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Optimising Sargable Conjunctive Predicate Queries in the Context of Big Data

By Veronica V.N. Akwukwuma & Patrick O. Obilikwu
Benue State University

Abstract- With the continued increase in the volume of data, the volume dimension of big data has become a significant factor in estimating query time. When all other factors are held constant, query time increases as the volume of data increases and vice versa. To enhance query time, several techniques have come out of research efforts in this direction. One of such techniques is factorisation of query predicates. Factorisation has been used as a query optimization technique for the general class of predicates but has been found inapplicable to the subclass of sargable conjunctive equality predicates. Experiments performed exposed a peculiar nature of sargable conjunctive equality predicates based on which insight, the concatenated predicate model was formulated as capable of optimising sargable conjunctive equality predicates. Equations from research results were combined in a way that theorems describing the application and optimality of the concatenated predicate model were derived and proved.

Keywords: concatenated predicate, conjunctive equality predicate, sargable predicate, query, factorisation, database, software applications.

GJCST-C Classification: I.2.4

Strictly as per the compliance and regulations of:

Optimising Sargable Conjunctive Predicate Queries in the Context of Big Data

Veronica V.N. Akwukwuma [α] & Patrick O. Obilikwu [σ]

Abstract- With the continued increase in the volume of data, the volume dimension of big data has become a significant factor in estimating query time. When all other factors are held constant, query time increases as the volume of data increases and vice versa. To enhance query time, several techniques have come out of research efforts in this direction. One of such techniques is factorisation of query predicates. Factorisation has been used as a query optimization technique for the general class of predicates but has been found inapplicable to the subclass of sargable conjunctive equality predicates. Experiments performed exposed a peculiar nature of sargable conjunctive equality predicates based on which insight, the concatenated predicate model was formulated as capable of optimising sargable conjunctive equality predicates. Equations from research results were combined in a way that theorems describing the application and optimality of the concatenated predicate model were derived and proved. The theorems proved that the novel concatenated predicate model transforms a sargable conjunctive equality predicate such that the resultant concatenated predicate is an optimal equivalent of the sargable conjunctive equality predicate from which it is derived. The model enhances conjunctive sargable equality queries making our results capable of application in software applications, majority of whose queries are of the conjunctive query type. The results are equally useful in optimising query time within the context of Big Data where the continuous increase in the volume dimension of data calls for query structures that enhance query time.

Keywords: concatenated predicate, conjunctive equality predicate, sargable predicate, query, factorisation, database, software applications.

I. Background to Study

The fundamental Vs of Big Data are volume, velocity and variety [1]. Volume refers to the size of data being created, Velocity is the speed at which data is created, captured, extracted, processed, and stored while variety connotes different data types and sources ranging from structured, semi-structured to unstructured data. Of the three Vs, volume is most directly associated with big data and to put its importance in a perspective that emphasizes its relevance to query optimisation, volume may be redefined as voluminosity, vacuum, and vitality – three additional V-dimensions of data as exposed by [2]. Voluminosity states that there is already a very large set of data collected and even much more is available that can be harvested. Voluminosity speaks of a significant gap that can be filled by data yet to be collected. From the perspective of voluminosity, volume refers to the size of data being created from all sources in an organization including text, audio, video, social networks, research studies, medical data, space images, crime reports, weather forecasting and natural disaster [3].

The vacuum dimension of volume states that there is a strong requirement for storage to store large volumes of data. Due to the fact that the data is acquired incrementally, empty spaces will always be needed for use in the creation of room to store, process and manage tremendous data set as they are harvested from different sources. This dimension of volume pops up the research question about how much storage space is available for incoming data rather than how much data has already been stored. The process of creating storage space for incoming data is equally as challenging as it is with managing vast sets of already stored data. Empty spaces that serve this purpose are created by either augmenting storage devices or techniques used to compress the size of data [4].

Vitality may be defined as the survival of data in the storage environment and thus its reliability and usefulness. Data in the storage environment falls into the two categories, namely active served and unserved. In a large data bank, some data are actively used while some are not [4]. Vitality redefines volume as meaning that data and its subsets are used actively at different times. While a portion of data may be actively used data at a time or within a specific transaction, the rest are stored for future uses. There is the risk that data stored for future may take so long for it to be used which may lead to such sub-datasets to be abandoned or not properly maintained. As the risk of being abandoned gets higher, anything can happen to those datasets not currently in use. In other words, with less investment and attention to the unserved data, they are exposed to incidences of fire, earthquake, flood, war, and terrorist which are the prominent causes of data loss. Thus, vitality is a critical component of volume. The lack of vitality, in any case, is symptomatic of the absence of disaster management systems which decimates data reliability or can lead to complete data loss. Apart from reliability, vitality also describes flexibility, dependability,

Author α: Ph.D, Department of Computer Science, University of Benin, Benin City, Nigeria.
Author σ: Ph.D, Department of Computer Science, Benue State University, Makurdi, Nigeria. e-mail: poblikwu@gmail.com

and security which are all integral components of volume,

As data gets larger in the dimensions of big data, partitioning strategies have been used to reduce the data to smaller subsets over which queries become faster compared to the original dataset [5]. Popular among these partitioning strategies is the horizontal scaling (scaling out), Horizontal scaling refers to resource increment by the addition of complete and independent units that work in unison with an existing system. The additional units may be of smaller capacity, making it cheaper compared to the replacement of an existing single unit with one of larger capacity. The scale out effect of the horizontal partitioning strategy creates a hardware infrastructure platform on which partitioned data is then distributed across multiple units or servers, hence, reducing the excess load of the entire data set on a single machine [6,7]. This platform comes with the added advantage of keeping the entire system up even if some of the units go down, thus, avoiding the "single point of failure" problem associated with vertical scaling. The vertical scaling (scaling up) strategy refers to increasing the ability of a single hardware unit such as a server to handle the ever-increasing workload as a way of achieving resource increment. From the perspective of hardware, this includes adding memory and processing power to the single unit.

The horizontal scaling strategy is at the heart of the implementation of big data stores namely p-stores, c-stores and NoSql among others that have pioneered the paradigm shift of "No One Size Fits-All" proposed by Stonebraker and Çetintemel [8]. The horizontal scaling strategy partitions data such that queries can be fired selectively on the partitions with the aim of retrieving the desired data in optimal query time. As is applicable to all datasets, the desired data in a partition is indicated in a query using a boolean expression of conditions called predicates. Predicates are used in joins as well in search arguments of queries. A join predicate is a predicate that relates columns of two tables to be joined and the columns referenced in a join predicate are called join columns. When used in Search ARGuments (SARGs), predicates are referred to as sargable predicates [9]. A sargable predicate is one of the form (or which can be put into the form) "column comparison-operator value". Matalqa and Mustafa [5] experimentally demonstrated that restructuring big data into partitions produces query enhancement results. Using the theorem and axiom, Obilikwu, Kwaghtyo and Ogbuju [10] theoretically proved the result of [5] as follows:

Theorem: Given P_1, P_2 … P_n as the partitions of a relation R, then R = {P_1, P_2, …, P_n} where n = the number of distinct values in the value set associated with the partition key that generated P_1, P_2 … P_n

Axiom: The following axioms are applicable:

1. A partition key has a value set, V whose element cannot be null
2. The number of distinct values of V is n= number of partitions produced

Proof: Let $\boldsymbol{\sigma}$ be the partition predicate associated with a distinct value of V, then Arity($\boldsymbol{\sigma}$) is the arity of the tuples filtered by $\boldsymbol{\sigma}$.

Given any value of n, there exists σ_1, σ_2, ..., σ_n, where

σ_1 filters all tuples in P_1 from relation R,

σ_2 filters all tuples in P_2 from relation R, and

σ_n filters all tuples in P_n from relation R,

Since the elements of V cannot be null, then Arity(V) = Arity (R)

Since σ_1, σ_2, ..., σ_n filter the tuples of R according to the distinct values of V, it follows that

$\text{Arity(V)}=\text{Arity}(\sigma_1) + \text{Arity}(\sigma_2) + ... + \text{Arity}(\sigma_n)=\sum_i^n Arity(\sigma_i)$

This implies that $\sum_i^n Arity(\sigma_i)$= Arity (R) since n is the number of distinct values of V defined in R

This shows that R = {P_1, P_2, …, P_n} since σ_1, σ_2, ..., σ_n filter the tuples of R. QED.

The use of partitioning strategies makes queries faster [5]. This is because retrieving a record or a set of records from a relation is done relative to the number of the total number of records in the relation (R). Based on this relationship, query time can be computed as a ratio using equation 1.

$$q_t = \frac{t_R}{T_R} \quad \ldots \qquad (1)$$

where q_t is query time, t_R is the number of tuples retrieved from a relation R using a predicate σ and T_R is number of tuples in R. Equation 1 assumes an asymptotic value of t_R as well as the fact that other factors that affect query time are held constant. Among others, these other factors are processor speed, RAM and ROM size, communication traffic and code efficiency.

The implication of equation 1 is that an increase in volume implies an increase in query time. The query works with the DBMS as part of the algorithms that ensure data is retrieved seamlessly. While the DBMS suggests how the data can be located and retrieved, the query syntax tells what data is to be retrieved. These make up the two components of a database management system as depicted in Figure 1.

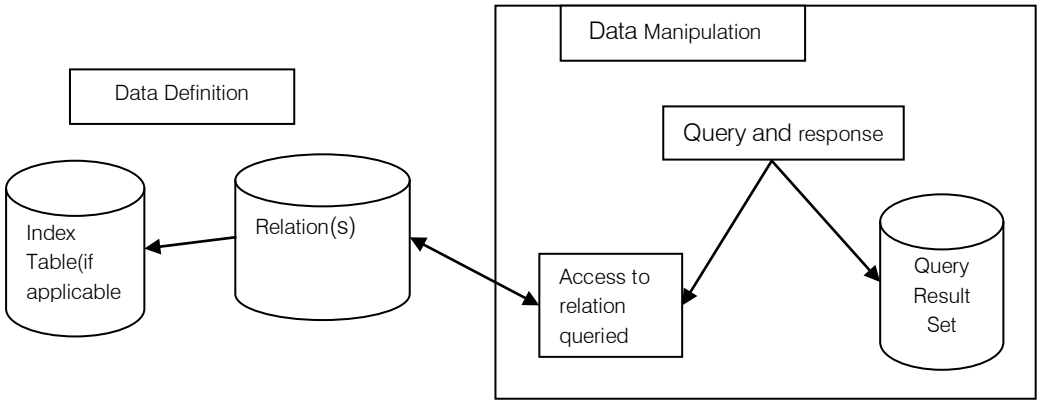


Figure 2: Architecture of Database Management

This paper is motivated by the critical need to optimise queries in the context of big data, big data being a development that has led to the ubiquitous incidence of big databases. The objectives of the paper are therefore as follows: (i) show that query time increases as the arity of database storage structures increase; (ii) show that optimising query time can be approached by organizing the storage structure using techniques like indexing and storage partitioning. It is also shown that queries can be modified or transformed to an equivalent form such that query time is reduced; (iii) use a combination of mathematical techniques to develop the concatenated predicate model thus enhancing the query time of sargable conjunctive equality predicates (iv) prove using mathematical induction and other applicable techniques that the concatenated predicate model optimises the sargable conjunctive equality predicate.

The rest of this paper is organized as follows: Section 2 reviews literature on the general concept of query optimisation and subsequently narrows the discussion down to the specific class of conjunctive predicates and how optimisation of predicates enhances query time. In Section 3, the product function is presented as a mathematical model to describe the product of atomic predicates, an operation also referred to as concatenation. Concatenation achieves literal minimisation as an alternative to factorization where there are no common atomic predicates. Concatenation in this paper to propose the concatenated predicate model. In Section 4, the results of this study are demonstrated using mathematical induction and other proofs. The proofs are discussed relative to the expected behavior of the concatenated predicate model.Finally, Section 5 concludes the paper and makes suggestions for future work.

II. Related Work

A predicate is that part of the query that filters records based on certain conditions. The properties of a predicate are multifarious and their study has exposed opportunities for optimising them, given that optimising them ultimately optimises database query time. Techniques for optimising queries are dependent on the query type.

a) Conjunctive Queries

Conjunctive queries represent one of the query languages used to retrieve data from relational databases [11,12,13] among other database models. Conjunctive queries correspond to the non-recursive Datalog rules [14]. In recursive datalog rules, conjunctive queries are of the form,

$$R_0(u_0) \leftarrow Ri_1(u_1) \wedge Ri_2(u_2) \wedge \ldots Ri_m(u_m)$$

where Ri_j is the relation name of the underlying database. R_0 is the output relation, and where each argument u_j is a list of $|u_j|$ variables, where $|u_j|$ is the arity of the corresponding relation Ri_j.

Conjunctive queries consist strictly of conjunctive predicates and they are the most widely used database queries in practice. It is against the background that optimising them makes a whole lot of sense [15,16,17,18,19]. The wide use of conjunctive queries are observable in not only their ubiquitous use in decision support systems based on relational databases but in other areas such as Description Language queries used to query knowledge representation (KR) systems , ontology-based queries and query answering frameworks in general [20,21]. Optimising a conjunctive query simply means optimising the conjunctive predicate component.

Heimel et al. [22] defined conjunctive predicates mathematically as

$$\theta = \Lambda_{i=1}^{m} \theta_i$$

where θ_i are atomic predicates joined by the AND relational operators and $i = 1,2, .., m$ are predicate terms (predicate literals, Boolean variables or atomic predicates) making up the conjunctive predicate. Sargable conjunctive predicates were defined by Yu X et al.[23] as conjunctive predicates of the form,

$$Q = P_1 \wedge P_2 \wedge \ldots \wedge P_m$$

where each component P_i, $i>0$ is an atomic predicate of the attribute value pair (*attribute* op *value*) with *op* being one of the comparison operators $<, \leq, =, \neq, \geq$ *or* $>$.

Practically speaking, conjunctive predicates are identified in the filter component of the project-select-join queries in the relational algebra, and in the where-clause of SQL queries having the general form SELECT . . . FROM . . . WHERE . . . where the where-clause is a predicate clause. Predicates are the conditions based on which database queries filter tuples in a relation or group of related relations. In its basic form, a query predicate is an atomic conditional expression also referred to as an atomic predicate. Several atomic predicates can be combined using logical operators to make up complex predicates [24] and the number of atomic predicates in a complex predicate is the boolean factor [9]. An atomic predicate has a boolean factor of 1. Boolean factors are notable because every tuple returned by a query must satisfy every boolean factor. A complex predicate made up of atomic predicates joined strictly using the AND logical operator is referred to as a conjunctive predicate. If all the atomic predicates in a complex predicate consist strictly of the equality operator, the complex predicate is referred to as a conjunctive equality predicate. Assuming the logical operator in the complex predicate is the OR logical operator then the resulting predicate will be a disjunctive predicate [25]. If the relational operator is the equality operator, then the complex predicate is a conjunctive equality predicate. If the conjunctive equality predicate is sargable, then it referred to as a sargable conjunctive equality predicate This paper is a study on how predicates of the class of sargable conjunctive equality predicates can be optimised.

b) Query Optimisation

Big data is resource-intensive and hence requires that both storage and query time are optimised for effective resource utilization. Resource optimisation, be it hardware or otherwise has been discussed within the larger context of solutions that we can never have enough of [26]. As a matter of fact the optimisation problem domain is one we are not yet done with [27]. Optimizing a number of running processes is considered an optimisation strategy though via software. Optimising query time by software (algorithms) is traditionally a function of the query optimizer, which is internal to the DBMS [9]. The algorithms associated with the query optimiser manipulate a query plan in its internal structure to choose an optimal plan for implementing a query. Query optimization gained research attention when the advantages of the relational data model in terms of user productivity and data independence became widely recognized in response to Codd's original ideas about the concept of relational databases [28]. Following this development, researchers began to ask questions about whether or not an automatic system can choose as efficient an algorithm for processing a complex query as a trained programmer would. System R, an experimental system was then constructed at the San Jose IBM Research Laboratory to demonstrate that a relational database system can incorporate the high performance and complete function, including automatic query optimisation required for everyday production use [9,29].

Query optimization has also been associated with modifying the structure of relations. In this regard, indexing can be said to be a pioneering effort at optimising query time from the dimension of database structure [30,31]. In processing a query that has a predicate, the attributes in the predicate are examined to find out if an index has been defined for any of the attributes, a concept referred to as index availability. The availability of an index makes searching relations faster compared to a full scan which is the search option used in the absence of an index. On the other hand, an index scan is used for the search if an index is available. The implementation of a full scan uses sequential search while an index scan is implemented using binary search. It is established in algorithmic theory that sequential search is of $O(n)$ and binary search is $O(\log n)$ making it obvious that an index scan is faster thereby enhancing query time.

Queries are also faster when relations are normalized. Partitioning relations also achieve good results. Optimising query operators, especially SELECTION and JOIN operators equally enhance query time. Incidentally, research into the optimisation of query operators has focused on joins and their ordering to the near neglect of research into the optimization of selection predicates [24]. Query optimisation is an open ended research question and hence it has been the object of research efforts over the years [26,9,15,32,33,34,35,36].

c) Predicate Optimisation

Query optimisation research efforts over the years in the specific area of predicate optimization have resulted in several optimization techniques notable among which are Predicate Pushdown [37], LDL approach [38,39], Predicate Move-around [12], Predicate Migration [40], By-Pass Predicate Processing [25], Optimising User-defined functions using Pruning Strategies [41,42]. Prominent among this technique is factorisation, a technique used to mininise the number of atomic predicates or terms in a complex predicate. Kemper et al. [43] and Chaudhuri et al.[24] used factorization to minimise atomic predicates in queries. The objective of factorization is to represent a Boolean function in a logically equivalent factored form having a

minimum number of literals [44]. The concept of minimizing atomic predicates (predicate literals) in a Boolean expression means that such expressions can be made simpler by reducing the terms in them. Predicate literals are found in the design of VSLI [45], compilers [46], and database query predicates [43,24] and minimization techniques of various types have been applied in each of these application areas thereby optimising the expressions involved. Factorisation is however only possible where the predicate expression has common atomic predicates.

Muralikrishna and DeWitt [47] established that the number of times a relation in a query is scanned is equal to the number of terms in which attributes of the relation are involved. This means that minimising the number of terms equally minimises the number of scans for each relation. Scanning constitutes a fundamental operation in query processing and thus a reduction in the number of scans done by a query equally reduces query time. Chaudhuri et al. [24] showed that factorization can be used to minimize predicate terms in scenarios where there are common atomic predicate factors. In this work, the sargable conjunctive equality predicates have been exposed as incidences of predicates where there are no common atomic predicate factors implying that factorization is inapplicable as a predicate minimisation technique. Sargable conjunctive equality predicates do not have common Boolean factors because an atomic predicate appearing more than once in a sargable conjunctive equality predicate duplicates such an atomic predicate. The duplicate atomic predicate is redundant and the result of such is unsatisfiable and evaluating them would lead to incorrect results [22]. This motivates the study of the nature of optimisation problems inherent in sargable conjunctive equality predicates. The experiments performed exposed interesting insights as to why existing predicate optimistion techniques, particularly factorization are inapplicable.

d) Nature of Optimisation Problem Posed by Sargable conjunctive equality predicates

To optimise sargable conjunctive equality predicates, there is need to understand the nature of optimisation problem posed by them. Series of experiments were conducted using a simulated data of students scores in an examination to expose what happens in terms of query time when the number of atomic predicates in a sargable conjunctive equality predicate is varied in a query. The experiments performed assumed that a number of students took an examination in the Department of Physics of a hypothetical University. The examination results are captured in a database relation, named studentscores. A schema is defined for the relation as studentscores(sno, studentID, level, courseCode, semesterID, sessionID, status, score) where the attributes are described as follows: sno (serial number); studentID (unique identifier for student); courseCode (semester course code); semesterID (identifier for semester); sessionID (identifier for session); status (semester course status) and score (an attribute for students score in the examination). Five instances of this schema are shown in Table 1.

Table 1: Instances of Examination Results Schema (Query Table)

Sno.	StudentID	Level	CourseCode	SemesterID	SessionID	Status	Score
1	SCN890178254	400	PHY412	1ST	2016/2017	C	94
2	SCN907524101	400	PHY412	1ST	2016/2017	C	65
3	SCN901782548	400	PHY412	1ST	2016/2017	C	76
4	SCN898888254	400	PHY412	1ST	2016/2017	C	35
5	SCN895428266	400	PHY412	1ST	2016/2017	C	58

The instances of the relational schema generated in Table 1 are five but the assumption is that as many students as there wrote the examination in the physics course (PHY412). The level is 400 (a course taken at the fourth year of study except when taken as a carry over). SessionID and semesterID are 2016/2017 and 1ST respectively. The semester course is a core course hence it has the code "C" for the status. A core course in this context is a course that is compulsory for all the students doing the same course of study or programme. Elective courses on the other hand are not compulsory. They are offered by students as a matter of choice.

The experiments conducted involved sargable conjunctive equality predicates and it involved varying the number of atomic predicates from two to five. Five atomic predicates are realistic enough to test the behavior of a complex predicate [41]. For each sargable conjunctive equality predicate, the number of schema instances was varied from 600,000 to 1,000,000. A data set of 1,000,000 records was assumed to be asymptotic (big data) and sufficient based on the use of the same number of records in a similar database experiment [41]. For a sargable conjunctive equality predicate to select an instance, the atomic conditions in the conjunct must all be true for the instance. For this reason, it is common in experiments testing conjunctive equality predicates to have record instances with repeated values [48]. The predicate attributes in the experimental data are grouped in terms of the number of predicates in the sargable conjunctive equality predicate and presented in Table 2.

Table 2: Predicates attributes used in the Sargable conjunctive equality predicates

Number of predicate attributes	Predicate attributes	Sargable conjunctive equality predicates
2	courseCode, semesterID	courseCode="PHY412" and semesterID='1st'
3	level, coursecode, semesterID	Level ="400" and courseCode="PHY412" and semesterID='1st'
4	courseCode,semesterID, sessionID, status	courseCode="PHY" and semesterID='1st' and sessionID = "2015/2016" and status = "C"
5	level,courseCode, semesterID, sessionID, status	Level ="400" and courseCode="PHY" and semesterID='1st' and sessionID = "2015/2016" and status = "C"

The predicate attributes listed in Table 2 are classified according to the number of their atomic predicate attributes beginning from 2 to 5 with their corresponding sargable conjunctive equality predicates. The combination of the predicate attributes of each sargable conjunctive equality predicate in any order or pattern is commutative and equivalent. They retrieve the same number of records and hence the order does not matter. The query times obtained from the experiment performed are shown in Table 3.

Table 3: Query times for sargable conjunctive equality predicates

Number of records (n)	Query time in microseconds according to number of predicates			
	2 predicates	3 predicates	4 predicates	5 predicates
600,000	7.653684139	8.102646112	8.527706862	8.91692996
800,000	9.84117198	10.45632792	10.88422108	11.36055684
1,000,000	12.14951396	12.91807389	13.49754906	14.54573202

The data obtained from the experiment exposed a pattern whereby the query times associated with sargable conjunctive equality predicates increase as the number of atomic predicates are varied from two through five which implies that when the atomic predicates are reduced, the query time is equally reduced. It is obvious from this observation that minimizing the number of atomic predicates of the sargable conjunctive equality predicate enhances query time. Figure 2 shows the query times of the sargable conjunctive equality predicates as a graph.

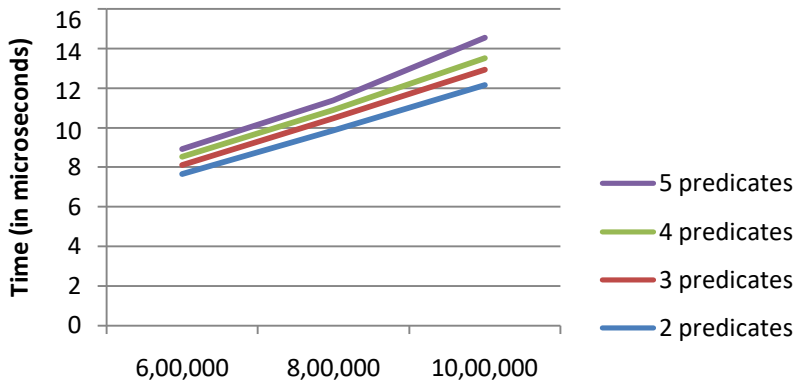


Figure 2: Query times of sargable conjunctive equality predicates

In the DBMS architecture shown in Figure 1, a query is a component of the DBMS that works in conjunction with the query optimiser to ensure queries run optimally. The query times obtained in Figure 2 includes every internal optimisation done by optimiser as well as any restructuring that can be done to the database such as indexing, partitioning, normalization and the introduction of primary keys. This scenario was earlier modeled in equation 1 as:

$$q_t = \frac{T_6}{T_R}$$

The point in this paper is that the query can be optimised even before it is submitted to the optimiser. A

typical case in point is the use of subqueries (nested queries also referred to as queries in predicates) in place of joins in scenarios where subqueries and joins are equivalent queries. An example is where the join retrieves a single tuple, then it is less costly to use subqueries than joins. The equivalent query that optimises query time introduces an optimisation factor, q_{opt} to equation 1 to produce equation 2.

$$q_t = \frac{T_6}{T_R} . q_{opt} \quad (2)$$

Where q_{opt} lies in the range $0 < q_{opt} < 1$. $q_{opt} = 1$ means there was no optimisation by the optimisation technique applied. The concept of optimizing sargable conjunctive predicates is rooted in theory of equivalent queries. For any sargable conjunctive predicates, there exists a corresponding sargable concatenated predicate.

On the basis of this insight from the experimental results in Figure 2, the concatenated Predicate model is formulated as consisting of a concatenated predicate and a corresponding surrogate index that is exploited by the concatenated predicate to enhance the query time of an equivalent sargable conjunctive equality predicate. The experiments are restricted to single table access and by implication, sargable conjunctive equality predicates [17]. Based on the experimental results, the methodology of this study consists of equations describing the product of terms (atomic predicates) which were subsequently used to formulate the concatenated predicate model. Theorems describing the application and optimality of the model as capable of optimising sargable conjunctive equality predicates are derived and also proved. It is hoped that the clarity of the concepts using the single table access will help in extrapolating the model to the other types of table access.

III. Methodology

The basic materials for this research were published literatures. Chaudhuri et al. [24] used the factorization technique to optimize a class of predicates that have common Boolean factors. The class of sargable conjunctive equality predicates on the other hand do not have common Boolean factors which makes factorization inapplicable to them. Motivated by this insight, this study unravelled some properties of the sargable conjunctive equality predicate which gave an insight on how this class of predicates can be optimised.

a) Mathematical Model

In describing the proposed model, mathematical models have been used extensively. Muralikrishna and DeWitt [47] referred to the product of the atomic predicates, P_i, i =1,2, .., m in a join or selection clause as,

$$\prod_{i=1}^{m} P_i, \quad m > 0$$

Each of the atomic predicates is referred to as a term. Assuming each atomic predicate, P_i to be of the form, $a_i = v_i$ and a_i denotes an attribute name of relation R and v_i is a value, then the predicate defined is an equality predicate. The product of terms operation is also referred to as the concatenation of the terms [49]. Since P_i in $\prod_{i=1}^{m} P_i$ is of the form, $a_i = v_i$, then $\prod_{i=1}^{m} P_i$ can be decomposed to become,

$$\prod_{i=1}^{m} a_i = \prod_{i=1}^{m} v_i$$

$\prod_{i=1}^{m} a_i$ is the product of attribute names which for ease of reference can be assigned a variable name, say C to get,

$$C = \prod_{i=1}^{m} v_i \quad \ldots \quad (3)$$

where $\prod_{i=1}^{m} v_i = v_1 . v_2, \ldots, v_m = v_1 v_2 \ldots v_m$ and equation (3), defining an atomic predicate can be referred to as the concatenated predicate.

b) The Concatenated Predicate Model

Concatenation amounts to finding the product of terms, the result of which is a single term. Given the equality predicates of a sargable conjunctive equality predicate as terms, concatenation can be used to find the product of the equality predicates which results in a single atomic predicate. Put differently, concatenation reduces (minimises) the number of terms (atomic predicates) in a sargable conjunctive equality predicate to one irrespective of the number of terms [49,11]. Concatenation in mathematics is the joining of two numbers by their numerals in contrast to arithmetic operations on numbers. Arithmetic operations such as addition, multiplication and all the others are based not only on the numerals but also on the magnitude of the numerals involved. Generalising, concatenation is an operation on the literals of an expression. If the term is a number, the literals are the numerals; the literals are alphabets or alphanumeric if the term is alphabetic or alphanumeric respectively.

Deen [50] exposed concatenation to be a very useful operation in computer programming and used it to generate surrogate keys as the product of an internal relation number (*irn*) and an effective key value (*ekey* value). The surrogate key generated is given by *surrogate* ::= <*irn*> <*ekey* value>. In Oracle noSQL, the concatenation of a *Major Key Path* and a *Minor Key Path* was used to generate record keys [51]. All records sharing a Major Key Path are co-located to achieve data locality. Within a co-located collection of Major Key Paths, the full key, comprising of both the Major and

Minor Key Paths, provides fast indexed lookups. Concatenation has also been applied in the theory of languages [52].

To use concatenation as a product of atomic predicates in a sargable predicate, the following conditions must be met:

1. The values of the atomic predicate attributes of the predicate must be exact and this can only be guaranteed by the equality relational operator
2. The predicate terms must not be less than two and each of them must be an atomic predicate in the predicate to be concatenated. This condition can only be guaranteed when the AND logical operator is used to join the atomic predicates

The second condition is a necessary condition because different values defined for the same atomic predicate attribute in a conjunctive equality predicate is unsatisfiable and evaluating them would lead to incorrect results [22]. In practical terms we cannot have A=12 and A=10 as a valid atomic predicates in a conjunctive equality predicate. The predicate attribute, A in a conjunctive equality predicate cannot have different values at the same time. Sargable conjunctive equality predicates meet the two conditions specified above hence concatenation is applicable to them as an optimisation technique. For every sargable conjunctive equality predicate, an equivalent concatenated predicate is derivable by concatenating the atomic predicates of the sargable conjunctive equality predicate.

The transformation of the sargable conjunctive equality predicate to the concatenated predicate can be shown diagrammatically using a logical plan tree, the height of which depends on the number of atomic predicate operations involved in the predicate. Considering a sargable conjunctive equality predicate having three atomic predicates, σ_1, σ_2 and σ_3 defined on relation, *R* for example, the plan tree will have three predicate operations as shown in Figure 3a. The equivalent concatenated predicate, say *C* is a product of the atomic predicates and hence it has a single predicate operation, σ defined on relation *R* as shown in Figure 3b.

Sargable conjunctive equality predicate Concatenated predicate

Figure 3: Equivalent Predicate Logical Plan Trees

In general, each atomic predicate in a sargable conjunctive equality predicate corresponds to a predicate operator (σ), on the logical plan tree. Each additional operator increases the height of 3a by 1 meanwhile the height of 3b remains constant. It is obvious from the logical plan trees that irrespective of the number of atomic predicates, the concatenated predicate has one atomic predicate and is assumed to be the transformation of an equivalent sargable conjunctive equality predicate.

The concatenated predicate is derived from the atomic predicate attributes of the sargable conjunctive equality predicate meaning that the atomic predicate attributes must be natural attributes of the relation queried by the sargable conjunctive equality predicate. The atomic predicate attributes are said to be sargable because they are used to search the relation. In a similar fashion, the concatenated predicate, being a product has a single attribute which it equally uses to search the relation. This also means that the concatenated predicate is also a sargable predicate. Sargable predicates search relations based on the value set of the attribute involved in the predicate. Incidentally, the attribute involved in the concatenated predicate is not a natural attribute in the relation and it has to be constructed as an artificial or surrogate attribute, call it *S*. The value sets of *S* are arrived at by concatenating the value sets of each of the natural attributes in the sargable conjunctive equality predicate as follows:

$$v(S) = \prod_{i=1}^{m} v_i{}^{pa} = v_1^{pa}.v_2^{pa},\ldots,v_m^{pa} = v_1^{pa}v_2^{pa}\ldots v_m^{pa} \quad \ldots \quad (4)$$

where *pa* is a predicate attribute of a sargable conjunctive equality predicate and v_i, i>0 is the value set with the natural fields involved in the sargable conjunctive equality predicate. It follows from this definition that, *S* is the artificial attribute whose value sets is the concatenation of all v_i for each value set of *m* atomic predicates in the sargable conjunctive equality predicate. This makes *S* one of the attributes defined for the query relation, *t* relative to which *t*(*S*) can be defines at the tuples of *S* in relation, *t* shown in equation (5).

$$t(S) = \prod_{i=1}^{m} v_i{}^{pa} \quad \ldots \quad (5)$$

making q_{opt} <1 in $q_t = \frac{T_6}{T_R}.q_{opt}$

The artificial attribute and its tuples referred to in equation 5 is a surrogate attribute and works very much like a user-defined index or surrogate value [30,53]. Surrogate indexes are very useful in database query optimisation [49,54,55]. The original concept of a surrogate value was to provide a unique identifier for each tuple (a kind of system primary key) that does not change irrespective of what the user chooses to do with the primary key value or the value of any of the other fields in terms of modifying them. These were called permanent surrogates. Deen [50] implemented the inpure type of surrogates in which the surrogate key changes if any of the values concatenated to generate the surrogate changes. The inpure surrogates were generated from the primary key using a hashing and a key compression algorithm, supported by an overflow mechanism. To effectively achieve this, surrogates are maintained using the following operations. When a tuple is inserted, a surrogate must be generated and the surrogate directory updated. This operation is referred to as surrogate generation. When a tuple is deleted, the surrogate directory must be updated, releasing the surrogate for possible re-use. This operation is referred to as surrogate release. Given the value of the attributes that make up the surrogate key value, the system should be able to find the surrogate. This operation is referred to as surrogate access. Given a surrogate, it should be possible to find the stored tuple. This operation is referred to as storage access and in this role, the surrogate serves the purpose of a data structure that can be exploited by predicates to locate records.

Diagrammatically, when the surrogate index is exploited by a concatenated predicate, tuples of the associated relation that match the predicate condition are fetched. The tuples defined in Equation (5) that are fetched by the concatenated predicate are depicted in Figure 4.

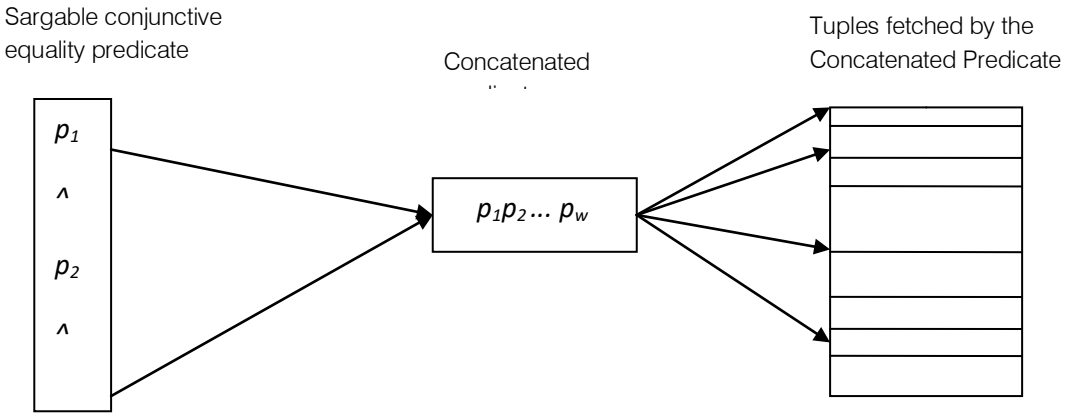


Figure 4: Tuples Returned by the Concatenated Predicate

Figure 4 is a diagrammatic representation of how the concatenated predicate is evaluated. Let θ be the concatenated predicate on the query table, *t,* then equation (6) models the evaluation of θ. A tuple, is returned from the query table by θ when the concatenated predicate is evaluated and the result is true.

$$\theta_i(t) = \begin{cases} true, & if\ t(S_i) = C \\ false, & otherwise \end{cases} \quad \ldots \quad (6)$$

IV. Results and Discussion

Resulting from the experiments performed in order to gain an understanding of the nature of optimisation problem posed by sargable conjunctive equality predicates, the Concatenated Predicate model was formulated as consisting of a concatenated predicate and a corresponding surrogate index that is exploited by the concatenated predicate to enhance the query time of an equivalent sargable conjunctive equality predicate. This result is proved using formal

methods for their correctness. The correctness of this result is discussed as a theoretical proof of the concatenated predicate model.

a) Proof of Existence of the Concatenated Predicate

Lemma 1: The equality condition in a sargable conjunctive equality predicate guarantees uniqueness of its atomic predicates because a conjunct of two different filters on the same attribute is unsatisfiable [22].

Theorem 1: For every sargable conjunctive equality predicate, there exists a product of its atomic conditions called the concatenated predicate

Proof: Theorem 1 follows from the work of [47] and [22]. Muralikrishna and DeWitt [47] and Heimel et al. [22] defined a conjunctive predicate as, P_1 AND P_2 AND... AND P_m and as being equivalent to a product of its atomic predicate terms expressed as $\prod_{i=1}^{m} P_i$, where each P_i in both the product term and the conjunctive term is strictly a Boolean expression and each P_i in the sargable conjunctive equality predicate is an equality predicate.

b) Proof of Equivalence

Lemma 2: Two query predicates (conditions) are equivalent if they return *the same records from a compatible database* [11].

Theorem 2: The product of atomic predicates (concatenation operation) is a bijective function on the set of concatenated predicates from the set of sargable conjunctive equality predicates thereby defining their equivalence

Proof: Let $a \in A$, where A is the set of sargable conjunctive equality predicates and $b \in B$, where B is the set of the concatenated predicates and *D* is a compatible database containing the surrogate field derived from the concatenation of the natural fields of D in the sargable conjunctive equality predicates. *Let f* represent the operation that concatenates the atomic conditions in *a* to get *b*. Then $a \equiv b$ if and only if, *f* is a bijective function. To be bijective, *f* must be onto as well as one-to-one:
f is onto because each concatenated predicate, *b* in B is in the image of *f* . That is,

$$\forall\, b \in B, \exists\, a \in A \text{ and } f(a) = b \quad \ldots$$

f is one-to-one because for a concatenated predicate, b ∈ B there is at most one *a* ∈ A such that *f*(a)= b . That is,

$$\forall\, a, a` \in A \text{ and } f(a) = f(a`) \text{ implies } a = a`$$

where *a`* is the inverse of *a* going by the concatenation operation implied by *f.*
Given that (3.10) and (3.11) holds, we conclude that
f: A → B is a bijective function on the set of concatenated equality predicates to the set of conjunctive predicates because *f* is both one-to-one and onto.
⇒ A ⇔ B and hence they return the same number of records for a compatible database

c) Proof of Optimisation

The generic optimisation model of a relational database query is described in terms of a relational algebra expression. The relational algebra corresponding to a query describes a set of operators whose number can be determined and their cost estimated. Based on either number of operators or estimated cost, two queries can be compared to ascertain that one optimizes the other. A relational algebra expression *e`* optimises another relational algebra, *e* if the following conditions are satisfied (1) *e`* is equivalent to *e* given a compatible database (2) the query time of *e`* is less than that of *e*. *e`* is optimal if a relational algebra expression that optimises *e`* does not exists.

Theorem 3: A predicate, *C`* is optimal relative to the conjunctive equality predicate, *C,* if (1) *C`* *is* equivalent to the conjunctive equality predicate, *C (2) C`* has fewer occurrences of equality predicates than *C,* and (3) there exists no other predicate, *p* that is equivalent to *C`* and has fewer occurrences of equality predicate than *C`*.

Proof:

The proof consists of a lemma and a proof by induction on $n(\sigma)$, the number of equality atomic conditions in the predicate, σ. The lemma establishes the equivalence of *C* to *C`*, while the poof by induction establishes the optimality of *C`* compared to *C* .

Lemma: Two query predicates (conditions) are said to be equivalent if they each *return the same records from a compatible database* [11].

Induction hypothesis: Consider the query execution tree in Figure 3 and let $n(\sigma)$= number of atomic predicates. Each atomic predicate in *C* corresponds to a predicate operator (σ_c=σ_1, σ_2, .. ., σ_m). In all circumstances, $n(\sigma)$=1 for *C`* since *C`* is a product of the terms of *C*. Being a product, the number of terms in *C`* is $m = 1$ and so $n(\sigma) = 1$ for *C`*.

Initial Induction Step: The height of the tree corresponding to *C* = number of atomic predicates in *C* = $n(\sigma_c) = m$, where *m* is the number of atomic predicates in *C* . Assuming $n(\sigma_c) = m = 5$ as the initial induction step, *m* is defined in subsequent induction steps as *m-i*, where *i* is the subsequent induction step, defined as 1,2, .., *m.* That is, in each subsequent induction step, we decrease the number of atomic predicates, *m,* by one at a time.

Subsequent Induction Steps:
When i=1, then m-i=5-1=4 implying $n(\sigma) = 4$
When i=2, then m-i=5-2=3 implying $n(\sigma) = 3$
When i=4, then m-i=5-4=1 implying $n(\sigma) = 1$

When i=5, then m-i=5-5=0 implying $n(\sigma) = 0$
When m approaches 0, $n(\sigma)$ approaches 0

This means that the number of operators, $n(\sigma)$ decreases in proportion to the number of atomic predicates, m. Mathematically,

$$n(\sigma) = m,\ m = 1, 2, .. , \infty$$

The query does not work and is undefined for when $m = 0$, $n(\sigma) = 0$. The query processes the least number of selection operators and hence does the least amount of work when $m = 1$, $n(\sigma) = 1$. Recall that $n(\sigma) = 1$ for predicate $C`$ . This means that $C`$ is optimal since any other reduction of the atomic predicates in C will result in a predicate that has $n(\sigma)<1$ and by the lemma, the equivalence of C since $C`$ is proved`

The proof assumes that the cost of execution of a predicate is directly proportional to the number of atomic predicates that makes it up. This was proved by previous experiments.

One of two equivalent predicates optimizes the other if the query time associated with the optimising predicate is lesser and both predicates are equivalent given a compatible database.

Proof: The relational algebra of the concatenate predicate and the sargable conjunctive equality predicate are made up of the same operator, the selection operation. The proof that the concatenate predicate, $C = \prod_{i=1}^{m} P_i$ has fewer occurrences of operators than *CP* the conjunctive predicate and hence optimizes the sargable conjunctive equality predicate is as follows

Let the compatible database be R and the relational algebra corresponding to the sargable conjunctive equality predicate be e and the algebra of the concatenated predicate be $e`$. Then $e= \sigma_{P1 \text{ and } P2, \text{ and...} \text{ and } Pm}(R)$ and $e`= \sigma_p(R)$ where the number of atomic predicates in e, $|e| = m$, $m>1$. Given that $e`$ is a product, it follows that $|e`| = 1$. Given $|e| = m$, $m>1$ and $|e`| = 1$, we can assume the minimum value of m=2 for $e$ resulting in $e= \sigma_{P1 \text{ AND } P2}(R)$. The relational algebra of $e$ and $e`$ consist of the selection operation, σ whose implementation uses either the sequential search (table scan) or the binary search (indexed scan). Since both e and $e`$ are made up of the same operation, it is convenient to assume that table scan has been used to implement them in the following algorithmic procedure.

The algorithmic steps corresponding to $e = \sigma_{P1 \text{ and } P2}(R)$ are:

1. Apply the selection operator σ_{P1} to get the intermediate relation, I_1
2. Apply the selection operator σ_{P2} to get the intermediate relation, I_2, the final result Analysis:

Assume the arity of the intermediate results, I_1 and I_2 to be approximately of the uniform value, n respectively. Let the total query time of e be $f_e(n)$. then $f_e(n)$ = query time of I_1 + query time of $I_2 = n+n = 2n$
The algorithmic steps corresponding to $e` = \sigma_p(R)$ are:

1. Apply the selection operator σ_p to get the intermediate relation, I the final result

Analysis

Assume the arity of the intermediate results, I_1 and I_2 to be approximately of the uniform value, n respectively. Let the total query time of $e`$ be $f_{e`}(n)$. then $f_{e`}(n)$ = query time of $I = n$
Clearly, $f_e(n) > f_{e`}(n)$, meaning that the query time of $e`$ is less than that of e implying that $e`$ optimises e.
The proof of optimality follows.

d) Proof of Optimality

Theorem 4: Given two equivalent relational algebras where one optimises the other, the one that optimises is optimal if a relational algebra expression that optimises it does not exist

Proof: The proof that the concatenated predicate, C that optimizes the conjunctive equality predicate, *CP* is optimal is proved by induction on m, the number of predicates in both C and *CP*.

Induction hypothesis: Each atomic predicate in *CP* corresponds to a predicate operator (σ) on the logical plan tree. Assuming m to be the number of atomic predicates in *CP* and $n(\sigma)$ be the number of predicate operators on the corresponding logical plan tree, then for every additional atomic predicate, $n(\sigma)$ increases by 1 such that $m = n(\sigma)$.

Induction Step: Assuming *CP* has a single atomic predicate, then $m = 1$ and $n(\sigma) = 1$. Proceeding with the induction steps, we increase the number of atomic predicates, m, by one at a time to get,
When $m = 2$, $n(\sigma) = 2$
When $m = 2$, $n(\sigma) = 3$
...When m approaches ∞, $n(\sigma)$ approaches ∞

This means that the number of operators, $n(\sigma)$ grows in proportion to the number of atomic predicates, m. Mathematically,
$n(\sigma) = m,\ m = 1, 2, .. , \infty$
But C is a product, implying that the number of terms in C is m = 1 and so $n(\sigma) = 1$ for every C corresponding to CP

For there to exist a predicate that optimises C, the occurrence of operators in such a predicate, m, must be zero, that is $m < 1$. If $m = 0$, then $n(\sigma)$ will also be zero. $n(\sigma) = 0$ defines a predicate that has no atomic predicate which is non-existent and hence a predicate that optimises C is non-existent. This means C having one operator has the least number of operators and hence it is optimal.

In this section, the proof of correctness of the concatenated predicate model has demonstrated. Table

3.4 shows the equations used to model the various components of the model.

Table 4: Summary of Model Equations

Model Component	Equation
Conjunctive Predicate	$\bigwedge_{i=1}^{m} \theta_i = P_1 \; AND \; P_2 \; AND \; \ldots AND \; P_m$
Conjunctive Predicate as a product of terms	$\prod_{i=1}^{m} P_i$
Concatenate Predicate	$C = \prod_{i=1}^{m} v_i$
Surrogate Index	$t(S) = \prod_{i=1}^{m} vs_i{}^{pa}$
Model Evaluation	$\theta_i(t) = \begin{cases} true, & if \; t(S_i) = C \\ false, & otherwise \end{cases}$

V. Conclusion and Suggestion for Further Work

The optimization of queries where complexity is due to a large number of joins has received a lot of attention in the database literature, but the optimization of complex selection predicates involving multiple ANDs (conjunctive predicates) and ORs (disjunctive predicates) has not been widely addressed [24]. In lieu of the dearth of research into selection predicates, the contribution to knowledge of this research effort can be said to be significant. Enhancing the query times of sargable conjunctive equality predicates is significant in the following ways:

1. The optimisation of the sargable conjunctive equality predicates within the context of big data minimises query time which tend to increase with big data
2. Sargable conjunctive equality predicates are widely used in applications involving data extraction, mining, matching and resolving data entities [56]. Enhancing these predicates directly improves the running times of applications designed to automate these operations.
3. Considering the very many other areas in which an improved query time can be of use, the research is of significance to software architects, software developers, the software industry and researchers.

The concatenated predicate model works very much like an index hence we can refer to it as a surrogate index. In our subsequent work, the concatenated predicate model will be experimentally validated and work on how to integrate the surrogate index into existing DBMS architecture studied.

References Références Referencias

1. Storey VC, Song I. Big data technologies and management: What conceptual modeling can do? Data & Knowledge Engineering. 2017; 108:50–67
2. Obilikwu, P., Ogbuju, E. (2020) A data model for enhanced data comparability across multiple organizations. *J Big Data* **7,** 95 (2020). https://doi.org/10.1186/s40537-020-00370-1
3. Khan MA, Uddin MF, Guptam N. Seven V's of Big Data: Understanding Big Data to extract value. Proceedings of 2014 Zone 1 Conference of the American Society for Engineering Education (ASEE Zone 1). 2014.
4. Patgiri R, Ahmed A. Big Data: The V's of the game changer paradigm. International Conference on High-Performance Computing and Communications. 2016.
5. Matalqa H. and Mustafa S.H. (2016): "The Effect of Horizontal Database Table Partitioning on Query Performance", The International Arab Journal of Information Technology, Vol. 13, No. 1A, 2016
6. Das, T.K. & Mohapatro, Arati. (2014). A Study on Big Data Integration with Data Warehouse. International Journal of Computer Trends and Technology. 9. 188-192. 10.14445/22312803/IJCTT-V9P137.
7. Tailor, U., and Patel, P. (2016). A Survey on Comparative Analysis of Horizontal Scaling and Vertical Scaling of Cloud Computing Resources. IJSART - Volume 2 Issue 6, ISSN [ONLINE]: 2395-1052.
8. Stonebraker, M., & Çetintemel, U. (2018). One size fits all: an idea whose time has come and gone. In Proceedings of the International Conference on Data Engineering (ICDE), 2-11. 10.1145/3226595.3226636.
9. Selinger, P. G, Astrahan, M.M, Chamberlin, D.D, Lorie, R.A, Price, T.G (1979): "Access Path Selection in a Relational Database Management System",

SIGMOD Conference 1979, Boston, Massachusetts, May 30 - June 01, pp. 23-34.
10. Obilikwu P.O., Kwaghtyo K.D and Ogbuju E. (2021), Enhancing Query Time Using a Volume-Adaptive Big Data Model OF Relational Databases, *The Journal of Basic Physical Research.* Department of Geological Sciences, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria
11. Chandra, A. K. and Merlin, P. M.(1977): "Optimal implementation of conjunctive queries in Relational Databases", Proceedings of the 9th ACM Symposium of Theory of Computing, Boulder, Colorado, USA, May 4th, pp. 77-90.
12. Levy, A.Y., Mumick, I. S. and Sagiv Y. (1994): "Query Optimization by Predicate Move-Around", Proceedings of the 20th VLDB Conference, Santiago, Chile, September 12-15, pp. 96-107.
13. Abiteboul, S., Hull R. and Vianu V.(1995): "Foundations of Databases", Addison_Wesley, Reading, MA. pp. 35 – 65.
14. Gottlob, G., Lee, S.T. and Valiant, G. (2012): "Size and Tree width Bounds for Conjunctive Queries", Journal of the ACM, Volume 59 Issue 3, Article No. 16
15. Swami, A. and Scheifer, K.B. (1993): "On Estimation of Join Result Sizes", Technical Report, IBM Research division, IBM Research Report RJ9569
16. Grohe, M., Schwentick, T. and Segoufin, L. (2001): "When is the evaluation of Conjunctive Queries Tractable", Proceeding of the 33rd Annual ACM symposium on Theory of Computing, Hersonissos, Greece, July 6-8, pp. 657 – 666
17. Mohan, C., Haderle, D. J., Wang, Y., and Cheng, J. M. (1990): "Single Table Access using Multiple Indexes: Optimization, execution, and concurrency control techniques", International Conference on Extending Database Technology (EDBT), Venice Italy, March 26-30, Volume 416 of LNCS, pp. 29–43.
18. Elmasri, R. and Navathe, S. B. (2011): "Fundamentals of Database Systems", 6th Edition, Pearson Education Inc, pp. 679 - 723
19. Garg, V.K. and Waldecker, B. (1994):"Detection of weak unstable predicates in distributed programs", IEEE Transactions on Parallel and Distributed Systems, Volume: 5, Issue: 3, Pp: 299 – 307
20. Mugnier M., Rousset M. and Ulliana F. (2016): "Ontology-Mediated Queries for NOSQL Databases", Association for the Advancement of Artificial Intelligence
21. Munir, K. and Anjum, M.S. (2017): "The use of ontologies for effective knowledge modelling and information retrieval", Applied Computing and Informatics (2017),http://dx.doi.org/10.1016/j.aci. 2017.07.003
22. Heimel, M., Markl, V. and Murthy, K. (2009): "A Bayesian Approach to Estimating the selectivity of Conjunctive Predicates", Proceedings of Datenbanken und Informationssysteme (DBIS), Münster, Germany, March 2-6, pp 47-56
23. Yu, X., Koudas, N., and Zuzarte, C. (2006): "HASE: A Hybrid Approach to Selectivity Estimation for Conjunctive Predicates", Advances in Database Technology - EDBT 2006, Springer International Publishing, AG, Volume 3896 of the series Lecture Notes in Computer Science pp. 460-477.
24. Chaudhuri, S., Ganesan, P. and Sarawagi, S. (2003): "Factorizing Complex Predicates in Queries to Exploit Indexes", ACM SIGMOD 2003, June 9-12, San Diego, CA. pp. 361-372
25. Kemper, A., Moerkotte, G., Peithner, K., and Steinbrunn, M. (1994): "Optimizing disjunctive queries with expensive predicates", ACM Intl. Conference on Management of Data (SIGMOD), Minneapolis, Minnesota, May 24-27, pp. 336–347.
26. Lohman, G. (2014): "Is Query Optimization a 'Solved' Problem?", ACM Special Interest Group on Management of Data blog, http://wp.sigmod.org/
27. Chaudhuri, S. (2012): "What next?: a half-dozen data management research goals for big data and the cloud", Proceeding PODS '12 Proceedings of the 31st ACM symposium on Principles of Database Systems, Scottsdale, Arizona, USA — May 21 - 23, pp. 1-4
28. Codd, E. F. (1970): "A Relational Model of Data for Large Shared Data Banks". Communications ACM 13(6): 377-387
29. Chamberlin, D.D., Astrahan, M.M., Blasgen, M. W., Gray, J. N., King, W. F., Lindsay B. G., Lorie, R., Mehl, J. W., Price, T, G., Putzolu, F., Selinger, P. G., Schkolnick, M., Slutz, D.R., Traiger, I. L., Wade, B. W., Yostet, R. A. (1981): "A History and Evaluation of System R", Communications of ACM 24(10): Pp. 632-646
30. Clough, L., Haseman, W.D. and So, Y.H.(1976): "Designing Optimal Data Structures", AFIPS national computer conference and exposition, New York, New York — June 07 - 10, pp. 829-837.
31. Codd, E. F. (1975): "Implementation of Relational Database Systems", Panel Discussion, NCC (AFIPS) 75, Anaheim.
32. Chaudhuri, S. (1998): "An Overview of Query Optimization in Relational Systems", Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems
33. Ioannidis, Y. (2003): "The History of Histograms", Proceedings of the 29th VLDB Conference, Berlin, Germany, September 9-12, pp. 19-30.
34. Cao, B. and Badia, A. (2005): "A Nested Relational Approach to Processing SQL Subqueries", SIGMOD 2005 June 14 - 16, 2005, Baltimore, Maryland, USA, pp. 191 – 202

35. Vellev, S. (2009): "Review of Algorithms for the Join Ordering Problem in Database Query Optimisation", Information Technologies and Control, pp. 32 – 40
36. Bamnote, G. R. and Agrawal, S.S. (2013): "Introduction to Query Processing and Optimization ", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, Pp. 53 – 56
37. Ullman, J. (1989): "Principles of Database and Knowledge Base System", Volume 1, Computer Science Press Inc., New York, p. 631.
38. Chimenti, D., Gamboa, R. and Krishnamurthy R. (1989): "Towards an open architecture for LDL", Proceedings of the fifth International VLDB Conference, Amsterdam, the Netherlands, August 22-25, pp 195-203
39. Chaudhuri, S. and Shim, K. (1993): "Query optimization in the presence of foreign functions", Proceeding of the 19th International VLDB Conference, Dublin, Ireland, August 24 – 27, pp. 529 – 542.
40. Hellerstein, J. M. and Stonebraker, M. (1993): "Predicate Migration: Optimising queries with Expensive Predicates", SIGMOD Conference, Washington DC, May 25-28, pp. 267–276
41. Chaudhuri, S. and Gravano, L (1996): "Optimizing queries over multimedia repositories", ACM International Conference on Management of Data (SIGMOD), Montreal, Quebec, Canada, June 4-6, pp 91–102.
42. Chaudhuri, S. and Shim, K. (1999): "Optimization of queries with user-defined predicates", ACM Transactions on Database Systems (TODS), 24(2), June 1-3, Seattle, Washington, USA, pp. 177–228.
43. Kemper, A., Moerkotte, G., and Steinbrunn, M.(1992): "Optimizing Boolean expressions in object Bases", Proceedings of the VLDB Conference, Vancouver, Canada, August 23-27, pp 79-90
44. Balasubramanian, P. and Arisaka R. (2007): "A Set Theory Based Factoring Technique and Its Use for Low Power Logic Design", World Academy of Science, Engineering and Technology, 3, pp. 446 – 456
45. Brayton, R.K., Rudell R. and Sangiovanni-Vincentelli, A. and Wang, A. (1987): "MIS: A multiple-level logic optimization system", IEEE Transactions. on CAD of Integrated Circuits and Systems, Vol 6, Issue 6, pp. 1062-1081.
46. Reinwald, L.T. and Soland, R.M. (1966): "Conversion of Limited-Entry Decision Tables to Optimal Computer Programs: Minimum Average Processing Time", JACM, 13(3), Pp 339-358
47. Muralikrishna, M. and DeWitt, D. J. (1988): "Optimization of multiple-relation multiple-disjunct queries, In Proceedings of the Seventh ACMSIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Austin, Texas, March 21 – 23, pp 263-275.
48. Hellerstein, J. M. (1994): "Practical Predicate Placement", Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data, Minneapolis, Minnesota, May 24-27, pp. 325-335
49. Deen, S. M. (1982): "An implementation of impure surrogates", International Conference on Very Large Databases, Mexico City, September 8-10, pp. 245-256.
50. Deen, S. M., Amin, R. R. and Taylor, M. C. (1994): "A Strategy for Decomposing Complex Queries in a Heterogeneous DDB", Proceedings of the Tenth International Conference on Very Large Databases, Singapore, August 27-31, pp. 397-400.
51. Oracle (2017): "Oracle Sharding Linear Scalability, Fault Isolation and Geo-distribution for Web-scale OLTP Applications", ORACLE White Paper, April 2017.
52. Sander-Bruggink, H.J., Konig, B. and Kupper S. (2013): "Concatenation and other Closure Properties of Recognizable Languages in Adhesive Categories", Proceedings of the 12th International Workshop on Graph Transformation and Visual Modeling Techniques, Mar 23 – Mar 24 2012, Rome, Italy
53. Lynch, C. and Stonebraker M. (1988): "Extended User-Defined Indexing with Application to Textual Databases", Proc. 14th International Conference on Very Large Datsbases, Los Angeles, August 29 – September 1, pp. 306 – 317
54. Harkins, S. (2011): "10 Tips for Choosing between a Surrogate and Natural Primary Key". Retrieved from www.techrepublic.com, pp 1-2
55. Valduriez, P. (1987): "Join Indices", ACM Transactions on Database Systems, Vol. 12, No. 2, June 1987, pp. 218-246.
56. Getoor, L. and Machanavajjhala, A. (2012): "Entity Resolution: Theory, Practice and Open Challenges", Proceedings of the VLDB Endownment, Istanbul, Turkey, August 27-31, Vol 5, No. 12, pp 2018 – 2019.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 22 Issue 1 Version 1.0 Year 2022
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Data Science and Management: A Study of Theoretical Approaches to Computer Systems with Organisation using Advanced Analytics

By Khemraj Dangi
Pacific Academy of Higher Education and Research University

Abstract- A firm's Data Management department is in charge of the corporate data capture, retention, security, management, and safety, as well as the formulation and execution of all data-related regulations inside that company. The Data Management team, on the other hand, merely maintains the data resources; it is underrecognized in the fundamental technological uses of the material. All data is owned by the Data Function of management. The Data Science department in an organisation, on either extreme, conceptualises, develops, executes, and practises all "terms of improving" of information assets. In this context, "technical implementations" refer to the research, technologies, skill, and business practises that use corporate data.

Keywords: *data management, data science, big data, techniques, computer system, organisation.*

GJCST-C Classification: *C.1.0*

Strictly as per the compliance and regulations of:

Data Science and Management: A Study of Theoretical Approaches to Computer Systems with Organisation using Advanced Analytics

Khemraj Dangi

Abstract- A firm's Data Management department is in charge of the corporate data capture, retention, security, management, and safety, as well as the formulation and execution of all data-related regulations inside that company. The Data Management team, on the other hand, merely maintains the data resources; it is underrecognized in the fundamental technological uses of the material. All data is owned by the Data Function of management. The Data Science department in an organisation, on either extreme, conceptualises, develops, executes, and practises all "terms of improving" of information assets. In this context, "technical implementations" refer to the research, technologies, skill, and business practises that use corporate data.

Data Management has recently reached unprecedented heights as a result of a revolution in the corporate perspective of data. Data Science has become an essential aspect of data administration, yet information management and data science are sometimes viewed as distinct tasks. Data scientists invest their efforts working with data professionals, computer scientists, and DBAs to set up the information system for data processing and competition analysis. However, in the expanding next-generation digital marketplace, Data Management combined insights will become the essential variables for commercial success, thus both Data Management and Data Science must collaborate.

Data science entails so much more than just data-mining techniques. Successful data researchers have to be equipped to see business challenges through the lens of data. There is a core framework to data-analytic reasoning, as well as theoretical aspects that must be recognised. Many "traditional" disciplines of study are included in data science. The fundamental concepts of causal analysis must be comprehended. There are also certain areas wherein perception, inventiveness, practical wisdom, and understanding of a specific technology must be applied. A data-science viewpoint offers professionals with organisation and rules that give the data analyst with a structure for taking good care of difficulties of extracting valuable insight from big datas.

Keywords: data management, data science, big data, techniques, computer system, organisation.

Author: Department of PG studies & department of cse Pacific academy of higher education and research university udaipur Raj.
e-mail: dangikhemraj40@gmail.com

I. Introduction

In recent years, various business organisations have seen a significant expansion in the use of Big Data Analytics (BDA). Need for BDA capacity in organisations is acknowledged as an information instrument to enable informed choice, but few research have expressed a grasp of BDA skills in a manner that may expand the practical knowledge of employing BDA in the organisational domain (Van Rijmenam., et al 2021)[1]. The findings increase the efficacy and adoption of BDA apps in diverse organisations[2]. The undertaken research paper is a study based on Big data management and data science. The primary aim of the paper is to conduct a detailed analysis upon the evaluation of Data Science and Management. The paper will critically discuss Theoretical Approaches to Computer Systems with Organisation Using Advanced Analytics.

II. Background

Organisations nowadays continually gather user data [e.g., data collecting] in order to enhance company efficiency and processes. Significant amounts of recorded datasets pertaining to online transactions are utilised to aid strategic planning, with administrators, regulators, and top executives increasingly frequently adopting innovative ways to turn this avalanche of original data into valuable, helpful information (Dubey., et al 2020). Data analysis is difficult, though one data-handling approach, "Big Data Analytics" (BDA), is extensively used. BDA is the use of sophisticated algorithms, such as data mining, statistics, and forecasting, on massive data as a new company intel activity. BDA transforms data into information which may be utilised to help decision-making using computational approaches. Big Data Analytics (BDA) is quickly becoming a popular method that many firms use to generate crucial results from BD. Businesses see the processes, including adoption and usage of BDA technologies, as a way to support business performance, despite its strategic potential to grow value for stakeholders and achieve a competitiveness over rival businesses.

III. Literature Reviews

According to Sivarajah (2020), BD practises and the use of BDA methodologies as given in a prescriptive chunk of literature explains that in its raw state, BD consisting of a huge raw data collection does not provide much value[3]. BD analytical approaches may be considered as a subsystem inside the larger method for extracting insights from BD[4]. Administrative problems associated with BD are a collection of issues experienced, for example, in obtaining, storing, and regulating data.

There are several types of BDA available to satisfy the particular decision-support needs of various businesses. Analytical methodologies are used by retail firms to obtain a competitive edge and organisation performance[5]. Contemporary corporations are constantly investing in BDA initiatives to save costs, make more accurate decisions, and plan for the future. Amazon, for instance, was the first online store and has retained its revolutionary BDA development and use[6].

Effective procedures are necessary to undertake activities like ongoing diagnosis, strategic planning, and the execution and assessment of BDA to aid organisation decision-making for development (Anshari., et al 2020). According to Organisational Development (OD) theory, processes have the purpose of transferring knowledge and expertise to an organisation, with the method primarily aimed at improving problem-solving capability and managing possible change. OD is defined as a company's inner dynamics, which include a team working together to increase organisational effectiveness, capacity, ability to do the job, and the ability to control culture, policies, practises, and procedural needs.

IV. Research Gap

The report critically assessed the gaps discovered in previous investigations. To resolve such gaps, the research has quickly illustrated the need of doing a systematic evaluation of Big Streaming data research employing robust and systematic methodologies to detect trends in Big Metadata instruments within various organisations within the computer system by analyses of techniques, innovations, and methods.

V. Research Question

1. How is the effectiveness of BDA procedures such as continual evaluation, objective setting, and organisational decision-making executed?
2. How Big Data Analytics to fulfil the distinct decision-support needs of various companies?
3. How is the body of research on Big Data analytics, its potential, and how businesses might use them?

a) Importance of the Study

The presented research paper is of utmost importance because it has briefly discussed Big data management and data science and the Theoretical Approaches to Computer Systems With Organisation Using Advanced Analytics. Findings highlight both strategic and practical implications related to decision making in organisations for top management, particularly in developing countries. This study attempts to contribute to the literature through novel findings and recommendations. These fallouts will help the top management during the key decision-making process and encourage practitioners who seek competitive advantage through enhanced organisational performance in SMEs.

VI. Research Objectives

1. To assess the effectiveness of BDA procedures such as continual evaluation, objective setting, and organisational decision-making execution.
2. To analyse Big Data Analytics to fulfil the distinct decision-support needs of various companies.
3. To address the body of research on Big Data analytics, its potential, and how businesses might use them.

VII. Scope and Limitation

The constraints presented in the study point to the application of Big Data processing and data performance in the field of company processes. The study of additional regulators in this situation might be a topic of future studies. Moreover, investigating the function of modifiers, such as quality management, throughout this setting may contribute positively to the research and yield unique insights. However this research demonstrates important insights into two important indicators of performance (i.e The object model of big data and analytics as well as organisational practises) in SMEs by evaluating the proposed structure, it is suggested that further studies be conducted to determine whether the suggested scheme varies in other industries and situations.

VIII. Research Methodology

a) Research Method & Design

Secondary sources, such as journals, books, articles, and web publications, will be used to supplement the research. The paper will critically examine the implementation of BDA at the organisational level using the interpretivism paradigm. The qualitative analysis approach was employed to gain broad access to the data and achieve the final purpose of the research work. This study's data is descriptive in nature, and the research approach is qualitative. Epistemology is the philosophy employed in the research (Ijab., et al 2020). The interpretivism technique

was used as the methodology in this investigation. In data analysis, the explanatory and descriptive techniques are employed to achieve results. All across the comprehensive study, the collected data will be compared to predefined criteria to ensure that the study goal is met. By utilising descriptive forms of research strategy, desk research was conducted to examine every part of the aim framework. This model was developed because it can aid in the establishment of linkages such as readiness and growth amongst dependent factors and their impact on achieving objectives.

b) Research Approach

The research strategy is the method of planning the study design. Because the comparative findings from the literature review investigates the perspectives of numerous datas from different origins here on study's topic, the paradigm for interpretivism is investigated to execute this inquiry. The idea of interpretivism was used to carry out this study because it is critical for secondary data collecting in order to obtain reliable data that'd be beneficial in achieving the research objectives. To get conclusions, data has been analysed utilising interpretivism and descriptive approaches. As a consequence of the thorough review, only relevant data is made available for inclusion in the findings. The data was gathered through Google Scholar, papers, journals, and relevant articles.

c) Analysis of Study

i. *How is the effectiveness of BDA procedures such as continual evaluation, objective setting, and organisational decision-making executed?*

Researchers develop and test a system that evaluates the link between the application of big data analytics & organisational performance (OP) in small and medium enterprises, relying on resource-based theory principles (SMEs). In addition, the mediating function of knowledge management practises (KMP) in connection to the ABDA and OP is investigated in this study[7]. A customised questionnaire was used to collect information from the respondents work in SMEs (Anshari., et al 20209)[8]. The Baron–Kenny technique is used to examine the mediation in this study. The ABDA had a favourable and significant influence on OP, according to the findings[9]. In addition, in SMEs, KMP has somewhat moderated the link between ABDA and OP. The dataset only included SMEs from Pakistan-controlled Kashmir, therefore it may not be representative of other locations. As a result, the findings' universal applicability is limited. The findings contribute to both conceptual and operational consequences for senior executives in firms, particularly in developing nations. This study aims to add to the literature by presenting new conclusions and discussions (Araz., et al 2020). These ramifications will aid senior management in making crucial decisions and will motivate practitioners seeking a competitive edge through organisational effectiveness in SMEs.

Administrative performance is linked to an industry's efforts to fulfil its objective including stakeholders' demands, along with market durability. It is also known as a process of measuring and evaluating an employee's performance in connection to its aims and goals, which comprises a comparison and projected results. The OP compares actual output or achievement of the company to the expected effect or goals. Better output is also contingent upon that business's capacity to interact with creative, secure scientific information systems, effectively applying everything in a way that favours the firm. Furthermore, OP may be defined as the process of ensuring that overall organisational commodities have been used properly, hence it encompasses all operations and responsibilities conducted by top managers.

Training programmes improve efficiency at which information is constructed, received, translated, and implemented. This encompasses information collection, preservation, transmission, and exploitation. Knowledge creation is a key component of KM theoretical approaches, covering four different stages to conversion that include explicit and tacit knowledge. Knowledge is a powerful instrument for overcoming organisational issues. Quality of service provided is the method of obtaining, transforming, studying, retrieval, and sharing intellectual assets in order to improve and maximise performance of the organisation, as well as to encourage development and economic growth. Businesses are generally concerned with expertise creation and maintenance in order to improve organisational effectiveness.

ii. *How Big Data Analytics to fulfil the distinct decision-support needs of various companies?*

Let's take a good look at some studies from 2016 - 2018 to discover if there was a predominant type of statistical analytics. For said 2016 International Big Data Survey: Key Decisions, upwards of 2,000 professionals were challenged to choose a group that better described their bank's judgement call procedure (Kambatla., et al 2014)[10]. In addition, the C-suite was informed which statistics they relied on the most. The shows the results: That quantitative research method topped (58 percent) in the "frequently informational judgement call" division; diagnoses analytics led (34 basis points) there in "somewhat statistics" area; or prescriptive modelling dominated in the "very statistics" paragraph (36 percent).The poll findings are consistent with ScienceSoft's hands-on research, highlighting the relevance of one or maybe more kinds of statistics at varying phases of such a business in the long term. Corporations that strived for intelligent decision, for illustration, regarded predictive analysis as

unsatisfactory and reinforced it with diagnostic testing assessment, or indeed went as far as known as a standard.

Analytics might well be categorised as follows[11]. We'll begin with one of the most simple or go to the highly difficult. As it happens, and the comprehensive the analysis, the lower the magnitude it produces.

d) Informative Analytics is a Type of Data Analysis that is Used to

Descriptive analytics provides an explanation for what occurred[12]. Let us use an example from ScienceSoft's experience: a manufacturer was able to answer a series of "what occurred" questions and choose target product categories after analysing monthly sales and income by product group, as well as the total quantity of metal parts produced each month.

e) Analytical Diagnostics

At such a point, past data may be compared to certain other data in order to determine how and why it occurred. For instance, users could see that a store may dig down into revenue and total profit to figure out why a company failed their net income objective in ScienceSoft's BI example. Another example from one of our data and analytics tasks: as in healthcare business, customer segmentation combined with multiple filters (such as diagnosis and medications prescribed) enabled for the identification of pharmaceutical impact.

f) Analytics that Predicts the Future

This analysis technique analyses what is mostly definitely going to happen. It uses exploratory and descriptive research analytics analysis to detect groupings and deviations, but also anticipate future occurrences, making it an effective forecasting technique (Dinh., et al 2020). See ScienceSoft's particular instance to know much about how powerful data analytics supported a famous FMCG organisation in forecasting what they would expect after revising marketing strategy.

g) Scenario Analysis

Scenario analysis's objective is to explain to you exactly how to use it in order to avoid future difficulties or profit on a steady increase. As a sample of Different scenarios from the capital projects, a multinational firm were able to discover possibilities for repeated purchase depending on customer analytics and sales data

i. *How is the Body of Research on Big Data Analytics, its Potential, and how Businesses Might Use them?*

Technologies that store the processing of the data are readily accessible at little cost[13]. Organisations, on the other hand, are already using methods to assess it at a completely different extreme, focusing on digital technologies to assure realistic, massive economic experimentation that educate regulators and analyse productivity information, commercial goals, and customer experience. In only certain circumstances, new trends might help businesses make major judgments (Dagilienė., et al 2019). These developments have the opportunity to usher in a seismic shift in science, development, and company's marketing. Several organisations, including Amazon, Google, and others, were early commandants, researching success variables to determine what boosted business income and user participation. Finance institutions are good experimenters, and therefore were among the first to develop their credit card consumer segmentation strategies.

Analytical information analysis is also being utilised by mortar and brick enterprise in order to adversely examine their capacity to inform customer information by assembling transaction - oriented information from millions of consumers via a reliable program; the data gathered is then used to analyse new possibilities, such as how to accomplish the most promote excellence for targeted customer segment and to make investment decision; and another organisations and firms utilising data analysis to gathering information via social media, such as Southwest Airlines[14].

Collected information is also used by brick & click businesses to intensively test their own ability to propose user information by designing and building payment relevant data from millions of customers through the use of a loyalty scheme; the collected data can be used to evaluate lots of opportunities, such as how to achieve the most promote excellence for targeted customer and make financial choices; and other enterprises utilising analysis of the data to gather intelligence.

Reengineering processes may have been used through companies to integrate data analytics in order to realise big data's possibilities & reap its rewards[15]. Big data analytics needs significant adaptation and segmentation of operational processes in collaboration with the institution's IT design in strengthening economic activities. Organisations should be related to data analyses now and in order to gain a competitive advantage since it has an effect on systems and applications.

IX. Results

Data science is a latest trend that appeared during the last couple of years, with so many intellectual organisations trying to implement big data analytics in order to stay competitive in the industrial environment. The idea here is to be agile in order to implement big data analytics to improve business. Several more companies failed to secure advanced analytics because they lacked the required infrastructure to implement Hadoop, while some others failed to take into account

the privacy licence by entering the business. The downside of someone using predictive analytics is obviously the confidentiality concerns; not much of the important information is free and accessible, therefore organisations need to examine the restrictions of collecting knowledge from other companies or even from individuals' personal accounts.

The handbook is genuine, information judgements lead toward the best moves, which tends to make supervisors start encouraging the said fact, and manufacturers which thus reveal how and when to integrate the specialist knowledge scope with big data analysis could well roll away from competition since some manufacturers may just not overpower this same computer aided to hold but also assess the irreplaceable relevant data, but instead they wouldn't have the comprehensive mastery but also practises to capture observation as well as derive benefit from huge amounts of data.

X. Future Scope

Any use of big data and analytics in modernization processes can increase modernization efficiency and overall effectiveness. The communication forward into big data and analytics coastline the efficiency prediction models, that either allow senior managers to use additional information in putting into consideration several more courses of action because once trying so hard for such a company's objectives. When entities use big data technologies, those that can ideally at least foresee now also wacky things, but rather strengthen performance of the process. Organisations realise the advantage of operational processes through cost reduction, the best operations plan, reduced inventory levels, the best organisational labour force, and the removal of unnecessary supplies. Companies also encourage improvements in operational excellence. Several capabilities of an organisation's advanced analytics (such as data pooling, retrieving, merging, and disseminating) and organisation characteristics (including such big data strategy) might enhance the optimal use of data analytics in processes and systems.

References Références Referencias

1. Dinh, L. T. N., Karmakar, G., & Kamruzzaman, J. (2020). A survey on context awareness in big data analytics for business applications. Knowledge and Information Systems, 62(9), 3387-3415.
2. Kangelani, P., & Iyamu, T. (2020). A model for evaluating big data analytics tools for organisation purposes. Responsible Design, Implementation and Use of Information and Communication Technology, 12066, 493.
3. Jha, M., Jha, S., & O'Brien, L. (2016, June). Combining big data analytics with business process using reengineering. In 2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS) (pp. 1-6). IEEE.
4. Anshari, M., & Sumardi, W. H. (2020). Employing big data in business organisation and business ethics. International Journal of Business Governance and Ethics, 14(2), 181-205.
5. Dubey, R., Gunasekaran, A., Childe, S. J., Bryde, D. J., Giannakis, M., Foropon, C., ... & Hazen, B. T. (2020). Big data analytics and artificial intelligence pathway to operational performance under the effects of entrepreneurial orientation and environmental dynamism: A study of manufacturing organisations. International Journal of Production Economics, 226, 107599.
6. Sivarajah, U., Irani, Z., Gupta, S., & Mahroof, K. (2020). Role of big data and social media analytics for business to business sustainability: A participatory web context. Industrial Marketing Management, 86, 163-179.
7. Walker, R. S., & Brown, I. (2019). Big data analytics adoption: A case study in a large South African telecommunications organisation. South African Journal of Information Management, 21(1), 1-10.
8. Malaka, I., & Brown, I. (2015, September). Challenges to the organisational adoption of big data analytics: A case study in the South African telecommunications industry. In Proceedings of the 2015 annual research conference on South African institute of computer scientists and information technologists (pp. 1-9).
9. Raut, R. D., Mangla, S. K., Narwane, V. S., Gardas, B. B., Priyadarshinee, P., & Narkhede, B. E. (2019). Linking big data analytics and operational sustainability practices for sustainable business management. Journal of cleaner production, 224, 10-24.
10. Hopkins, J., & Hawking, P. (2018). Big Data Analytics and IoT in logistics: a case study. The International Journal of Logistics Management.
11. Ijab, M. T., Wahab, S. M. A., Salleh, M. A. M., & Bakar, A. A. (2019, December). Investigating big data analytics readiness in higher education using the technology-organisation-environment (TOE) framework. In 2019 6th International Conference on Research and Innovation in Information Systems (ICRIIS) (pp. 1-7). IEEE.
12. Van Rijmenam, M., Erekhinskaya, T., Schweitzer, J., & Williams, M. A. (2019). Avoid being the Turkey: how big data analytics changes the game of strategy in times of ambiguity and uncertainty. Long range planning, 52(5), 101841.
13. Araz, O. M., Choi, T. M., Olson, D. L., & Salman, F. S. (2020). Data Analytics for Operational Risk Management. Decis. Sci., 51(6), 1316-1319.
14. Dagilienė, L., & Klovienė, L. (2019). Motivation to use big data and big data analytics in external auditing. Managerial Auditing Journal.

15. Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. Journal of parallel and distributed computing, 74(7), 2561-2573.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 22 Issue 1 Version 1.0 Year 2022
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Review on the Application of Machine Learning to Cancer Research

By Eunice C. Chibudike, Henry O. Chibudike, Nwaebuni E. Odega, Emeka E. Njokanma, Olubamike A. Adeyoju & Constance O. Ngige

Federal Institute of Industrial Research

Abstract- This study reviews the application of machine learning through different algorithms in cancer research. In recent years, the introduction of machine learning has been an exciting tool that enhances cancer research which has improved statistical method of speeding up both fundamental and applied research considerably.

The application of machine learning goes around in predicting the future events and outcomes with the available datasets. There is an indication that on yearly bases up to 14 million new cancer patients are diagnosed by Pathologists round the world, and they are people whose conditions are uncertain. Definitely, the diagnoses and prognoses of cancer have been performed by Pathologists. The research on machine learning flourished in 1980s and 1990s and information become digitalized through improved artificial network connectivity and computational power.

Keywords: machine learning, cancer research, cancer diagnoses, cancer predicting, and diagnosis.

GJCST-C Classification: F.1.1

Strictly as per the compliance and regulations of:

Review on the Application of Machine Learning to Cancer Research

Eunice C. Chibudike [α], Henry O. Chibudike [σ], Nwaebuni E. Odega [ρ], Emeka E. Njokanma [ω], Olubamike A. Adeyoju [¥] & Constance O. Ngige [§]

Abstract- This study reviews the application of machine learning through different algorithms in cancer research. In recent years, the introduction of machine learning has been an exciting tool that enhances cancer research which has improved statistical method of speeding up both fundamental and applied research considerably.

The application of machine learning goes around in predicting the future events and outcomes with the available datasets. There is an indication that on yearly bases up to 14 million new cancer patients are diagnosed by Pathologists round the world, and they are people whose conditions are uncertain. Definitely, the diagnoses and prognoses of cancer have been performed by Pathologists. The research on machine learning flourished in 1980s and 1990s and information become digitalized through improved artificial network connectivity and computational power. This shifted the effect of machine learning from artificial intelligence to solving practically natural problems. From there, shortly the potential of machine learning became obvious in medical science by scientists and gained its ground in medical specialties such as radiology, cardiology, mental health and pathology. In health care machine learning is used to interpret data hence speed up workflow, reduce medical error and promote human health. Pathologists are accurate at diagnosing cancer but have an accuracy rate of only 65% when predicting the development of cancer. Computed tomography, mammography, magnetic resonance imaging (MRI), or histopathology have been derived from imaging datasets over decades for diagnoses and staging prognosis of various cancers. The development of novel computational tools for stratification, grading, prognostication of patients with the goal of improving patient care has been achieved through the impact of machine learning.

Keywords: machine learning, cancer research, cancer diagnoses, cancer predicting, and diagnosis.

Author α §: Planning, Technology Transfer and Information Management, Federal Institute of Industrial Research, Oshodi, F.I.I.R.O., Lagos, Nigeria.

Author σ: Department of Chemical, Fiber and Environmental Technology, Federal Institute of Industrial Research, Oshodi, F.I.I.R.O., Lagos, Nigeria.

Author ρ: Nigerian Upstream Petroleum Regulatory Commission (NUPRC).

Author ω: Chevron Nigeria Limited.

Author ¥: Production, Analytical and Laboratory Management, Federal Institute of Industrial Research, Oshodi, F.I.I.R.O., Lagos, Nigeria.

e-mail: henrychibudike@gmail.com

I. Introduction

In recent years, the availability of large datasets combined with the improvement in algorithms and the exponential growth in computing power led to an unparalleled surge of interest in the topic of machine learning *(Khan Academy, 2018)*. Nowadays, machine learning algorithms are successfully employed for classification, regression, clustering, or dimensionality reduction tasks of large sets of especially high-dimensional input data *(Sunil Ray, 2017)*. In fact, machine learning has proved to have superhuman abilities in numerous fields (such as prediction, self-driving cars, image classification, 4 medical diagnoses etc.). As a result, huge parts of our daily life, for example, image and speech recognition, web-searches, fraud detection, email/spam filtering, credit scores, report extraction and many more are powered by machine learning algorithms *(Jonathan Schmidt, et al; 2019)*.While data-driven research and more specifically machine learning, have already a long history in biology or chemistry, they only rose to prominence recently in the field of cancer research. A first computational revolution in cancer research was fueled by the advent of computational methods, especially magnetic resonance imaging (MRI) *(Mandeep Kaur 2019)*. The constant increase in computing power and the development of more efficient codes also allowed for computational high-throughput studies of large samples in order to screen for the ideal experimental candidates.

Over decades, cancer researchers have researched into cancer to identify causes and dive into measures for its prevention, diagnosis, treatment and cure. The epidemiology, molecular bioscience to the performance of clinical trials have been evaluated and compared for the application of their various treatments; *(Susan A. Nadin-Davis, in Rabies (Second Editon), 2007)*. It could be applied in surgery, immunotherapy, hormone therapy, chemotherapy, radiation therapy and combined treatment modalities such as chemo-radiotherapy. In the mid-1990s the clinical cancer research shifted to therapies and this was derived from biotechnology research such as immunotherapy and gene therapy. Cancer research is done in academia, research institutes, and corporate environments, and is largely government funded, according to Martin Stumpe (AI and Data Science, MI, USA 2019), and collaborators

developed a deep-learning system (DLS) 2019. However, the challenges and interesting tasks of physicians are the accurate prediction outcomes of diseases. For this reason, Machine Learning methods have taken over in medical research as a popular tool. This review has an indication of some of the models that have been developed for cancer biopsies and prognoses. For instance, there a model that predicts cancer susceptibility; Craig Mermel (Google AI Healthcare, CA, USA 2019). The model was built to discriminate tumors as either malignant or benign in the midst of breast cancer patients. In this model, the completion of the tasks was done by ANN.

The building of this model was with a large number of hidden layers that could generalize data better. As thousands of mammographic data were fed in the model to obtain and learn the difference between benign and malignant tumors. Before being inputted, all the data was reviewed by radiologists. An approach by Regina Barzilay (MGH, MA, USA) 2019. The causes of cancer have been researched into many different disciplines including genetics, diet, environmental factors (i.e. chemical carcinogens). During the investigation of causes and also potential therapy targets, the route with data derived from clinical observations, basic research commences, and once convinced and independently obtained results are confirmed, proceeds with clinical research, which involves appropriate designed trials on consenting human subjects, with the goal to ascertain safety and efficiency of the therapeutic intervention method; Connie Lehman at Massachusetts General Hospital (MGH, MA, USA) 2019. One of the important parts of basic research is characterization of the potential mechanisms of carcinogenesis, having in mind the types of genetic and epigenetic changes that are associated with cancer development. The use of mouse is like a model for mammalian manipulation of the function of genes that play a role in tumor formation, while basic aspects such as bacteria and mammalian cells are assayed on cultures for tumor initiation, such as mutagenesis.

II. Methodology

Image filtering: In this review we examined a few of the most widely used image processing algorithms, then move on to machine learning implementation in image processing. At a glance is as follows:

- Feature mapping using the scale-invariant feature transform (SIFT) algorithm.
- Image registration using the random sample consensus (RANSAC) algorithm.
- Image Classification using artificial neural networks.
- Image classification using convolutional neural networks (CNNs).
- Image Classification using machine learning.
- Important Terms

Dynamic Contrast enhancement: Conventional contrast-enhanced magnetic resonance imaging (MRI) displays a single snapshot of tumor enhancement after contrast administration; although the anatomical information derived from such images is valuable, it lacks functional information ((National Institute of Health, 2017)). Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI), which relies on fast MRI sequences obtained before, during and after the rapid intravenous (IV) administration of a gadolinium (Gd) based contrast agent is analogous to a movie and is an emerging imaging method to assess tumor angiogenesis. To investigate whether a combination of radionics and automatic machine learning applied to dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) of primary breast cancer can non-invasively predict axillary sentinel lymph node (SLN) metastasis

Image Segmentation and Radiomic Feature Extraction: Axial DCE-MRI Digital Imaging and Communications in Medicine (DICOM) images were archived from the Picture Archiving and Communication System (PACS) (Stefan Leger et al 2019). The calculation of time signal intensity curves for tumor lesions in the DCE-MRI images were done using a GE Advanced Workstation ADW4.4 (Jan C. Peeken et al; 2019). Based on these curves, the volumes of interest (VOIs) were delineated on the whole tumor in the images with the strongest enhanced phase. The VOIs were determined manually by a radiologist with 10 years of experience who was blinded to the clinical information of the patients, and all contours were reviewed by another senior radiologist with 20 years of experience (Pan Sun et al 2019). If the discrepancy was ≥5%, the senior radiologist determined the tumor borders. Cohen's kappa method was used to assess inter-reader agreement (lanna Vial1 et al, 2018). In general, the (pre- processing of images are often the first step to later extraction of the features that would be used to train a machine learning classifier. Signal processing can be used to improve or eliminate properties of the image that could enhance the performance of the machine learning algorithm.

Classification of effectiveness of model: In machine learning, classification models are often used to get a predicted result of population data. Classification is one of the two sections of supervised learning deals with data from different categories (Manojit Chattopadhyay et al; 2017). The training dataset trains the model to predict the unknown labels of population data. There are multiple algorithms, namely, Logistic regression, K-nearest neighbour, Decision tree, Naive Bayes etc. All these algorithms have their own way of execution and different methods of prediction. But, at the end, we need to find the effectiveness of an algorithm (qbal H. Sarker, et al; 2019). To find the most suitable algorithm for a particular problem, there are model evaluation techniques. In this article several model evaluation techniques will be discussed.

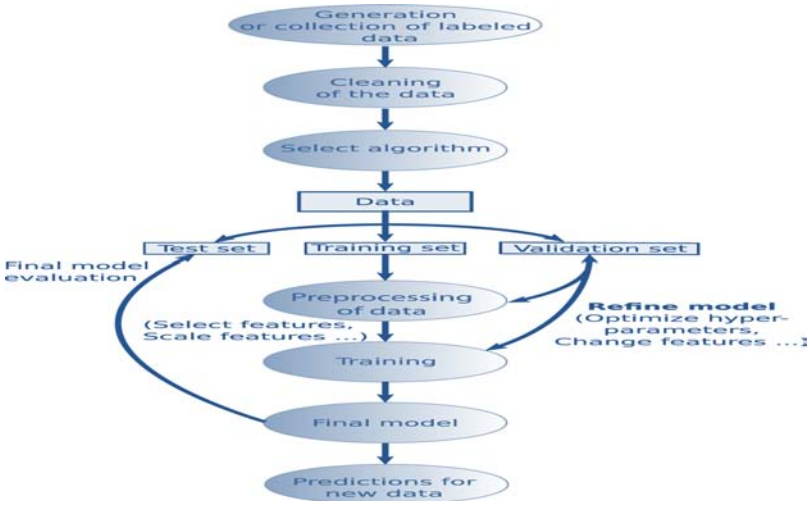


Figure 1: Supervised learning workflow(Jonathan Schmidt et al; 2019)

Figure 1: Depicts the workflow applied in supervised learning. One generally chooses a subset of the relevant population for which values of the target property are known or creates the data if necessary. This process is accompanied by the selection of a machine learning algorithm that will be used to fit the desired target quantity (Jonathan Schmidt et al; 2019).

a) Evaluation

One of the evaluations to conduct during prediction is Confusion matrix in the image. A confusion matrix is a table that describes the performance of a classifier/classification model. It contains information about the *actual and prediction classifications* done by the classifier and this information is used to evaluate the performance of the classifier. Here is the sample of a *Confusion Matrix (Banso D. Wisdom 2017).*

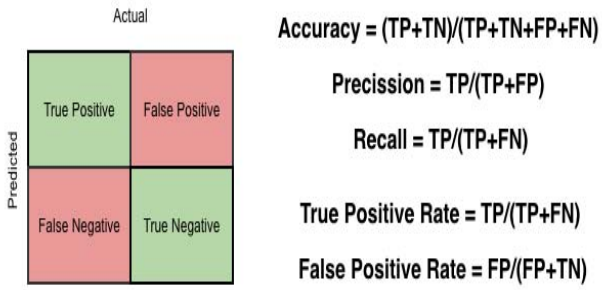


Figure 2: Confusion Matrix (Banso D. Wisdom 2017)

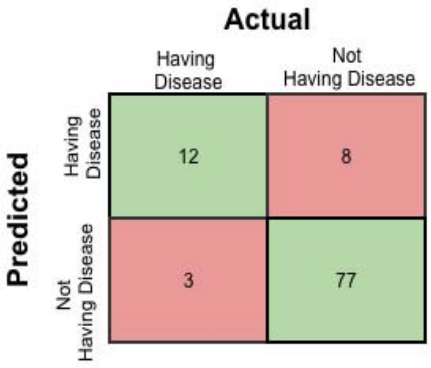


Figure 3: Confusion Matrix of predicting a disease

Confusion matrix is the image given above. This is a matrix representing the results of any binary testing. For example, let us take the case of predicting a disease. You have done some medical testing and with the help of the results of those tests, you are going to predict whether the person is having a disease. So, actually you are going to validate if the hypothesis of declaring a person as having disease is acceptable or not. Say, among 100 people you are predicting 20 people to have the disease. In actual only 15 people to have the disease and among those 15 people you have diagnosed 12 people correctly. So, if I put the result in a confusion matrix, it will look like the following.

As observed in figure 3

True Positive: 12 (You have predicted the positive case correctly!)

1. *True Negative:* 77 (You have predicted negative case correctly!)
2. *False Positive:* 8 (You have predicted these people as having disease, but in actual they do not have. There is no course for alarm; this can be rectified during further medical analysis. So, this is a low risk error. This is type-II error in this case.)
3. *False Negative*: 3 (You have predicted these three poor fellows as fit. But actually they have the disease. This is dangerous! Be careful! This is type-I error in this case.)

Now, this is the accuracy of the prediction model was followed to get this results; i.e the ratio of the accurately predicted number and the total number of people which is (12+77)/100 = 0.89. There is need for you to study the confusion matrix thoroughly so as to find the following things.

b) Test for specificity and sensitivity

In medical diagnosis, the term test sensitivity is the reliability of a test to correctly identify those affected with the disease (true positive rate), while test specificity is the ability of the test to correctly identify those that are not affected with the disease (true negative rate).

Table 1: Test for specificity and sensitivity (Dr. Aaron Swanson, 2011).

	Sensitivity	Specificity
Definition	Proportion of patients with a disease who test positive	Proportion of patients without the disease who test negative
100% (1.0) Means	The test correctly identify every person who has the target disorder	The test correctly identify every person who does not have the target disorder
Statistical Outcome	True Positive	True Negative
Ideal Test Result	Negative Test Result	Positive Test Result
Test Interpretation	They are definitely not positive → They DON'T have it	They are definitely not negative → They DO have it
The Rule	Rule Out (SnOut)	Rule In (SpIn)

Sensitivity and specificity are great values to lead you in your fair clinical examination. It gives more information regarding the patient and guide to a better assessment and authentic diagnosis. Keep in mind that there is always the possibility of false positives and negatives. Special tests should never be the only sign to determine a patient's pathology. It is merely a piece of the clinical examination and assessment (Dr. Aaron

Table 2: Attribute Information Swanson, 2011).

Sample code number	Id number
Clump Thickness	1 – 10
Uniformity of Cell Size	1 – 10
Uniformity of Cell Shape	1 – 10
Marginal Adhesion	1 – 10
Single Epithelial Cell Size	1 – 10
Bare Nuclei	1 – 10
Bland Chromatin	1 – 10
Normal Nucleoli	1 – 10
Mitoses	1 – 10
Class	(2 for benign, 4 for malignant)

III. Results and Discussion

a) Parameters for cancer dictation

In the development of metastases there is a negative prognostic parameter for the clinical result of breast cancer. Bone consists of the first site of distant metastases for several affected women. The idea of this attribute information is to perform an exploratory analysis of the information contained in the dataset, figuring out ways of making the dataset tidier. The ultimate objective is to, in the end, build and compare models to predict if a given tumor is benign or malignant (breast cancer) using the information available on the dataset in Table 2 below.

Table 3: A Sample of Analysis and Modeling of Breast Cancer Data (Random Forest model) from (*ml-repository* '@' *ics.uci.edu)*.

The analysis shows that, with a Random Forest model, we can predict if a given tumor is malignant with 97.86% of Accuracy. This result is 1.96% higher than the Accuracy of 95.90% reported in the UCI Machine Learning as the highest for this dataset (*ml-repository* '@' *ics.uci.edu*). We also conclude that the most important information for this prediction is the 'uniformity of the cell size'. The idea is to perform an exploratory analysis of the information contained in the dataset, figuring out ways of making the dataset tidier. The ultimate objective is to, in the end, build and compare models to predict if a given tumor is available on this dataset. The analysis show that, with a Random Forest model, we can predict if a given tumor is malignant or benign for (breast cancer) using the information (*ml-repository* '@' *ics.uci.edu*).

Table 4: A sample of Dataset (ml-repository '@' ics.uci.edu)

ID	ID number	Radius mean	Texture mean	Perimeter mean	Smoothness mean	Compactness mean	Concavity mean	Concave points mean	Symmetry mean	Fractal dimension mean	Diagnosis
3	842302	19.69	10.38	122.8	0.118	0.2776	0.3001	0.1471	0.2419	0.07871	1
4	842517	20.57	17.77	132.9	0.085	0.07864	0.0869	0.07017	0.1812	0.05667	1
5	84300903	19.69	21.25	130	0.11	0.1599	0.1974	0.1279	0.2069	0.05999	1
6	84348301	11.42	20.38	77.58	0.143	0.2839	0.2414	0.1052	0.2597	0.09744	1
7	84358402	20.29	14.34	135.1	0.1	0.1328	0.198	0.1043	0.1809	0.05883	1
8	843786	12.45	15.7	82.57	0.128	0.17	0.1578	0.08089	0.2087	0.07613	1
9	844359	18.25	19.98	119.6	0.095	0.109	0.1127	0.074	0.1794	0.05742	1
10	84458202	13.71	20.83	90.2	0.119	0.1645	0.09366	0.05985	0.2196	0.07451	1
11	844981	13	21.82	87.5	0.127	0.1932	0.1859	0.09353	0.235	0.07389	1
12	84501001	12.46	24.04	83.97	0.119	0.2396	0.2273	0.08543	0.203	0.08243	1
13	845636	16.02	23.24	102.7	0.082	0.06669	0.03299	0.03323	0.1528	0.05697	1
14	84610002	15.78	17.89	103.6	0.097	0.1292	0.09954	0.06606	0.1842	0.06082	1
15	846226	19.17	24.8	132.4	0.097	0.2458	0.2065	0.1118	0.2397	0.078	1
16	846381	15.85	23.95	103.7	0.084	0.1002	0.09938	0.05364	0.1847	0.05338	1
17	84667401	13.73	22.61	93.6	0.113	0.2293	0.2128	0.08025	0.2069	0.07682	1
18	84799002	14.54	27.54	96.73	0.114	0.1595	0.1639	0.07364	0.2303	0.07077	1
19	848406	14.68	20.13	94.74	0.099	0.072	0.07395	0.05259	0.1586	0.05922	1
20	84862001	16.13	20.68	108.1	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	1

The diagnosis of breast tissue(1 = malignant, 0 = benig)

b) Datasets and their Features

In table 4 above, when it comes to classification, there is a need of dataset to classify. Dataset is a statistical matrix which represents different features. It is a matrix where all the information about different features is given. Each column of the dataset represents the feature of the tumorous tissue and each row represents the number of instances. Table 4 is the details of attributes found in *WDBC dataset (19) (Vania V Estrela et al; 2019):* ID number, Diagnosis (M=Malignant, B=Benign) and ten real real-valued features are computed for each cell nucleus: radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave points, Symmetry and Fractal dimension *(20) (Anirban Banerji, 2013*). These features are computed from digitized image of a fine needless aspirate (FNA) of a breast mass *(ml-repository* '@' *ics.uci.edu).* They described characteristics of the cell nuclei present in the *image (21)(Tula Neilson 2012).* When the radius of an individual nucleus measured by averaging the length of the radial line segments, it is defined by the centroid of the snake and the individual snake points. The Nuclear Perimeter constitutes of the total distance between consecutive snake points.

c) *Exploratory Analysis*

To explore this data and later also be able to create models correctly, we need to separate our data into *train* and *test* data. This is to achieve a simulated real world dataset (test) that have class information that has not been used in anyway during the analysis (instead we use train). This ensures that our test dataset is really simulating real world data, since it has not been seen during exploration or modeling (*Prasad Patil 2018)* For this purpose, the R package caTools, as displayed below

```
library (caTools)
set. Seed(1000)
split=sample. Split (cancer$Class, Split Ratio=0.80)
train=subset(cancer, split==TRUE)
test=subset(cancer,split==FALSE)
```

d) *Cancer Research*

Cancer research is a research into the cause of cancer, prevention, diagnosis, treatment and cure which involves many diverse disciplines including genetics, diet, environmental factors (i.e. chemical carcinogens). The ranges of cancer research are from epidemiology, molecular bioscience to the performance of clinical trials to make evaluation and comparison of the application of various cancer treatments (Douglas Hanahan et al 2011). Cancer research has been on for ages. In the early years of research, the focus was on the causes of cancer. The first identification of environmental trigger (chimney soot) for cancer was PercivallPott in 1775 and identification of cigarette smoking as a cause of lung cancer in 1950. The treatment of cancer was early focused on enhancing surgical techniques for removing tumors. Radiation therapy took hold in the 1900s (Douglas Hanahan et al; 2011.) The development and definition of Chemotherapeutics were done throughout the 20th century. Cancer research involves various types and interdisciplinary areas of research. Scientists in cancer research may get their trainings in areas such as epidemiology, chemistry, biomedical engineering, molecular biology, medical physics, physiology and biochemistry. Research principles and mechanisms were always clarified at basic research level. Translational search aims to discover the mechanisms of cancer development and progression and convert n basic scientific results into ideas that can be applied to the treatment and prevention of cancer (The Hallmarks of Cancer, published in 2000). The development of pharmaceuticals, surgical procedures, and medical technologies for the eventual treatment of patients are achieved through clinical research

e) *Genes involved in cancer*

The aim of *oncogenomics* is to discover new *oncogenes* or *tumor suppressor genes* that may provide new knowledge into diagnosing cancer, predicting clinical outcome of cancers, and an update targets for cancer therapies *(John Carpten et al on RNASEL: M. Sprinsky/AACR 2002)*. As the Cancer Genome Project stated in a 2004 review article, "a central aim of cancer research has been to identify the mutated genes that are causally implicated in oncogenesis (cancer genes). The project of Cancer Genome Atlas is a related effort which focused in investigating the genomic changes that relates to cancer, while the genetic mutations from hundreds of thousands of human cancer samples were acquired from COSMIC cancer database documents. In the cause of several literature reviews, there is an indication that projects have been carried out, involving about 350 different types of cancer, have identified ~130,000 mutations in ~3000 genes that have been mutated in the tumors. The majority occurred in 319 genes, of which 286 were tumor suppressor genes and 33 oncogenes *(American Association for Cancer Research, Databases for oncogenomic research)*. Some hereditary factors can shoot up the chance of cancer-causing mutations that includes activating oncogenes or inhibiting tumor suppressor genes. The functioning of various oncogenes and tumor suppressor genes can be interrupted at different levels of tumor progression. Gene's mutations can be used to classify the malignancy of a tumor. In some stages, tumors can form a resistance to cancer treatment. The understanding of tumor progression and treatment success is achieved when identification of oncogenes and tumor suppressor genes done. The function of a given gene in cancer progression may differ tremendously, as it depends on the stage and type of cancer involved (*the National Cancer Institute 2017)*

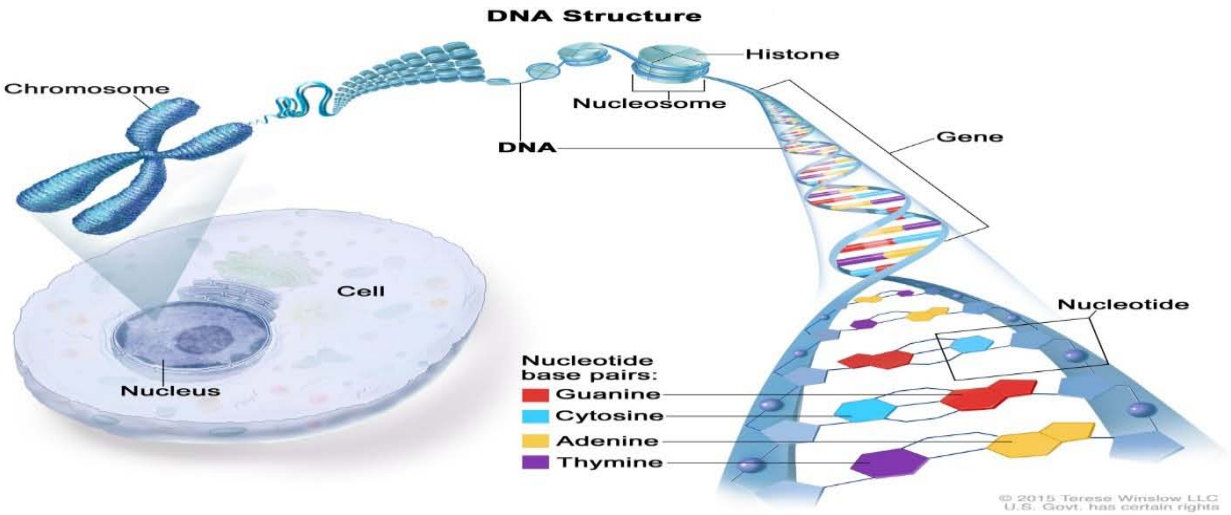


Diagram: Structure of DNA (*the National Cancer Institute 2017*)

It has been ascertained that most DNA is seen inside the nucleus of a cell, where it forms the chromosomes. Chromosomes acquire proteins called histones that join to DNA. DNA has two strands that fickle into the shape of a spiral ladder called a helix. DNA is made up of four building blocks called nucleotides: adenine (A), thymine (T), guanine (G), and cytosine (C) *(the National Cancer Institute 2017)*. The nucleotides attach to each other (A with T, and G with C) to form chemical bonds called base pairs, which connect the two DNA strands *(the National Cancer Institute 2017).* Genes are short pieces of DNA that carry specific genetic information *(the National Cancer Institute 2017).*

IV. Cancer Detection

It is advisable to dictate cancer early so as to avert the difficulty of treating it in later stages. Accuracy in detection of cancer is paramount because false positives can cause harm owing to unnecessary medical procedures *(Joensuu, Heikki et al; 2013)*. Some screening procedures are not accurate currently (such as prostate-specific antigen testing). In some other cases like a colonoscopy or mammogram are unpleasant and gives room for some patients to opt out. Active research is to address all these problems *(Andrew Mckeon et al; 2016).*

Three main ways cancer cells can spread.

1. Through the blood vessels: This is known as hematogenous spread. Cancerous cells invade blood vessels and use the flow of blood cells as transportation.
2. Through nearby tissue: This is known as transcoelomic spread. Cancerous cells penetrate the surfaces of peritoneal cavities in the body.
3. Through the Lymphatic system: This is known as lymphatic spread. Cancerous cells invade the lymph nodes and use the lymphatic system to travel.

V. Application of Machine Learning to Cancer Research

There are two ways to cancer, Prediction/ Prognosis and Detection/Diagnosis. *In cancer Prediction/ Prognosis there are three core points:*

- *Prediction of cancer susceptibility (i.e. risk assessment)*
- *Prediction of cancer recurrence*
- *Prediction of cancer survivability*
- Risk assessment is predicting the probability of developing a type of cancer prior to the occurrence of the disease *(Wolters Kluwer Health, Inc. 2003).* The prediction of cancer recurrence is about trying hard to discover the likelihood of re-developing cancer after to the apparent resolution of the disease. The predicting of cancer survivability is determining outcomes like life expectancy, progression, survivability, tumor-drug sensitivity after the diagnosis of the disease *(Joseph A. Cruz, 2006)*. The quality of the diagnosis and other factors determines the success of the prognostic prediction. However, a medical diagnosis and a prognostic prediction must take into account more than just a simple diagnosis before disease prognosis can takes place.

Experts in cancer research have already compiled a list of features to dictate cancer cells, which is preferably to adding chemicals to blood samples that destroys cells *(Cancer Informatics 2006: 2 59–78)*. Refractive indices is an example of what data is used to help the machine predict and diagnose cancer and by using that, it tells us how much light slows down when passing through cells. It helps in light absorption, scattering properties as well as morphological features *(Joseph A. Cruz et al; 2006)*. The input is an image, then the neural networks help identify the cancer cells by learning the relationships of what values of the features

leads to cancer cells. The deep learning algorithm makes use of these features to classify cells based on learning the values of each feature that leads to a cancerous cell. Metastasized detection requires highly.

a) *Artificial Intelligence*

Artificial Intelligence manages more comprehensive issues of automating a system. This computerization should be possible by utilizing anything any field such as image processing, cognitive science, neural systems, machine learning etc. Most recent updates in Artificial Intelligence (AI) are due to application of machine learning to very large data sets. Artificial Intelligence is when computer algorithm does intelligent work. Artificial intelligence is the superset of machine learning i.e all the machine learning is artificial intelligence but not all the AI is machine learning. *Machine learning (ML*) manages and influences user's machine to gain from the external environment. This external environment can be sensors, electronic segments, external storage gadgets and numerous other devices. Machine Learning enables computers to learn by themselves. With the aid of modern computers, it is easier manipulating large data sets *(Japan AI Experience in 2017)*. The algorithms detect patterns and learn ways to make predictions and recommendations by processing data and experiences, instead of being explicitly written in program. *Machine Learning* is made up of three major types: *Supervised* In which data is labeled. The model is to identify the labels and put them in groups accordingly. In other words, the input is provided to the model and the desired output is offered. This process is done countless times until the desired output is obtained. *Unsupervised* In which data is not labeled *(B.J. Copeland; 2019)*. Different features and classifications have to be identified based on the distinct characteristics through the model. In this case, the input is given, but there is no expected output. The logical classifications or groupings are made by computer. *Reinforcement*: This learning treats the problem of finding optimal or sufficiently good actions for a situation in order to maximize a reward. In other words, it learns from interactions *(JAKE FRANKENFIELD; 2019).*

VI. Algorithm for Cancer Cell Development Prediction

Machine learning algorithms have already revolutionized other fields, such as image recognition. However, the development from the first perception up to modern deep convolutional neural networks was a long and tortuous process. In order to produce significant results in cancer research, one necessarily has not only to play to the strength of machine learning techniques but also apply the lessons already learned in other fields (*Konstantina Kourou et al; 2015).*

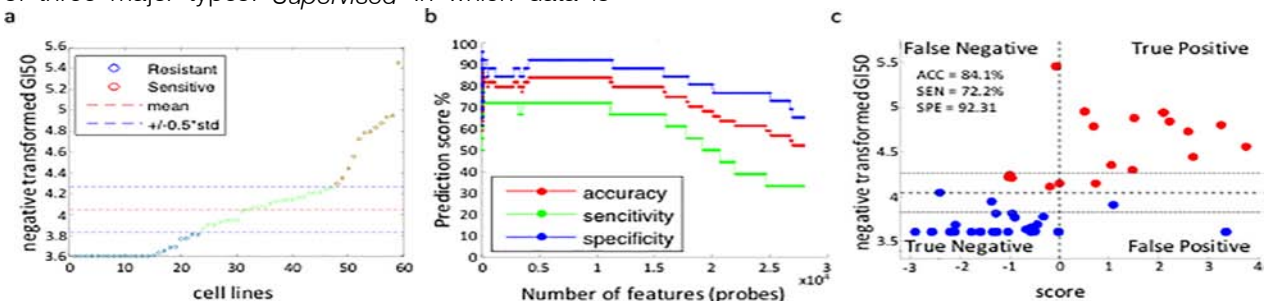


Figure 1: A Sample Algorithm for Cancer Prediction (*Konstantina Kourou[a] et al; 2015).*

VII. Conclusion

In conclusion, there has been an estimated 100 plus types of cancerous cells. This imposes difficulty in curing cancer. For example, if a certain group of similar cancer cells accepts a particular drug or treatment, it could have a peculiar different effect on another group.

- Skilled pathologists or radiologists that perform manual segmentation, which is time-consuming and prone to error, particularly in cases where tumors are few or there are no tumors. Deep learning networks have significantly enhanced accuracy on a wide range of computer vision tasks such as object detection, image recognition, and semantic segmentation.
- Machine learning is now a veritable tool used in cancer research labs to classify tumors based on growth characteristics; features such as where they grow, how fast they grow and size etc. and they are classified into groups based on similar range of predictive outcomes. The reason being that, one can create a controlled environment by picking a classified group and perform desired experiments to see the effect.

References Références Referencias

1. W.H. Wolberg, &, O.L. Mangasarian `(1990). In Proceedings of the National Academy of Sciences, 87, 9193--9196

2. J. Zhang, (1992). Selecting typical instances in instance-based learning. (pp. 470--479). Aberdeen, Scotland: Morgan Kaufmann.
3. PA. Futreal, L. "Early Theories about Cancer Causes – American Cancer Society".Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, Stratton MR (2004). "A census of human cancer genes". Nat. Rev. Cancer. 4(3): 177–83. doi:10.1038/nrc1299. PMC 2665285. PMID 14993899.
4. Forbes S, Clements J, Dawson E, Bamford S, Webb T, Dogan A, Flanagan A, Teague J, Wooster R, PA. Futreal, MR Stratton (2006). "COSMIC 2005". Br J Cancer. 94 (2): 31822. doi:10.1038/sj.bjc.6602928. PMC 2361125. PMID 16421597.
5. Gavin Brown. Diversity in Neural Network Ensembles. The University of Birmingham. 2004.
6. BabackMoghaddam and Gregory Shakhnarovich. Boosted Dyadic Kernel Discriminants. NIPS. 2002.
7. Krzysztof Grabczewski and Wl/odzisl/aw Duch. Heterogeneous Forests of Decision Trees. ICANN. 2002.
8. AndrásAntos and BalázsKégl and Tamás Linder and Gábor Lugosi. Data-dependent margin-based generalization bounds for classification. Journal of Machine Learning Research, 3. 2002. [View Context].
9. P. Kristin Bennett and AyhanDemiriz and Richard Maclin. Exploiting unlabeled data in ensemble methods. KDD. 2002.
10. P. Kristin Bennett and Erin J. Bredensteiner. A Parametric Optimization Method for Machine Learning. INFORMS Journal on Computing, 9. 1997.

This page is intentionally left blank

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 22 Issue 1 Version 1.0 Year 2022
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Design of Automated Database System for Storage and Management of Reports on Mycotoxins Contaminated Agricultural Products in Sub-Saharan Africa

By Eunice C. Chibudike, Henry O. Chibudike, Nwaebuni E. Odega, Emeka E. Njokanma, Olubamike A. Adeyoju & Constance O. Ngige

Federal Institute of Industrial Research

Abstract- This paper discusses the idea and the design of an automated system for storage and management of mycotoxins reports for decision making. Mycotoxins are poisonous chemical compounds produced by certain fungi. Mycotoxins are fungal secondary metabolites that contaminate various feedstuffs and agricultural crops. The contamination of food by mycotoxins can occur before production, during storage, processing, transportation or marketing of the food products. High temperature, moisture content and water activity are among the factors that facilitate the production of mycotoxins in food. The five major mycotoxins produced in food and feedstuffs are Aflatoxins, ochratoxins, fumonisins, deoxynivalenol and zearalenone.

Keywords: sub-saharan africa, health hazards, mycotoxinsautomated data base system, and agricultural products.

GJCST-C Classification: H.2

Strictly as per the compliance and regulations of:

Design of Automated Database System for Storage and Management of Reports on Mycotoxins Contaminated Agricultural Products in Sub-Saharan Africa

Eunice C. Chibudike [α], Henry O. Chibudike [σ], N waebuni E. Odega [ρ], Emeka E. Njokanma [ω]
Olubamike A. Adeyoju [¥] & Constance O. Ngige [§]

Abstract- This paper discusses the idea and the design of an automated system for storage and management of mycotoxins reports for decision making. Mycotoxins are poisonous chemical compounds produced by certain fungi. Mycotoxins are fungal secondary metabolites that contaminate various feedstuffs and agricultural crops. The contamination of food by mycotoxins can occur before production, during storage, processing, transportation or marketing of the food products. High temperature, moisture content and water activity are among the factors that facilitate the production of mycotoxins in food. The five major mycotoxins produced in food and feedstuffs are Aflatoxins, ochratoxins, fumonisins, deoxynivalenol and zearalenone. In Africa, mycotoxin contamination is considered to be a major problem with implications that causes human and animal health hazards and poor economy. Aflatoxin-related hepatic diseases are reported in many African countries. Ochratoxin and fumonisin toxicity in humans and animals is widespread in Africa. The available and updated information on the incidence of mycotoxin is not collectively vivid for policy making. A complete automated system allows to monitor the statistical report of mycotoxins stored in agricultural products. This study involves analytical Service conducted onMycotoxins such as Mold Culture and Identification and Chemical Analysis which involvesmicrobiological Culturing; Microscopic or biochemical identification, enzyme linked Immunosorbent (ELISA), tin layer Chromatography (TLC), high Performance Liquid Chromatography (HPLC), and gas Chromatography /Mass Spectroscopy. The design and development of Mycotoxins Automated Database System (MADAS) makes provision for easy access and acknowledgment of mycotoxins in differentgrains, fruits, vegetables and foods in Sub-Saharan Africa. It also enhances robust data collection, management, andanalysis, a secure and protected data environment, error reduction and data storage to facilitate regulatory compliance, improved maintainability, standardization, control, predictability, and traceability of data and lower costs due to automation of labor intensive tasks and elimination of redundant work.

Keywords: *sub-saharan africa, health hazards, mycotoxinsautomated data base system, and agricultural products.*

Author α, §: Planning, Technology Transfer and Information Management, Federal Institute of Industrial Research, Oshodi, F.I.I.R.O., Lagos, Nigeria.
Author σ: Department of Chemical, Fiber and Environmental Technology, Federal Institute of Industrial Research, Oshodi, F.I.I.R.O., Lagos, Nigeria.
Author ρ: Nigerian Upstream Petroleum Regulatory Commission (NUPRC).
Author ω: Chevron Nigeria Limited
Author¥: Production, Analytical and Laboratory Management, Federal Institute of Industrial Research, Oshodi, F.I.I.R.O., Lagos, Nigeria.
e-mail: henrychibudike@gmail.com

I. Introduction

In Sub-Saharan Africa, work on mycotoxins covering field cases, acute exposures and chronic effects related to dietary intake is reviewed. Mycotoxins have been implicated in the etiology of diseases like kwashiorkor, marasmic kwashiorkor, hepatocellular carcinoma in humans, encephalopathy and other acute diseases in animals. Mycotoxins are poisonous chemical compounds produced by certain fungi. There are many such compounds, but only a few of them are regularly found in food and animal feedstuffs such as grains and seeds. Nevertheless, those that do occur in food have great significance in the health of humans and livestock. Since they are produced by fungi, mycotoxins are associated with diseased or moldy crops, although the visible mold contamination can be superficial. The effects of some food-borne mycotoxins are acute, symptoms of severe illness appearing very quickly. Other mycotoxins occurring in food have longer term chronic or cumulative effects on health, including the induction of cancers and immune deficiency. Information about food-borne mycotoxins is far from complete, but enough is known to identify them as a serious problem in many parts of the world, causing significant economic losses. The economic and health hazards of mycotoxin contamination in crops and food products present a huge challenge, especially in Sub-Saharan Africa, where there is limited data to ascertain the degree of harm caused by these toxins. Tackling this problem needs a multi-factorial approach. A workable strategy would be the systematic development of centers of research expertise, and building research capacities aimed at establishing a database on mycotoxins found in different grains and seeds at each

given time and health-related risks caused by mycotoxins. Growing the interest of the African scientific community towards increasing the research output in the region is imperative. To this end, building an automated system on mycotoxicology is a good starting point. This will enable a better collation of data which will aid decision making. This research work will also aid the access and acknowledgment of mycotoxins reports of agricultural products in Sub-Saharan Africa. To aid policy makers in having an overview of mycotoxin reports of agricultural products in Sub-Saharan Africa. This research work will be relevant to research officers especially those in the area of mycotoxins and related research topics to have quick access to referencing data and to make comparisons as desired for better results. It will as well aid interaction between research scientists and farmers for updates. This will enable governments to make adequate policies that will help to improve and secure human and animal health.

II. Methodology

a) Materials

This study adopts a case study of some FIIRO scientists in Food, Biotechnology and CEFT departments. Verbal interview was conducted randomly to ascertain some facts about Mycotoxins. The following material were involved: Paper, pen, computer system, flash, and printer.

b) Methods

The analytical services used in testing for mycotoxins and the methods used are presented in tables and figure below. This gave us the insight in the designing and development of the database. The system will be able to display the services carried out and also the methods used for each agricultural product. Figure 1 shows the flow of the method used.

Table 1: Examples of Analytical Service on Mycotoxins

Services	Method
Mold Culture and Identifications.	➢ Microbiological Culturing; Microscopic or biochemical identification test.
Chemical Analysis	➢ Enzyme linked Immunosorbent (ELISA). ➢ Tin layer Chromatography (TLC). ➢ High Performance Liquid Chromatography (HPLC). ➢ Gas Chromatography /Mass Spectroscopy.

Table 2 : Examples of Food Contamination with Aflatoxins in Sub-Saharan Africa

Country	Food	Year(s)	Sample source	AF types	N	+ves (%)	Range (ppb)	Mean (ppb)	Ref
Nigeria	Maize	2001	Preharvested	AFB1	103	18	3-130	22	16
	Dry roasted groundnuts	(2005)	Retail	Total	106	64	5-165	25.5	17
Ghana	Kenkey (fermented maize)	(2000)	Processing sites	Total	15	53	2-662	176	18
		1996	Processing sites	Total	12	100	0.7-313	135.4	19
	Kenkey (cooked fermented maize)	1996	Processing sites	Total	16	94	0.7-313	50.9	19
Botswana	Groundnut	2001	Retail Outlets	Total	120	78	12-329	118	20

Year: year in which the survey was carried out, while the years in parentheses are years in which studies were published

AF types: aflatoxins types; total-AFB1+AFB2+AFG1+ AFG2

N: Total number of samples analyzed

+ves (%): percentage of samples contaminated

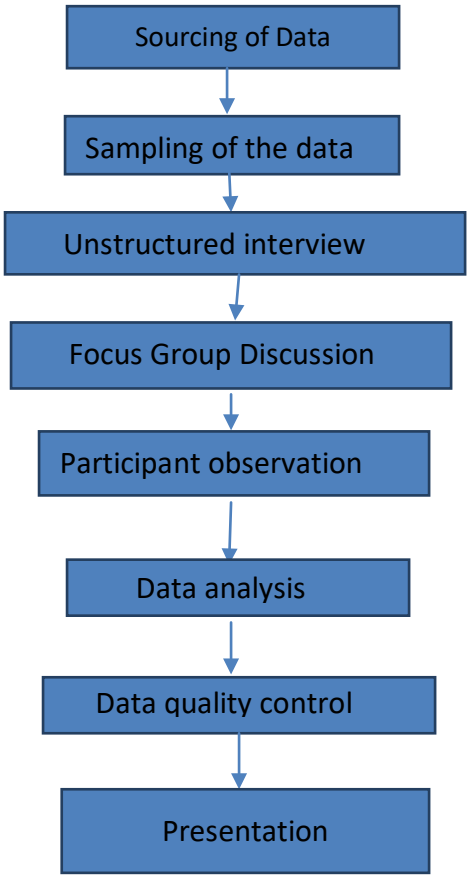


Figure 1: Flow Diagram of the Research Methods Adopt

III. Results and Discussion

An unstructured interview was conducted with fifteen (15) respondents in the Federal Institute of Industrial Research Oshodi, Lagos as presented in table 3. The socio-economic characteristics of the respondents and the opinions of Mycotoxins were gathered.

Table 3: Research Scientists (Case study)

Department	Number	Percentage (%)
Biotechnology	60	40
Food Technology	40	26.7
CFET	50	33.3

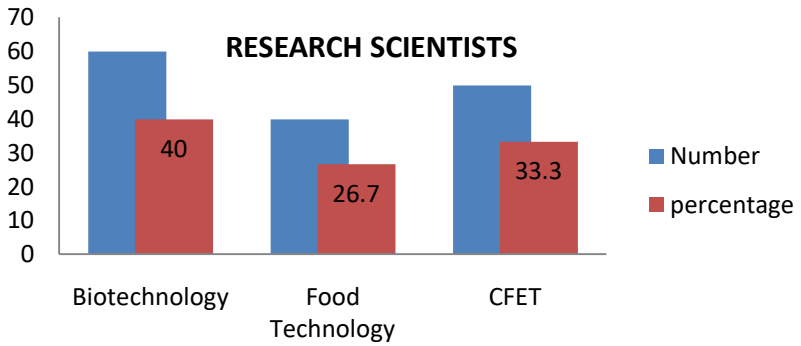


Figure 2: The Research Scientists

From Table 3, fifteen scientists responded, 6 from department of Biotechnology, 4 from Food technology department and 5 from CEFT department with 40, 26.7 and 33.3% respectively. The socio-economic characteristics of these respondents were gathered based on UN standard, which formed Table 2.

Table 4: Socio-economic characteristics of respondents

VARIABLES	CATEGORIES	NUMBER	PERCENTAGE (%)
Age	15-25	0	0
	25-34	50	33.3
	34-45	30	20
	45-54	60	40
	54-64	10	6.7
	65 Above	0	0
Sex	Male	70	46.7
	Female	80	53.3
Education Level	B.Sc	20	13.3
	M.Sc	60	40
	Ph.D	70	46.7
Marital status	Married	90	60
	Single	50	33.3
	Divorce	0	0
	Widow/widower	10	6.7

The issue of Mycotoxins in agricultural products is very vital in food and biotech industry. These caused health hazards in humans and animals. The opinions of these scientists in Table 5, formed our decision in designing an automated database system

Table 5: Mycotoxins in agricultural products

Variables	Yes	No	Percentage Yes	Percentage No
Relevant Issue	150	0	100	0
Positive Effect	10	140	6.7	93.3
Negative Effect	140	10	93.3	6.7
Harmful to Health	140	10	93.3	6.7
Controlled	50	100	33.3	66.7
Research on-going	130	20	86.7	13.3
Involved in the research	90	60	60	40
Need referencing data	150	0	100	0
Internet as data source	90	60	60	40
Database needed	150	0	100	0
Data will quicken the solution	150	0	100	0

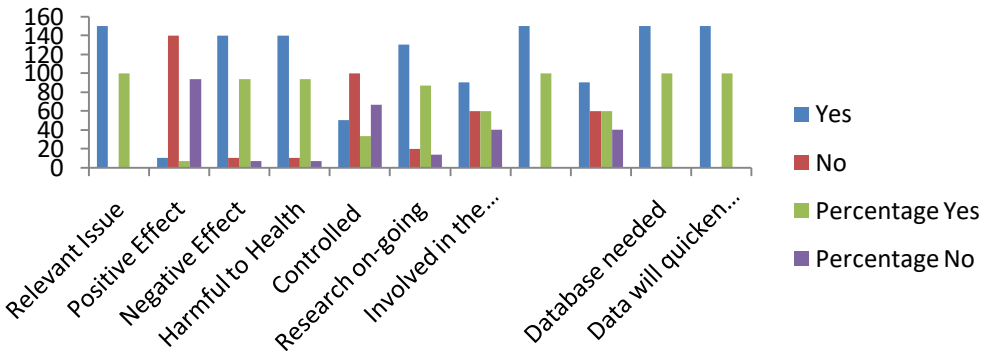


Figure 3: Mycotoxins in Agricultural products

Mycotoxins Automated Database System (MADAS)

A database has been developed for users to showthe Mycotoxins statistical reports of agricultural products that are registered and available here in Sub-Saharan Africa.

- The languanges used are Microsoft.Net framework, C# at front end and SQL/my SQL
- *Searching the database*

In order to search the database, the following instructions must be followed:

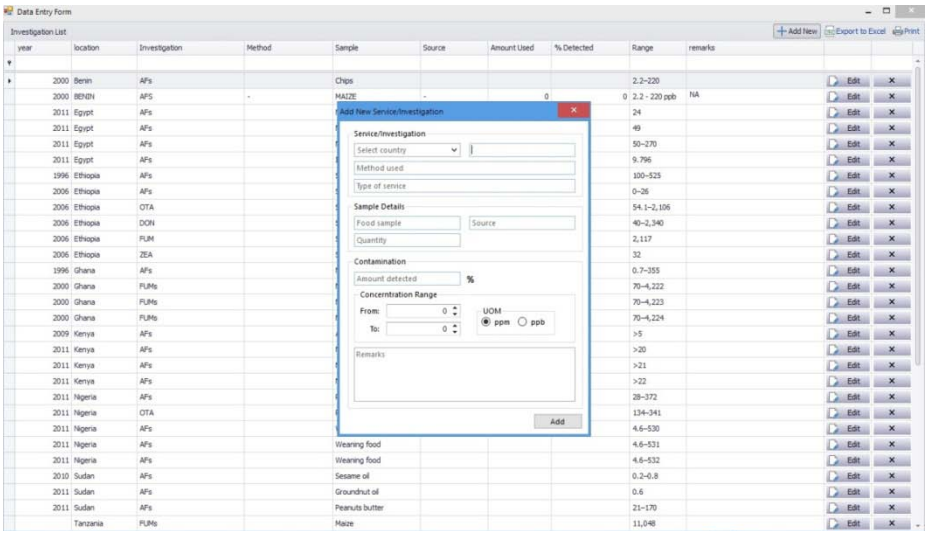
The database is available on the website of the Federal institute of Industrial Research oshodi, (FIIRO), Lagos.

http://www.mycotoxin.gov.ng.ie/biotech/crops/mycotoxinscertification/mycotoxinsdatabase/

- *To Navigating the program:* - Launch the Application
- from the investigation grid view, Click on Export to excel or pdf to print
- *Add new Service/Investigation:* - Launch the Application

- Click Add New- "To open the service/investigation dialog box

Table 6: The created interface of data capturing

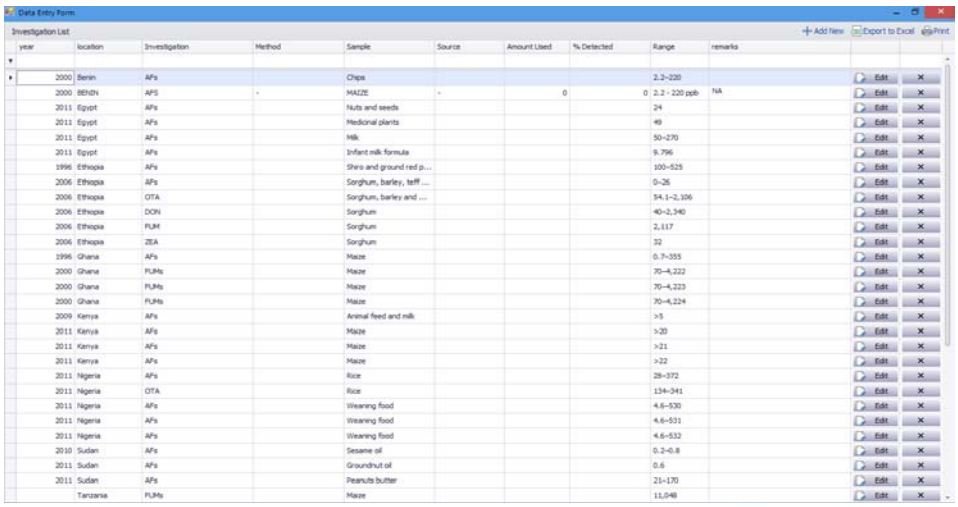


- Make necessary inputs

- Click the Add button, when done

- Corfirm the new entry on the grid view list

Table 7: The view of the captured data in the database



IV. Benefits

The automated database system provides the client significant time and cost savings, improved its ability to analyze data, and aided decisions in the issue of mycotoxins in crops.

As a result, the solution offered the following benefits:

- Robust data collection, management, and analysis methods
- A secure and protected data environment
- Reduction in errors caused by insufficient or inconsistent data
- Data storage to facilitate regulatory compliance
- Data storage practices that offer scalability and reduced testing rework
- Improved maintainability, standardization, control, predictability, and traceability of data
- Enhanced decision-making capabilities in choosing better variety of crop seed and better quality end products
- Better audit and control procedures
- Lower costs due to automation of labor intensive tasks and elimination of redundant work
- Overall cost reduction due to streamlining of the traditional way of data sourcing.
- Reduced time-to-conclude analysis of mycotoxins in agricultural products

V. Conclusion

The presence of mycotoxins in grains and other staple foods and feedstuffs has serious implications for human and animal health. Many countries have enacted regulations stipulating maximum amounts of mycotoxins permissible in food and feedstuffs. Most developed countries will not permit the import of commodities containing amounts of mycotoxins above specified limits. Mycotoxins therefore have implications for trade between nations. Prevention of fungal invasion of commodities is by far the most effective method of avoiding mycotoxin problems. The role Information Technology cannot be over emphasized in this matter. For accuracy in monitoring and management of mycotoxins and its related diseases, a pool of data must be in place. Therefore, the automated database system will be of great help which will lead to a sustainable development.

VI. Recommendation

- ❖ The Mycotoxins association should organise training programme to create awareness of the automated database system to Sub-Saharan Africa.
- ❖ Government should fund the research of mycotoxin, for this is a necessity for life security. Since this involves crops and feedstuff, the lives of animals and humans need to be assured. This is also major case when we talk of sustainability development.
- ❖ The ICT centers should be made available for research officers and the farmers to ensure communication and proper interaction via the database.

References Références Referencias

1. S..Bankole 2006. Institute of Animal Nutrition, and university of Hehenhein, Emi-wolf-str 10,70599stuttgart, Germany.
2. Abarca, M. L., M. R. Bragulat, G. Sastella, and F. J. Cabanes. 1994. Ochratoxin A production by strains of Aspergillus niger var. niger. Appl. Environ. Microbiol. 60:2650-2652. [PMC free article] [PubMed]
3. Abramson, D., E. Usleber, and E. Marlbauer. 2001. Immunochemical method for citrinin. p. 195-204. In M. W. Trucksess and A. F. Pohland (ed.), Mycotoxin protocols. Humana Press, Totowa, N.J.
4. American Academy of Pediatrics. 1998. Toxic effects of indoor molds. Pediatrics 101:712-714. [PubMed]
5. Anderson, S. J. 1995. Compositional changes in surface mycoflora during ripening of naturally fermented sausages. J. Food Protect. 58:426-429.
6. Barrett, J. 2000. Mycotoxins: of molds and maladies. Environ. Health Perspect. 108:A20-A23. [PMC free article] [PubMed]
7. Bayman, P., J. L. Baker, M. A. Doster, T. J. Michailides and, N. E. Mahoney. 2002. Ochratoxin production by the Aspergillus ochraceus group and Aspergillus alliaceus. Appl. Environ. Microbiol. 68:2326-2329. [PMC free article] [PubMed]
8. Kahn J.M., Katz R.H., Pister K.S.J. Next Century Challenges: Mobile Networking for “Smart Dust”. Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking; Seattle, WA, USA. August 1999; pp. 271–278.
9. Akyildiz I.F., Su W., Sankarasubramaniam Y., Cayirci E. Wireless sensor networks: a survey. Comput. Netw. 2002;38:393–422.
10. Fukatsu T., Hirafuji M. Field monitoring using sensor-nodes with a Web server. J. Rob. Mechatron. 2005;17:164–172.
11. Otuka A., Sugawara K. A labor management application using handheld computers. Agric. Inf. Res. 2003;12:95–104.
12. Bange M.P., Deutscher S.A., Larsen D., Linsley D., Whiteside S. A handheld decision support system to facilitate improved insect pest management in Australian cotton systems. Comput. Electron. Agric. 2004;43:131–147.
13. Yokoyama K. Promoting the Good Agricultural Practice Movement through Interactive and Seamless Communication based on User-friendly Mobile Information Technology. Proceedings of the

International Seminar on Technology Development for Good Agricultural Practice in Asia and Oceania; Tsukuba, Japan. October 2005; pp. 213–219.

Global Journals Guidelines Handbook 2022

www.GlobalJournals.org

Memberships

Fellows/Associates of Computer Science Research Council

FCSRC/ACSRC Memberships

Introduction

FCSRC/ACSRC is the most prestigious membership of Global Journals accredited by Open Association of Research Society, U.S.A (OARS). The credentials of Fellow and Associate designations signify that the researcher has gained the knowledge of the fundamental and high-level concepts, and is a subject matter expert, proficient in an expertise course covering the professional code of conduct, and follows recognized standards of practice. The credentials are designated only to the researchers, scientists, and professionals that have been selected by a rigorous process by our Editorial Board and Management Board.

Associates of FCSRC/ACSRC are scientists and researchers from around the world are working on projects/researches that have huge potentials. Members support Global Journals' mission to advance technology for humanity and the profession.

FCSRC

Fellow of Computer Science Research Council

FELLOW OF COMPUTER SCIENCE RESEARCH COUNCIL is the most prestigious membership of Global Journals. It is an award and membership granted to individuals that the Open Association of Research Society judges to have made a 'substantial contribution to the improvement of computer science, technology, and electronics engineering.

The primary objective is to recognize the leaders in research and scientific fields of the current era with a global perspective and to create a channel between them and other researchers for better exposure and knowledge sharing. Members are most eminent scientists, engineers, and technologists from all across the world. Fellows are elected for life through a peer review process on the basis of excellence in the respective domain. There is no limit on the number of new nominations made in any year. Each year, the Open Association of Research Society elect up to 12 new Fellow Members.

Benefit

To the institution

Get letter of appreciation

Global Journals sends a letter of appreciation of author to the Dean or CEO of the University or Company of which author is a part, signed by editor in chief or chief author.

Exclusive Network

Get access to a closed network

A FCSRC member gets access to a closed network of Tier 1 researchers and scientists with direct communication channel through our website. Fellows can reach out to other members or researchers directly.They should also be open to reaching out by other.

Career | Credibility | Exclusive | Reputation

Certificate

Certificate, LoR and Laser-Momento

Fellows receive a printed copy of a certificate signed by our Chief Author that may be used for academic purposes and a personal recommendation letter to the dean of member's university.

Career | Credibility | Exclusive | Reputation

Designation

Get honored title of membership

Fellows can use the honored title of membership. The “FCSRC” is an honored title which is accorded to a person’s name viz. Dr. John E. Hall, Ph.D., FCSRC or William Walldroff, M.S., FCSRC.

Career | Credibility | Exclusive | Reputation

Recognition on the Platform

Better visibility and citation

All the Fellow members of FCSRC get a badge of "Leading Member of Global Journals" on the Research Community that distinguishes them from others. Additionally, the profile is also partially maintained by our team for better visibility and citation. All fellows get a dedicated page on the website with their biography.

Career | Credibility | Reputation

Future Work

Get discounts on the future publications

Fellows receive discounts on future publications with Global Journals up to 60%. Through our recommendation programs, members also receive discounts on publications made with OARS affiliated organizations.

Career | Financial

GJ Account

Unlimited forward of Emails

Fellows get secure and fast GJ work emails with unlimited forward of emails that they may use them as their primary email. For example, john [AT] globaljournals [DOT] org.

Career | Credibility | Reputation

Premium Tools

Access to all the premium tools

To take future researches to the zenith, fellows receive access to all the premium tools that Global Journals have to offer along with the partnership with some of the best marketing leading tools out there.

Financial

Conferences & Events

Organize seminar/conference

Fellows are authorized to organize symposium/seminar/conference on behalf of Global Journal Incorporation (USA). They can also participate in the same organized by another institution as representative of Global Journal. In both the cases, it is mandatory for him to discuss with us and obtain our consent. Additionally, they get free research conferences (and others) alerts.

Career | Credibility | Financial

Early Invitations

Early invitations to all the symposiums, seminars, conferences

All fellows receive the early invitations to all the symposiums, seminars, conferences and webinars hosted by Global Journals in their subject.

Exclusive

Publishing Articles & Books

Earn 60% of sales proceeds

Fellows can publish articles (limited) without any fees. Also, they can earn up to 70% of sales proceeds from the sale of reference/review books/literature/publishing of research paper. The FCSRC member can decide its price and we can help in making the right decision.

Exclusive | Financial

Reviewers

Get a remuneration of 15% of author fees

Fellow members are eligible to join as a paid peer reviewer at Global Journals Incorporation (USA) and can get a remuneration of 15% of author fees, taken from the author of a respective paper.

Financial

Access to Editorial Board

Become a member of the Editorial Board

Fellows may join as a member of the Editorial Board of Global Journals Incorporation (USA) after successful completion of three years as Fellow and as Peer Reviewer. Additionally, Fellows get a chance to nominate other members for Editorial Board.

Career | Credibility | Exclusive | Reputation

And Much More

Get access to scientific museums and observatories across the globe

All members get access to 5 selected scientific museums and observatories across the globe. All researches published with Global Journals will be kept under deep archival facilities across regions for future protections and disaster recovery. They get 10 GB free secure cloud access for storing research files.

ASSOCIATE OF COMPUTER SCIENCE RESEARCH COUNCIL

ASSOCIATE OF COMPUTER SCIENCE RESEARCH COUNCIL is the membership of Global Journals awarded to individuals that the Open Association of Research Society judges to have made a 'substantial contribution to the improvement of computer science, technology, and electronics engineering.

The primary objective is to recognize the leaders in research and scientific fields of the current era with a global perspective and to create a channel between them and other researchers for better exposure and knowledge sharing. Members are most eminent scientists, engineers, and technologists from all across the world. Associate membership can later be promoted to Fellow Membership. Associates are elected for life through a peer review process on the basis of excellence in the respective domain. There is no limit on the number of new nominations made in any year. Each year, the Open Association of Research Society elect up to 12 new Associate Members.

Benefit

To the institution

Get letter of appreciation

Global Journals sends a letter of appreciation of author to the Dean or CEO of the University or Company of which author is a part, signed by editor in chief or chief author.

Exclusive Network

Get access to a closed network

A ACSRC member gets access to a closed network of Tier 2 researchers and scientists with direct communication channel through our website. Associates can reach out to other members or researchers directly.They should also be open to reaching out by other.

Career | Credibility | Exclusive | Reputation

Certificate

Certificate, LoR and Laser-Momento

Associates receive a printed copy of a certificate signed by our Chief Author that may be used for academic purposes and a personal recommendation letter to the dean of member's university.

Career | Credibility | Exclusive | Reputation

Designation

Get honored title of membership

Associates can use the honored title of membership. The “ACSRC” is an honored title which is accorded to a person’s name viz. Dr. John E. Hall, Ph.D., ACSRC or William Walldroff, M.S., ACSRC.

Career | Credibility | Exclusive | Reputation

Recognition on the Platform

Better visibility and citation

All the Associate members of ACSRC get a badge of "Leading Member of Global Journals" on the Research Community that distinguishes them from others. Additionally, the profile is also partially maintained by our team for better visibility and citation.

Career | Credibility | Reputation

Future Work

Get discounts on the future publications

Associates receive discounts on future publications with Global Journals up to 30%. Through our recommendation programs, members also receive discounts on publications made with OARS affiliated organizations.

Career | Financial

GJ Account

Unlimited forward of Emails

Associates get secure and fast GJ work emails with 5GB forward of emails that they may use them as their primary email. For example, john [AT] globaljournals [DOT] org.

Career | Credibility | Reputation

Premium Tools

Access to all the premium tools

To take future researches to the zenith, associates receive access to all the premium tools that Global Journals have to offer along with the partnership with some of the best marketing leading tools out there.

Financial

Conferences & Events

Organize seminar/conference

Associates are authorized to organize symposium/seminar/conference on behalf of Global Journal Incorporation (USA). They can also participate in the same organized by another institution as representative of Global Journal. In both the cases, it is mandatory for him to discuss with us and obtain our consent. Additionally, they get free research conferences (and others) alerts.

Career | Credibility | Financial

Early Invitations

Early invitations to all the symposiums, seminars, conferences

All associates receive the early invitations to all the symposiums, seminars, conferences and webinars hosted by Global Journals in their subject.

Exclusive

Publishing Articles & Books

Earn 30-40% of sales proceeds

Associates can publish articles (limited) without any fees. Also, they can earn up to 30-40% of sales proceeds from the sale of reference/review books/literature/publishing of research paper.

Exclusive | Financial

Reviewers

Get a remuneration of 15% of author fees

Associate members are eligible to join as a paid peer reviewer at Global Journals Incorporation (USA) and can get a remuneration of 15% of author fees, taken from the author of a respective paper.

Financial

And Much More

Get access to scientific museums and observatories across the globe

All members get access to 2 selected scientific museums and observatories across the globe. All researches published with Global Journals will be kept under deep archival facilities across regions for future protections and disaster recovery. They get 5 GB free secure cloud access for storing research files.

Associate	Fellow	Research Group	Basic
$4800 **lifetime designation**	$6800 **lifetime designation**	$12500.00 **organizational**	APC **per article**
Certificate, LoR and Momento **2** discounted publishing/year **Gradation** of Research **10** research contacts/day **1 GB** Cloud Storage **GJ** Community Access	**Certificate,** LoR and Momento **Unlimited** discounted publishing/year **Gradation** of Research **Unlimited** research contacts/day **5 GB** Cloud Storage **Online Presense** Assistance **GJ** Community Access	**Certificates,** LoRs and Momentos **Unlimited** free publishing/year **Gradation** of Research **Unlimited** research contacts/day **Unlimited** Cloud Storage **Online Presense** Assistance **GJ** Community Access	**GJ** Community Access

Preferred Author Guidelines

We accept the manuscript submissions in any standard (generic) format.

We typeset manuscripts using advanced typesetting tools like Adobe In Design, CorelDraw, TeXnicCenter, and TeXStudio. We usually recommend authors submit their research using any standard format they are comfortable with, and let Global Journals do the rest.

Alternatively, you can download our basic template from https://globaljournals.org/Template.zip

Authors should submit their complete paper/article, including text illustrations, graphics, conclusions, artwork, and tables. Authors who are not able to submit manuscript using the form above can email the manuscript department at submit@globaljournals.org or get in touch with chiefeditor@globaljournals.org if they wish to send the abstract before submission.

Before and during Submission

Authors must ensure the information provided during the submission of a paper is authentic. Please go through the following checklist before submitting:

1. Authors must go through the complete author guideline and understand and *agree to Global Journals' ethics and code of conduct,* along with author responsibilities.
2. Authors must accept the privacy policy, terms, and conditions of Global Journals.
3. Ensure corresponding author's email address and postal address are accurate and reachable.
4. Manuscript to be submitted must include keywords, an abstract, a paper title, co-author(s') names and details (email address, name, phone number, and institution), figures and illustrations in vector format including appropriate captions, tables, including titles and footnotes, a conclusion, results, acknowledgments and references.
5. Authors should submit paper in a ZIP archive if any supplementary files are required along with the paper.
6. Proper permissions must be acquired for the use of any copyrighted material.
7. Manuscript submitted *must not have been submitted or published elsewhere* and all authors must be aware of the submission.

Declaration of Conflicts of Interest

It is required for authors to declare all financial, institutional, and personal relationships with other individuals and organizations that could influence (bias) their research.

Policy on Plagiarism

Plagiarism is not acceptable in Global Journals submissions at all.

Plagiarized content will not be considered for publication. We reserve the right to inform authors' institutions about plagiarism detected either before or after publication. If plagiarism is identified, we will follow COPE guidelines:

Authors are solely responsible for all the plagiarism that is found. The author must not fabricate, falsify or plagiarize existing research data. The following, if copied, will be considered plagiarism:

- Words (language)
- Ideas
- Findings
- Writings
- Diagrams
- Graphs
- Illustrations
- Lectures

- Printed material
- Graphic representations
- Computer programs
- Electronic material
- Any other original work

Authorship Policies

Global Journals follows the definition of authorship set up by the Open Association of Research Society, USA. According to its guidelines, authorship criteria must be based on:

1. Substantial contributions to the conception and acquisition of data, analysis, and interpretation of findings.
2. Drafting the paper and revising it critically regarding important academic content.
3. Final approval of the version of the paper to be published.

Changes in Authorship

The corresponding author should mention the name and complete details of all co-authors during submission and in manuscript. We support addition, rearrangement, manipulation, and deletions in authors list till the early view publication of the journal. We expect that corresponding author will notify all co-authors of submission. We follow COPE guidelines for changes in authorship.

Copyright

During submission of the manuscript, the author is confirming an exclusive license agreement with Global Journals which gives Global Journals the authority to reproduce, reuse, and republish authors' research. We also believe in flexible copyright terms where copyright may remain with authors/employers/institutions as well. Contact your editor after acceptance to choose your copyright policy. You may follow this form for copyright transfers.

Appealing Decisions

Unless specified in the notification, the Editorial Board's decision on publication of the paper is final and cannot be appealed before making the major change in the manuscript.

Acknowledgments

Contributors to the research other than authors credited should be mentioned in Acknowledgments. The source of funding for the research can be included. Suppliers of resources may be mentioned along with their addresses.

Declaration of funding sources

Global Journals is in partnership with various universities, laboratories, and other institutions worldwide in the research domain. Authors are requested to disclose their source of funding during every stage of their research, such as making analysis, performing laboratory operations, computing data, and using institutional resources, from writing an article to its submission. This will also help authors to get reimbursements by requesting an open access publication letter from Global Journals and submitting to the respective funding source.

Preparing your Manuscript

Authors can submit papers and articles in an acceptable file format: MS Word (doc, docx), LaTeX (.tex, .zip or .rar including all of your files), Adobe PDF (.pdf), rich text format (.rtf), simple text document (.txt), Open Document Text (.odt), and Apple Pages (.pages). Our professional layout editors will format the entire paper according to our official guidelines. This is one of the highlights of publishing with Global Journals—authors should not be concerned about the formatting of their paper. Global Journals accepts articles and manuscripts in every major language, be it Spanish, Chinese, Japanese, Portuguese, Russian, French, German, Dutch, Italian, Greek, or any other national language, but the title, subtitle, and abstract should be in English. This will facilitate indexing and the pre-peer review process.

The following is the official style and template developed for publication of a research paper. Authors are not required to follow this style during the submission of the paper. It is just for reference purposes.

Manuscript Style Instruction (Optional)

- Microsoft Word Document Setting Instructions.
- Font type of all text should be Swis721 Lt BT.
- Page size: 8.27" x 11'", left margin: 0.65, right margin: 0.65, bottom margin: 0.75.
- Paper title should be in one column of font size 24.
- Author name in font size of 11 in one column.
- Abstract: font size 9 with the word "Abstract" in bold italics.
- Main text: font size 10 with two justified columns.
- Two columns with equal column width of 3.38 and spacing of 0.2.
- First character must be three lines drop-capped.
- The paragraph before spacing of 1 pt and after of 0 pt.
- Line spacing of 1 pt.
- Large images must be in one column.
- The names of first main headings (Heading 1) must be in Roman font, capital letters, and font size of 10.
- The names of second main headings (Heading 2) must not include numbers and must be in italics with a font size of 10.

Structure and Format of Manuscript

The recommended size of an original research paper is under 15,000 words and review papers under 7,000 words. Research articles should be less than 10,000 words. Research papers are usually longer than review papers. Review papers are reports of significant research (typically less than 7,000 words, including tables, figures, and references)

A research paper must include:

a) A title which should be relevant to the theme of the paper.
b) A summary, known as an abstract (less than 150 words), containing the major results and conclusions.
c) Up to 10 keywords that precisely identify the paper's subject, purpose, and focus.
d) An introduction, giving fundamental background objectives.
e) Resources and techniques with sufficient complete experimental details (wherever possible by reference) to permit repetition, sources of information must be given, and numerical methods must be specified by reference.
f) Results which should be presented concisely by well-designed tables and figures.
g) Suitable statistical data should also be given.
h) All data must have been gathered with attention to numerical detail in the planning stage.

Design has been recognized to be essential to experiments for a considerable time, and the editor has decided that any paper that appears not to have adequate numerical treatments of the data will be returned unrefereed.

i) Discussion should cover implications and consequences and not just recapitulate the results; conclusions should also be summarized.
j) There should be brief acknowledgments.
k) There ought to be references in the conventional format. Global Journals recommends APA format.

Authors should carefully consider the preparation of papers to ensure that they communicate effectively. Papers are much more likely to be accepted if they are carefully designed and laid out, contain few or no errors, are summarizing, and follow instructions. They will also be published with much fewer delays than those that require much technical and editorial correction.

The Editorial Board reserves the right to make literary corrections and suggestions to improve brevity.

Format Structure

It is necessary that authors take care in submitting a manuscript that is written in simple language and adheres to published guidelines.

All manuscripts submitted to Global Journals should include:

Title

The title page must carry an informative title that reflects the content, a running title (less than 45 characters together with spaces), names of the authors and co-authors, and the place(s) where the work was carried out.

Author details

The full postal address of any related author(s) must be specified.

Abstract

The abstract is the foundation of the research paper. It should be clear and concise and must contain the objective of the paper and inferences drawn. It is advised to not include big mathematical equations or complicated jargon.

Many researchers searching for information online will use search engines such as Google, Yahoo or others. By optimizing your paper for search engines, you will amplify the chance of someone finding it. In turn, this will make it more likely to be viewed and cited in further works. Global Journals has compiled these guidelines to facilitate you to maximize the web-friendliness of the most public part of your paper.

Keywords

A major lynchpin of research work for the writing of research papers is the keyword search, which one will employ to find both library and internet resources. Up to eleven keywords or very brief phrases have to be given to help data retrieval, mining, and indexing.

One must be persistent and creative in using keywords. An effective keyword search requires a strategy: planning of a list of possible keywords and phrases to try.

Choice of the main keywords is the first tool of writing a research paper. Research paper writing is an art. Keyword search should be as strategic as possible.

One should start brainstorming lists of potential keywords before even beginning searching. Think about the most important concepts related to research work. Ask, “What words would a source have to include to be truly valuable in a research paper?” Then consider synonyms for the important words.

It may take the discovery of only one important paper to steer in the right keyword direction because, in most databases, the keywords under which a research paper is abstracted are listed with the paper.

Numerical Methods

Numerical methods used should be transparent and, where appropriate, supported by references.

Abbreviations

Authors must list all the abbreviations used in the paper at the end of the paper or in a separate table before using them.

Formulas and equations

Authors are advised to submit any mathematical equation using either MathJax, KaTeX, or LaTeX, or in a very high-quality image.

Tables, Figures, and Figure Legends

Tables: Tables should be cautiously designed, uncrowned, and include only essential data. Each must have an Arabic number, e.g., Table 4, a self-explanatory caption, and be on a separate sheet. Authors must submit tables in an editable format and not as images. References to these tables (if any) must be mentioned accurately.

Figures

Figures are supposed to be submitted as separate files. Always include a citation in the text for each figure using Arabic numbers, e.g., Fig. 4. Artwork must be submitted online in vector electronic form or by emailing it.

Preparation of Eletronic Figures for Publication

Although low-quality images are sufficient for review purposes, print publication requires high-quality images to prevent the final product being blurred or fuzzy. Submit (possibly by e-mail) EPS (line art) or TIFF (halftone/ photographs) files only. MS PowerPoint and Word Graphics are unsuitable for printed pictures. Avoid using pixel-oriented software. Scans (TIFF only) should have a resolution of at least 350 dpi (halftone) or 700 to 1100 dpi (line drawings). Please give the data for figures in black and white or submit a Color Work Agreement form. EPS files must be saved with fonts embedded (and with a TIFF preview, if possible).

For scanned images, the scanning resolution at final image size ought to be as follows to ensure good reproduction: line art: >650 dpi; halftones (including gel photographs): >350 dpi; figures containing both halftone and line images: >650 dpi.

Color charges: Authors are advised to pay the full cost for the reproduction of their color artwork. Hence, please note that if there is color artwork in your manuscript when it is accepted for publication, we would require you to complete and return a Color Work Agreement form before your paper can be published. Also, you can email your editor to remove the color fee after acceptance of the paper.

Tips for writing a good quality Computer Science Research Paper

Techniques for writing a good quality computer science research paper:

1. Choosing the topic: In most cases, the topic is selected by the interests of the author, but it can also be suggested by the guides. You can have several topics, and then judge which you are most comfortable with. This may be done by asking several questions of yourself, like "Will I be able to carry out a search in this area? Will I find all necessary resources to accomplish the search? Will I be able to find all information in this field area?" If the answer to this type of question is "yes," then you ought to choose that topic. In most cases, you may have to conduct surveys and visit several places. Also, you might have to do a lot of work to find all the rises and falls of the various data on that subject. Sometimes, detailed information plays a vital role, instead of short information. Evaluators are human: The first thing to remember is that evaluators are also human beings. They are not only meant for rejecting a paper. They are here to evaluate your paper. So present your best aspect.

2. Think like evaluators: If you are in confusion or getting demotivated because your paper may not be accepted by the evaluators, then think, and try to evaluate your paper like an evaluator. Try to understand what an evaluator wants in your research paper, and you will automatically have your answer. Make blueprints of paper: The outline is the plan or framework that will help you to arrange your thoughts. It will make your paper logical. But remember that all points of your outline must be related to the topic you have chosen.

3. Ask your guides: If you are having any difficulty with your research, then do not hesitate to share your difficulty with your guide (if you have one). They will surely help you out and resolve your doubts. If you can't clarify what exactly you require for your work, then ask your supervisor to help you with an alternative. He or she might also provide you with a list of essential readings.

4. Use of computer is recommended: As you are doing research in the field of computer science then this point is quite obvious. Use right software: Always use good quality software packages. If you are not capable of judging good software, then you can lose the quality of your paper unknowingly. There are various programs available to help you which you can get through the internet.

5. Use the internet for help: An excellent start for your paper is using Google. It is a wondrous search engine, where you can have your doubts resolved. You may also read some answers for the frequent question of how to write your research paper or find a model research paper. You can download books from the internet. If you have all the required books, place importance on reading, selecting, and analyzing the specified information. Then sketch out your research paper. Use big pictures: You may use encyclopedias like Wikipedia to get pictures with the best resolution. At Global Journals, you should strictly follow here.

6. Bookmarks are useful: When you read any book or magazine, you generally use bookmarks, right? It is a good habit which helps to not lose your continuity. You should always use bookmarks while searching on the internet also, which will make your search easier.

7. Revise what you wrote: When you write anything, always read it, summarize it, and then finalize it.

8. Make every effort: Make every effort to mention what you are going to write in your paper. That means always have a good start. Try to mention everything in the introduction—what is the need for a particular research paper. Polish your work with good writing skills and always give an evaluator what he wants. Make backups: When you are going to do any important thing like making a research paper, you should always have backup copies of it either on your computer or on paper. This protects you from losing any portion of your important data.

9. Produce good diagrams of your own: Always try to include good charts or diagrams in your paper to improve quality. Using several unnecessary diagrams will degrade the quality of your paper by creating a hodgepodge. So always try to include diagrams which were made by you to improve the readability of your paper. Use of direct quotes: When you do research relevant to literature, history, or current affairs, then use of quotes becomes essential, but if the study is relevant to science, use of quotes is not preferable.

10.Use proper verb tense: Use proper verb tenses in your paper. Use past tense to present those events that have happened. Use present tense to indicate events that are going on. Use future tense to indicate events that will happen in the future. Use of wrong tenses will confuse the evaluator. Avoid sentences that are incomplete.

11. Pick a good study spot: Always try to pick a spot for your research which is quiet. Not every spot is good for studying.

12. Know what you know: Always try to know what you know by making objectives, otherwise you will be confused and unable to achieve your target.

13. Use good grammar: Always use good grammar and words that will have a positive impact on the evaluator; use of good vocabulary does not mean using tough words which the evaluator has to find in a dictionary. Do not fragment sentences. Eliminate one-word sentences. Do not ever use a big word when a smaller one would suffice.

Verbs have to be in agreement with their subjects. In a research paper, do not start sentences with conjunctions or finish them with prepositions. When writing formally, it is advisable to never split an infinitive because someone will (wrongly) complain. Avoid clichés like a disease. Always shun irritating alliteration. Use language which is simple and straightforward. Put together a neat summary.

14. Arrangement of information: Each section of the main body should start with an opening sentence, and there should be a changeover at the end of the section. Give only valid and powerful arguments for your topic. You may also maintain your arguments with records.

15. Never start at the last minute: Always allow enough time for research work. Leaving everything to the last minute will degrade your paper and spoil your work.

16. Multitasking in research is not good: Doing several things at the same time is a bad habit in the case of research activity. Research is an area where everything has a particular time slot. Divide your research work into parts, and do a particular part in a particular time slot.

17. Never copy others' work: Never copy others' work and give it your name because if the evaluator has seen it anywhere, you will be in trouble. Take proper rest and food: No matter how many hours you spend on your research activity, if you are not taking care of your health, then all your efforts will have been in vain. For quality research, take proper rest and food.

18. Go to seminars: Attend seminars if the topic is relevant to your research area. Utilize all your resources.

19. Refresh your mind after intervals: Try to give your mind a rest by listening to soft music or sleeping in intervals. This will also improve your memory. Acquire colleagues: Always try to acquire colleagues. No matter how sharp you are, if you acquire colleagues, they can give you ideas which will be helpful to your research.

20. Think technically: Always think technically. If anything happens, search for its reasons, benefits, and demerits. Think and then print: When you go to print your paper, check that tables are not split, headings are not detached from their descriptions, and page sequence is maintained.

21. Adding unnecessary information: Do not add unnecessary information like "I have used MS Excel to draw graphs." Irrelevant and inappropriate material is superfluous. Foreign terminology and phrases are not apropos. One should never take a broad view. Analogy is like feathers on a snake. Use words properly, regardless of how others use them. Remove quotations. Puns are for kids, not grunt readers. Never oversimplify: When adding material to your research paper, never go for oversimplification; this will definitely irritate the evaluator. Be specific. Never use rhythmic redundancies. Contractions shouldn't be used in a research paper. Comparisons are as terrible as clichés. Give up ampersands, abbreviations, and so on. Remove commas that are not necessary. Parenthetical words should be between brackets or commas. Understatement is always the best way to put forward earth-shaking thoughts. Give a detailed literary review.

22. Report concluded results: Use concluded results. From raw data, filter the results, and then conclude your studies based on measurements and observations taken. An appropriate number of decimal places should be used. Parenthetical remarks are prohibited here. Proofread carefully at the final stage. At the end, give an outline to your arguments. Spot perspectives of further study of the subject. Justify your conclusion at the bottom sufficiently, which will probably include examples.

23. Upon conclusion: Once you have concluded your research, the next most important step is to present your findings. Presentation is extremely important as it is the definite medium though which your research is going to be in print for the rest of the crowd. Care should be taken to categorize your thoughts well and present them in a logical and neat manner. A good quality research paper format is essential because it serves to highlight your research paper and bring to light all necessary aspects of your research.

Informal Guidelines of Research Paper Writing

Key points to remember:

- Submit all work in its final form.
- Write your paper in the form which is presented in the guidelines using the template.
- Please note the criteria peer reviewers will use for grading the final paper.

Final points:

One purpose of organizing a research paper is to let people interpret your efforts selectively. The journal requires the following sections, submitted in the order listed, with each section starting on a new page:

The introduction: This will be compiled from reference matter and reflect the design processes or outline of basis that directed you to make a study. As you carry out the process of study, the method and process section will be constructed like that. The results segment will show related statistics in nearly sequential order and direct reviewers to similar intellectual paths throughout the data that you gathered to carry out your study.

The discussion section:

This will provide understanding of the data and projections as to the implications of the results. The use of good quality references throughout the paper will give the effort trustworthiness by representing an alertness to prior workings.

Writing a research paper is not an easy job, no matter how trouble-free the actual research or concept. Practice, excellent preparation, and controlled record-keeping are the only means to make straightforward progression.

General style:

Specific editorial column necessities for compliance of a manuscript will always take over from directions in these general guidelines.

To make a paper clear: Adhere to recommended page limits.

Mistakes to avoid:

- Insertion of a title at the foot of a page with subsequent text on the next page.
- Separating a table, chart, or figure—confine each to a single page.
- Submitting a manuscript with pages out of sequence.
- In every section of your document, use standard writing style, including articles ("a" and "the").
- Keep paying attention to the topic of the paper.
- Use paragraphs to split each significant point (excluding the abstract).
- Align the primary line of each section.
- Present your points in sound order.
- Use present tense to report well-accepted matters.
- Use past tense to describe specific results.
- Do not use familiar wording; don't address the reviewer directly. Don't use slang or superlatives.
- Avoid use of extra pictures—include only those figures essential to presenting results.

Title page:

Choose a revealing title. It should be short and include the name(s) and address(es) of all authors. It should not have acronyms or abbreviations or exceed two printed lines.

Abstract: This summary should be two hundred words or less. It should clearly and briefly explain the key findings reported in the manuscript and must have precise statistics. It should not have acronyms or abbreviations. It should be logical in itself. Do not cite references at this point.

An abstract is a brief, distinct paragraph summary of finished work or work in development. In a minute or less, a reviewer can be taught the foundation behind the study, common approaches to the problem, relevant results, and significant conclusions or new questions.

Write your summary when your paper is completed because how can you write the summary of anything which is not yet written? Wealth of terminology is very essential in abstract. Use comprehensive sentences, and do not sacrifice readability for brevity; you can maintain it succinctly by phrasing sentences so that they provide more than a lone rationale. The author can at this moment go straight to shortening the outcome. Sum up the study with the subsequent elements in any summary. Try to limit the initial two items to no more than one line each.

Reason for writing the article—theory, overall issue, purpose.

- Fundamental goal.
- To-the-point depiction of the research.
- Consequences, including definite statistics—if the consequences are quantitative in nature, account for this; results of any numerical analysis should be reported. Significant conclusions or questions that emerge from the research.

Approach:

- Single section and succinct.
- An outline of the job done is always written in past tense.
- Concentrate on shortening results—limit background information to a verdict or two.
- Exact spelling, clarity of sentences and phrases, and appropriate reporting of quantities (proper units, important statistics) are just as significant in an abstract as they are anywhere else.

Introduction:

The introduction should "introduce" the manuscript. The reviewer should be presented with sufficient background information to be capable of comprehending and calculating the purpose of your study without having to refer to other works. The basis for the study should be offered. Give the most important references, but avoid making a comprehensive appraisal of the topic. Describe the problem visibly. If the problem is not acknowledged in a logical, reasonable way, the reviewer will give no attention to your results. Speak in common terms about techniques used to explain the problem, if needed, but do not present any particulars about the protocols here.

The following approach can create a valuable beginning:

- Explain the value (significance) of the study.
- Defend the model—why did you employ this particular system or method? What is its compensation? Remark upon its appropriateness from an abstract point of view as well as pointing out sensible reasons for using it.
- Present a justification. State your particular theory(-ies) or aim(s), and describe the logic that led you to choose them.
- Briefly explain the study's tentative purpose and how it meets the declared objectives.

Approach:

Use past tense except for when referring to recognized facts. After all, the manuscript will be submitted after the entire job is done. Sort out your thoughts; manufacture one key point for every section. If you make the four points listed above, you will need at least four paragraphs. Present surrounding information only when it is necessary to support a situation. The reviewer does not desire to read everything you know about a topic. Shape the theory specifically—do not take a broad view.

As always, give awareness to spelling, simplicity, and correctness of sentences and phrases.

Procedures (methods and materials):

This part is supposed to be the easiest to carve if you have good skills. A soundly written procedures segment allows a capable scientist to replicate your results. Present precise information about your supplies. The suppliers and clarity of reagents can be helpful bits of information. Present methods in sequential order, but linked methodologies can be grouped as a segment. Be concise when relating the protocols. Attempt to give the least amount of information that would permit another capable scientist to replicate your outcome, but be cautious that vital information is integrated. The use of subheadings is suggested and ought to be synchronized with the results section.

When a technique is used that has been well-described in another section, mention the specific item describing the way, but draw the basic principle while stating the situation. The purpose is to show all particular resources and broad procedures so that another person may use some or all of the methods in one more study or referee the scientific value of your work. It is not to be a step-by-step report of the whole thing you did, nor is a methods section a set of orders.

Materials:

Materials may be reported in part of a section or else they may be recognized along with your measures.

Methods:

- Report the method and not the particulars of each process that engaged the same methodology.
- Describe the method entirely.
- To be succinct, present methods under headings dedicated to specific dealings or groups of measures.
- Simplify—detail how procedures were completed, not how they were performed on a particular day.
- If well-known procedures were used, account for the procedure by name, possibly with a reference, and that's all.

Approach:

It is embarrassing to use vigorous voice when documenting methods without using first person, which would focus the reviewer's interest on the researcher rather than the job. As a result, when writing up the methods, most authors use third person passive voice.

Use standard style in this and every other part of the paper—avoid familiar lists, and use full sentences.

What to keep away from:

- Resources and methods are not a set of information.
- Skip all descriptive information and surroundings—save it for the argument.
- Leave out information that is immaterial to a third party.

Results:

The principle of a results segment is to present and demonstrate your conclusion. Create this part as entirely objective details of the outcome, and save all understanding for the discussion.

The page length of this segment is set by the sum and types of data to be reported. Use statistics and tables, if suitable, to present consequences most efficiently.

You must clearly differentiate material which would usually be incorporated in a study editorial from any unprocessed data or additional appendix matter that would not be available. In fact, such matters should not be submitted at all except if requested by the instructor.

Content:

- Sum up your conclusions in text and demonstrate them, if suitable, with figures and tables.
- In the manuscript, explain each of your consequences, and point the reader to remarks that are most appropriate.
- Present a background, such as by describing the question that was addressed by creation of an exacting study.
- Explain results of control experiments and give remarks that are not accessible in a prescribed figure or table, if appropriate.
- Examine your data, then prepare the analyzed (transformed) data in the form of a figure (graph), table, or manuscript.

What to stay away from:

- Do not discuss or infer your outcome, report surrounding information, or try to explain anything.
- Do not include raw data or intermediate calculations in a research manuscript.
- Do not present similar data more than once.
- A manuscript should complement any figures or tables, not duplicate information.
- Never confuse figures with tables—there is a difference.

Approach:

As always, use past tense when you submit your results, and put the whole thing in a reasonable order.

Put figures and tables, appropriately numbered, in order at the end of the report.

If you desire, you may place your figures and tables properly within the text of your results section.

Figures and tables:

If you put figures and tables at the end of some details, make certain that they are visibly distinguished from any attached appendix materials, such as raw facts. Whatever the position, each table must be titled, numbered one after the other, and include a heading. All figures and tables must be divided from the text.

Discussion:

The discussion is expected to be the trickiest segment to write. A lot of papers submitted to the journal are discarded based on problems with the discussion. There is no rule for how long an argument should be.

Position your understanding of the outcome visibly to lead the reviewer through your conclusions, and then finish the paper with a summing up of the implications of the study. The purpose here is to offer an understanding of your results and support all of your conclusions, using facts from your research and generally accepted information, if suitable. The implication of results should be fully described.

Infer your data in the conversation in suitable depth. This means that when you clarify an observable fact, you must explain mechanisms that may account for the observation. If your results vary from your prospect, make clear why that may have happened. If your results agree, then explain the theory that the proof supported. It is never suitable to just state that the data approved the prospect, and let it drop at that. Make a decision as to whether each premise is supported or discarded or if you cannot make a conclusion with assurance. Do not just dismiss a study or part of a study as "uncertain."

Research papers are not acknowledged if the work is imperfect. Draw what conclusions you can based upon the results that you have, and take care of the study as a finished work.

- You may propose future guidelines, such as how an experiment might be personalized to accomplish a new idea.
- Give details of all of your remarks as much as possible, focusing on mechanisms.
- Make a decision as to whether the tentative design sufficiently addressed the theory and whether or not it was correctly restricted. Try to present substitute explanations if they are sensible alternatives.
- One piece of research will not counter an overall question, so maintain the large picture in mind. Where do you go next? The best studies unlock new avenues of study. What questions remain?
- Recommendations for detailed papers will offer supplementary suggestions.

Approach:

When you refer to information, differentiate data generated by your own studies from other available information. Present work done by specific persons (including you) in past tense.

Describe generally acknowledged facts and main beliefs in present tense.

The Administration Rules

Administration Rules to Be Strictly Followed before Submitting Your Research Paper to Global Journals Inc.

Please read the following rules and regulations carefully before submitting your research paper to Global Journals Inc. to avoid rejection.

Segment draft and final research paper: You have to strictly follow the template of a research paper, failing which your paper may get rejected. You are expected to write each part of the paper wholly on your own. The peer reviewers need to identify your own perspective of the concepts in your own terms. Please do not extract straight from any other source, and do not rephrase someone else's analysis. Do not allow anyone else to proofread your manuscript.

Written material: You may discuss this with your guides and key sources. Do not copy anyone else's paper, even if this is only imitation, otherwise it will be rejected on the grounds of plagiarism, which is illegal. Various methods to avoid plagiarism are strictly applied by us to every paper, and, if found guilty, you may be blacklisted, which could affect your career adversely. To guard yourself and others from possible illegal use, please do not permit anyone to use or even read your paper and file.

Criterion for Grading a Research Paper (Compilation) by Global Journals Inc. (US)

Please note that following table is only a Grading of "Paper Compilation" and not on "Performed/Stated Research" whose grading solely depends on Individual Assigned Peer Reviewer and Editorial Board Member. These can be available only on request and after decision of Paper. This report will be the property of Global Journals Inc. (US).

Topics	Grades		
	A-B	**C-D**	**E-F**
Abstract	Clear and concise with appropriate content, Correct format. 200 words or below	Unclear summary and no specific data, Incorrect form Above 200 words	No specific data with ambiguous information Above 250 words
Introduction	Containing all background details with clear goal and appropriate details, flow specification, no grammar and spelling mistake, well organized sentence and paragraph, reference cited	Unclear and confusing data, appropriate format, grammar and spelling errors with unorganized matter	Out of place depth and content, hazy format
Methods and Procedures	Clear and to the point with well arranged paragraph, precision and accuracy of facts and figures, well organized subheads	Difficult to comprehend with embarrassed text, too much explanation but completed	Incorrect and unorganized structure with hazy meaning
Result	Well organized, Clear and specific, Correct units with precision, correct data, well structuring of paragraph, no grammar and spelling mistake	Complete and embarrassed text, difficult to comprehend	Irregular format with wrong facts and figures
Discussion	Well organized, meaningful specification, sound conclusion, logical and concise explanation, highly structured paragraph reference cited	Wordy, unclear conclusion, spurious	Conclusion is not cited, unorganized, difficult to comprehend
References	Complete and correct format, well organized	Beside the point, Incomplete	Wrong format and structuring

INDEX

save our planet

Global Journal of Computer Science and Technology

Visit us on the Web at www.GlobalJournals.org | www.ComputerResearch.org
or email us at helpdesk@globaljournals.org

ISSN 9754350