

### SOFTWARE & DATA ENGG.

Prim's  
Algorithm

Software Bug  
Prediction

Highlights

Business  
Management  
Systems

Integrating  
Machine  
Learning

Discovering Thoughts, Inventing Future

VOLUME 26 / 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 26 ISSUE 1 (VER. 1.0)

OPEN ASSOCIATION OF RESEARCH SOCIETY

© Global Journal of Computer Science and Technology. 2026

All rights reserved.

This is a special issue published in version 1.0 of "Global Journal of Computer Science and Technology" By Global Journals Inc.

All articles are open access articles distributed under "Global Journal of Computer Science and Technology"

Reading License, which permits restricted use. Entire contents are copyright by of "Global Journal of Computer Science and Technology" unless otherwise noted on specific articles.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without written permission.

The opinions and statements made in this book are those of the authors concerned. Ultraculture has not verified and neither confirms nor denies any of the foregoing and no warranty or fitness is implied.

Engage with the contents herein at your own risk.

The use of this journal, and the terms and conditions for our providing information, is governed by our Disclaimer, Terms and Conditions and Privacy Policy given on our website <http://globaljournals.us/terms-and-condition/menu-id-1463/>

By referring / using / reading / any type of association / referencing this journal, this signifies and you acknowledge that you have read them and that you accept and will be bound by the terms thereof.

All information, journals, this journal, activities undertaken, materials, services and our website, terms and conditions, privacy policy, and this journal is subject to change anytime without any prior notice.

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](mailto:local@globaljournals.org)

### *eContacts*

Press Inquiries: [press@globaljournals.org](mailto:press@globaljournals.org)  
Investor Inquiries: [investors@globaljournals.org](mailto:investors@globaljournals.org)  
Technical Support: [technology@globaljournals.org](mailto:technology@globaljournals.org)  
Media & Releases: [media@globaljournals.org](mailto: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. Corina Sas*

School of Computing and Communication  
Lancaster University Lancaster, UK

*Dr. Sotiris Kotsiantis*

Ph.D. in Computer Science, Department of Mathematics,  
University of Patras, Greece

*Dr. Diego Gonzalez-Aguilera*

Ph.D. in Photogrammetry and Computer Vision Head of  
the Cartographic and Land Engineering Department  
University of Salamanca Spain

*Dr. Yuanyang Zhang*

Ph.D. of Computer Science, B.S. of Electrical and  
Computer Engineering, University of California, Santa  
Barbara, United States

*Dr. Osman Balci, Professor*

Department of Computer Science Virginia Tech, Virginia  
University Ph.D. and M.S. Syracuse University, Syracuse,  
New York M.S. and B.S. Bogazici University, Istanbul,  
Turkey

*Dr. Kwan Min Lee*

Ph. D., Communication, MA, Telecommunication,  
Nanyang Technological University, Singapore

*Dr. Khalid Nazim Abdul Sattar*

Ph.D, B.E., M.Tech, MBA, Majmaah University,  
Saudi Arabia

*Dr. Jianyuan Min*

Ph.D. in Computer Science, M.S. in Computer Science, B.S.  
in Computer Science, Texas A&M University, United States

*Dr. Kassim Mwitondi*

M.Sc., PGCLT, Ph.D. Senior Lecturer Applied Statistics/  
Data Mining, Sheffield Hallam University, UK

*Dr. Kurt Maly*

Ph.D. in Computer Networks, New York University,  
Department of Computer Science Old Dominion  
University, Norfolk, Virginia

*Dr. Zhengyu Yang*

Ph.D. in Computer Engineering, M.Sc. in  
Telecommunications, B.Sc. in Communication Engineering,  
Northeastern University, Boston, United States

*Dr. Don. S*

Ph.D in Computer, Information and Communication  
Engineering, M.Tech in Computer Cognition Technology,  
B.Sc in Computer Science, Konkuk University, South  
Korea

*Dr. Ramadan Elaiess*

Ph.D in Computer and Information Science, University of  
Benghazi, Libya

*Dr. Omar Ahmed Abed Alzubi*

Ph.D in Computer and Network Security, Al-Balqa Applied  
University, Jordan

*Dr. Stefano Berretti*

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

*Dr. Lamri Sayad*

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

*Dr. Hazra Imran*

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

*Dr. Nurul Akmar Binti Emran*

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

*Dr. Anis Bey*

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

*Dr. Rajesh Kumar Rolan*

Ph.D in Computer Science, MCA & BCA - IGNOU, MCTS & MCP - MICROSOFT, SCJP - SUN MICROSYSTEMS, Singhania University, 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. Chutisant Kerdvibulvech*

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

*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. Tauqeer Ahmad Usmani*

Ph.D in Computer Science, Oman

*Dr. Magdy Shayboub Ali*

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

*Dr. Asim Sinan Yuksel*

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

*Alessandra Lumini*

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

*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. Federico Tramarin*

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

*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

---

- i. Copyright Notice
  - ii. Editorial Board Members
  - iii. Chief Author and Dean
  - iv. Contents of the Issue
- 
1. Integrating Machine Learning into Business Management Systems: The Rbox+ API. *1-14*
  2. Blockchain Challenges: Advantages and Algorithms. *15-26*
  3. Towards Optimized K means Clustering using Nature-Inspired Algorithms for Software Bug Prediction. *27-36*
  4. Comparison of Primand Kruskal's Algorithm *37-43*
- 
- v. Fellows
  - vi. Auxiliary Memberships
  - vii. Preferred Author Guidelines
  - viii. Index



# Integrating Machine Learning into Business Management Systems: The Rbox+ API

By Antonios Konomos & Spiros Chountasis

**Abstract-** This paper presents an innovative software interface for the utilization of widely used machine learning algorithms in a unified Python/R programming environment. This study makes two contributions. First, a more comprehensive and specialized architecture is made available for integrating machine learning into enterprise information systems. Second, a novel software model, Rbox+, is proposed to execute machine learning algorithms by jointly leveraging the capabilities of the Python and R programming languages through an Application Programming Interface (API). The proposed API is tested and evaluated using a publicly available benchmark dataset for regression analysis (Car-sales dataset, available on Kaggle), applying multiple machine learning models and comparative performance metrics. The obtained results demonstrate improved computational efficiency and scalability, with the execution of multiple models completed within a short processing time on standard hardware. Unlike conventional machine learning APIs or isolated ERP analytics tools, Rbox+ enables transparent, language-independent execution and validation of machine learning models while exposing the underlying source code. The proposed approach supports practical applications in enterprise analytics, reproducible research, and machine learning education, enhancing interoperability between ERP systems, analytics platforms, and statistical programming environments.

**Keywords:** *software interface, software applications, computer technologies, machine learning.*

**GJCST-C Classification:** LCC Code: QA76.9.A43



*Strictly as per the compliance and regulations of:*



# Integrating Machine Learning into Business Management Systems: The Rbox+ API

Antonios Konomos <sup>α</sup> & Spiros Chountasis <sup>σ</sup>

**Abstract-** This paper presents an innovative software interface for the utilization of widely used machine learning algorithms in a unified Python/R programming environment. This study makes two contributions. First, a more comprehensive and specialized architecture is made available for integrating machine learning into enterprise information systems. Second, a novel software model, Rbox+, is proposed to execute machine learning algorithms by jointly leveraging the capabilities of the Python and R programming languages through an Application Programming Interface (API). The proposed API is tested and evaluated using a publicly available benchmark dataset for regression analysis (Car-sales dataset, available on Kaggle), applying multiple machine learning models and comparative performance metrics. The obtained results demonstrate improved computational efficiency and scalability, with the execution of multiple models completed within a short processing time on standard hardware. Unlike conventional machine learning APIs or isolated ERP analytics tools, Rbox+ enables transparent, language-independent execution and validation of machine learning models while exposing the underlying source code. The proposed approach supports practical applications in enterprise analytics, reproducible research, and machine learning education, enhancing interoperability between ERP systems, analytics platforms, and statistical programming environments.

**Keywords:** software interface, software applications, computer technologies, machine learning.

## I. INTRODUCTION

### a) ML Overview and Challenges

Machine Learning (ML) is the process of creating software applications that enables a system to gain new knowledge from training data rather than through explicit programming. It employs several algorithms that learn from data iteratively to enhance, characterize, and predict outcomes. More accurate models based on the training data can then be created as the algorithms ingest training data. ML can be categorized into supervised, unsupervised learning and reinforcement learning. Many effective ML applications, from information-filtering systems to data-mining techniques, have been developed in recent years.

### b) Python and R Ecosystems

The frameworks and algorithms that are the key components of this field have also undergone substantial modifications. Today's technological advancements have updated forms of software engineering and computer architecture [7,15,31]. Furthermore, large volume of data management stresses the need for more efficient visualization and presentation tools [9,19]. Leading open-source programming languages like Python and R, are at the forefront of requisite quantitative computer programs [8,14,22]. Both programming languages have their own distinctive characteristics that encompass a wide range of ML algorithms, statistical controls, and strong data visualization features.

Being able to correlate data to detect patterns and anomalies can help an organization predict outcomes and improve its operations of business. Appropriate datasets must be applied for the learning process of an ML technique. Because the time at which training converges is unexpected, it is challenging to anticipate in advance the effort of training a model. The model training data set, learning parameter settings, and random variables all affect convergence [11,27].

### c) MLaaS and Cloud Deployment

Nowadays, big data helps to improve the accuracy of the ML models and makes it feasible to virtualize data, so that it may be kept in the cloud or on-premises in the most effective and economical way possible. Major cloud providers deliver ML as a Service (MLaaS), allowing users to train predictive models on this data even if they lack ML expertise or infrastructure [16,18]. Customers may outsource their costly ML operations to reliable servers using MLaaS, yet there are still no firm guarantees of service accuracy. The user may merely note that the trained model or prediction result is well-formed without knowing if the delivered result is accurate [35].

### d) The Rbox+ Initiative

Web APIs have become the preferred choice for the seamless integration of heterogeneous software systems. The primary objective of this research is to propose the design and implementation of an API (Rbox+), which is responsible for sending data to the R or Python programming environments, requesting the execution of various ML algorithms. Furthermore, statistical tests and graphical representations may also

Author <sup>α</sup>: Intrasoft International SA, Athens, Greece.

e-mail: akonomos@gmail.com

Author <sup>σ</sup>: Independent Power Transmission Operator, Athens, Greece.

e-mail: schountasis@admie.gr

be performed in the R programming language returning the computation results in Python/R Jupyter Notebook [29] format.

The Rbox+ facilitates the integration of both Python and R programming environments with different types and content systems such as Enterprise Resource Planning (ERP) or Customer Relationship Management (CRM) which can consume Web services. The current study is intended to further develop an interface for integrating Python/R with the SAP S/4 Hana ERP system. SAP S/4 Hana ERP is an ERP business suite based on the SAP HANA in-memory database [4].

#### e) *Rbox+ Features*

The Rbox+ is designed as an “open box” that is receptive to further adjustments, modifications, and improvements. Validation of ML models is very crucial, providing a parameter for software integration and system implementation. The user can review and compare the source code of both Python and R by examining the results of the requested ML and statistical tests. This feature further enhances the transparency of the results by modifying and improving the ML algorithms if needed, as their validation becomes more critical than ever before.

#### f) *Organization of the Paper*

The outline of this paper is organized as follows. Section 2 presents the Rbox+ interface and its utilization, emphasizing the implementation of ML algorithms. Section 3 introduces and analyzes the software architecture of the proposed system from a Web service perspective. Section 4 presents simulation results and code implementation based on a use-case scenario involving multiple ML regression algorithms. Finally, Section 5 concludes the paper with a discussion and an outlook on future work.

## II. THE PYTHON/R INTEGRATION FOR THE RBOX+ UTILIZATION

#### a) *Rbox+ Overview*

Machine learning data processing and visualization tools are rarely embedded directly into information systems that are commonly used for managing an organization’s core operations [1]. Representative examples of such systems include Microsoft Excel and SAP S/4 Hana ERP. SAP S/4 Hana represents SAP A. G.’s ERP solution and integrates critical enterprise activities such as Financial Accounting, Sales and Distribution, Materials Management, Production Planning, and Human Resources [34]. Despite its widespread adoption, SAP S/4 Hana lacks native advanced analytics, machine learning execution, and data visualization capabilities. Instead, these needs are typically addressed through separate analytical platforms such as SAP BI or SAP Analytics Cloud [30].

Rbox+ is designed to bridge this gap by providing a lightweight and flexible Web API that enables enterprise systems to directly invoke machine learning and statistical algorithms implemented in Python and R. Unlike traditional ERP analytics solutions or MLaaS platforms, which often operate as closed or proprietary environments, Rbox+ allows enterprise applications to trigger ML computations while retaining full access to the underlying algorithmic logic and execution flow.

From the perspective of system integration, Rbox+ simplifies the interaction between heterogeneous software environments. Through a single web service call, client applications can submit datasets and analysis requests without requiring native support for Python or R. This design enables ML-driven analytics to be embedded into existing enterprise workflows without altering the underlying ERP or spreadsheet-based infrastructure.

#### b) *Integration with Python and R*

Rbox+ is implemented as a generic API that enables the Python and R programming environments to be seamlessly incorporated into any application capable of issuing HTTP requests [24]. From an integrator’s perspective, Rbox+ simplifies the integration process by minimizing the required implementation effort. The integration consists of a single step, which involves preparing and executing a direct service call through Rbox+’s dispatcher method. This design simplicity and flexibility make Rbox+ highly user-friendly.

Most modern programming languages provide built-in mechanisms for converting two-dimensional datasets into JSON structures that can be transmitted as parameters to web services. As a result, invoking Rbox+ is comparable to making a standard function call in most programming environments. A key feature of Rbox+ is its generic calling mechanism, which remains consistent regardless of the specific machine learning algorithm, statistical test, or graphical output requested. Integrators are therefore only required to construct a JSON request containing the dataset and the selected analysis, while the remaining integration code can be reused. This approach ensures efficient, flexible, and reusable calls to Rbox+, thereby enhancing the overall integration process.

#### c) *Output and Reproducibility via Jupyter Notebook*

From an end-user perspective, Rbox+ generates output in the form of Jupyter Notebook files. Jupyter Notebook is an open-source web application that supports interactive computing by combining executable code, narrative text, equations, and visualizations within a single document [13]. Rather than detailing the general functionality of Jupyter Notebook, this work focuses on its specific role in enhancing the usability, transparency, and reproducibility of Rbox+.

By producing results directly in Jupyter Notebook format, Rbox+ preserves the complete computational workflow, including the Python /R source code, intermediate outputs, and final results. This integration enables users to review, validate, and reproduce analyses without relying on opaque result summaries. In addition, the notebook format allows analysts to extend or modify the generated code, supporting iterative experimentation and a deeper understanding of the applied ML techniques [10,21].

The reproducibility facilitated by Rbox+ is particularly important in research and enterprise analytics contexts, where model validation and auditability are critical. Unlike many ERP analytics tools or MLaaS platforms, which typically provide only aggregated results, Rbox+ exposes the entire execution logic. This allows users to verify assumptions, inspect parameter settings, and assess model robustness, thereby enhancing trust in ML-driven decision support and aligning with best practices in reproducible research [23].

Furthermore, the notebook-based output makes Rbox+ suitable for educational use. It can serve as a structured template for teaching ML concepts, enabling students to explore algorithms, visualize results, and connect theoretical concepts with executable code. The ease of use and standardized output format make Rbox+ applicable to both introductory ML coursework and advanced analytical research.

In summary, Rbox+ combines generic Web API architecture with direct Python/R execution and reproducible Jupyter-based output. This combination distinguishes it from existing ERP analytics solutions, MLaaS platforms, and standalone scripting approaches by offering transparency, interoperability, and extensibility within a unified integration framework.

*d) Machine Learning Process in Rbox+*

A wide range of companies, including Google Cloud AI [20], Amazon Web Services (AWS) AI [2], IBM Watson [17], and Microsoft Azure [25], provide cloud-based machine learning services. These platforms offer collections of APIs that execute inference computations using industry-trained deep neural networks (DNNs) on powerful cloud infrastructures, without requiring developers to possess in-depth knowledge of ML algorithms or resource provisioning.

Despite their advantages, third-party ML APIs still present several challenges, particularly when ML applications must be integrated into larger software systems. One approach to making ML more powerful, practical, and interoperable is the development of interfaces that enable algorithm execution independently of programming language and data format. ML algorithms are implemented in diverse ways and are designed to operate on heterogeneous datasets. Owing to differences in semantics, data requirements, and accuracy–performance trade-offs, ML APIs can be difficult to use correctly and efficiently.

In this study, these limitations are addressed through the introduction of Rbox+, a web-service-based interface that enables researchers and developers to perform statistical analyses and execute ML algorithms in the R and Python environments. Rbox+ provides a simplified and uniform access layer for ML execution, facilitating integration and comparative analysis across platforms. Tables 1 and 2 present the Rbox+ templates developed using the Python and R programming language.

*Table 1:* ML Algorithms in Python

Template	ML Task	Options / Hyperparameters
Decision Trees	Classification/Regression	Descriptives Correlation matrix k-fold Cross Validation
Advanced Trees	Classification/Regression	Descriptives Correlation matrix Bagging Random Forest Extremely Randomized Trees k-fold Cross Validation
Boosting Trees	Classification/Regression	Descriptives Correlation matrix AdaBoost XG Boost Light GBM k-fold Cross Validation
Gaussian Mixture	Clustering	Descriptives Correlation matrix
Mini Batch K. means	Clustering	Descriptives Correlation matrix



Naive Bayes	Classification	Descriptives Correlation matrix k-fold Cross Validation
Stochastic Gradient Descent	Classification/Regression	Descriptives Correlation matrix k-fold Cross Validation
Support Vector Machines	Classification/Regression	Descriptives Correlation matrix Linear kernel RBF kernel Sigmoid kernel Polynomial kernel k-fold Cross Validation

Table 2: ML & Statistical Algorithms in R

Template	ML Task	Options/Hyperparameters
k-means	Clustering	Descriptives Silhouette method Gap statistic method Anova
K-prototypes	Clustering	Descriptives Silhouette method Anova Chi-square tests
Hierarchical Clustering	Clustering	Descriptives Anova
Partitioning Around Medoids (PAM)	Clustering	Descriptives Silhouette method Gap statistic method Anova Chi-square tests
Clustering Large Applications (CLARA)	Clustering	Descriptives Silhouette method Gap statistic method Anova Chi-square tests
K-Nearest Neighbors	Classification/ Regression	Descriptives Correlation matrix Elbow method ROC curve
Linear Regression	Regression	Descriptives Pearson's Correlation R.square value Residuals distribution NCV test Spread Level Plot
Linear Multiple Regression	Regression	Descriptives Correlation matrix Collinearity test K-fold cross-validation Stepwise Regression
Logistic Regression	Classification	Descriptives Correlation matrix Collinearity test Hosmer test (Goodness of fit) Predictive ability ROC curve
Multinomial Logistic Regression	Classification	Descriptives Correlation matrix Evaluate Collinearity Stepwise Regression



Principal Component Analysis (PCA)	Dimensionality Reduction	Descriptives Correlation matrix Correlations Vars – PCs
PCA mix	Dimensionality Reduction	Descriptives Plots
Random Forest	Classification / Regression	Descriptives Correlation matrix
t-Test	Association	Descriptives Kolmogorov-Smirnov Levene's test Mann-Whitney U Boxplots Histogram
Paired t-Test	Association	Descriptives Pearson's Correlation Kolmogorov-Smirnov Mann-Whitney Histograms
One-way Anova	Association	Descriptives Tukey HSD (honest significance test) Levene's test (homoskedasticity test) Diagnostic plots Kruskal Wallis
Two-way Anova	Association	Descriptives Tukey HSD Pairwise t-tests Levene's test Normality test
$\chi^2$ test	Association	Descriptives Crosstabs Balloon plot Barplot

e) *The Rbox+ Learning Aspect*

Rbox+ has been designed as a distinctive learning and analysis tool, acknowledging the researcher's fundamental motivation to understand how algorithmic calculations are performed. By displaying the generated Python/R source code at every stage of the computational workflow, Rbox+ ensures full transparency throughout the entire data processing cycle, including data upload, data manipulation, statistical testing, data analysis, and visualization of the results. This characteristic transforms Rbox+ into a powerful instructional tool that enables developers and researchers to gain a deeper understanding of machine learning algorithms and their execution context, while also supporting the development of Python/R programming skills.

Data analysis in a machine learning context involves multiple stages, including data collection and exploration, data cleaning and preprocessing, feature selection, model training and evaluation, and iterative refinement. The objective is to understand the data, improve its quality, and optimize it for effective model training [32]. The iterative nature of this process allows for continuous refinement of the analytical pipeline and the resolution of emerging challenges. Ultimately, the goal is to deploy a well-performing model and monitor its performance over time [26]. Within this framework,

Rbox+ provides an accessible platform for exploring complex statistical and machine learning techniques, supporting the learning process of its users.

III. SYSTEM ARCHITECTURE

a) *The Three-Tier Design*

This section describes the System Architecture, which is defined independently of the implementation details. The Rbox+ API operates on a web server that is connected to both the R and Python programming environments and is implemented using the PHP programming language. The machine learning methods included in Rbox+ are accessible to any application capable of handling HTTP requests. The design of Rbox+ follows a conventional three-tier architecture [24]. More specifically, it is based on a client-server model consisting of three distinct processing layers, as illustrated in *Figure 1*:

- a) Client Tier (service consumer) concerns the calling applications of Rbox+.
- b) API Tier (service provider) handles the requests towards the R and Python environments.
- c) Processing Tier executes the requested ML and statistical computations.

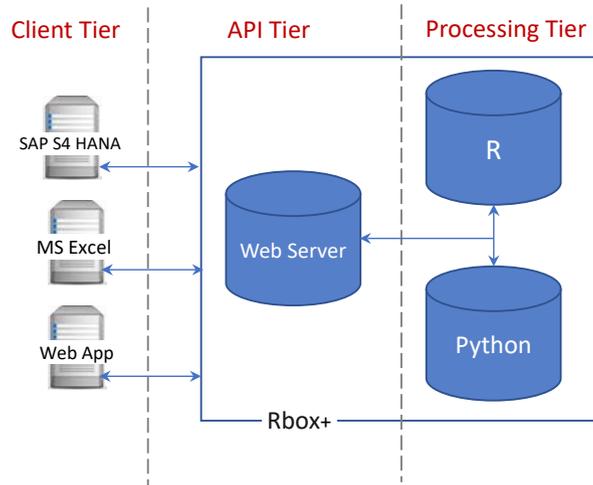


Figure 1: Rbox+ Three-tier Architecture

b) Request JSON Structure

All three tiers (Processing, API, and Client tier), operate in accordance with the previously proposed Rbox API [24] and are briefly described in the following section. The API tier functions as a web server that

handles HTTP requests via POST method. Each request carries a JSON structure (Table 1) containing information about the requested machine learning algorithm (implemented in Python/R), as well as the dataset to which the model will be applied.

Table 3: JSON Structure

Variable	Description	JSON sample for t-Test
template	ML / Statistical test	<pre>{   "template": "P.NaiveBayes",   "numvars": 2,   "numopts": 6,   "options": [1,1,1,1,1,1],   "variables": [     {       "id": "X0",       "name": "VEHICLE_TYPE",       "datatype": "factor",       "values": ["Car", "Passenger", ...]     },     {       "id": "X1",       "name": "PRICE",       "datatype": "numeric",       "values": ["21.50", "28.40", ...]     }   ] }</pre>
numvars	Variables counter	
numopts	Options counter	
options	Options on/off	
variables	Variables list	
id	$X_1 \dots X_n$	
name	Variable name	
datatype	Variable type	
values	Value list	

After the execution of the corresponding script, a response is returned to the calling system. This response essentially includes a URL pointing to the results page. The machine learning outputs are presented in Jupyter Notebook format [29], allowing end users to view both the generated Python/R source code and the corresponding computational results within a single HTML document.

The JSON element template specifies the requested machine learning technique or statistical test. As shown in Table 1, the options parameter includes on/off switches that control the hyperparameters of the ML algorithm or enable additional tests associated with each method. The variables element of the JSON structure contains the attributes id, name, and datatype, along with the corresponding list of values for each variable. The id attribute represents the unique identifier of each variable and ranges from  $x_0$  to  $x_n$ . The numvars parameter specifies the total number of variables, while

numopts indicates the number of available on/off switches within the options array.

Through this mechanism, Rbox+ dynamically enables or disables hyperparameters for each ML method, thereby reducing the execution time of the corresponding Python/R script and minimizing the overall response time of the HTTP request.

IV. SIMULATION RESULTS AND CODE IMPLEMENTATION

a) Sample Dataset and use-case Scenario

In this section, indicative results produced by Rbox+ are presented. A publicly available dataset, commonly used for educational purposes, is employed to demonstrate the flexibility of Rbox+ in executing multiple ML algorithms and selecting the best-fitting model through comparative analysis. The car sales data set is available in the Kaggle repository and contains information on vehicle sales across different

manufacturers. In addition to sales volume, the dataset includes several explanatory variables, such as resale value, price, engine size, horsepower, wheelbase, width, length, fuel capacity, fuel efficiency, and power performance factor.

For the purposes of the presented use-case scenario, a range of machine learning algorithms supported by Rbox+ is applied to predict vehicle sales volume. The explanatory variables used in the analysis include engine size, horsepower, wheelbase, width, length, fuel capacity, and fuel efficiency. This type of prediction problem is addressed using multiple regression models, and the predictive performance of each method is evaluated using the R<sup>2</sup> and adjusted R<sup>2</sup> metrics [3].

b) Model Selection

After importing the dataset in the Rbox+ Excel Client, the researcher can select the Data Analysis model (Regression, Classification, Clustering or Dimensionality Reduction) and the desired ML method. At this step, it is important to specify the response variable as well as the independent variables. Additional options can also be enabled, performing further calculations or plots useful during the analysis stage as shown in Figure 2. Moreover, Figure 3 illustrates the option of using the SAP S4Hana Client. The analyst can easily explore all the available ML Regression methods and trigger the execution accordingly.

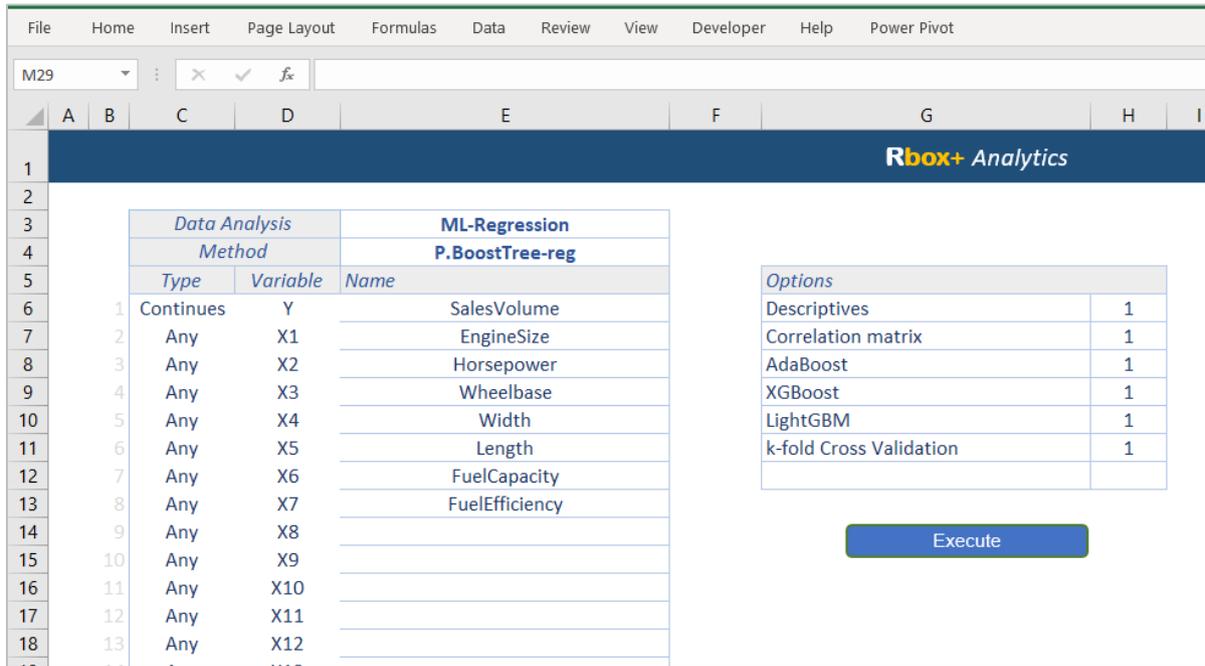


Figure 2: Rbox+ Excel Client



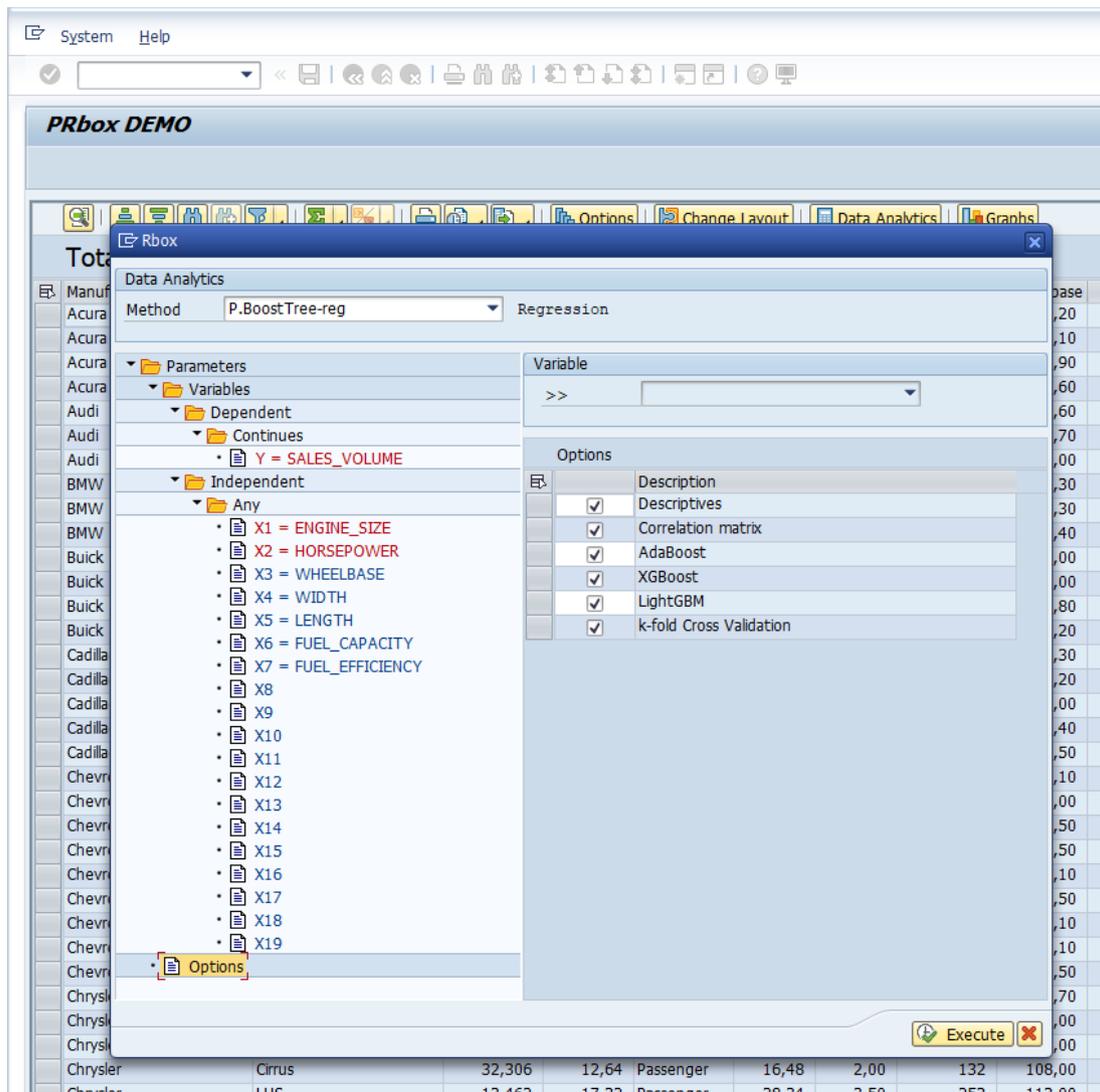


Figure 3: Rbox+ SAP S4 Hana Client

c) Results Page

The calculation outcomes are presented in Jupyter Notebook format within the user’s web browser, as illustrated in Figures 4 and 5. Both the source code and the extracted results appear on the same page, allowing the analyst to review the outcomes and compare the goodness of fit across different models.

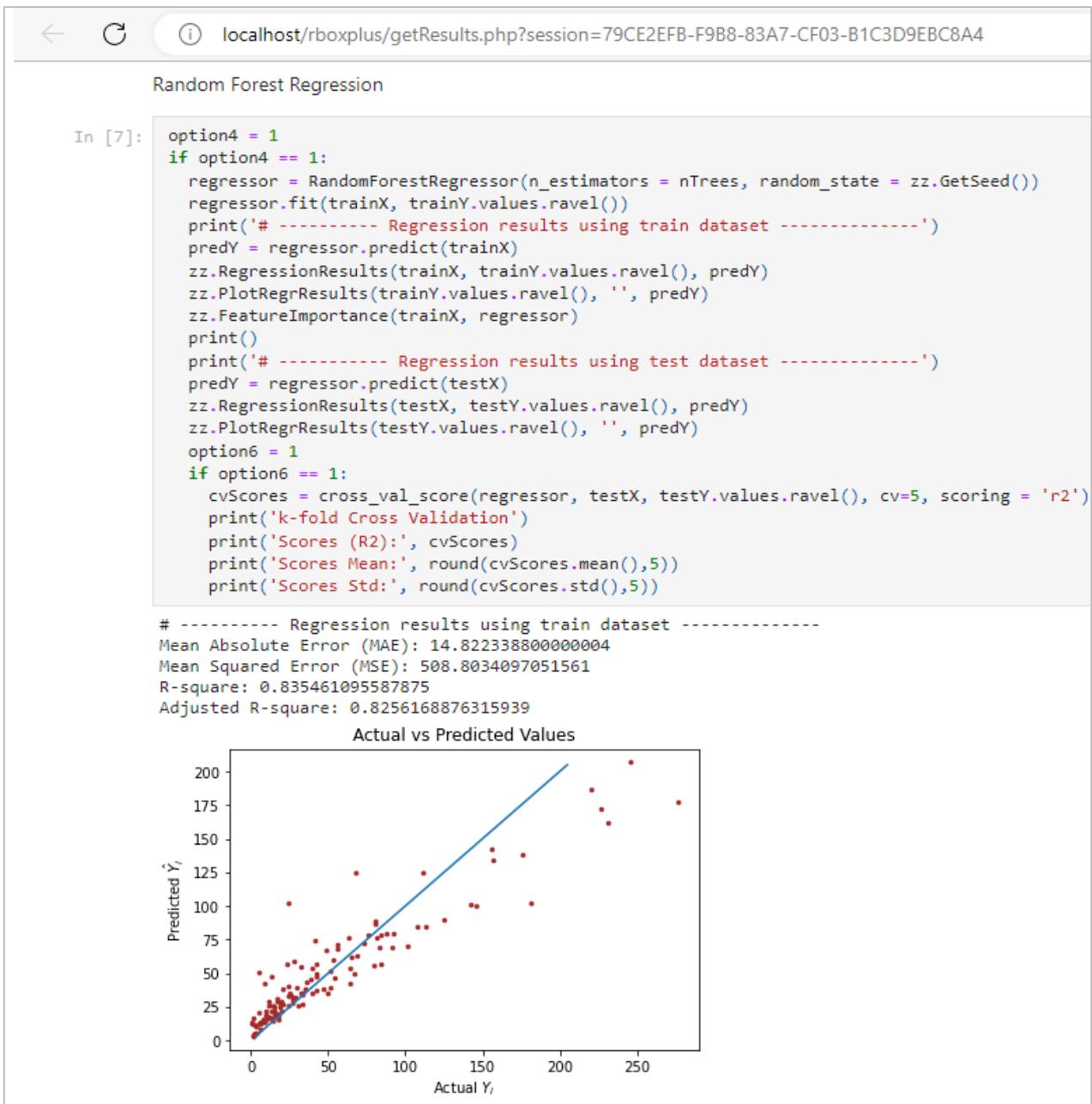


Figure 4: Rbox + Results: Jupyter Notebook

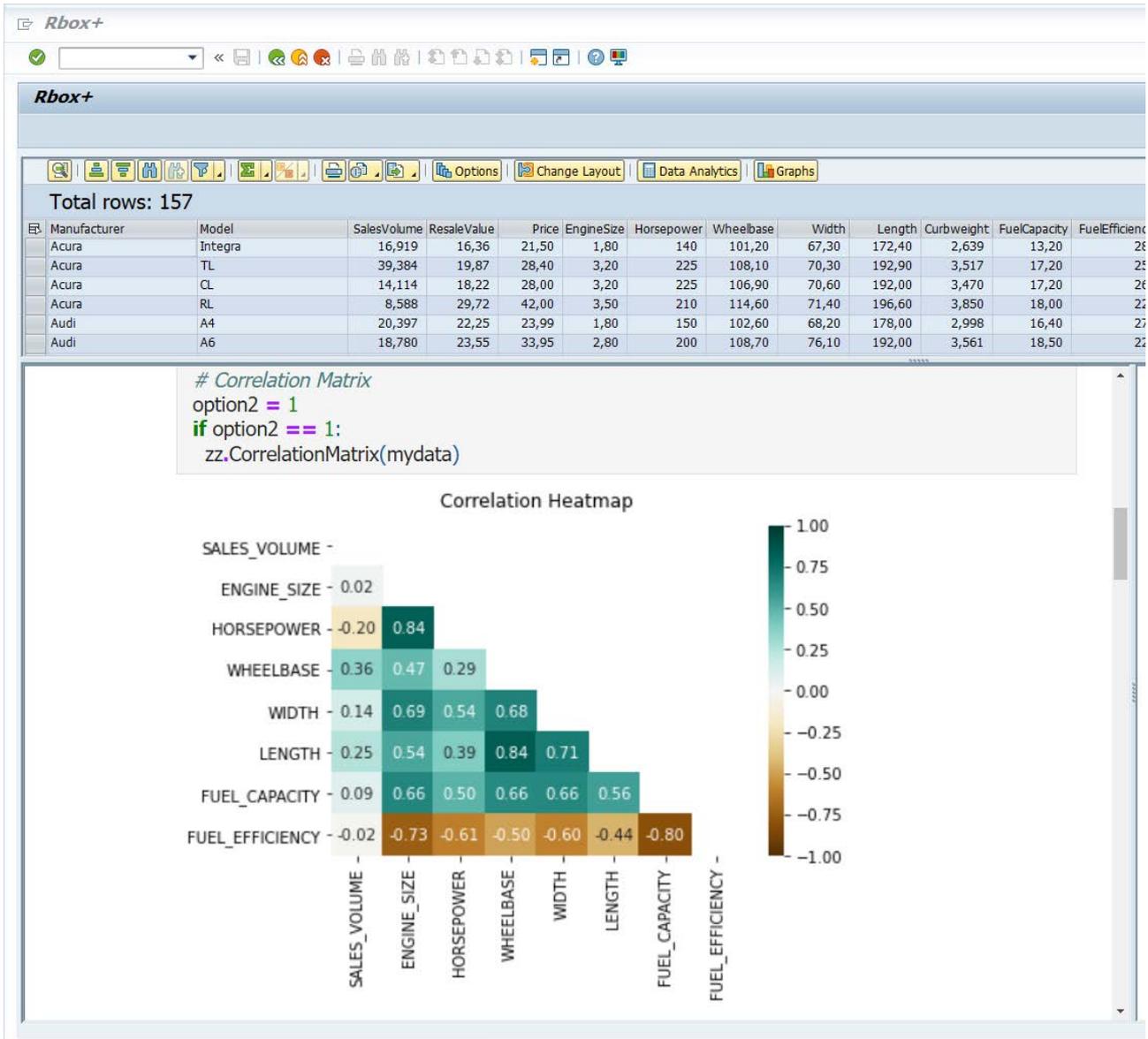


Figure 5: Rbox + SAP S4 Hana output

In the Excel worksheet, a table is displayed that contains hyperlinks to the results page, the generated source code, and the corresponding Jupyter Notebook file. Through these links, the analyst can access the

complete history of the performed computations and revisit or reassess the associated results at any time, as illustrated in Figure 6.

Result	Source code	Jupyter Notebook
<a href="#">P.SGD-reg</a>	<a href="#">P.SGD-reg</a>	<a href="#">P.SGD-reg</a>
<a href="#">R.KNN-reg</a>	<a href="#">R.KNN-reg</a>	<a href="#">R.KNN-reg</a>
<a href="#">P.DcsTree-reg</a>	<a href="#">P.DcsTree-reg</a>	<a href="#">P.DcsTree-reg</a>
<a href="#">P.AdvTree-reg</a>	<a href="#">P.AdvTree-reg</a>	<a href="#">P.AdvTree-reg</a>
<a href="#">P.BoostTree-reg</a>	<a href="#">P.BoostTree-reg</a>	<a href="#">P.BoostTree-reg</a>
<a href="#">P.SVM-reg</a>	<a href="#">P.SVM-reg</a>	<a href="#">P.SVM-reg</a>
<a href="#">R.MultipleRegression</a>	<a href="#">R.MultipleRegression</a>	<a href="#">R.MultipleRegression</a>
<a href="#">R.KNN-reg</a>	<a href="#">R.KNN-reg</a>	<a href="#">R.KNN-reg</a>

Figure 6: Rbox + API hyperlinks

The users can export the entire Python or R source code, as depicted in Figures 7 and 8, and subsequently edit and process it in their preferred development environment (e.g., Jupyter Notebook, VS Code, Eclipse).

```
localhost/rboxplus/getSource.php?session=79CE2EFB-F9B8-83A7-CF03-B1C3D9EBC8A4

# Advanced Trees Regression
# Rbox+ output
# 2023-07-01

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import zlib as zz
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import BaggingRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import ExtraTreesRegressor
from sklearn.model_selection import cross_val_score

myjson = zz.GetJSONdata('79CE2EFB-F9B8-83A7-CF03-B1C3D9EBC8A4')
options = myjson['options']

mycateg = zz.BuildCategories(myjson)
mydata = zz.BuildDataFrame(myjson)
print(mydata.head())

# =====
# Descriptive Statistics
option1 = 1
if option1 == 1 :
    zz.DescriptiveStats(mydata, mycateg)
```

Figure 7: The Source Code

```
localhost/rboxplus/getSource.php?session=79CE2EFB-F9B8-83A7-CF03-B1C3D9EBC8A4&ftype=ipynb

{
  "cells": [
    {
      "cell_type": "markdown",
      "metadata": {},
      "source": [
        "Advanced Trees Regression<br>\n",
        "Rbox+ output<br>\n",
        "2023-07-01"
      ]
    },
    {
      "cell_type": "code",
      "execution_count": null,
      "metadata": {},
      "outputs": [],
      "source": [
        "import pandas as pd\n",
        "import numpy as np\n",
        "import matplotlib.pyplot as plt\n",
        "import zlib as zz\n",
        "from sklearn.tree import DecisionTreeRegressor\n",
        "from sklearn.ensemble import BaggingRegressor\n",
        "from sklearn.ensemble import RandomForestRegressor\n",
        "from sklearn.ensemble import ExtraTreesRegressor\n",
        "from sklearn.model_selection import cross_val_score"
      ]
    }
  ],
}
```

Figure 8: Jupyter Notebook (.ipynb) file

d) Performance Evaluation

For the purposes of the selected use-case scenario, a wide variety of ML regression algorithms were executed, and the results are presented in Table 4.

Table 4: ML Regression Algorithms and their Results

ML Regression	Method	MSE	R <sup>2</sup>	Adj.R <sup>2</sup>
Support Vector Machines	Linear kernel	0.804	0.195	0.147
	Polynomial kernel	0.830	0.169	0.120
	RBF kernel ( <i>Gaussian Kernel Radial Basis</i> )	0.767	0.232	0.186
Boosting Trees	AdaBoost ( <i>Adaptive Boosting</i> )	270.738	0.912	0.907
	XGBoost ( <i>eXtreme Gradient Boosting</i> )	101.074	0.967	<b>0.965</b>
	LightGBM ( <i>Light Gradient Boosting Machine</i> )	1678.109	0.457	0.424
Advanced Trees	Bagging Regression	507.963	0.835	0.825
	Random Forest	508.803	0.835	0.825
	ExtraTrees ( <i>Extremely Randomized Trees</i> )	100.603	0.967	0.964
KNN	K-Nearest Neighbors	1.073	0.067	-0.216
SGD	Stochastic Gradient Descent	0.742	0.257	0.213
Decision Tree		1559.749	0.495	0.465
Multiple Regression		56.010	0.331	0.291

Seven discrete ML regression algorithms were evaluated, some with multiple variations, resulting in a total of 13 distinct result sets. The predictive performance of each model can be assessed using the R<sup>2</sup> and Adjusted R<sup>2</sup> metrics. Notably, the execution of all 13 models on a dataset with 8 variables and 157 observations required less than two minutes on a conventional commercial laptop equipped with an Intel Core i5 processor and 8 GB of RAM.

The analyst can readily identify the best-fit model by comparing the extracted results and determining the next steps. For the Car-sales dataset, the XGBoost (eXtreme Gradient Boosting) regression model produced the best fit (Adj.R<sup>2</sup>= 0.965) among the tested algorithms.

e) Findings and Key Take aways

Exploring datasets with Rbox+ across multiple ML models provides several advantages. It enables researchers to identify the model that best fits a given dataset, supporting informed decision-making while offering insight into the relative strengths and limitations of each approach [33]. Because every ML model carries inherent assumptions and potential biases [6], evaluating diverse algorithms helps assess the robustness of results and ensures that observed patterns are consistent across methods, thereby enhancing the generalizability of findings.

Additionally, models may assign varying levels of importance to features within the dataset, and comparing feature rankings across algorithms allows researchers to identify the most influential variables and gain insight into underlying data dynamics [36].

Employing multiple models also aids in detecting overfitting, a common challenge in which models perform well on training data but fail to generalize to unseen data [5]; this detection informs necessary adjustments or regularization strategies. Furthermore, combining models through ensemble or stacking techniques can improve predictive performance by mitigating individual model weaknesses and leveraging complementary strengths [28].

Finally, each ML model embodies unique assumptions, and exploring a range of algorithms provides alternative perspectives that can reveal hidden patterns, thereby enrich the depth of analysis and contribute to a more comprehensive understanding of the data [12].

f) Limitations and Future Research

While Rbox+ offers significant advantages, certain constraints remain. Current testing has focused on moderate-sized datasets; consequently, further research is required to assess scalability regarding high-frequency data streams and "Big Data" contexts. Additionally, the complexity of security protocols and access control in highly regulated environments necessitate deeper investigation. To address these gaps, future development could transition the "open box" framework toward more sophisticated capabilities, including automated hyperparameter tuning and distributed processing to manage large-scale data. A planned emphasis on deeper cloud infrastructure integration to enhance the system's flexibility and deployment readiness must also be considered.

## V. CONCLUSIONS

This paper introduces Rbox+, an innovative API designed as a comprehensive framework for the integration of machine learning systems. By bridging the gap between enterprise information systems and the Python/R programming environments, Rbox+ streamlines the analytics design cycle, minimizes implementation complexity, and enhances the operational utility of ML algorithms. The framework provides a unified, language-independent interface that allows complex statistical analyses to be executed seamlessly within existing corporate infrastructures.

Empirical evaluation indicates that Rbox+ efficiently executes multiple ML models within a single analytical workflow. In a demonstrated use-case, thirteen regression models were applied to a moderate-sized dataset, completing processing tasks on standard consumer hardware and thereby confirming the practical feasibility of the approach. Furthermore, the framework's ability to directly compare model performance using established metrics facilitates more informed model selection and validation.

The primary contributions of this study are fourfold. First, Rbox+ establishes a unified Web API that provides a generic and extensible interface, enabling enterprise systems to invoke ML algorithms without requiring native support for statistical languages. Second, the framework promotes transparency and reproducibility by exposing generated source code and delivering results in Jupyter Notebook format, which fosters greater trust in ML-driven decision support. Third, the practical applicability of the system is demonstrated through successful integration with SAP S/4Hana and Microsoft Excel. Finally, Rbox+ supports comparative ML evaluation by enabling the parallel execution of multiple algorithms, allowing analysts to efficiently assess predictive performance and robustness.

The identified limitations and future research directions for Rbox+ focus on four primary areas: scalability, data governance, security protocols, and functional expansion.

In conclusion, Rbox+ offers a transparent, flexible, and extensible approach to integrating machine learning into enterprise systems. By combining interoperability with reproducibility, the framework helps bridge the gap between advanced ML techniques and operational enterprise analytics, acting in parallel as an educational and learning tool.

### Conflict of Interest Statement

Spiros Chountas and Antonios Konomos declare no conflicts of interest.

## REFERENCES RÉFÉRENCES REFERENCIAS

- O. Ali, A. Nassif, L. Capretz: *Business intelligence solutions in healthcare a case study: Transforming OLTP system to BI solution*, IEEE 2013 Third International Conference on Communications and Information Technology (ICCIT), (2013), 209-214.
- Amazon, *Amazon Artificial Intelligence Service*, <https://aws.amazon.com/machine-learning/ai-services>, 2020.
- A. L. Arias, P. H. Westfall, *Understanding Regression Analysis: A Conditional Distribution Approach*, United States: CRC Press, 2020.
- D. Bardhan, A. Baumgartl, N. Choi, M. Dudgeon, *SAP S/4 HANA An Introduction*, SAP Press, Waldorf, 2021.
- M. M. Bejani, M. Ghatee: *A systematic review on overfitting control in shallow and deep neural networks*, Artificial Intelligence Review, 54(8), (2021), 6391-6438.
- A. Bellet, A. Habrard: *Robustness and generalization for metric learning*, Neurocomputing, 151(1), (2015), 259-267.
- E. Borgia: *The Internet of Things vision: Key features, applications and open issues*, Computer Communications, 54(2014), 1-31.
- P. Bruce, A. Bruce, P. Gedeck, *Practical Statistics for Data Scientists: 50+ Essential Concepts using R and Python*, O'Reilly, Sebastopol 2020.
- L. Cao: *Data Science: A Comprehensive Overview*, ACM Computing Surveys, 50(2017), 1-42.
- A. Cardoso, J. Leitão, C. Teixeira: *using the Jupyter notebook as a tool to support the teaching and learning processes in engineering courses*, The Challenges of the Digital Transformation in Education: Proceedings of the 21<sup>st</sup> International Conference on Interactive Collaborative Learning (ICL2018), Springer, 2, (2019), 227-236.
- Y. Chen, Z. Yuan, B. Chen: *Process optimization with consideration of uncertainties – An overview*, Chinese Journal of Chemical Engineering, 26(8), (2018), 1700-1706.
- V. Cox: *Translating Statistics to Make Decisions: A Guide for the Non-Statistician*, Apress, California 2017.
- J. Hao, T. K. Ho: *Machine learning made easy: a review of scikit-learn package in python programming language*, Journal of Educational and Behavioral Statistics, 44(3), (2019), 348-361.
- T. Haslwanter, *An Introduction to Statistics with Python*, Springer, Linz, 2016.
- B. Hazen, C. Boone, J. Ezell, D. Jones-Farme: *Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and application*, International Journal of Production Economics, 154(2014), 72-80.

16. J. C. Huang, K. M. Ko, M. H. Shu, B. M. Hsu: *Application and comparison of several machine learning algorithms and their integration models in regression problems*, Neural Computing and Applications, 32, (2020), 5461-5469.
17. IBM, *IBM Watson*, Online document <https://www.ibm.com/14icros>, 2020.
18. R. Ihaka, R. Gentleman: *A Language for Data Analysis and Graphics*, Journal of Computational and Graphical Statistics, 5(1996), 299-314.
19. A. Gandomi, M. Haider: *Beyond the hype: Big data concepts, methods, and analytics*, International Journal of Information Management, 35(2015) 137 – 144.
20. Google, *Google Cloud AI*, Online document <https://cloud.google.com/products/ai>, 2020.
21. B. E. Granger, F. Pérez: *Jupyter: Thinking and storytelling with code and data*, Computing in Science & Engineering, 23(2), (2021), 7-14.
22. N. Jukic, A. Sharma, J. Nestorov B.: *Augmenting Data Warehouses with Big Data*, Information Systems Management, 32(2015), 200-209.
23. M. B. Kery, M. Radensky, M. Arya, B. E. John, B. A. Myers: *The story in the notebook: Exploratory data science using a literate programming tool*, Proceedings of the 2018 CHI conference on human factors in computing systems, (2018), 1-11.
24. A. Konomos, S. Chountasis: *Rbox: A web API for software integration with the R programming language*, Computer Applications in Engineering Education, DOI: 10.1002/cae.2262, (2023).
25. Microsoft, *Microsoft Azure Cognitive Services*, Online document <https://azure.microsoft.com/en-us/services/cognitive-services>, 2020.
26. S. A. Naghibi, H. R. Pourghasemi: *A comparative assessment between three machine learning models and their performance comparison by bivariate and multivariate statistical methods in groundwater potential mapping*, Water resources management, 29(2015), 5217-5236.
27. A. Neumann, N. Laranjeiro, J Bernardino: *An Analysis of Public REST Web Service APIs*, IEEE Transactions on Services Computing, 14(4) (2021), 957-970.
28. B. Pavlyshenko, *using stacking approaches for machine learning models*, IEEE Second International Conference on Data Stream Mining & Processing (DSMP) (2018), 255-258
29. Project Jupyter, *Jupyter Documentation*, <https://docs.jupyter.org/en/latest/>
30. L. Richardson, S. Ruby, *RESTful Web Services*, O'Reilly Media, Cambridge, 2008.
31. W. Richert, L. P. Coelho, *Building Machine Learning Systems with Python*, Packt Publishing, 2013.
32. P. Y. Simard, S. Amershi, D. M. Chickering, A. E. Pelton, S. Ghorashi, C. Meek, G. Ramos, J. Suh, J. Verwey, M. Wang, J. Wernsing, *Machine teaching: A new paradigm for building machine learning systems*, (2017), arXiv preprint arXiv:1707.06742,
33. A. Singh, N. Thakur, A. Sharma: *A review of supervised machine learning algorithms*, 3rd International conference on computing for sustainable global development (INDIACom), (2016), 1310-1315.
34. H. Tanuwidjaja, R. Choi, S. Baek, K. Kim: *Privacy-preserving deep learning on machine learning as a service – A comprehensive survey*, International Journal of Computer Science and Network Security, 21(6) (2020), 137-42.
35. L. Zhao, Q. Wang, C. Wang, Q. Li, C. Shen, B. Feng: *VeriML: Enabling Integrity Assurances and Fair Payments for Machine Learning as a Service*, IEEE Transactions on Parallel and Distributed Systems, 30(10) (2021), 2524-2540.
36. A. Zien, N. Krämer, S. Sonnenburg, G. Rätsch: *The feature importance ranking measure*, Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD Bled, Slovenia, September 7-11, 2009, Proceedings, (2009) 694-709.



# GLOBAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY: C SOFTWARE & DATA ENGINEERING

Volume 26 Issue 1 Version 1.0 Year 2026

Type: Double Blind Peer Reviewed International Research Journal

Publisher: Global Journals

Online ISSN: 0975-4172 & Print ISSN: 0975-4350

## Blockchain Challenges: Advantages and Algorithms

By Zeel B Dabhi & Aishwarya

*Ajeenkya DY Patil University*

**Abstract-** Cryptocurrency is the innovation that has changed the way of life most significantly over the past ten years. Bitcoins is a term that often comes up when discussing the blockchain system. Although they are not identical, Ethereum and Cryptocurrency nevertheless remain widely misunderstood. Innovative technologies had to be created as a result from rising degrees of globalization. These groundbreaking innovations improve the speed of international trade. There are many technical experiments; some of them were successful, whereas others died or required development. The decentralized ledger technology, its benefits, and methods for consensus are described on this article.

**Keywords:** *the pros and disadvantages, and consensus method of the digital currency blockchain.*

**GJST-C Classification:** LCC: QA76.9.B58



*Strictly as per the compliance and regulations of:*



© 2026 Zeel B Dabhi & Aishwarya. This research/review article is distributed under the terms of the Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0). You must give appropriate credit to authors and reference this article if parts of the article are reproduced in any manner. Applicable licensing terms are at <https://creativecommons.org/licenses/by-nc-nd/4.0/>.

# Blockchain Challenges: Advantages and Algorithms

Zeel B Dabhi<sup>α</sup> & Aishwarya<sup>ο</sup>

**Abstract-** Cryptocurrency is the innovation that has changed the way of life most significantly over the past ten years. Bitcoins is a term that often comes up when discussing the blockchain system. Although they are not identical, Ethereum and Cryptocurrency nevertheless remain widely misunderstood. Innovative technologies had to be created as a result from rising degrees of globalization. These groundbreaking innovations improve the speed of international trade. There are many technical experiments; some of them were successful, whereas others died or required development. The decentralized ledger technology, its benefits, and methods for consensus are described on this article.

**Keywords:** the pros and disadvantages, and consensus method of the digital currency blockchain.

## I. INTRODUCTION

Innovative innovations had to be created as a result of the rising degrees of globalization. These novel technologies improve the efficiency of international trade. Here are many technical experiments; a handful of them were successful, while others died or required development. But without the advent of blockchain technology, a number of significant turning points are being reached, particularly in terms of computational innovations. The techniques that were utilized to produce this research centered on review of literature, analyses of the most commonly quoted projects, pattern findings, readings of reports, analysis of advances in technology, and research of the priorities of the major IT firms.

Cryptocurrency is a decentralized, trustworthy, and challenging to utilize for unlawful type of record keeping. On the opposite hand, Cryptocurrency is a type of electronic money that conducts operations between peers to peers using an open database called the Blockchain, or distributed ledger. Blockchain-based solutions is used in a number of different industries, including Cryptocurrency and Hyperledger's and intelligent contracts. Thus, a wide range of possibilities can be made using the technology of blockchain. The distributed ledger called Blockchain is undoubtedly a new sort of store. Although it can address one of the major issues relating to banking, this type of technology

**Author α:** MCA DS, School of Engineering, Ajeenkya DY Patil University, Pune (MH). e-mail: zeel.dabhi@adypu.edu.in

**Author ο:** Ajeenkya DY Patil University, Pune (MH). e-mail: Aishwarya@inurture.co.in

is quite intriguing to people. Ethereum is an innovation that combines a number of various technologies and tools, including consensus, networks of peers, the use of encryption, and arithmetic.

## II. BLOCKCHAIN TECHNOLOGY

It is a sort of modern technology in which a computerized register is employed to track operations throughout a decentralized computing infrastructure in order to prevent the operations throughout the machines from being changed retrospectively. Here, every member in the shared ledger gathers the details of each deal the other person engages in. The site does feature a scheduling mechanism, but it provides no way to delete an operation through it one time it has been officially completed.

### a) The definition of the Blockchain technology

Among the most commonly used descriptions of the Blockchain system, that was created by John & Alan Tapscott, who says it is "an infallible electronic register that records economic event that can be configured to store not only financial activities but nearly anything of significance" [1].

### b) The structure of the Blockchain technology

The digital ledger is made up of pieces that are introduced to the channel in a straight line at scheduled times [1]. However, the date, purchase, and hashes are present in all Ethereum implementations. The data included in the blocks of data varies on the distributed network of Blockchain.

A single block includes the preceding item's digital digest (Fig. 1).

Since every piece of knowledge in the hash algorithm is created on its own, it is impossible to alter any of its components.

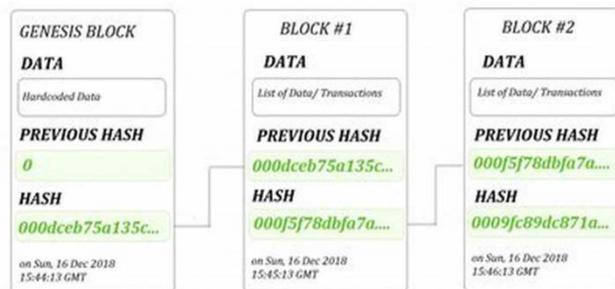


Fig. 1: The sequence of the hash value in the Blockchain

The authentication procedure, including applying the personal key and license, is shown in Fig. 2. Verification begins after the authentication step is complete. (Fig. 3). If the resulting hash results are identical, the check is successful.

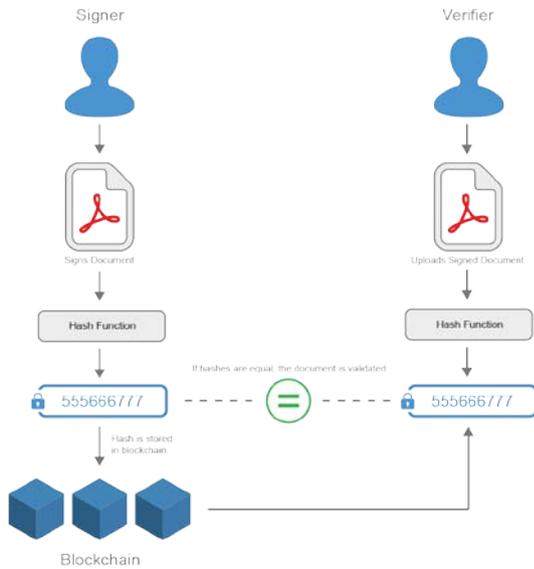


Fig. 2: The signing process in the Blockchain



Fig. 3: The verification process in the Blockchain

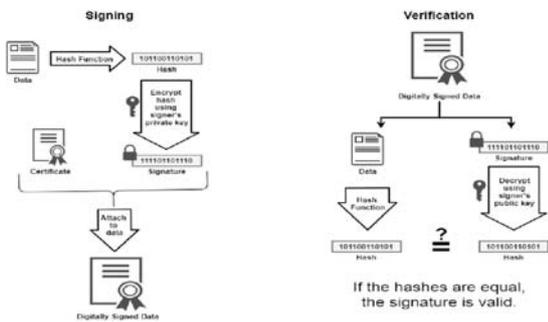


Fig. 4: The very simple shows, how blocks signing and verification processes work in the Blockchain

Fig. 5 illustrates the building block formation procedure. Every block in the current instance had the nonce, the Merkle shape, the date, the time, and the prior hashing value.

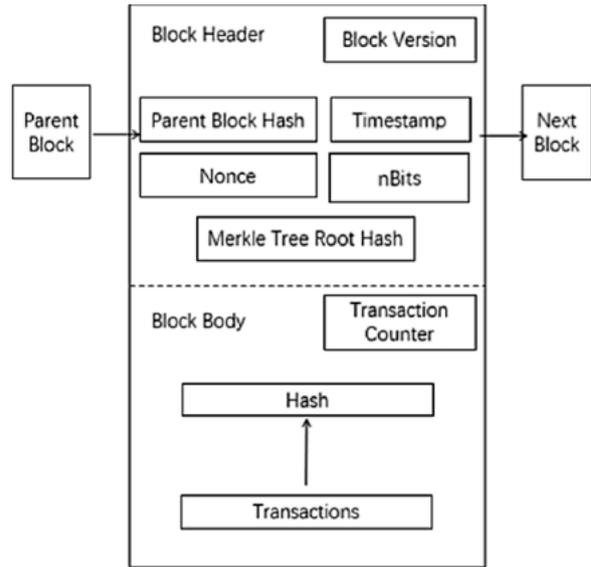


Fig. 5: The structure of the Blockchain

c) What are the key components of blockchain technology?

Distributed ledger:

- A distributed ledger is the shared database in the blockchain network that stores the transactions, such as a shared file that everyone in the team can edit. In most shared text editors, anyone with editing rights can delete the entire file. However, distributed ledger technologies have strict rules about who can edit and how to edit. You cannot delete entries once they have been recorded.

Smart contracts:

- Companies use smart contracts to self-manage business contracts without the need for an assisting third party. They are programs stored on the blockchain system that run automatically when predetermined conditions are met. They run if-then checks so that transactions can be completed confidently. For example, a logistics company can have a smart contract that automatically makes payment once goods have arrived at the port.

Public key cryptography:

- Public key cryptography is a security feature to uniquely identify participants in the blockchain network. This mechanism generates two sets of keys for network members. One key is a public key that is common to everyone in the network. The other is a private key that is unique to every member. The private and public keys work together to unlock the data in the ledger.

- For example, John and Jill are two members of the network. John records a transaction that is encrypted with his private key. Jill can decrypt it with her public key. This way, Jill is confident that John made the transaction. Jill's public key wouldn't have worked if John's private key had been tampered with.

### III. CHALLENGES

Additionally, blockchain science has applications in a number of commercial sectors. The medical field is one intriguing area where the blockchain technology is being used. Through the use of Ethereum for transferring costs via the digital currency, this meets everyone involved including medical facilities, medical facilities, and public health regulators by revealing consumers' expectations and maintaining the confidentiality of patients. If the public wanted to view the medical records of a person under the traditional framework, they'd have to fill out an inquiry form and send it to the registration location for permission. The details The buyer must pay an extra charge to the bookkeeper and get an invoice of payment after getting clearance. The receipt is subsequently presented by the details of the user to the registrations offices in order to get an electronic version of the medical records of the individual. But a patient's medical files. can be misplaced or copies made for nefarious intentions. Figure 6 illustrates the idea of a blockchain-based electronic medical record system.

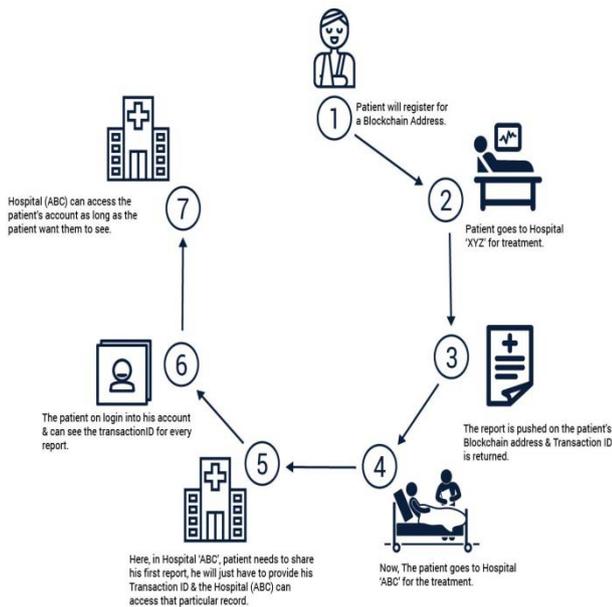


Fig. 6: The structure of the Blockchain

The Payment will indeed be put when such a data client asks a provider (a inpatient care facility) for a specific patient and the publisher approves. A main surgeon's and also the participant's consent are

required before providing a participant's medical files to an info buyer so that just particular documents, such as medical files, are provided. The specifics of this procedure will be described in later study.

#### a) Lack of adoption

Ledgers work better and more efficiently when used by a large network of users. For instance, a cryptocurrency environment may require suppliers as well as users to sign up for such platform. On the opposite hand, according to APQC, only 29% of businesses are actively testing with or using cryptocurrency. Blockchain networks will continue inefficient & unmanageable without widespread usage.

There are certain signs, nevertheless, that its popularity of blockchain may grow. Businesses are progressively organising cooperative blockchain collaborations to address related issues and offer answers that can be helpful to everybody without sharing private data.

This reduces dependency on physical logs with adheres data set and helps to protect the integrity of the distribution network.

#### b) The rising cost of blockchain implementation

The key is making early stock assets. Costs related to execution may be too high for some businesses. Despite the fact that the majority of latest methods are free, hiring skilled software engineers who specialise in blockchains, paying licencing fees in the event that one wishes to switch to a for-profit program version, book keeping, and other costs all require a large investment. It is among the most significant challenges experienced for cryptocurrency.

#### c) Scalability

Durability is the key issue with it's own deployment. Cryptocurrency is not practical for sizable apps because, despite the fact that purchase networks could really control thousands and thousands of simultaneous transactions without experiencing any problems, atm machines for Crypto currency (about 3–7 exchanges per second) and Cryptocurrency (about 15-20 exchanges per moment) takes a long time.

#### d) Security and privacy challenges

Now, numerous organizations must abide by legal restrictions. With regards to important information, their clients have faith in them. Nevertheless, this information won't be completely private if it's all preserved on a public ledger. Here, blockchains in corporate or consortium settings may be used. Your personal data would still be protected, and you would only have access to what you needed.

Cybersecurity is a further essential component. Yet, only a small number of circumstances have robust processes that can deal with this. Even while blockchain-based apps, systems, and businesses are

more secure than traditional pcs, criminals may still be able to access them.

He doesn't just want the authorities to safeguard our privacy. Ethereum ego IDs may enable us to gather and control our info. although there We've put in plenty of efforts towards creating new privacy procedures, including confirmation of negligible, but a new identity structure continues to be a long way off.

To learn whether cryptocurrency and MI may be utilized for safe storage of information, visit the page on cryptocurrency and AI secure data handling. It is among the most significant challenges experienced for cryptocurrency.

#### e) *Regulations*

The very next aspect that you might run into trouble is with the lack of regulations. It is possible for fraud and price fixing to lead to a world economic catastrophe. As a reason, Cryptocurrency is the subject of a lot of unfavourable media coverage.

Although some nations have openly prohibited cryptocurrency, someone else has made vain attempts to control blockchain platform.

#### f) *Criminal activities*

The proliferation of bogus enterprises as well as other bad actors looking to take advantage of naive participants has been encouraged by the absence of strict regulations and indeed the notion that blockchains is still in its infancy. A number of prominent crypto trading scams have indeed occurred, along with the notorious Largest Cryptocurrency bitcoin hack in 2014 that almost brought out the entire sector.

#### g) *Energy consumption*

The fact that evidence of work, the most popular compromise technique, consumes a great deal of electricity seems to be another cause for concern. This makes it difficult for average users to access Distributed consensus networks, promotes the creation of big mining pools, inhibits decentralisation by pressuring users to join these pools, and creates global pollution.

#### h) *51% attacks*

Block chain technology have an amazing feature. Certain are safer than in others. For instance, compared to centralised blockchain technologies, decentralised ledgers seem to be more susceptible to 51% attacks. For cryptocurrency traders who desire to hold financial funds on decentralised channels, this is what has created a few problems.

Several cryptocurrency platforms have been hampered by 51% assaults, wherein the criminals seize more than half of the channel's processing capacity. They take use of a flaw in decentralised systems that gives people access to over 51% of the processing capacity, giving them control over a chain. On systems that employ the concrete evidence paradigm, this

frequently occurs. The architecture of blockchain technologies is distinct. Some are more secure than others. The decentralized blockchains, for example, are more vulnerable to 51% attacks than the centralized ones.

#### i) *Low workforce availability*

These nonfungible currency and Describe businesses have experienced a sharp increase in nonfungible assets & enterprises over the past year, which has caused problems in the labour market. As per current data, as startups and existing businesses search out best players, the demand for blockchain talent has surged by more than 300%.

#### j) *Interoperability*

One of the most important issues that must be addressed is interoperability, as this is one of the primary reasons businesses are yet hesitant to embrace blockchain technology. Most blockchains are maintained in isolation and do not communicate with other peer networks since they cannot transmit and receive data from a different blockchain-based system.

#### k) *Lack of standardization*

What standards does ethereum now follow? Despite the abundance of connectivity, there is no global standard. As a result of no global standard, there are issues with accessibility, rising costs, and complicated processes. Blockchain technology has no specific version, which discourages investment opportunities and entrepreneurs from getting involved entering the market.

#### l) *Integration with legacy systems*

Another issue is how to integrate blockchain solutions with an existing system. If a business chooses to use cryptocurrencies, they must typically entirely replace their outdated system or create a plan to properly connect the two techniques.

Additional problem is that businesses without software engineers limit access to the skill pool needed to take part in this undertaking. Reliance on an external source may make this issue worse. Yet, to implement the majority of market mechanisms, the business must commit a substantial amount of time and money.

#### m) *Private key issues*

In a decentralised environment, credentials that people hold in a centralised environment may have become exposed.

After a wallet has been created, they allow access to all of its data. If stolen, it puts everyone wealth and personal information in danger. If the wallet is stolen or annihilated, access is permanently gone. That is one of the riskiest impediments for the cryptocurrency Destroyed. It is one of the most dangerous blockchain implementation challenges.

#### IV. ADVANTAGES

The blockchain technology's main benefit is that it is not regulated. What does it mean for our lifetimes? Merely expressed, there is no requirement for cooperation only with formal leader or a third party organisation. This suggests there's not a mediator in the design and that decisions are made by all owners of such virtual cash. A system software keeps a record of material, so must be adequately safeguarded as there's a danger that now the documentation may well be compromised if an institution works with other companies and might end up in the wrong hands due to misuse. There is a potential that the process of securing the data will be time-consuming and expensive. Using Btc Whenever transfers have been conducted It has the potential that its process of hiding the data will be time-consuming & expensive. When using Bitcoins Utilizing similar hardware is possible to be avoided because send or receive with Bitcoin generate their own evidence of legitimacy and power can enforce the restrictions. Moreover, it suggests that the actions might be verified and managed as a single The primary advantage of the distributed ledger tech is its decentralized nature. What is it significant to our lives? It is not needed to collaborate with the central administrator or an outside company, to put it simply. Every action taken is saved on the the distributed ledger, and the data in these records is accessible to all users and is unable to altered or removed. The Bitcoin's openness, constancy, and reliability are demonstrated by the outcomes resulting from this documentary.

- Each activity is recorded just on shared database, where the data is obtainable and cannot be changed or withdrawn. The outcomes of this film serve as evidence of the Blockchain's transparency, consistency, and dependability.
- Its Bitcoin relies on the belief of a number of people who are strangers to one another for its reliability. The key concept is that these are genuine, valuable interactions seen between unidentified parties. Since there might be additional provides valid and information, trust may be enhanced even more.
- When operations are approved and disseminated across the the distributed ledger, immutability is guaranteed. It isn't feasible to modify or remove an operation once it is linked with the Internet. It also hinges on the type of structure; if anything's centrally managed it might be modified or removed since a single individual makes a choice. But with a decentralized system, like the the distributed ledger, each purchase that is connected to the system gets duplicated to each computer in the community. This feature renders the public ledger technology impermeable and unbreakable. Data integrity is ensured when activities are authorised and spread

throughout the public ledger. After an action has been uploaded to the web, it is impossible to change or delete it So it depends on the organization's system; if something is locally administered, it could have to be changed or eliminated because only one person can decide. And even though, in a decentralised system, such as the public blockchain, every buy linked towards the framework was mirrored across all of the computers with in group. The underlying blockchain has been rendered impenetrable and indestructible by this function.

- The Chain provides its clients with the ability to control every transaction and data point. Whenever a hacker has access to advanced technology, they can alter or delete the data on the Bitcoins It has enough processing power to change or erase every piece of content just on country's notebooks, along with the content of the electronic record, prior to the preceding batch is placed in. Should there be few Contrarily, with numerous notebooks, your connection seems to be more transparent and safer. Companies on the Ledger, that is, the technology, are much more susceptible to attack.
- The blockchain technology is designed to be able to detect any problems and, if necessary, correct them. Staying identifiable is a feature of the Cryptosystem. The interaction between technologies and the represent the data achieves a substantial degree of safety upon every user's admission towards the network. As a result, each Bitcoin user is given a unique identifier which is linked to the account. The reliable encoded lanyard is yet another element contributing toward the Channel's security.
- The year and day the document was created, as well as the person's ID. The present incarnation of mined includes its merkel's stem, it contains information on earlier buys and associated hashes. The component automatically returns that amount. In this case, it is impossible to alter any particular part of the hash function.
- That advantage is really the fast performance. It normally takes a lengthy time to finish and initiate a contribution further into banking institution. By using cryptocurrency technologies, the cleaning and initialising procedure can be completed in a fraction of the time—from around five days to just few seconds or less.

##### a) *Immutability*

Blockchain technology enables data integrity, making it hard to alter or change data that has already been committed. As a result, the cryptocurrency stops data manipulation on the internet.

Conventional data are not impervious to change. The CRUD approach makes it simple to erase and replace data, while the traditional database

employs creation, access, update, and delete at the primary level to ensure proper application performance. Such information is vulnerable to modification by malicious employees or outside hackers. Blockchain supports immutability, meaning it is impossible to erase or replace recorded data. Therefore, the blockchain prevents data tampering within the network.

#### b) *Transparency*

Blockchain is decentralized, meaning any network member can verify data recorded into the blockchain. Therefore, the public can trust the network.

On the other hand, a traditional database is centralized and does not support transparency. Users cannot verify information whenever they want, and the administration makes a selected set of data public. Still, however, individuals cannot verify the data.

#### c) *Censorship*

Blockchain technology is free from censorship since it does not have control of any single party. Therefore, no single authority can interrupt the operation of the network.

Meanwhile, traditional databases have central authorities regulating the operation of the network, and the authority can exercise censorship. For instance, banks can suspend users' accounts.

#### d) *Traceability*

Blockchain creates an irreversible audit trail, allowing easy tracing of changes on the network.

The traditional database is neither transparent nor immutable; hence, no permanent trail is guaranteed.

#### e) *Open*

One of the major advantages of blockchain technology is that it is accessible to all means anyone can become a participant in the contribution to blockchain technology, one does not require any permission from anybody to join the distributed network.

#### f) *Verifiable*

Blockchain technology is used to store information in a decentralized manner so everyone can verify the correctness of the information by using zero-knowledge proof through which one party proves the correctness of data to another party without revealing anything about data

#### g) *Permanent*

Records or information which is stored using blockchain technology is permanent means one needs not worry about losing the data because duplicate copies are stored at each local node as it is a decentralized network that has a number of trustworthy nodes.

#### h) *Free from Censorship*

Blockchain technology is considered free from censorship as it does not have control of any single

party rather it has the concept of trustworthy nodes for validation and consensus protocols that approve transactions by using smart contracts.

#### i) *Tighter Security*

Blockchain uses hashing techniques to store each transaction on a block that is connected to each other so it has tighter security. It uses SHA 256 hashing technique for storing transactions.

#### j) *Immutability*

Data cannot be tampered with in blockchain technology due to its decentralized structure so any change will be reflected in all the nodes so one cannot do fraud here, hence it can be claimed that transactions are tamper-proof.

#### k) *Transparency*

It makes histories of transactions transparent everywhere all the nodes in the network have a copy of the transaction in the network. If any changes occur in the transaction, it is visible to the other nodes.

#### l) *Efficiency*

Blockchain removes any third-party intervention between transactions and removes the mistake making the system efficient and faster. Settlement is made easier and smooth.

#### m) *Cost Reduction*

As blockchain needs no third man it reduces the cost for the businesses and gives trust to the other partner.

## V. DISADVANTAGES

Although the account individuals has advantages, additionally there are downsides or problems with such kind of technology as well. The main problem with Crypto currency is how much electricity it consumes. For a continuous registration to be maintained, power must be provided.

If a new block appears, it engages with the rest of the nodes simultaneously. On this basis, truth is generated. The show's hidden content are engaged on a variety of problems every moment during the day in an effort to check functioning. They are using a lot of CPU resources. Every node offers high levels of dependability, ensures quality support, and renders data stored on the decentralized system dissent & unchangeable for all time. Even while the digital ledger provides benefits, there are drawbacks or difficulties with this form of technology. The significant power use of the Ethereum is its biggest drawback. Power usage is required to maintain an ongoing register.

Such processes waste precious resources because every node must replicate the action. Bit coin Bit coin has expanded as a result of the addition of additional bits towards the network and rising computational demands. Not all nodes have the national

resources offering. There really are two issues: the shorter logbook is the primary one since Both preservation and transparency of a Network are broken since the sites are unable to maintain the complete block chain; additionally, the Block chain shifts to a more centralised form of consent. A significant disadvantage of Ethereum is its excessive price. The average cost of the transfer is between \$75 to \$160, without electricity consumption taking up most of the costs [12]. One of the contributing factors to this situation has previously been highlighted. Additional reason is that it requires a substantial initial expenditure.

a) *Speed and performance*

Blockchain is considerably slower than the traditional database because blockchain technology carries out more operations. First, it performs signature verification, which involves signing transactions cryptographically. Blockchain also relies on a consensus mechanism to validate transactions. Some consensus mechanisms, such as proof of work, have a low transaction throughput. Finally, there is redundancy, where the network requires each node to play a crucial role in verifying and storing each transaction.

b) *High implementation cost*

Blockchain is costlier compared to a traditional database. Additionally, businesses need proper planning and execution to integrate blockchain into their process.

c) *Data modification*

Blockchain technology does not allow easy modification of data once recorded, and it requires rewriting the codes in all of the blocks, which is time-consuming and expensive. The downside of this feature is that it is hard to correct a mistake or make any necessary adjustments.

d) *Scalability*

It is one of the biggest drawbacks of blockchain technology as it cannot be scaled due to the fixed size of the block for storing information. The block size is 1 MB due to which it can hold only a couple of transactions on a single block.

e) *Immaturity*

Blockchain is only a couple-year-old technology so people do not have much confidence in it, they are not ready to invest in it yet several applications of blockchain are doing great in different industries but still it needs to win the confidence of even more people to be recognized for its complete utilization.

f) *Energy Consuming*

For verifying any transaction, a lot of energy is used so it becomes a problem according to the survey it is considered that 0.3 percent of the world's electricity had been used by 2018 in the verification of transactions done using blockchain technology.

g) *Time-Consuming*

To add the next block in the chain miners, need to compute nonce values many times so this is a time-consuming process and needs to be speed up to be used for industrial purposes.

h) *Legal Formalities*

In some countries, the use of blockchain technology applications is banned like cryptocurrency due to some environmental issues they are not promoting to use blockchain technology in the commercial sector.

i) *Storage*

Blockchain databases are stored on all the nodes of the network creates an issue with the storage, increasing number of transactions will require more storage.

j) *Regulations*

Blockchain faces challenges with some financial institution. Other aspects of technology will be required in order to adopt blockchain in wider aspect.

## VI. CONSENSUS ALGORITHMS

Each participant of the decentralized network must agree on the validity of a payment, its participation or deletion first from log, along with the subsequent blocks to just be recorded. The issue at hand is what each of these people can concur here on proper scenario given the facts on the tape until they all reach consensus. Any crypto currencies organization needs to agree on the past of operations because distributed systems lack governance or confidence amongst information system.

→ *Objectives of Blockchain Consensus Mechanism:*

a) *Unified Agreement*

- One of the prime objectives of consensus mechanisms is attaining unified agreement.
- Unlike centralized systems where having a trust on authority is necessary, users can operate even without building trust in each other in a decentralized manner. The protocols embedded in the Distributed blockchain network ensures that the data involved in the process is true and accurate, and the status of the public ledger is up-to-date.

b) *Align Economic Incentive*

- When it comes to building a trustless system that regulates on its own, aligning the interests of participants in the network is a must.
- A consensus blockchain protocol, in this situation, offers rewards for good behavior and punishes the bad actors. This way, it ensures regulating economic incentives too.

c) *Fair & Equitable*

- Consensus mechanisms enable anyone to participate in the network and use the same basics. This way, it justifies the open-source and decentralization property of the blockchain system.

d) *Prevent Double Spending*

- Consensus mechanisms works on the basis of certain algorithms that ensures that only those transactions are included in the public transparent ledger which are verified and valid. This solves the traditional problem of double-spending, i.e., the problem of spending a digital currency twice.

e) *Fault Tolerant*

- Another characteristic of the Consensus method is that it ensures that the blockchain is fault-tolerant, consistent, and reliable. That means, the governed system would work indefinite times even in the case of failures and threats.
- Currently, there are a plethora of Blockchain consensus algorithms in the ecosystem and many more are heading to enter the marketplace. This

makes it imperative for every Blockchain Development Company and enthusiastic Entrepreneur to be familiar with the factors that defines a good consensus protocol, and the possible effect of going with a poor one.

f) *The Bottom Line*

- Consensus mechanisms have become an essential aspect of distributed ledgers, databases, and blockchains because much of the world is becoming more digital. Ownership of physical assets is being tokenized on ledgers and blockchains, people without access to financial services have access through blockchains, and businesses need data security more than ever.
- Consensus mechanisms verify data inputs and outputs, which translates to automatically auditing the digital transactions that are common today—without human oversight or intervention. They create an environment where you don't need to trust that the other party in a transaction is honest because they ensure the information is unalterable and secure.

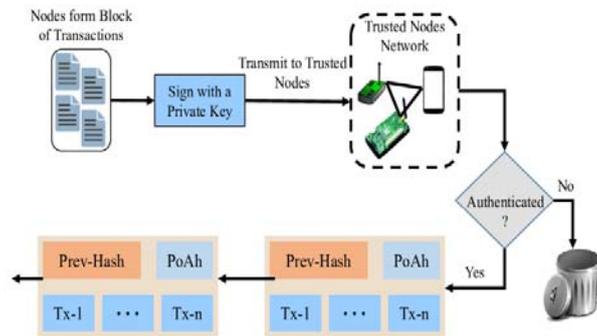
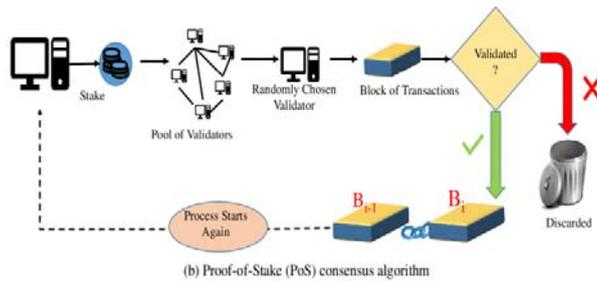
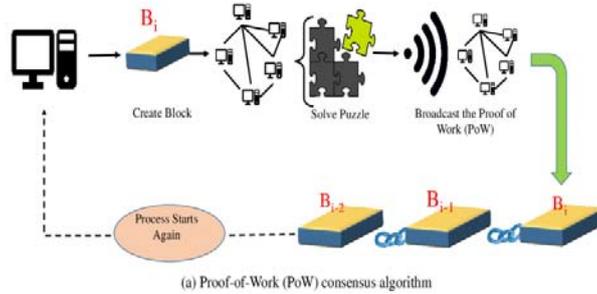


Fig. 7: Various consensus algorithms which can be used in a blockchain technology.

g) *Proof of work*

- There will be proof of promise soon. This system composed of crypto currencies uses the early acceptance process known as confirmation of labour. The argument for such working notion is founded on the notion that every organisation in some kind of a node hierarchy competes to figure out the right hashing value in order to add a new transactions to the block chain and get the payment, as seen in image 8.
- There are numerous unique sorts of employment certification in literature. Demonstrating work is a great method to negotiate an agreement, but it has a major economic disadvantage. Costly computer-based labor evaluation. This motivates us to propose fresh agreement strategies to deal with the evidentiary labor.

```

run:
Please, guess anumber between 1 and 100.
Type 'y' If my guess matches.
Type 'n' If my number does not match.
Type 's' to start the game
s
I guess the number 90
Am I right
n
I guess the number 53
Am I right
n
I guess the number 75
Am I right
n
I guess the number 9
Am I right
y
Hurrah! I guessed your number 4 in attempts!
BUILD SUCCESSFUL (total time: 10 seconds)
    
```

Fig. 8: Proof of Work consensus pseudo code

h) *Proof of stake*

- Ledgers use the proof of stake (Pops) consensus process. It establishes which individual or persons authenticate brand-new transactional chunks yet are rewarded by being rewarded successfully. Undeservedly, crypto currency seems to have a notoriety for just being difficult to understand and impregnable. proof of stake consensus algorithm there is less computation performed.
1. Delegated Proof of Stake (DPoS)
    - One decision matrix is delegation to demonstrate stake. The primary concept behind it is that the shareholders should be allowed to choose a manager who will support you and possibly pass along other benefits as well. These leadership can indeed be elected or removed at multiple times over time, and they create bricks since round form instead of in a sequence.

- In Dopes, the miners stake their coin and vote for a particular number of delegates, in a way that, the more they invest, the more precedence they receive. They get rewards in terms of coins or transaction fees.
  - In DPoS, there are 21-100 delegates charged periodically and assigned to deliver their blocks. Having fewer delegates allows for an efficient organization to design time slots for publishing blocks in the network. In case of, insufficient, invalid, or missing block publishing, the miners vote them out to be replaced with other selected delegates.
  - As DPoS works on the stake-weighted voting system, it has become one of the fastest growing and adapted blockchain consensus models.
2. Leased Proof of Stake (LPoS)
    - LPoS operates on Waves' blockchain platform and is an advanced version of PoS.
    - In LPoS, users lease crypto tokens to the node that wants to act as a block producer for the network. A node with the maximum number of staked tokens is more likely to be selected for the next block generation as well as receive rewards.
    - It also helps users with smaller tokens who might not have been eligible for participating as the blockchain creator in the traditional proof of Stake process in pooling their assets while enhancing their chances of receiving network transaction fees' share.
    - The leased proof of stake consensus algorithm is best for networks with high high-technical requirements for operating full nodes capable of verifying and validating transactions.

i) *Delegated proof of stake*

Delegated proof of stake is another consensus algorithm. Its idea main about the stakeholders able to select a leader who votes for them and potentially passes some rewards as well. These leaders can be voted in or out at different times and they produce blocks in around robin fashions so they do not get to put them all in a row.

1. Leased Proof-Of-Stake (LPoS)
  - The leasing stake evidence is also another variation on the traditional evidence of stake. Its Wave propagation network exposed us to the based on block chain technology consensus method. Similar to every other system for block chains, Waters wants to make sure to provide a better catch with less electricity usage.
  - There were several restrictions on staking in the original proof of stake. Those with a little number of coins might never truly take part in the staking. The infrastructure is capable of being managed to keep by a small number of people who have more currencies to donate...

#### j) Proof of activity

Proof of activity consensus algorithm is proposed, it is a hybrid approach that includes proof of work and proof of stake. It starts with a proof of work allowing miners to mine empty template without any transactions then it switched to proof of stake where validators select a block to sign and rewards get split between both proof of work minor and the stake.

#### k) Proof of authority

The Evidence of Activity electricity usage was also resolved by this protocol. Di Gooseneck faucet . the faucet et al. presented an official proof of unanimity, with the idea focusing on auditors or credentials that are approved or public identity, requiring them to manage what is known as an authorities station.

#### l) Proof of space or proof of capacity

The distinction between several work evidence and then this methodology is that here, the network node reserves a certain amount of storage or storage in to resolve challenging phrases in so it can get towards the ability to add another transaction, as opposed to the use of computational power or calculation power. It is a wise strategy to apply proof of work in more depth with additional resources.

#### m) Proof of importance

It is a more comprehensive variant of stake verification, with the premise that could for ought to be included in addition towards the serious risk or number of coins. The drawback of it is the energy wastage it causes.

It overcomes Transaction processing restrictions by giving priority to miners according to the amounts of transactions matching to each denomination. There in context of Poi, the greater the number of payments between and within a user's crypto currency, the greater the likelihood that even these users will be awarded crypto currency coal mines.

#### n) Proof of burn

It is a more comprehensive variant of stake evidence, with the premise that some measures will be factored in in addition towards the serious risk or number of bit coins. The drawback that it brings is the environmental wastage it causes.

#### 1. The Pros and Cons of the Proof-of-Burn Algorithm

- The main objective of burning the coins is to find out the strength. We are aware that lengthy players constantly keep coins for a very long time so as to benefit.
- By providing less strong currency with huge commitment, our system benefits such protracted stakeholders. Besides which, this improves decentralization and develops a more equitably spread networks.

#### 2. Practical byzantine fault tolerance

A useful bureaucratic fault detection technique addresses the problem of hostile nodes within a network. The dispersed network device may well be enabled to reach an accord regardless of certain nodes that are failing or delivering false info because the virtual pad uses a replicate technique to accommodate bureaucratic defects.

- PBFT makes an effort to offer a robust logic replicating that functions even in the existence of malicious nodes.
  - A primary node (or the leader) and numerous subsidiary nodes are successively arranged in micro services with bet (or the backups). Any eligible node in the system has the ability to switch form intermediate to major in the case of a major node failure. All trustworthy locations can participate using the majority rule.
- #### 3. Delegated Byzantine Fault Tolerance (dBFT)
- Daft is a consensus method that offers outstanding fraud protection. The mining are given the duty of voting for the members but are not contingent on our participation quantities, which is somewhat similar to the Dopes approach. The only requirements for becoming a deputy yourself are the appropriate tools, biometric identification, and 10,000 GAS.
  - Genuine voting is used by identification to choose the consensus algorithm, this enhances the method and cuts down on condition that makes for payments. The chosen team of auditors then generates new blocks using the BFT process.

#### o) Ripple

Inside one bigger system, the ripples protocol performs well. It can make use of the existence of reliable thread. Instead of operating in a competition fashion, it operates cooperatively; the nodes collaborate to determine the authenticity and sequence of events with in channel.

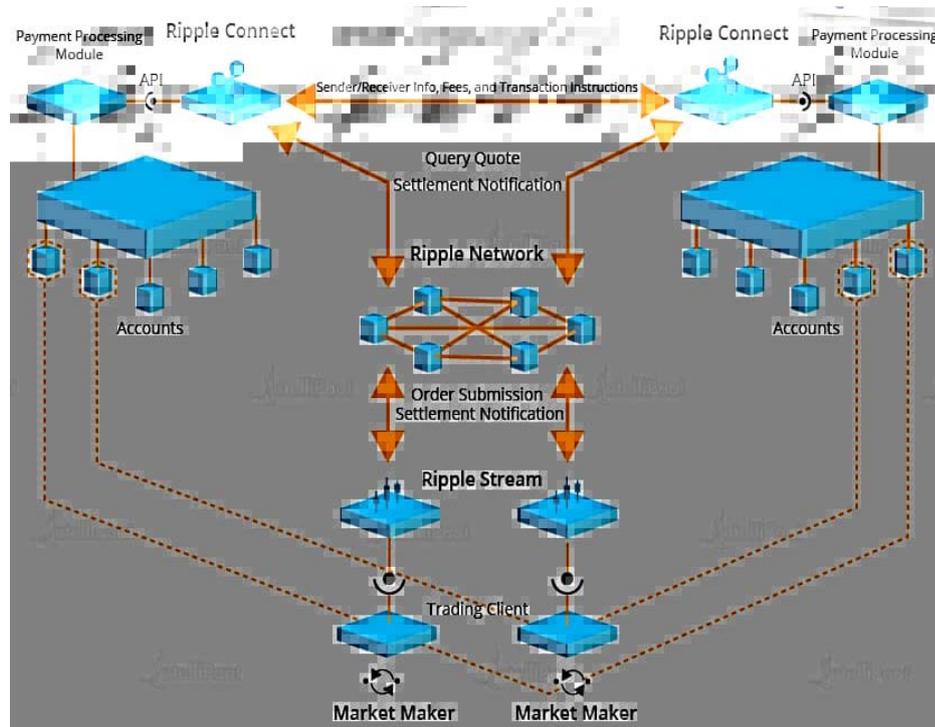


Fig. 8: Ripple Blockchain

p) Proof of Elapsed Time (PoET)

- Among the top voting systems is Poet. This specific technique is primarily utilized in public block chain networks, whereby networking access requires authorization. Those authorization systems must choose ballot or oil rights policies.
- The Poet techniques employ a specific strategy for masking openness over the entire system to guarantee that everything operates as planned. Since the networks required identity before allowing a user to engage the mining, the Approval methods also guarantee a secure entrance into in the platform.

→ Properties of a Good Blockchain Consensus Mechanism:-

1. Safety

- ✚ Together all stations in a competent consensus algorithm have able to produce outcomes that seem to be legitimate in accordance with the program's requirements.

Inclusive

- ✚ A strong consensual smart contract makes sure that each specific system node gets involved in the election system.

Participatory

- ✚ A decent convention on Ethereum entails a system in which all sites engage and help to maintaining datasets. updating databases on Blockchain is called a good consensus model.

Egalitarian

- ✚ Giving each ballot obtained first from network similar worth and importance seems to be another quality of a successful process.

→ Consequences of Choosing a Bad Consensus Protocol:

2. Blockchain Forks

- ✚ Using a subpar consensus mechanism technique makes the network more susceptible. Ethereum conflicts are but one weakness that block chain enthusiasts and developers must deal with. In simpler terms, a crypto currency splitting occurs when one or so more chains split off into another one or more.

- ✚ In the video that is posted below, a thorough explanation of block chain forks and their varieties is provided. In the video that is posted beneath, a comprehensive description of crypto currency forks and their varieties is provided.

- ✚ That whenever a fork in the Ethereum happens, the app comes to operate erratically, leading to two or more diverging nodes forward.

3. Poor Performance

- ✚ That whenever a poor block chains method is taken into account, whether the node malfunctions or experiences net division. As a result, the software's latencies grows and the operation of transmitting letters amongst servers takes longer, lowering its level of play.

#### 4. Consensus Failure

Majority failures is a result of using a poor trust model in your company strategy. In this case, a small percentage of nodes refuse to engage inside any transaction, and without their voting, the agreement is unable to produce the intended and correct results.

### VII. CONCLUSION

It is clear that crypto currency has several benefits and applications, including the capacity to operate in a decentralized mentor net devoid of a centralized government and to send money over the globe more cheaply. Digital medical records using block chains. It will take into account how a foreign entity can utilize or seek a child's health history from a medical facility or other body while violating the participant's right to privacy.

Crypto currency is dependable and unbreakable due to its benefits, including visibility, confidence, extra copies of activities, and a decentralized database. The aforementioned threats might only affect how the system functions, not the innovation itself.

We examined lightning network in this study and underlined the most recent research on

In this survey, we presented a survey of blockchain technology and highlighted the latest studies in blockchain and consensus algorithms.

Given the number and complexity of these blockchain issues, it would be unrealistic to think they are not major roadblocks to its adoption. In general, though, many of blockchain's greatest obstacles reflect growing pains typical with any new technology. Blockchain advocates will need to persuade their organizations to take similar risks, establish comparable relationships, and make similar trade-offs in other business areas to make a business case for adoption.

### ACKNOWLEDGMENT

We would like to thank our Professor Aishwarya, for his invaluable guidance and support throughout the research process. His expertise and insights were instrumental in shaping the direction and focus of our research. We are also grateful to the MCA Data Science at Ajeenkya DY Patil University for providing us with the resources and support we needed to complete this project.

### REFERENCES RÉFÉRENCES REFERENCIAS

1. Julija Golosova, Andrejs Romanovs., the Advantages and Disadvantages of the Blockchain Technology, November 2018.
2. Pinyaphat Tasatanattakool, Chian Techapanupreeda, Blockchain: Challenges and Applications, January 2018.
3. Stefan Forsstrom, Blockchain Research Report, December 2018.
4. Samar Al-Saqqa, Blockchain Technology Consensus Algorithms and Applications: A Survey, 30 December 2019.
5. Soliman Soliman, Yao Xu, Yifei Shen, blockchain technology, | April 2019.
6. Nakamoto, Satoshi, and A. Bitcoin. "A peer-to-peer electronic cash system." Bitcoin pdf. 2008
7. Popov, Serguei. "A probabilistic analysis of the nxt forging algorithm." Ledger. 2016.
8. Zheng, Zibin, Shaoan Xie, Hongning Dai, Xiangping Chen, and Huaimin Wang. "An overview of blockchain technology: Architecture, consensus, and future trends." In IEEE International Congress on Big Data. 2017.
9. Blockchain, April. 10, 2020
10. R Bhme, N Christin, B Edelman, Bitcoin: Economics, technology, and governance, 2015.
11. M Iansiti, KR Lakhani, The truth about blockchain, 2017.
12. M Crosby, P Pattanayak, S Verma, V Kalyan Raman, Blockchain technology: Beyond bitcoin, 2016.
13. "History of bitcoin," Nov 2018. Available: [https://en.wikipedia.org/wiki/History\\_of\\_bitcoin](https://en.wikipedia.org/wiki/History_of_bitcoin).
14. CNNMoney, What is Bitcoin, [online] <http://money.cnn.com/infographic/technology/what-is-Bitcoin/>.
15. Vitalik Buterin, Ethereum and The Decentralized Future, April 2015.
16. Christian C., Elli A., Angelo De Caro, Andreas K., Mike O., Simon S., Alessandro S., Marko V., et al, Blockchain, cryptography, and consensus, June 2017.
17. Ripple, Ripple Net, [online], <https://ripple.com>.
18. Eray Eliaçık, Blockchain challenges, May 30 2017.



# Towards Optimized K means Clustering using Nature-Inspired Algorithms for Software Bug Prediction

By Kajal Tameswar, Geerish Suddul & Kumar Dookhitram

*University of Technology (UTM)*

**Abstract-** In today's software development environment, the necessity for providing quality software products has undoubtedly remained the largest difficulty. As a result, early software bug prediction in the development phase is critical for lowering maintenance costs and improving overall software performance. Clustering is a well-known unsupervised method for data classification and finding related patterns hidden in datasets. However, the k-means algorithm has the tendency to converge to local optima due to its sensitivity to its initial partition and random initialization of clusters centers. On the other hand, Nature-inspired algorithms (NIAs) are known for their general ability to establish global optima while searching around the whole search place. When these algorithms are combined with the K-means clustering mechanism, the novel hybrids are projected to yield outstanding results in terms of enhancing clustering quality by avoiding local optima and uncovering global optima. This study shows that the hybrid clustering of the Coral reefs algorithm outperforms the typical K-means specification in terms of prediction accuracy.

**Keywords:** data clustering, K-means algorithm, Nature-inspired algorithms, software bug detection, coral reefs.

**GJCST-C Classification:** DDC Code: 005.1 LCC Code: QA76.76.D47



Strictly as per the compliance and regulations of:



# Towards Optimized K means Clustering using Nature-Inspired Algorithms for Software Bug Prediction

Kajal Tameswar<sup>α</sup>, Geerish Suddul<sup>σ</sup> & Kumar Dookhitram<sup>ρ</sup>

**Abstract-** In today's software development environment, the necessity for providing quality software products has undoubtedly remained the largest difficulty. As a result, early software bug prediction in the development phase is critical for lowering maintenance costs and improving overall software performance. Clustering is a well-known unsupervised method for data classification and finding related patterns hidden in datasets. However, the k-means algorithm has the tendency to converge to local optima due to its sensitivity to its initial partition and random initialization of clusters centers. On the other hand, Nature-inspired algorithms (NIAs) are known for their general ability to establish global optima while searching around the whole search space. When these algorithms are combined with the K-means clustering mechanism, the novel hybrids are projected to yield outstanding results in terms of enhancing clustering quality by avoiding local optima and uncovering global optima. This study shows that the hybrid clustering of the Coral reefs algorithm outperforms the typical K-means specification in terms of prediction accuracy.

**Keywords:** data clustering, K-means algorithm, Nature-inspired algorithms, software bug detection, coral reefs.

## 1. INTRODUCTION

In an era of technological disruption, the demand for software adoption has accelerated. They are a part of our society and play an important role in shaping it. Our modern society is becoming increasingly reliant on complex software systems. Thus, it is critical to build reliable and trustworthy systems in a cost-effective and timely manner. The presence of defective modules in a software drives up development and maintenance expenses, leading to customer dissatisfaction. The need for quality assurance has inevitably remained the biggest challenge in today's software development environment. Hence, software bug prediction is an important task to help developers locate bugs more efficiently.

Software bug prediction is an imperative task in Software Development Life cycle (SDLC) as it pertains to the overall success of software. One method in this direction is to use machine learning (ML) methods to predict defects in software. In addition, implementing

this method earlier in the SDLC process enhances quality of the product and lowers the cost of software maintenance. Many researchers have applied different theories and methodologies in the field of software bug prediction. Two things are clear from the literature when it comes to defect prediction. Initially, no single prediction approach dominates (Lessmann et al., 2008), and next, the employment of various set of data, data pre-processing, validation systems, and performance statistics makes it challenging to make sense of the multiple prediction outcomes (Myrtveit et al., 2005). There are two common ML model used for prediction based on dataset availability. The first, known as supervised approach, in which a software defect prediction model is built from training set of data and then tested on a testing dataset. Secondly, unsupervised approach, in which the defect prediction model for software is built from scratch using the present testing dataset without training the dataset.

Clustering algorithms have been commonly used to evade the lack of training datasets available being a constraint. Cluster analysis groups things into clusters based on their similarity to create a visual representation of data (Jain and Dubes, 1998). As pointed out by Kaur, 2010, one of the better instances of unsupervised learning is K-means clustering. Clustering is beneficial because it makes it easier to obtain or locate relevant information at a faster rate. Among the different clustering approaches that already exist, the K-means methodology is obviously fairly popular. (Gayathri et al., 2015). The preliminary values of the initial centroids, which are generated randomly each time the algorithm is run, have a significant impact on the performance of k-means. K-means frequently fall into local optima that produce poor clustering results. Obtaining a globally optimal clustering result involves a time-consuming, exhaustive approach that tests all partitioning choices. A heuristic approach to the problem is to use an optimization algorithm to search for global optima in each computer iteration.

Our unsupervised approach uses the k-means approach to divide the unlabeled dataset into defective and non-defective non-overlapped clusters for bug prediction. The goal of this research is to verify the hybrids' efficacy as well as to quantify the quality of results produced by each clustering hybrid model. In

Author <sup>α</sup> <sup>σ</sup> <sup>ρ</sup>: University of Technology (UTM), Pointe-aux-Sables, Mauritius. e-mails: ktameswar@umail.utm.ac.mu, g.suddul@umail.utm.ac.mu, kdookhitram@umail.utm.ac.mu

this study, we have applied the k-means clustering algorithm, an unsupervised algorithm with different NIAs including Genetic algorithm (GA), Bat algorithm (BA), Particle Swarm Optimization (PSO), Coral Reefs Optimization (CRO), Cuckoo Search optimization (CSO) algorithm, Ant colony optimization (ACO), Firefly algorithm (FA) and Grey Wolf Optimizer (GWO) for software bug prediction. The rest of this paper is organized as follows. Section 2 presents a discussion of the related work in software bug prediction. An overview of the methodology, consisting of the algorithms used are presented in Section 3. Section 4 describes the proposed method. Section 5 describes the Dataset and Data Processing method. The evaluation methodology is discussed in section 6. The results and discussion part is discussed in Section 7. Section 8 discusses the practical implications followed by conclusions and future works in section 9.

## II. RELATED WORKS

K-means clustering is a well-known partitioned clustering algorithm that has been used in a variety of applications. In the literature, several variations of K-means have been proposed to improve its performance for the broad clustering problem. Fong et al. (2012) studied the integration of bio-inspired optimization methods into K-means clustering for software bug prediction in order to assess clustering performance. The main optimization algorithms tested include the Firefly algorithm, Cuckoo search algorithm, Bat algorithm, Wolf and Ant Colony Optimization (ACO) algorithms. Results show that the combination of these algorithms acquired improved performance accuracy compared with ordinary k-means, at the same time accelerating the search process and avoid local optima. Zhong et al., 2004 compared the k-means algorithm to natural-gas algorithms. The natural gas algorithm outperformed the k-means algorithm in terms of mean square error values. However, this method necessitates the use of a software expert to determine whether the software is appropriate.

Annisa et al., 2020, came up with an improved version of k-means algorithm for software bug prediction, that locate the initial centroid of the k-means algorithm and determine the number of clusters present. Because it produces better accuracy than the simple K-Means method, this proposed method could be useful for clustering other data types. Seliya and Khoshgoftaar, 2007 proposed K-means for software failure prediction. Their method iteratively labels clusters as fault-prone or not using expert domain knowledge as a restriction.

The k-means algorithm based on quad tree was proposed by Bishnu and Bhattacharjee, 2012 and it was compared to some clustering algorithms. Their proposed algorithm has error rates that are comparable to k-means, Linear Discriminant Analysis and Naive

Bayes. Catal et al. 2009 used the x-means clustering algorithm to create faulty and non-faulty clusters based on software metrics. Lines of code, cyclomatic complexity, operand and operator are the metrics. If the metric values are complex than the threshold, the software entity is predicted to be defective, and vice versa. Almayyan, 2021 used dataset from the NASA repository and used three clustering algorithms, Farthest First, X-means and Self-organizing map. This article presents a comparison of software defect prediction algorithms based on Bat, Cuckoo, Grey Wolf Optimizer (GWO), and Particle Swarm Optimization (PSO) in order to evaluate different feature selection algorithms. The Farthest First clustering algorithm was found to be effective in predicting software faultiness, and Bat and Cuckoo were found to be useful in comparison to all other metaheuristic algorithms.

Though several academics have sought to merge K-means clustering with nature-inspired algorithms (NIAs), their efforts have been restricted to almost identical group movements, such as the Firefly, Artificial Bee Colony (ACO), and Particle Swarm Optimization (PSO) algorithms (Jensi and Jiji, 2015). In addition, only a few bio-inspired optimization methods that are integrated with K-means are provided in the previous studies. Only 7 of the 28 NIAs hybridized with K-means (Genetic Algorithm, Particle Swarm Optimization, Bat Algorithm, Artificial Bee Colony, Differential Evolution, Harmony Search, and Symbiotic Organism Search) dedicated their hybridization to solving automatic clustering problems, accounting for 20.6 percent of the total (Ikotun et al., 2021). In general, it can be seen that the rate of publishing on K-means hybridization with specific NIAAs is minimal. More research is needed in this area to see if there are any other ways to improve the performance of the existing hybridization algorithm. This suggests that combining K-means with these other NIAs to solve automatic clustering problems should be investigated.

The purpose of this research is to look into the mechanics of incorporating certain NIAs into the K-means clustering algorithm. The optimization function adds to the existing best solution by progressively improving it with a new solution from an unknown fragment of the search space. When a new solution is identified to be better than the present one, the searching agents replace the solutions and continue searching until some stopping criteria are fulfilled.

## III. METHODOLOGY

### a) *K means Clustering Algorithm*

The K-means clustering algorithm is a partitioned clustering technique that divides a dataset into k number of clusters using a certain fitness measure. Due to the large amount of data objects in real-world datasets, distributing data items into

appropriate clusters to obtain an ideal cluster outcome is computationally expensive and time-consuming (Ikotun et al.2021).

Given a dataset  $X = \{x_i\}$ , where  $i = 1, 2, \dots, n$  of  $d$ -dimension data points of size  $n$ ,  $X$  is partitioned into 'k' clusters such that

$$J(C_k) = \sum_{x_i \in c_k} \|x_i - \mu_k\|^2 \quad (1)$$

With the objective function: minimize the sum of the square error over all the  $k$  clusters. That is, minimize

$$J(C) = \sum_{k=1}^K \sum_{x_i \in c_k} \|x_i - \mu_k\|^2 \quad (2)$$

When assigning  $N$  objects to  $k$  clusters, the purpose of the clustering algorithm is to limit the number of potential possibilities. This can be expressed numerically as:

$$S(N, K) = 1/K! \sum_{i=0}^K (-1)^{K-i} \binom{K}{i} i^N \quad (3)$$

#### b) Nature-inspired algorithms (NIAs)

Nature-inspired computation has gained popularity in the previous two decades and has been used in practically every field of research and engineering (Yang et al.2013). NIAs are global optimization strategies for solving difficult real-world issues (Okwu et al. 2020). NIAs have successfully provided suboptimal solutions to automatic clustering problems in a reasonable amount of time (Hruschka et al. 2009). The population is used for the exploration of search space in the nature-inspired metaheuristic, ensuring a higher possibility of finding optimal cluster partitions (Nanda and Panda, 2014). It has been discovered that combining K-means with NIAs for automatic clustering improves the performance of algorithms when dealing with cluster analysis. In most circumstances, the automatic cluster number determination aids in the selection of near-optimal starting cluster centroids for the clustering process rather than the normal random selection (Zhou et al. 2017).

#### c) Combination of k-means with Nature-Inspired Algorithms (NIAs)

Clustering using NIAs is now as simple as assigning combinations of centroids to the searching agents, allowing them to heuristically find the best answer. Though the specifics of conducting a heuristic search vary depending on which nature-inspired optimization algorithm technique is used, the initialization stage and the finishing step, where the quality of the discovered solution is evaluated as a stopping condition, are both comparable.

$S$  is defined as the solution space that contains a finite number of  $x_i$ , where  $i$  is the solution's index, in the initialization construct. The search agents represent the solutions  $x$ , each of which holds a set of centroids, regardless of the types of bio-inspired optimization methods used.

Typically, a large population of searching agents,  $N$ , is utilized to collaboratively search for the best feasible cluster configurations (as expressed by the locations of the optimal centroids).  $K$  is the number of clusters that must be formed, which is generally a user-defined figure.  $D$  is the dimension of the search space, which is the number of attributes a data point possesses.

To find the optimal configuration of centroids we let  $cen_{j,v}$  be the centroids at the  $j^{\text{th}}$  cluster and the  $v^{\text{th}}$  attribute. To obtain the centroid location, the following formula is used:

$$cen_{j,v} = \sum_{i=1}^S w_i x_i^j / \sum_{i=1}^S w_i, \quad \text{Where } j=1 \dots K, v=1 \dots K * D \quad (4)$$

In our concept, the matrix  $cen_{j,v}$  contains all of the cluster centers and is a two-dimensional matrix with  $K * D$  characteristics.

$$F(cen) = \sum_{j=1}^K \sum_{i=1}^S W_i \sum_{v=1}^{K * D} (X_{iv} - cen_{j,v})^2 \quad (5)$$

The calculation method loops  $K * D$  times to analyze the values of all the attributes of  $x$  in each cluster  $v$  to calculate the distance between each  $x$  and the centroid.

Cluster centers can be designated by data points. For example, in a two-cluster clustering task, the objective function requires three variables. As a result, there are three dimensions.

Three variables, and hence three-dimensional spaces, are required, and the  $i^{\text{th}}$  data point may be written as  $x_i = (i, [x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4}, x_{i,5}, x_{i,6}])$ .

The clustering strategy can be formulated as follows:

$$clmat_{i,j} = \min_{k \in K} \{ \|X_i - cen_k\| \} \quad (6)$$

Where  $i=1 \dots N$ ,  $j=1 \dots S$ ,  $k=1 \dots K$ . Equation (3) tells us that the  $i^{\text{th}}$  data point belongs to the  $k^{\text{th}}$  cluster. The equation is an objective function with a lower value indicating better performance.

Sets of functional parameters must be defined in order to execute the bio-inspired optimization algorithms. Despite the fact that some of their parameters are shared, each set of parameters for the hybrid bio-inspired clustering algorithms is designed independently. The six models investigated are K means with Genetic Algorithm, K means with Bat algorithm, K means with Ant colony algorithm, K means with Cuckoo Search Algorithm, K means with Firefly Algorithm and K means with Coral reefs algorithm. The most significant variations are in how the global optimal exploration is carried out for all these algorithms. The evaluation stage comes right after the exploration construct, and it compares if the new solution is better than the current best one.

#### d) Genetic Algorithm

Genetic Algorithm (Ga) are randomized heuristic search algorithms that are based on natural

selection and genetic principles (Goldberg, 1989). The genetic operators used in the combination of K-means and GA are selection, distance-based mutation, and the K-means operator. The parameters have been set according to the study of Bouhmala et al. 2015.

P (0) is chosen at random as the starting population. Each allele in the population can be given a cluster number from the uniform distribution over the set  $\{1, \dots, K\}$  at random.

According to the distribution given by, the selection operator selects a chromosome from the preceding population at random as follows:

$$P(s_i) = F(s_i) / \sum_{j=1}^N F(s_j) \quad (7)$$

Where  $F(s_i)$  represents fitness value of the string  $s_i$  in the population.

The possibility of solutions surviving in the future population is ranked in the current population. Each solution in the population must be assigned a figure of merit or a fitness value.

$$F(s_w) = \begin{cases} g^{(sw)}_0 & ; \text{if } g(s_w) \geq 0 \\ \text{otherwise.} & \end{cases} \quad (8)$$

#### e) Bat Algorithm (BA)

Bat echolocation is used in the bat algorithm (BA), which is a heuristic optimization tool (Yang, 2010). The four basic parameters of a BA are pulse frequency, pulse rate, velocity, and a constant. The parameters have been set according to the study (Huang and Ma, 2020).

The frequency, velocity, and position for each bat are initialized. The virtual bats' movement is described by updating their velocity and position using the equations below for each time step  $t$ , where  $T$  is the iteration limit.

$$f_i = f_{\min} + (f_{\max} - f_{\min})\beta \quad (9)$$

$$V_i^{t+1} = v_i^t + [X_i^t + X^*]f_i \quad (10)$$

$$X_i^{t+1} = X_i^t + v_i^t \quad (11)$$

Where  $V_i^t$  and  $X_i^t$  are the velocity and position at time  $t$ ,  $V_i^{t+1}$  and  $X_i^{t+1}$  are the velocity and position at time  $t+1$ , and  $\beta$  is a random number between 0 and 1.

A random number is generated when the bat positions are updated; if the random number is greater than the pulse emission rate, a new location is formed around the current best solutions, as shown in the equation below.

$$X_{\text{new}} = x_{\text{old}} + EA^t \quad (12)$$

Where  $E$  is a random number  $A^t$  represents the average loudness of all bats at time  $t$ .

#### f) Ant Colony Optimization (ACO)

The ACO heuristic was inspired by investigations of ant foraging behavior in real colonies, which indicated that ants can often figure out the shortest path between food source and nest (Zheng et

al. 2003). The parameters have been set according to the study (Tang et al. 2012).

When the ant moves from  $i$  to  $j$ , the path node at the start can set as  $A$ ,  $A = \{0, 1, \dots, n-1\}$ . This reflects the role of pheromones accumulated by ants during exercise during ant migration and reveals the relative relevance of the trajectory. The larger  $\alpha$  is, it indicates the high probability for subsequent ants to choose this path.

The probability of the ant moving from  $i$  to  $j$  is computed using the following formula:

$$P_{ij}^k(t) = r_{ij}^k(t) n_{ij}^\beta(t) / \sum_{r_{ij}^\alpha(t)} n_{ij}^\beta(t) \quad (13)$$

Where pheromone is  $\rho$ , which is a constant that represents weight. The time of iteration is  $Nc$  and the initial setting is  $\phi$ . The predicted heuristic factor is  $\beta$ , which demonstrates the relevance of visibility relative to other factors. It also represents the significance of the heuristic component in the entire path of the ant's movement.

#### g) Firefly Algorithm (FA)

Firefly algorithm is a very strong technique for solving restricted optimization and NP-hard problems (Apostolopoulos and Vlachos, 2011). The parameters have been set according to the study (Tang et al. 2012).

The attractiveness of a firefly  $i$  on a firefly  $j$  is determined by the degree of the firefly  $i$ 's brightness and the distance  $r_{ij}$  between the firefly  $i$  and the firefly  $j$ , as shown below:

$$I(r) = I_0 / r^2 \quad (14)$$

Consider the case when there are  $n$  fireflies and the solution for firefly  $i$  is  $x_i$ . The brightness of the firefly  $i$  is linked to the objective function  $f(x_i)$ .

$$I = f(x_i) \quad (15)$$

Each firefly has an attraction value, and the less dazzling (attractive) one is drawn to the brighter one and transferred there. The attractiveness value  $\beta$  is relative based on the distance between fireflies.

$$\beta(r) = \beta_0 e^{-\gamma r^2} \quad (16)$$

Where  $\beta_0$  is the firefly attraction value at  $r = 0$  and  $\gamma$  is the media light absorption coefficient.

#### h) Cuckoo Search (CS) Algorithm

Yang and Deb, 2009, developed the Cuckoo Search algorithm which is based on some cuckoo species' brood parasitism. The parameters have been set according to the study (Fong et al. 2014).

An initial population of  $n$  nests is randomly generated at the positions,  $X = \{x_1^0, x_2^0, \dots, x_n^0\}$ , to evaluate the objective values to find the current global best  $g_t^0$ .

The new position is updated accordingly by performing a Lévy flight:

$$x_i^{(t+1)} = x_i^{(t)} + \alpha \oplus \text{Lévy}(\lambda), \quad (17)$$

where  $\alpha > 0$  denotes the step size, which should be connected to the problem's scales. In most circumstances, we can use  $\alpha = 1$ .

#### i) Coral Reefs Optimization Algorithm (CRO)

CRO is another nature-inspired algorithm, based on an artificial simulation of the process of coral reef formation and reproduction (Sanz et al. 2014). The CRO algorithm has never been utilized in the realm of software bug detection to our knowledge. Corals reproduce at each iteration step in the CRO algorithm, producing new individuals. The parameters have been set according to the study (Medeiros et al., 2015).

By allocating a coral to each square (i j), the CRO algorithm generates a  $N \times M$  square grid in which each square (i,j) may represent an alternate solution to a problem (or colony of corals). The formation of coral is the second phase. After three phases, the entire collection of existing corals in the reef is graded according to their level of healthiness (broadcast spawning, brooding, and larvae setting).

#### j) Particle Swarm Optimization (PSO)

The behavior of particles in a swarm is the central concept of the PSO. Each particle has its own location in a multidimensional space and communicates with the others. To move about in space, the particles employ social and cognitive information. When the algorithm comes to a halt, the best solution has been discovered (Koochi and Groza, 2014). The parameters have been set according to the study (Rana et al., 2010).

The inertia weight balances the algorithm's local and global search abilities. The proportional contribution of the prior velocity to the current velocity is defined by the inertia weight.

$$V_i^{k+1} = wv_i^k + c1 \text{rand} (p_{\text{best}i} - x_i^k) + c2 \text{rand} (g_{\text{best}} - X_i^k) \quad (18)$$

$$X_i^{k+1} = X_i^k + v_i^{k+1} \quad (19)$$

#### k) Grey Wolf Optimizer (GWO)

The Grey Wolf Optimizer (GWO) is a simple, population-based, flexible, and derivative-free metaheuristic optimization method that intelligently avoids stagnation in local optima spots of the search space. It simulates the social behaviors of grey wolves in the aspects of their hierarchical leadership and hunting movement (Mirjalili et al., 2013). Grey wolves' leadership and haunting mechanism help to design a new metaheuristic algorithm with three steps: searching prey, encircling prey, and attacking prey.

During the GWO operation, the position of the wolves is continuously updated, with appropriate mathematical formulas (Hou et al., 2022). The parameters have been set according to the study (Wang et al., 2019).

## IV. PROPOSED METHOD

To address the curbs of the K-means clustering approach in generating globally optimum clusters, the suggested method uses the k-means algorithm together with a range of NIAs for software bug prediction. By adding an exploration function to the k-means algorithm, the combination of these strategies may improve the model. The exploration function improves the existing solution by examining regions outside of its immediate vicinity, and if a new, better solution than the current best one is discovered, the search agents will move toward it. The exploring procedure will continue until certain stopping criteria are met. Nature-inspired algorithms are metaheuristic algorithms, which means they have the ability to explore the combinatorial search space heuristically rather than exhaustively. The integration methods are based on representing the search agents as a combination of centroid locations, then the search agents explore the search space for the best solution.

The purpose of clustering is to discover a proper set of centroids using the metaheuristic of the nature-inspired method as a guide. The metaheuristic will always insist on centroids being moved in a progressive manner in each phase, with the goal of finding the best grouping. The ideal group's ultimate result should be that the data points inside each cluster are closest to their centroid. During the search, the centroids move around in the search space, following the swarming pattern of the nature-inspired optimization method, until no further progress is seen. It comes to a halt when there is no other possible relocation that will yield a better result. Along with the success of employing nature-inspired metaheuristic algorithms to solve automatic clustering problems, it has been discovered that combining two or more metaheuristics for the same objective improves clustering performance. The performance of hybrid algorithms, according to Nanda and Panda 2014, is superior to that of separate algorithms in terms of robustness, effectiveness, and accuracy.

## V. DATASET AND DATA PROCESSING

The dataset was collected from the online PROMISE repository. AR1, AR3, AR4, AR5, AR6, KC1, KC2, JM1, CM1, PC1 and PC5 were used respectively. With reference to the paper, by Shepperd et al. 2013, data cleaning is mandatory before using any datasets available. Indeed, we noted a huge class imbalance issue with the available datasets (faulty, non-faulty), and all data inconsistencies, missing and null values were removed. Each dataset selected represents a NASA software system that includes various metrics. Each dataset is made up of a number of software modules and attributes. Modules with defects are classified as prone to faults, whereas those without defects are

classified as non-fault prone. For the training purpose, the entire dataset is used except for the last column (output column), only columns consisting of numerical values were considered.

Table 1: Summary of dataset

Dataset	Modules	Defective modules	Software metrics (Attributes)
AR1	121	9	29
AR3	63	8	29
AR4	107	20	29
AR5	36	8	29
AR6	101	15	29
KC1	2109	1783	22
KC2	522	107	21
JM1	7782	1672	21
CM1	327	42	37
PC1	705	61	37
PC5	1711	471	38

## VI. EVALUATION

### a) Experimental Setup

The main goal of this research is to demonstrate the utility of the k-means algorithm with different NIAs, which we accomplished using Tensorflow to train the model. TensorFlow is an open-source machine learning platform to build and deploy prediction models. Google Colab was also used to run the results, which allowed the code to run with no configuration and free GPU access. Each dataset is performed 10 times in the trials to find the average CPU time and objective function values/best fitness value.

The clustering results of the new hybrid clustering algorithms are compared to the K-means, which serve as a benchmarking reference. The full dataset is used for training, and cluster formation is referred to until perfection is attained using the entire set of data. The ultimate clustering result's quality is determined by each cluster's integrity, which is represented by the objective function's final fitness value.

The hardware configuration used for all experiments in this study is as follows: Corei7-6500U CPU @2.50 GHz 2.60 GHz, Windows 10, 64-bit operating system, x64 based processor, RAM: 8 GB DDR4, and Hard Disk: SSD.

### b) Performance Evaluation Measures

In order to assess the effectiveness of combining the k-means algorithm and optimization algorithms in the prediction of software bugs, the evaluation metrics, accuracy and F-measure have been calculated accordingly as shown in the Equation (1):

$$Accuracy = (TP+TN) / (TP+TN+FP+FN), \quad (20)$$

Where TP = true positive, TN = true negative, FN = false negative and FP = false positive.

On the other hand, the external metric used to determine the accuracy of the clustering findings, known as the F-measure, is also computed.

The F-measure, which is the average of precision and sensitivity performance, is calculated as follows:

$$F = 2 * P * Sensitivity / P + Sensitivity, \quad (21)$$

Where P refers to precision and sensitivity is calculated by finding the non-defective modules that were accurately categorized.

## VII. RESULTS AND DISCUSSIONS

Table 2: Accuracy of algorithms

Datasets	AR1	AR3	AR4	AR5	AR6	KC1	KC2	JM1	CM1	PC1	PC5
k-Means	88.90	88.00	89.01	88.85	88.43	89.10	89.00	88.80	89.00	89.19	89.99
K-Means +GA	90.50	90.58	91.28	91.55	90.11	90.00	90.54	90.53	91.25	90.00	90.05
K-Means +BAT	90.00	91.59	91.00	92.34	92.00	92.98	91.34	90.00	91.25	92.56	92.00
K-Means +PSO	92.50	92.65	92.87	93.01	93.00	92.99	94.10	92.67	92.89	93.10	93.58
K-Means +Coral Reefs	94.00	94.54	94.56	94.87	94.00	95.96	95.66	96.88	95.01	95.04	95.54
K-Means +Cuckoo	94.50	94.58	94.58	94.00	94.56	95.45	95.88	95.67	95.44	94.56	94.78
K-Means + ACO	94.00	93.56	93.50	94.10	93.78	93.03	93.56	93.44	93.89	94.01	94.52
K-Means +Firefly	92.56	92.67	93.00	93.44	93.02	93.56	94.78	93.67	94.88	94.34	94.54
K-Means + GWO	90.09	92.47	94.65	93.22	92.00	92.60	93.00	92.50	94.50	94.12	94.13

From the table above, K-means clustering is optimized using the various NIAs. We can see that all of the proposed algorithms perform better than the traditional standalone k-means algorithm. K-means appears to take the shortest computation time in any of

the tests, maybe because it stops early in local optima (Table 3). This is evident from the accuracy obtained from the table above. NIAs speed up the process of clustering centroids and illustrate that all partitioning clustering methods can be linked with the natural search

process to prevent local optima. Secondly, simple K-means were applied to the robust nature of GA, which shows adequate prediction accuracy for all datasets. Even though GA may converge to the global optimum due to mutation, GA faces the issue in terms of computational challenges. The application of k means with the Bat algorithm apparently yields the same accuracy. This hybrid algorithm improves the convergence speed of BA and helps the k means algorithm independent of the initial centers. Next, K means is combined with PSO. The PSO method is used to start the process because of its fast convergence, and then the K-Means algorithm is used to refine the PSO algorithm's outcome to near-optimal solutions. The hybridization of these two methods yields effective results in terms of efficiency and precision. The PSO algorithm can be used to generate good initial cluster centroids for the K-Means.

Furthermore, K means and Coral reefs algorithm are combined. The results for this combined method are quite promising since they show that using the CRO method for a clustering application can produce better results to using hybrid genetic algorithms, which is the most often used clustering optimization technique. To best of our knowledge, CRO has not been used with clustering for software bug detection. The hybrid model of k means with Cuckoo Search algorithm shows significant accuracy, likewise CRO algorithm. Cuckoo search is used to provide a robust initialization, whereas K-means is utilized to construct solutions faster. K means is also combined with Ant Colony Optimization algorithm. The suggested method's learning mechanism is based on the use of a defined parameter termed pheromone, which eliminates

undesirable K-means algorithm solutions. The suggested method improves the K-means algorithm by making it less reliant on starting parameters such as randomly picked beginning cluster centers, resulting in a more stable algorithm. K means with firefly also produce near accuracy with CRO and Cuckoo search algorithm. This is because fireflies with high similarity are dispersed, resulting in a more diverse distribution of the entire swarm in search space. K means with GWO has also shown rapid convergence. This improvement is caused by the fact that K-means significantly affects the GWO population and separates it into two clusters. Because GWO often operates as three clusters and has three wolves in the search space, K-means is advantageous for GWO. As a result, it can be concluded that K-means combined with GWO increased GWO's effectiveness to some extent.

High clustering accuracy and efficiency were obtained from the hybrid clustering of Coral reefs and Cuckoo Search Algorithm. Hybrid clustering of Coral reefs algorithm has never been applied in the field of software bug detection and has indeed shown promising results. Hybrid clustering of Coral reefs algorithm locate cluster centroids without causing premature convergence. The findings of the evaluation results add evidence that NIAs can indeed speed up the process and avoid local optima. Because fewer iterations are required to achieve the best cluster outcome, selecting the number of clusters enhances the hybridized clustering method's convergence speed. The computational time for each algorithm is computed as shown in Table 3. Less computational time was noted when K means was integrated with Coral reefs and Cuckoo Search algorithm respectively.

*Table 3:* Computational Time for all Algorithms

Datasets	AR1	AR3	AR4	AR5	AR6	KC1	KC2	JM1	CM1	PC1	PC5
k-Means	79.88	80.01	79.23	80.10	80.15	79.99	79.98	80.65	80.00	79.80	80.34
K-Means +GA	159.99	157.85	162.89	161.00	162.00	172.45	170.55	169.87	172.99	169.00	170.03
K-Means +BAT	162.40	161.00	160.88	165.54	166.87	164.34	157.88	169.90	161.45	162.34	165.10
K-Means +PSO	165.78	155.00	168.98	172.99	170.00	169.99	159.00	159.90	172.78	172.00	169.56
K-Means +Coral Reefs	148.89	152.77	148.54	150.00	149.90	155.42	150.09	151.23	147.77	148.99	149.00
K-Means +Cuckoo	146.67	146.90	149.45	148.00	145.78	148.00	149.99	148.45	150.45	146.88	149.00
K-Means+ ACO	162.67	166.34	159.90	155.67	161.88	160.10	160.00	168.89	159.45	158.45	158.00
K-Means +Firefly	150.23	152.90	150.00	151.90	155.67	150.45	152.56	154.78	152.89	155.90	154.78
K-Means+ GWO	151.45	151.00	156.40	156.34	156.12	154.98	153.10	154.88	154.00	151.17	154.97

For statistical performance, the F1 score has been calculated for all the algorithms as shown in Table 4. Again, the F1 Score shows that K-means with Coral reefs resulted in dependable and significant performance that can be used to predict software

defects. When a good validity measure is applied, most metaheuristic algorithms can automatically divide datasets into an appropriate number of clusters, according to Gbaje et al.2019.

Table 4: Statistical Performance Analysis of Algorithms- F1 Score

Datasets	AR1	AR3	AR4	AR5	AR6	KC1	KC2	JM1	CM1	PC1	PC5
k-Means	0.66	0.79	0.82	0.80	0.75	0.81	0.80	0.81	0.82	0.82	0.80
K-Means+GA	0.84	0.83	0.83	0.80	0.83	0.84	0.84	0.85	0.82	0.81	0.85
K-Means +BAT	0.83	0.81	0.83	0.86	0.86	0.86	0.85	0.85	0.85	0.87	0.85
K-Means +PSO	0.85	0.85	0.87	0.87	0.86	0.85	0.87	0.85	0.87	0.87	0.87
K-Means +Coral Reefs	0.86	0.86	0.86	0.85	0.86	0.86	0.87	0.88	0.86	0.87	0.88
K-Means +Cuckoo	0.89	0.85	0.88	0.89	0.86	0.89	0.86	0.89	0.89	0.87	0.88
K-Means+ ACO	0.84	0.83	0.86	0.85	0.84	0.86	0.85	0.85	0.86	0.85	0.86
K-Means +Firefly	0.86	0.85	0.83	0.87	0.87	0.85	0.83	0.85	0.86	0.88	0.85
K-Means+ GWO	0.82	0.82	0.81	0.86	0.84	0.83	0.79	0.85	0.84	0.84	0.85

## VIII. PRACTICAL IMPLICATIONS

Metaheuristics algorithms have shown to be effective optimizers. This research found that each of the hybrid K means based-nature-inspired optimization algorithm models outperformed the standalone K means algorithm in terms of accuracy and F1 score. Following the intrinsic limitations of K-means design and the virtues of Nature-inspired optimization techniques, it seems feasible to integrate them, allowing them to complement and function together. The algorithms' successful integration gives reason to believe that more advanced optimization mining techniques can be developed. This study can be used as a roadmap for researchers who want to incorporate other new emerging NIAs into improved clustering methods in the field of software bug detection.

## IX. CONCLUSION AND FUTURE WORKS

Prediction of defect-prone software modules is an important goal in software engineering. The traditional clustering algorithm usually gets trapped in the problem of local optima. As a result, the nature-inspired method provides an alternative technique for solving clustering problems using its searching capabilities. This study's main contribution is combining the clustering algorithm with the different NIAs for software bug detection. To the authors' knowledge, only PSO, Cuckoo, Bat, and GWO (Grey Wolf Optimizer) algorithms were applied with clustering algorithms for software bug detection (Almayyan, 2021). The results are improved significantly when clustering algorithms are combined with bio-inspired optimization methods, apparently for the hybrid model of k means clustering with Coral reefs algorithm, achieving an accuracy of 96%. For future work, this work can be replicated with other related datasets for the analysis of bug prediction in software.

## ACKNOWLEDGMENTS

This study received no formal support from public, private, or not-for-profit funding organizations.

## REFERENCES RÉFÉRENCES REFERENCIAS

1. J.B.Macqueen, "Some methods for classification and Analysis of Multivariate Observations". Proceedings of 5<sup>th</sup> Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1967, pp 281-297
2. Rui Tang, Simon Fong, Xin-she Yang and Suash Deb, Integrating Nature inspired Optimization algorithms to k-means clustering, University of Macau, Aug 2012
3. RiskiAnnisa, DidiRosiyadi, Dwiza Riana, Improved point center algorithm for k-means clustering to increase software defect prediction, International Journal of Advances in Intelligent Informatics, Vol 6, No.3, November 2020, pp.328-339
4. S. Lessmann, B. Baesens, C. Mues, and S. Pietsch, "Benchmarking classification models for software defect prediction: A proposed framework and novel findings," IEEE Transactions on Software Engineering, vol. 34, no. 4, pp. 485–496, 2008.
5. I. Myrtveit, E. Stensrud, and M. Shepperd, "Reliability and validity in comparative studies of software prediction models," IEEE Transactions on Software Engineering, vol. 31, no. 5, pp. 380–391, 2005.
6. Jain, A.K. and R.C. Dubes, Algorithms for clustering data. 1988: Prentice-Hall, Inc.
7. Gayathri, R., Cauveri, A., Kanagapriya, R., Nivetha, V., Tamizhselvi, P., & Kumar, K. P. (2015, March). A Novel Approach for Clustering Based On Bayesian Network. In Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015) (p. 60). ACM.
8. Zhong S, Khoshgoftaar TM, Seliya N. Unsupervised learning for expert-based software quality estimation. In: Proceedings of the eighth IEEE international conference on high assurance systems engineering HASE 2004; 2004. p. 149–55. <https://doi.org/10.1109/HASE.2004.1281739>.

9. Bishnu PS, Bhattacharjee V. Software fault prediction using quad tree-based k-means clustering algorithm. *IEEE Trans Knowl Data Eng.* 2012; 24(6):1146–50. <https://doi.org/10.1109/TKDE.2011.163>.
10. Catal C, Sevim U, Diri B. Software fault prediction of unlabeled program modules. In: *Proceedings of the world congress on engineering WCE 2009; 2009.* p. 1–6.
11. YunlongZhu ,Xiaohui Yan, Wenping Zou and Liang Wang, (2012), "A new approach for data clustering using hybrid artificial bee colony algorithm", *Neuro computing*, Vol. 97 , pp.241–250
12. Yi-Tung Kao, ErwieZahara and I-Wei Kao, (2008), "A hybridized approach to data clustering", *Expert Systems with Applications*, Vol.34, No. 3, pp.1754–1762
13. Waheeda Almayyan, Towards Predicting software defects with clustering techniques, *International Journal of Artificial Intelligence and Application (IJAAI)*, Vol 12, No 1, January 2021
14. M. Shepperd, Q.Song, Z.Sun, and C.Mair, –Data quality: *Some Eng.*, vol. 39, no. 9, pp. 1208–1215, 2013.
15. Z. Tóth, P. Gyimesi, and R. Ferenc, "A Public Bug Database of GitHub Projects and Its Application in Bug Prediction," in *Computational Science and Its Applications -- ICCSA 2016*, Cham, 2016, pp. 625–638: Springer International Publishing.
16. Simon Fong, Suash Deb, Xin-She Yang, and Yan Zhuang, Towards Enhancement of Performance of K-Means Clustering Using Nature-Inspired Optimization Algorithms, *Computational Intelligence and Metaheuristic Algorithms with Applications*, 2014, <https://doi.org/10.1155/2014/564829>
17. Deepinder Kaur, Arashdeep Kaur, "Fault Prediction using K-Canberra Means Clustering", *CNC 2010*[in Press]
18. Abiodun M. Ikotun, Mubarak S. Almutari and Absalom E. Ezugwu, K-Means-Based Nature-Inspired Metaheuristic Algorithms for Automatic Data Clustering Problems: Recent Advances and Future Directions, *Appl. Sci.* 2021, 11, 11246. <https://doi.org/10.3390/app112311246>
19. Okwu, M.O.; Tartibu, L.K. *Metaheuristic Optimization: Nature-Inspired Algorithms Swarm and Computational Intelligence, Theory and Applications*; Springer Nature: Berlin/Heidelberg, Germany, 2020; Volume 927.
20. Hruschka, E.; Campello, R.J.G.B.; Freitas, A.A.; de Carvalho, A. A Survey of Evolutionary Algorithms for Clustering. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* 2009, 39, 133–155.
21. Nanda, S.J.; Panda, G. A survey on nature inspired metaheuristic algorithms for partitional clustering. *Swarm Evol. Comput.* 2014, 16, 1–18. [CrossRef]
22. Zhou, X.; Gu, J.; Shen, S.; Ma, H.; Miao, F.; Zhang, H.; Gong, H. An Automatic K-Means Clustering Algorithm of GPS Data Combining a Novel Niche Genetic Algorithm with Noise and Density. *ISPRS Int. J. Geo-Inf.* 2017, 6, 392
23. Mousa, A.A., El-Shorbagy, M.A. and Abd El-Wahed, W.F. (2012) Local Search Based Hybrid Particle Swarm Optimization for Multiobjective Optimization. *Swarm and Evolutionary Computation*, 3, 1-14.
24. N. Bouhmal, A. Viken, and J. B. Lønnum, Enhanced Genetic Algorithm with K-Means for the Clustering Problem, *International Journal of Modeling and Optimization*, Vol. 5, No. 2, April 2015
25. D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, New York, 1989 X.-S. Yang, "A new metaheuristic bat-inspired algorithm," *Nature Inspired Cooperative Strategies for Optimization*, vol. 284, pp. 65–74, 2010
26. Jianqiang Huang and Yan Ma, Bat Algorithm Based on an Integration Strategy and Gaussian Distribution, Volume 2020 | Article ID 9495281 | <https://doi.org/10.1155/2020/9495281>
27. Zheng, H., Zheng, Z., Xiang, Y., The application of ant colony system to image texture classification [textute read texture], In: *Proceedings of the 2nd International Conference on Machine Learning and Cybernetics*, Vol. 3, Xi'an, China, (2003) 1491-1495
28. R. Tang, S. Fong, X. Yang and S. Deb, "Integrating nature-inspired optimization algorithms to K-means clustering," *Seventh International Conference on Digital Information Management (ICDIM 2012)*, 2012, pp. 116-123, doi: 10.1109/ICDIM.2012.6360145.
29. Apostolopoulos, T. and Vlachos, A. (2011). Application of the Firefly Algorithm for Solving the Economic Emissions Load Dispatch Problem. *International journal of Combinatorics*. doi:10.1155/2011/523806
30. X.-S. Yang and S. Deb, "Cuckoo search via Lévy flights," in *Proceedings of the World Congress on Nature & Biologically Inspired Computing (NABIC '09)*, pp. 210–214, Coimbatore, india, December 2009.
31. Inacio G. Medeiros and Joao C. Xavier- Junior, Anne M. P. Canuto, Applying the Coral Reefs Optimization Algorithm to Clustering Problems, *Conference Paper · July 2015 DOI: 10.1109/IJCNN.2015.7280845*
32. Salcedo-Sanz, S. and Del Ser, J. and Gil-Lpez, S. and Landa-Torres, I. and Portilla-Figueras, J. A., "The Coral Reefs Optimization Algorithm: A Novel Metaheuristic for Efficiently Solving Optimization Problems," *The Scientific World Journal*, Volume 2014, Hindawi Publishing Corporation, 2014.
33. I. Koochi and V. Z. Groza, "Optimizing Particle Swarm Optimization algorithm," *2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering*

- (CCECE), 2014, pp. 1-5, doi: 10.1109/CCECE.2014.6901057.
34. R.Jensi and G.WiselinJiji, HYBRID DATA CLUSTERING APPROACH USING K-MEANS AND FLOWER POLLINATION ALGORITHM, *Advanced Computational Intelligence: An International Journal (ACIJ)*, Vol.2, No.2, April 2015
  35. Agbaje, M.B.; Ezugwu, A.E.; Els, R. Automatic Data Clustering Using Hybrid Firefly Particle Swarm Optimization Algorithm.IEEE Access 2019, 7, 184963–184984.
  36. Yang, X. S., and et al." Swarm Intelligence and Bio-Inspired Computation: Theory and Applications, Elsevier Science Publishers B. V. Amsterdam, The Netherlands,(2013)
  37. S. Mirjalili, S. M. Mirjalili and A. Lewis, Grey wolf optimizer, *Adv. Eng. Softw.* 69 (2014), 46–61.10.1016/j.advengsoft.2013.12.007
  38. Hou, Y.; Gao, H.; Wang, Z.; Du, C. Improved Grey Wolf Optimization Algorithm and Application. *Sensors* 2022, 22, 3810. <https://doi.org/10.3390/s22103810>
  39. Jie-Sheng Wang, Shu-Xia Li, An Improved Grey Wolf Optimizer Based on Differential Evolution and Elimination Mechanism, *SciRep* 9, 7181 (2019). <https://doi.org/10.1038/s41598-019-43546-3>
  40. Sandeep Rana, Sanjay Jasola , Rajesh Kumar, A hybrid sequential approach for data clustering using K-Means and particle swarm optimization algorithm, *International Journal of Engineering, Science and Technology* Vol. 2, No. 6, 2010, pp. 167-176



# GLOBAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY: C SOFTWARE & DATA ENGINEERING

Volume 26 Issue 1 Version 1.0 Year 2026

Type: Double Blind Peer Reviewed International Research Journal

Publisher: Global Journals

Online ISSN: 0975-4172 & Print ISSN: 0975-4350

## Comparison of Prim and Kruskal's Algorithm

By Rohit Maurya & Rahul Sharma

*Ajeenkya D Y Patil University*

**Abstract-** The goal of this research is to compare the performance of the common Prim and the Kruskal of the minimum spanning tree in building up super metric space. We suggested using complexity analysis and experimental methods to evaluate these two methods. After analysing daily sample data from the Shanghai and Shenzhen 300 indexes from the second half of 2005 to the second half of 2007, the results revealed that when the number of shares is less than 100, the Kruskal algorithm is relatively superior to the Prim algorithm in terms of space complexity; however, when the number of shares is greater than 100, the Prim algorithm is more superior in terms of time complexity. A spanning tree is defined in the glossary as a connected graph with non-negative weights on its edges, and the challenge is to identify a minimum weight spanning tree. Surprisingly, the greedy algorithm yields an answer. For the problem of finding a minimum weight spanning tree, we propose greedy algorithms based on Prim and Kruskal, respectively. Graham and Hell provide a history of the issue, which began with Czekanowski's work in 1909. The information presented here is based on Rosen.

**Keywords:** *kruskal, prim's, graph, minimal spanning trees, complexity.*

**GJCST-C Classification:** *FOR Code: 080201*



*Strictly as per the compliance and regulations of:*



© 2026 Rohit Maurya & Rahul Sharma. This research/review article is distributed under the terms of the Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0). You must give appropriate credit to authors and reference this article if parts of the article are reproduced in any manner. Applicable licensing terms are at <https://creativecommons.org/licenses/by-nc-nd/4.0/>.

# Comparison of Prim and Kruskal's Algorithm

Rohit Maurya<sup>α</sup> & Rahul Sharma<sup>ο</sup>

**Abstract-** The goal of this research is to compare the performance of the common Prim and the Kruskal of the minimum spanning tree in building up super metric space. We suggested using complexity analysis and experimental methods to evaluate these two methods. After analysing daily sample data from the Shanghai and Shenzhen 300 indexes from the second half of 2005 to the second half of 2007, the results revealed that when the number of shares is less than 100, the Kruskal algorithm is relatively superior to the Prim algorithm in terms of space complexity; however, when the number of shares is greater than 100, the Prim algorithm is more superior in terms of time complexity. A spanning tree is defined in the glossary as a connected graph with non-negative weights on its edges, and the challenge is to identify a min weight spanning tree. Surprisingly, the greedy algorithm yields an answer. For the problem of finding a min weight spanning tree, we propose greedy algorithms based on Prim and Kruskal, respectively. Graham and Hell provide a history of the issue, which began with Czeckanowski's work in 1909. The information presented here is based on Rosen.

**Keywords:** *kruskal, prim's, graph, minimal spanning trees, complexity.*

## I. INTRODUCTION

A minimum spanning tree of an undirected graph can be readily obtained using Prim or Kruskal's classical algorithms. To enumerate all spanning trees in an undirected graph, a number of algorithms have been suggested. These algorithms' main worries are good time and space complexities. A minimum spanning tree of an undirected graph can be readily obtained using Prim or Kruskal's classical algorithms. A number of algorithms have been suggested to count all of an object's spanning trees.

A spanning tree of a connected graph can be constructed including all the vertices with minimum possible no of edges. If there are  $n$  vertices in the graph, then each spanning tree has  $n-1$  edges. A connected weighted graph where all the vertices are interlinked by some weighted edges can contain multiple numbers of spanning trees. A minimum spanning tree of an undirected graph can be easily obtained using Prim or Kruskal's classical algorithms. A number of algorithms for enumerating all spanning trees in an undirected graph have been proposed. These algorithms' main concerns are good time and space complexities. The majority of algorithms generate spanning trees by utilising some fundamental cut or circuit. The cost of the tree is not considered during the generation process.

This paper presents an algorithm for generating spanning trees of a graph in decreasing cost order. New opportunities emerge by generating spanning trees in increasing cost order. This method can be used to find the second smallest or, more broadly, the  $k$ -th smallest spanning tree. The smallest spanning tree satisfying some additional constraints can be found by checking whether these constraints are satisfied at each generation. Our algorithm is based on Murty's (1967) algorithm, which enumerates all solutions to an assignment problem in increasing cost order. The complexities of time and space are discussed.

The network is undirected. These algorithms' main worries are good time and space complexities. The goal of this research is to compare the performance of the common Prim and the Kruskal of the minimum spanning tree in constructing super metric space. To evaluate these two methods, we suggested using complexity analysis and experimental methods. After analysing daily sample data from the Shanghai and Shenzhen 300 indexes from the second half of 2005 to the second half of 2007, the results revealed that when the number of shares is less than 100, Kruskal algorithm is relatively superior to Prim algorithm in terms of space complexity; however, when the number of shares is greater than 100, Prim algorithm is more superior in terms of time complexity.

## II. PRIM'S ALGORITHMS

Vojtech Jarnak, a Czech mathematician, created the Prim algorithm in 1930. Robert C. Prim rediscovered it in 1957, and Edsger W. Dijkstra republished it in 1959. To determine the minimal spanning tree (MST) of a given linked weighted graph, Prim's algorithm uses a greedy approach. When the graph is dense, this algorithm is recommended. When there are many edges in a graph, the graph is said to be dense. Only undirected linked graphs can use this approach, and there must not be any edges with a negative edge weight. The algorithm is pretty effective in this situation. There will always be a shortest path because there are no cycles with nonnegative weights.

It begins by choosing a random vertex to serve as the tree's root. Then the shortest edge from any vertex in the tree to the new vertex is added in order to extend the tree, as is the edge closest to the current vertex. Once all vertices have been added to the tree, the procedure ends.

Author <sup>α</sup>: Ajeenkya D Y Patil University, Pune, Maharashtra, India.  
e-mail: rohit221107@gmail.com

### III. THE FOLLOWING ARE THE STEPS TO FIND THE MINIMUM SPANING TREE USING PRIM'S ALGORITHMS

1. If a graph has loops and parallel edges, remove those loops and parallel edges.
2. Select any node at random, labelling it with a distance of 0 and all other nodes as. The chosen node is considered the current node and has been visited. All other nodes are assumed to be unvisited.
3. Locate all unvisited nodes that are currently linked to the current node. Determine the distance between the unvisited nodes and the current node.
4. Label each vertice with its corresponding weight to the current node, but relabel a node if it is less than the previous label value. The nodes are labelled with their weights each time; keep track of the path with the smallest weight.
5. Color over the current node to indicate that it has been visited. We don't need to look at a vertex again once we've visited it.
6. Among all unvisited nodes, find the one with the lowest weight to the current node; mark this node as visited and treat it as the current working node.
7. Repeat steps 3-5 until all nodes have been visited.

Prim's Algorithms:

```

PRIM (Graph, c, t)
P ← N[Graph]
For every m ∈ P
do key[m] ← ∞
key[t] ← 0
[t] ← null
while P ≠ ∅
do m ← EXTRACTMINNODE(P)
for every n ∈ Adjacent[m]
do if n ∈ P and c(m,n) < key[n]

[n] ← m
key[n] ← c(m,n)
    
```

### IV. KRUSKAL'S ALGORITHMS

This algorithm, designed by Joseph Kruskal, was published for the first time in the Proceedings of the American Mathematical Society in 1956. The algorithm begins by creating an ordered set of edges by weights and proceeds through the ordered set by adding an edge to the partial MST if the new edge does not form a cycle. The algorithm takes a greedy approach, in which it finds the path with the least weight in each iteration stage and includes it in the growing spanning tree.

Sort all edges of the given graph in increasing order using Kruskal's algorithm. If the newly added edge does not form a cycle, it continues to add new edges and nodes to the MST. It selects the minimum weighted edge first, followed by the maximum weighted edge.

Thus, in order to find the optimal solution, it makes a locally optimal choice in each step. As a result, this is a Greedy Algorithm.

### V. THE STEPS FOR DETERMINING MST USING KRUSKAL ALGORITHMS ARE AS FOLLOW

Sort the edges in non-descending order of weight.

Choose the smallest edge. Check to see if it forms a cycle with the spanning tree that has been formed thus far. Include this edge if the cycle is not formed. Otherwise, throw it away.

Step 2 should be repeated until the spanning tree has (V-1) edges.

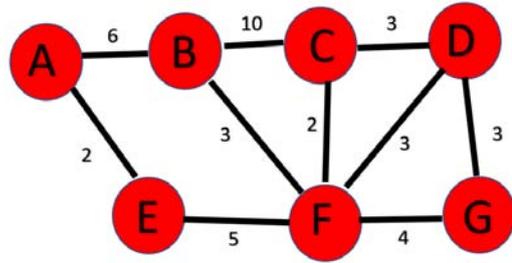
Kruskal Algorithms

```

Kruskal(Graph):
T = Empty;
For every node n ∈ G.N:
CreateSet(n)
For every path (m, n) ∈ G.E arranged by
increasing
weights(m,n):
if NewSet(m) ∩ NewSet(n) = ∅
T = T ∪ {(m, n)}
UNION (m, n)
return T
    
```

Problem Statement

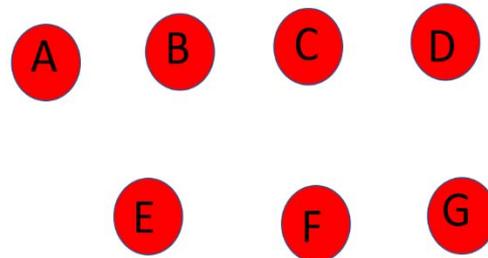
Find the MST through Prim's Algorithm and Explain it step wise.



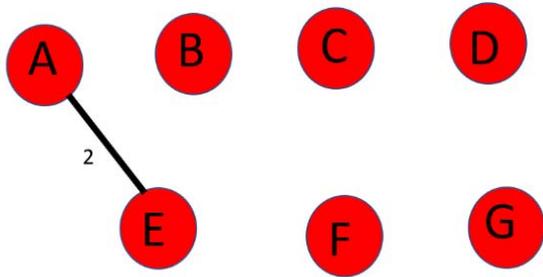
Solution:

Step 1:

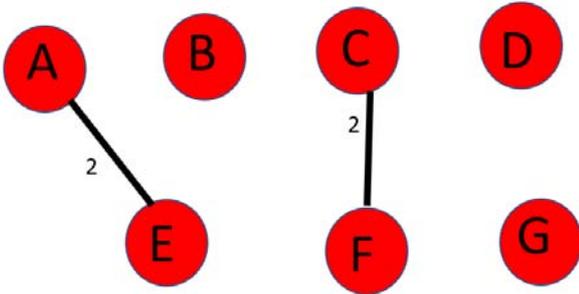
First write all edges weight in Ascending order:  
2,2,3,3,3,3,4,5,6,10



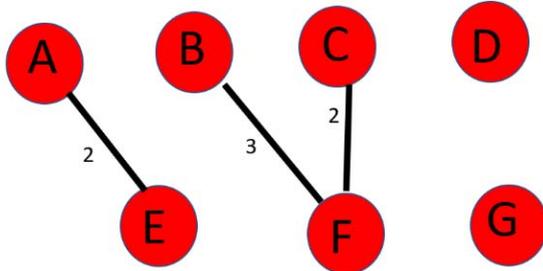
Step 2:



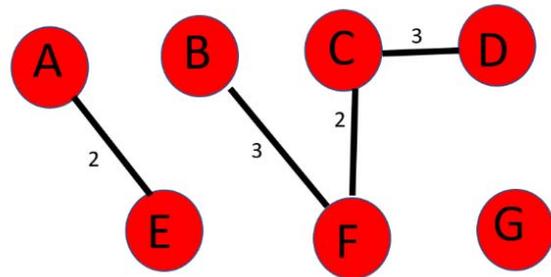
Step 3:



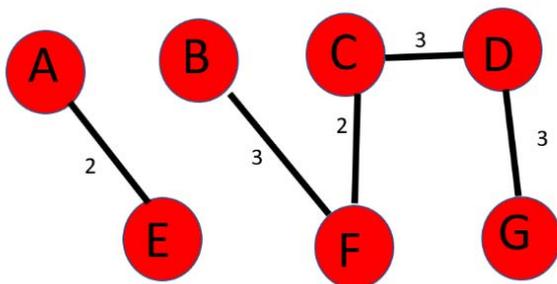
Step 4:



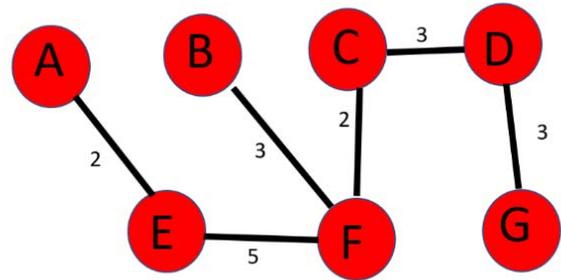
Step 5:



Step 6:

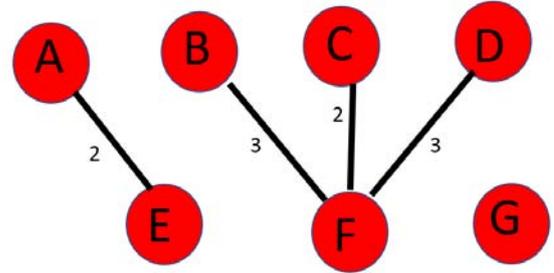


Step 7:

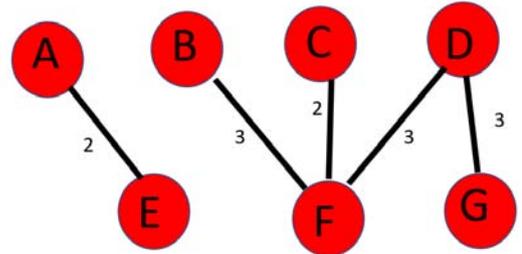


Total weight of MST is 18 and one more path or another MST

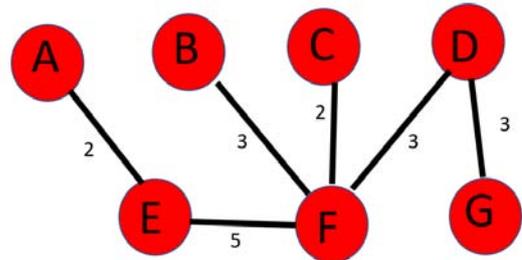
Step 8:



Step 9:



Step 10:



It also have weigh is 18 so both way are right.

### VI. FIND THE TIME COMPLEXITY OF PRIM'S ALGORITHM WE FOLLOW STEP BY STEP

1. Set up a minimum priority queue (heap) and add the first vertex to it.
2. Do the following while the queue is still not empty:
  - a. Take the queue's minimum weight edge.
  - b. If the edge connects a vertex that has not yet been visited, add it to the minimum spanning tree and mark the vertex as visited.

c. Add all adjacent edges of the newly visited vertex to the queue if they connect to unvisited vertices.

Assume the input graph contains  $V$  vertices and  $E$  edges. Prim's algorithm's time complexity can be calculated as follows:

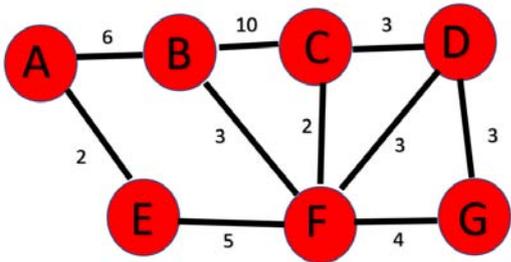
- It takes  $O(V)$  time to initialize the priority queue.
- It takes  $O(\log V)$  time to extract the minimum weight edge from the queue.
- It takes constant time to check if a vertex has been visited.
- Adding an edge to the minimum spanning tree takes an infinite amount of time.
- It takes constant time to mark a vertex as visited.
- Because each edge can only be added to the queue once, adding adjacent edges to the queue takes  $O(E \log V)$  time.

Prim's algorithm has a total time complexity of  $O((V-1) \log V + E \log V)$  because step 2 is repeated  $V-1$  times. In practise, the term  $E \log V$  dominates the time complexity, so Prim's algorithm has an overall time complexity of  $O(E \log V)$ .

It should be noted that this analysis assumes that the priority queue is implemented using a binary heap. The time complexity may differ slightly if a Fibonacci heap or another data structure is used.

**Problem Statement**

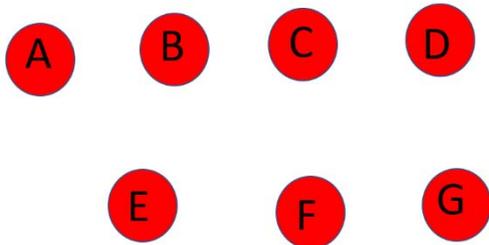
Find the MST through Kruskal's Algorithm and Explain it step wise.



Solution:

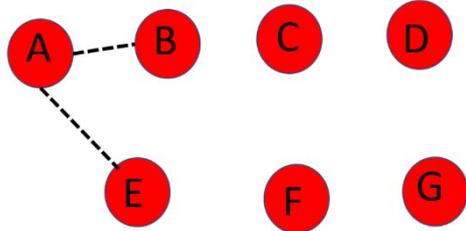
Step 1:

In the Kruskal algorithm first we assign first a node as head or starting point to start the finding MST and show all possible way from that node

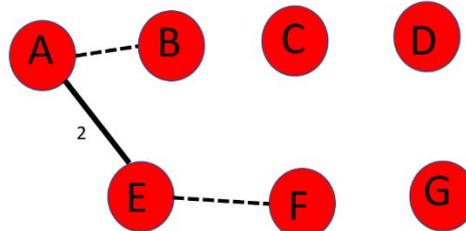


Here I take as a head or starting point node is A

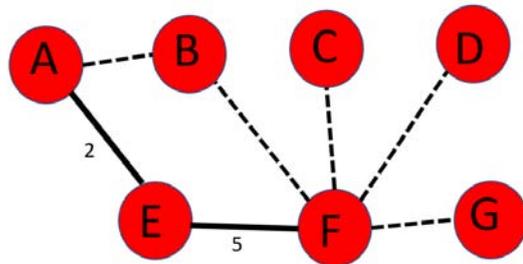
Step 2:



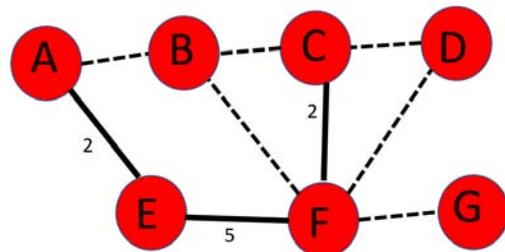
Step 3:



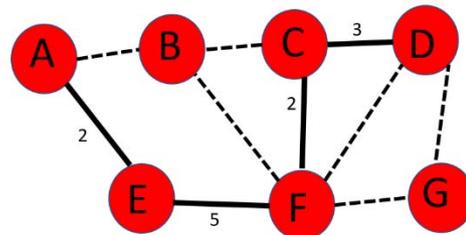
Step 4:



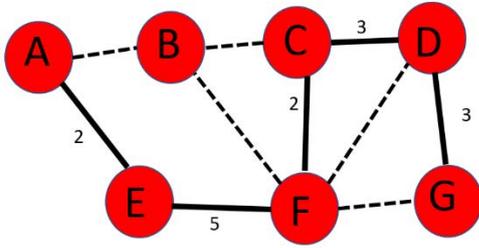
Step 5:



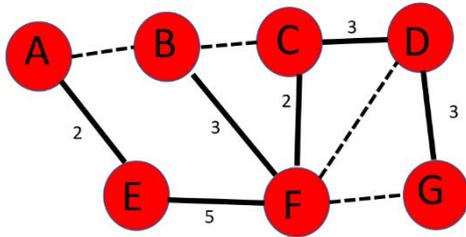
Step 6:



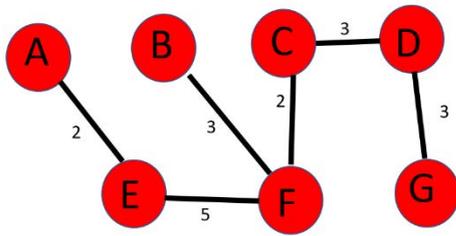
Step 7:



Step 8:



We get MST through it which weight is 18  
Final view is this



### VII. FIND THE TIME COMPLEXITY OF KRUSKAL ALGORITHM WE FOLLOW STEP BY STEP

1. Sort the edges in increasing weight order.
2. To keep track of connected components, create a disjoint-set data structure.
3. Do the following for each edge, in increasing order of weight:
  - a. Add the edge to the minimum spanning tree and merge the two components if it connects two vertices that are not in the same connected component.
  - b. Otherwise, throw away the edge.

Assume the input graph contains  $V$  vertices and  $E$  edges. Kruskal's algorithm's time complexity can be calculated as follows:

- It takes  $O(E \log E)$  time to sort the edges.
- It takes  $O(V)$  time to initialise the disjoint-set data structure.
- Using the disjoint-set data structure, determining whether two vertices are in the same connected component takes  $O(\log V)$  time.
- Using the disjoint-set data structure, merging two connected components takes  $O(\log V)$  time.

- Adding an edge to the minimum spanning tree takes an infinite amount of time.
- It takes time to discard an edge.

Kruskal's algorithm has a total time complexity of  $O(E \log E + V \log V)$  because step 3 is repeated  $E$  times. In practise, the term  $E \log E$  dominates the time complexity, so Kruskal's algorithm has an overall time complexity of  $O(E \log E)$ .

It should be noted that this analysis assumes that the edges are sorted using a standard sorting algorithm such as quicksort or mergesort. The time complexity may differ slightly if a radix sort or another algorithm is used. Furthermore, the time complexity of the disjoint-set data structure is determined by the implementation used.

### VIII. TO FIND THE SPACE COMPLEXITY OF PRIM'S AND KRUSKAL'S ALGORITHM

*The Prim Algorithm:*

- To store edges, the minimum priority queue (heap) requires  $O(E)$  space.
- The boolean array used to mark visited vertices takes up  $O(V)$  space.
- The minimum spanning tree takes up  $O(V)$  space.
- Prim's algorithm has a total space complexity of  $O(V + E)$ .

*The Kruskal Algorithm:*

- The disjoint-set data structure used to keep connected components together takes up  $O(V)$  space.
- The array used to store the edges takes up  $O(E)$  space.
- The minimum spanning tree takes up  $O(V)$  space.
- Kruskal's algorithm has an overall space complexity of  $O(V + E)$ .

It should be noted that the above space complexity calculations assume a standard implementation of each algorithm. However, depending on the implementation, the space complexity can vary. Additionally, the input graph may require space, but this is typically considered a separate factor and is not included in the algorithm's space complexity calculation.

### IX. COMPARISON BETWEEN PRIM'S AND KRUSKAL'S ALGORITHM

1. Prim's algorithm is a greedy algorithm that starts with a single vertex and gradually adds edges to form a minimum spanning tree. Kruskal's algorithm is also a greedy algorithm that begins with the edge with the smallest weight and gradually adds edges to form a minimum spanning tree.
2. Time complexity: The time complexity of Prim's algorithm is  $O(E \log V)$ , where  $E$  is the number of edges and  $V$  is the number of vertices in the graph.

The time complexity of Kruskal's algorithm is  $O(E \log E)$  or  $O(E \log V)$ , depending on the implementation.

3. Prim's algorithm has a space complexity of  $O(V)$ , where  $V$  is the number of vertices in the graph. Kruskal's algorithm has a space complexity of  $O(E)$ , where  $E$  is the number of edges in the graph.
4. Prim's algorithm always generates a connected minimum spanning tree. If the graph is not connected, Kruskal's algorithm can generate a forest of minimum spanning trees.
5. Edge selection: Prim's algorithm chooses the edges with the lowest weight among all those that connect the tree to a non-tree vertex. Kruskal's algorithm chooses edges based on the lowest weight among all edges that have yet to be added to the tree.
6. Performance: Prim's algorithm performs better on dense graphs (where  $E$  is close to  $V^2$ ), while Kruskal's algorithm performs better on sparse graphs (where  $E$  is much less than  $V^2$ ).

Overall, both the Prim and Kruskal algorithms are effective and widely used for determining minimum spanning trees. The algorithm chosen is determined by the characteristics of the input graph and the specific requirements of the problem.

*Time Complication:*

- The time complexity of Prim's algorithm is  $O(E \log V)$ , where  $E$  is the number of edges and  $V$  is the number of vertices in the graph. In dense graphs where  $E$  is close to  $V^2$ , this complexity outperforms Kruskal's algorithm.
- In sparse graphs, where  $E$  is much less than  $V^2$ , Kruskal's algorithm has a time complexity of  $O(E \log E)$ , which is better than Prim's algorithm.

*Space Complexity:*

- Prim's algorithm takes up  $O(V + E)$  space, and Kruskal's algorithm takes up  $O(V + E)$  space as well. As a result, the space complexity of both algorithms is comparable.

There are several alternative algorithms for determining the minimum spanning tree, each with its own time and space complexities:

- Boruvka's algorithm has a time complexity of  $O(E \log V)$ , making it faster in dense graphs than Prim's algorithm. However, it requires  $O(E \log V)$  space, which is greater than that required by Prim's algorithm.
- Although the Reverse-Delete algorithm has an  $O(E^2)$  time complexity, it performs well on sparse graphs. It has an  $O(V + E)$  space complexity.
- The time complexity of Randomized Prim's algorithm is similar to that of Prim's algorithm, but it can be faster in practise due to the randomised nature of its implementation.

Overall, the algorithm chosen is determined by the properties of the graph being processed. Prim's or Boruvka's algorithms may be preferable for dense graphs. Kruskal's algorithm or the Reverse-Delete algorithm may be preferable for sparse graphs. Randomized Prim's algorithm is another viable option in practise.

## X. CONCLUSION

Finally, Prim's and Kruskal's algorithms are two well-known algorithms for determining the minimum spanning tree of a weighted, connected graph.

Prim's algorithm employs a greedy approach, beginning with a single vertex and expanding the minimum spanning tree one edge at a time. At each step, the algorithm maintains a priority queue to select the edge with the lowest weight. Prim's algorithm has a time complexity of  $O(E \log V)$ , where  $E$  is the number of edges and  $V$  is the number of vertices in the graph.

Kruskal's algorithm, on the other hand, employs a greedy approach but works by adding edges to the minimum spanning tree in increasing weight order while avoiding cycles. To keep connected components and check for cycles, the algorithm employs a disjoint-set data structure. Kruskal's algorithm has a time complexity of  $O(E \log E)$ , where  $E$  is the number of edges in the graph.

Both algorithms require  $O(V + E)$  space in terms of complexity.

Overall, the algorithm chosen is determined by the properties of the graph being processed. Prim's algorithm is preferable for dense graphs, whereas Kruskal's algorithm is preferable for sparse graphs. Other algorithms, such as Boruvka's algorithm and the Reverse-Delete algorithm, can also be used depending on the problem's specific requirements.

## REFERENCES RÉFÉRENCES REFERENCIAS

1. B.Hughes. Trees and ultra metric spaces: a categorical equivalence [J]. *Advances in Mathematics*, 2004, 189(1):148-191
2. M. J. Naylor, L.C.Rose, B. J.Moyle. Topology of foreign exchange markets using hierarchical structure methods [J]. *Physica A: Statistical Mechanics and its Applications*, 2007, 382(1):199-208.
3. J.G. Brida, W. A. Risso. Multidimensional minimal spanning tree: The Dow Jones case [J]. *Physica A: Statistical Mechanics and its Applications*, 2008, 387(21):5205-5210.
4. Martel. The expected complexity of Prim's minimum spanning tree algorithm [J]. *Information Processing Letters*, 2002, 81(4):197-201.
5. Yang Guo Hui, Zhou Chun Guang. An algorithm for clustering gene expression data using minimum

- spanning tree [J]. Journal of Computer Research and Development, 2003, 40(10):1431-1435.
6. Feixue Huang. Comparison of Prim and Kruskal on Shanghai and Shenzhen 300 Index hierarchical structure tree, 2009,237-240
  7. Michael Laszlo and Sumitra Mukherjee, Member, IEE . Minimum Spanning Tree Partitioning Algorithm for Micro aggregation, July 2005 ,902-904
  8. Peace Ayegba. A Comparative Study of Minimal Spanning Tree Algorithms, 2020
  9. Jogamohan Medak. Review and Analysis of Minimum Spanning Tree Using Prim's Algorithm, 2018
  10. Kenneth Sorensen. An Algorithm to Generate all Spanning Trees of a Graph in Order of Increasing Cost, 2005
  11. Arogundade O.T. Prim Algorithm to Improving Local Access Network in Rural Areas, 2011
  12. Harvey J. Greenberg. Greedy Algorithm for Minimum Spanning Tree, 1998



# GLOBAL JOURNALS GUIDELINES HANDBOOK 2026

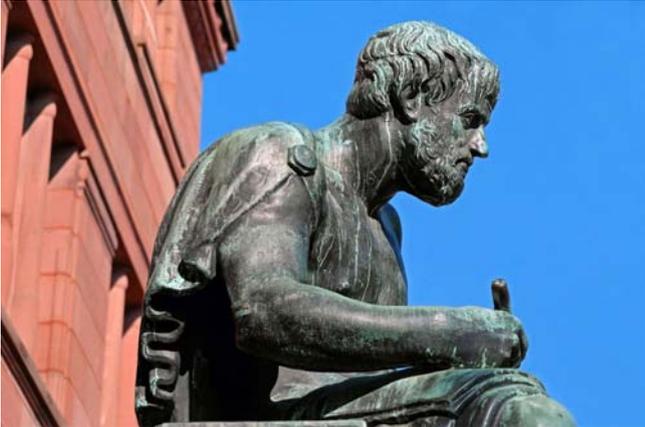
---

[WWW.GLOBALJOURNALS.ORG](http://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.



## BENEFITS

### 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

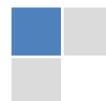
Financial

## 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.



## BENEFITS

### 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
<p>\$4800 lifetime designation</p> <hr/> <p>Certificate, LoR and Momento 2 discounted publishing/year Gradation of Research 10 research contacts/day 1 GB Cloud Storage GJ Community Access</p>	<p>\$6800 lifetime designation</p> <hr/> <p>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</p>	<p>\$12500.00 organizational</p> <hr/> <p>Certificates, LoRs and Momentos Unlimited free publishing/year Gradation of Research Unlimited research contacts/day Unlimited Cloud Storage Online Presense Assistance GJ Community Access</p>	<p>APC per article</p> <hr/> <p>GJ Community Access</p>



# 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](mailto:submit@globaljournals.org) or get in touch with [chiefeditor@globaljournals.org](mailto: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.

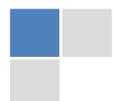
**15. Never start at the last minute:** Always allow enough time for research work. Leaving everything to the last minute will degrade your paper and spoil your work.

**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 through 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.*

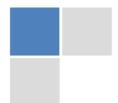
- Fundamental goal.
- To-the-point depiction of the research.
- Consequences, including definite statistics—if the consequences are quantitative in nature, account for this; results of any numerical analysis should be reported. Significant conclusions or questions that emerge from the research.

### **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:**

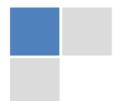
If you put figures and tables at the end of some details, make certain that they are visibly distinguished from any attached appendix materials, such as raw facts. Whatever the position, each table must be titled, numbered one after the other, and include a heading. All figures and tables must be divided 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
<i>Abstract</i>	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
<i>Introduction</i>	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
<i>Methods and Procedures</i>	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
<i>Result</i>	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
<i>Discussion</i>	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
<i>References</i>	Complete and correct format, well organized	Beside the point, Incomplete	Wrong format and structuring



# INDEX

---

---

## **A**

Accordance · 6

---

## **B**

Biases · 12

---

## **C**

Correlate · 1

---

## **D**

Depicted · 11  
Distinctive · 1, 5

---

## **E**

Explicit · 1

---

## **J**

Jupyter · 2, 3, 6, 8

---

## **K**

Kaggle · 1, 6

---

## **O**

Opaque · 3

---

## **R**

Reproducible · 1, 3  
Requisite · 1

---

## **S**

Stacking · 12, XIV



save our planet



# Global Journal of Computer Science and Technology

---

Visit us on the Web at [www.GlobalJournals.org](http://www.GlobalJournals.org) | [www.ComputerResearch.org](http://www.ComputerResearch.org)  
or email us at [helpdesk@globaljournals.org](mailto:helpdesk@globaljournals.org)



ISSN 9754350