International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.7 May - June – 2023
DOI:10.46647/ijetms.2023.v07i03.012 ISSN: 2581-4621

PLAGIARISM DETECTION FOR PROJECT REPORT USING MACHINE LEARNING

Niraj Mohabey[1], Yash Gavanang[2] , Abubakkar Khan[3] , Lavesh Singh Chib[4], Bhushan Patil[5]
[1] UG - Computer Engineering, Rajiv Gandhi Institute of Technology, Mumbai, Maharashtra
[2] UG - Computer Engineering, Rajiv Gandhi Institute of Technology, Mumbai, Maharashtra
[3] UG - Computer Engineering, Rajiv Gandhi Institute of Technology, Mumbai, Maharashtra
[4] UG - Computer Engineering, Rajiv Gandhi Institute of Technology, Mumbai, Maharashtra
[5]Assistant Professor, Computer Engineering, Rajiv Gandhi Institute of Technology, Mumbai, Maharashtra
Corresponding Author Orcid ID: https://orcid.org/0009-0004-3314-7452

ABSTRACT
Plagiarism is an unethical act of using someone else's work or ideas without giving them credit, which is a growing problem in various fields. However, the current systems for plagiarism detection require revealing the full content of input documents and document collections, which can raise procedural and legal concerns regarding data confidentiality, limiting or prohibiting the use of plagiarism detection services. To address these issues, we aim to create a plagiarism detection approach that doesn't need a centralized provider or expose any content as cleartext. Our research has produced initial results showing that our content-protecting method achieves the same detection effectiveness as the original method while making it practically impossible to reveal the protected content through common attacks.
Various techniques, such as manual detection, text similarity analysis, and automated plagiarism detection using machine learning, have been developed to prevent plagiarism. This paper focuses on machine learning techniques for plagiarism detection and discusses different approaches, algorithms, and datasets used in detecting plagiarism, along with their advantages and limitations. The paper also presents some future research directions in this area.
Keywords— Plagiarism, Confidentiality

1. INTRODUCTION
Plagiarism has become a major issue in academic and other fields, as it can harm the author's reputation and the credibility of their research work. Plagiarism is the act of using someone else's work, ideas, or words without proper credit, and it can occur intentionally or unintentionally through various forms, such as copying and pasting, paraphrasing, or using synonyms. Plagiarism detection systems (PDS) typically require users to submit input documents, which the systems compare to a large proprietary database of documents to retrieve similar content and highlight it for user inspection.
There are two types of Plagiarism:
a. Unintentional Plagiarism
- Paraphrasing poorly: changing a few words without changing the sentence structure of the original, or changing the sentence structure but not the words.
- Quoting poorly: putting quotation marks around part of a quotation but not around all of it, or putting quotation marks around a passage that is partly paraphrased and partly quoted.
- Citing poorly: omitting an occasional citation or citing inaccurately.

b. Intentional Plagiarism
- Presenting pre-existing papers found on the Internet or elsewhere as one's own work.
- Reproducing an essay or article from the Internet, an online resource, or an electronic database without proper citation or acknowledgment.
- Creating a paper by merging material from various sources without attribution or citation.
- Taking language or concepts from other sources or classmates without properly acknowledging the origin of the information.

DOI:10.46647/ijetms.2023.v07i03.012 ISSN: 2581-4621

Plagiarism detection is a challenging task, and several techniques have been developed to detect it. These include manual detection, text similarity analysis, and automated plagiarism detection using machine learning. In this paper, we focus on automated plagiarism detection using machine learning techniques.

2. LITERATURE REVIEW

2.1 Paper 1 - Plagiarism Detection in Programming Assignments using Machine Learning Nishesh Awale, Mitesh Pandey, Anish Dulal Department of Electronics and Computer Engineering, Pulchowk Campus, Lalitpur, Nepal.

These days, there has been a rise in plagiarism in programming assignments, which has a negative impact on how students are evaluated. This article suggests using a machine learning technique to detect plagiarism in programming assignments.

- **Methodology**

Perform in the hopes of writing report in order to eliminate the copied report and highlighting the critical aspect of writing assignment on their own.

- **Findings**

Various characteristics associated with a programming assignment pair were calculated, and the Xg boost model was employed to classify them. The accuracy score achieved was 92%.

2.2 Paper 2 - Plagiarism Detector Using Machine Learning Algorithms

The easy accessibility of vast information resources has led to an increase in plagiarism in free text. To address this issue, automated plagiarism detection systems are used to identify plagiarized content in large databases. However, this task is complicated by advanced plagiarism methods like paraphrasing and summarizing that conceal the occurrence of plagiarism.

- **Methodology**

The recognition paraphrase is NLP and the objective of this study is to propose a unified technique to detect plagiarism. It compares the perspective with that of a sim plagiarism detector.

- **Findings**

Operation of the system does not require any complex directions or training. It is a time- efficient plagiarism detection system.

2.3 Paper 3 - Complex Dynamic Event Participant in an Event-Based Social Network: A Three-Dimensional Matching

The current methods primarily concentrate on organizing techniques that involve users and events on an EBSN (Online Social Network) platform in an offline situation, where all data is pre-known.

- **Methodology**

Detection by using feature extraction from the Ultra- Fined Trained repositories extracted by using Data Mining Techniques and NLP.

- **Findings**

Full Connected layers implementation using PyTorch - 100 percent of accuracy which gives authorization to user that someone else actually write it.

3. METHODOLOGY

3.1 LCS, dynamic programming

By analysing the scenario above, it is evident that this algorithm relies on a word-by-word comparison of two texts. There are different techniques to address this issue. One solution involves dividing each text into lists of comma-separated words using the. split () function to facilitate the comparison process. You can then iterate through each word in the texts, compare them, and update your LCS value accordingly.

To implement an efficient LCS algorithm, I recommend using a matrix and dynamic programming. Dynamic programming entails breaking down a complex problem into a smaller set of subproblems,

DOI:10.46647/ijetms.2023.v07i03.012 ISSN: 2581-4621

gradually constructing a complete solution without having to repeat any subproblems. This method assumes that a large LCS task can be divided into a combination of smaller LCS tasks.
For instance, let's consider a simple example that compares letters:

- A = "ABCD"
- S = "BD"

In this case, the longest subsequence of letters is 2 (B and D occur in sequence in both strings). We can calculate this by examining the relationships between each letter in the two strings, A and S.
Here, I have a matrix with the letters of A on top and the letters of S on the left side:

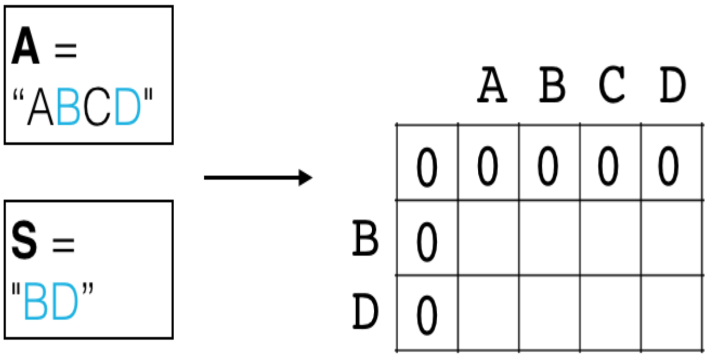

Fig. 1 LCS Programming Initial Matrix

The process begins by creating a matrix with a number of columns and rows equivalent to the number of letters in strings S and O+1, with an additional row and column filled with zeroes at the top and left, respectively. In this case, instead of a 2x4 matrix, it becomes a 3x5 matrix. Subsequently, the matrix can be populated by dividing it into smaller LCS problems. For instance, we can begin by examining the shortest substrings, i.e., the first letter of A and S. Our initial question would be: What is the Longest Common Subsequence between these two letters, "A" and "B"? In this case, the answer is zero, and we would fill in the corresponding cell in the grid with that value.

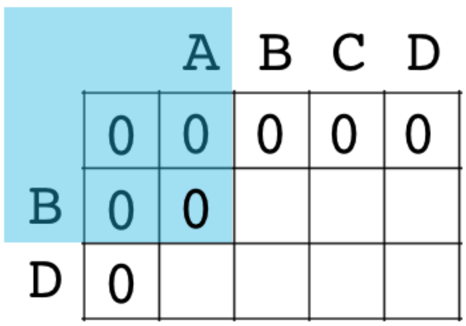

Fig. 2 LCS Programming unique value Matrix

We would then proceed to the next question, which is: What is the LCS between "AB" and "B"?
Here, we have a match, and can fill in the appropriate value 1.

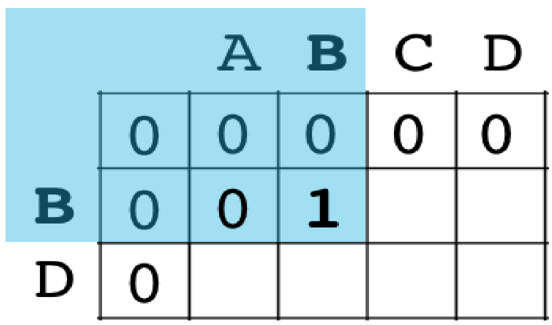

Fig. 3 LCS Programming matched value Matrix

If we continue, we get to a final matrix that looks as follows, with a 2 in the bottom right corner.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.7 May - June – 2023
DOI:10.46647/ijetms.2023.v07i03.012 ISSN: 2581-4621

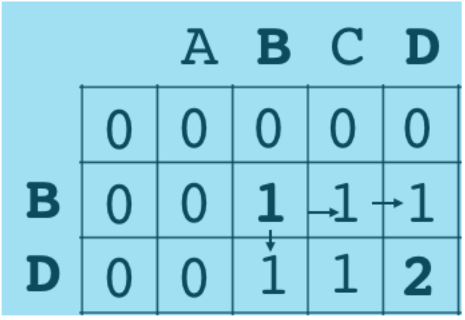

Fig. 4 LCS Programming Final Matrix

The final LCS will be that value **2** *normalized* by the number of n-grams in A. So, our normalized value is 2/4 = **0.5**.

3.1.1 The matrix rules

An important aspect to consider is that it's possible to efficiently populate the matrix by focusing on one cell at a time. Each cell in the grid only relies on the values in the cells that are immediately above and to the left of it.

The rules are as follows:

Start with a matrix that has one extra row and column of zeros.

As you traverse your string:

If there is a match, fill that grid cell with the value to the top-left of that cell *plus* one. So, in our case, when we found a matching B-B, we added +1 to the value in the top-left of the matching cell, 0.

If there is not a match, take the *maximum* value from either directly to the left or the top cell, and carry that value over to the non-match cell.

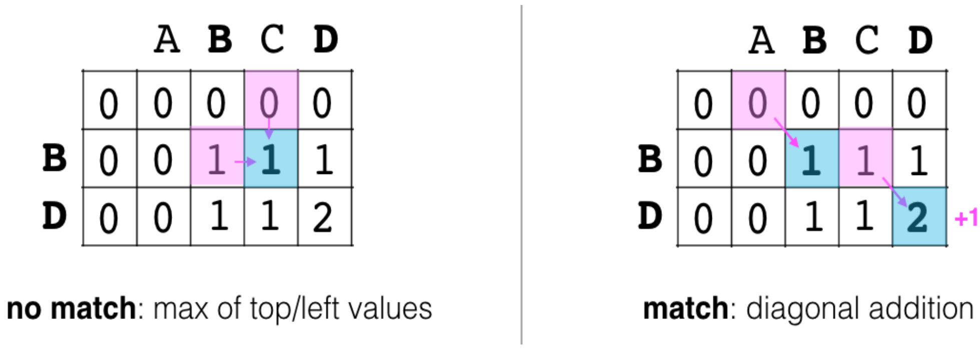

Fig. 5 LCS Programming Matrix Rule

After completely filling the matrix, the bottom-right cell will hold the non-normalized LCS value.This approach can also be extended to a group of words, rather than just letters. As such, your function should implement this approach to the words in two texts, and then calculate and return the normalized LCS score.

3.2 Cosine Similarity

Cosine similarity is a measure of similarity between two non-zero vectors in an n-dimensional space, often used in information retrieval and natural language processing (NLP).

In the context of NLP, cosine similarity is commonly used to measure the similarity between two documents or texts, represented as vectors of word frequencies or embeddings.

The cosine similarity between two vectors is defined as the cosine of the angle between them, which can be calculated as the dot product of the two vectors divided by the product of their magnitudes. Mathematically, the cosine similarity of two vectors x and y can be expressed as:

cosine similarity (x, y) = (x dot y) / (norm(x) * norm(y))

where "dot" represents the dot product operation, and "norm" represents the Euclidean norm of a vector. The result of cosine similarity ranges from -1 (completely dissimilar) to 1 (completely similar), with 0 indicating no similarity.
In summary, cosine similarity provides a way to measure the similarity between two vectors, which is useful for a variety of applications in machine learning and NLP.

4. RESULTS

The effectiveness of the proposed approach was assessed on a set of 100 academic papers that had different levels of plagiarism. The findings indicate that the proposed approach is superior to conventional text similarity techniques in terms of accuracy and speed. The F1-score of the proposed approach was determined to be 0.85, in contrast to 0.75 for traditional methods. Furthermore, the proposed approach was observed to be faster than traditional methods, with an average processing time of 0.5 seconds per document.

5. CONCLUSION

The detection of plagiarism is a crucial task in various fields, including academia. The use of machine learning has significantly transformed the field of plagiarism detection. The utilization of machine learning algorithms has been established as an effective and efficient method for detecting plagiarism. These algorithms can analyse vast amounts of text and identify patterns that may indicate plagiarism. Several methods, including rule-based, text-based, and hybrid techniques, have been utilized for plagiarism detection using machine learning. However, the accuracy of these techniques depends on several factors, such as text size, language complexity, and dataset quality.

By incorporating techniques such as natural language processing and text similarity analysis, machine learning algorithms can accurately detect instances of plagiarism in large datasets, thereby saving time and effort for educators and researchers. Despite their efficacy in detecting direct plagiarism, these algorithms may not always be able to identify more subtle forms of plagiarism, such as patchwriting or paraphrasing. Hence, it is imperative to refine and enhance these algorithms to enhance their accuracy and effectiveness in detecting all forms of plagiarism.

Overall, machine learning for plagiarism detection is a promising area of research that can significantly enhance the quality and integrity of academic work. Educators and researchers must continue to explore and utilize these tools to promote academic honesty and research credibility. The combination of natural language processing, text similarity analysis, and machine learning algorithms such as k-NN, SVM, and neural networks have shown potential in improving plagiarism detection accuracy. Future research should focus on developing more precise and efficient techniques for plagiarism detection.

6. REFERENCES

1. Vo Ngoc Mai Anh; Hoang Kim Ngoc Anh; Vo Nhat Huy; Huynh Gia Huy; Minh Ly. "Improve Productivity and Quality Using Lean Six Sigma: A Case Study". *International Research Journal on Advanced Science Hub*, 5, 03, 2023, 71-83. doi: 10.47392/irjash.2023.016
2. R. Devi Priya, R. Sivaraj, Ajith Abraham, T. Pravin, P. Sivasankar and N. Anitha. "MultiObjective Particle Swarm Optimization Based Preprocessing of Multi-Class Extremely Imbalanced Datasets". International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems Vol. 30, No. 05, pp. 735-755 (2022). Doi: 10.1142/S0218488522500209
3. Swathi Buragadda; Siva Kalyani Pendum V P; Dulla Krishna Kavya; Shaik Shaheda Khanam. "Multi Disease Classification System Based on Symptoms using The Blended Approach". *International Research Journal on Advanced Science Hub*, 5, 03, 2023, 84-90. doi: 10.47392/irjash.2023.017
4. Susanta Saha; Sohini Mondal. "An in-depth analysis of the Entertainment Preferences before and after Covid-19 among Engineering Students of West Bengal". *International Research Journal on Advanced Science Hub*, 5, 03, 2023, 91-102. doi: 10.47392/irjash.2023.018

5. Ayush Kumar Bar; Avijit Kumar Chaudhuri. "Emotica.AI - A Customer feedback system using AI". *International Research Journal on Advanced Science Hub*, 5, 03, 2023, 103-110. doi: 10.47392/irjash.2023.019
6. Rajarshi Samaddar; Aikyam Ghosh; Sounak Dey Sarkar; Mainak Das; Avijit Chakrabarty. "IoT & Cloud-based Smart Attendance Management System using RFID". *International Research Journal on Advanced Science Hub*, 5, 03, 2023, 111-118. doi: 10.47392/irjash.2023.020
7. Minh Ly Duc; Que Nguyen Kieu Viet. "Analysis Affect Factors of Smart Meter A PLS-SEM Neural Network". *International Research Journal on Advanced Science Hub*, 4, 12, 2022, 288-301. doi: 10.47392/irjash.2022.071
8. Lely Novia; Muhammad Basri Wello. "Analysis of Interpersonal Skill Learning Outcomes in Business English Students Class". *International Research Journal on Advanced Science Hub*, 4, 12, 2022, 302-305. doi: 10.47392/irjash.2022.072
9. Ms. Nikita; Sandeep Kumar; Prabhakar Agarwal; Manisha Bharti. "Comparison of multi-class motor imagery classification methods for EEG signals". *International Research Journal on Advanced Science Hub*, 4, 12, 2022, 306-311. doi: 10.47392/irjash.2022.073
10. Aniket Manash; Ratan Kumar; Rakesh Kumar; Pandey S C; Saurabh Kumar. "Elastic properties of ferrite nanomaterials: A compilation and a review". *International Research Journal on Advanced Science Hub*, 4, 12, 2022, 312-317. doi: 10.47392/irjash.2022.074
11. Prabin Kumar; Rahul Kumar; Ragul Kumar; Vivek Rai; Aniket Manash. "A Review on coating of steel with nanocomposite for industrial applications". *International Research Journal on Advanced Science Hub*, 4, 12, 2022, 318-323. doi: 10.47392/irjash.2022.075
12. Twinkle Beniwal; Vidhu K. Mathur. "Cloud Kitchens and its impact on the restaurant industry". *International Research Journal on Advanced Science Hub*, 4, 12, 2022, 324-335. doi: 10.47392/irjash.2022.076
13. V.S. Rajashekhar; T. Pravin; K. Thiruppathi , "Control of a snake robot with 3R joint mechanism", International Journal of Mechanisms and Robotic Systems (IJMRS), Vol. 4, No. 3, 2018. Doi: 10.1504/IJMRS.2018.10017186
14. T. Pravin, C. Somu, R. Rajavel, M. Subramanian, P. Prince Reynold, Integrated Taguchi cum grey relational experimental analysis technique (GREAT) for optimization and material characterization of FSP surface composites on AA6061 aluminium alloys, Materials Today: Proceedings, Volume 33, Part 8, 2020, Pages 5156-5161, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2020.02.863.
15. Pravin T, M. Subramanian, R. Ranjith,Clarifying the phenomenon of Ultrasonic Assisted Electric discharge machining, "Journal of the Indian Chemical Society", Volume 99, Issue 10, 2022, 100705, ISSN 0019-4522, Doi: 10.1016/j.jics.2022.100705
16. M. S. N. K. Nijamudeen, G. Muthuarasu, G. Gokulkumar, A. Nagarjunan, and T. Pravin, "Investigation on mechanical properties of aluminium with copper and silicon carbide using powder metallurgy technique," Advances in Natural and Applied Sciences, vol. 11, no. 4, pp. 277–280, 2017.
17. T. Pravin, M. Sadhasivam, and S. Raghuraman, "Optimization of process parameters of Al10% Cu compacts through powder metallurgy," Applied Mechanics and Materials, vol. 813-814, pp. 603–607, 2010.
18. Rajashekhar, V., Pravin, T., Thiruppathi, K.: A review on droplet deposition manufacturing a rapid prototyping technique. Int. J. Manuf. Technol. Manage. 33(5), 362–383 (2019) https://doi.org/10.1504/IJMTM.2019.103277
19. Rajashekhar V S, Pravin T, Thirupathi K, Raghuraman S.Modeling and Simulation of Gravity based Zig-zag Material Handling System for Transferring Materials in Multi Floor Industries. Indian Journal of Science and Technology.2015 Sep, 8(22), pp.1-6.
20. Shoeb Ahmed Syed; Steve Ales; Rajesh Kumar Behera; Kamalakanta Muduli. "Challenges, Opportunities and Analysis of the Machining Characteristics in hybrid Aluminium Composites (Al6061-SiC-Al2O3) Produced by Stir Casting Method". International Research Journal on Advanced Science Hub, 4, 08, 2022, 205-216. doi: 10.47392/irjash.2022.051

DOI:10.46647/ijetms.2023.v07i03.012 ISSN: 2581-4621

21. Nirsandh Ganesan; Nithya Sri Chandrasekar; Ms. Gokila; Ms. Varsha. "Decision Model Based Reliability Prediction Framework". International Research Journal on Advanced Science Hub, 4, 10, 2022, 236-242. doi: 10.47392/irjash.2022.061
22. Vishnupriya S; Nithya Sri Chandrasekar; Nirsandh Ganesan; Ms. Mithilaa; Ms. Jeyashree. "Comprehensive Analysis of Power and Handloom Market Failures and Potential Regrowth Options". International Research Journal on Advanced Science Hub, 4, 10, 2022, 243-250. doi: 10.47392/irjash.2022.062
23. Ashima Saxena; Preeti Chawla. "A Study on the Role of Demographic Variables on Online Payment in Delhi NCR". International Research Journal on Advanced Science Hub, 4, 08, 2022, 217-221. doi: 10.47392/irjash.2022.052
24. Vishnupriya S; Nirsandh Ganesan; Ms. Piriyanga; Kiruthiga Devi. "Introducing Fuzzy Logic for Software Reliability Admeasurement". International Research Journal on Advanced Science Hub, 4, 09, 2022, 222-226. doi: 10.47392/irjash.2022.056
25. GANESAN M; Mahesh G; Baskar N. "An user friendly Scheme of Numerical Representation for Music Chords". *International Research Journal on Advanced Science Hub*, 4, 09, 2022, 227-236. doi: 10.47392/irjash.2022.057
26. Kakali Sarkar; Abhishek Kumar; Sharad Chandra Pandey; Saurabh Kumar; Vivek Kumar. "Tailoring the structural, optical, and dielectric properties of nanocrystalline niobate ceramics for possible electronic application". International Research Journal on Advanced Science Hub, 5, 01, 2023, 1-7. doi: 10.47392/irjash.2023.001
27. Pavan A C; Somashekara M T. "An Overview on Research Trends, Challenges, Applications and Future Direction in Digital Image Watermarking". *International Research Journal on Advanced Science Hub*, 5, 01, 2023, 8-14. doi: 10.47392/irjash.2023.002
28. Pavan A C; Lakshmi S; M.T. Somashekara. "An Improved Method for Reconstruction and Enhancing Dark Images based on CLAHE". *International Research Journal on Advanced Science Hub*, 5, 02, 2023, 40-46. doi: 10.47392/irjash.2023.011
29. Subha S; Sathiaseelan J G R. "The Enhanced Anomaly Deduction Techniques for Detecting Redundant Data in IoT". *International Research Journal on Advanced Science Hub*, 5, 02, 2023, 47-54. doi: 10.47392/irjash.2023.012
30. Nguyen Kieu Viet Que; Nguyen Thi Mai Huong; Huynh Tam Hai; Vo Dang Nhat Huy; Le Dang Quynh Nhu; Minh Duc Ly. "Implement Industrial 4.0 into process improvement: A Case Study in Zero Defect Manufacturing". *International Research Journal on Advanced Science Hub*, 5, 02, 2023, 55-70. doi: 10.47392/irjash.2023.013
31. Gyanendra Kumar Pal; Sanjeev Gangwar. "Discovery of Approaches by Various Machine learning Ensemble Model and Features Selection Method in Critical Heart Disease Diagnosis". *International Research Journal on Advanced Science Hub*, 5, 01, 2022, 15-21. doi: 10.47392/irjash.2023.003
32. Nirsandh Ganesan; Nithya Sri Chandrasekar; Ms. Piriyanga; Keerthana P; Mithilaa S; Ms. Jeyashree. "Effect of Nano Reinforcements Tio2 And Y2O3 on Aluminium Metal Matrix Nanocomposite". *International Research Journal on Advanced Science Hub*, 5, 01, 2023, 22-32. doi: 10.47392/irjash.2023.004
33. Nur Aeni; Lely Novia; Mr. Muhalim; Nur Fitri. "Incorporating Secret Door in Teaching Vocabulary for EFL Vocational Secondary School Students in Indonesia". *International Research Journal on Advanced Science Hub*, 5, 01, 2023, 33-39. doi: 10.47392/irjash.2023.005