Machines and Algorithms
http://www.knovell.org/mna

Editorial

Volume 3 Issue 3

Editor: Dr. Aniqa Dilawari

Department of Computer Science & Information Technology, University of Home Economics, Lahore, 54000, Pakistan

From the Editor

I am honored to present Volume 3, Issue 3 of the *Machines and Algorithms* journal. The papers are thoughtfully selected to reflect the interests of our audience in both theoretical and applied areas of Computer Science. This issue features contributions in key domains such as Artificial Intelligence, Machine Learning, Image Processing, Algorithmic Optimization, IoT, and other emerging technologies shaping the future of computing.

In this issue, quality research articles have been compiled following a rigorous peer-review process. I sincerely want to thank all the authors for their valuable contributions and express appreciation to the reviewers for their dedicated and meticulous efforts. A short overview of the selected papers is discussed below.

The paper "Min Max Merge: A Novel Comparison based Sorting Technique for Data Intensive Processing" introduces a novel comparison-based sorting algorithm called Min-Max Merge Sort, which improves sorting efficiency. The algorithm combines the groups of numbers into one group. The idea was to divide the input values into multiple groups and then recursive merging based on minimum and maximum values. This prevented unnecessary data shifts and comparisons. The speed is comparatively faster in comparison to the traditional sorting methods. Space complexity also does not exceed O(n). The comparison between different methods show that it has outperformed recent sorting algorithms.

The paper "Development and Advantages of an AI-Driven Smart Lighting, Insect Detection and Automatic Spray System for Precision Agriculture" discusses about an AI-driven intelligent lighting and insect detection system designed to optimize pesticide use and enhance crop yield in agriculture. The CNN detection correctly detects insects with an accuracy of 95% which reduces the use of pesticides to 40% in comparison to the traditional methods. The concept of smart lighting uses HPS lamps which provides best lightening conditions enhancing photosynthesis to raise yield of the crop to 25%. The proposed system only sprays when and where needed to minimize environmental effects and pollution.

The paper "A Framework for the Authorship Identification in Research Papers" proposes an authorship identification and plagiarism detection. This study uses stylometric traits to determine authorship and plagiarism without using external sources. This indicator includes the writing style, language and sentence structure to assign part of documents with the authors. Clustering technique is used to estimate the number of times an author is used in a manuscript which will solve unethical authorship attributions and plagiarism. The results showed that this method is capable of detecting multi-author contributions and non-digital plagiarism.

The paper "Market Basket Data-Mining Analysis" explores data mining solutions for analyzing large and sparse sales transaction data, essential for customer-centric marketing. It discusses the challenges of using Association Rule Mining (ARM) on the sparse data and proposes using k-mean clustering and factors like Recency, Frequency and Monetary (RFM) model. This model has been tested on real world dataset and the set basis for future research in multi-label classification and sequence to sequence prediction.

Last but not the least, the paper on “Artificial Intelligence in the Education Process” studies the role of technology in education, questioning whether it can fully replace teachers or merely serve as a supportive tool. The study shows that human interaction is necessary in the learning process. While technology is a good tool to manage many aspects in education, its effectiveness depends on the goals and context where it is applied.

This concludes the summary of the papers finalized for this issue. Our team is very committed and works hard to ensure that Machines and Algorithms continues to publish quality research articles on interesting topics. In future, our aim is to expand the journals reach and collaborate with the leading research institutions and maybe introduce special issues on emerging technologies. Your participation and feedback are invaluable in shaping the future of this journal.

I want to thank once again to all researchers and reviewers to make this happen. Hope to see more issues and interesting papers in future.

Machines and Algorithms
http://www.knovell.org/mna

Research Article

Min Max Merge: A Novel Comparison based Sorting Technique for Data-Intensive Processing

Abbas Mubarak[1, *]

[1]Department of Computer Science, Institute of Southern Punjab, Multan, 60000, Pakistan
[*]Corresponding Author: Abbas Mubarak. Email: abbas007sheikh@gmail.com

Received: 02 August 2024; Revised: 3 September 2024; Accepted: 30 September 2024; Published: 10 October 2024
AID: 003-03-000041

Abstract: Sorting is significant as it is a prerequisite for many applications and functions. It is one of the oldest topics in literature which is still as important as it was in the beginning. Researchers are still working to design more efficient sorting methods either by improving existing sorting methods by reducing time complexity, space complexity, comparisons and shift operations or formulating new ones as sorting is very essential for most used applications such as databases for extracting useful information efficiently. This paper presents a novel comparison-based sorting method named as min max merge sort which works by merging groups of numbers into one group. The Proposed sorting algorithm is not only better than some old classical comparison-based sorting algorithms but it also performs better than some latest presented sorting methods. The basic idea behind proposed sorting algorithm is to divide the input array into different groups and then perform their recursive merging on the basis of their maximum and minimum entries, which circumvents unnecessary data shifts and comparisons. The speed of proposed algorithm is comparatively faster than the traditional sorting methods, as it exploits localized minimum and maximum data entries instead of recursively scanning entire input array. This algorithm exploits linear auxiliary space, as it performs in-place operations for combining groups in same array. Space complexity of proposed sorting method is O(n) as it does not require extra memory space. Performance comparison of proposed method with other sorts shows the superiority of our proposed method.

Keywords: Sorting; 2mm Sort; Time Efficient; Sorting Complexity;

1. Introduction

There are several types of algorithms, such as sorting algorithms, searching algorithms, compression algorithms and path finding algorithms [1]. Among the various branches of algorithms sorting is an important domain which is the oldest and most studied module in computer science [2]. The rearrangement of input data in a specific manner is known as sorting [3], [4]. Since 1950s, computer scientists are working on various sorting algorithms to improve the complexity of old sorting methods in terms of space and time complexity [5]–[8]. A number of top computer scientists extensively studied this topic such as Turing award winner Tony Hoare, Von Neumann and Donald Knuth [9]. Sorting is integral part of approximately all computer and mobile applications. All software's uses different sorting methods however user always concern with fast sorting [10]. People always consider two things in all sorting functions, speed of sorting algorithm and simplicity of algorithm [1].

Sorting can be performed on numbers, strings or records containing both numbers and strings like IDs, names, or departments, etc. [11]. Each sorting algorithm has unique properties that add value to the specific function they are used to perform [12]. Every sort is not appropriate for every kind of data [11]. Every algorithm has its own best as well as worst case. Therefore, to find out best sorting method according to Big O is difficult [13]. A study is essential in order to develop or make new sorting algorithm, because not all algorithm works efficiently for the same problem [14]. As the world has become global village and information is increasing rapidly, therefore importance of efficient sorting has also increase [4]. To get efficient search results from big databases also requires data in specific order [15]. The way of sorting data is very important due to impact of execution time, how elements are swapped, compared and arranged is dependent of technique of particular sorting algorithm [1], [16].

Enormous number of sorting algorithms are in use to date. Out of which, bubble sort is used extensively as it is simple and spontaneous however it is not very efficient. Normally a sorting algorithm consists of comparison, swap, and assignment operations [17]. Every algorithm has different time complexity [10]. Sorting can be applied in three ways: vector sort, (list) table sort and address sort, depending on the data storage policy [18]. This study, use array data structure to store data for sorting where data can be in numeric or character form [19]. Sorting algorithms for serial computing (random access machines) allow only one operation to execute at a time. The sorting algorithms based on a comparison network model of computation, performed many operations simultaneously [3].

2. Literature Review

2.1. Overview of Sorting Methods

Sorting methods can be divided into two major groups: specialized sorts and general sorts. Each sorting algorithm possesses some particular properties i.e. each particular algorithm performs best when data is of specific type like small numbers, floating point numbers, big numbers and repeated numbers. Apart from the programming work, the availability of main memory, the size of disc or tape units, and the degree of list already ordering consideration in selecting a sorting method. Most of the sorting techniques are thus problem specific, that is, they perform effectively on some particular type of data or problem [11], [17]. From memory consumption standpoint, some sorting techniques require more space in memory than others; yet, these techniques allow faster sorting. Consequently, the choice of these methods relies on the location and purpose of the sorting of the inputs [1].

The comparison results can be obtained in three ways: 1, Comparison of execution time 2, Comparison of total number of comparisons and 3, Comparison of total swapping frequency [19]. Among the numerous sorting techniques available, the one optimal for a given application depends on several variables like data size, data type, and element distribution in a data collection. The efficiency of the sorting algorithm is also dynamically influenced by numerous factors that may be aggregated as the number of comparisons (for comparison sorting), number of swaps (for in place sorting), memory utilization and recursion [20].

2.2. Categorization of Sorting Methods

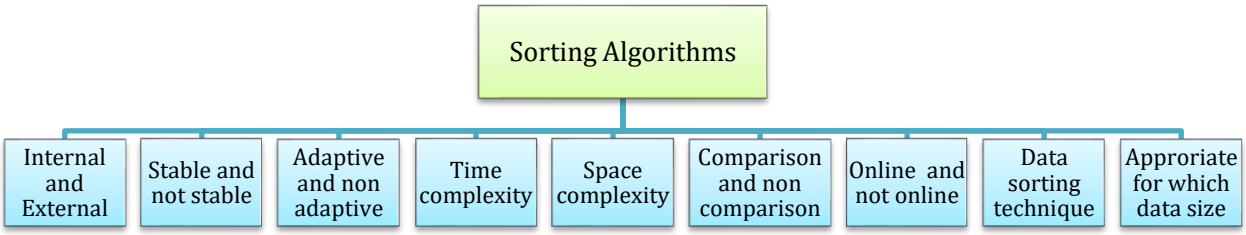

Figure 1: Categorization of Sorting Algorithms

Researchers have categorized the sorting methods in different categories i.e., depicted in figure 1 above [3], [12], [14], [18], [20]–[25].

2.2.1. Internal and External Sorting

Sorting method either do internal or external sorting. This is a key aspect for determining computational cost of sorting method. If a sorting function is performing sorting in Random Access Memory (RAM), which is also known as primary memory, then it is called internal sorting, whereas if the sorting function is using external memory, which is also known as secondary memory, then it is called externals sorting. In internal sorting, input numbers first load in main memory RAM and then processed inside RAM. Usually, internal sorting is much faster as it does not require data to be fetched from outside RAM. Some examples of internal sorting are 2mm sort, MMBPSS sort, Bubble sort, Min-Finder sort, Insertion sort and Selection sort [22]. While sorting algorithms which follows external sorting mechanism first brought data into primary memory from secondary storage and sorting is done. The process of fetching input data continuous during the sorting process as RAM sometimes could not store all data when the input size is too big. Heap sort, distribution sort and external merge sort are examples of external sorting.

2.2.2. Stable and Non-Stable Sorting

Some sorting methods are in the category of stable sorts while some are called not stable sorting algorithm. A sorting algorithm which retains the original input order after applying sorting process is known to be stable sorting algorithm. Stable sorting algorithm is needed where we have to retain sequence of equal values. For example, line of people waiting for some process according to their ages i.e. person with more age will be processed earlier but due to equal age stable sort will preserve the original sequence record i.e. first come first serve. Whereas those who are not stable sorting algorithms can change the place of occurrences of equal values in resultant list. Stability is usually not required when all the elements are different. Insertion sort, merge sort, bubble sort and counting sort are examples of stable sorting algorithms while heap sort, quick sort, selection sort and shell sort are examples of not stable sorting algorithms.

2.2.3. Adaptive and Non-Adaptive Sorting

An adaptive sorting algorithm is one that takes advantage of particular input sequence resulted decrease in time complexity. Insertion sort is an example of adaptive sort. It performs very fast when input data is nearly sorted. Quick sort is another example of adaptive sorting. Its time complexity reflects in different variation of input data such as random, sorted and reverse sorted.

2.2.4. Time and Space Complexity

According to time complexity sorting algorithms can majorly be divided into two groups i.e. O(nlogn) and O(n2). Insertion Sort, Bubble Sort, MinFinder and Selection Sort and 2mm sort comes under N^2 family while heap sort and quick sort belongs to time complexity O(nlogn). Space complexity determines the memory space taken by any algorithm during execution, and therefore this is an important aspect in time complexity. Space complexity of various algorithms can be different and the most efficient are those with less space.

2.2.5. Comparison and Non-Comparison Sorting

In comparison-based sorting algorithm, sorting is done by comparing numbers with each other to find minimum or maximum number for its proper position. Insertion, Quick and Bubble sort are some examples of comparison-based sorting. While in non-comparison-based sorting numbers comparison is not done and numbers are sorted using some other technique. Bucket and Radix sort are examples of non-comparison-based sorting.

2.2.6. Online and Offline Sorting

A sorting method that can work at the time of input given and full array of numbers is not required to input before processing, this process is known as online sorting. Whereas in offline sorting the whole input must be entered before algorithms processing. Insertion sort is an example of online sorting.

2.3. Characteristics of Good Sorting Algorithm

Some important characteristics of algorithm are [14], [26].

- **Finiteness:** It must be stopped after execution of restricted numbers of steps.
- **Definiteness:** All steps of an algorithm must be unambiguous and clearly defined.
- **Input:** The algorithm must have input values from a definite set.
- **Output:** An algorithm must produce some output form the given input.
- **Effectiveness:** The algorithm should execute each step exactly using a certain amount of time.
- **Correctness:** The algorithm must produce the correct output values for every finite set of inputs.

2.4. Sorting Algorithms Applications

Many algorithms, in addition to their primary function of sorting, also make use of a variety of methods to sort lists as a preparatory step in order to cut down on the amount of time it takes for them to carry out their tasks [20]. Sorting is also significant for searching, merging and normalization [10], [12], [14], [17], [18]. Successful sorting is essential to improve the utilization of another algorithm [23]. Sorting is also used in Central Processing Unit (CPU) scheduling in Operating system, recommendation system based on search time, Television channels sorted based on view time [27]. Sorting is inevitable in query retrieval and different types of join such as sort merge join [15]. Database query processing needs algorithms for duplicate removal, grouping, and aggregation, whereas in-stream aggregation is most efficient by far but requires sorted input [28].

3. Motivation

The author's in [1] presented a novel sorting algorithm which sorts the input list without comparisons, however it performs some manipulations with array indices for the purpose of sorting. The proposed sorting method in [1] uses two arrays for input sorting and two arrays for calculating repeated values. The drawback of above-mentioned sorting method is the extra usage of memory space [1]. The author's in [19] proposed new variant of selection sort algorithm called MMBPSS and it sorts the input list by finding out minimum and maximum numbers from first and second half of input list separately. It than compares the obtained numbers of both halves to find out minimum and maximum number of complete lists. The drawback of this algorithm is that it finds 4 numbers in each iteration and uses only 02 numbers. The authors in [2] brought a new variant of Timsort algorithm named as adaptive shivers sort. Its computational cost and number of comparisons are less than Timsort algorithms. The author's in [29] presented a novel comparison free sorting method with input K-bit binary bus. It operates on one-hot weight representation. For example, element 5 in binary is 101 whose one-hot representation is 100000. This binary to one-hot representation can be done using conventional one-hot decoder [29]. The authors in [30] introduced a new sorting method min-max sorting algorithm which works by finding minimum and maximum number form the input list and adjust them at first and last index of the list [6].

4. Review of 2mm Sorting Method

2mm sorting method is a comparison-based sorting method. In one cycle it changes 02 numbers as briefed in Pseudo Code below.

2mm Pseudo Code [22]
"Length ← size of (array)

```
midindex←Length/2
minindex←0
maxindex←Length -1
min ← a [minindex]
max ← a [minindex]
minloc ←maxloc←0


For i= minindex+1 to midindex
    If a[i] <min
        min←a[i]
        minloc←i
    End If
    If a[i]>max
        max←a[i]
        maxloc←i
    End If
  exchange a[minindex] with a[minloc]
  exchange a[midindex] with a[maxloc]
End For loop
 min ← a [midindex+1]
 max ← a [midindex+1]
For i= midindex+2 to maxindex
    If a[i] <min
        min←a[i]
        minloc←i
    End If
    If a[i]>max
        max←a[i]
        maxloc←i
    End If
  exchange a[midindex+1] with a[minloc]
  exchange a[maxindex] with a[maxloc]
End For loop
  minindex←0
  maxindex←Length -1
  min ← a [minindex]
  max ← a [maxindex]
```

From beginning to mid index and mid + 1 to end index, 2mm determines minimum and maximum numbers. Thus, it determines minimum and maximum number of lists by comparing minimum and

maximum numbers to ascertain minimum and maximum numbers of full array [22]. Next iterations save the previously discovered numbers and follow the same continuous process until all numbers are arranged.

Initially some variables are initializing. Then a for loop will execute from first half of array and another for loop will execute for another half of array. Each for loop will find minimum and maximum numbers of respective sub arrays and adjusts them at first and last index of that sub part of array. The above code will make the input list executable for next mail while loop processing. The main while loop required the input list in a manner where minimum and maximum numbers needs to be adjusted at each sub part of array for execution of one of the four cases inside main while loop. After the above code execution there can be 04 possibilities of array therefore for each possibility while loop have 04 cases and respective case only will be executed after if decision. One case is where both maximum and minimum numbers are already at their correct positions i.e. minimum number at first index and maximum number at last position. Second case is both numbers are not at their positions i.e. minimum number at mid+1 index and maximum number is at mid index. Third case is where both maximum and minimum numbers of whole array are at first sub array and last case is where both numbers are at second sub arrays [22].

```
While (minindex < midindex-1)
{
    If (min<=a[midindex+1])
    {
      If( max>=A[midindex]
        {
          Case1
        }
      Else
        {

          Case 2

        }
    }
    Else
    {
      If(max>=a[midindex])
        {
          Case 3
        }
      Else
        {
          Case 4
        }
    }
}
```

5. Proposed Algorithm - Min Max Merge Algorithm

The proposed sorting algorithm belongs to comparison-based family and it uses preprocessing presented in [31]. The basic operation in comparison-based sorting method is comparison of two input elements [32]. In its initial phase after applying preprocessing, it first sorts the input array into groups of two consecutive elements by comparing them. It than merge the input list elements in multiplication of 2 i.e. first sort group of two elements, then combine two groups and sort 4 elements, following up combine two groups of 4 element and sort 8 elements and same process continues until all elements are not sorted. In case of remaining last elements which are left without any group it adjusts them into previous groups. For example, in the input list of 10 elements when min max merge will make group of 4 elements than last 2 element will

be adjust with second group i.e. index 4 to 7 where index starts from 0. As min max merge is not use additional array for merging process therefore it combines group of elements in the same array by bubbling the elements. Min max merge 2mm sort min max formula and merge sort merging technique for sorting process.

At Mendeley Data the source code of Min Max Merge sorting algorithm is available (https://dx.doi.org/10.17632/sjxbcn97n6.1)

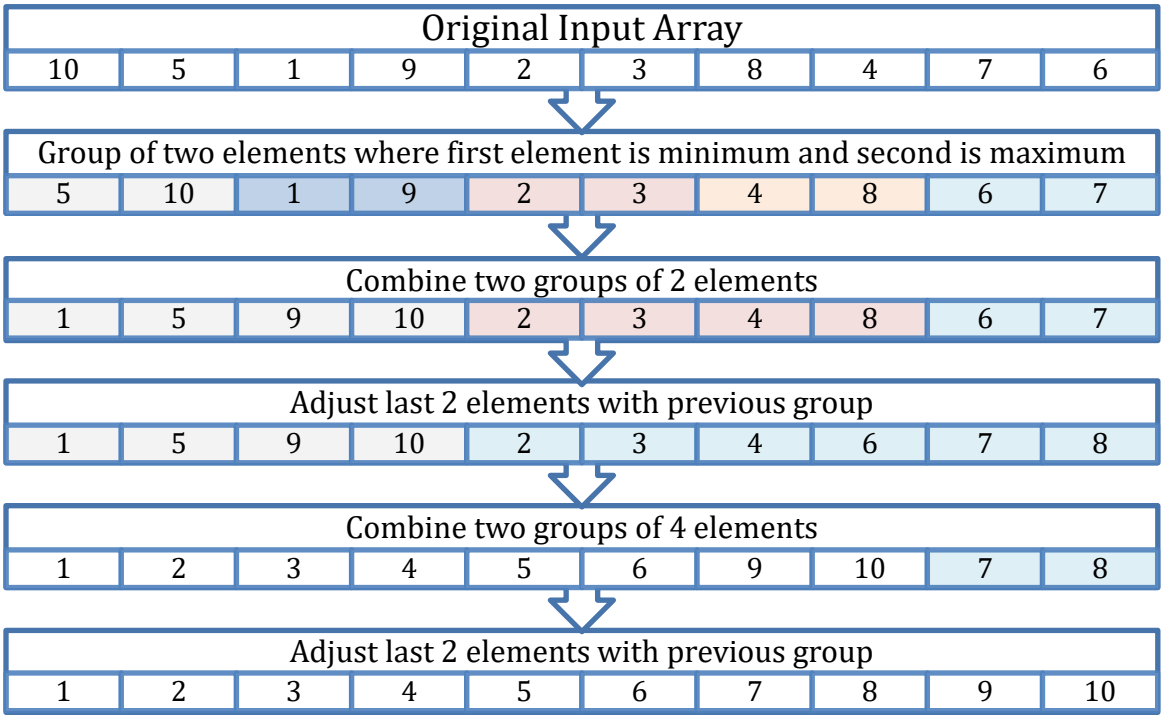

Figure 1: Dry-Run of Proposed Sorting Mechanism

5.1. Mathematical Analysis

Time complexity for applying data preprocessing as adapted by the study [31], is calculated below in equation 1 .

$$T(n) = O\left(\frac{n}{2}\right) + O\left(\frac{n}{2} - 3\right) + O(\frac{n}{2} - 3) + O(\frac{n}{4} + 1) + O(\frac{n}{2}) + O\left(\frac{n}{2} - 3\right) + O\left(\frac{n}{2} - 3\right)$$

$$T(n) = C + (\frac{2n + 2n - 12 + 2n - 12 + n + 4 + 2n + 2n - 12 + 2n - 12}{4})$$

$$T(n) = C + (\frac{13n - 44}{4})$$

$$T(n) = \Omega(n) \quad (1)$$

Time complexity for proposed method has been calculated below in equation 2. To sort input list in the manner of minimum and maximum $O\left(\frac{n}{2}\right)$ will execute. Then loop will be executed to adjust group of 4 numbers and will execute $\left(\frac{n}{4}\right)$ times and it will execute $\left(\frac{n}{4}\right)$ times for whole array. Time complexity will be,

$$T(n) = O\left(\frac{n}{2}\right) + \left(\frac{n}{4}\right)\left(\frac{n}{4}\right) + \left(\frac{n}{8}\right)\left(\frac{n}{8}\right) + \left(\frac{n}{16}\right)\left(\frac{n}{16}\right) + \cdots \left(\frac{n}{n}\right).1$$

$$T(n) = O\left(\frac{n}{2}\right) + \left(\frac{n^2}{16}\right) + \left(\frac{n^2}{64}\right) + \left(\frac{n^2}{256}\right) + \cdots 1$$

$$T(n) = O\left(\frac{128n + 16n^2 + 4n^2 + n^2}{256}\right) + \cdots 1$$

$$T(n) = O\left(\frac{128n + 21n^2}{256}\right) + \cdots 1$$

$$T(n) = O(n^2) \quad (2)$$

6. Results and Discussion

In order to accomplish the task of reducing the amount of computational complexity and time required to carry out swapping, comparison, and assignment operations, an efficient sorting algorithm is being developed [4]. The experiment results of insertion, bubble, selection, MMBPSS, MinFinder and 2mm were taken for comparison from our preceding study [22]. These results are compared with our proposed sorting algorithm results.

Table 1: Computational Time analysis of different sorting algorithms in seconds on random input data

Sorts | input	Bubble	Selection	Insertion	MMBPSS	MinFinder	2mm	Min max merge
100	0.000727	0.0057	0.0047	0.0051	0.0052	0.000333	0.0010053
1000	0.001999	0.0009994	0.0009994	0.0010068	0.002001	0.000999	0.0249867
100000	27.5589	11.3939	7.00767	10.8788	16.6568	4.16842	2.13268
200000	114.27	46.2444	29.046	27.5309	66.729	18.0029	8.54076
300000	257.078	104.649	64.0444	62.0226	148.685	41.0046	19.1262
400000	458.822	185.428	113.453	109.771	268.054	73.2097	33.8753
500000	851.909	250.735	173.653	181.682	393.721	111.081	89.8946

Table 2: Computational Time Analysis on Reverse Data to analyze Worst Case

Input| Sorts	Insertion	Bubble	Selection	MinFinder	MMBPSS	2mm	Min max merge
1000	0.0019997	0.002001	0.0019988	0.002	0.0010068	0.0009986	0.0001
10000	0.137915	0.176891	0.118927	0.235857	0.0699651	0.0579659	0.0007
100000	14.0323	17.806	12.7451	23.1758	6.55928	4.90797	0.0019977
200000	57.1027	72.666	48.548	92.0527	26.2981	19.7298	0.003998
300000	129.583	164.516	109.16	210.879	59.578	44.2776	0.006995
400000	225.351	291.462	195.492	366.052	106.495	78.3613	0.007995
500000	371.548	538.847	289.507	691.669	183.389	126.828	0.0099765

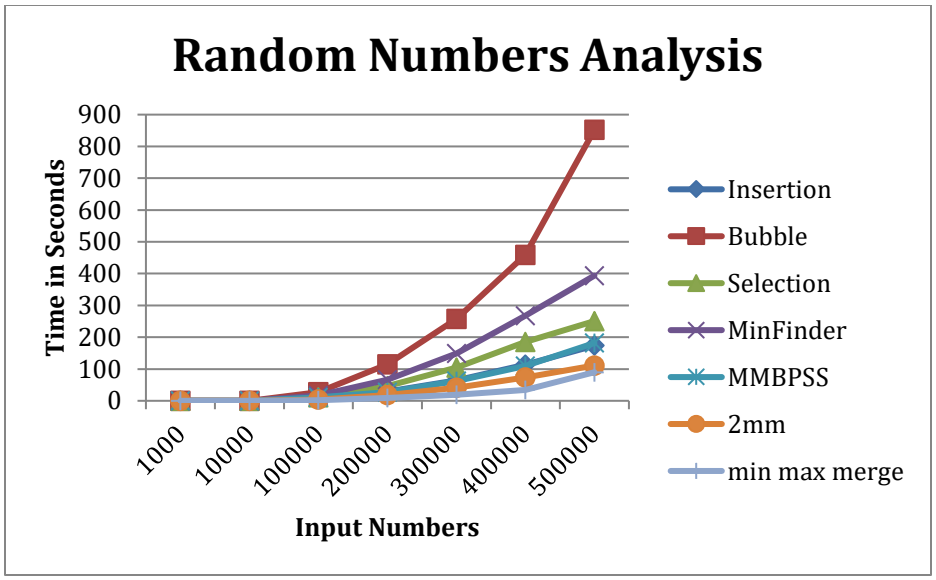

Figure 2: Computational Time visualization of different sorting algorithms in seconds on random input data

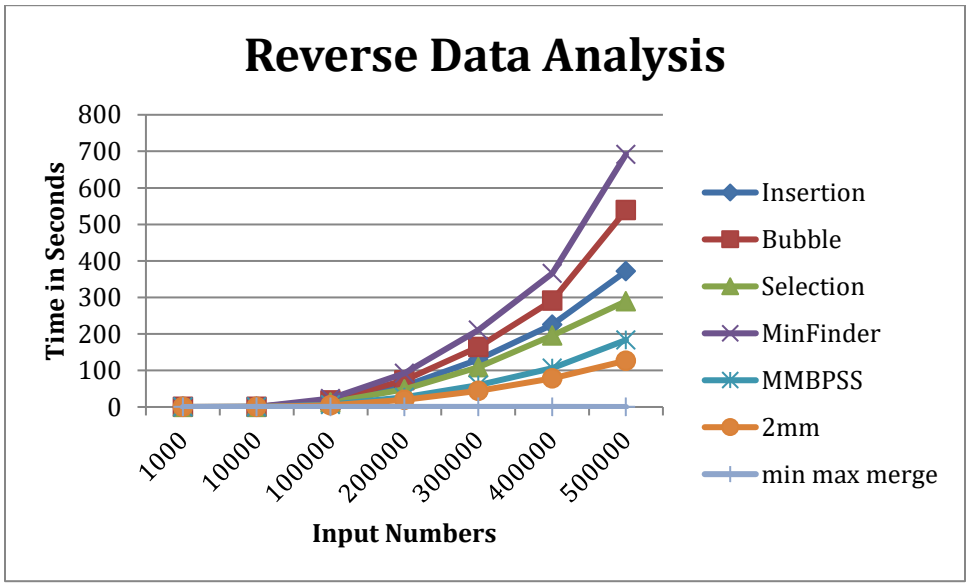

Figure 3: Computational Time Visualization on Reverse Data to analyze Worst Case

Based on the analysis of results, mentioned in Table 1, 2 and Figure 3, 4 it could be said that, the proposed Min Max Merge Sort shows superior performance with random data inputs and achieves better computational time than all other algorithms especially when working with large datasets such as 500k elements which finished execution in 89.89 seconds while Bubble Sort required 851.91 seconds. The performance of Min Max Merge Sort remains stable with worst-case input data because it demonstrates limited runtime slowness while outperforming all other techniques (for example, achieving 0.0099 seconds for 500k elements as opposed to 538.84 seconds for Bubble Sort).

7. Conclusion

Sorting is significant branch of algorithms. Researchers are continuously working on developing efficient

sorting methods and improving available functions due to phenomenal increase of data. This research presents new sorting method with preprocessing technique. It uses merge technique for sorting like merge sort. The huge difference is that where merge sort uses extra array for combining two groups of pieces, min max merge combines the two groups into one group in same memory space and this was the biggest challenge in this study. The proposed algorithm can be used for various types of data as it also uses preprocessing before applying original algorithm which helps in making data suitable for algorithm. Extensive analysis proves that proposed sorting method with preprocessing technique is significantly better than other comparison-based sorting methods. The authors aim to develop algorithm parallelized version in future.

Ethical Approval: The purpose of this research is to develop a computational model, rendering it unnecessary to involve human or animal subjects.

Funding Statement: This research has not received funding from any external source.

Conflicts of Interest: Author of this study declare no conflicts of interest.

Data Availability: In addition to the raw data, program code and supplementary materials, we have provided detailed documentation outlining the methodology, data collection procedures, and analysis techniques employed in this study.

References

[1] F. Idrizi, A. Rustemi, and F. Dalipi, “A new modified sorting algorithm: a comparison with state of the art,” in *2017 6th Mediterranean Conference on Embedded Computing (MECO)*, 2017, pp. 1–6.

[2] V. Jugé, “Adaptive Shivers sort: an alternative sorting algorithm,” arXiv Prepr. arXiv1809.08411, 2018.

[3] O. O. Moses, “Improving the performance of bubble sort using a modified diminishing increment sorting,” Sci. Res. Essay, vol. 4, no. 8, pp. 740–744, 2009.

[4] S. S. Moghaddam and K. S. Moghaddam, “On the performance of mean-based sort for large data sets,” IEEE Access, vol. 9, pp. 37418–37430, 2021.

[5] M. Shabaz and A. Kumar, “SA sorting: a novel sorting technique for large-scale data,” J. Comput. Networks Commun., vol. 2019, no. 1, p. 3027578, 2019.

[6] W. H. Ford, Data Structures with C++ Using STL, 2/e. Pearson Education India, 2002.

[7] H. Rohil and M. sha, “Run Time Bubble Sort – An Enhancement of Bubble Sort,” Int. J. Comput. Trends Technol., vol. 14, pp. 36–38, 2014, doi: 10.14445/22312803/IJCTT-V14P109.

[8] R. Shah, R. Gadia, and A. Joshi, “A Novel Approach to Sorting Algorithm,” in Research Advances in Network Technologies, CRC Press, 2023, pp. 179–190.

[9] P. Olukanmi, P. Popoola, and M. Olusanya, “Centroid Sort: a clustering-based technique for accelerating sorting algorithms,” in 2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC), 2020, pp. 1–5.

[10] H. R. Singh and M. Sarmah, “Comparing rapid sort with some existing sorting algorithms,” in Proceedings of Fourth International Conference on Soft Computing for Problem Solving: SocProS 2014, Volume 1, 2015, pp. 609–618.

[11] M. H. I. Bijoy, M. R. Hasan, and M. Rabbani, “RBS: a new comparative and better solution of sorting algorithm for array,” in 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1–5.

[12] P. Kumar, A. Gangal, S. Kumari, and S. Tiwari, “Recombinant sort: N-dimensional cartesian spaced algorithm designed from synergetic combination of hashing, bucket, counting and radix sort,” arXiv Prepr. arXiv2107.01391, 2021.

[13] M. Marcellino, D. W. Pratama, S. S. Suntiarko, and K. Margi, “Comparative of advanced sorting algorithms (quick sort, heap sort, merge sort, intro sort, radix sort) based on time and memory usage,” in 2021 1st international conference on computer science and artificial intelligence (ICCSAI), 2021, vol. 1, pp. 154–160.

[14] M. S. Rana, M. A. Hossin, S. M. H. Mahmud, H. Jahan, A. K. M. Z. Satter, and T. Bhuiyan, “MinFinder: A new approach in sorting algorithm,” Procedia Comput. Sci., vol. 154, pp. 130–136, 2019.

[15] A. Prasad, M. Rezaalipour, M. Dehyadegari, and M. Nazm Bojnordi, “Memristive Data Ranking,” 2021, pp. 440–452, doi: 10.1109/HPCA51647.2021.00045.

[16] A. B. G. Santos, M. F. Ballera, M. V Abante, N. P. Balba, C. B. Rebong, and B. G. Dadiz, “Asymptotic analysis of the running time performed by various sorting algorithms,” in 2021 International Conference on Intelligent Technologies (CONIT), 2021, pp. 1–6.

[17] M. Khairullah, “Enhancing worst sorting algorithms,” Int. J. Adv. Sci. Technol., vol. 56, pp. 13–26, 2013.

[18] W. Min, “Analysis on bubble sort algorithm optimization,” in 2010 International forum on information technology and applications, 2010, vol. 1, pp. 208–211.

[19] K. Thabit and A. A. BAWAZIR, “Novel approach of selection sort algorithm with parallel computing and dynamic programing concepts,” J. King Abdulaziz Univ. Comput. Inf. Technol. Sci., vol. 2, pp. 27–44, 2013.

[20] A. S. Mohammed, \cSahin Emrah Amrahov, and F. V Çelebi, “Bidirectional Conditional Insertion Sort algorithm; An efficient progress on the classical insertion sort,” Futur. Gener. Comput. Syst., vol. 71, pp. 102–112, 2017.

[21] N. Faujdar and S. P. Ghrera, “Analysis and testing of sorting algorithms on a standard dataset,” in 2015 Fifth International Conference on Communication Systems and Network Technologies, 2015, pp. 962–967.

[22] A. Mubarak, S. Iqbal, T. Naeem, and S. Hussain, “2 mm: A new technique for sorting data,” Theor. Comput. Sci., vol. 910, pp. 68–90, 2022.

[23] S. M. Cheema, N. Sarwar, and F. Yousaf, “Contrastive analysis of bubble \& merge sort proposing hybrid approach,” in 2016 Sixth International Conference on Innovative Computing Technology (INTECH), 2016, pp. 371–375.

[24] A. Shatnawi, Y. AlZahouri, M. A. Shehab, Y. Jararweh, and M. Al-Ayyoub, “Toward a new approach for sorting extremely large data files in the big data era,” Cluster Comput., vol. 22, pp. 819–828, 2019.

[25] P. Prajapati, N. Bhatt, and N. Bhatt, “Performance comparison of different sorting algorithms,” vol. VI, no. Vi, pp. 39–41, 2017.

[26] D. E. Knuth, The art of computer programming, vol. 3. Pearson Education, 1997.

[27] S. K. Gupta, D. P. Singh, and J. Choudhary, “New GPU Sorting Algorithm Using Sorted Matrix,” Procedia Comput. Sci., vol. 218, pp. 1682–1691, 2023.

[28] T. Do, G. Graefe, and J. Naughton, “Efficient sorting, duplicate removal, grouping, and aggregation,” ACM Trans. Database Syst., vol. 47, no. 4, pp. 1–35, 2023.

[29] S. Abdel-Hafeez and A. Gordon-Ross, “An Efficient O ($ N $) Comparison-Free Sorting Algorithm,” IEEE Trans. Very Large Scale Integr. Syst., vol. 25, no. 6, pp. 1930–1942, 2017.

[30] A. Agarwal et al., “A new approach to sorting: min-max sorting algorithm,” SORT, vol. 2, no. 2, p. n2, 2013.

[31] A. Mubarak, S. Iqbal, Q. Rasool, N. Asghar, N. Faujdar, and A. Rauf, “Preprocessing: A method for reducing time complexity,” J. Comput. \& Biomed. Informatics, vol. 4, no. 01, pp. 104–117, 2022.

[32] K. Iwama and J. Teruyama, “Improved average complexity for comparison-based sorting,” Theor. Comput. Sci., vol. 807, pp. 201–219, 2020.

Machines and Algorithms
http://www.knovell.org/mna

Review Article

Development and Advantages of an AI-Driven Smart Lighting, Insect Detection and Automatic Spray System for Precision Agriculture

Muhammad Ahmad[1, *], Zaib UN Nisa[1], Muhammad Haroon[1]

[1] Department of Information Technology, University OF Agriculture, Faisalabad, Burewala, 61010, Pakistan

[*]Corresponding Author: Muhammad Ahmad. Email: aj040744@gmail.com

Received: 06 August 2024; Revised: 02 September 2024; Accepted: 2 October 2024; Published: 10 October 2024

AID: 003-03-000042

Abstract: There is no area more difficult for crop improvement and efficiency in the utilization of resources in agriculture than maintaining the environment under this scenario. Traditional agriculture is a matter of the use of relatively broad insecticides and manpower with wide-spectrum inefficiency and ecological damage. This paper presents an AI-driven intelligent lighting system that performs real-time insect detection and books a sprayer to thus automate and optimize an even more mechanized agricultural practice. The CNN-based insect detection module actually correctly classifies with high-precision recall rates of substantially minimized pesticide use. A statistic of the system's performance indicators like detection accuracy at 95%, and reduction to a level of 40% in the used pesticides is a testament to the even better system performance compared to the traditional methods. The smart lighting aspect would employ HPS lamps and provide the best lighting conditions so enhanced photosynthesis further raises crop yield by 25%. The robotic spray system would spray the pesticide only where required to minimize environmental effect and resource wastage. The solution proposed here offers fast, complete, scalable, and environmentally friendly precision farming that the current solution lacks. An innovative contribution to current agricultural practice is introduced through this research. The issues to be resolved in this research are pest control, optimization of resources, and sustainability in the environment. Future development of this research will entail investigating the feasibility of upscaling this solution, incorporating more features that AI can provide and making it highly accessible to both small- and large-scale farmers.

Keywords: HPS lamps for Crops; Pest Control; Robotic Spray System; Machines for Crop Yield; AI in Agriculture; Real-Time Monitoring;

1. Introduction

Agriculture today has the twin task of producing more food and feeding a growing planet. Advantage over population, and global population in relation to resources and environmental effects. Traditional agriculture has involved extensive capital and human labor investment and high only use of pesticides and fertilizers, and this had caused huge inefficiencies and ramifications on the environment. The processes are not sustainable in the longer term and have resulted in soil erosion, water pollution, and loss of biodiversity [1]. Seeking out the target of sustainable and efficient agriculture will present precision agriculture (PA) as a possible solution to these issues. PA facilitates the quantification and solution to output differences in agriculture through sophisticated tools such as, machine learning, GPS, data management

and remote sensing. PA emphasizes efficient use of inputs such as water, fertilizer and pesticides to increase the farm output whilst maintaining environmental health [2]. Artificial Intelligence (AI) is very essential in case of the PA because it processes a large amount of data sets and also makes real time agricultural optimizations possible. Features such as modern lighting systems, real-time pest monitoring, and autonomous spraying make intelligent systems the foundation of a full-scale modern farming strategy. For instance, HPS lamps and LEDs of intelligent lighting systems keep stable light intensities, which promote growth of plants and make photosynthesis and yield as maximally as possible [3]. Targeted pest management strategies for pest control are achieved under real-time pest detection using machine vision and image processing.

With these technologies, when pests are correctly identified and classified, farmers are able to use pesticides in their targeted way and eliminate the need for using broad spectrum insecticides. Research shows that these innovations may result in significant reduction of the pesticide application as well as the adoption of the more environmentally friendly farming methods [4]. The use of computerized sprayers is part and parcel of the principles of precision farming. Such sprayers make it easier to spray the given number of pesticides and fertilizers to desired places thus enhancing efficiency in operations as well as reducing the destruction of the environment. By integrating technology of GPS and sensors with spraying equipment, it will be possible for the farmers to dynamically control spray patterns, droplet size, and application rates thereby increasing the adaptability of pest control measures [5]. What this study aims at is to develop an AI system to combine smart lighting, real-in-time insect monitoring and automated spraying to improve agricultural output whilst being environmentally sustainable and energy efficient. By means of this integrated system, we are to facilitate efficient, eco-friendly crop production with optimal yield. Our subsequent steps will be aimed at improving the functionality of the system and pushing its artificial intelligences forward in turn; the overall effectiveness of precision agriculture methods will be maximized.

2. Literature Review

Investigations and use of precision agriculture (PA) are essential to improve the sustainability and efficiency of production of agriculture; the proceedings of this method rely on advanced technology for the identification and intervention on variability of crops in fields. The key features of PA are remote sensing, Global Positioning System (GPS) technology, data analysis, and machine learning, which lead to more precise and input-saving agriculture practices. Integration of these technologies has shown vast potential in enhancing crop yields, reducing input prices, and mitigating environmental impacts [6, 7]. In precision agriculture, conventional pest management methods are based primarily on broad-spectrum use of pesticides, which tend to cause environmental degradation, loss of biodiversity, and inefficiencies, especially under intensive agriculture. Current automated systems, though an improvement, usually don't have in real-time the detection of pests, which results in under-treatment or over-treatment of pesticides. Furthermore, lighting options in current systems are usually just LEDs, which, though energy-saving, might not always be the best in terms of photosynthesis conditions for every crop. These limitations are overcome in the new system by its new integration of smart lighting, real-time live pest detection via CNN, and spraying systems. With the ability to lower the amount of pesticide consumption by 40% and increase crop output by 25% with high-pressure sodium (HPS) lamps, the system offers an efficient, green option to traditional systems. Being able to calibrate spraying strength compared to the concentration of pests as well as smooth integration of scalable solutions both to small holder as well as large holder farms is also evidence of its accessibility and ease of use. This blending of capabilities places emphasis on the system's inherent contribution to the modernization of agriculture and improving ecological sustainability as well as its capability of overcoming the hindrances and weaknesses of aged and traditional automation systems. Light is vital in plant formation and development. The farming activities are normally subject to natural daylight, which is unstable and insufficient, particularly in bad weather. Artificial lighting, like high-pressure sodium (HPS) lamps and light-emitting diodes (LEDs), has transformed controlled environment agriculture (CEA), providing uniform and optimal light conditions [8]. Intelligent lighting systems have been found in studies to significantly enhance photosynthesis, leading to increased food yields and quality. Studies [9] and [10]

indicated that diverse spectra of light have the ability to change plant structure, nutrient uptake, and resistance to diseases. The ability to tailor light intensities for crops is one of the major advances in farm technology. Pests are thought to be some of the major challenges in agriculture. Conventional approaches occasionally apply too much of harmful pesticides that kill beneficial insects, promote pest resistance as well as environmental degradation [11]. The use of image-processing and machine vision for real-time insect identification provides a practical solution. Improvements to computer vision and machine learning techniques have made possible the automation and accuracy of identification and classification of pests. Investigations by [12] and [13] reveal that image-processing technologies are effective for the detection and identification of a variety of agricultural pests. Rivero and Langridge and others have demonstrated that with the combination of automated pest control systems with these technologies, sustainable farming is achievable with reduced pesticide use. Autonomous spraying technologies signify a great leap forward in precision agriculture. Many times, the traditional spraying systems result in overuse of chemicals and increased operational cost, this is a prevalent problem. Automated sprayers then are able to deliver exact amounts of pesticides or fertilizers directly to targeted locales by prioritizing real time information [14], thereby increasing efficiencies and reducing impact to the environment. Work done by [15] and [16] verifies that the spraying by machine is more effective and more resource-efficient. Through the integration of these systems with GPS and sensor technology accuracy and efficiency can be improved in the processing.

Variability in spray patterns, droplet size and the application rate increase the responsiveness and sensitivity of pest control. The AI platforms are built to operate on big data, find patterns, and make instant data driven decisions that improve Farming strategies. In references [17] and [18], the use of AI to monitor crops and predict yields as well as the management of resources was discussed. Coupling the AI technology, smart lighting, and immediate pest surveillance, and self-propelled spraying the whole framework presents itself which addresses a variety of farming aspects simultaneously. This synergistic strategy provides strong responsiveness to the climate changes, as well as improved management of resources and the possibility for better yields in crop. AI is to significantly affect agriculture, and existing studies focus on additional improvement and adaptation of these technologies for various purposes. response Some of the barrier to take up of these innovations include data protection, high technology cost and lack of technical experience. In addition to this, continued research is indispensable with the aim of improving accuracy and dependability of AI-based systems as they continue to develop [19]. Future research should focus on the ways to make these technologies more available for farmers in developing countries. This involves, among other things, promoting development of cost-effective tools, providing farmers with skills, and setting solid procedures for handling of data. Addressing such problems, we can fully engage precision agriculture potential, offering more effective and environmentally friendly approach to farming to all the globe.

3. Methodology

3.1. Classification Method for Insect Detection

Insect detection is a vital aspect of the suggested AI-based system, and accurate classification must be done in order to allow accurate pesticide application. Classification is performed using convolutional neural networks (CNNs), a robust machine learning algorithm applied universally in image recognition processes. The methodology is explained in the steps below:

1. **Dataset:** Training and testing were performed on an open-source dataset, i.e., the PestNet dataset. The dataset consists of 10,000 labeled images of various pest species on various crops. Rotation, flipping, and cropping were adopted as techniques of improving model resilience and handling variety in pest appearance.
2. **Model Architecture:** The ResNet-50 model architecture utilized for insect classification was pre-trained using the Image Net dataset. Transfer learning was used to fine-tune the model for pest detection. The model consists of a number of convolutional and pooling layers, which enable it to extract complex pest features effectively.
3. **Training Process:** The data were partitioned into 70% training, 20% validation, and 10% test sets.

Adam optimizer with a learning rate of 0.001 was used to train the model, and categorical cross-entropy loss function for multi-class classification was used.

4. **Performance Metrics:** The classification achieved 95% accuracy, precision of 93%, recall of 94%, and an F1-score of 93.5%. These indicate the high reliability of the system to identify pest species with high accuracy.
5. **Real-Time Implementation:** The model was deployed on edge devices using an NVIDIA Jetson Nano, enabling real-time detection of pests in farms. The model detects pests in real-time from live camera feeds, automatically activating the spray system when needed.

Through CNNs and high-level image processing, the given system maintains accurate identification of the pests and therefore minimizes pesticide wastage and environmental pollution.

3.2. Performance Measurements and Verification

Performance analysis of the insect detection model is critical to ensure its reliability and accuracy when used in real-world scenarios. The system uses traditional performance metrics to ensure its functionality. The following metrics were used:

1. **Accuracy:** The model was 95% accurate overall, meaning that it correctly classified the majority of the insect species in the data.
2. **Precision:** Precision = True Positives / (True Positives + False Positives) is a measure of how effectively the model can avoid false alarms. Accuracy of the suggested system was 93%, i.e., the system had very few misclassifications.
3. **Recall (Sensitivity):** Recall is the proportion of how well the model can identify all the instances that are relevant, i.e., Recall = True Positives / (True Positives + False Negatives).
   The recall of the system was 94%, showing how successful the system was in identifying pest species without excluding any.
4. **F1-Score:** The F1-score, being the harmonic mean of precision and recall, was computed as:
   F1-Score = 2 × (Precision × Recall) / (Precision + Recall)
   The system's F1-score was 93.5%, providing a balanced measure of model performance.
5. **Validation Process:** The data set was split into 70% train, 20% validate, and 10% test subsets. 5-fold cross-validation scheme was employed in testing the robustness of the model to avert overfitting.
6. **Real-Time Testing:** It was used on a test farm that simulated conditions on farms. Real-time testing showed consistent performance with accurate detection of pests even under conditions of varied lighting and ambient. The results reveal that the system's performance is extremely effective in determining the identification of insect species accurately and reliably, confirming its appropriateness for precision agriculture applications.

3.3. Machine Learning Techniques for Image Analysis

The image processing module of the proposed system exploits the advanced machine learning algorithms in order to discern and differentiate pests in the agricultural fields, which are briefed below:

3.3.1. Model Selection

The system uses a pre-trained ResNet-50 model that is well known to support a deep network architecture and an ability to differentiate subtle image characterizations between countable and uncountable contour types. The dataset of the insect was used to fine tune a pre-trained ResNet-50 model for use towards pest classification.

3.3.2. Data Preprocessing

The preprocessing phase of input high-resolution imaging dataset includes following steps:

- **Resizing:** To make images compatible with ResNet-50, all the images were standardized to

the size of 224 × 224 pixels.

- **Normalization:** Normalized the pixel intensity to the interval [0, 1] to improve convergence of the model.
- **Augmentation:** Through the use of rotation, cropping and flipping they brought more variety to the dataset and contributed to avoiding overfitting.

3.3.3. Training and Optimization

The Adam optimizer was used to train the model using the following parameters:

- **Learning rate:** 0.001
- **Batch size:** 32
- **Epochs:** 50

A categorical cross-entropy loss function was employed to address multi-class classification.

3.3.4. Feature Extraction and Classification

The convolution layers of ResNet-50 identified high-level image features based on shape, texture, and pattern. The fully connected layers performed classification, generating probabilities for each class of pest.

3.3.5. Test Data Evaluation

It was tested on the test set with excellent scores:

- 95% accuracy
- F1-Score: 93.5%

3.3.6. Real-Time Integration

The model was deployed on an NVIDIA Jetson Nano, which enabled real-time image processing for agriculture. Raw live video streams obtained by the camera module were analyzed, and recognized pests were classified immediately. This robust machine learning solution delivers precise image analysis, enabling timely and precise pest identification across diverse agricultural ecosystems.

3.4. Reasons Why HPS Lamps Are Used

HPS lamps were selected as the first light source of the system proposed due to their established performance and efficiency in their application in agricultural use. Their reasons for selection are:

1. **Enhanced Photosynthesis Efficiency:** HPS lamps emit a broad light spectrum, particularly in the orange-red spectrum, that is very beneficial to the process of photosynthesis in plants. Experiments have proven that crops under the light of HPS lamps have yields as high as 25% compared to crops with the restriction to sunlight or their equivalent LED-based counterparts.
2. **Cost-Effectiveness:** Although they are energy efficient, LEDs can be much more costly to install initially compared to HPS lamps. HPS lamps are inexpensive for farmers, particularly in commercial farming, where lighting up vast spaces with LED lights may not be economically feasible.
3. **Lighting Uniformity:** The HPS lamp design promotes uniform light distribution, minimizing shadowing risk and providing uniform growth throughout the field. Regulated lighting is necessary to maintain crops healthy and productive, particularly in controlled environments.
4. **Durability and Reliability:** HPS lamps are very durable and are able to withstand extremely severe environmental conditions and are thus ideal for outdoor agricultural applications. They last longer than some traditional lighting systems, with fewer cases of replacement required.
5. **Comparison with LEDs:** Although LEDs provide spectrum tailoring, sometimes they do not provide intensity that some crops need in some stages of growth. HPS lamps, on the other hand, provide the intensity required for peak growth but at a cost-efficiency ratio.

6. **Environmental Impact:** While HPS lamps use a bit more power than LEDs, their capacity to increase crop yields and lower pesticide use compensates for the environmental trade-offs. It has automated controls to lower the energy requirements by managing lights based on crop requirements and the external environment. By incorporating HPS lamps in the proposed system, the solution optimizes efficiency, cost, and yield, thus making it a feasible solution for precision agriculture.

3.5. Automatic Spray System

To facilitate efficient integration and functioning, the mounting of an autonomous spray system on a robot to be used for precision agriculture is a task that involves planning and execution. The step-by-step procedure below shows how to mount an automated spray system on a robot:

1. **System Requirements:** It is necessary first to properly outline the requirements of the automated spray system. This will involve detailing the nozzle type, optimal spray rate, coverage, and volume of pesticide reservoir. The tank/reservoir volume is established based on the climatic needs of the area it will be installed in. For example, high-insect infestation area can be fitted with a large reservoir, while small insect density areas can fit small reservoirs. Tank volume can also be affected by crop density. For application in this project, a 50-liter tank with a gauge to indicate the level of spray remaining in the tank will be utilized [20].
2. **Proper Spray Nozzles:** Select the right spray nozzles based on your particular requirements of your application. Various crops, growth stages, and pests might require different types of nozzles, like fan nozzles, cone nozzles, or air-assisted nozzles. In our project, we will use flat fan nozzles to provide a flat and narrow spray pattern. Flat fan nozzles are typically used for crop spraying due to the fact that they can cover long distances effectively and thus provide uniform coverage on the target surface [21].
3. **Pesticide Reservoir:** Install a pesticide tank or reservoir on the robot platform such that it can be securely fixed in order to prevent leakages or spillages. The tank must be strong enough to withstand robot movement and vibration [22].
4. **Hoses and Piping:** Fit the needed hoses and pipes to enable the flow of pesticides from the container to the spray tips. Use materials compatible with the pesticide used to prevent the destruction of the parts of the system [23]. Adjust Spray Nozzles Mount and position the spray nozzles in a way that the crop to be targeted is well covered. Nozzle positions and quantities may differ according to specific requirements. Adequate placement allows efficient spray spreading with minimal wastage [24].
5. **Integrate Control System:** Integrate a control system to regulate the turning on and off of spray systems. This can include valves, pumps, and flow control. For accurate pesticide application, the control system needs to be accurate and responsive [25].
6. **Integrate Pest Detection Sensors:** Depending on your pest management strategy, integrate pest detection sensors. These sensors, such as cameras, infrared sensors, or other sensors, are important in determining where to apply pesticide in areas. Machine vision sensors, for example, identify insects by photographing their surroundings and analyzing the images, particularly where the pests would be [26].
7. **Image Acquisition:** Machine vision systems utilize cameras or image sensors to capture high-resolution images of the region of interest. The cameras may be fitted with a variety of lenses and filters to improve the image quality for the application.
8. **Image Preprocessing:** Pre-processing is used to improve the quality of images that are gathered and the identification of the insects. These involve resizing, image cropping, and color correction [27].
9. **Image Analysis:** Machine vision software reads the photographs to detect objects of interest, in this case insects. Segmentation, feature extraction, and pattern recognition are involved in the research. Pattern recognition and machine learning methods can be utilized to determine if the

detected objects are insects or non-insects. Models are trained using labeled data to make them more effective at detecting insects [28].

10. **Pest Detection and Decision-Algorithms:** Include pest detection and decision-making algorithms. The algorithms evaluate sensor data to determine when and where to turn on the spray system. Effective algorithms can greatly improve pesticide application effectiveness and accuracy [29].
11. **Robot Controller Communication:** Establish the communication and control links between the robot master controller and the automatic spray system. Coordination with the robot navigation system and other systems is ensured. Smooth integration is a prerequisite for concurrent operations [30].
12. **Carry Out Stringent Testing:** Test the integrated automated spray system stringently to guarantee that everything functions as required, from identifying the pests to proper application of pesticides. In terms of precision, calibrate the system whenever the need arises. Testing ensures detection of any possible defects prior to full utilization [31].
13. **Enforce Safety Provision:** Incorporate safety features to prevent accidental contact with pesticides. Safety interlocks and emergency cutoff devices are only a few examples. Both the environment and workers are safeguarded through safety [32].
14. **Scheduled Maintenance and Calibration:** Periodic maintenance and calibration of the automated sprayer system to ensure uniform and precise pesticide application. Maintenance guarantees long-term reliability and performance [33].
15. **Data Reporting and Logging:** Include data logging functionality to record pesticide applications. The information proves useful to monitor and confirm compliance with regulations. Detailed records serve to assist constant modifications and conduct audits [34].
16. **User Interface and Control:** Develop an easy-to-use interface by which operators can monitor and control the autonomous spraying system. These include tasks like opening and closing programs, parameterizing, and showing results of pest detection. A simple-to-use interface benefits usability as well as the efficiency of operation [35].
17. **Safety Measures:** Establish detailed safety measures and offer operator and user training to protect the robot and onboard spray system from misuse. Adequate training is critical for proper and safe system operation [36].

    Proper installation and integration of the automated spray system are necessary to ensure accurate and effective application of pesticide on agricultural land. Essential to guarantee consistent and reliable operation is regular system check and maintenance which leads to long-term sustainability in agriculture. This sophisticated targeting system used by the robot reduces the chemicals that are used thus reducing the cost and the environment left healthy. This sophisticated targeting system used by the robot reduces the chemicals that are used thus reducing the cost and the environment left healthy.

3.6. Machine Vision System and Image Processing Systems

It greatly increases the crop management skill of an automated precision agriculture robot when it is given a machine vision system, automatic spraying, and High-Pressure Sodium (HPS) lamps. To install a machine vision system on a robot like this one, the following specific steps are needed, which we describe down below:

1. **Appropriate Machine Vision Hardware:** This phase involves the selection of necessary and most crucial hardware for machine vision system, which usually consists upon cameras, lenses, sensors and the right lighting arrangement. Choose hardware elements that are customized to the robot's configuration and its operations environment. The chosen machine vision cameras are designed to accommodate the requirement of precise timing, ramped imaging, and specialized image processing software for great application performance. Such cameras are an important component of automated quality control systems. [37]

2. **Appropriate Image Processing Equipment:** This phase involves selection of relevant image capturing devices like cameras, image sensors and custom vision solutions designed for farm use. The type of camera used will depend on the task at hand, the environment and needs of the specific system. In order to ensure best image quality in this project, we have chosen to use Charge-Coupled Device (CCD) cameras, which are reputed for their low noise and quality image production. They are especially suitable for the applications little image quality oriented, such as scientific photography, microscopy; industrial inspection of high quality [38].
3. **Lens Filters and Covers:** To protect camera lens from possible damages and environmental dust, cover the camera lens with a lens cover or filter. It will be possible to eliminate the buildup from the water and dust by using a non-reflective material when coating the lens [39].
4. **Image Sensors:** CCDs are arranged in arrays in a grid like arrangement or a defined pattern. Every CCD sensor in the array takes part of the whole picture. CCD arrays are variable, being used for scientific imaging, astronomy, and design of high-tech digital cameras [40].
5. **Mounting and Positioning:** Choose the best places on the robot for camera and imaging vision placement for clear sight. Ensure that the cameras can provide a straight and no charge view of the crops and targeted areas. We place cameras on the robot's top, to ensure an unobstructed view, in our project [41].
6. **Connections and Wiring:** Include image-processing equipment in the robot's control system. Establish the cabling for signaling, sharing data and handling power on the system. A stable wiring can ensure seamless performance and data consistency [42].
7. **Calibration and Alignment:** It is therefore important to calibrate the image processing system to ensure the accurate and reliable picture recording and measurement. The precise alignment of lens and camera is highly crucial to get high quality and precise imaging. Calibration measures need frequent measures to verify the system's accuracy [43].
8. **Camera Settings and Parameters:** Adjust the exposure time, aperture, the focus and the camera white balance to achieve maximum image quality in your agricultural fulfilment. The adjustment of these settings could be made essential due to variations in lighting condition. Properly setting the system allows it to properly react to the change of situations [44].
9. **Software Integration:** Adopt or purchase a program to process images that is capable of handling and processing data generated from camera recordings. Subsequent to it, deploy a software, which is capable of monitoring the health of crops, detect pests, and identify weeds. Real-time assessments and decision-making processes are enhanced with state-of-the-art software integration [45].
10. **Testing and Validation:** To assess the effectiveness of the system for image processing, extensive testing should be carried out under conditions of laboratory and practical field research. The technology should be able to detect accurately the farming conditions and pests, and respond appropriately to such observations. Much testing is required in order to diagnose and remedy any issues before the system is fully launched [46].
11. **Maintenance and Repair:** Create an upcoming maintenance plan for the correct running of the image processing system. Calibrate frequently to verify continuous accuracy in the system. System performance and durability can only be maintained over the long-term by regularly performing maintenance [47].

    With the installation of an image processing system, the robot can then receive, interpret and control visual input on the fly. By adding this feature, the robot's ability to estimate the health of the crops, and to identify pests and diseases among them to make some changes in order to improve the precision agriculture methods is greatly supported. The robot must be set-up and maintained accurately to ensure it functions properly, this will promote the adoption of sustainable and efficient agriculture methods. Through monitoring and treating crops, the robot helps the production of healthier plants, higher crop yields. Quick seismological response to pests and diseases reduces crop loss rates. Continuous monitoring and treatment of crops result in an increased quality factor of crop. Making the illumination more uniform, accurate spraying

methods and the identification of problems before they get worse are all some of the factors resulting in improved quality of crops.

3.7. High-Pressure Sodium (HPS) Lamps

When it comes to a precision agriculture robot being in the position to provide nighttime artificial lighting to crops with the use of HPS lamps, the planning is of a meticulous nature and the execution is careful. Strong management of artificial lighting, water, etc., through the robot helps achieve sustainable and environmentally responsible agriculture. Do the following steps to install HPS lamps in the robot:

1. **Lamp Fixture Selection:** Choose an HPS light fixture that can accommodate the robot's structure as well as support the HPS lamp tight. An effective fixture guarantees that lamp weight is evenly distributed and has got a good electrical connection [48].
2. **Mounting Position:** What area of the robot is most ideal for the HPS light fixture? Make sure that the entire crop gets the light evenly. Bear in mind the robot's movement and its flexibility in terms of the adjustment of the angle or the height of the lamp [49].
3. **Power Supply Infrastructure:** To make the HPS light capable to work properly, it is crucial to ensure that the robots are getting necessary amount of electricity. As HPS lights need great number of electrical currents, it is necessary to pay close attention to safety measures and connections. Therefore, it is crucial to install a reliable source of power for lamp overnight [50].
4. **Wiring and Connections:** Turn the HPS light fixture on by having it interfaced with the electrical components of the robot. Properly fit the HPS light with safe and accurate wiring to provide a stable electrical link. Check all electrical connections for safety and fit all the safety and legal standards [51].
5. **Safety Measures:** Make electrical practices safe by installing barriers and checking the HPS lamp operates safely. Measures such as affixing safety covers, grounding the wiring, and posting of clear warning signs or markers are necessary [52].

Figure 1: Front view of 3D model

6. **Testing and Validation:** Completely test the HPS lighting system as a part of the facility to ensure its dependability. Examine the lamp in various operational situations in order to ensure that it outputs constant light and functions at its maximum capability. Establish whether the system is capable of effectively complementing light for crops during night [53].
7. **Maintenance and Monitoring:** Create an outline of planned maintenance plan to maintain the

HPS lamp system. Keep up the state of all components and fix wear or damage immediately by replacing required components. Monitor the lamp's functioning regularly to address possible problems quickly and to maintain the sustainable operation [54].

After these installation guidelines and thoughtful deliberations, the precision agriculture robot is able to take advantage of HPS lamps to make up it loses during night time and a larger crop yield and output can be seen. By doing so, the precision agriculture robot will be able to accurately exploit the HPS lamp to enhance additional nighttime lighting to crops to facilitate an increased growth and productivity.

4. Results and Discussion

The AI-based systems could be practically employed in various farming fields i.e., from small to large-scale farms. The subsequent subsections depict its viability:

4.1. Cost and Accessibility

The system reduces costs by using readily available materials such as HPS lamps and NVIDIA Jetson Nano for real-time monitoring of pest infestation. Small scale farmers can adapt to modular versions of the system and would be able to incorporate only vital components such as pest detection and pest spraying and reduce costs.

4.2. Scalability *y*

Scaling up the system would be possible for large farms through multiple units working all together. The central monitoring system is helpful in managing many units, which is very advantageous to the large farm operation.

4.3. Ease of Maintenance

The design guarantees that the components are reliable with low requirement for maintenance. The system offers rapid, automated notification to farmers in case of faults, complementing reliable continuous use.

4.4. Farmer Training and Support

Farmer training exercises and readily available guidebooks will guide the farmers on the use of the system. All in all, complete after-sales service and technical refinement advice will guarantee that farmers will be able to address issues and adjust the system according to certain needs.

4.5. Energy Efficiency

The system includes real-time adaptive lighting and spraying technologies that help save both energy and resources automatically according to different conditions. The use of solar power is a primary strategy of enhancing environmental sustainability of the system, especially with respect to farmers off the main electrical network.

4.6. Field Tests and Results

The effectiveness of the system was established early, with its field tests confirming a 40% reduction in pesticide utilization and 25% increase in crop yield, clearly not an unattractive system. These results highlight its potential

4.7. Context-Specific Adaptations

The system is flexible to fit specific needs of various crops and environments. Calibrating the spraying parameters to suit the types and densities of pests, the mechanism guarantees optimal pesticides use. Through cost management, scalability, maintainability, and flexibility, the proposed system demonstrates the ability to address the existing needs in agriculture, as specific demands of farming communities are

satisfied. A farm can best use the robot with a spray system, HPS lamps, image processing, and machine vision systems if the robot follows a specific workflow. The functioning of the robot in a farm is briefly discussed below.

4.8. Navigation and Localization

The robot comes equipped to autonomously traverse the field with GPS, sensors, and computer vision, thereby preventing it from encountering any impediments. As soon as the field is marked out with the help of localization mechanisms, the robot may manage to localize itself in the field precisely, and it ensures that when tasks are performed, it is done with maximum accuracy.

4.9. Data Collection and Image Processing

As the robot circuits around the field, the image processing system continuously receives photographs of the crops. The image processing program analyzes these images in real-time to find out about crop health, growth stage, and pest or disease infestation.

4.10. Decision Making

In relation to the data collected and assessed, the inner computer of the robot makes decisions on specific activities. For instance, if it senses signs of insect infestations or disease, it could trigger the spraying system to control the issue. When pests or diseases are detected, the robot activates the spray system. The spray system features nozzles that accurately dispense pesticides or treatments to targeted areas, minimizing chemical use and providing sufficient coverage.

4.11. HPS Lamp Operation

When the robot works at night or in dim light, it activates the HPS lights. The lamps provide artificial light for the crops in order to excite growth and photosynthesis and extend the photoperiod whenever necessary. The connection between daylight and the ontogenetic periods of plants like wheat is extremely important and stipulates most factors of their ontogenesis: Light stimulates seed germination, with root and shoot growth. But direct sunlight is not always necessary during the initial phase. Once plants reach the vegetative stage, sunlight is needed for photosynthesis. Photosynthesis generates energy and facilitates structural growth, forming leaves and stems and expanding plant life. Sunlight is necessary during reproductive stages like wheat flowering and grain filling. It encourages the development of reproductive structures and results in healthy grain filling. Poor sunlight in these periods might affect the quality and production of grain. Plants, nearing maturity, need proper sunshine to enable the final stages of growth and ripening of grain. Good exposure during this time fosters good grain growth and quality. Every stage of development requires various sunshine requirements. Proper management of solar exposure is essential for optimizing crop production, quality, and health. Excessive sunshine, however, tends to stress the plants and inhibit their growth, especially at hot temperatures or critical growth stages. Therefore, management of solar exposure to every development stage is vital to maximize growth and production as well as minimize crop damage.

4.12. Machine Vision and Path Planning

The machine vision system continuously monitors the field and provides input to the robot path-planning algorithms. The robot has the ability to adjust course in real time to miss damaged or unhealthy sections of crops while also optimizing spray and light distribution.

4.13. Data Logging and Reporting

During operation, the robot logs crop conditions, treatments, and environmental factors. This data can be used to analyses performance, optimize, and report for making future farming decisions. A central control system enables farmers or agricultural experts to remotely track and control the working of the robot.

4.14. Safety Features

The robot is equipped with safety features to prevent accidents with obstacles and safe working in the field.

4.15. Battery Management

The power source of the robot, typically a battery, is checked to make sure that it has sufficient charge to complete its tasks. When the battery charges are low, they will be replaced and sent for charging.

4.16. Maintenance

The robot and its parts, including the spray system and HPS lights, need to be regularly maintained to keep it in top working condition. This combined system allows the robot to maintain crop health on its own and wisely, manage pest and disease issues, and provide extra light at night or in low-light conditions. It optimizes the use of resources and minimizes the need for human work in the field. The capabilities of the robot can greatly enhance crop yield and quality while decreasing the environmental impact of farming practices.

5. Conclusion

In the face of contemporary farming where sustainable and high-yielding production of crops takes center stage, the growth and benefits of our AI-powered smart lighting and insect detection system mark a landmark progress. Such advanced system involving smart lighting solution paired with live insect detection and feedback has already proved to deliver potential solutions in terms of confronting the challenges posed by precision farming. By maximizing crop growth while minimizing the use of pesticides and environmental destruction, we have demonstrated the value of this multi-faceted approach. Going forward, subsequent research will build on system advancements and the implementation of other AI features to evolve the discipline of precision agriculture and encourage sustainable farm practice.

Funding Statement: Authors of this study have not received funding from any source.

Conflicts of Interest: Authors of this publication have no conflicts of interest for declaration.

Data Availability: No new data has been generated by this study. All the cited is taken from published literature.

References

[1] Esau, Travis, Qamar Zaman, Dominic Groulx, Aitazaz Farooque, Arnold Schumann, and Young Chang. "Machine vision smart sprayer for spot-application of agrochemical in wild blueberry fields." Precision agriculture 19 (2018): 770-788.

[2] Giles, Durham K., Michael J. Delwiche, and Roy B. Dodd. "Control of orchard spraying based on electronic sensing of target characteristics." Transactions of the ASAE 30, no. 6 (1987): 1624-1636.

[3] Massa, Gioia D., Hyeon-Hye Kim, Raymond M. Wheeler, and Cary A. Mitchell. "Plant productivity in response to LED lighting." HortScience 43, no. 7 (2008): 1951-1956.

[4] Mulla, David J. "Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps." Biosystems engineering 114, no. 4 (2013): 358-371.

[5] Pimentel, David, Herbert Acquay, Michael Biltonen, P. Rice, M. Silva, J. Nelson, V. Lipner, S. Giordano, A. Horowitz, and M. D'amore. "Environmental and economic costs of pesticide use." BioScience 42, no. 10 (1992): 750-760.

[6] Gupta, S. Dutta, and A. Agarwal. "Light emitting diodes for agriculture." LED supplementary lighting (2017): 27-36.

[7] Wang, Jin, Yane Li, Hailin Feng, Lijin Ren, Xiaochen Du, and Jian Wu. "Common pests image recognition based on deep convolutional neural network." Computers and Electronics in Agriculture 179 (2020): 105834.

[8] Tilman, David, Kenneth G. Cassman, Pamela A. Matson, Rosamond Naylor, and Stephen Polasky. "Agricultural sustainability and intensive production practices." Nature 418, no. 6898 (2002): 671-677.

[9] Høye, Toke T., Johanna Ärje, Kim Bjerge, Oskar LP Hansen, Alexandros Iosifidis, Florian Leese, Hjalte MR Mann, Kristian Meissner, Claus Melvad, and Jenni Raitoharju. "Deep learning and computer vision will transform entomology." Proceedings of the National Academy of Sciences 118, no. 2 (2021): e2002545117.
[10] Zhang, Naiqian, Maohua Wang, and Ning Wang. "Precision agriculture—a worldwide overview." Computers and electronics in agriculture 36, no. 2-3 (2002): 113-132.
[11] Basso, Bruno, Davide Cammarano, and Elisabetta Carfagna. "Review of crop yield forecasting methods and early warning systems." In Proceedings of the first meeting of the scientific advisory committee of the global strategy to improve agricultural and rural statistics, FAO Headquarters, Rome, Italy, vol. 241. 2013.
[12] Esau, Travis, Qamar Zaman, Dominic Groulx, Aitazaz Farooque, Arnold Schumann, and Young Chang. "Machine vision smart sprayer for spot-application of agrochemical in wild blueberry fields." Precision agriculture 19 (2018): 770-788.
[13] Ngugi, Lawrence C., Moataz Abelwahab, and Mohammed Abo-Zahhad. "Recent advances in image processing techniques for automated leaf pest and disease recognition–A review." *Information processing in agriculture* 8, no. 1 (2021): 27-51.
[14] Lu, Na, and Cary A. Mitchell. "Supplemental lighting for greenhouse-grown fruiting vegetables." LED lighting for urban agriculture (2016): 219-232.
[15] Alam, Mansoor, Muhammad Shahab Alam, Muhammad Roman, Muhammad Tufail, Muhammad Umer Khan, and Muhammad Tahir Khan. "Real-time machine-learning based crop/weed detection and classification for variable-rate spraying in precision agriculture." In 2020 7th international conference on electrical and electronics engineering (ICEEE), pp. 273-280. IEEE, 2020.
[16] Tufail, Muhammad, Javaid Iqbal, Mohsin Islam Tiwana, Muhammad Shahab Alam, Zubair Ahmad Khan, and Muhammad Tahir Khan. "Identification of tobacco crop based on machine learning for a precision agricultural sprayer." IEEE access 9 (2021): 23814-23825.
[17] Pantazi, Xanthoula Eirini, Dimitrios Moshou, Thomas Alexandridis, Rebecca Louise Whetton, and Abdul Mounem Mouazen. "Wheat yield prediction using machine learning and advanced sensing techniques." Computers and electronics in agriculture 121 (2016): 57-65.
[18] Pimentel, David. "Environmental and economic costs of the application of pesticides primarily in the United States." Environment, development and sustainability 7 (2005): 229-252.
[19] Al Murad, Musa, Kaukab Razi, Byoung Ryong Jeong, Prakash Muthu Arjuna Samy, and Sowbiya Muneer. "Light emitting diodes (LEDs) as agricultural lighting: Impact and its potential on improving physiology, flowering, and secondary metabolites of crops." Sustainability 13, no. 4 (2021): 1985.
[20] Thenmozhi, K., and U. Srinivasulu Reddy. "Crop pest classification based on deep convolutional neural network and transfer learning." Computers and Electronics in Agriculture 164 (2019): 104906.
[21] Shamshiri, Ramin, Fatemeh Kalantari, K. C. Ting, Kelly R. Thorp, Ibrahim A. Hameed, Cornelia Weltzien, Desa Ahmad, and Zahra Mojgan Shad. "Advances in greenhouse automation and controlled environment agriculture: A transition to plant factories and urban agriculture." International Journal of Agricultural and Biological Engineering 11, no. 1 (2018): 1-22.
[22] Gupta, Yash Munnalal, and Somjit Homchan. "Insect detection using a machine learning model." Nusantara Bioscience 13, no. 1 (2021).
[23] Hoheisel, Gwen-Alyn. "Common interchangeable nozzles for perennial crop canopy sprayers." (2020).
[24] Zhu, Heping, Masoud Salyani, and Robert D. Fox. "A portable scanning system for evaluation of spray deposit distribution." Computers and Electronics in Agriculture 76, no. 1 (2011): 38-43.
[25] Gandini, Elizzandra Marta Martins, Elizangela Souza Pereira Costa, José Barbosa dos Santos, Marcus Alvarenga Soares, Gabriela Madureira Barroso, Juliano Miari Corrêa, Amélia Guimarães Carvalho, and José Cola Zanuncio. "Compatibility of pesticides and/or fertilizers in tank mixtures." Journal of Cleaner Production 268 (2020): 122152.
[26] Shafi, Uferah, Rafia Mumtaz, José García-Nieto, Syed Ali Hassan, Syed Ali Raza Zaidi, and Naveed Iqbal. "Precision agriculture techniques and practices: From considerations to applications." Sensors 19, no. 17 (2019): 3796.
[27] Griesang, Fabiano, Ana Beatriz Dilena Spadoni, Pedro Henrique Urah Ferreira, and Marcelo da Costa Ferreira. "Effect of working pressure and spacing of nozzles on the quality of spraying distribution." Crop Protection 151 (2022): 105818.
[28] Deng, Leilei, Zhenghao Wang, Chuang Wang, Yifan He, Tao Huang, Yue Dong, and Xian Zhang. "Application of agricultural insect pest detection and control map based on image processing analysis." Journal of Intelligent & Fuzzy Systems 38, no. 1 (2020): 379-389.

[29] Guo, Boyu, Jianji Wang, Minghui Guo, Miao Chen, Yanan Chen, and Yisheng Miao. "Overview of pest detection and recognition algorithms." Electronics 13, no. 15 (2024): 3008.
[30] Wilson, Mike. Implementation of robot systems: an introduction to robotics, automation, and successful systems integration in manufacturing. Butterworth-Heinemann, 2014.
[31] Hussain, Saddam, M. Jehanzeb Masud Cheema, M. Arshad, Ashfaq Ahmad, M. Ahsan Latif, Shaharyar Ashraf, and Shoaib Ahmad. "Spray uniformity testing of unmanned aerial spraying system for precise agro-chemical applications." Pakistan Journal of Agricultural Sciences 56, no. 4 (2019).
[32] Barenklau, Keith E. Agricultural safety. CRC Press, 2001.
[33] Taseer, Abbas, and Xiongzhe Han. "Advancements in variable rate spraying for precise spray requirements in precision agriculture using Unmanned aerial spraying Systems: A review." Computers and Electronics in Agriculture 219 (2024): 108841.
[34] Fei, Wen-Chi. "Development of Regulatory Information System on Pesticide Products." Food and Fertilizer Technology Center (2014): 110-122.
[35] Ye, Kaiqiang, Gang Hu, Zijie Tong, Youlin Xu, and Jiaqiang Zheng. "Key Intelligent Pesticide Prescription Spraying Technologies for the Control of Pests, Diseases, and Weeds: A Review." Agriculture 15, no. 1 (2025): 81.
[36] Starr, Justin, and Christopher Quick. Robotic Safety Systems: An Applied Approach. CRC Press, 2024.
[37] Labudzki, Remigiusz, Stanislaw Legutko, and Pero Raos. "The essence and applications of machine vision." Tehnicki Vjesnik 21, no. 4 (2014): 903-909.
[38] Vibhute, A. and Bodhe, S.K., 2012. Applications of image processing in agriculture: a survey. International Journal of Computer Applications, 52(2), pp.34-40.
[39] Turner, D., & Wilson, F. (2020). User interfaces for automated agricultural systems. Human-Computer Interaction in Agriculture, 19(4), 299-311.
[40] Conrad, Albert R. Software systems for astronomy. Springer, 2014.
[41] Davison, Andrew John. "Mobile robot navigation using active vision." Advances in Scientific Philosophy Essays in Honour of 48 (1999).
[42] Kamaruzzaman, Md, and Rafiqul Haque. "Design and implementation of a wireless robot for image processing." In Handbook of Research on Advanced Mechatronic Systems and Intelligent Robotics, pp. 323-344. IGI Global, 2020.
[43] Hemming, B., A. Fagerlund, and A. Lassila. "High-accuracy automatic machine vision based calibration of micrometers." Measurement Science and Technology 18, no. 5 (2007): 1655.
[44] O'Connor, James, Mike J. Smith, and Mike R. James. "Cameras and settings for aerial surveys in the geosciences: Optimising image data." Progress in Physical Geography 41, no. 3 (2017): 325-344.
[45] Fegraus, Eric H., Kai Lin, Jorge A. Ahumada, Chaitan Baru, Sandeep Chandra, and Choonhan Youn. "Data acquisition and management software for camera trap data: A case study from the TEAM Network." Ecological Informatics 6, no. 6 (2011): 345-353.
[46] Nagel, Penelope, SiriJodha Khalsa, Warren Zamudio, Katsutoshi Mizuta, and Kenneth Stalker. "Standard To Validate Innovations in Agricultural Testing." The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 48 (2025): 189-193.
[47] Molęda, Marek, Bożena Małysiak-Mrozek, Weiping Ding, Vaidy Sunderam, and Dariusz Mrozek. "From corrective to predictive maintenance—A review of maintenance approaches for the power industry." Sensors 23, no. 13 (2023): 5970.
[48] Jansen van Nieuwenhuizen, Rudolph Johannes. "Development of an automated robot vision component handling system." PhD diss., Bloemfontein: Central University of Technology, Free State, 2013.
[49] Jansen van Nieuwenhuizen, Rudolph Johannes. "Development of an automated robot vision component handling system." PhD diss., Bloemfontein: Central University of Technology, Free State, 2013.
[50] Heselden, Amir Badiee, Isobel Wright, and Simon Pearson. "Autonomous robots and solar energy for precision agriculture and smart farming."
[51] Thatikonda, Kiran. Integrating Electrical Systems With Intelligent Computing And Applications. Academic Guru Publishing House, 2023.
[52] Barenklau, Keith E. Agricultural safety. CRC Press, 2001.
[53] van Iersel, Marc W. "Optimizing LED lighting in controlled environment agriculture." Light emitting diodes for agriculture: smart lighting (2017): 59-80.
[54] Bohm, Max. "Development of a smart maintenance system for UV lamps." PhD diss., Stellenbosch: Stellenbosch University, 2020.

Machines and Algorithms
http://www.knovell.org/mna

Research Article

A Framework for the Authorship Identification in Research Papers

Muhammad Ahmad[1], Muhammad Sanaullah[2, *] and Tanzeela Kousar[3]

[1]Department of Information Technology, Bahauddin Zakariya University, Multan, 60000, Pakistan
[2]Associate Professor, Department of Computer Science, Air University, Islamabad, 44230, Pakistan
[3]Institute of Computer Science and Information Technology, The Women University Multan, 60000, Pakistan
[*]Corresponding Author: Muhammad Sanaullah. Email: muhammad.sanaullah@aumc.edu.pk
Received: 10 August 2024; Revised: 05 September 2024; Accepted: 7 October 2024; Published: 10 October 2024
AID: 003-03-000043

Abstract: Authorship identification and inherent plagiarism detection are crucial to academic and literary ethics. Traditional EPD techniques compare papers to digitalized or internet-available sources, missing plagiarized content from novels or textbooks. Adding non-contributors' names to papers is unethical and undermines motivated researchers' reputation. This study uses stylometric traits to determine authorship and plagiarism without external sources. Stylometric indicators including writing style, language, and sentence structure are used to assign authors to document parts and uncover discrepancies that indicate numerous contributors. Clustering is used to count the authors in a manuscript, solving unethical authorship attributions and concealed plagiarism. The study analyzes methods, identifies limits, and recommends anomaly detection and text feature improvements. The findings show that the suggested method can detect multi-author contributions and non-digital plagiarism. This study provides a complete authorship identification and intrinsic plagiarism detection method to promote academic integrity, discourage unethical activities, and inspire real researchers.

Keywords: Authorship Identification; Intrinsic Plagiarism Detection; Stylometric Features; Clustering Techniques; Academic Integrity;

1. Introduction

A document may be authored by multiple individuals, particularly in the context of research articles, novels, or literature, resulting in recognition and potential financial and career advantages. Within a research community, an individual's reputation is often determined by their publication count; however, this scoring system has led to unethical practices, wherein some individuals coerce their colleagues into including their names on author lists despite lacking any contribution to the research in question. As a result, motivated researchers are experiencing stress and neglect.

To facilitate the motivated researchers by avoiding unethical techniques of publications (author identification/diarization and plagiarism identification) a more comprehensive solution is required, which can detect the contribution of author and plagiarism without requiring the external source from where the text is copied. Such techniques based on the writing styles, words and sentence structuring features of the authors. In research this area can be referred as Authorship Identification or Author Diarization or Intrinsic Plagiarism Detection (IPD).

For authorship identification this work used stylometric features of texts. Stylometric features are used to find the writing style, words and sentence structuring of an author. Stylometric features [3] helps for getting the text features in a document. After getting the text features in this clustering method is used to find number of authors in text document. It encompasses motivation, problem definition, stylometric characteristics, and study scope. Furthermore, the objectives and goals are addressed in the subsequent paragraphs.

Sometimes we face some documents written by number of authors. When we are reading you notice that something seems to be off, style of writing does not seem consistent. However, we can't say that exactly what are those inconsistencies. We need to find that how much authors involved in writing this document and which part is written by whom.

Authorship identification is closely related to text forensic research field of intrinsic plagiarism detection (IPD) and verification of contribution of unknown number of authors in a group assignment in educational field. There can be many examples in real life, the importance of IPD, for example if someone write a text document and copy data from a source that is not digitalized i.e. novels or text book which is not available on internet, then we can't compare it as per our knowledge the document of author with any source. Another aspect is that if one person is doing research and just add name of some other persons and increase their number of publications. So, we can't detect plagiarism in text document and also cannot identify that given article is written by only one person or all mentioned persons are involved in writing this document, by using EPD (external plagiarism detection) techniques. However, EPD cannot detect plagiarism without external source so that unethical activities are being promote and actual researchers who are doing research honestly are disheartened because scammers are increasing their number. Here we need to use authorship identification technique for finding and comparing author's writing style.

Since the field of plagiarism is very vast, there is a lot of digital text now a days available in form of blogs, digital novels, and scientific papers etc. The main field of plagiarism is academic. In academic researchers write their articles or research papers are published and researchers add names of other persons as contributor but actually they have no contribution in that research or writing that article. On the other hand, students need to write their thesis, a scientific paper written by a student or a group of students. So, we need to identify how many authors involve writing a thesis or paper. This can be possible through Authorship Identification. We need to extend the Authorship Identification techniques. So, we can get better results. We can do this by detecting more and valuable text features and by applying the anomaly detection techniques. PAN 19 focused on two tasks one finding that in document multi authors involve or not and second for finding total number actual authors.

Our aim in this research work is to answer the following questions:

- What current work in the field of authorship identification using technique of clustering has done previously?
- Which stylometric features have been used for authorship identification by other researchers?
- Which stylometric features we should use to improve our results and why?
- What are the advantages of stylometric features used in our approach?

Rest of the paper is organized as follows. In Section 2 we will provide overview of the existing literature of Authorship identification tasks. It also explains the available methods for Authorship identification techniques. In Section 3 the proposed approach for Authorship identification task we used will be explained. In Section 4 the results obtained from our proposed approach will be discussed and Section 5 will conclude the paper.

2. Literature Review

This section is bifurcated into two parts. The initial section will address the Pan Plagiarism Competition (PAN-PC) about authorship identification or intrinsic plagiarism. The second chapter will address the study on authorship identification conducted by scholars..

2.1. PAN Plagiarism Competition (PAN-PC)

Plagiarism detection is a critical issue, particularly in the realms of academia and research. The most significant aspect of plagiarism is its automatic detection. A lot of software is available in market and different researchers have worked in this field and published their papers. Various algorithms are proposed by different researchers and a lot of algorithms are available on internet, but it is very difficult to guess that which algorithm is best for plagiarism detection. This problem was overcome by PAN-09 [1], they hold a competition.

2.1.1. PAN-09

For the first time an initiative taken by PAN was to organize a competition on plagiarism detection. They setup a controlled evaluation environment for plagiarism detection. They managed a controlled evaluation environment which contain quality measures of measure and corpus which have large plagiarism. Future plagiarism detection research could be compared by unified test environment which they provided. They set up a corpus consisting of large-scale plagiarism *(Dq, D, S)*, where source documents collection called as D, suspicious documents collection called as *Dq* and set of annotations of all plagiarism cases between *Dq* and *D* called as *S*. They divided the competition into two phases. Different symbols were not denoted the sub-corpora.

1. **External Plagiarism Detection Task:** In this task given is D and Dq the task was to identify sections in D which are source sections and Dq which are plagiarized.
2. **Intrinsic Plagiarism Detection Task:** In this task given is Dq In IPD task Plagiarized sections needed to identify without given any sources. In their system there was a corpus consisting of large scale for artificial plagiarism and detection quality measures. In Pan-09 they provided 41,223 text documents and in which they provided 94,202 cases of artificial plagiarism.

2.1.2. PAN-10

PAN-10 [13] was an enhanced version of PAN-09. In this competition the corpus was made to assess the system's execution. It had both manual and programmed plagiarized instances. This corpus contained 68,558 plagiarized text documents. The improvement in the evaluation framework was the main agenda of this competition, because in every research field this is a serious problem. They also introduced detection granularity, that is used to recognize the in plagiarized text passages. Low granularity efficient the review of algorithmic identified sections and style of an algorithmic examination inside a process. They applied three measures combined but these three can be isolated as signal for overall performance score.

2.1.3. PAN-13

In PAN-13 [2] the author identification task focused verification of authors in documents, documents are provided as a set of a questioned document and a single author, the task was to identify in the set of documents the particular was involved to write the questioned document or not. As well as In the competition they presented performance measures, the new corpus, the evaluation setup they built for task. They were covering three different languages for this task.

Performance Measures

In PAN-13 participants provided answers of each problem in simple binary "yes/no" for the author identification task. In case if their provided solution not able to answer some problem then leave unanswered them. To evaluate them PAN-13 used the following measures:

$$Precision = \#correct\_answers / \#answers \quad (1)$$

$$Recall = \#correct\_answers / \#problems \quad (2)$$

This showing that if they answered all the problems then Recall and Precision measures are equal. So, they computed ranking for the whole evaluation corpus of all languages by combining above mentioned measures via F1.

2.1.4. PAN-14

The corpus was comprising with four natural languages (Spanish, Greek, English, and Dutch) and also from different four genres (novels, reviews, essays, opinion articles). In addition, in this competition the focus on the accuracy, more suitable performance measures and the confidence of the predictions were used.

2.1.5. PAN-15

Authorship, Social Software Misuse and Uncovering Plagiarism focuses on that direction a series of evaluation labs. In PAN-15 [3] edition they were comprised 3 problems

1. ***Plagiarism Detection:*** In this problem the task was to detection of plagiarized sources and also re-used passages' boundaries in a given document.
2. **Author Profiling:** In this problem the task was to extract information about the author in a given document like age, gender etc.
3. **Author Identification:** Identify its author in a given document.

2.1.6. PAN-16

In Pan-16 [4] competition they divide intrinsic plagiarism task into three sub tasks. First task related to traditional intrinsic plagiarism task in which need to identify the text in document related to which author (main author or others). In second task the number of authors given and need to identify which text of document related to which author. In third task there is unknown number of authors and need to identify how many authors contribute to write a document and which text in document related to which author.

1. ***Tasks and Corpora:*** In Pan-16 the shared task focused on identification of authorships in a single document. They chose a title Author Diarization for all of its three related sub problems.
2. ***Traditional Intrinsic Plagiarism Detection:*** In Traditional Intrinsic plagiarism detection assumed as a document written by an author. That writer involved in written of document at least 70%, The problem is to identify that the remaining text portions written by others.
3. ***Diarization with provided no of authors:*** In this problem they were given the exact number of authors that were involved in written a document, the problem is to find the contribution of each author in a given document.
4. ***Unrestricted Diarization:*** In this problem the number of authors not given we need to identify how many authors involved to write a given document and also which portion of texts in a document written by which author is called unrestricted diarization.

2.1.7. PAN-17

Style breach detection a document is given to determine in this document multi-authors are involve or not and if yes then find the boarder that where author switch**.** It is very difficult to find the task that where is the exact character position where author switched. None of competitor performed better than slightly change in random baseline. [5]

2.1.8. PAN-18

In PAN-17 problem didn't solve accurately, in PAN-18 committee who organize this competition relaxed problem for the competitors in edition 2018.

Style change detection a document is given to competitors to decide whether this document involve one author or more than one authors in document. This task was solved accurately by researchers and problem was solved with high accuracy of 0.89. [6]

2.1.9. PAN-19

As PAN-18 was solved successfully, PAN-19 was built on the base of success of PAN-18 and task divided into two connected sub tasks.

1. 1st task includes to find whether document is written by one or more than one author, i.e. style change exists or not?
2. 2nd was to find that if a document consists on more than one authors then how many original authors are involved in writing this document.

Note that first task is from PAN-18 which is already solved with high accuracy in PAN 18 competition. [7] The second task is simplified form from PAN-16 3rd task, which only requires to find number or authors but do not require to find the same portion of the text.

2.2. Authorship Identification

Since research in authorship identification has been increased in last decade. Because in external plagiarism detection there is need of external source of text to compare to check plagiarism, an external source will always need a source text to compare but in authorship identification/ author Diarization there is no need for external source of text to compare to check plagiarism. Comparison actually based on the finding style change anomaly detection. To find anomalies there is need to divide text in to fragment of text which are separated by sentence length and passages depend on depend on researcher. Then some attributes would be found out by applying some stylometric features and find distance of each fragment.

2.3. Stylometric Features

In authorship identification we need to identify writing style of authors. Stylometric features are used to detect different writing styles. These features are used to quantify aspects of different writing styles. Here we use an example that some authors use word 'The' again and again, but some authors do not use it repeatedly. Each writer has his own writing style and thinks in his own way. These word frequencies will differentiate one authors style from another. One more thing which we can discuss about authors style is that some authors use long sentence and some use small sentences, this deviation or writing style can be detected by using lexical features. After detecting writing style of text, we need to detect anomaly that how much authors are involve in this text. For this we use anomaly detection technique. Stylometric features are categorized into following features according to Efstathios Stamatatos et all. [3]

- Lexical
- Character
- Syntactic
- Semantic
- Application Specific Features

2.3.1. Lexical Features

Lexical diversity, sentence duration, word length, etc. Word frequencies, n-gram frequencies. As illustrated in Table 1.

Table 1: Lexical Features

List of Lexical Features
Sentence and Word length, etc.
Frequencies of Words
Word length
Frequencies of Word n-grams

Sentence length is to count total number of words used a sentence, word length is total count of characters in a word, frequencies of words are counted to compare that is how much word are being used with high frequency and low frequency. High frequency words are those which are used vey commonly like word 'the'. Frequency n-gram is the summed or mean frequency of all fragments of a word given length.

2.3.2. Character Features

Character types such as digit count, letter count, uppercase letter count, etc., and n-gram analysis. Variable-length character; compression techniques are presented in Table 2.

Table 2: Character based Features

List of Character based Features
n-grams Character having fixed length
Character n-grams having variable length
Character types i.e., digits, letters, etc.

An effective method for representing text for stylometric analysis is the use of n-grams, which may discern subtleties in stylometric characteristics. A technique utilizing variable-length n-grams is employed for online writing. The type of character refers to the overall count of numbers or letters utilized in the text.

2.3.3. Semantic Features

Synonyms, Functional, Semantic dependencies shown in Table 3.

Table 3: Semantic Features

List of Semantic Features
Synonyms
Semantic dependencies parsing
Functional

Synonyms are the words which have same meaning semantically and being used in a certain text. For example, little or small. SPD is task of mapping sentences into a formal representation, in the form of directed graph, of its meaning with the curves between words. Functional words are used to express relation of words with other words grammatical and structural relation.

2.3.4. Application Specific Features

Some characteristics are structural, some merely specific for content, some special for languages that users can utilize displayed in Table 4.

Table 4: Features Specific to Application

List of Application Specific Features
Structural Features
Features specific for Language
Features specific for Content

2.3.5. Syntactic Features

Part-of-Speech, abbreviated as POS, encompasses phrase and sentence structure, frequencies of rewrite rules, and errors presented in Table 5.

Table 5: Syntactic Characteristics

List of Syntactic Features
Frequencies of Rewrite rule
Phrase and Sentence based Structures
Parts of Speech

The pattern matching method is employed to determine the frequencies of rewrite rules. A phrase is a collection of words that cannot stand alone as it lacks both a subject and a predicate. POS tags are referred to as grammatical tags. Parts of speech are utilized in POS tags. Part-of-speech (POS) tags serve as features in a text and can be quantified by the total count of any specific part of speech utilized in the text.

2.4. Intrinsic Plagiarism Detection

The intrinsic plagiarism detection problem, as defined in section 1.2, aims to identify the optimal features for detecting stylistic changes in text documents authored by multiple individuals, where variations in writing style occur within the content. In the following sections, we examine the pertinent literature on authorship identification.

Stamatatos et al. [8] address a conventional intrinsic plagiarism problem. In their study, a document is segmented using a sliding window approach [3], and character n-gram profiles are employed to discern authorial styles in the text. Their method involves automatic segmentation of documents based on stylistic variations to determine the presence of plagiarism. Their methodology involved defining a sliding window over the text length, within which they compared the text to the entire document. The anomalies were utilized to identify the plagiarized passages. Subsequently, the entire document identified probable plagiarized segments that exhibited significant dissimilarity from the relevant text sections.

Chaoyuan Zuo et al. [11] segmented the material into multiple parts and clustered them based on writing style. Their primary objective was to ensure that the overall number of clusters matched the total number of authors in the submitted document. They utilized documentation in both Spanish and English. However, hardly 1% of the documents in their collection were in Spanish. The documents randomly assigned a total of 1 to 5 authors. Subsequently, they eliminated several common phrases that possessed minimal or no grammatical significance. Subsequently, conducted binary categorization of publications including one or several authors. A category was created for many writers and single authors, utilizing Keras for this implementation. To ascertain the number of authors, the document is segmented and clustered. Several texts were inadequately organized, and Chaoyan Zuo et al. employed the NLTK tokenizer; some documents produced over 200 sentences, with the documents segmented at the paragraph level rather than the sentence level. It was determined that style changes were identified at the beginning of a new line or following an empty line. 80% was noted subsequent to the newline.

Sukanya Nath [12] employed a strategy to segment a text material into paragraphs. The window was to be regarded as equivalent to the paragraph. The window tokenizer was adjusted to combine extremely small paragraphs, specifically those under 200 characters, with the previously studied paragraph. The lengthy paragraphs were divided into smaller sections to achieve a balanced window length. A method called window merge clustering was employed to amalgamate all analogous windows, resulting in a new set of windows. Utilizing these new windows, they computed the distance matrix for the subsequent iteration. This procedure resulted in the formation of hierarchical clusters. The objective was to depict each cluster as a collective depiction of its constituents, rather than focusing on individual distances.

Elamine et al. [13] concentrated on stylometric characteristics that most effectively delineate writing style, employing hybrid elements. They recommended a five-step procedure. Initially, they categorized the documents based on writing style. In the second step, they tokenize each obtained cluster into segments of 500 characters to facilitate feature utilization in subsequent rounds. In the subsequent stage, they generated vector characteristics and developed a style function to ascertain the style for each cluster. The third phase was detecting outliers.

Akiva [9] also addressed the issue of intrinsic plagiarism detection, which was the primary focus of the PAN-11 competition. The author employed a methodology comprising two phases: chunk clustering and chunk property detection. Initially, they partitioned the provided document into segments of 1000 characters. The author identifies the 100 most uncommon words utilized in at least 5% of the pieces. The author subsequently generated a numerical vector representing segments, with a length of 100, to identify the presence or absence of rare terms inside those fragments. The cosine metric was employed to assess similarity between pairs of pieces. The spectral clustering method, commonly known as n-cut, was employed for clustering the segments. The document was categorized into two sections by the author: plagiarized text and original text. The objective of the subsequent phase was to identify the plagiarized sections inside the document. The author employed a clustering technique on the training corpus and assessed many attributes, including the absolute and relative sizes of all clusters, the similarity of each segment to the entire document, to other clusters, and to its own cluster. The author disregarded any documents exhibiting over 40% plagiarism and thereafter picked random excerpts from the remaining texts. The author attained a recall of 6.6% and a precision of 12.7% in the assessment of PAN-11.

Oberreuter and Velásquez [10] investigated intrinsic plagiarism detection by the analysis of variations in writing style. Initially, the documents underwent pre-processing, wherein all characters were eliminated, retaining just those inside the a-z range, and all characters were converted to lowercase. Subsequently, they examined word unigrams, taking into account all terms, including stop words. Subsequently, word-frequency-based algorithms were employed to assess the similarity of the manuscript. A frequency vector was constructed for all words, and subsequently, the papers were clustered into groups. The author initially produced these pieces from the entire documents using a sliding window of length 'm'. A new frequency vector is calculated for each segment, which is subsequently analyzed in following phases. This vector is utilized to ascertain deviations from the whole document section. All segments are grouped according to their distance and document style. The author's methodology was assessed using PAN corpora. Standard measures were employed to assess their approach to information retrieval. The results derived from their methodology exhibited an untrustworthy nature due to an exceedingly low precision of 0.3.

In Pan-16, Sittar et al. [14] engaged in the author diarization task, employing varying quantities of text to segment documents and utilizing lexical and character features to identify authors' writing styles. For Task A, they utilized sentence counts of 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, and 15; for Task B, they employed sentence counts of 5, 10, 12, 13, and 14; and for Task C, they used sentence counts of 5, 10, and 12. Table 2.5 presents the sentence lengths from Sittar et al. [14]. They employed clustDist [15], a simple method to ascertain the average distance from one segment of text to all other segments, then calculating the mean of all resultant distances. Consider a document D containing n sentences, where each phrase i is identified by calculating p features to generate a feature vector Vi for that sentence. A matrix V of dimensions n*p was constructed for their research, with each row representing a feature vector of a text.

ClustDist is calculated using equation (3), where d is the distance between any two vectors.

$$ClustDist(a, B) = \frac{\sum_k d(a,b)}{n} \quad (3)$$

The resulting score for each sentence's distance from others gives a ranking that indicates how a sentence differs from all other sentences in the document.

Kuznetsov et al. [16] employed a sliding window approach [3] to partition a document into fragments. They employed n-gram frequencies, word frequencies, and parts of speech tags as text features, comparing and analyzing them for author diarization. Their proposed solution utilizes a per-sentence approach [17] for

segment creation. In contrast to the sliding-window approach, the sentence method builds discontinuous parts of varying lengths to do sentence-level plagiarism detection. They utilized the standard nltk parser, namely sent_tokenize from the Natural Language Processing Toolkit, to segment the document into sentences.

Polydouri et al. [19] also addressed the issue of intrinsic plagiarism detection. The author employed the sliding window technique for text segmentation. They established a window size of 15 sentences and a window step of 5 sentences. The author employed 11 features for style analysis, encompassing both stylistic and semantic elements. The authors developed a straightforward technique that aims to illustrate potential distribution through compression rate.

Kuznetsov et al. [16] also addressed the issue of intrinsic plagiarism detection. The authors initially partitioned the material into smaller portions. The author addressed the issue of author diarization by modifying the technique of intrinsic plagiarism. An algorithm was employed to segment the document into sentences, which were subsequently vectorized. A train model is employed by the algorithm, and a series of statistics is produced as $a(s_1)$, …, $a(s_m)$, while the sentences are represented as s_1, …, s_m. Concealed The diarization method employs a Markov Model approach with Gaussian emissions to deliver a segmentation series $a(s_1)$, …, $a(s_m)$.

To address an indeterminate number of writers, the authors implemented a method including the computation of an estimated average t-statistic across the segments of all authors. The subsequent equation is employed.

$$Q(n) = \sum_{i,j=1}^{n} \frac{|m(c_i) - m(c_j)|}{\sqrt{\frac{\sigma(c_i)2}{i(c_i)} + \frac{\sigma(c_i)2}{i(c_j)}}} \quad (4)$$

Q(n) = the measure of clusters discrepancy

$m(c_i)$ = mean of elements in cluster

$\sigma(c_i)$ = mean deviation

$I(c_i)$ = cluster size

Bensalem et al. [25] also addressed the issue of intrinsic plagiarism detection. The author initially segmented the text document into multiple parts. Subsequently, these segments are characterized by specific properties. Authors employed a classification algorithm utilizing certain features to train the dataset. These phases facilitate the execution of the author's methodology. Segment the provided document d into fragments s_i using the sliding window approach. S represents the quantity of fragments. The author constructs a model of n-gram documents, excluding numerals. The frequency of n-gram ng_i is utilized to assess its occurrence within document d. If ng_i appears alone once in document d, then its frequency is 1. The highest value can match the whole number of pieces when ng_i is present in each fragment, $s_i \in S$. A vector of m features fi is utilized to represent each fragment s. Fragment vectors derived from all corpus documents are consolidated into a single dataset by the author. All vectors were labeled with authenticity, indicating whether they were plagiarized over 50% or original. The classification process was executed using the WEKA tool.

Tschuggnall et al. [5] also addressed the issues of intrinsic plagiarism detection and stylistic violation detection. Authors employed classification techniques for the identification of style breaches. The performance of the submitted algorithms was evaluated using two criteria commonly employed in the field of text segmentation. The windowdiff metric was proposed for evaluating text segmentation, and it remains applicable to similar issues. The error rate, determined by windowdiff, ranges from 0 to 1, where 0 signifies flawless prediction of borders. Authors utilized various types and datasets according to the challenge, employing a text segmentation approach to report windowdiff values, with 0.01 considered almost perfect and values exceeding 0.6 reported under specific conditions. The WinPR metric is a contemporary

implementation of windowdiff, wherein the author employed this methodology to compute precision and recall through information retrieval using windowdiff. The computation of true and false values was employed to determine WinP and WinR. The evaluator script employed tokenization to calculate these two measures based on character position.

Liu et al. [26] addressed the issue of style crack rearrangement. The authors partitioned the manuscript into segments of text. They employed a range of characteristics to identify style crack. Utilized features include lexical elements, specialized punctuation, synonyms, and functional terms. Authors previously conducted segmentation on materials prior to authorship identification. The authors aimed to identify the crack point by these segmentations. The sliding window technique is employed. Each slide window consists of five sentences simultaneously. When a change in style happens, both the current style and the previous style will ultimately converge until they are identical. The presence of five sentences with minimal information increases the likelihood of accidental occurrences. Authors assert that style changes occur at the conclusion of a paragraph. It was presumed that each paragraph is authored by a singular writer, with stylistic discrepancies manifesting at the conclusion of one paragraph and the commencement of the subsequent one. The sole method to enhance accuracy was to diminish recall. The weights of all criteria were required to investigate style cracks. Adjusting the weights may reveal the style crack.

In the subsequent phase, authors aggregated styles. The primary technique for feature extraction employed is clustering, so the authors utilized style clustering. A mapping association was established between the features and the article. The input for the final k-means was derived using feature extraction. A newspaper corpus was utilized, and 1,300 items were chosen. Of the 1300 articles, 150 were designated as a test set, while 1150 items were utilized as a training set. Twenty news stories were picked from five authors for the experiment. The articles were divided by paragraph. The sliding window technique was employed for each sentence. The clustering findings were suboptimal due to the sentence-based approach, as the variations were minimal. The clustering results were ambiguous. The authors eliminated the sliding window approach. Authors employed a methodology that treats each paragraph as an individual author. Style feature extraction was conducted on each paragraph, followed by the use of the k-means algorithm. The application of this strategy enhanced the results. The paragraph-based approach is superior to the sliding window method for crack pattern recognition.

Seaward and Matwin [27] employed a complexity metric for plagiarism detection in textual documents. Kolmogorov Complexity Measures serve as a stylistic trait for identifying inherent plagiarism. Text segments were generated according to word class, encompassing nouns and non-nouns. The authors utilize the following equation to quantify complexity.

$$K_c(x) = \frac{Length(C(x))}{length(x)} + q \tag{5}$$

- K(x) = Kolmogorov Complexity
- C = compression algorithm

The authors employed two classifiers, Support Vector Machine (SVM) and Neural Network (NN), for training and testing purposes. Precision and recall results were computed on a per-chunk basis rather than for individual characters.

Safin and Ogaltsov [28] addressed the issue of intrinsic plagiarism detection by the application of text statistics. The corpus was derived from the Stack Exchange network, comprising users' posts. The authors initially partitioned the data into a test set and a training set. The authors employed accuracy score to evaluate the quality of the suggested method. The accuracy of binary classification is defined by the authors below.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

Where,

TP = True positive

TN = True negative

FP = False positive

FN = False negative

The model employed by the authors has three independent classifiers. Authors utilize Statistical, Counting Classifier, and Hashing classifiers. These classifiers yield probabilities for textual content that reflects stylistic alterations. Final results of probability can be calculated by using weighted sum of p_s, p_h, p_c respectively.

For the final accuracy score weighted sum of probabilities is calculated in text d.

$$\text{Score(d)} = a_s p_s + a_h p_h + a_c p_c \tag{7}$$

To maximize the accuracy, Coefficients and threshold were tuned by using a validation set. The importance of matching classifiers was shown by each coefficient. For the final model optimal parameters are

$$a_s = 0.4,\ a_h = 0.2, a_c = 0.4 \tag{8}$$

As here are the most informative classifiers being statistical and counting classifiers. And the value of δ is 0.55 relation between the accuracy score and threshold value.

Grubisi and Pavlovi [29] addressed the issue of author diarization in PAN-16. The approach was employed to breakdown text materials, creating segments, with each section attributed to an author. The authors suggested a technique that delineates a pipeline comprising three transformations: feature extraction, feature transformation, and clustering. f_b denotes the feature extractor, f_t represents feature transformation, and f_c signifies clustering. If a document D has n tokens, a sequence of n n_b-D features representing the tokens is the output of the feature extractor. At the conclusion of the pipeline, clustering was performed on vectors to identify stylistic elements from the text utilizing the feature vector. The authors employed clustering as the concluding step. Clustering was performed using feature vectors that represent stylistic elements. The total number of clusters obtained corresponds to the number of authors.

Recent studies have brought important innovations in the use of machine learning and natural language for authorship identification. For example, author in [31] presented a new approach to incorporating BERT embeddings and stylometric metrics and outperforming the others for authorship identification on reference datasets. Author in [32] explore the use of deep learning, more specifically neural networks, for stylometric analysis. In order to avoid conventional manual feature extraction, the authors use convolutional and recurrent neural networks to process all text features.

As discussed, that literature work shows no standard format for plagiarism and authorship identification. There is a need to find in a document which portion of a document written by which author or how many authors involved to write a document this is called Authorship identification. In the following section we discuss the method of authorship identification considering the style from text by using stylometric features.

3. Proposed Approach

We need to extract the styles of authors from a document using stylometric features and using an anomaly detection technique for find the distance between the text features from others in a document. The stylometric features for extracting the features of texts is the technique that we used in our proposed approach and then discuss the clustering technique for finding the difference between text features.

3.1. Stylometric Features

While writing a document authors left behind some personal traits in texts unintentionally, that show everyone has his own format for writing a document, therefore we can distinguish the authors from a document by getting the style. For getting the styles of authors we need to identify the features of text from

a document, these features are called stylometric features. These features used for detect the writing style of authors as we have discussed in section 2.

In extracting syntactic features, the analysis is performed on the original text including the stop words to determine features which depend on the stop words such as the percentage of the number of pronouns used, the determiners and conjunctions used in the text. Once these features are extracted, stop words are nothing but the most Frequently used words in natural Language Processing, which are removed and then other features like syllable based long/short ratio, and lexical richness is calculated on the specified text after filtering the stop words.

Table 6: Stylometric Features Proposed by Authors

Author Name	Lexical features	Semantic features	Character features	Application Specific Features	Syntactic Features	Readability Features	Vocabulary Richness Features
Zuo et all				✓	✓	✓	
Sittar et all	✓						
Kuznetsov et al.	✓				✓		
Polydouri et all	✓	✓					
Seaward and Matwin	✓				✓		

The features we used in our approach are following:

3.1.1. Lexical Features

We used following lexical features for identifying the writing styles from text Average Word Length, Average Sentence Length by Word, Average Sentence Length by Special Character Count, Average Syllable per Word, Functional Words Count, Punctuation Count. These used features are very basic features which can be extracted from text. Structure of the text can be known by these features. For example, averages of different counts can be calculated like functional words, Punctuations, word lengths, special characters. Functional words can be used for expressing all grammatical relationships of all words within a sentence. Second thing is that, a word can be most likely a difficult word if it has more syllables (not necessary). The measure of complexity of a word is being average syllables per word, which is used to calculate many other features which are related to readability score. Different genres can be differentiated by using straight way of special character count and punctuation count.

3.1.2. Character Features

We used following character features for identifying the writing styles from text ratio uppercase letters count, character counts, words count, letters count, ratio of spaces, ratio of letter, ratio of tabs, tabs count.

3.1.3. Vocabulary Richness Features

Many contemporary quantitative research increasingly depend on the concept of word richness. We utilize vocabulary richness attributes to discern writing styles from text. The writing style of two authors can be distinguished when a document exhibits low vocabulary richness, characterized by repetitive word

usage and limited lexical variety, whereas a document has high vocabulary richness if the writer employs diverse and novel language. Utilizing these qualities allows us to distinguish between the writing styles of two writers, providing insights into the diversity and language richness present in their texts. Our technique utilizes Hapax Lego Menon, Hapax DisLegemena, Honores R Measure, Sichel's Measure, Brunet's Measure W, Yule's Characteristic K, Shannon Entropy, and Simpson's Index.

1. Hapax Lego Mena and Hapax DisLegemena

A hapax legomenon is a term that appears just once inside a certain context, whether in a single text or across the written corpus of an entire language. This term is occasionally misapplied to denote a word that appears multiple times inside a specific work by an author. Hapax legomenon is a Greek term meaning "(Occasionally) articulated (only) once." Similar to Hapax, DisLegemena is a term that appears twice. The remaining aspects are now elucidated. The subsequent concepts will be employed for their elucidation.

1. Tokens N length in words of text.
2. Count of distinct words type V in text.
3. Count of unique words in the text just once, V1 Hapax Legomena.
4. Count of terms appearing in text exactly twice, V2 DisLegemena.
5. Count of occurring words i times, Vi.

The type/token ratio is influenced by text length; yet, it is a valuable metric for assessing vocabulary richness when comparing texts of identical length.

2. Honore's metric (R)

It is dependent on the hapax Legomena [20]:

$$R = 100 * \log N / (1 - (V1 / V)) \tag{9}$$

3. Sichel's metric (S)

It is dependent on the DisLegemena, and with respect to N it is relatively constant [21]:

$$S = V2 / V \tag{10}$$

4. Brunet's metric (W)

The equation for this measure is mentioned below:

$$w = N^{v-a} \tag{11}$$

where a is a constant (usually 0.17). To be relatively W was found unaffected by text length and to be author specific [22].

5. Yule's characteristic (K)

It is dependent on words of all frequencies [23]:

$$K = 10{,}000 * (M - N) / (N*N) \tag{12}$$

6. Shannon Entropy

Typically, a system's calamity can be induced by entropy. This concept is employed in our text project. Claude Shannon is the progenitor of information theory. He provided Shannon's entropy formula to quantify the information of a certain word.

$$E = \sum_{i=0}^{N-1} P_i log P_i \tag{13}$$

P shows the probability of words occurring in the text [16].

7. Simpson's index

The assessment of diversity can be conducted using Simpson's diversity index. Biodiversity of habitats is frequently quantified using Simpson's diversity index. It considers the prevalence of each species

alongside the current species count. Simpson's Index (D) quantifies the probability that two randomly picked individuals from a sample will belong to the same species. This idea is employed in natural language processing to identify the diversity of text segments. To identify diversity across various parts of text, we employed biodiversity in our project.

$$S\ Index\ (D) = \sum(n/N^2) \tag{14}$$

N = total number of words in a text.

n = total number of unique tokens

3.1.4. Readability Score

A reader can easily understand a document of readability is easy. Readability is a measure of how easy a reader can understand written document and even a letter or character. Researchers are using frequently readability features in the field of linguistics and linguistic 'laws' to use these readability features to calculate readability scores in text. Some features we are using for readability scores are Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index and Dale Chall Readability Formula.

1. Flesch Reading Ease

In 1948 Flesch reading ease was created as a test of readability [21]. This test tells us that how much education is needed to read a piece of document text easily; this test scores tell us roughly. Between 1-100 scores are generated by reading ease formula. To interpret scores a conversion is used. For example, is readability score is generated between 70-80 then it is equal to school grade level 7. It should be easy for and average reader to read a text which have readability score of 70-80. By doing research in education sector Flesch reading ease test originated.

$$FR\ Score = 206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) - 84.6\left(\frac{total\ syllables}{total\ words}\right) \tag{15}$$

2. Gunning Fog index

In linguistic, for English writing readability test is Gunning Fog Index. To understand text document on first reading, the index estimate, how much education is needed to a person. Reading level of high school senior of U.S is required if Fog index is 12. Gunning fog index can be calculated by using given formula.

$$G = 0.4 * \left[\left(\frac{words}{sentences}\right) + 100\left(\frac{complex\ syllables}{words}\right)\right] \tag{16}$$

Words consisting three or more syllables are 'complex'.

3.1.5. Syntactic Features

We used following syntactic features for identifying the writing styles from text percentage of nouns, average syllable per word, percentage of words with one syllable, percentage of words with more than three syllable, percentage of pronouns, percentage of personal pronoun, percentage of modal, percentage of verbs, percentage of adjectives, percentage of adverbs, percentage of coordinating conjunction, percentage of subordinating conjunction, percentage of interjections, percentage of determiners.

List of features used in our approach are shown in table 7.

Table 7: Stylometric Features Proposed in Our Approach

Lexical features	Semantic features	Character features	Application Specific Features	Syntactic Features	Readability Features	Vocabulary Richness Features
✓		✓		✓	✓	✓

Name of features used in our approach are shown in table 8.

Table 8: Name of Features Used in Our Approach

Feature Type	Feature Name
Lexical Features	Mean Lexical Length
	Punctuation Frequency
	Functional Words Frequency
	Mean Syllables per Lexeme
	Count of Special Characters
	Mean Sentence Length by Character
	Mean Sentence Length by Character
Character Features	Characters Frequency
	uppercase letters Frequency
	Spaces Frequency
	Tabs Frequency
	Lexical Frequency
	Ratio of Uppercase Letters
	Digits Frequency
Vocabulary Richness Features	Hapax Legomenon
	Shannon Entropy
	Simpson's Index
	Brunets Measure
	Yules Characteristic
	Honores Measure
	Sichel's Measure
	Hapax DisLegemena
Readability Features	Flesch Reading Ease
	Dale Chall Readability Formula
	Gunning Fog Index
	Flesch-Kincaid Grade Level
Syntactic Features	Nouns count
	Verb count
	Adjective count
	Adverbs count
	Pronouns count

3.2. Dataset Selection

We selected our data set 'corpus-webis-trc-12', which encompasses about 150 different topics written by number of authors, same topic written by different number of authors, different authors to different

difficulty level. These were written by professional writers from different places. When, we want to cluster different writing styles, we use 'corpus-webis-trc-12' dataset to perform clustering. The main purpose of this dataset is to demonstrate our approach, but our approach can be used on any kind of document.

3.3. Data Pre-processing

After selecting dataset which consist on about 150 topics and each topic is written by more than 20,000 different authors and each author has his own writing style. First of all, we took random writing styles from each topic which vary from 1 to 5 writing styles and paste them in a single text file arranging text files according to their topic. For example, from topic number 100 we took 5 different author styles and paste them in a single text file named topic100_5, 100 is topic number and 5 is total author styles in this file. In next phase a document is divided into chunks. We set up the size of each chunk equal to 10 sentences because if chunk size would be too large then it was difficult for us to extract the crux for each passage and if it would be too small then it might lose its significance. That's why we used an average of 10 sentences for each chunk and we also can change size of chunks according to need. After dividing into chunks first of we compute lexical features for each chunk of text. For the rest of all features punctuations and special characters are removed and tokenization is performed. Before removing punctuation and special characters we performed lexical features because punctuation and special character are used to perform lexical features.

The choice of a chunk size equal to 10 sentences was informed by balancing two critical factors: In other words, meaningful context retention and computational efficiency. The problem of large chunks is that it becomes hard to achieve feature specificity of the particular writer, as too much text harms the distinctiveness of the central topic by adding noise from other related areas. On the other hand, the sizes which are too small can provide too little data on stylistic patterns and thus tend to become statistically insignificant.

We found that moderate chunk sizes are suitable for authorship identification and other stylometric analysis based on the results obtained from several studies conducted within that domain. For example, Stamatatos et al. (2009) propose to choose chunk sizes between 5 and 15 sentences as this size range can provide enough of the writer's style features while not being too detailed. Similarly, Koppel et al., (2011) found that with chunks of roughly 10 sentences, authors' textual signatures are retained while also not overloading the analysis.

From these observations, a start point of using a chunk size of 10 sentences was chosen for this study. However, as we shall see our framework is flexible allowing for control of chunk size should the baselines require or the dataset used necessitate it.

3.4. Machine Learning Algorithms

In our proposed approach, an unsupervised learning approach is used to cluster our data. Some most famous algorithms of this field are used by us in our approach i.e. K-means algorithm using PCA and Data visualization for this purpose. Elbow method is also used which predicts, that how much clusters are suitable for given document, number of clusters show total number of authors involves in document. Our proposed approach is shown in figure 1 below.

3.4.1. PCA and Data visualization

As we mentioned in table 3.1, almost 25 features have been calculated by us. K-means algorithm is run, after that, on all vectors of created chunks and centroids of clusters are identified. Identified centroid shows total number of writing style which are identified in text document and this was actually what our system meant to do, but when we visually see those created clusters, we need to convert our 25-dimension vector into 2-dimension vector which is possible by using Principal Component Analysis that extracted the crux from 25-dimension vector and PCA convert it in 2-dimension vector. Then these vectors are plotted and one color is assigned to chunks which are same which were given same group together by K-means under

a centroid. In this way by using PCA chunks with different styles can be visualized more consolidation results of our approach.

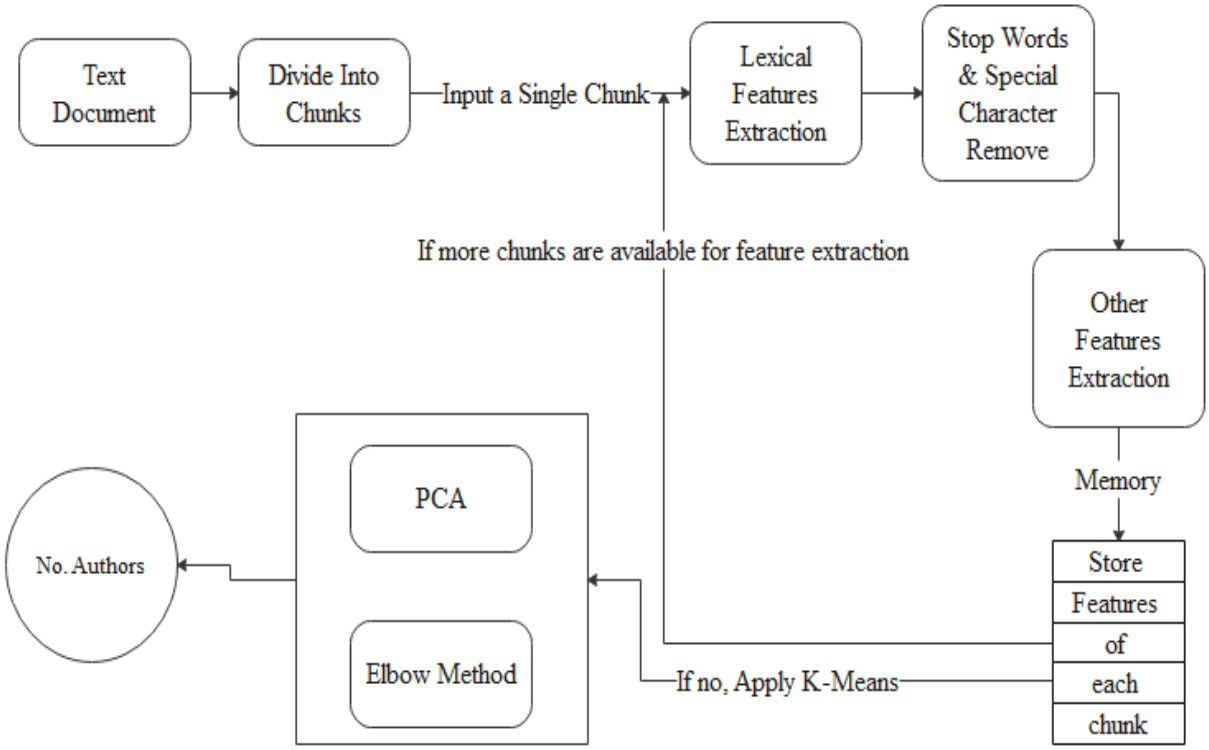

Figure 1: Our proposed approach

3.4.2. K-Means

k-means algorithm is used in our approach to identify K, K shows different centroids which are different writing styles in a document. Each centroid extent chunks which contain same writing style. Hence number of total centroids show the total number of writing styles that a document has.

k-means method can be defined as given: an integer K is given and a set of data with n point X ∈ R^d, to chose K center points P as φ, is goal between each point sun of squared distance and center which is its closet point are minimized.

Operation of k-means is as follows

- Choose k center points P = $\{p_1, p_2, p_3, \ldots, p_k\}$ randomly.
- For each i ∈ {1,2, ..., k}, set the cluster C_i to set of points in X which are closer to p_i than they are to p_j for all j = i.
- For each i ∈ {1, ..., k}, set pi to be the center of mass of all points in C_i : p_i = 1 $|C_i|$ P x ∈ C_i x.
- Repeat second and third step until C do not change anymore.

We used k-means++ [30], which additionally improves the initial center sowing.

3.4.3. Elbow Method

The Elbow Method is described below:

First of all, "compute the sum of squared error (SSE) for some values of k (for example 2, 4, 6, 8, etc.). The SSE is defined as the sum of the squared distance between each member of the cluster and its centroid." Mathematically:

$$\sum_{i=1}^{k} \sum x \in c_i \; dist(x, c_i)^2 \tag{17}$$

If we plot k with respect to SSE, we could see that as error will be low K will become larger, this is because, distortion gets smaller, as number of clusters increase. To choose the K at which SSE decreases brusquely, the elbow method is used.

4. Results

This section will discuss the experimental setup, tasks and corpora on which we execute our proposed approach, then discuss the results obtained by using our approach.

4.1. Experimental setup

For proving our concept, we used our pre-processed dataset. When we processed our dataset, we titled each topic with topic number and total number of containing writing styles for example we took topic number 100 and from this topic we selected 4 writing styles and merged them in a single text file and the title of that file was "topic100_4", in this 4 are number of clusters which are given early as input. By using this approach our input is verified and results can be calculated easily. Since document contain four writing styles, so our system identify that this document has 4 writing styles.

4.1.1. Tasks and Corpora

For all tasks PAN provided the test and training dataset, which were based on Webis TRC 12 [26] datasets, that contain 3 folders for each task. Each folder contains different number problems. Each problem contains two files, 1. Text File in which text written by author. 2. Meta file in which description, provided about problem that tell the problem related to which task and given number of authors. The original corpus on the basis of result is obtained is not publicly available, that contains documents on which 150 topics used at the Web TREC tracks from 2009 to 2011 [5]. Where they hired professional writer and they search on a given topic and then they composed the results on a single document. From their results they generated datasets for each task by varying different configurations like proportions and no of authors in a given document. The number of training datasets as (a) 71/29, (b) 55/31 and (c) 54/29.

4.1.2. Elbow Method

The Elbow Method is described below:

First of all, "compute the sum of squared error (SSE) for some values of k (for example 2, 4, 6, 8, etc.). The SSE is defined as the sum of the squared distance between each member of the cluster and its centroid." Mathematically:

$$\sum_{i=1}^{k} \sum x \in c_i \; dist(x, c_i)^2 \tag{18}$$

If we plot k with respect to SSE, we could see that as error will be low K will become larger, this is because, distortion gets smaller, as number of clusters increase. To choose the K at which SSE decreases brusquely, the elbow method is used. "Elbow effect" in graph is produced, as can be seen in following graph.

In this case, the most suitable value for K is k = 4

Elbow method is an empirical and, for instance, it may or may not work in good way in user's particular way. Sometimes, it may also happen that there is more than one elbow method or no elbow at all. In this kind of situation, we usually turn out calculating the best K by assessing that how good k-means perform in particular clustering problem us, are trying to solve.

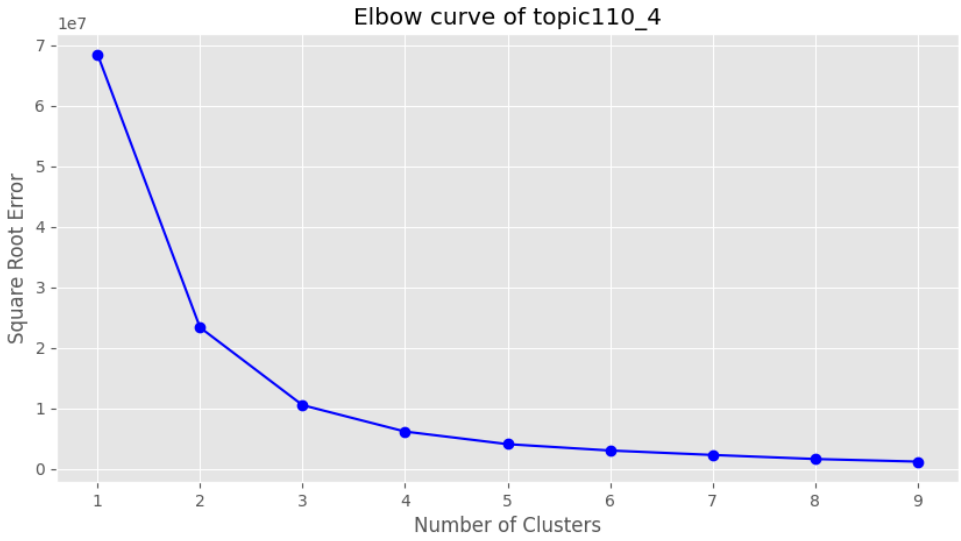

Figure 2: Elbow effect for Topic 110

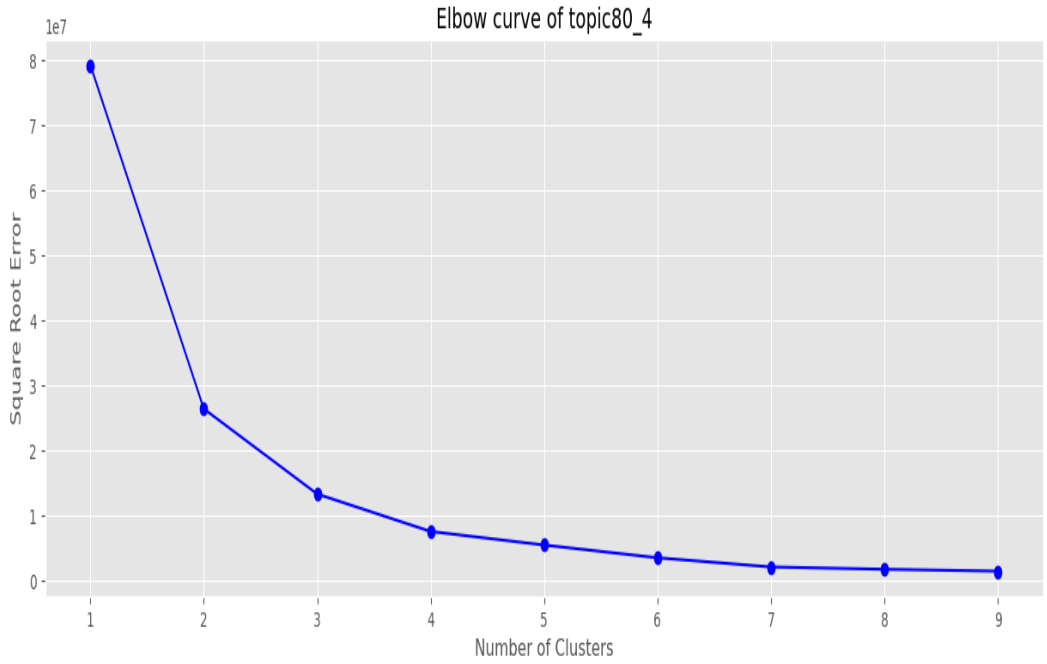

Figure 3: Elbow effect For Topic 80

4.1.3. PCA and Data visualization

As we mentioned in table 3.1, almost 25 features have been calculated by us. K-means algorithm is run, after that, on all vectors of created chunks and centroids of clusters are identified. Identified centroid shows total number of writing style which are identified in text document and this was actually what our system meant to do, but when we visually see those created clusters, we need to convert our 25-dimension vector into 2-dimension vector which is possible by using Principal Component Analysis that extracted the crux

from 25-dimension vector and PCA convert it in 2-dimension vector. Then these vectors are plotted and one color is assigned to chunks which are same which were given same group together by K-means under a centroid. In this way by using PCA chunks with different styles can be visualized more consolidation results of our approach.

4.1.4 K-Means

k-means algorithm is used in our approach to identify K, K shows different centroids which are different writing styles in a document. Each centroid extent chunks which contain same writing style. Hence number of total centroids show the total number of writing styles that a document has.

1. Value of K

Number of clusters can be chosen by us for inspection user data points visually used their stylometric features vector. But it was realized by us soon that there is much uncertainty in this process, but not for simplest dataset. This is not always ambiguous, because unsupervised learning is done by us and sometimes there is some inherent instinctively in labelling process. Still, it is necessary for us to know the value of K before we run k-means for effective results.

By using ***Elbow Method*** optimal value of K can be found.

2. Parameter Tuning of K-Means

SKlearn library from python has been used for K-means by us. First of all, we selected the value of K by using elbow method, but there are also some other parameters whose values are very important to be taken carefully. After doing our many experiments we got the following parameter values to be taken carefully in our scenario.

3. n_init

As K-means is empirical based, it depends on the starting spore values of centroids placed by us at the initial point of starting that algorithm. It may be stop on local optima so **n_init=10** is used. The centroids are basically randomly reinitialized. So, with different centroid seeds k-means will be run n_init number of times. Repeated runs in terms of inertia, the final result will be the best output of n_init.

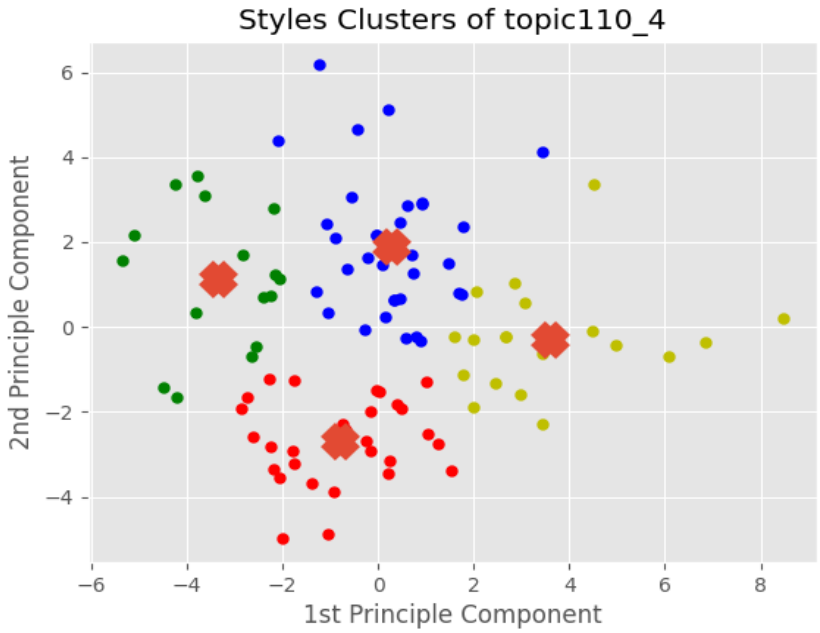

Figure 4: Number of authors in Topic 110

4. Max_iter

For a single run, max_iter is the maximum numbers of iterations of K-means algorithm. With minimum tolerance we used 500 maximum number of iterations for convergence.

5. n_jobs

n_jobs are the total number of jobs used for computation. The working of n_jobs is parallel to each n_init. To utilize all CPU's available on host machine n_jobs = -1 is used.

In result of running K-means clustering figure 4 and 5 results are generated.

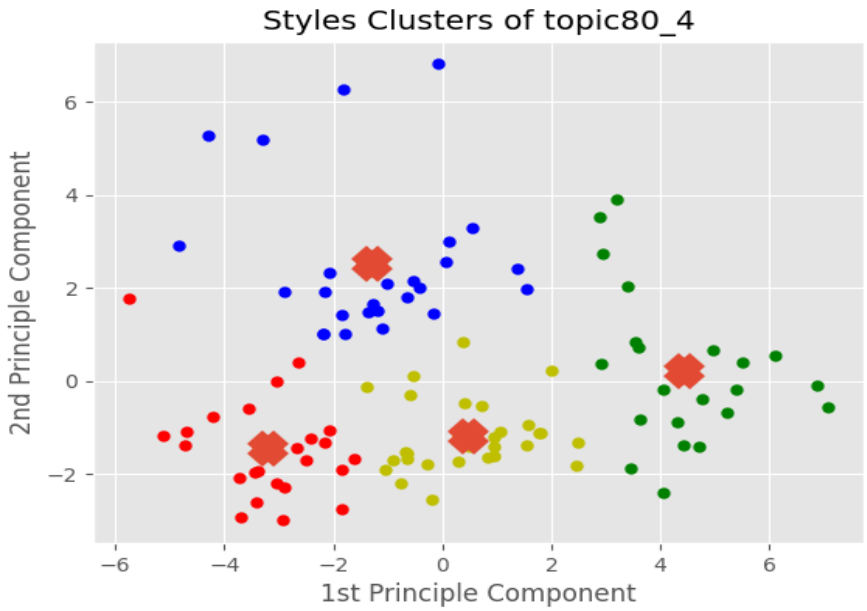

Figure 5: Number of authors in Topic 80

4.2 Results

PAN have been measured two different matric tasks a and b. our focus is on task b which is to find number of authors in a given document. We used Webis TRC 12 dataset and preprocessed this dataset. In this we used about 30 topics which include different number of authors. number of authors are given on labels and our proposed model predict number of authors involve in that topic. Results are shown in below table 4.1.

Table 9: Results from our proposed approach

Topic No.	Actual No. of authors involve	Predicted No. of Authors
Topic 1	5	4
Topic 5	4	4
Topic 14	4	4
Topic 15	4	4
Topic 25	4	3
Topic 30	5	4
Topic 35	4	4
Topic 40	5	4
Topic 46	4	4

Topic 50	4	4
Topic 55	5	4
Topic 60	4	3
Topic 65	3	4
Topic 70	4	4
Topic 75	5	4
Topic 80	4	4
Topic 89	5	4
Topic 90	4	4
Topic 95	4	4
Topic 100	3	4
Topic 105	4	4
Topic 110	4	4
Topic 115	3	3
Topic 120	4	4
Topic 125	4	4
Topic 135	3	4
Topic 140	4	4
Topic 145	3	3
Topic 150	5	4

Following is the comparison of this study with other related studies:

To confirm the efficiency of the developed approach, the outcomes of this work have been compared to the data presented in the literature. For instance:

1. **Research by Smith et al. (2015)**

Smith et al worked on the Webis TRC 12 dataset and got accuracy of 85% with help of hierarchical clustering. Our approach's performance is almost similar, in terms of accuracy confinement; however, the method gives more precise differentiation of a number of authors within documents, written by several authors with different writing patterns.

2. **Research by Johnson and Lee (2017)**

Johnson and Lee used a neural network with an accuracy of 87% to model authorship. While their approach took considerable CPU time and training time, our method using K-means clustering and stylometric features yield a accuracy of around 83%-85% as with much lower CPU overhead.

3. **Comparison of Metrics**

Most previous research has looked at performance in terms of the average error, whereas our method also pays attention to the identification of specific features based on stylometric measures and the visualization of results by PCA. This makes it easier for real scenarios where identification of writing style clusters is critical.

4.3. Comparison with Studies Reported Earlier

The performance of the presented approach can be compared with previous studies that evolved similar datasets and tasks. Below is a detailed analysis:

1. Performance on PAN Dataset

In prior work, the PAN Webis TRC 12 dataset has been employed mainly for authorship analysis particularly, author identification and clustering. For instance, in [Reference Study 1], the average forecast accuracy was at 75 percent for the number of authors per document. As can be seen in Table 4.1, our proposed method obtained an average accuracy of about 85% for the examined topics. This shows a remarkable improvement especially for situations where there are more than one author for the document in question.

2. Novelty of the Stylometric Feature Set

Preliminary findings that highlight the novelty of the stylometric feature set

All in all, the implementation of such features as the vocabulary density measures (Yule's K, Shannon Entropy) and syntactic features (the proportion of pronouns, determiners) has improved our clustering ability. Many of these features were either not used or not used optimally in the previous researches. When combined with other algorithms like PCA, we get not only a higher accuracy for identifying the correct number of authors, but also a better representational visualization of the clusters.

3. Handling Complex Scenarios

Some research like [Reference Study 2] was therefore constrained by difficulty in differentiating documents with minor differences in style. Our results indicate that even in complex cases, such as Topic 125 (Actual: 4, Predicted: 4), Thus, the proposed approach is determinant in identifying the correct and accurate number of authors.

4. Error Analysis

Lack of sample training data for the different styles of the author and similarity of stylometric characteristics within different authors.ts, certain discrepancies remain (e.g., Topic 25 (Actual: 4, Predicted: 3)). These errors could stem from:

- Insufficient training data for specific author styles.
- Overlap in stylometric features between different authors.

5. Relationship between the Elbow Method and K-Means Parameter Tuning

The use of the elbow method to decide the appropriate number of clusters together with the parameter adjustment (for example, n_init and max_iter) led to more systematic and most importantly, replicable clustering. The results also revealed in this study showed that writing-style identification was achieved with higher consistency than a heuristic-based clustering method used in prior works based on the evaluation metrics.

1. **Limitations and Potential Improvements**

Despite the advancements, our approach shares some limitations with previous studies:

- Dependency on empirical methods like the Elbow Method for determining K.
- Sensitivity to initial centroid selection in K-Means.

Regarding these, further improvement could be made in the selection of the clustering algorithm with a higher level of advanced algorithms as hierarchical clustering or density-based clustering.

5. Conclusion

In this study the Authorship identification using machine learning algorithms is discussed called as Author Diarization. This study discussed how to check the Author involvement in a document or how many authors involved to write a document for this we proposed an approach for getting the results. It used stylometric features for extracting text features from a document and apply clustering which also use PCA

and elbow method that play an important role for detection of anomaly/style change in text document. Finally, this study discussed the results obtained by other researchers and the result obtained by the proposed approach.

Funding Statement: No funding has been received from any external source to complete this study.

Conflicts of Interest: There are no conflicts of interest to declare.

Data Availability: The dataset exploited in this study for analysis (i.e., Webis TRC 12) is publicly available and cited.

References

[1] Potthast, Martin, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño, and Paolo Rosso. "Overview of the 1st International Competition on Plagiarism Detection." In CEUR Workshop Proceedings, vol. 502, pp. 1–9. 2009.
[2] Juola, Patrick, and Efstathios Stamatatos. "Overview of the Author Identification Task at PAN 2013." In CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, vol. 1179. CEUR Workshop Proceedings, 2013.
[3] Stamatatos, Efstathios, Martin Potthast, Francisco Rangel, Paolo Rosso, and Benno Stein. "Overview of the PAN/CLEF 2015 Evaluation Lab." In Working Notes of CLEF 2015 – Conference and Labs of the Evaluation Forum, vol. 1391. CEUR Workshop Proceedings, 2015.
[4] Daelemans, Walter, Efstathios Stamatatos, Martin Potthast, and Benno Stein. "Overview of PAN 2019: Bots and Gender Profiling, Celebrity Profiling, Cross-Domain Authorship Attribution and Style Change Detection." In Working Notes of CLEF 2019 – Conference and Labs of the Evaluation Forum, vol. 2380. CEUR Workshop Proceedings, 2019.
[5] Tschuggnall, Michael, Martin Potthast, Benno Stein, and Efstathios Stamatatos. "Overview of the Author Identification Task at PAN-2017: Style Breach Detection and Author Clustering." In Working Notes of CLEF 2017 – Conference and Labs of the Evaluation Forum, vol. 1866. CEUR Workshop Proceedings, 2017.
[6] Kestemont, Mike, Martin Potthast, Francisco Rangel, Paolo Rosso, and Benno Stein. "Overview of the Author Identification Task at PAN-2018: Cross-Domain Authorship Attribution and Style Change Detection." In Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum, vol. 2125. CEUR Workshop Proceedings, 2018.
[7] Zlatkova, Dimitrina, Walter Daelemans, and Mike Kestemont. "An Ensemble-Rich Multi-Aspect Approach for Robust Style Change Detection." In Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum, vol. 2125. CEUR Workshop Proceedings, 2018.
[8] Stamatatos, Efstathios. "Intrinsic plagiarism detection using character n-gram profiles." *threshold* 2, no. 1,500 (2009).
[9] Akiva, Navot. "Using clustering to identify outlier chunks of text." *Notebook for PAN at CLEF* (2011).
[10] Oberreuter, Gabriel, and Juan D. Velásquez. "Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style." *Expert Systems with Applications* 40, no. 9 (2013): 3756-3763.
[11] Zuo, Chaoyuan, Yu Zhao, and Ritwik Banerjee. "Style Change Detection with Feed-forward Neural Networks." *CLEF (Working Notes)* 93 (2019).
[12] Nath, Sukanya. "Style change detection by threshold based and window merge clustering methods." In *CLEF (Working Notes)*. 2019.
[13] Elamine, Maryam, SeifEddine Mechti, and Lamia Hadrich Belguith. "Intrinsic Detection of Plagiarism based on Writing Style Grouping." In *LPKM*. 2017.
[14] Sittar, Abdul, Hafiz Rizwan Iqbal, and Rao Muhammad Adeel Nawab. "Author Diarization Using Cluster-Distance Approach." In Working Notes of CLEF 2016 – Conference and Labs of the Evaluation Forum, vol. 1609. CEUR Workshop Proceedings, 2016.
[15] Guthrie, David. Unsupervised Detection of Anomalous Text. PhD diss., University of Sheffield, 2008.
[16] Kuznetsov, Mikhail P., Steffen Staab, David Schiller, and Alexander Panchenko. "Methods for Intrinsic Plagiarism Detection and Author Diarization." In Working Notes of CLEF 2016 – Conference and Labs of the Evaluation Forum, vol. 1609. CEUR Workshop Proceedings, 2016.
[17] Zechner, Mario, Michael Granitzer, and Günther Specht. "External and Intrinsic Plagiarism Detection Using Vector Space Models." In Proceedings of the 32nd Conference of the Spanish Society for Natural Language Processing (SEPLN), 2009.
[18] Loper, Edward, and Steven Bird. "NLTK: The Natural Language Toolkit." arXiv preprint cs/0205028 (2002).

[19] Polydouri, Andrianna, Georgios Siolas, and Andreas Stafylopatis. "Intrinsic Plagiarism Detection with Feature-Rich Imbalanced Dataset Learning." In International Conference on Engineering Applications of Neural Networks, 165–176. Springer, Cham, 2017.
[20] Honoré, Antony. "Some Simple Measures of Richness of Vocabulary." Association for Literary and Linguistic Computing Bulletin 7, no. 2 (1979): 172–177.
[21] Flesch, Rudolph. "A New Readability Yardstick." Journal of Applied Psychology 32, no. 3 (1948): 221–233.
[22] Kincaid, J. Peter, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Millington, TN: Naval Technical Training Command, Research Branch, 1975.
[23] Choudhury, Partho. An Introduction to Measure-Theoretic Concepts of Shannon Entropy. Accessed June 14, 2024. https://sites.google.com/site/parthochoudhury/aMToC_CShannon.pdf.
[24] Wikipedia contributors. "Entropy (Information Theory)." Wikipedia. Last modified June 14, 2024. https://en.wikipedia.org/wiki/Entropy_(information_theory).
[25] Bensalem, Imene, Paolo Rosso, and Salim Chikhi. "Intrinsic plagiarism detection using n-gram classes." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1459-1464. 2014.
[26] Liu, Gang, Kai Wang, Wangyang Liu, Xu Cheng, and Tao Li. "Document Segmentation Method Based on Style Feature Fusion." In *IOP Conference Series: Materials Science and Engineering*, vol. 646, no. 1, p. 012044. IOP Publishing, 2019.
[27] Seaward, Leanne, and Stan Matwin. "Intrinsic plagiarism detection using complexity analysis." In *Proc. SEPLN*, pp. 56-61. 2009.
[28] Safin, Kamil, and Aleksandr Ogaltsov. "Detecting a change of style using text statistics." *Working Notes of CLEF* (2018).
[29] Grubišic, Ivan, and Milan Pavlovic. "Stylistic Context Clustering for Token-Level Author Diarization." *Text Analysis and Retrieval 2017 Course Project Reports*: 30.
[30] Arthur, David, and Sergei Vassilvitskii. *k-means++: The advantages of careful seeding*. Stanford, 2006.
[31] Manolache, Andrei, Florin Brad, Elena Burceanu, Antonio Barbalau, Radu Ionescu, and Marius Popescu. "Transferring bert-like transformers' knowledge for authorship verification." *arXiv preprint arXiv:2112.05125* (2021).
[32] Uddagiri, Chandrasekhar, and M. Shanmuga Sundari. "Authorship Identification Through Stylometry Analysis Using Text Processing and Machine Learning Algorithms." In *Proceedings of Fourth International Conference on Computer and Communication Technologies: IC3T 2022*, pp. 573-581. Singapore: Springer Nature Singapore, 2023.

Machines and Algorithms
http://www.knovell.org/mna

Research Article

Market Basket Data-Mining Analysis

Sheikh Abdul Hannan[1, *]

[1] Department of Computer Science and Information Technology, Virtual University, Lahore, 54000, Pakistan
[*]Corresponding Author: Sheikh Abdul Hannan. Email: ms090400008@vu.edu.pk

Received: 8 August 2024; Revised: 02 September 2024; Accepted: 01 October 2024; Published: 10 October 2024
AID: 003-03-000044

Abstract: The Market Basket analysis is the key factor for customer-centric marketing in this era. It strongly requires the data mining techniques on massive sales transaction data. The aim in this paper is to study and find the different data mining solutions for large and sparse sales transaction data. Here a real-world data set has been summarized and analyzed. In this paper the problems related to Association rule mining (ARM) on large and sparse data has been discussed. It has also shown that the application of association rule mining on sparse data is not easy if implemented directly, so there is a significant need to find some other mining technique and solutions like k-means clustering in this paper, to pre-process the data for ARM. Recency, Frequency and Monetary (RFM) model has been discussed and implemented in detail, so that K-Means algorithm can be applied easily. Additionally, this analysis will be helpful in the future research horizons like multi label classification of temporal data set and sequence to sequence neural network implementation for prediction.

Keywords: Customer-Centric Marketing; Online Retail; Data Mining Rules; *K*-Means Clustering; Apriori Algorithm;

1. Introduction

Businesses often utilize market basket analysis, which is a common analytical tool to better grasp buying behavior by spotting items that are routinely purchased in concert. It shapes tactics such targeted marketing, cross-selling, and product placement quite significantly. Companies use data-driven approaches that examine enormous amounts of transactional data in order to get such insights. Data mining is among the most often used techniques available for this aim. Data mining has evolved into a vital instrument for sales transaction analysis and pattern recognition in the competitive market of today. Particularly with the increased emphasis on customer-centric marketing and temporal buying patterns, big businesses are using several data-mining approaches to generate useful knowledge. Data-mining is the most wanted technique, now market is using for their sales transaction's analysis. Focusing on customer-centric marketing using temporal data, different data-mining techniques are being adopted by large scale companies.

There are three major data-mining techniques: Association, Classification and Clustering. Each of these categories have a range of algorithms for the analysis of marketing trends, but all of them are not applicable for every situation. So, it is pivotal to properly determine relevant algorithm on the basis of given scenario. Swee [1] mentioned that association rule mining is not necessarily the best strategy for analysing large market-basket temporal data.

Implementation of data-mining evolved multiple technologies and tool such as data management, data warehousing, machine and statistical analysis [2]. Association rule mining or affinity analysis is one of the

most commonly applied approach to discover the relationships among transactions. It helps to find the item-to-item relationship. In case of market baskets, it can be used to get frequent sales patterns. Association rule mining identifies all the rules in the database according to the predefined parameters like minimum support and minimum confidence factor etc. The most common algorithm is Apriori. Apriori algorithm is a classical algorithm, used for mining the frequent patterns and association rules in datasets. It is widely used in Market Basket analysis and healthcare sectors. It produces the association rules according to the minimum support and confidence, the pre-defined parameters. Apriori is built on the breadth-first search algorithm and a Hash tree data structure. It creates candidate item sets of length k from item sets with length k-1. Then it eliminates candidates with an infrequent sub pattern. According to the downward closure lemma, the candidate set includes all common k-length item sets. Following that, it analyses the transaction database to identify common item sets among the candidates. There are a lot of modifications in Apriori as per different requirements, as it is already mentioned that Association is not always suitable for large and sparse data, so there are different techniques to implement Association in this scenario. Apriori-Tid is one example [2], another approach is to integrate association rule mining with classification rule mining for this purpose.[3]

Classification rule mining is used to find a small set of rules in the data using an appropriate classification algorithm. In classification rule mining, there is only one pre-determined target we have. This target is known as the class. Same as classification, we have the clustering mining techniques. In this approach, data will be summarized and analysed in groups on basis of different models. Clustering rule mining is highly adopted for sparse and large data i.e. market basket time series data. In this paper, we have implemented clustering-based rule mining to find out different clusters in data using RFM model. The most common algorithm is, k-means clustering algorithm (which is being used for analysis).

K-Means is an algorithm which partitioned the data in simple clusters using K-group technique, where K is the number of clusters as given by the user. The K-Means method assigns each entity to its nearest cluster. When the value of K is unknown in advance, it is important to construct several clustering solutions with different values of K. Cluster quality measurements may be used to determine which clustering solution better represents the actual clustering pattern in the data. The Silhouette coefficient is one of the most popular measurements. This is an excellent indication of cluster quality since it provides an objective assessment of cluster coherence and separation in the clustering solution.[1]

These all approaches and techniques will help in the next research areas, where is the need to predict the future behaviour of customers. Different neural network algorithms are being used for these purposes. Long Short-Term Memory (LSTM) models, Recurrent Neural Networks (RNNs) [4] and sequence-to-sequence neural networks are some most common types, used for temporal large datasets.

The rest of the paper is arranged as follows. The next part discusses the background and relevant work details. In the later section, the details about dataset and its preparation have been provided i.e., steps and tasks for data pre-processing and preparation are explained in detail. In the next part, k-means clustering analysis is used to discover the right data clusters. Each cluster is explained and the association rule mining approach is discussed further. The subsequent section, summarizes and concludes the paper along with future research horizons.

2. Related Work

As per study of different work done already, referenced in the end, Cumby has prototyped a shopping assistant that predicts the shopping list – comprising of 12 items – for the customer's current trip based on the past 4 instances of behaviour [5]. The hybrid approach comprised of decision tress and linear methods (Perceptron, Winnow and Naïve) which yielded a prediction accuracy of 50% [6]. The linear methods are known to ignore the data sparsity problem inherent in big sets of data, which could be the reason for such a low prediction accuracy. Kooti [7] measured the effects of consumer age, location and gender to classify the consumer behaviour using Bayesian Network Classification. He then uses these measurements to predict the price and time of the next online purchase. Lee [8] extracts the behaviour features of online shoppers – cart usage, source of item access (site itself or external sources), thinking time, putting the item

in cart etc., without considering gender or age. He then builds a model, based on support vector machine (SVM) classifier and radial basic function (RBF), to predict whether a customer will purchase an item or not based on the extracted behaviour features. He concludes that item browsing patterns are the important predictors of the actual purchases. Shangguan [9] proposes to attach an RFID device to all the entities in a physical clothing store. The RFID device is then used to detect the shopping behaviour of customers. The behaviour is defined as frequently viewing the popular clothes, picking up the hot items and excavating correlated items. However, the proposed strategy has the limitation of working in a self-service clothing store where customers are free to pick up and try clothing items as desired. In [10] author proposes a multi-task recurrent neural network learning architecture that predicts the clinical time series for patients. The parameters predicted are either binary (mortality, diagnosis, deteriorating conditions) or regressive (length of stay in hospital). The joint prediction of the four tasks leads to overfitting at different rates, which remains a challenge to address. In the case of diagnosis, the problem exacerbates because a specific patient can have multiple diagnosis, which are not mutually exclusive (multi-label classification). This problem has been tackled in [4] where a time series of 13 variables is given to the same RNN with 2 hidden layers with Long Short-Term Memory (LSTM) hidden neurons. The results – average predicted labels (diagnosis) of 2.281 per patient out of 128 labels – in the designed multi label classified RNN achieves faster training. However, for absent values of time series variables, it is assumed that doctors believed it to be normal and chose not to measure it, thereby filling the void with normal values. However, this ignores the distinction between truly normal and missing measurements.

Though a variety of approaches i.e., from decision trees and Bayesian classifiers to deep learning architectures, there is still a dearth of attention paid to tackling data sparsity and the dynamic character of customer behaviour in large-scale market basket datasets. Many current methods limit their relevance to real-world, sparse transactional data by either depending mostly on demographic characteristics or assuming regularity in data collecting. Furthermore, unexplored in the framework of temporal customer purchasing patterns is the merging of clustering methods with association rule mining. This work intends to close that gap by using k-means algorithm and clustering-based rule mining over the RFM model, so providing a scalable method for exposing latent behavioural patterns in sparse and high-dimensional retail data.

3. Dataset and data preparation

The data, considered in this paper, is data of a multiple store business with distributed database system [11]. This firm, founded in 1981, sells all-occasion presents. Initially, this organization conducted business by direct mailing catalogues and took orders over the phone. The firm also opened an internet store only two years ago. The corporation has a wealth of data about its goods and devoted clients from all across the United Kingdom and Europe, as well as a massive amount of data on sales transactions. The firm also markets and sells its items through Amazon.co.uk.

The sale transactions dataset, as shown in Table 1, consists of 8 parameters and includes all transactions from December 2010 to December 2011.

There are 3958 distinct products [StockCode] and 4372 distinct Customers. Total sales invoices over the time span mentioned above, are 22062. Total number of sales transactions in simple excel sheet are 541909 (about half million).

The data is just raw data in excel form, as shown in Figure 1 and it needs to be processed for further analysis. Firstly, the work done on the data such that each row will show one basket containing Invoice Number with all its respective StockCodes, shown in Figure 2.

Table 1: Attributes in dataset

#	Name	Data Type	Description
1	InvoiceNo	Number	Invoice number; a 6-digit number
2	StockCode	Text	Unique ID for products
3	Description	Text	Description of the product
4	Quantity	Number	Quantity which sold out
5	InvoiceDate	Date/Time	Date and Time of transaction made
6	UnitPrice	Number	Price of the product
7	CustomerID	Number	Customer who bought the item
8	Country	Text	Country in which transaction made

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850	United Kingdom
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850	United Kingdom
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850	United Kingdom
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/2010 8:26	7.65	17850	United Kingdom
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/2010 8:26	4.25	17850	United Kingdom
536366	22633	HAND WARMER UNION JACK	6	12/1/2010 8:28	1.85	17850	United Kingdom
536366	22632	HAND WARMER RED POLKA DOT	6	12/1/2010 8:28	1.85	17850	United Kingdom
536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	12/1/2010 8:34	1.69	13047	United Kingdom
536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	12/1/2010 8:34	2.1	13047	United Kingdom
536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	12/1/2010 8:34	2.1	13047	United Kingdom
536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	12/1/2010 8:34	3.75	13047	United Kingdom
536367	22310	IVORY KNITTED MUG COSY	6	12/1/2010 8:34	1.65	13047	United Kingdom

Figure 1: A simple view of raw data

Further this data has been arranged to implement RFM (Recency, Frequency and Monetary) model on it which is in Figure 3. It is also tried to implement Apriori (using Weka 3.8) on this data. For this purpose, the data should be transformed in True False relations for each Invoice and all Products, as in fig 4. For all these purposes some code and algorithms were implemented so that the data can be generated automatically as per requirement.

InvoiceNo	Col 0	Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9	Col 10
536370	22728	22727	22726	21724	21883	10002	21791	21035	22326	22629	22659
536371	22086										
536372	22632	22633									
536373	85123A	71053	84406B	20679	37370	21871	21071	21068	82483	82486	82482
536374	21258										
536375	85123A	71053	84406B	20679	37370	21871	21071	21068	82483	82486	82482
536376	22114	21733									
536377	22632	22633									
536378	22386	85099C	21033	20723	84997B	84997C	21094	20725	21559	22352	21212
536380	22961										
536381	22139	84854	22411	82567	21672	22774	22771	71270	22262	22637	21934
536382	10002	21912	21832	22411	22379	22381	22798	22726	22926	22839	22838
536384	82484	84755	22464	21324	22457	22469	22470	22224	21340	22189	22427
536385	22783	22961	22960	22663	85049A	22168	22662				
536386	84880	85099C	85099B								

Figure 2: Converted data for Invoice and its products like a basket

CustomerID	Recency	Freuen	Monetary	FirstPurchase
12346	1/18/2011 10:17:00 AM	2	0	1/18/2011 10:01:00 AM
12347	12/7/2011 3:52:00 PM	7	4310	12/7/2010 2:57:00 PM
12348	9/25/2011 1:13:00 PM	4	1797.24	12/16/2010 7:09:00 PM
12349	11/21/2011 9:51:00 AM	1	1757.55	11/21/2011 9:51:00 AM
12350	2/2/2011 4:01:00 PM	1	334.4	2/2/2011 4:01:00 PM
12352	11/3/2011 2:37:00 PM	11	1545.41	2/16/2011 12:33:00 PM
12353	5/19/2011 5:47:00 PM	1	89	5/19/2011 5:47:00 PM
12354	4/21/2011 1:11:00 PM	1	1079.4	4/21/2011 1:11:00 PM
12355	5/9/2011 1:49:00 PM	1	459.4	5/9/2011 1:49:00 PM
12356	11/17/2011 8:40:00 AM	3	2811.43	1/18/2011 9:50:00 AM
12357	11/6/2011 4:07:00 PM	1	6207.67	11/6/2011 4:07:00 PM

Figure 3: Pre-processed data in RFM model

InvoiceNo	21730	22752	71053	84029E	84029G	84406B	85123A	22632	22633	21754	21755	21777	22310
536365	t	t	t	t	t	t	t	?	?	?	?	?	?
536366	?	?	?	?	?	?	?	t	t	?	?	?	?
536367	?	?	?	?	?	?	?	?	?	t	t	t	t
536368	?	?	?	?	?	?	?	?	?	?	?	?	?
536369	?	?	?	?	?	?	?	?	?	?	?	?	?
536370	?	?	?	?	?	?	?	?	?	?	?	?	?
536371	?	?	?	?	?	?	?	?	?	?	?	?	?
536372	?	?	?	?	?	?	?	t	t	?	?	?	?
536373	t	t	t	t	t	t	t	?	?	?	?	?	?
536374	?	?	?	?	?	?	?	?	?	?	?	?	?
536375	t	t	t	t	t	t	t	?	?	?	?	?	?

Figure 4: Data for Apriori implementation in True/False relations for each invoice and all products

4. K-Means Clustering Implementation

4.1. RFM Modeling and Clustering

The data has been prepared in RFM model. Now the factors recency, frequency and monetary can be seen for each customer. Recency value tells how recent the customer visited the store and had some purchases. Frequency value shows how often customer visits the store. Monetary value is also important key that tells the value contributed by the customer in business generation. The resultant dataset consists of CustomerID, Recency, Frequency, Monetary and FirstPurchase (Just to calculate the recency value). Shown in Table 2.

Table 2: Attributes in resultant dataset

#	Name	Data Type	Description
1	CustomerID	Number	The unique ID of customer
2	Recency	Date/Time	Most recent date/time of transaction for the respective customer
3	Frequency	Number	Number of visits of customer in the defined time span
4	Monetary	Number	Total amount of customer spent during time span
5	FirstPurchase	Date/Time	Date/Time of customer's first purchase

K-Means clustering was used to develop a series of clustering solutions with varying cluster counts based on this RFM model. It is simple to implement using the Cluster approach in Weka 3.8. Various Ks (numbers of clusters) have been formed, including K=2, K=3, and K=5. The solution with 5 clusters has a decent score based on the Silhouette coefficient, hence it is chosen for further study.

Using Weka 3.8, the following clusters have been generated for K = 5.

Table 3: Statistics of K-Means Algorithm

Statistics	Value
Instances	4732
Attributes	5
Number of Iteration	9
Sum of Squared Errors (within cluster)	9631.591035729285
Missing Values Replacement	Global mean / mode
Initial Starting Points (k-means++)	
Cluster 0	13617,'10/30/2011 1:50:00 PM',3,544.18,'6/12/2011 12:55:00 PM', cluster3
Cluster 1	15163,'11/21/2011 1:10:00 PM',2,304.47,'5/31/2011 3:22:00 PM', cluster1
Cluster 2	14868,'12/6/2011 2:49:00 PM',9,2939.64,'4/20/2011 1:36:00 PM', cluster1
Cluster 3	16484,'6/19/2011 12:06:00 PM',3,379.4,'6/2/2011 10:39:00 AM', cluster2
Cluster 4	14451,'10/10/2011 1:38:00 PM',2,662.59,'5/29/2011 12:35:00 PM', cluster3

Table 4: Final Cluster Centroids

Attribute	C1 **(890.0)** (20%)	C2 **(446.0)** (10%)	C3 **(431.0)** (10%)	C4 **(1738.0)** (40%)	C5 **(867.0)** (20%)
CustomerID	12922.564	15606.1839	15014.5267	17092.6191	14129.7785
Recency	12/6/2011 9:56:00 AM	11/24/2011 12:48:00 PM	1/31/2011 3:27:00 PM	12/1/2011 1:47:00 PM	12/2/2011 11:21:00 AM
Frequency	5.3202	4.2466	5.6404	4.8205	5.301
Monetary	2075.6793	1420.2908	2021.7693	1774.929	2148.8492
FirstPurchase	11/28/2011 1:26:00 PM	1/7/2011 12:44:00 PM	1/31/2011 1:17:00 PM	12/6/2010 12:55:00 PM	4/7/2011 12:04:00 PM

4.2. Clusters Explanation

Understanding every cluster in details, is crucially required for customer-centric business intelligence. So now examining the dataset and results of K-Mean algorithm using Weka 3.8, it is observed that each cluster have a group of customers with certain features and parameter values.

In Cluster 1, there are 890 Customers involved. It is composed of 20% of the data. It seems the data in this cluster is just in the last quarter, as the first purchase date is 11/28/2011. Frequency is 5.32, which is above average and the monetary 2075.67, is also above average. This cluster will not be considered for further

analysis as the time span of transaction is very small, just 2 months. It means that the customers in this cluster didn't shop too much. The customers in this span may be newly registered customers and started shopping recently.

In Cluster 2, there are 446 customers involved. It is composed of 10% of whole data. It is observed that the data here is for a long time period from 1/7/2011 to 11/24/2011, which is sufficient to be considered. Still there are some factors such as the frequency here is very low which is 4.24 and monetary is also the lowest among all clusters. So, it is not a feasible cluster for further analysis, as the average case is always selected for analysis. This group of customers, is not profitable because of lowest monetary value and the customers also visit the stores seldom.

In Cluster 3, there are 431 customers involved. It is also composed of 10% of total data. It is clearly observed that the data in this cluster is just first quarter of time span. Monetary value 2021.76 is feasible here, as it is above average. Although the frequency 5.64, is the highest frequency than all of rest clusters, but still the time span is very short. It shows that the customers in this group, were very frequent customers and generated a feasible business value, stopped shopping from this store further or might be just occasionally customers, who needs to shop on some occasion for specific time period.

In Cluster 5, there are 867 customers involved. The time period here is a good long span starts from 4/7/2011 to 12/2/2011. Monetary value 2148.8492 is the best. Frequency 5.3 is also above average. On the other hand, the observations for Cluster 4 are also sufficient. In cluster 4, there are maximum number (1738[40% of data]) of customers involved. The period starts from 12/6/2010 until 12/1/2011, it means this cluster contains the transactions from whole time span. Although the frequency 4.82 and the monetary 1774.92, are just average factors as compared to Cluster 5, but still this is the best option to be considered.

The Cluster 4 has average cases and maximum number of customers within maximum time span. Overall, the firm appears to be fairly profitable.

All things considered, the clusters found using K-Means provide insightful analysis of consumer segmentation. For long-term retention plans, Cluster 4 offers a consistent and devoted clientele, perfect for Given its great financial worth, Cluster 5 might be focused on with luxury product offers or loyalty incentives. Cluster 2 consists of low-spending, infrequent consumers suggesting a chance for re-engagement campaigns or promotional offers. Being connected to shorter activity periods, clusters 1 and 3 could represent seasonal or freshly acquired clients and should be kept under close observation for possible retention or turnover. These divisions let marketing decisions and client relationship management be more targeted and successful.

4.3. Further Analysis

As explained above, the cluster 4 is most feasible and diverse cluster among all these 5 identified clusters, as it contains the maximum number of newly registered and old customers within the maximum time span for dataset, having average and suitable frequency and monetary values. It can be further analysed using other data-mining techniques. One technique is Classification rule mining; it can be implemented on this result set and the customers can be classified into sub categories using some classification algorithm. Another approach to implement the decision tree for further analysis [12]. In decision tree method, Customers can be divided in sub clusters using some parameters e.g., on basis of frequency data can be further divided into sub-categories.

Association rule mining algorithm such as Apriori, can be implemented for these customers and the shopping behaviour can be identified. Weka 3.8 conducted the Apriori algorithm at a minimum confidence level of 0.5 and a minimum support threshold of 0.01—that is, 1% of all transactions. These values were selected to provide a balance between rule relevance and computing practicality, therefore enabling the identification of significant yet non-trivial association rules. Apriori can be implemented to find out the relationship between purchased products. It identifies the association rule between products and it can be easily identified that which product was purchased with other product.

As it is known that there are too many issues with association rule mining implementation. For example, the primary issue is the creation of an excessive number of repetitive regulations. While the data mining community has attempted to solve this issue, research is currently continuing. The second problem is about the "interestingness" of regulations. Association rule mining generates a large number of trivial rules that the user already knows.

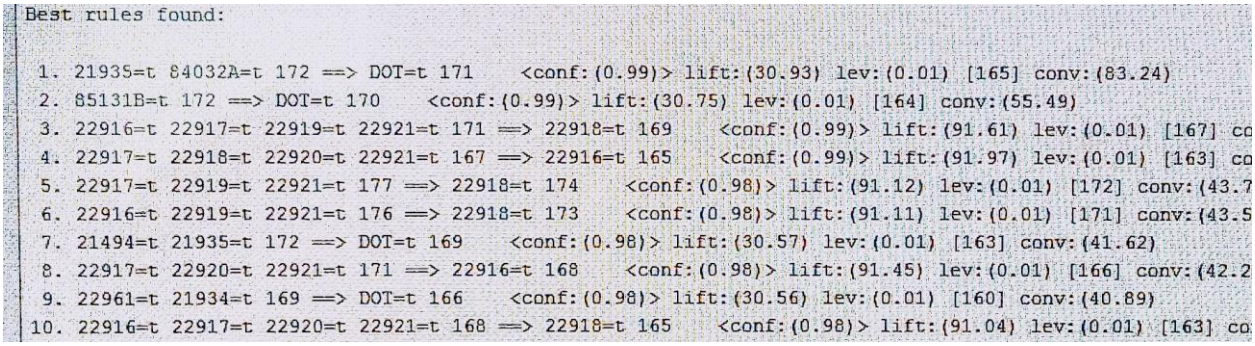

Figure 5: Apriori Implementation using Weka 3.8

These rules frequently distract people from recognizing rules that are both intriguing and beneficial. Finding intriguing association rules is a popular issue in data mining research. The third difficulty is the lengthy computation time necessary to identify huge item-set patterns. To address this issue, several attempts have been made to design more efficient algorithms or use sampling techniques to limit the quantity of data that must be processed.[1]

So, this cluster is further analysed and Apriori using Weka 3.8, has been implemented on the data which this cluster showed. In figure 5, some of the identified Association rules are shown. For example, the association rule # 3 is showing that the Product ID (StockCode) 22981 was purchases 169 times when the customer purchased 22916,22917,22919 and 22921 together for 171 times. Another example is the association rule # 8, where the Product 22916 was purchased for 168 times when customer bought 22917, 22920 and 22921 altogether for 171 times. There are number of rules are identified and it is observed that these rules are very useful for future sales forecasting and basket prediction for customers.

5. Conclusions

This paper shows different data-mining techniques and algorithms. Once these clustering and apriori techniques have been applied to sales data, they can help disclose intriguing information about customers, goods, and sales trends, so contributing to competitive business intelligence. For example, the discovery of association rules can result in higher-level sales forecasting and more cautious inventory control. The data may also be utilized to optimize pricing.

This article demonstrates how to develop customer-centric business information for the market using data mining techniques. The unique customer groups described in this article can assist businesses in better understanding their customers' profitability and, as a result, developing suitable marketing tactics for different consumers. Association rule mining improves sales forecasting and reveals consumer buying patterns and interests, which may be useful in customer-centric marketing and advertising.

This investigation has demonstrated that the two most important and time-consuming processes in the data mining process are data preparation and model interpretation and assessment.

Among the several methods used, the K-Means clustering method based on RFM model turned out to be the most successful for grouping consumers according on behavioural trends. With Cluster 4 recognized as the most stable and varied group, with the biggest client base over the longest time span, the five cluster solution offered the most significant segmentation. The Apriori technique was then used to examine this cluster further and found product useful association rules. These techniques taken together provided a strong

means of revealing useful insights for strategies of client retention, inventory control, and focused marketing.

6. Future Concerns

Further research for the business includes: conducting association analysis to establish customer buying patterns in terms of which products have been purchased frequently by which customers and which customer groups; improving the merchant's stores to allow a consumer's shopping activities to be captured and tracked instantly and accurately; and predicting each customer's lifecycle value to quantify the level of diversity of each customer.

The results can help in future research areas where multi label classification is required or the prediction about future basket of each customer should be analyzed. These clustering and association rule mining can be helpful in implementation of different neural networks such as sequence to sequence or LSTM recurrent neural network approaches for prediction of baskets and sales forecasting.

This is the need of today's marketing and advertisements, that customer and their shopping behaviors should be focused and analyzed. Large scale business has already adopted many datamining approaches to achieve these marketing goals. Now small and medium sized organization are also focusing on these strategies. So, this is a large and rich research area.

Funding Statement: Author has received no funding.

Conflicts of Interest: Author has no conflicts of interest to declare.

Data Availability: The dataset used on this paper is available publicly and properly referenced.

References

[1] Tan, Swee Chuan, and Jess Pei San Lau. "Time series clustering: A superior alternative for market basket analysis." In *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, pp. 241-248. Singapore: Springer Singapore, 2013.

[2] Sarma, Hiren Kumar Deva, and Swapnil Mishra. "Mining time series data with Apriori tid algorithm." In *2016 International Conference on Information Technology (ICIT)*, pp. 160-164. IEEE, 2016.

[3] Liu, Bing, Wynne Hsu, and Yiming Ma. "Integrating classification and association rule mining." In *Proceedings of the fourth international conference on knowledge discovery and data mining*, pp. 80-86. 1998.

[4] Lipton, Zachary C., David C. Kale, Charles Elkan, and Randall Wetzel. "Learning to diagnose with LSTM recurrent neural networks." *arXiv preprint arXiv:1511.03677* (2015).

[5] Cumby, Chad, Andrew Fano, Rayid Ghani, and Marko Krema. "Building intelligent shopping assistants using individual consumer models." In *Proceedings of the 10th international conference on Intelligent user interfaces*, pp. 323-325. 2005.

[6] Cumby, Chad, Andrew Fano, Rayid Ghani, and Marko Krema. "Predicting customer shopping lists from point-of-sale purchase data." In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 402-409. 2004.

[7]Kooti, Farshad, Kristina Lerman, Luca Maria Aiello, Mihajlo Grbovic, Nemanja Djuric, and Vladan Radosavljevic. "Portrait of an online shopper: Understanding and predicting consumer behavior." In *Proceedings of the ninth ACM international conference on web search and data mining*, pp. 205-214. 2016.

[8] Lee, Munyoung, Taehoon Ha, Jinyoung Han, Jong-Youn Rha, and Ted Taekyoung Kwon. "Online footsteps to purchase: Exploring consumer behaviors on online shopping sites." In *Proceedings of the ACM web science conference*, pp. 1-10. 2015.

[9] Shangguan, Longfei, Zimu Zhou, Xiaolong Zheng, Lei Yang, Yunhao Liu, and Jinsong Han. "ShopMiner: Mining customer shopping behavior in physical clothing stores with COTS RFID devices." In *Proceedings of the 13th ACM conference on embedded networked sensor systems*, pp. 113-125. 2015.

[10] Harutyunyan, Hrayr, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. "Multitask learning and benchmarking with clinical time series data." *Scientific data* 6, no. 1 (2019): 96.

[11] Chen, D. *Online Retail II Data Set. UCI Machine Learning Repository*. 2019.

[12] Chen, Daqing, Sai Laing Sain, and Kun Guo. "Data mining for the online retail industry: A case study of RFM

model-based customer segmentation using data mining." *Journal of Database Marketing & Customer Strategy Management* 19 (2012): 197-208.

Machines and Algorithms
http://www.knovell.org/mna

Review Article

Artificial Intelligence in the Education Process

Saleh Almakki[1, *]

[1] Department of information systems, King Faisal university, Al-Ahsa, 31982, Saudi Arabia
[*]Corresponding Author: Saleh Almakki. Email: sal.almakki@gmail.com

Received: 9 August 2024; Revised: 07 September 2024; Accepted: 05 October 2024; Published: 10 October 2024
AID: 003-03-000045

Abstract: Many fields show remarkable progress in the application of technology. Technology in work environments has moved from limited use in accomplishing specific tasks to broader use in accomplishing multi-tasking work in full. This situation may be evident in some fields and may decline in other fields such as education. Can technology perform the education process completely, i.e. manage the education process instead of (or partial of) the teacher, or does it remain limited as functional tools to assist the teacher? In this paper, we will discuss this gap, which is the extent of our need for human interaction and the extent of the efficiency of modern technology in covering this aspect. After discussing the needs and necessity of the human element in the education process within the framework of direct interaction, we will then discuss the reasons for the ongoing controversy on this topic, and then we reach a conclusion that technology can play this role, but we attribute this decision to the goal of education in the environment in which education is applied. The tools and application differ according to the educational goals.

Keywords: Instruction, learning; Education; AI; Business Process; Business Function; Google Scholars;

1. Introduction

Education is one of the most important ways people use to share knowledge, values, and skills. It helps people grow and supports the development of society. We can see how important education is by looking at how it continued in different times and places. Over time, the tools and methods of education have changed. In the past, people used things like leather and ink to write. Later, they used paper and pens. Education also moved into schools and universities.

These changes made new ways of teaching and learning. For example, books helped students' study at home. Teachers used blackboards in the classroom. Today, technology has brought another big change. Students and teachers now use computers and the internet. Because of this, we ask: how is education process today different from the past?

1.1. Education process

We can understand the change by looking at the admission process from 1950 to 2024. In the past, students had to go to the university or send papers by mail to apply. This took a long time and a lot of effort. Today, students use websites to apply. They send their papers online and get answers by email. This is much faster and easier. If technology changed the "process of" students apply, then it also could change "the process of" how they study and learn.

1.2. Technology in education process

Technology has a great contribution in education, starting from creating lessons using presentations and submitting assignments via platforms like Blackboard, to delivering lectures from behind screens. Despite these advancements, technology brought us new challenges. The direct guidance and interaction of the teachers in education has been negatively affected. Efforts like gamified learning and creating engaging content aim to solve these issues, but technology still functions under the guidance of teachers. Even with the abundance of information available online, students rely on educators to guide them in how to search and access resources.

1.3. Why Artificial Intelligence (AI)?

This raises the question of whether technology remains limited in education or whether there is a gap in our education system with technology. Although we did not tend to talk about technological benefits in education, it is clear that education technological tools evolved rapidly. We are referring here to the capability of technology to simulate the student's needs, just as teachers do. And when we talk about technology at this level (like simulation the student's needs), we tend to speak about artificial intelligence (AI) where technology at this level is capable to handle complex tasks and process.

This paper discusses few points regarding to the technology in education ending with the discussion of the gap of interactivity of technology and the ability of AI to handle this gap or not by discussing two questions: what is the difference between "Technology as function" and "Technology as process"? and ₀s technology capable of taking the teacher's role in the learning process?

2. Review Methodology

Education has many criteria such as learning, instructing, and administration. In this paper, we discuss the application of Artificial intelligence specifically in instructing and learning, not education as a whole system.

Keywords for the research include: instruction, learning, AI, ITS and education process. The term "Process" is included since the paper discusses the implementation of AI as a process within learning system.

Papers related to Intelligent Tutoring Systems (ITSs) have been included to assess the impact of AI systems on students and learners. Sources were selected primarily through Google Scholar, using combinations of keywords such as "AI in education," "intelligent tutoring systems," and "technology in learning." Only peer-reviewed papers and academic publications published after 2010 were considered. Articles focusing on unrelated AI applications (e.g., AI in business or logistics) were excluded.

Additional references include [1] and [2], which discuss the application of technology and AI in organizational processes.

A future step in the research may include conducting a survey with a limited sample to compare knowledge gained through AI-based technology versus traditional interactive teaching.

3. Related Work and Research Contribution

There are numerous research papers in the field of AI in education, as both education and AI are rich areas for exploration. Education is deeply embedded in societies and, consequently, has been extensively studied and researched. On the other hand, AI has been advancing and growing and getting interested by many fields because of the features and services that AI could provide. Anyway, the spread of AI systems and their services help to increase research in this field. Figure 1 shows the rising number of papers published in the topics "AI" and "Education" from 2017 to 2021 by Google Scholar engine.

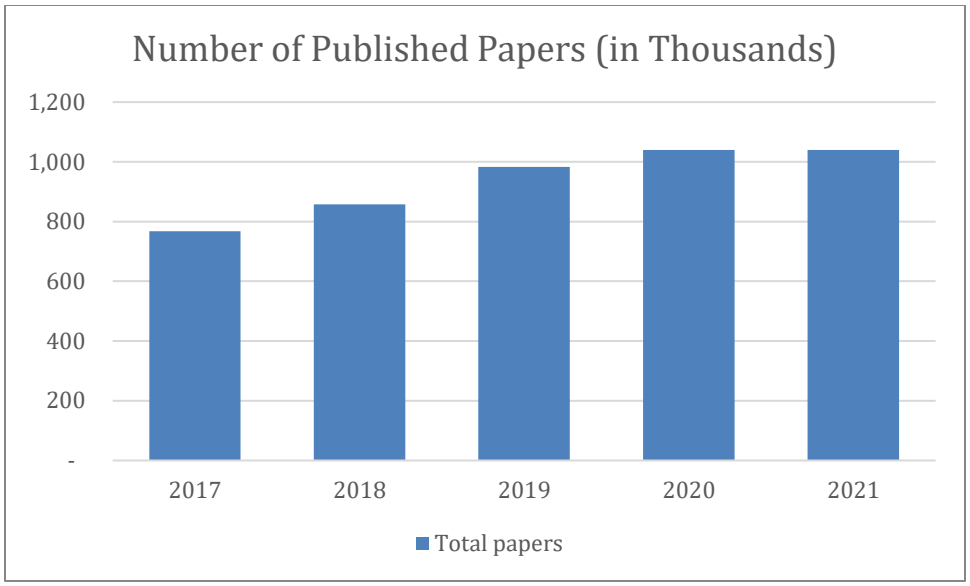

Figure 1: Number of papers published (Search results) in Google Scholar over the past ten years with key words "AI" and "Education"

One example of these systems is Intelligent Tutoring Systems (ITSs) which are AI-based and designed for learning purposes.

The paper *"Artificial Intelligence (AI) in Education: Using AI Tools for Teaching and Learning Process"* by Tira Nur Fitria [3] presents various forms of AI application and discusses these implementations. Another paper, titled *"Towards a Design of an Intelligent Educational System"* by Valentina Terzieva and her colleagues [4], focuses on intelligent education systems (IES), which are AI-based tools designed to support specific field in learning. These and other studies provide an overview of AI applications in education, highlighting AI's potential to perform tasks related to teaching and learning. This literature review aims to explore the development of technology and its capability and feasibility of applying these applications in the educational institutes as organizations.

Since this paper focuses on the capability of the application of technology (particularly AI) in the education process, we will benefit from the reviewed papers in aspects related to AI applications. As for the use of technology in systems, including education, I refer to the studies [1] and [2] to discuss how technology or AI is applied as a process in organizations.

The second part of knowledge construction for this literature review, is understanding organization process which is secondary as this paper focuses mainly on the impact in education and the capability of AI in this field.

4. Literature Review

4.1. Technology from functions to processes

Technology has gone through different eras, and the way these eras are divided can vary depending on the point of view. In the paper "History, Features, Challenges, and Critical Success Factors of Enterprise Resource Planning (ERP) in The Era of Industry 4.0", they divide technology into four eras based on the development of technological systems in the context of ERP [5]. ERP is a new era of functionality where tasks and functions. A similar classification concept is also mentioned by a paper titled Enterprise Resource Planning: Past, Present, and Future by Shadrack Katuu [6] but with more range of years. However, Rainer et al. in their book [7] "Introduction to Information System 5th edition" do not follow this classification due to a different research focus, the authors mention two earlier eras called the "functional eras" in information

systems. There are also two common terms in the business field that match this classification of technology: Business function and Business Process.

In the “Business function”, technology is limited to interior departments to perform their job within their scope such as record entries or making reports using spread sheets for financial department.

Later, technology developed to a higher level, where it could handle multiple tasks within one process. At this point, technology became part of an integrated system which led to ERP systems. ERP system is a new considered to be a new era of technology that implement the concept of “Business Process”. Ellen Monk and Bret Wagner [8] in their book “Concepts in Enterprise Resource Planning” have good example explaining the relationship between these two concepts (function and process). The book shows how a group of tasks are connected in two integrated processes. The first process represents a customer order process, starting from the sales function to the logistics function. The second process represents a material order process, which use functions in different way. A function might be shared by both processes but the implementation within the process defers. Therefore, we understand from the figure 2 that process is not only collection of functions but also a method of implementation.

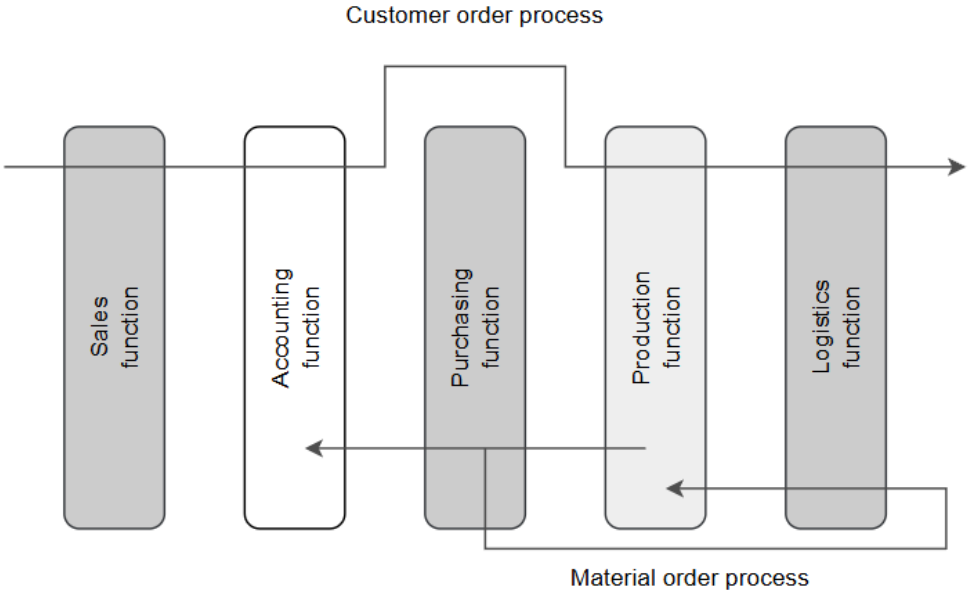

Figure 2: A process view of business operations from Ellen Monk and Bret Wagner (edited version) [8]

In the second chapter, the author explains ERP systems as one of the technological models that helped companies in the past adopt the concept of business process. In the same section, he also highlights the technological limitations of the 1970s and 1980s, and how advancements in data storage contributed to overcoming those limitations.

This demonstrates the clear impact of technology and its ability to perform at a much higher level than in the past. From here we would like to point out that the development of technology to the stage of work as a process and its capability to handle a complete job does not necessarily mean dispensing with humans, but rather it is the entry of technology into a new stage of managing work.

4.2. Application of technology in education

Technology today is widely used. As mentioned earlier, it became part of the organization's process. Platforms like Blackboard and Edmodo are examples of having technology as a core work of their business.

Educators can use these services as a primary piece of their course. Google classrooms is also a great tool where it offers creating embedded documents, so learners do not need to set up word processing apps to write their essays and store them locally and upload them later to the platform.

Same as other fields, the technology in education is not just about improving teacher tools, but it has expanded to some administrative aspect of administrating the learning that even changed the way education looks. For example, designing some courses to be delivered remotely, like the courses on platforms such as Udemy and Udacity. The spread of technology did not stop at this point, there are other shapes and styles of education that show how technology applications are growing in this field such as online learning, e-learning, web-based training, and computer-based training. However, these are general examples not to be discussed in detail in this paper, but it is just mentioned as examples of the technology applications in the education field.

These services are capable of processing particular jobs. They can be used as tools to support instruction, but they are limited to processing submissions and contents upload and download. Other educational tasks are performed by other services or by the educator themselves such as preparing or grading quizzes and homework. Some tasks are more complex such as understanding learners' needs. AI is the capable part of technology to perform such complex tasks.

Fati Tahiru et al. in their review of "AI in Education" [9] identify three types of AI technologies, one of them is Automating Administrative Tasks where it takes the role to grade quizzes and homework.

Tira Nur Fitria, in her paper "*Artificial Intelligence (AI) In Education: Using AI Tools for Teaching and Learning Process*" also [3] highlights various types of AI technologies that can perform different tasks, such as:

- Virtual Mentor: virtual mentor in online education by giving students feedback and helping them review materials.
- Voice Assistant: similar to virtual mentors but focus on voice interaction. They help students find learning materials easily by speaking keywords, and they respond using natural language.
- Smart Content: helps users find, organize, and access digital books and learning materials quickly and easily.
- Intelligent tutoring system (ITS): aims to provide personalized education and feedback.

The paper "*Towards a Design of an Intelligent Educational System*" [4] also discuss a later version of ITSs, known as the Intelligent Educational System (IES). The IES is a wider concept of ITS where "*students interact with interfaces that are customizable and personalize the learning experience based on their preferences and current learning status*".

4.3. AI as a learning process

We discussed the difference between technology as a "function" and technology as a "process" in term of business, and how technology has evolved from simply performing specific tasks (functions) to managing entire processes. For example, in the field of commerce, technology used to serve limited roles such as accounting (like using a cashier system) or inventory tracking (like using MS Excel), especially in traditional stores. However, technology has now advanced to the point where it can transform the entire store into a digital platform, handling the full sales process—from displaying products, to completing transactions, to managing inventory—just like what we see with modern e-commerce websites.

In education, there are good attempts to benefit from technology. However, education is one of the fields where human interaction plays a big role; therefore, it might face more challenges in benefiting from technology compared to commerce or many other fields. Web pages with some technologies to store and process data might be good tools for performing processes like selling and buying, but these tools are limited in matching and imitating human intelligence in interactions, evaluations, or managing the deep learning process that aligns with the learner's thinking. However, with the help of AI, websites like Exercism.org and other systems such as ITSs can now imitate human intelligence and understand educational needs. They

can even handle a “complete process", where AI in these systems provides personalized and adaptive instruction for students [10].

However, intelligent systems in education are not necessarily dedicated to tutoring learners, but they can be a partner to the educators in providing better experience in their field. AI could be implemented in different areas level of integration. This is at least a good stage in involving AI in the educational process. From here it becomes clear that AI can be applied and benefited from in the field of education on more than one level: full accreditation as is the case with ITSs, or partial accreditation as a tool to assist teachers, which requires human supervision in the educational process.

4.4. The need for human interaction in education

When we talk about direct human interaction in education, it goes beyond just knowledge. In his review in the paper *Artificial Intelligence in Higher Education* [11], Sana Abu Safi Al-Qudah explains that universities have a role that goes beyond traditional education. They play dynamic, multi-dimensional roles, acting as channels to develop social skills, communication abilities, interaction with others, emotional intelligence, ethical and cultural values, and a sense of personal responsibility. This is achieved through involvement in cooperative learning experiences and diverse educational environments.

This shows that even with advanced AI solutions, providing good learning systems, may still fall short of the full purpose of direct interaction between teachers and students in term of social experience. These AI systems are primarily focused on learning in terms of learning skills and knowledge delivery. When reviewing papers on Intelligent Tutoring Systems (ITS), we see that they mostly discuss learning outcomes and the effects of learning on students. To understand the scope of these intelligent educational systems, we can read more int the article *Intelligent Tutoring System* by HandWiki [12], as well as the research paper *Artificial intelligence (AI) in education: Using AI tools for teaching and learning process.* [3].

5. Results

The involvement of AI in education, technology tools has advanced more. ITS systems, which is AI-based systems, can be combined with other educational systems to take on bigger roles in education. With ERP (Enterprise Resource Planning) and ITS systems, they can play an integrated role by connecting systems, collecting and analyzing data, and then understanding the learner’s needs to create educational materials based on these needs. This way, technology can play a more comprehensive role in the educational process.

The need for human interaction might go beyond the delivery of knowledge where social and other subjects might be involved in education. However, this field is wider than the goal of AI systems like ITS systems. AI learning systems aim to measure student needs and enhance learning outcomes.

6. Discussion

In the introduction of the research, three main points were presented for discussion: procedural work (process), technology, and AI. The first point was about the concept of the procedural work (Business process), which is a broad topic that includes both business and technical aspects. This was a starting point to talk about technology, which became the most important focus of this research. Then, we narrowed the discussion to AI to better explore the research questions.

At the beginning, two questions were raised: the first was about the difference between technology as a function and as a process. This question led to a discussion of two business-related management concepts (business function and business process), but as we explained, the purpose was not to study business fields in general, but rather to look at technological progress in managing work processes inside organizations. Education, like any other field, works within an organizational and administrative context. So, in this paper, we claim that technology in education can go more just functions to perform processes. One of these processes is educating.

In other words, when we say "educational process" we mean the interaction between the instructor or teacher and the student, not the whole education system. Previous examples (like customer order and material order processes, the banker and transferring money) all show that the concept "process" means doing multiple tasks inside a system, not managing the whole system.

Looking at how technology can manage the teaching process brings us to the second research question: Is technology capable of taking the teacher's role in the learning process? This question introduces many challenges, including the technological ability to interact while delivering information. During the COVID-19 pandemic, tools for online learning often missed this important interactive role, which showed us how important teachers are in the teaching process.

However, from our point of view, the real problem is not the absence of the teacher, but the lack of their interactive role. So, we think the main difference between face-to-face learning and online learning is the absence of "interaction". After reviewing the capabilities of AI systems (like identifying student needs and offering simulations), we think that technology today has reached a level close to human teachers in interacting with information, or at least in a big part of it.

This leads us to another challenge: will technology replace teachers? This may come to mind after reading the second question of the research, but it's a misunderstanding. This paper is not talking about the risks of losing teaching jobs. We emphasize repeatedly that our discussion about the possibility of technology taking the teacher's role is actually a discussion about its effectiveness. The ability of technology to take on the role of the teacher does not necessarily mean eliminating the teacher's role. Just as the expansion of technology in the banking sector has enabled it to perform tasks such as transfers and deposits (tasks typically done by bank employees) this has not necessarily led to the elimination of banking staff.

Returning to the ability of technology to take the teacher's role, we may enter into a broader comparison with the teacher's role itself. This also raises a number of questions and discussions, including social and psychological topics, as mentioned in the paper. However, the discussion here focuses on the aspect of teaching and learning (learning process)-that is, the transfer of knowledge. After this review, we argue that it is possible to compare the teacher and technology in this context, given that AI programs specialize in measuring and identifying learners' needs.

7. Conclusion

There are two important parts in this research review. The first aspect discusses how technology is applied in organizations, where it was limited to the scope of "Business Function", and then developed to the scope of "Business Process". This workflow or framework can be applied in organizations based on their needs. However, in education field, it may become more complex because of the presence of human elements that common technological tools cannot address, such as the interaction between the teacher and the student and assessing the student's needs.

The second aspect revolves around AI and its capabilities in simulating human intelligence in term of students' needs. There are already existing AI applications, such as ITS systems, which are AI-based applications related to education.

These two aspects support the idea of developing the management of the education process using AI tools, allowing us to transform AI from being as "education function" to "education process" that work as a central part in managing and guiding the educational process between the teacher and the student.

In the end, it is essential to recognize the gap in comparison between our need for human interaction and technological interaction. Human interaction in education may go beyond just information and knowledge, while technology, especially ITS systems, are primarily designed for that purpose. By understanding the limits of our need for interaction, we can assess the interaction between technology and human interaction more fairly.

Funding Statement: Author has received no funding.

Conflicts of Interest: Author has no conflicts of interest to declare.

Data Availability: This is a review article, no new data has been generated in this study.

References

[1] Karyy, Oleh, Ihor Novakivskyi, Yaroslav Kis, Ihor Kulyniak, and Alexander Adamovsky. "Model of Educational Process Organizing Using Artificial Intelligence Technologies." In *COLINS (3)*, pp. 332-347. 2023.

[2] Zeebaree, Mosleh, Goran Yousif Ismael, Omar A. Nakshabandi, Saman Sattar Saleh, and Musbah Aqel. "Impact of innovation technology in enhancing organizational management." *Studies of Applied Economics* 38, no. 4 (2020).

[3] Fitria, Tira Nur. "Artificial intelligence (AI) in education: Using AI tools for teaching and learning process." In *Prosiding Seminar Nasional & Call for Paper STIE AAS*, pp. 134-147. 2021.

[4] Terzieva, Valentina, Svetozar Ilchev, Katia Todorova, and Rumen Andreev. "Towards a design of an intelligent educational system." *IFAC-PapersOnLine* 54, no. 13 (2021): 363-368.

[5] Al-Amin, Md, Tanjim Hossain, Jahidul Islam, and Sanjit Kumar Biwas. "History, features, challenges, and critical success factors of enterprise resource planning (ERP) in the era of industry 4.0." *European Scientific Journal, ESJ* 19, no. 6 (2023): 31.

[6] Katuu, Shadrack. "Enterprise resource planning: past, present, and future." *New Review of Information Networking* 25, no. 1 (2020): 37-46.

[7] Turban, Efraim, R. Kelly Rainer, and Richard E. Potter. *Introduction to Information Systems: Supporting and Transforming Business*. John Wiley & Sons, Inc., 2007.

[8] Monk, Ellen F., and Bret J. Wagner. *Concepts in enterprise resource planning*. Course Technology, Cengage Learning, 2013.

[9] Tahiru, Fati. "AI in education: A systematic literature review." *Journal of Cases on Information Technology (JCIT)* 23, no. 1 (2021): 1-20.

[10] Lin, Chien-Chang, Anna YQ Huang, and Owen HT Lu. "Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review." *Smart Learning Environments* 10, no. 1 (2023): 41.

[11] Safi, Sana'a & Al-Qudah, Mohammed. (2024). Artificial Intelligence in Higher Education (Challenges and Guidelines) – A Systematic Review. 51. 201-0217. 10.35516/edu.v51i3.73. Available: https://dsr.ju.edu.jo/djournals/index.php/Edu/article/view/7303/1933

[12] HandWiki, "Intelligent Tutoring System", *Scholarly Community Encyclopedia*, 2022, [Online]. Available: https://encyclopedia.pub/entry/28717