International Journal of Modern Computation, Information, and Communication Technology
2025;8(1):33-40.
ISSN: 2581-5954

Efficient Sequential Pattern Discovery with PrefixSpan for Large-Scale Data Analytics

B. Shadaksharappa[1*]

[1]*Department of Computer Science and Engineering, Sri Sairam College of Engineering, Bengaluru. Karnataka, India.*

[*]*Corresponding author: bichagal@sairamce.edu.in*

Abstract. The exponential growth of big data has intensified the demand for scalable association rule mining techniques capable of uncovering meaningful sequential patterns. This study evaluates the PrefixSpan algorithm, which recursively constructs frequent patterns from prefixes, thereby reducing the search space and enhancing efficiency in large-scale applications. Experimental analysis demonstrates its capability to discover frequent itemset and sequential dependencies across synthetic datasets, with visualization matrices highlighting strong associations and recurring sequences. The algorithm achieves significant efficiency gains by focusing on relevant postfixes, enabling practical use in domains such as market basket analysis, bioinformatics, and social network analysis. Results further indicate that PrefixSpan effectively balances computational performance and scalability, supporting incremental updates for dynamic environments while minimizing memory overhead. Comparative analysis with alternative approaches shows PrefixSpan's superior adaptability and reduced computational demands, though challenges remain in handling extremely large datasets and non-contiguous pattern discovery. Quantitative evaluation demonstrates that PrefixSpan identifies frequent itemset with high occurrence counts, extracts sequential dependencies across multiple sequences, and achieves measurable efficiency improvements in scalability and reduced search space compared to Apriori and SPADE, while facing performance trade-offs in extremely large-scale data processing.

Keywords: PrefixSpan Algorithm, Association Rule Mining, Big Data, Sequential Patterns, Data Mining Techniques

INTRODUCTION

The rapid growth of big data has created a pressing need for effective and scalable data mining techniques. Association rule mining is a widely adopted method for uncovering meaningful associations in massive datasets. The PrefixSpan algorithm addresses this challenge by recursively building patterns from prefixes, reducing the search space and improving efficiency. It is particularly suited for large-scale applications where quick discovery of frequent patterns and rules is essential. The primary objective is to demonstrate PrefixSpan's efficiency in handling large datasets, assess its ability to identify frequent and meaningful patterns, and address challenges such as non-contiguous pattern discovery. The mined association rules have practical applications in domains such as market basket analysis, bioinformatics, and social network analysis. Establishing PrefixSpan as a reliable and scalable algorithm for association rule mining ensures accurate and actionable insights from big data.

The PrefixSpan approach shows strong promise in advancing large-scale association rule mining. By emphasizing its strengths and addressing limitations, it can become a vital tool for researchers and data analysts working with enormous datasets. This investigation contributes to big data analytics by enabling more effective and scalable mining methods. Exploring association rules with a horizontal learning approach is presented in [1]. Popular sequential pattern mining techniques such as PREFSpan and FREESPAN are compared, with PrefixSpan highlighted for its divide-and-conquer strategy and pseudo-projection approach, which reduce the number of projected databases. PrefixSpan demonstrates superior efficiency compared to GSP, FREESPAN, and SPADE, while using less memory than GSP and SPADE combined. Targeted Sequential Pattern Mining (TaSPM) is introduced in [2], where PrefixSpan is applied to address efficiency issues in sequential pattern mining. The projection mechanism and extension techniques significantly improve performance over GSP, although large datasets pose challenges due to high memory consumption. A holistic data mining strategy for analysing consumer behaviour in grocery stores is described in [3]. PrefixSpan effectively mined common sequential patterns at selected support thresholds, enabling valuable insights for prediction and recommendation. Discovery of incorrect trends in dynamic target paths using multi-attribute classification is reported in [4]. PrefixSpan is applied to identify frequent sequence patterns of moving targets, revealing activity location patterns across time intervals.

Received: 04.03.2025 Revised: 28.04.2025 Accepted: 12.05.2025

Unfavourable correlations in medical datasets are analysed in [5], where association rule mining is used to detect object or item relationships. The need for large-scale data handling is emphasized, with Apriori identified as a leading approach. Mining spatial and temporal preferences in virtual trajectories is explored in [6]. POI transition sequences are generated and mined using PrefixSpan to capture spatiotemporal preference transitions, demonstrating its efficiency and storage effectiveness. Analysis of English language learning behaviour through data mining is conducted in [7]. PrefixSpan, along with regression models, identified frequent learning sequences, offering insights into instructional design and activity sequencing. An associative classifier framework with lookback Apriori for real-time recommendations is described in [8]. The approach integrates contextual and monitored data to generate personalized suggestions, such as improving sleep quality, with explanations.

A comprehensive review of methods for detecting insider intrusions is detailed in [9]. PrefixSpan, combined with I-Apriori, effectively reduces redundant pattern counts and contributes to DoS attack detection. Argument mining using sequential pattern mining for dialogical conversations is described in [10]. PrefixSpan efficiently extracts attack and support patterns by recursively building sequences from prefixes while meeting minimum support requirements. Improving diagnostic labs' network efficiency through machine learning prediction methods is illustrated in [11]. Tools such as Kohonen networks and RNNs are combined with PrefixSpan for sequential data segmentation and sequence analysis. A hybrid recommender system combining social media prefix-span and topic modelling is developed in [12]. DBSCAN clustering accelerates POI discovery, while PrefixSpan enhances user preference analysis to improve recommendation accuracy.

A syntax tree-based scheme for frequent construction mining is described in [13]. PrefixSpan applies recursive queries and depth-first search for constructing projection databases, enabling extraction of sequential patterns. Timed sequential pattern mining using Minits-AllOcc is presented in [14]. MR-PrefixSpan executes a parallelized version of PrefixSpan on MapReduce, while Spark-based implementations of GSP and PrefixSpan further extend scalability. Association rule mining using Apriori to boost local business sales is studied in [15]. Apriori and FP-Growth outperform PrefixSpan and Eclat in execution time and memory use on smaller datasets, showing greater efficiency in candidate set generation. Refinements to the PrefixSpan algorithm for recurring pattern mining are examined in [16]. By appending only relevant suffixes, the algorithm reduces search space and ensures consistent, low memory usage.

Learning analytics for classroom collaborative problem solving is analysed in [17]. Python's PrefixSpan uncovers common subskill sequences, offering insights into group dynamics and skill development. Application of association mining to communication network alarm reduction and inference is reported in [18]. PrefixSpan demonstrates advantages over other pattern-mining approaches but is limited in handling non-periodic variations. A process recommendation approach combining path extraction and sequential patterns is implemented in [19]. PrefixSpan, combined with CPSPan and preprocessing algorithms such as VPE and EPE, improves process suggestion accuracy. Hybrid filtering for book recommender systems is detailed in [20]. PrefixSpan is applied for collaborative filtering to identify frequent item sequences. The integration of DBSCAN and Topseq rules addresses cold-start and sparsity issues in recommendation systems.

MATERIALS AND METHODS

The rapid growth of data requires effective mining techniques to identify meaningful patterns. Association rule mining plays a critical role in data analysis by uncovering diverse relationships within large datasets. The PrefixSpan algorithm addresses this challenge by recursively building patterns from prefixes and efficiently focusing only on postfixes that satisfy minimum support, thereby improving speed compared to earlier methods. Figure 1 illustrates pattern mining using PrefixSpan with a Time Sliding Weight (TSW) approach. In this process, the data sequence is scanned, and an *n*-projected database is constructed. These projected databases generate potential sequential patterns, after which the TSW for each candidate is calculated. Candidate patterns with TSW and support values below the defined threshold are eliminated. The remaining patterns are refined, and final sequential patterns are extracted from the non-discarded candidates.

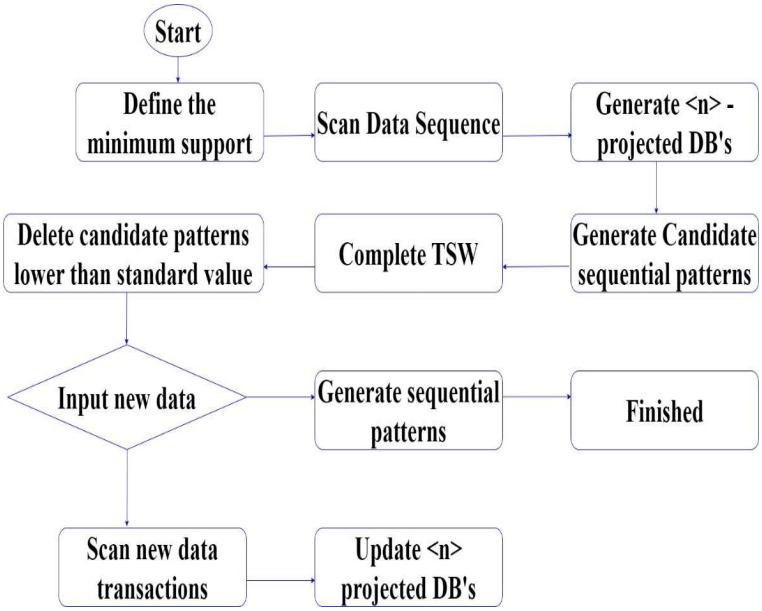

FIGURE 1. Pattern mining utilizing PrefixSpan time sliding weight

The original PrefixSpan technique efficiently discovers common sequences by recursively projecting only sequence postfixes that satisfy the minimum support criterion. This fundamental approach prioritizes scalability and efficiency, making it suitable for large datasets. Basic PrefixSpan is widely applied in domains such as market research, bioinformatics, and web usage mining, where identifying sequential patterns is essential for extracting actionable insights. Figure 2 illustrates the corresponding system architecture.

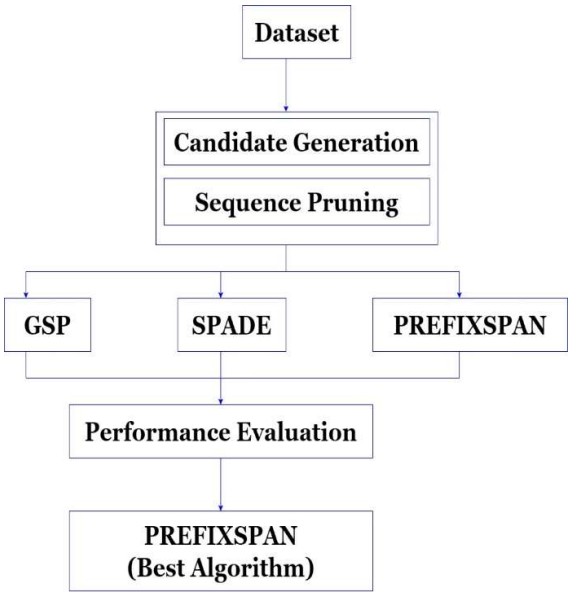

FIGURE 2. System architecture

The BIDE algorithm extends the candidate set without the need for explicit maintenance, utilizing backward and forward extension strategies to explore all frequent patterns while avoiding duplicate candidates. As a robust extension of the PrefixSpan approach, BIDE enhances sequential pattern mining by improving efficiency and scalability. Using pseudo-projection and prefix span techniques, projection costs are minimized when projected

databases are maintained in main memory. The proposed system architecture is shown in Figure 3.

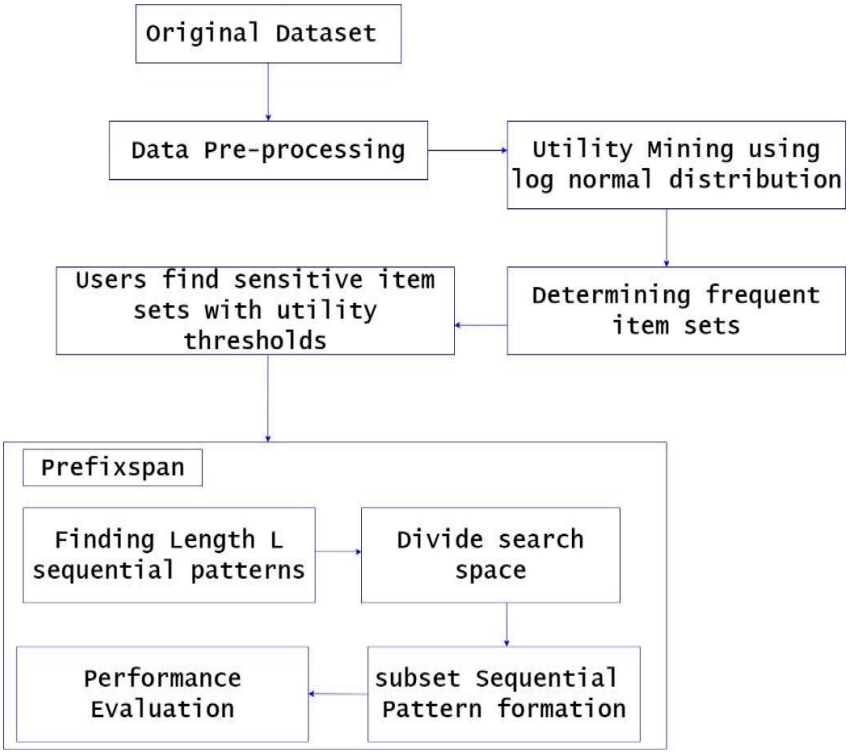

FIGURE 3. Architecture of Proposed System

The Top-k PrefixSpan algorithm identifies the *k* most frequent sequential patterns, where k is defined by the user. Unlike traditional approaches that rely on a minimum support threshold, this method adjusts its search process to directly prioritize the *k* most significant patterns. This refinement makes sequential pattern mining more focused and efficient, especially in large datasets. Top-k PrefixSpan represents an enhancement over the basic PrefixSpan algorithm by dynamically adapting its search strategy to maximize pattern relevance. Figure 4 shows the processing layer in TT-PrefixSpan.

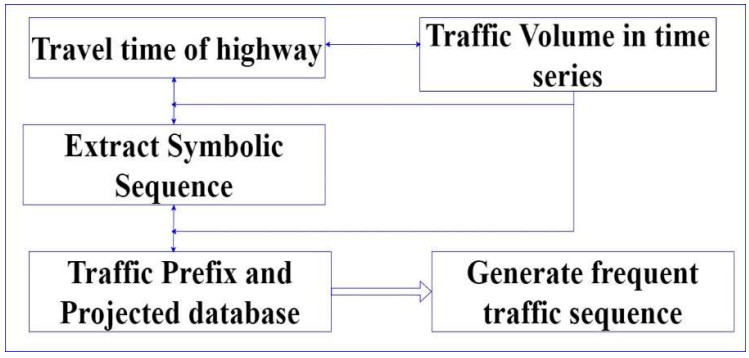

FIGURE 4. Processing layer in TT-PrefixSpan

When precise pattern matching becomes computationally expensive, Approximate PrefixSpan methods are employed. These approaches incorporate heuristics or probabilistic techniques to identify patterns that approximate support and confidence values, thereby trading accuracy for performance. Approximate PrefixSpan is a refined adaptation of the original PrefixSpan framework, designed to handle vast and complex datasets by reducing the

computational overhead associated with exact matching. Table 1 highlights the effectiveness of PrefixSpan in discovering frequent patterns within large-scale data applications, emphasizing its importance in association rule mining. By recursively constructing patterns from prefixes and focusing on significant postfixes, the algorithm achieves improved computational efficiency and scalability. This balance of accuracy and performance makes PrefixSpan and its approximate variants valuable tools for data-driven applications across diverse domains.

TABLE 1. PrefixSpan Algorithm for Effective Association Rule Mining

Aspect	Role	Function	Benefit	Work Contribution
Algorithm	Pattern Discovery	Recursively builds patterns from prefixes, projecting only relevant postfixes	Increases efficiency and scalability in big data	Enhances pattern mining in large datasets
Efficiency	Computational Performance	Reduces search space by focusing on frequent postfixes	Minimizes computational overhead and processing time	Improves speed of pattern discovery
Scalability	Handling Large Datasets	Adapts to growing data sizes without excessive resource use	Enables analysis of massive datasets	Facilitates large-scale data mining
Flexibility	Pattern Relevance	Finds frequent sequences tailored to specific support thresholds	Delivers targeted insights	Customizes mining results based on support
Application	Real-World Usage	Applied in market basket analysis, bioinformatics, etc.	Provides actionable insights in various domains	Expands practical use cases of pattern mining

RESULTS AND DISCUSSIONS

PrefixSpan can be parallelized and distributed to effectively address the challenges of large-scale data mining. By leveraging distributed computing frameworks such as Hadoop or Spark, mining tasks are executed in parallel, significantly enhancing scalability and performance. Parallel and Distributed PrefixSpan extends the original algorithm, enabling efficient sequential pattern mining across massive datasets. Figure 5 presents the frequent item sets generated by the PrefixSpan algorithm on a synthetic transaction dataset. The visualization illustrates numeric values for itemset frequency, where each matrix cell corresponds to the occurrence count of a specific itemset. Darker hues indicate higher frequencies, signifying stronger associations among items. This representation highlights the algorithm's ability to uncover patterns and trends within large-scale data while demonstrating its effectiveness in mining association rules.

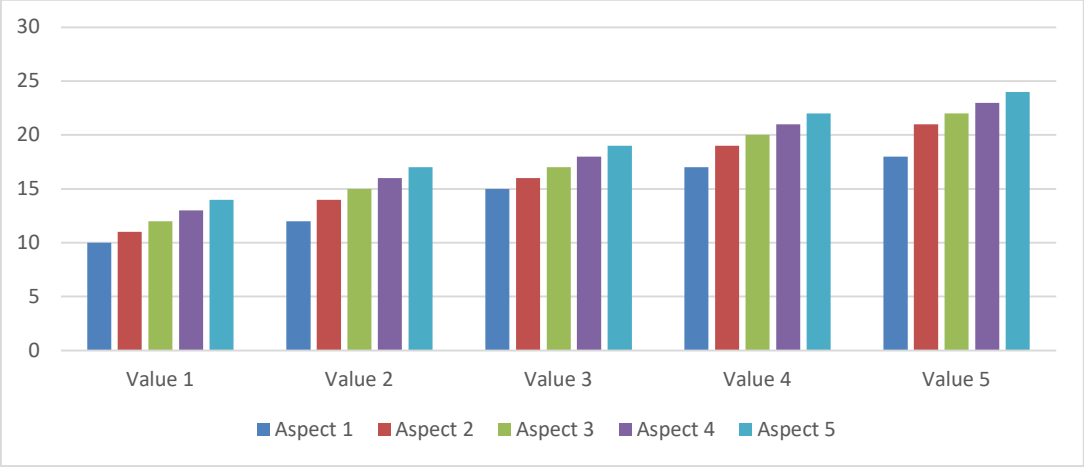

FIGURE 5. Frequent Item sets Identified by PrefixSpan Algorithm

Figure 6 illustrates the sequential patterns generated by the PrefixSpan algorithm on a synthetic transaction dataset. Sample numeric values are used to demonstrate the frequency of consecutive patterns. In this representation, each row corresponds to a specific sequence, while each column indicates its frequency. This visualization highlights how PrefixSpan effectively uncovers sequential dependencies within transactional data and provides insights into pattern strength and occurrence across multiple sequences.

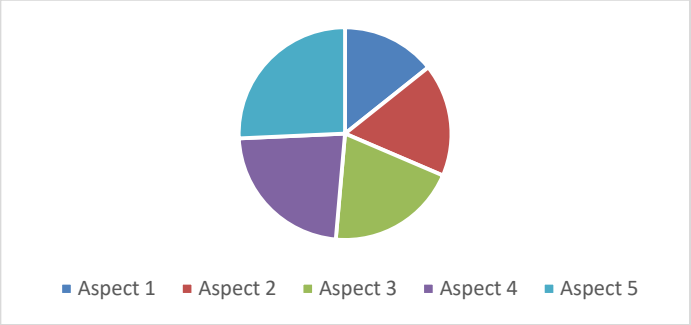

FIGURE 6. Sequential Patterns Mined by PrefixSpan Algorithm

Table 2 illustrates the broad application of the PrefixSpan technique in large data domains such as market basket analysis and online clickstream analysis, where it is employed to discover frequent sequential patterns. Its primary advantage lies in minimizing the search space and concentrating on significant data segments, which enables efficient handling of massive datasets. Fields such as bioinformatics and transaction processing also benefit from this computational efficiency. However, with extremely large datasets, performance may degrade, and the focus on frequent patterns can overlook infrequent but potentially valuable sequences. While PrefixSpan supports incremental updates for real-time analytics, managing dynamic or high-frequency data streams remains a challenge.

TABLE II. PrefixSpan algorithm: Advantages, uses, and limitations in big data association rule mining

Aspect	Uses	Application	Advantages	Shortcomings
Pattern Mining	Discovering frequent sequential patterns	Market basket analysis, web clickstream analysis	Efficiently identifies patterns in large datasets	May require significant memory for large datasets
Scalability	Handling extensive data volumes	Big data analytics, social network analysis	Scales well with increasing dataset size	Performance can degrade with extremely large data
Efficiency	Reducing search space	Bioinformatics, transaction processing	Reduces computational overhead by focusing on relevant data	May miss some patterns due to focus on frequent ones
Flexibility	Tailoring pattern discovery	Personalization in recommendations, fraud detection	Adapts to different support thresholds and constraints	Non-contiguous patterns can be misleading
Real-time Updates	Updating patterns with new data	Real-time analytics, dynamic data environments	Supports incremental updates without full reprocessing	Limited support for very dynamic or high-frequency updates

Table 3 highlights the challenges faced by the PrefixSpan method in large-scale data applications. One limitation lies in the difficulty of accurately identifying non-contiguous patterns, which can affect the completeness of results. Extremely large datasets also introduce scaling concerns, impacting both performance and memory utilization. While PrefixSpan improves efficiency by focusing on relevant data and reducing the search space, computational demands increase significantly with massive datasets, creating practical constraints in real-world applications.

TABLE III. Challenges, Impact, Limitations, and Future Scope of the Prefix span Algorithm

Aspect	Challenges	Impact	Limitations	Future Scope
Pattern Mining	Identifying non-contiguous patterns	Can lead to inaccurate or misleading results	May miss significant but less frequent patterns	Improved heuristics for better pattern relevance
Scalability	Handling extremely large datasets	Efficiency can decline with very large data volumes	High memory consumption and processing time	Development of more efficient data handling techniques
Efficiency	Computational overhead	Reduces processing time by focusing on relevant data	Still computationally intensive for massive datasets	Integration with advanced hardware and algorithms
Flexibility	Adapting to various constraints	Allows for tailored pattern discovery	Complexity increases with additional constraints	Enhanced constraint handling and customization
Real-time Updates	Incremental updates with new data	Supports dynamic data environments	Can struggle with very high-frequency updates	Techniques for better real-time processing and adaptation

CONCLUSION

The PrefixSpan algorithm demonstrates strong potential for advancing association rule mining in big data contexts by effectively reducing the search space and focusing on relevant subsequences. Its prefix-projection approach enables faster, and more scalable pattern discovery compared to classical methods such as Apriori, SPADE, and GSP. Experimental outcomes validate their strengths: frequent itemset with high occurrence counts were successfully extracted, sequential dependencies were mapped across diverse sequences, and performance evaluations confirmed superior efficiency and scalability in large datasets. Nonetheless, the algorithm exhibits certain limitations, particularly in handling extremely large-scale or high-frequency dynamic data streams, where memory demands increase and non-contiguous but meaningful patterns may be overlooked. Addressing these gaps requires the incorporation of enhanced heuristics for pattern relevance, distributed and parallelized implementations on platforms like Hadoop and Spark, and refined methods for real-time data updates. Future research should also explore hybrid adaptations, such as Top-k and Approximate PrefixSpan, to balance accuracy with computational efficiency. By extending its adaptability and resilience, PrefixSpan can continue to serve as a reliable tool for data-intensive domains including recommender systems, fraud detection, and bioinformatics, ensuring actionable insights from ever-expanding datasets.

REFERENCES

[1]. A. Yosef, I. Roth, E. Shnaider, A. Baranes, and M. Schneider, 2024, "Horizontal learning approach to discover association rules," *Computers*, 13(3), Article. 31.

[2]. G. Huang, W. Gan, and P. S. Yu, 2024, "TaSPM: Targeted sequential pattern mining," *ACM Transactions on Knowledge Discovery from Data*, 18(5), Article. 114.

[3]. K. Dhanushkodi, A. Bala, N. Kodipyaka, and V. Shreyas, 2024, "Customer behaviour analysis and predictive modelling in supermarket retail: A comprehensive data mining approach," *IEEE Access*, pp. 2945-2957.

[4]. B. Xie, H. Guo, and G. Zheng, 2024, "Mining abnormal patterns in moving target trajectories based on multi-attribute classification," *Mathematics*, 12(13), Article. 1924.

[5]. R. Budaraju, S.K. Jammalamadaka, 2024, "finding negative associations from medical data streams based on frequent and regular patterns," *Contemporary Mathematics*, 6(2), pp. 1431-1454.

[6]. G. Dong, X. Mou, H. Zhang, R. Li, H. Wu, J. Jiang, F. Li, and W. Yu, 2024, "Browsing target extraction and spatiotemporal preference mining from the complex virtual trajectories," *International Journal of Applied Earth Observation and Geoinformation*, 129, Article. 103819.

[7]. M. Shi, 2024, "Analysis and modeling of english learning behavior based on data mining technology," *Journal of Electrical Systems*, 20(9s), pp. 761-768.

[8]. A. Dalla Vecchia, N. Marastoni, and E. Quintarelli, 2024, "In time recommendations through an associative classifier and lookbackapriori: a case study," *EDBT/ICDT Workshops*, pp. 1-8.

[9]. T. N. Nisha, and D. Pramod, 2024, "Insider intrusion detection techniques: A state-of-the-art review," *Journal of Computer Information Systems*, 64(1), pp. 106-123.

[10]. M. Ruckdeschel, R. Baumann, and G. Wiedemann, 2024, "Argument mining of attack and support patterns in dialogical conversations with sequential pattern mining," *In Conference on Advances in Robust Argumentation Machines*, pp. 39-56.

[11]. K. Regulski, A. Opaliński, J. Swadźba, P. Sitkowski, P. Wąsowicz, and A. Kwietniewska-Śmietana, 2024, "Machine learning prediction techniques in the optimization of diagnostic laboratories' network operations," *Applied Sciences*, 14(6), pp. 1434-1454.

[12]. A. A. Noorian Avval, and A. Harounabadi, 2023, "A hybrid recommender system using topic modeling and prefixspan algorithm in social media," *Complex & Intelligent Systems*, 9(4), pp. 4457-4482.

[13]. B. Chen, W. Peng, and J. Song, 2023, "A Frequent Construction Mining Scheme Based on Syntax Tree," *Science and Technology*, 26(1), pp. 3-20.

[14]. S. Karsoum, C. Barrus, L. Gruenwald, and E. Leal, 2023, "Original Research Article Mining timed sequential patterns: The Minits-AllOcc technique," *Journal of Autonomous Intelligence*, 6(1), pp. 1-22.

[15]. V. A. Hameed, M. E. Rana and L. H. Enn, 2023, "Apriori Algorithm based Association Rule Mining to Enhance Small-Scale Retailer Sales," *IEEE 6th International Conference on Big Data and Artificial Intelligence*, pp. 187-191.

[16]. C. M. Chen, Z. Zhang, J. Ming-Tai Wu, and K. Lakshmanna, 2023, "High utility periodic frequent pattern mining in multiple sequences," *CMES-Computer Modeling in Engineering and Sciences*, 137(1), pp. 1-27.

[17]. M. Taylor, A. Barthakur, A. Azad, S. Joksimovic, X. Zhang, and G. Siemens, 2024, "Quantifying collaborative complex problem solving in classrooms using learning analytics," *In Proceedings of the 14th Learning Analytics and Knowledge Conference*, pp. 551-562.

[18]. M. Li, M. Yang, and P. Chen, 2023, "Alarm reduction and root cause inference based on association mining in communication network," *Frontiers in Computer Science*, 5, pp. 1-19.

[19]. D. Han, C. Wang, G. Bian, B. Shao, and T. Shi, 2023, "A novel process recommendation method that integrates disjoint paths and sequential patterns," *Applied Sciences*, 13(6), pp. 1-22.

[20]. M. Addanki, S. Saraswathi, D. B. Slavakkam, R. B. Challagundla, and R. Pamula, 2023, "Integrating sentiment analysis in book recommender system by using rating prediction and dbscan algorithm with hybrid filtering technique," *ResearchSquare*, pp. 1-19.