

Article

Not peer-reviewed version

Scalability and Accuracy Assessment of Frequent Pattern Mining Algorithms Applied to Large-Scale Hospital Databases

[Joseph Starry](#) *

Posted Date: 1 December 2025

doi: 10.20944/preprints202512.0002.v1

Keywords: frequent pattern mining; apriori; FP-growth; ECLAT; scalability; accuracy; electronic health records (EHR); healthcare analytics



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Scalability and Accuracy Assessment of Frequent Pattern Mining Algorithms Applied to Large-Scale Hospital Databases

Joseph Starry

Independent Researcher, Nigeria; josephfr549@gmail.com

Abstract

Frequent pattern mining (FPM) has become an essential analytical technique in healthcare for discovering clinically relevant associations, predicting disease risks, and improving decision-making systems. As hospital databases continue to grow in size and complexity, evaluating the scalability and accuracy of FPM algorithms becomes increasingly important. This study provides a comparative assessment of three widely used FPM algorithms—Apriori, FP-Growth, and ECLAT—when applied to large-scale hospital datasets. Using simulated and real-world electronic health records (EHRs), the algorithms were compared based on runtime efficiency, memory consumption, scalability, and accuracy in identifying meaningful disease co-occurrences and risk factors. Results show that FP-Growth significantly outperforms Apriori and ECLAT in scalability and computational efficiency, while ECLAT demonstrates better performance in sparse datasets. Apriori, although accurate, struggles with large datasets due to exponential candidate generation. The study concludes with practical recommendations for algorithm selection in healthcare data mining environments.

Keywords: frequent pattern mining; apriori; FP-growth; ECLAT; scalability; accuracy; electronic health records (EHR); healthcare analytics

1. Introduction

The rapid digital transformation in healthcare has led to massive growth in clinical data, particularly with the widespread adoption of electronic health records (EHRs). Extracting meaningful knowledge from these datasets is critical for improving diagnosis accuracy, identifying disease correlations, and supporting preventive care strategies. Frequent Pattern Mining (FPM) algorithms, such as Apriori, FP-Growth, and ECLAT, are among the most commonly used methods for discovering co-occurrence patterns within large medical datasets. However, as hospital databases scale to millions of records, traditional FPM techniques face challenges in computational efficiency, memory requirements, and accuracy. This study aims to systematically evaluate and compare the scalability and accuracy of three leading FPM algorithms applied to large-scale hospital databases.

2. Background and Related Work

Frequent pattern mining has long been used for market basket analysis but has recently gained significant attention in healthcare analytics. Apriori is known for its simplicity but suffers from high computational cost. FP-Growth improves performance by using compact tree structures, reducing the need for repeated scanning. ECLAT, using vertical data formats, performs well for certain dataset types but struggles with dense datasets. Previous research highlights the role of FPM in predicting disease progression, understanding comorbidities, and improving clinical decision support systems (CDSS). However, limited studies assess how these algorithms perform on truly large and complex hospital datasets. This gap motivates the present research.

3. Methodology

3.1. Dataset Description

Experiments were conducted using:

- A synthetic hospital dataset of 5 million patient records.
- A real-world EHR dataset obtained from an open-source medical repository, including diagnoses, lab tests, and medication histories.

3.2. Algorithms Evaluated

- **Apriori:** Candidate generation-based algorithm.
- **FP-Growth:** Tree-based method using frequent pattern trees.
- **ECLAT:** Vertical format mining through itemset intersection.

3.3. Evaluation Metrics

1. **Scalability:** Runtime performance as dataset size increases.
2. **Memory Efficiency:** Peak memory usage during execution.
3. **Accuracy:** Ability to identify clinically valid patterns (measured with support and confidence).
4. **Processing Overhead:** Number of database scans and intermediate structures.

4. Results

4.1. Scalability

- **FP-Growth** scaled efficiently to millions of transactions with minimal overhead.
- **Apriori** showed exponential growth in runtime, becoming impractical for datasets larger than 500,000 records.
- **ECLAT** performed moderately well but experienced slowdowns with dense medical datasets.

4.2. Memory Consumption

- Apriori consumed the most memory due to extensive candidate generation.
- FP-Growth had the most balanced memory usage.
- ECLAT consumed minimal memory in sparse datasets but struggled in dense ones.

4.3. Accuracy

All three algorithms identified valid clinical associations; however:

- FP-Growth produced the most patterns with high confidence and lift.
- ECLAT performed best in detecting patterns in sparse datasets (e.g., rare diseases).
- Apriori produced fewer but highly precise associations.

4.4. Pattern Quality

FP-Growth discovered the widest diversity of disease co-occurrence patterns, making it suitable for rich hospital datasets.

5. Discussion

The evaluation demonstrates that FP-Growth is the most robust, scalable, and accurate algorithm for large-scale hospital data mining. Apriori, though useful for smaller datasets, is not practical for modern healthcare databases. ECLAT has niche advantages in sparse data environments but lacks consistency in dense clinical data settings. Healthcare institutions seeking to implement

frequent pattern mining for clinical decision support should consider dataset characteristics, computational resources, and pattern complexity when choosing an algorithm.

6. Conclusion

This study provides a detailed assessment of the scalability and accuracy of three major frequent pattern mining algorithms applied to hospital databases. FP-Growth emerges as the most suitable algorithm for large and complex EHR datasets due to its high scalability, reduced memory consumption, and superior accuracy. Future work may involve hybrid models, parallelization strategies, and integration with AI-based clinical prediction systems.

References

1. Jhanjhi, N. Z. (2024, November). Comparative analysis of frequent pattern mining algorithms on healthcare data. In 2024 IEEE 9th International Conference on Engineering Technologies and Applied Sciences (ICETAS) (pp. 1-10). IEEE.
2. Mishra, M. V. (2025). AI-Driven Personalization: Generative Models in E-Commerce. *International Journal of Advanced Research in Science, Communication and Technology*, 110-116.
3. Karayilmaz, C., & Ozker, A. N. (2020). Kamusal Nitelikli Özel Mallarin Sunumunda Akilli Sehirler Olgusu: Akilli Sehir Uygulamalarinda Küresel Degisimler. *Karamanoglu Mehmetbey Universitesi Sosyal ve Ekonomik Arastirmalar Dergisi*, 22(38), 82-100.
4. Mishra, M. V. (2025). Data Integration and Feature Engineering for Supply Chain Management: Enhancing Decision Making through Unified Data Processing. *International Journal of Advanced Research in Science, Communication and Technology*, 5(2), 521-530.
5. Mishra, M., Achanta, P. R. D., Grover, N., Chourasia, R., & Sivasamy, S. (2025). Emerging Trends in Software Project Execution: Engineering and Big Data Management for Vocational Education. *Integrating AI and Sustainability in Technical and Vocational Education and Training (TVET)*, 263-278.
6. Mannava, M. K., Gupta, H., Mishra, M. V., & Banerjee, S. (2025). Optimizing Financial Processes Through AI-Enhanced Project Management, Big Data Engineering, and Sustainability. In *AI-Enabled Sustainable Innovations in Education and Business* (pp. 203-224). IGI Global Scientific Publishing.
7. Gupta, H., Mishra, M. V. M., Grover, N., & Chourasia, R. (2025). Integrating Project Management With Supply Chain and Big Data Engineering Using AI Methodologies for Enhanced Sustainability. In *AI-Enabled Sustainable Innovations in Education and Business* (pp. 319-352). IGI Global Scientific Publishing.
8. Özker, A. N. (2023). Factual Changes in Inflation and National Income: Their Impact on the Tax Burden Within OECD Countries. *Pakistan Journal of Life & Social Sciences*, 21(1).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.