Unveiling Hidden Patterns in Clinical Databases: A Novel Approach Using Level-By-Level Association Rule Mining

Bartolome Ortiz-Viso[1,3,*][0000−0003−2181−0734], Carlos Fernandez-Basso[1,2,3][0000−0002−8809−8676], M. Dolores Ruiz[1,3][0000−0003−1077−3173], and Maria J. Martin-Bautista[1,3][0000−0002−6973−477X]

[1] Research Centre for Information and Communications Technologies (CITIC-UGR), University of Granada, Granada 18014, Spain
[2] Causal Cognition Lab, Division of Psychology and Language Sciences, University College London, London, United Kingdom
[3] Department of Computer Science and Artificial Intelligence, University of Granada, Granada 18014, Spain
[4] * Corresponding author bortiz@ugr.es

Abstract. Medical data stands out as one of the most valuable sources of information in contemporary data mining research. These datasets encapsulate diverse information, from patient records to current situations. Simultaneously, this wealth of information poses a challenge for data mining due to its diversity and the requirement for context to discern which data holds value and what conclusions truly contribute to generating new knowledge. In this work, we leverage a database of medical records from patients' visits to several hospitals across several years, enhancing it with hospital information and a disease ontology. This allows us to identify a medical diagnosis at various levels of semantic depth. Based on this information, we propose information mining centered around extracting association rules at progressive specific levels concerning diagnoses. Subsequently, we present the initial results of this study for different sets of diseases and suggest the most relevant steps for development based on these outcomes.

Keywords: Association rules · medicine applications · data mining · medical records

1 Introduction

Emerging technologies and analytical methods in healthcare have sparked a significant shift in how we utilize clinical databases. This paper introduces a level-by-level approach to Association Rule Mining (ARM) for clinical databases to enhance decision-making by uncovering hidden patterns and relationships.

Despite their richness in patient data, treatment plans, and health outcomes, clinical databases are underutilized due to their size, complexity, and standardization issues. Traditional techniques fall short of fully harnessing these vast

datasets. ARM offers a solution by effectively extracting hidden correlations, leading to impact healthcare interventions and policies.

The unique, tiered ARM approach fits well with the hierarchical structure of clinical databases and their multi-dimensional data points. It enables healthcare practitioners to uncover valuable insights about treatment effectiveness and predictors of health outcomes.

This paper highlights that a hierarchical ARM approach can be used to address computational efficiency and interpretability challenges. By segmenting the analysis into different layers of conceptual specificity, the methodology improves the interpretability of the results for clinicians and ensures the direct applicability of the findings to healthcare delivery and policy development.

In summary, exploring clinical databases using level-by-level association rules opens new avenues for understanding patient care complexities and improving healthcare outcomes. Through this research, we aim to demonstrate the efficacy and applicability of ARM in clinical settings by applying the ARM analysis performed progressively with more specific categories, gathering novel insights on the usefulness of medical records, and contributing to the advancement of data-driven healthcare solutions.

Following the introduction, we delve into various related works documented in the scientific literature in Section 2. Section 3 introduces the data sources and the methodology essential for conducting the experiments. The results are comprehensively presented in Section 4. In Section 5, we expand on their implications, analyzing the improvements, technologies, and processes that can help or improve this process. The paper concludes by summarizing our findings and insights in Section 6.

2 Related works

A substantial body of research has demonstrated success in mining patterns in clinical databases, employing many data mining techniques, including Association Rule Mining (ARM). Its proven efficacy in deciphering hidden patterns in voluminous data sets has become a crucial instrument in health informatics.

Historically, one of the inaugural applications of rule-based methods stems from the work [4], which initially aimed to manage vast data volumes in retail scenarios like supermarkets. However, the potential of ARM has reached beyond this sector, finding its foothold in various areas, notably in the healthcare and medical domains. Further advancements in this area were marked by the work of Han, Pei, and Yin [19], who introduced the Frequent Pattern (FP)-Growth algorithm, significantly improving the computational efficiency of ARM.

Parallelly, a critical yet often underestimated aspect of big-data research, particularly in healthcare, is data preprocessing. Recognized by Kurgan et al. [20] and Fernandez-Basso et al. [14], this foundational step increases the performance and reliability of data mining techniques like ARM. It is unequivocally instrumental in resolving prevalent issues in raw healthcare data, such as inconsistent data, redundancy, outliers, and missing data points. Numerous studies [12, 6] show how strategic preprocessing approaches, including data normalization, discretization, cleaning, and transformation, remarkably improved data analysis

outcomes. While its significance is irrefutable, the nature of the data and the specific objectives of the analysis must guide the choice of data preprocessing methods.

Data enrichment is a noteworthy advancement that considerably contributes to the success of clinical database analysis. This practice implies augmenting primary data with context-specific, supplementary data from various sources. These could include patient medical histories, lab results, physician notes, genomics, and imaging data, amongst others [8]. By aggregating these diverse data through knowledge representation methods and tools, a more detailed, comprehensive patient view can be garnered, thus unveiling complex patterns and associations. Natural language processing and health informatics have increased the ability to extract, analyze, and interpret medically relevant information [9, 25]. Despite its challenges due to its high complexity and dimensionality, medical data enrichment propels personalized medicine and improved patient care.

Within the healthcare landscape, utilization of ARM has particularly been exemplified in the electronic health records (EHR) domain. Notably, [22] postulated that the interaction of ARM with machine learning classification methods could unveil beneficial rules within a dataset, thus proving the value-laden potential of intelligent data analysis in healthcare.

Innovation in recent years has seen the advent of techniques like level-by-level Association Rule Mining, which have shown promising results. To illustrate, [3] used this method to unveil diabetes risk factors, thus providing insightful inputs for disease management and prevention. [15] conducted parallel work and were able to scrutinize associations between hospital operational performance and medical error incidents.

On the other side of the horizon, specific research work has focused on mitigating the limitations of ARM, such as its huge rule sets and the challenge of false associations [13, 28]. Statistical techniques and Bayesian inferential strategies have been employed in these same efforts to enhance the algorithm's behavior.

Despite the rich array of studies in this area, ripe opportunity for exploration remains, especially with the constant evolution in computational hardware, which propels the augmentation of clinical databases. Therefore, efforts towards improving ARM and employing innovative practices such as level-by-level association rule mining are both immediate and mandatory. This will enable further insight into hidden patterns in healthcare data, potentially leading to improved patient risk stratification or even novel predictive models for disease progression.

3 Methodology

The primary objective of this work is to collect and standardize medical documents generated when users visit any of the associated hospital centers. For this purpose, we used a dataset collected from the Clinical Hospital "San Cecilio" between the years 2016 and 2020. This initial dataset follows the structure outlined in what is known as the Minimum Basic Dataset (CMDB) in the hospitals of the public health system of Andalusia, information mandated to be generated in hos-

pitals, Community Therapeutic Health Mental Hospitalization Centers (CTE), Surgical Day Hospitals (HDQ), and Medical Day Hospitals (HDM).

Subsequently, to address privacy concerns, this data is anonymized by the European and Spanish regulations regarding the protection of personal data. Within this framework, we find information about the patient and their hospital stay through specific dates and codes that refer to each of their characteristics (e.g., the coded discharge reason among the 13 possible reasons). More detailed information on this framework can be found in [2].

Since our initial interest is to understand potential relationships between patients' illnesses and their severity, the dataset contains two main points: the diagnoses associated with a specific patient (called *"Diagnoses"*) and the related hospital stay (called *"Patients time in hospital"*). We describe these two areas with greater precision and outline the transformations applied to each of them, with a graphic summary in Figure 1:

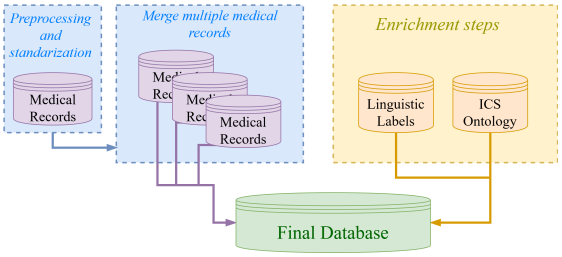

Fig. 1. Database creation process and posterior enrichment

– **Patients time in hospital:** Firstly, our interest lies in understanding the duration of patients' stay in the hospital. For this purpose, we require information from three specific sources: the total hospitalization time, whether the patient at any point needed to transition from general hospitalization to the Intensive Care Unit (ICU) of the hospital (reserved for patients undergoing necessary surgical procedures or dealing with more severe cases) and if so, the Time spent in the ICU. The hospitalization time is calculated for these three variables, excluding the Boolean-coded ICU stay. The ICU stay is then calculated and transformed into different categories, each labeled with specific linguistic labels.
In creating the ICU categories, we designate a category for cases with no hospitalization in the ICU and another for those related to a brief visit or a quick surgical procedure lasting one day. For cases involving more delicate surgical procedures (e.g., myocardial infarction), the ICU stay is categorized between 2 and 4 days. Finally, cases exceeding four days in the ICU are considered the most severe. According to medical literature [21], it is possible further to subdivide the spectrum beyond those initial four days (with two additional categories between 4 and 13 days as prolonged and 14 days onwards as very prolonged). However, due to the limited number of cases supporting these rules, a decision is made to encompass these cases.

A similar procedure is followed for the hospital stay, where seven linguistic labels express the nature of the stay. These labels range from "intervention" for stays of 1 to 3 days, typically after a simple surgical intervention or childbirth, to "very long stays" lasting more than two and a half months. All labels are described in Figure 2.

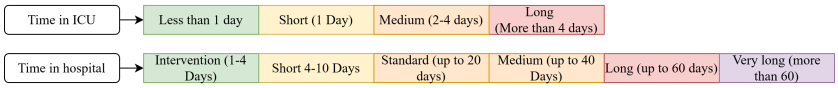

Fig. 2. Semantic labels associated with different stay durations

– **Diagnoses and diagnoses enrichment:** Another differentiating aspect of our analysis is the depth of diagnoses we can evaluate with the rules. Properly encoding and mining the diagnosis section in our datasets is necessary to achieve this. Various classification indices have been proposed for disease classification, with the most widely used being the International Classification of Diseases (ICD) [1], created and managed by the World Health Organization (WHO). The ICD serves a broad range of global uses, providing crucial knowledge about the extent, causes, and consequences of human disease and death worldwide through data reported and coded using the ICD. The diagnostic guidance linked to ICD categories standardizes data collection and enables large-scale research.
However, this protocol is not static and is continuously evolving, with new diseases or diagnoses being evaluated and updated, leading to new versions of the index. This evolution can result in discrepancies if the appropriate version is not considered. For Spain, the ICD classification as of 2024 refers to the second edition of the CIE-10-ES [11], which is derived from the International Classification of Diseases, 10th Revision, Clinical Modification (ICD-10-CM), initially published by the United States government. The ICD-10-CM was developed by the National Center for Health Statistics (NCHS), part of the Department of Health of the U.S. federal government (DHHS), and the clinical modification is based on the original classification by the World Health Organization (WHO). In our case, we use the ICD-10-CM classification available through the NHI. This collection can be used for all diagnoses correlated until 2020.
In that year, the emergence of COVID-19 is classified as a new and unclassified disease in Spanish records. In subsequent revisions, it would precisely align with a classification following scientific consensus and be categorized under respiratory diseases. For this study, we have utilized this initial classification, separating its processing by years, which may influence potential differences in the number of ICU patients not yet included with a respiratory diagnosis. This topic is addressed in Section 5.
In terms of information, the ICD-10 (International Classification of Diseases, 10th Revision) or CIE-10-ES consists of two distinct parts: the Alphabetical Index and the Tabular List. Our analysis focused on the Tabular List, an

alphanumeric listing of codes organized into chapters based on body systems or medical entities. Our classification, derived from the Tabular List, captures the details of the most precise diagnosis (the one stated in the dataset as diagnosis) and ascends three levels in the ontology of diseases, transforming the codes into linguistic labels of progressive generality. This approach enables us to establish a broad classification of the affected apparatus with two additional levels of specificity before reaching the specific diagnosis. An illustration of this process can be found in Figure 3.

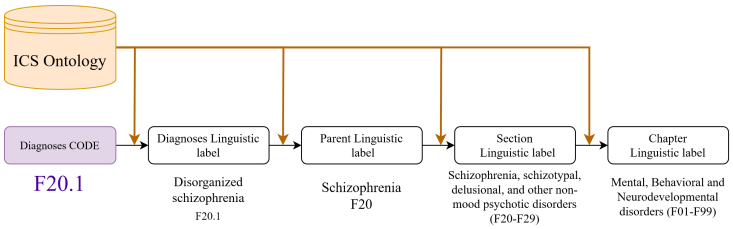

Fig. 3. Diagnoses enrichment via ICS10, example with a schizophrenia diagnostic.

After this process, we obtained 426,557 records of enriched patient data. We then selected the data fields related to diagnoses and Time within the hospital/ICU.

For the associated rule mining, we used the package `mlxstend` [24], where several association rules mining algorithms already established in literature can be employed. To conduct practical data mining, it was necessary to categorize diagnoses by extracting boolean values from each type. This process was initially applied across three levels of depth and gradually eliminated as we delved deeper levels. In simpler terms, the boolean categorization of a pathology's section was discarded when selecting specific pathologies within a particular section.

The selection of chapters, sections, and pathologies for the study is driven by two main factors: first, the type of pathology and its societal relevance (ensuring sufficient literature exists to validate the importance and validity of rules), and second, the notable proportion of cases within a specific section or chapter, as well as the percentage of these cases spending multiple days in the ICU. For instance, we excluded those related to traumas and associated surgeries in favor of pathologies like heart attacks or infections that result in ICU stays.

Regarding the chosen algorithms, given the information sets at our disposal, most algorithms can produce interesting results within an acceptable timeframe. Future steps and adaptations at this juncture are discussed in the discussion section. For our experimentation, we initially analyzed using the Apriori algorithm [5] and the FP-growth algorithm [19]. Both yielded similar results in terms of performance and rules generated, requiring extensive algorithms in this initial phase of the study rather than those exploring space probabilistically.

During our experimentation, we relaxed the support limit to 0.1 to ensure capturing the majority of rules the algorithm might find, even those with few occurrences, for this preliminary feasibility analysis. Subsequently, we filtered

out simple rules that did not contribute to any relationships and those that only related information showcasing semantic relationships across different diagnostic levels. After that, we also filtered the results enforcing in the antecedent and consequent to respect the ontology levels; that is, no rules should connect a more restricted category (i.e., diagnoses) to a broader category in the consequent (i.e., section-level).

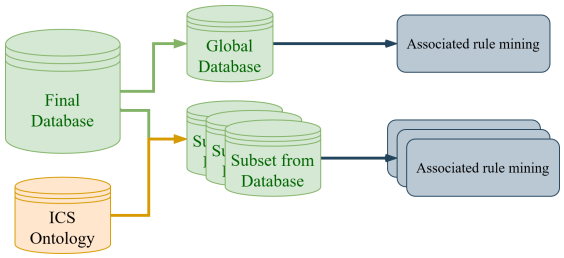

Fig. 4. Level-by-level approach: Once the database is built we run different Association rule mining algorithms ranging from more general level to more specific diagnoses.

4 Results

Throughout our study, several significant patterns emerged from the clinical database analyses. We effectively mitigated the limitations inherent to traditional ARM by utilizing level-by-level association rules, which offered tangible benefits in computational efficiency and the trade-off between memory usage and running time. This is due to the fact that selecting specific levels reduces the overall search space, avoiding the need to consider every possible combination of diagnoses simultaneously. This targeted approach optimizes resource allocation and facilitates parallelization, enabling efficient processing of large datasets commonly encountered in pathology ontologies.

4.1 Cardiac pathologies

Cardiac diseases represent the third most frequent category of pathologies in our dataset, encompassing over 30,000 patients whose primary reason for admission is a consultation due to issues related to them. It is also a widely studied category and interesting because of the effect of long ICU stays from cardiac pathologies or interventions [26, 27]. In our case, to test the system's effectiveness, we have conducted experiments considering the number of admitted patients and their stay in the ICU. For this purpose, we have selected Cardiac diseases, ranging from more general to more specific. In the deeper levels, we focused our analysis on *"ST elevation (STEMI) myocardial infarction"* belonging to the broader category of *"Acute myocardial infarction"* and, in turn, to *"Ischemic heart diseases (I20-I25)"* and *"Diseases of the circulatory system (I00-I99)"*. On the rules extracted, we offer in Table 1 a selection backed by the current medical literature. In the table, the level-by-level approach can be seen within the different objectives and information extracted.

In the context of STEMI, two rules of particular interest stand out, and these can be referenced in Table 1 at the section level. These rules illustrate that right atrial infarcts demonstrate a stronger association with a shorter timeframe than left atrial infarcts. Thus, we can infer that left atrial infarcts are potentially more severe due to the extended duration of ICU stays and hospital admissions associated with them. This could be attributed to the left atrium's role—it receives oxygen-enriched blood from the lungs and forwards it to the left ventricle, which subsequently pumps this oxygen-rich blood to the body. Hence, any impairment or damage to the left atrium might significantly affect the heart's capacity to provide an efficient blood supply to the body.

Antecedents	Consequent	Support	Confidence	Lift
Chapter-level				
Parent: *Aortic aneurysm and dissection*, Days in UCI: *Short*	AGe > 80 y.o.	0.0181	0.5763	1.7840
Section: *Diseases of arteries, arterioles and capillaries (I70-I79)*, Days in UCI: *Short*	Parent: Aortic aneurysm and dissection'	0.0314	0.7763	9.2762
Section-level				
Diagnoses: *ST elevation (STEMI) myocardial infarction involving left anterior descending coronary artery*, Time in hospital: *Standard*, Parent: *Acute myocardial infarction*	Time in UCI: *Long*	0.0136	0.7222	7.0454
Age: *50-68 y.o.*, Parent: *Acute myocardial infarction*, Diagnoses: *ST elevation (STEMI) myocardial infarction involving right coronary artery*, Days in UCI: *Short*	Time in hospital: *Intervention*	0.0188	0.75	2.0486
Parent-level				
Diagnoses: ST elevation (STEMI) myocardial infarction involving left anterior descending coronary artery, Time in hospital: Standard	Time in UCI: Long	0.0143	0.7222	6.8953

Table 1. Cardiac rules examples by different levels.

5 Discussion

With the previously highlighted results and our initial approach, numerous study aspects emerge that can be the subject of successive expansions in our work. In this section, we develop these aspects and discuss their implications.

Firstly, the data we have can be further enriched by hospital records. Works, such as [15], emphasize the extraction of association rules from medical records, including seasonal factors not only as background but also as a means of filtering

future rules. In this case, we would discuss the transformation of additional records to create data fields related to seasonality, location, or type of hospital. Moreover, this level-by-level extracting can be a very promising approach to get data insights not only on the different diagnosis levels proposed but also with specific sets of initial conditions. In areas like pregnancy, when we reproduce the level-by-level approach selecting certain patient characteristics as *Age*, we can delve into more detailed rules centered on age-dependent pathologies.

Another potential enrichment could come from secondary diagnoses. Hospital records contain secondary diagnoses that point to specific patient illnesses unrelated to the reason for their visit to the hospital. These secondary illnesses could provide another interesting point of discussion to enrich the data and obtain relationships conditioned by the patient's medical history. While this additional diagnosis offers an interesting enrichment, up to 20 additional new data fields encode the patient historical data. As these records increase the volume of data for rule generation, it becomes necessary to use optimized algorithms for large datasets [17], [10].

This would also allow us to explore the application of generalized association rules [7], [23]. It is also worth noting that applications of generalized rule mining in medicine, especially in cases involving associated complications, as we aim to study in this work, may overlook essential data or generate associations that lose significant information. For instance, malignant variants with a worse prognosis than benign ones may share close categorization of pathologies, leading to potential oversights. However, this factor may be useful for including secondary previous diagnoses, where detailed information may not be as crucial as for primary diagnoses. Although this data is encoded with a specific diagnosis, as depicted in Figure 3, a multilevel approach can be achieved.

Throughout this work, we have utilized aspects such as ICU stay duration or hospitalization time through linguistic labels. These labels condense information into categories that may or may not apply to the patients we have considered. However, it is arguable that this distinction and categorization of data are limited and lack reasonable flexibility when dealing with less rigid variables such as Time. Therefore, the natural extension of this approach could benefit from a shift in the treatment of these variables to one that allows for more flexibility: the fuzzy approach. Aspects like age, ICU or hospitalization time, or some potential additional enrichment data mentioned earlier take on a more specific significance when using a fuzzy approach where the degree of membership enables a categorization that better reflects reality. The fuzzification of the system also involves using mechanisms capable of working with fuzzy rules beyond a mere crisp transformation. To achieve this, some interesting works in this area have been developed as [18, 16].

6 Conclusions

This study has showcased the application of level-by-level Association Rule Mining (ARM) as a novel approach to unveil hidden patterns within clinical databases. Our methodology has settled the way for managing the unique challenge of mining high-dimensional and complex medical data, uncovering signifi-

cant associations that can pave the avenue for more effective healthcare provision and policy formation.

Our findings from the database analyses have strengthened the importance and relevance of the ARM approach and, more specifically, the level-by-level ARM approach in healthcare data interpretation. This level-by-level approach presents a practical solution to the challenges in traditional ARM, particularly in terms of interpretability and computational efficiency.

Moreover, our results underscore the need for data preprocessing and the inclusion of a wide range of data sources in data enrichment. These critical steps in the processing pipeline have a direct and notable impact on the quality of the resulting mined rules, highlighting the importance of investing in these areas alongside developing advanced analytical techniques.

However, as with all progress in the data science domain, this study is not without its limitations and opportunities for future growth. Key among the challenges faced is the robustness of ARM for handling excessively large rule sets and preventing false associations. This matter warrants further research and exploration to enhance the usability and reliability of the findings.

Acknowledgements

We would like to acknowledge support for this work from FederaMed project: Grant PID2021-123960OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by ERDF/EU. And from DesinfoScan project: Grant TED2021-129402B-C21 funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. In addition, the Ministry of Universities has partially supported this research through the EU-funded Margarita Salas programme NextGenerationEU and the pre-competitive project of the Plan Propio of the "University of Granada".We would like to thank to Clinical Hospital San Cecilio for the data. Finally, the research reported in this paper is also funded by the European Union (BAG-INTEL project, grant agreement no. 101121309).

References

1. International Classification of Diseases (ICD)
2. Manual de instrucciones del Conjunto Mínimo Básico de Datos de Andalucía. 2024
3. Abdelhamid, A.A., Eid, M.M., Abotaleb, M., Towfek, S., et al.: Identification of cardiovascular disease risk factors among diabetes patients using ontological data mining techniques. Journal of Artificial Intelligence and Metaheuristics **4**(2), 45–53 (2023)
4. Agrawal, R., Imielinski, T., Swami, A.: Database mining: A performance perspective. IEEE transactions on knowledge and data engineering **5**(6), 914–925 (1993)
5. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: Proc. 20th int. conf. very large data bases, VLDB. vol. 1215, pp. 487–499. Santiago, Chile (1994)
6. Andreu-Perez, J., Poon, C.C., Merrifield, R.D., Wong, S.T., Yang, G.Z.: Big data for health. IEEE journal of biomedical and health informatics **19**(4), 1193–1208 (2015)
7. Baralis, E., Cagliero, L., Cerquitelli, T., Garza, P.: Generalized association rule mining with constraints. Information Sciences **194**, 68–84 (Jul 2012)

8. Bellazzi, R.: Big data and biomedical informatics: a challenging opportunity. Yearbook of medical informatics **23**(01), 08–13 (2014)
9. Demner-Fushman, D., Elhadad, N., Friedman, C.: Natural language processing for health-related texts. In: Biomedical Informatics: Computer Applications in Health Care and Biomedicine, pp. 241–272. Springer (2021)
10. Dolores, M., Fernandez-Basso, C., Gómez-Romero, J., Martin-Bautista, M.J.: A big data association rule mining based approach for energy building behaviour analysis in an IoT environment. Scientific Reports **13**(1), 19810 (Nov 2023), number: 1 Publisher: Nature Publishing Group
11. Edicion-Enero: Clasificacion Internacional de Enfermedades, 10. Revision. Modificacion Clinica
12. Elhoseny, M., Abdelaziz, A., Salama, A.S., Riad, A.M., Muhammad, K., Sangaiah, A.K.: A hybrid model of internet of things and cloud computing to manage big data in health services applications. Future generation computer systems **86**, 1383–1394 (2018)
13. Evfimievski, A., Srikant, R., Agrawal, R., Gehrke, J.: Privacy preserving mining of association rules. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 217–228 (2002)
14. Fernandez-Basso, C., Gutiérrez-Batista, K., Morcillo-Jiménez, R., Vila, M.A., Martin-Bautista, M.J.: A fuzzy-based medical system for pattern mining in a distributed environment: Application to diagnostic and co-morbidity. Applied Soft Computing **122**, 108870 (2022)
15. Fernandez-Basso, C., Gutiérrez-Batista, K., Morcillo-Jiménez, R., Vila, M.A., Martin-Bautista, M.J.: A fuzzy-based medical system for pattern mining in a distributed environment: Application to diagnostic and co-morbidity. Applied Soft Computing **122**, 108870 (Jun 2022)
16. Fernandez-Basso, C., Ruiz, M.D., Martin-Bautista, M.J.: A fuzzy mining approach for energy efficiency in a big data framework. IEEE Transactions on Fuzzy Systems **28**(11), 2747–2758 (2020)
17. Fernandez-Basso, C., Ruiz, M.D., Martin-Bautista, M.J.: Spark solutions for discovering fuzzy association rules in Big Data. International Journal of Approximate Reasoning **137**, 94–112 (Oct 2021)
18. Fernandez-Basso, C., Ruiz, M.D., Martin-Bautista, M.J.: Spark solutions for discovering fuzzy association rules in big data. International Journal of Approximate Reasoning **137**, 94–112 (2021)
19. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data mining and knowledge discovery **8**, 53–87 (2004)
20. Kurgan, L.A., Musilek, P.: A survey of knowledge discovery and data mining process models. The Knowledge Engineering Review **21**(1), 1–24 (2006)
21. Laupland, K.B., Kirkpatrick, A.W., Kortbeek, J.B., Zuege, D.J.: Long-term Mortality Outcome Associated With Prolonged Admission to the ICU. CHEST **129**(4), 954–959 (Apr 2006), publisher: Elsevier
22. Patel, P., Sivaiah, B., Patel, R., Choudhary, R.: Association rule mining for healthcare data analysis. In: Computational Intelligence in Healthcare Informatics, pp. 127–139. Springer (2024)
23. Percin, I., Yagin, F., Guldogan, E., Yologlu, S.: ARM: An Interactive Web Software for Association Rules Mining and an Application in Medicine (2019)
24. Raschka, S.: Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. The Journal of Open Source Software **3**(24) (Apr 2018)

25. Rehman, A., Naz, S., Razzak, I.: Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities. Multimedia Systems **28**(4), 1339–1371 (2022)
26. Shah, V., Ahuja, A., Kumar, A., Anstey, C., Thang, C., Guo, L., Shekar, K., Ramanan, M.: Outcomes of Prolonged ICU Stay for Patients Undergoing Cardiac Surgery in Australia and New Zealand. Journal of Cardiothoracic and Vascular Anesthesia **36**(12), 4313–4319 (Dec 2022)
27. Stoppe, C., Ney, J., Lomivorotov, V.V., Efremov, S.M., Benstoem, C., Hill, A., Nesterova, E., Laaf, E., Goetzenich, A., McDonald, B., Peine, A., Marx, G., Fehnle, K., Heyland, D.K.: Prediction of Prolonged ICU Stay in Cardiac Surgery Patients as a Useful Method to Identify Nutrition Risk in Cardiac Surgery Patients: A Post Hoc Analysis of a Prospective Observational Study. JPEN. Journal of Parenteral and Enteral Nutrition **43**(6), 768–779 (Aug 2019)
28. Veloso, A., Jr, W.M., Cristo, M., Gonçalves, M., Zaki, M.: Multi-evidence, multi-criteria, lazy associative document classification. In: Proceedings of the 15th ACM international conference on information and knowledge management. pp. 218–227 (2006)