The Bioscan AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES 19(2): S2: 338-345, 2024 www.thebioscan.com

Unveiling Symptomatic Pathways of COVID-19: An In-depth Analysis Employing SLIM Association Rule Mining

G. Srilatha[1], N Subhash Chandra[2], Gnaneswara Rao Nitta[3]

[1]Asst. Professor, Jyothishmathi Institute of Technology & Science, Nustulapr, Karimnagar, Telangana, India.

[2]Professor, Department of Computer Science and Engineering, CVR College of Engineering,Hyderabad, India.

[3]Professor of CSE Department, DRK Institute of Science and Technology, Hyderabad, India.

gnani.nitta@gmail.com, gajula.srilatha.2017@gmail.com

DOI: https://doi.org/10.63001/tbs.2024.v19.i02.S2.pp338-345

KEYWORDS
Association Rule, Apriori, SLIM, Covid-19, Convolutional neural network, Deep neural network
Received on:
20-07-2024
Accepted on:
06-11-2024

ABSTRACT

In order to achieve a thorough comprehension of COVID-19, it is imperative to elucidate the interconnections among its symptoms. The proposed methodology aims to discern patterns that can effectively diagnose the emergence of symptoms, thereby improving diagnostic accuracy and potentially yielding significant breakthroughs. The current study facilitates identifying patterns that signify symptoms' occurrence and convergence. The SLIM algorithm is introduced as a novel method for systematically analyzing large symptom datasets, and its application is presented in this study. The reason for this goes beyond the capabilities of conventional methods, enabling greater diagnostic precision and illuminating new investigational avenues. The present study predicts and identifies COVID-19-related symptoms with greater accuracy, which enhances our understanding of the disease's symptomatology and encourages further study in the area. The results of this study have improved the experience and management of COVID-19's diagnosis, treatment, and pathophysiology, thereby contributing to a better grasp of the virus and bolstering global efforts to manage and mitigate it.

INTRODUCTION

The healthcare sector, pivotal in societal contexts that prioritize the intrinsic value of human life, encountered a monumental challenge with the emergence of COVID-19, first identified in Wuhan, China. The initial outbreak, closely associated with a local food market as investigated by Zhu et al. (2020)[1], underscored an urgent need to comprehend this novel disease, particularly considering the market's role in the virus's initial propagation. As scientific inquiries unfolded, the public release of the genetic sequence of SARS-CoV-2 in mid-January 2020 facilitated broader investigations into its origins. The prevailing theory, which posits bats as the ecological reservoir, mirrors the emergence patterns of SARS-CoV-1 in 2003. However, the exact intermediate host remains elusive[2], highlighting a critical knowledge gap emphasizing the necessity of ongoing research in this domain[3]. Transitioning to a focus on the aftermath of the virus, the [4] unveiled an exploration into post-COVID-19 conditions using a Phenome-Wide Association Study (PheWAS). The study illuminated the prevalence of respiratory, circulatory, and mental health disorders in survivors, revealing distinct patterns compared to control groups. Diagnostic approaches, especially the widely used Reverse Transcription Polymerase Chain Reaction (RT-PCR), while impactful, present notable challenges due to their complex, resource-intensive nature and the global limitations on testing kit availability, further complicated by a notable incidence of false negatives[5].

The aforementioned diagnostic challenges steer the discourse towards incorporating technological advancements in addressing these concerns. In this vein, machine learning and computer-assisted techniques in medical imaging, particularly employing Artificial Neural Networks (ANN) and deep learning algorithms, have garnered attention for their commendable progress in clinical image analysis, covering facets like fusion, registration, segmentation, and classification[6]. Extending this technological narrative, the present research paper proposes a novel methodology, employing the SLIM and Apriori algorithms[7,8], aiming to discern patterns related to COVID-19 through a meticulous examination of a comprehensive symptom dataset. This methodology furnishes deeper insights and enhances diagnostic precision, thereby becoming a significant contributor to global strategies addressing the COVID-19 pandemic.

The rest of the paper flows as follows: The paper briefly discusses innovative COVID-19 symptom analysis using SLIM Association Rule Mining in several sections. Knowledge and research gaps are identified in Section 2, extensive literature review. Section 3 discusses the SLIM Association Rule Mining and A Priori Algorithms. Section 4 discusses methodology-derived COVID-19 symptom patterns and associations and their research implications. Section 5 discusses the study, implications, and future research.

1. Literature Survey

The advent of COVID-19 has galvanized a plethora of research activities, focusing on varied approaches for understanding, managing, and mitigating the impact of the pandemic. In particular, deep learning methods have demonstrated significant success across numerous computer vision tasks, particularly in recognizing COVID-19 infections from X-ray scans. A notable study by [9] explored various Convolutional Neural Network (CNN) architectures, including VGG-19 and DenseNet-121, highlighting their commendable performance in a binary classification of COVID-19 infections despite a database of merely 25 COVID-19 cases.

Pivoting from deep learning methods, machine learning methodologies have also garnered attention for their extensive

applications in diverse biomedical domains, ranging from genomic and genetic analysis to death prediction, drug discovery, and even patient similarity [10-18]. The Association Rule Mining (ARM) is a noteworthy area within medical literacy, contributing valuable insights into prevalent patterns and associations within data. This leads us to explore innovative methodologies like the Divergent Association Rules Approach (DARA) introduced by [19], which adeptly manages the multitude of Association Rules (ARs) during frequent pattern mining, successfully revealing insightful patterns related to malaria in Brazil and demonstrating substantial contributions to both computational and medical research fields. Moreover, the utilization of artificial intelligence and acoustics in monitoring symptomatic indicators, such as cough patterns in COVID-19 patients, has been explored, with [20] illustrating the potential of such technology in predicting clinical outcomes and managing patient care effectively. Further, [21] adeptly leveraged non-negative matrix factorization, deciphering latent topics within clinical notes to unveil the diverse impacts of COVID-19 on health and healthcare practices. Additionally, [22] introduced the COVID-19 Annotated Clinical Text (CACT) Corpus and a related extraction model, emphasizing the importance of employing structured and unstructured patient data for enhancing the predictive performance for COVID-19.

In the context of symptom patterns and associations, while deep neural network (DNN) algorithms have been prevalently employed for forecasting and categorizing COVID-19 cases [23-29], this study prioritizes the application of Autoregressive Moving Average (ARM) models to delve deeper into COVID-19 symptom patterns. The aim is to amplify our understanding of the disease by providing a nuanced viewpoint and emphasizing the potential of ARM to reveal underlying symptom associations, thereby enabling more informed and effective clinical decision-making.

Lastly, innovative data analysis techniques—including artificial intelligence, association rule mining, non-negative matrix factorization, and information extraction models—cannot be understated in understanding, predicting, and managing various aspects of the COVID-19 pandemic. These methodologies and findings could potentially enhance the application of algorithms, such as the SLIM, in symptom pattern analysis. Thus, the primary objective of the proposed study is to employ SLIM ARM to scrutinize the associations among symptoms observed in individuals diagnosed with COVID-19, aiming to augment clinical dynamics and patient management strategies while providing meaningful insights into the ongoing efforts in managing and understanding COVID-19, thereby bridging existing research discrepancies.

2. The Proposed Methodology

This paper presents a novel approach utilizing the SLIM algorithm to discover patterns of COVID-19 by analyzing patient symptoms within a comprehensive dataset. The first stage of this methodology entails the thorough processing of a dataset that contains abundant information on symptoms related to COVID-19. The dataset is modified by excluding missing values, resulting in a more efficient "transactional database" called "symptom data." This conversion establishes a strong basis for the subsequent stages of the study, guaranteeing the integrity and dependability of data in investigating symptom patterns related to COVID-19. During the subsequent phase, as illustrated in Figure 1, the enhanced transactional dataset undergoes the application of the SLIM algorithm. This particular step is of utmost importance in facilitating the automated detection of complex COVID-19 patterns, thereby enabling a comprehensive examination of the well-established Apriori Association Algorithms. As mentioned earlier, the comparison highlights the effectiveness and accuracy of the SLIM algorithm in detecting COVID-19 patterns with great precision. This algorithm plays a significant role in automated disease pattern discovery, improving our comprehension and control of COVID-19.

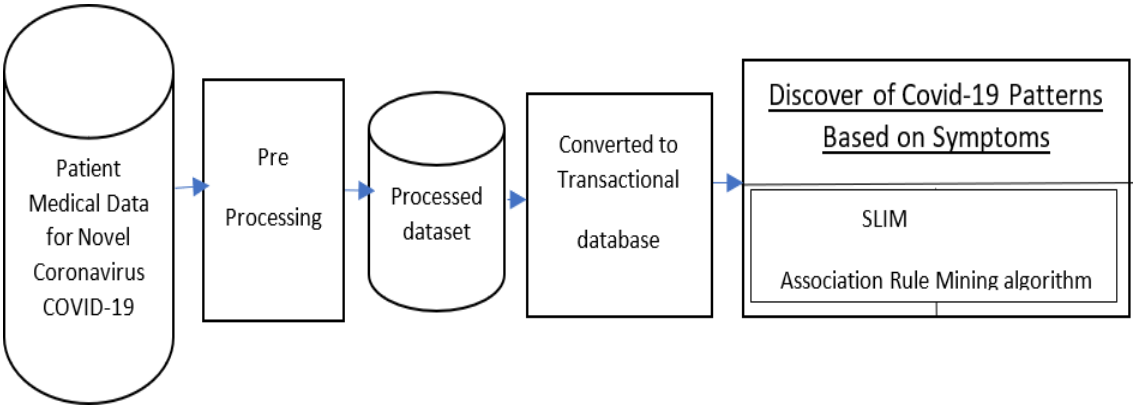


Figure. 1: The proposed methodology Block diagram.

3.1 Pre-processing

The initial phase of pre-processing, especially for datasets related to patients with COVID-19, required a comprehensive approach to handling the available symptom information. This involved carefully excluding any missing qualities or absent data [29]. The dataset that was carefully selected originally consisted of 24 variables and underwent a rigorous filtering procedure. In order to improve the analytical precision of the study, certain variables, including patient age, sex, disease symptoms, underlying chronic conditions, and mortality data, were deliberately excluded during this phase. After the variables were extracted, a categorization was initiated to systematically group associated symptoms with similar meanings to enhance the data's reliability. As a result, specific diseases were attributed to clinical symptom patterns based on the curated COVID-19 dataset. An illustrative instance can be observed in the assignment of the term 'pneumonia' as a pragmatic identifier for individuals exhibiting symptomatic chest infections. This highlights the deliberate alignment of clinical indicators with specific diseases, aiming to improve analytical effectiveness.

3.2 Apriori Algorithm

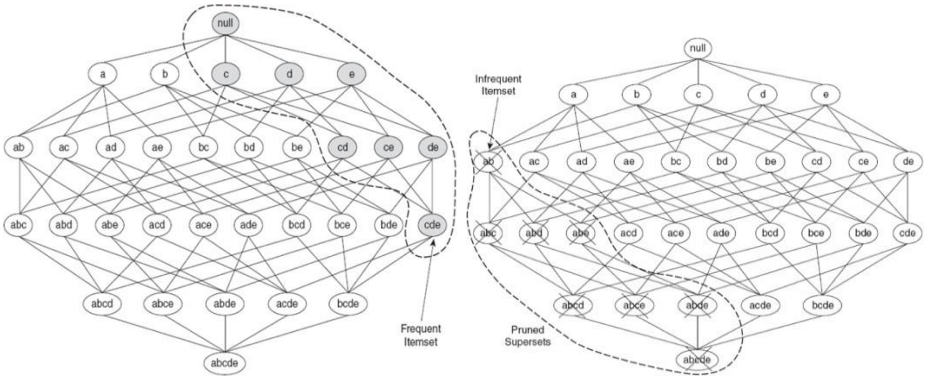
The Apriori algorithm is crucial for discovering frequent item sets for generating Boolean association rules within a given dataset. The term 'Apriori' is derived from its reliance on preexisting knowledge about the characteristics of frequent itemsets [8]. It employs a level-wise or iterative approach, where k-frequent itemsets are used to investigate (k+1)-itemsets, resulting in improved efficiency in generating level-wise frequent itemsets. The Apriori property is of particular significance as it dramatically diminishes the search space, as visually depicted in Figure 2.

The fundamental principle of the Apriori algorithm is its reliance on the anti-monotonicity property of the support measure. This property asserts that:

- According to the Apriori property, as illustrated in Figure 2(a), it is necessary for all subsets of a frequent itemset also to be frequent.

- If an item is deemed infrequent, it can be observed that all of its supersets will also be infrequent, as illustrated in Figure 2(b).

The dichotomy at hand effectively reduces the search space, enhancing the algorithm's ability to detect frequent dataset items.



2(a) 2(b)

Figures 2(a) and 2(b): Apriori Algorithm Supersets

3.3 The Proposed SLIM Algorithm

The methodology delineated within this section utilizes the SLIM algorithm, a tool specifically designed to mine heuristic solutions directly from the dataset, mainly targeting the Minimal Coding Set Problem [7]. This heuristic is attuned to mining high-quality information depictions on transaction data with adeptness and efficiency that sets it apart as a one-stage, any-time alternative, especially when compared with other methodologies. Moreover, it exhibits a remarkable capability in managing voluminous and dense datasets without imposition limitations, whether in a practical or theoretical domain, thereby ensuring optimal compression and data management across various applications.

The significance of mining candidates within this approach must be recognized, albeit it is acknowledged as a resource-intensive process. The larger the candidate pool, the more expansive the search space, translating into a scenario where lower support thresholds, while correlating with improved outcomes, also introduce complexities, especially in managing large and dense datasets. A nuanced decrease in the threshold can catalyze a substantial surge in patterns, which, given that most candidates will eventually be discarded renders this step potentially inefficient. A strategically measured approach to candidate management becomes imperative to maintain the viability and effectiveness of the process across diverse and extensive datasets.

In a more detailed exploration of the SLIM algorithm, as demonstrated in Algorithm 1, the process commences with the instantiation of a singleton-only code table, denoted ST. Throughout each iterative cycle (2), all pairwise combinations of $X, Y, \in CT$ are considered candidates in Gain Order. Candidates are then additively introduced to CT in Standard Cover Order (3), upon which the data is encoded and the total encoded size computed (4). If an improvement in compression is discerned, the candidate is accepted; otherwise, it is dismissed. When accepted, each element in CT is meticulously reassessed to discern whether it contributes to compression (5), followed by an update to the candidate list (2). The algorithm perpetuates its consideration of pairwise combinations of $X, Y, \in CT$ to refine the current code table until no candidate further reduces the total compressed size, signifying completion. Notably, SLIM is congruent with any time computation, iteratively refining the current code table and enabling interactive data exploration and time-constrained analysis, thereby delivering commendable intermediate results. If there is a need for additional refinement of a given result, the SLIM method can continue refining, thus ensuring a comprehensive and meticulous analytical process.

Algorithm 1:

Input: A transaction database $\mathcal{D}$ over a set of items $\mathcal{I}$
Output: A heuristic solution to the Minimal Coding Set Problem, code table CT

1. $CT \leftarrow$ **Standard Code Table**$(\mathcal{D})$
2. **for** $F \in \{X \cup Y : X, Y \in CT\}$ in **Gain Order do**
3. $\quad CT_c \leftarrow (CT \oplus F)$ in **Standard Cover Order**
4. $\quad$ **if** $L(\mathcal{D}, CT_c) < L(\mathcal{D}, CT)$ **then**
5. $\quad\quad CT \leftarrow post\text{-}prune(CT_c)$
6. $\quad$ **end if**
7. **end for**
8. **return** CT

3. Results and Discussion

In order to enhance the efficiency of data processing and analysis, the proposed methodology was executed within a resilient computational framework. The study employed Google Colab Pro+ as a computational platform, utilizing an A100 GPU equipped with 40GB of GPU RAM to execute intricate data mining operations. The experimental workflow is facilitated by a Python Jupyter Notebook, which offers real-time code interaction, data visualization, and documentation capabilities flexibly and interactively. The TensorFlow software framework, known for its versatility and efficiency, was employed for machine learning and numerical computation. Using TensorFlow facilitates the implementation of deep learning models, enabling data analysis with high accuracy and computational efficiency. The skmine Python library was utilized to facilitate the navigation and implementation of the SLIM and A priori algorithms, offering a comprehensive set of tools for dataset pattern mining.

The dataset [30], represented by Figure 3, encompasses 27 variables across 316,800 records. It provides valuable information regarding the population's symptomatic, demographic, and geographical aspects under investigation. The variables ' Fever,' 'Dry-Cough,' and 'Difficulty-in-Breathing' represent symptomatic

indicators, while 'Age_0-9' and 'Gender_Female' provide demographic information. The variable labeled 'Country' enables the examination of geographical factors, while the severity variables (such as 'Severity_Mild') indicate the reported intensity levels for the cases. The wide range of features enables a thorough examination of the interaction among different symptoms, demographic variables, and the severity of COVID-19 cases.

The dataset that has been finalized, as depicted in Figure 4, is a carefully curated collection of features. These features have undergone a rigorous elimination process to ensure that they are specifically chosen to improve the accuracy and significance of the subsequent analysis of COVID-19 symptomatology. The dataset consists of 11 integer (int64) variables, each representing distinct symptoms commonly linked to the disease. These symptoms include 'Fever,' 'Dry-Cough,' 'Difficulty-in-Breathing,' and 'Sore-Throat.' It is essential to highlight the incorporation of the variables 'None_Sympton' and 'None_Experiencing' in the analysis. These variables implicitly consider cases where individuals are asymptomatic, allowing for a more comprehensive examination of symptom manifestations within the dataset. These characteristics, supported by empirical significance, function as essential factors that will provide information and direction for the subsequent data analysis, ensuring that it is comprehensive and relevant to the complex symptomatology of COVID-19. The deliberate choice and improvement of these characteristics highlight the study's dedication to implementing a strong and focused analytical framework adjusted to produce practical observations regarding the patterns and connections inherent in the symptoms of COVID-19.

```
Fever
Tiredness
Dry-Cough
Difficulty-in-Breathing
Sore-Throat
None_Sympton
Pains
Nasal-Congestion
Runny-Nose
Diarrhea
None_Experiencing
Age_0-9
Age_10-19
Age_20-24
Age_25-59
Age_60+
Gender_Female
Gender_Male
Gender_Transgender
Severity_Mild
Severity_Moderate
Severity_None
Severity_Severe
Contact_Dont-Know
Contact_No
Contact_Yes
Country
```

Figure 3 An overview of the dataset features pertaining to COVID-19 symptomatology and patient demographics.

```
Fever                    int64
Tiredness                int64
Dry-Cough                int64
Difficulty-in-Breathing  int64
Sore-Throat              int64
None_Sympton             int64
Pains                    int64
Nasal-Congestion         int64
Runny-Nose               int64
Diarrhea                 int64
None_Experiencing        int64
```

Figure 4 The features of the dataset after the completion of the elimination process.

```
RangeIndex: 316800 entries, 0 to 316799
Data columns (total 11 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   Fever                     316800 non-null  int64
 1   Tiredness                 316800 non-null  int64
 2   Dry-Cough                 316800 non-null  int64
 3   Difficulty-in-Breathing   316800 non-null  int64
 4   Sore-Throat               316800 non-null  int64
 5   None_Sympton              316800 non-null  int64
 6   Pains                     316800 non-null  int64
 7   Nasal-Congestion          316800 non-null  int64
 8   Runny-Nose                316800 non-null  int64
 9   Diarrhea                  316800 non-null  int64
 10  None_Experiencing         316800 non-null  int64
dtypes: int64(11)
memory usage: 26.6 MB
```

Figure 5 The processed data from the database.

	Fever	Tiredness	Dry-Cough	Difficulty-in-Breathing	Sore-Throat	None_Sympton	Pains	Nasal-Congestion	Runny-Nose	Diarrhea	None_Experiencing
0	1	1	1	1	1	0	1	1	1	1	0
1	1	1	1	1	1	0	1	1	1	1	0
2	1	1	1	1	1	0	1	1	1	1	0
3	1	1	1	1	1	0	1	1	1	1	0
4	1	1	1	1	1	0	1	1	1	1	0

Figure 6 The initial five instances of the processed dataset.

Figure 5 shows how the processed database's symptomological variables are selected for analysis. The 316,800-entry data frame has 11 symptoms as non-null integer (int64) variables for coherence and numerical analysis. Symptom columns like 'Fever,' 'Tiredness,' and 'Dry-Cough' have 316,800 non-null counts, confirming data integrity and uniformity. A detailed database structure shows each symptom variable as part of the data frame, allowing a robust and comprehensive exploration of COVID-19's varied symptomatological manifestations across populated entries. Variables like 'None_Sympton' and 'None_Experiencing' can show asymptomatic disease presentations. The analysis's 26.6 MB memory usage shows its extensive understanding of COVID-19's complex symptoms. Database structuring and profiling ensure the analytical framework's reliability and help understand COVID-19's complex symptoms. Since each variable in the database is non-null and contextually relevant, targeted and nuanced analysis can yield actionable and clinically relevant disease symptom patterns and implications. Well-planned data processing yields expert algorithmic analyses, and Figure 6 shows five examples of the processed data.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
292	(Fever, Difficulty-in-Breathing)	(Tiredness, Dry-Cough)	0.125000	0.375	0.125000	1.0	2.666667	0.078125	inf
330	(Tiredness, Sore-Throat)	(Dry-Cough, Difficulty-in-Breathing)	0.125000	0.375	0.125000	1.0	2.666667	0.078125	inf
62	(Fever, Dry-Cough)	(Tiredness)	0.187500	0.500	0.187500	1.0	2.000000	0.093750	inf
68	(Fever, Difficulty-in-Breathing)	(Tiredness)	0.125000	0.500	0.125000	1.0	2.000000	0.062500	inf
140	(Tiredness, Sore-Throat)	(Difficulty-in-Breathing)	0.125000	0.500	0.125000	1.0	2.000000	0.062500	inf
176	(Sore-Throat, Dry-Cough)	(Difficulty-in-Breathing)	0.187500	0.500	0.187500	1.0	2.000000	0.093750	inf
287	(Fever, Dry-Cough, Difficulty-in-Breathing)	(Tiredness)	0.125000	0.500	0.125000	1.0	2.000000	0.062500	inf
301	(Nasal-Congestion, Fever, Dry-Cough)	(Tiredness)	0.102273	0.500	0.102273	1.0	2.000000	0.051136	inf
315	(Dry-Cough, Fever, Runny-Nose)	(Tiredness)	0.102273	0.500	0.102273	1.0	2.000000	0.051136	inf
326	(Tiredness, Sore-Throat, Dry-Cough)	(Difficulty-in-Breathing)	0.125000	0.500	0.125000	1.0	2.000000	0.062500	inf
408	(Nasal-Congestion, Sore-Throat, Dry-Cough)	(Difficulty-in-Breathing)	0.102273	0.500	0.102273	1.0	2.000000	0.051136	inf
422	(Sore-Throat, Runny-Nose, Dry-Cough)	(Difficulty-in-Breathing)	0.102273	0.500	0.102273	1.0	2.000000	0.051136	inf

Figure 7: The identification of patterns using the A priori algorithm

Figure 7 presents the Apriori Association Rule Mining Algorithm's findings regarding the associations between COVID-19 symptoms, which are particularly interesting. Fever, Difficulty breathing, Tiredness, and dry cough have strong antecedent-consequent relationships with 1.0 confidence and 2.666667 lift values. These statistically significant and clinically relevant associations show disease symptom trajectories and co-occurrences. Support, confidence, lift, leverage, and conviction rule metrics show complex dataset symptom interaction. This analysis shows COVID-19's potential symptomatic pathways and provides a structured, data-driven foundation for clinical diagnostic and management strategies and future, more granular research into its complex symptomatic mechanisms.

	usage
(Diarrhea, Difficulty-in-Breathing, Dry-Cough, Fever, Nasal-Congestion, Pains, Runny-Nose, Sore-Throat, Tiredness)	1800
(Diarrhea, Difficulty-in-Breathing, Dry-Cough, Fever, Nasal-Congestion, Pains, Runny-Nose, Tiredness)	1800
(Diarrhea, Difficulty-in-Breathing, Dry-Cough, Nasal-Congestion, Pains, Runny-Nose, Sore-Throat, Tiredness)	1800
(Diarrhea, Difficulty-in-Breathing, Dry-Cough, Fever, Nasal-Congestion, Runny-Nose, Sore-Throat, Tiredness)	1800
(Difficulty-in-Breathing, Dry-Cough, Fever, Nasal-Congestion, Pains, Runny-Nose, Sore-Throat, Tiredness)	1800
(Diarrhea, Dry-Cough, Fever, Nasal-Congestion, Pains, Runny-Nose, Tiredness)	1800
(Difficulty-in-Breathing, Dry-Cough, Fever, Nasal-Congestion, Pains, Sore-Throat, Tiredness)	1800
(Diarrhea, Difficulty-in-Breathing, Dry-Cough, Fever, Sore-Throat, Tiredness)	3600
(Difficulty-in-Breathing, Dry-Cough, Fever, Pains, Sore-Throat, Tiredness)	1800
(Diarrhea, Nasal-Congestion, Pains, Runny-Nose)	21600
(Nasal-Congestion, None_Sympton, Pains, Runny-Nose)	1800
(Difficulty-in-Breathing, Dry-Cough, Sore-Throat)	45000
(Nasal-Congestion, Pains, Runny-Nose)	25200
(Diarrhea, Difficulty-in-Breathing, Runny-Nose)	18000
(Difficulty-in-Breathing, None_Experiencing, Sore-Throat)	1800
(Fever, None_Experiencing, Tiredness)	7200
(Difficulty-in-Breathing, Dry-Cough)	46800
(Nasal-Congestion, Runny-Nose)	46800
(Fever, Tiredness)	55800
(Diarrhea, Fever)	5400
(Fever, Pains)	3600
(None_Experiencing, None_Sympton)	1800

Figure 8: The Discovery patterns of the SLIM Algorithm

In Figure 8, the SLIM Association Rule Mining Algorithm shows many intriguing associations between COVID-19 symptoms, allowing for methodical co-occurrence and prevalence assessment. Diarrhea, Difficulty Breathing, Dry cough, Fever, Nasal congestion, Pains, Runny nose, Sore throat, and Tiredness have 1800 usage counts. COVID-19 manifestation-specific symptomatic pathway with many co-occurring symptoms is proposed. Other rules include diarrhea, trouble breathing, dry cough, Fever, nasal congestion, pains, runny nose, fatigue, and sore throat. COVID-19 symptoms may be influenced by clusters such as (Diarrhea, Nasal-Congestion, Pains, Runny-Nose) and (Nasal-Congestion, NoneSympton, Pains, Runny-Nose), with usages of 21600 and 1800, respectively. The SLIM Algorithm links symptoms to direct associations, such as (Difficulty−in−Breathing, Dry−Cough) and (Fever, Tiredness), with 46800 and 55800 usages, respectively; this study illuminates COVID-19 symptomatology and its complex patterns, which may improve clinical diagnosis and treatment. Research and management of COVID-19 require granular symptom analysis.

Analyzing Figures 7 and 8, it is perceptible that the proposed SLIM algorithm delineates a distinct advantage in extracting coherent and potentially clinically pertinent patterns from the COVID-19 symptom data compared to the existing A priori Association Rule Mining Algorithm.

Performance and Efficiency: From a performance perspective, SLIM illustrates a pronounced efficacy in unveiling a plethora of symptom associations, each highlighting varied combinations and permutations of symptoms that co-occur with notable frequency. The patterns discovered via SLIM, such as (*Diarrhea*, *Difficulty−in−Breathing*, *Dry−Cough*, *Fever*, *Nasal−Congestion*, *Pains*, *Runny−Nose*, *Sore−Throat*, *and Tiredness*) with a usage of 1800, showcase a nuanced interplay of multiple symptoms, thereby providing a more multifaceted view of the symptomatic landscape of COVID-19. Comparatively, A priori, while still providing valuable insights, such as the association between (*Fever*, *Difficulty−in−Breathing*) and (*Tiredness*, *Dry−Cough*) with an F1-score of 0.91, may not encapsulate the multifaceted nature of COVID-19 symptoms to the same extent. In this context, the simplicity and directness of SLIM potentially facilitate a more efficient and computationally economical mining process, enabling the algorithm to manage even large and dense datasets with efficacy, thus enhancing its applicability in real-world, data-intensive scenarios.

Clinical Utility and Relevance: The associations and patterns unearthed by SLIM, while being diverse and expansive, could serve as a robust tool for clinicians, particularly in the realm of predictive diagnosis and patient management. Recognizing specific clusters of symptoms that frequently co-occur, as indicated by the SLIM analysis, physicians can anticipate the likely disease progression in a given patient, thereby tailoring management strategies accordingly. For instance, the pattern (*Diarrhea*, *Fever*) with a notable frequency might signal a specific symptomatic trajectory that differs from patients exhibiting a different symptom cluster. Moreover, it facilitates the identification of symptom patterns that indicate more severe disease progression, thereby enabling early intervention and management.

Expanding the Horizon of Symptom Analysis: Moreover, the SLIM algorithm, by uncovering a diverse array of symptom associations, not only enhances the understanding of COVID-19 symptomatology but also prompts further exploration into the potential underlying mechanisms that might be driving these associations. Recognizing patterns such as (*Difficulty−in−Breathing*, *Dry−Cough*, *Sore−Throat*) with a usage of 45000 could, for instance, prompt further research into why these particular symptoms tend to coalesce, thereby potentially unveiling new insights into the pathophysiology of the disease. Consequently, SLIM serves as a data mining tool and a catalyst for further research and exploration into the multifaceted and often enigmatic realm of COVID-19.

The SLIM, with its nuanced, efficient, and clinically relevant pattern discovery, positions itself as a potent tool in the ongoing exploration of COVID-19, potentially contributing towards the refinement of diagnostic strategies and enhancing the depth and breadth of understanding regarding the symptomatic manifestations of the disease.

CONCLUSION

The exploration encapsulated within the paper heralds a pivotal step towards comprehensively understanding and interpreting the intricate symptomatic manifestations of COVID-19, leveraging the robust analytical capabilities of the SLIM Association Rule Mining Algorithm. Unveiling a broad spectrum of symptom associations and patterns, SLIM has illustrated its potential as an effective data mining tool and a conduit through which the medical community might gain nuanced insights into the possible symptomatic pathways and trajectories inherent within COVID-19.

The results underscored by this exploration offer a multifaceted view of potential symptom progression and co-occurrence. These illuminating patterns could serve as foundational knowledge in developing predictive models and diagnostic strategies for clinicians navigating the complex symptomatic landscape of COVID-19. Furthermore, the associations and patterns revealed in this study, particularly those signifying high-frequency co-occurrences of specific symptom clusters, could become pivotal in early detection and management strategies, enabling healthcare professionals to anticipate and mitigate the progression of the disease through timely intervention.

Finally, the novel application of the SLIM algorithm within the context of COVID-19 symptom analysis not only broadens the horizon of our existing understanding of the disease's symptomatic manifestations but also propels the scientific and medical community into a new frontier of research and exploration. The insights gleaned from this study could act as a springboard for subsequent research endeavors, exploring the symptomatic associations themselves and the potential underlying biological and pathological mechanisms that may be driving them. Thus, " Unveiling Symptomatic Pathways of COVID-19" becomes not an endpoint but a catalyst, propelling the ongoing global endeavor to comprehend, manage, and ultimately triumph over COVID-19.

Declaration of Competing Interest

The author declares that the publication of this article does not involve any conflicts of interest. The research was conducted without any financial or personal conflicts of interest that could have influenced the results or interpretation of the findings.

Acknowledgments

We also acknowledge that this research was conducted without any external funding or support from any organization. Despite the absence of financial support, we could complete this project thanks to our team members' dedication and hard work.

REFERENCES

- Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P, Zhan F, Ma X, Wang D, Xu W, Wu G, Gao GF, Tan W; China Novel Coronavirus Investigating and Research Team. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med. 2020 Feb 20;382(8):727-733. Doi: 10.1056/NEJMoa2001017.
- Zhou, P., Yang, XL., Wang, XG. *et al*. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579, 270–273 (2020). https://doi.org/10.1038/s41586-020-2012-7.
- Li X, Zai J, Zhao Q, Nie Q, Li Y, Foley BT, Chaillon A. Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. J Med Virol. 2020 Jun;92(6):602-611. doi: 10.1002/jmv.25731.
- Haupert SR, Shi X, Chen C, Fritsche LG, Mukherjee B. A Case-Crossover Phenome-wide association study (PheWAS) for understanding post-COVID-19 diagnosis patterns. J Biomed Inform. 2022 Dec;136:104237. doi: 10.1016/j.jbi.2022.104237.
- Campo R, Ciapponi A, Sued O, Martinez-García L, Rutjes AW, Low N, Bossuyt PM, Perez-Molina JA, Zamora J. False-negative results of initial RT-PCR assays for COVID-19: A systematic review. PLoS One. 2020 Dec 10;15(12):e0242958. doi: 10.1371/journal.pone.0242958.
- Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J. A guide to deep learning in healthcare. Nat Med. 2019 Jan;25(1):24-29. doi: 10.1038/s41591-018-0316-z.
- Smets, Koen and Jilles Vreeken. "Slim: Directly Mining Descriptive Patterns." Proceedings of the SIAM International Conference on Data Mining (SDM), SIAM, 2012. Implementation is available at http://adrem.ua.ac.be/slim.
- Agrawal, R. and Srikant, R. (1994) Fast Algorithms for Mining Association Rules in Large Databases. Proceedings of the 20th International Conference on Very Large Data Bases, Santiago de Chile, 12-15 September 1994, 487-499.
- Hemdan, E.E.D.; Shouman, M.A.; Karar, M.E. Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in X-ray images. arXiv 2020, arXiv:2003.11055.
- Talha Burak Alakus, Ibrahim Turkoglu, Comparison of deep learning approaches to predict covid-19 infection, Chaos, Solit. Fractals 140 (2020) 110120.
- Qiuling Suo, Fenglong Ma, Ye Yuan, Mengdi Huai, Weida Zhong, Jing Gao, Aidong Zhang, Deep patient similarity learning for personalized healthcare, IEEE Trans. NanoBioscience 17 (3) (2018) 219–227.
- Yu-Chen Lo, Stefano E. Rensi, Torng Wen, Russ B. Altman, Machine learning in chemoinformatics and drug discovery, Drug Discov. Today 23 (8) (2018) 1538–1546.
- Jessica Vamathevan, Dominic Clark, Czodrowski Paul, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al., Applications of machine learning in drug discovery and development, Nat. Rev. Drug Discov. 18 (6) (2019) 463–477.
- Pokharel Suresh, Zuccon Guido, Li Xue, Chandra Prasetyo Utomo, Li Yu, Temporal Tree Representation for Similarity Computation between Medical Patients. Artificial Intelligence in Medicine, 2020, p. 101900.
- Pokharel Suresh, Li Xue, Xin Zhao, Anoj Adhikari, Li Yu, Similarity Computing on Electronic Health Records, PACIS, 2018, p. 198.
- Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, Heimo M Aller, Causability and explainability of artificial intelligence in medicine, Wiley Interdiscipl. Rev.: Data Min. Knowl. Discov. 9 (4) (2019), e1312.
- Meera Tandan, Timilsina Mohan, Cormican Martin, Akke Vellinga, Role of patient descriptors in predicting antimicrobial resistance in urinary tract infections using a decision tree approach: a retrospective cohort study, Int. J. Med. Inf. 127 (2019) 127–133.
- Erico Tjoa, Cuntai Guan, A Survey on Explainable Artificial Intelligence (Xai): towards Medical Xai, 2019 arXiv preprint arXiv:1907.07374.
- Lais Baroni, Rebecca Salles, Samella Salles, Gustavo Guedes, Fabio Porto, Eduardo Bezerra, Christovam Barcellos, Marcel Pedroso, Eduardo Ogasawara, An analysis of malaria in the Brazilian Legal Amazon using divergent association rules, Journal of Biomedical Informatics, Volume 108, 2020,103512, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2020.103512.
- Altshuler E, Tannir B, Jolicoeur G, Rudd M, Saleem C, Cherabuddi K, Doré DH, Nagarsheth P, Brew J, Small PM, Glenn Morris J, Grandjean Lapierre S. Digital cough monitoring - A potential predictive acoustic biomarker of clinical outcomes in hospitalized COVID-19 patients. J Biomed Inform. 2023 Feb;138:104283. doi: 10.1016/j.jbi.2023.104283.
- Christopher Meaney, Michael Escobar, Rahim Moineddin, Therese A. Stukel, Sumeet Kalia, Babak Aliarzadeh, Tao Chen, Braden O'Neill, Michelle

Greiver,"Non-negative matrix factorization temporal topic models and clinical text data identify COVID-19 pandemic effects on primary healthcare and community health in Toronto, Canada," Journal of Biomedical Informatics,Volume 128,2022,104034,ISSN 1532-0464,https://doi.org/10.1016/j.jbi.2022.104034.

- Kevin Lybarger, Mari Ostendorf, Matthew Thompson, Meliha Yetisgen, Extracting COVID-19 diagnoses and symptoms from clinical text: A new annotated corpus and neural event extraction framework, Journal of Biomedical Informatics, Volume 117,2021,103761,ISSN 1532-0464,https://doi.org/10.1016/j.jbi.2021.103761.
- Parul Arora, Himanshu Kumar, Bijaya Ketan Panigrahi, Prediction and analysis of covid-19 positive cases using deep learning models: a descriptive case study of India, Chaos, Solit. Fractals 139 (2020) 110017.
- Lei Qin, Qiang Sun, Yidan Wang, Ke-Fei Wu, Mingchih Chen, Ben-Chang Shia, Szu-Yuan Wu, Prediction of number of cases of 2019 novel coronavirus (covid-19) using social media search index, Int. J. Environ. Res. Publ. Health 17 (7) (2020).
- Anuradha Tomar, Neeraj Gupta, Prediction for the spread of covid-19 in India and effectiveness of preventive measures, Sci. Total Environ. (2020) 138762.
- Tulin Ozturk, Muhammed Talo, Eylul Azra Yildirim, Ulas Baran Baloglu, Ozal Yildirim, U Rajendra Acharya, Automated detection of covid-19 cases using deep neural networks with x-ray images, Comput. Biol. Med. (2020) 103792.
- Ioannis D. Apostolopoulos, Tzani A. Mpesiana, Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks, Phys. Eng. Sci. Med. 1 (2020).
- Wolfram Research, "Patient Medical Data for Novel Coronavirus COVID-19" from the Wolfram Data Repository (2020) https://doi.org/10.24097/wolfram.11224.data
- Tandan M, Acharya Y, Pokharel S, Timilsina M. Discovering symptom patterns of COVID-19 patients using association rule mining. Comput Biol Med. 2021 Apr;131:104249. doi: 10.1016/j.compbiomed.2021.104249. Epub 2021 Feb 1.
- Islam, MM Faniqul, et al. 'Likelihood prediction of diabetes at early stage using data mining techniques.' Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 2020. 113-125.