Exploring Data Mining Techniques in COVID-19 Research

Alexandros Ioannou[#1], Cameron Anderson[$2], Lubna Al Rifaie[*3], Luke Aikman[^4]
[1]ioan9620@mylaurier.ca, [2]ande9930@mylaurier.ca, [3]alri1590@mylaurier.ca, [4]aikm2750@mylaurier.ca

***Abstract*—The COVID-19 pandemic posed unprecedented and unpredictable challenges across the world, requiring the deployment of complex data mining techniques to manage and mitigate the impact of the virus. This paper reviews recent research works that apply various data mining methods, such as Natural Language Processing (NLP), supervised learning techniques, clustering algorithms, frequent itemsets, and association rules, to address the big data challenges within the healthcare sector during the COVID-19 crisis. By analyzing these methods' applications in tracking the virus spread, predicting outcomes, and enhancing healthcare responses, we aim to identify existing data mining problems and propose viable solutions. The insights from this study show the potential of data mining in revolutionizing healthcare monitoring and information technology, particularly in managing pandemic situations.**

***Keywords*—Data mining(DM), COVID-19, Healthcare monitoring, Clustering algorithms, Frequent itemsets, Association rules, Big Data, Predictive analytics.**

I. Introduction

The emergence of the COVID-19 pandemic has not only brought up a global health crisis but also led to an increased dependence on data science and informatics in public health responses. The vast amounts of data produced by health monitoring systems, contact tracking, and case reporting offer both opportunities and challenges. Understanding the complexity of COVID-19 data has become more dependent on data mining techniques, which have historically been essential in extracting relevant information from huge datasets. To support well-informed healthcare decisions and public health policies this study explores the various applications of these techniques – ranging from NLP and supervised learning to association rules – in filtering through data linked to pandemics.

One of the methods in the reviewed papers was the use of Natural Language Processing. This method was a key tool for tracking misinformation, analyzing sentiment on social media, and compiling clinical data to comprehend the dynamics of the pandemic. Similarly, significant effectiveness was shown by supervised learning algorithms in forecasting infection rates, and patient outcomes, and identifying high-risk groups, which allowed for more focused interventions. Also, clustering algorithms have made it possible to stratify patient groups and implement individualized treatment plans by providing insights into patient symptomatology and disease development.

However, the deployment of data mining in healthcare, particularly in a crisis of this magnitude, is full of challenges. Data quality and availability remain significant obstacles, with the heterogeneity of data sources and formats complicating analysis efforts. The ethical considerations around data privacy and the potential for bias in data-driven decisions emphasize the need for strict guidelines and transparent methodologies.

Through the integration of multiple data mining methodologies, this paper assesses the benefits of each technique's application in COVID-19 research, as well as its ability to handle diverse datasets, predict outcomes, and recognize patterns in the healthcare domain.

Additionally, this paper also highlights the limitations of certain data mining techniques, as illustrated by the occasional misinterpretation of data by K-Means Clustering or the challenges in data accessibility. It also draws attention to the emerging field of long-term COVID research, where, despite the lack of comprehensive studies, data mining has started to unravel the intricate relationship between demographic parameters and long-term symptomatology.

In this extended analysis, we aim to provide a comprehensive view of the potential of data mining in revolutionizing healthcare monitoring and information technology in the face of a global health emergency. The insights gained from this exploration are expected to make a significant contribution to the ongoing discourse on leveraging advanced data analytics for pandemic response and beyond.

II. Related Work (Literature Review)

A. Using data mining techniques for deep analysis and theoretical investigation of the COVID-19 pandemic [3]

This article covers the use of K-Means Clustering as a method for isolating COVID-19 and determining the overall disease link. Throughout this article, several different data mining techniques are used to achieve this with an emphasis on K-Means. These techniques include converting data, feature extraction, cleaning the data, and more. These data mining techniques are used to focus on the COVID-19 outbreak in Iraq where the first case was confirmed on February 21, 2020.

The aim is to apply K-Means to samples of human body data to produce coronavirus cases and infections using preexisting data. K-Means is chosen as the primary method because of the simplicity the model has as well as the scalability in terms of categorization. It is a very modular method allowing for simple or complex problems to be relatively easy to solve. The parameters for the model will stay smaller as the difficulty of assessing a model with a large

number of parameters properly becomes too much of a challenge for the problem at hand.

Through the literature review within this article, a lot of information is digested where other authors fail to extract COVID-19 using similar approaches because of noisy data [8] or authors have denied proposals or produced inaccurate results [9]. Lots are learned here as solutions including modified threshold values, suppressed background noise, and limiting data exposure using histogram equalization are improvements that can increase the chance of improved results.

As the normal life in Iraq changed variables were added accordingly including a component related to the fact that a lot of Iraq would be in isolation. With all these variables and parameters, K-Means still has a chance to evaluate information that is not relevant or bypass information that is but K-Means is still used. K-Means is the most effective method for comparing outcomes based on results. As shown in Figure 1, the outcomes show that K-Means was a beneficial choice for data mining. The incidence rate increases in proportion to the transmission rate and all the data is taken from a large pool of situations where the illumination varies.

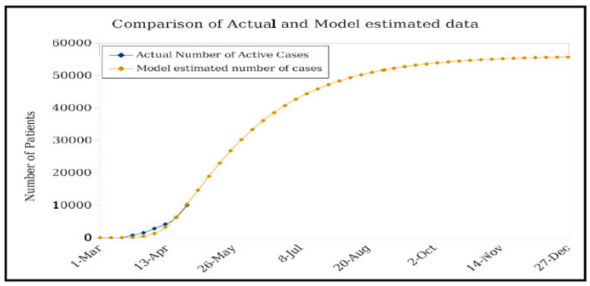

Fig. 1. A curve showing the actual active cases in comparison to the model's estimated number of cases.

As displayed in Figure 1, there is a strong correlation between the actual number of active cases versus the predicted cases. This scenario can be altered very easily where parameters like the spread of the disease, mortality rate, recovery rate and isolation can change as time passes. Still, the model using K-Means was successful in predicting and giving some insight into how COVID-19 was to pan out in Iraq.

B. Mining Big Healthcare Data to Predict Long COVID Cases [4]

This paper focuses on long-term COVID-19, a common term for a condition post-COVID-19 and developing a potential solution to predict the development of long-term COVID-19 in patients. The authors aim to discover associations in long COVID demographic data and the grouping of common symptoms.

Long COVID, to be more specific, is defined as the development of new symptoms closer to three months after initially contracting COVID-19. These symptoms can last upwards of 2 months and can affect anybody regardless of their age or how severe their initial symptoms from COVID-19 were.

The authors' solution takes the demographic information to create a frequent itemset of characteristics to then produce association rules, creating relationships between these characteristics. Data processing will then be carried out like cleaning and reading solutions into a data frame taking an array form. After the preprocessing, attributes like age range, race, birth sex, symptom severity, and many more will be used in demographic analysis, pattern mining and also classification and prediction.

Once the demographic information is analyzed frequent patterns are determined with attribute values frequently co-occurring with COVID or long COVID conditions. Here, a frequent pattern is found if its support meets the minimum support threshold which is called *minsup* here. It can be calculated like this: $minsup = 0.1 + \frac{1}{e^x}$ where x increases where the number of dataset rows increases. Once the frequent patterns are found association rules can be created where the rules are deemed interesting if they meet a minimum confidence level that the authors have set at 0.3. The association rules are then split into two categories long_COVID_0 and long_COVID_1 where no long COVID and long COVID are the consequent Y respectively.

There were very few datasets and information the authors could pull from to gather knowledge between demographic and symptom information which was one of the big challenges the authors faced.. The few datasets they were able to use containing long COVID data included survey data from the UN Office for the Coordination of Humanitarian Affairs [10] and US Census Bureau Data [11].

With all this information gathered, some results can be created and put into graphs and tables showing attributes like ethnicity ratios and long COVID patients by ethnicity (see Figure 2) to COVID and long COVID based on symptom severity (see Figure 3).

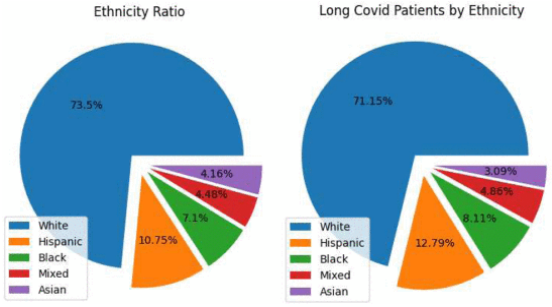

Fig. 2. Ethnicity ratio and long-term COVID patients

Severity	COVID-19	Long COVID	Changes
Severe	14.48%	22.40%	+7.92%
Moderate	41.33%	47.31%	+5.98%
No symptom	5.30%	3.09%	-2.21%
Mild	38.90%	28.80%	-10.10%

Fig. 3. Symptom Severity

Overall, with the information from demographic analysis of both COVID-19 and long-term COVID patients, the authors can successfully mine frequent itemsets and create interesting association rules and produce results for predicting long-term COVID in patients.

C. Big data analytics for preventive medicine [5]

The integration of modern data analytics into the healthcare sector holds the potential to revolutionize the way healthcare providers approach disease prevention, diagnosis, and treatment. The capability to analyze vast amounts of unstructured, heterogeneous, non-standard, and incomplete medical data can unveil patterns that are vital for enhancing patient care quality and reducing healthcare costs. This project report delves into the advancements of data analytics methods tailored for disease prevention, providing a structured overview of both traditional and contemporary methodologies, their respective merits, limitations, and the criteria for selecting appropriate models for specific scenarios.

Medical data presents both a valuable resource and a significant challenge due to its complexity and volume. Efficiently mining this data for actionable insights can lead to breakthroughs in patient care and substantial reductions in healthcare expenses. The focus of this study is on exploring the advancements in data analytics that facilitate disease prevention. By offering a comprehensive review of various analytics algorithms and their applications, this report aims to highlight the role of data analytics in the ongoing evolution of healthcare services.

Disease Prevention: Challenges and Traditional Methodologies

Disease prevention encompasses strategies and practices aimed at reducing the risk of developing health problems. This section outlines the inherent challenges in disease prevention, such as the need for early detection, lifestyle and environmental factors, and genetic predispositions to certain diseases. It also reviews traditional methodologies, which primarily rely on public health campaigns, vaccination programs, and lifestyle modifications, underscoring their limitations in addressing complex health issues that require personalized approaches.

Advancements in Data Analytics for Disease Prevention
Data Analytics Algorithms

The core of modern disease prevention strategies lies in the application of advanced data analytics algorithms. This section provides an in-depth analysis of

Classification Algorithms: Used for categorizing data into predefined groups, which is essential for diagnosing diseases based on symptoms and test results.

Clustering Algorithms: Focus on identifying unusually high incidences of diseases within a population, aiding in the early detection of outbreaks.

Anomaly Detection: Critical for recognizing abnormal patterns that may indicate the onset of a disease, enabling proactive intervention.

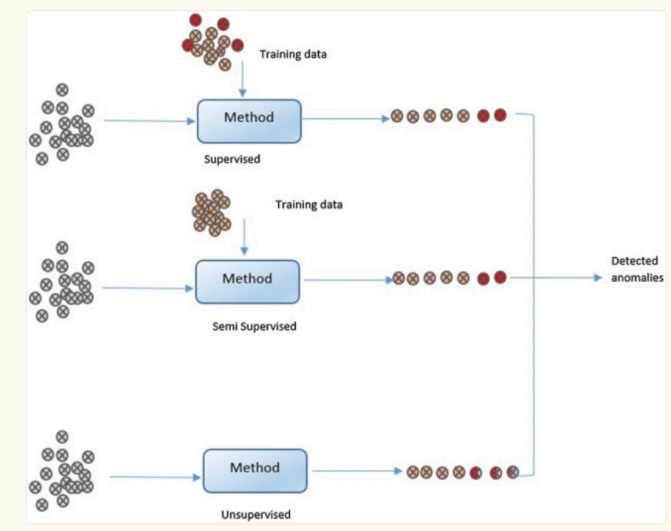

Fig. 4. Architecture of healthcare data analytics

Association Rule Learning: Helps in uncovering relationships between different variables in medical data, which can predict disease risk factors.

Each algorithm's advantages, drawbacks, and selection guidelines are discussed to inform healthcare providers about their application in various scenarios.

Recent Developments and Applications

This section highlights recent advancements in data analytics for disease prevention, including the integration of machine learning and artificial intelligence (AI). Successful case studies are presented to demonstrate how these technologies have been applied to predict outbreaks, improve diagnosis accuracy, and personalize treatment plans, thereby significantly impacting patient outcomes.

Challenges and Future Directions

Despite the promising developments in data analytics for disease prevention, several open research challenges remain. These include data privacy and security concerns, the need for standardized data formats, and the integration of analytics into existing healthcare systems. The report concludes with recommendations for future research directions, emphasizing the importance of interdisciplinary collaboration, the development of robust data governance frameworks, and ongoing innovation in analytics algorithms.

The advancement of data analytics in disease prevention represents a critical frontier in healthcare. By leveraging cutting-edge algorithms and technologies, healthcare providers can unlock valuable insights from complex medical data, leading to improved patient care, early disease detection, and more efficient healthcare systems. As this field continues to evolve, researchers and practitioners must address the existing challenges and explore new opportunities to maximize the potential of data analytics in transforming healthcare.

D. The State of the Art of Data Mining Algorithms for and Predicting the COVID-19 Pandemic [12]

The COVID-19 pandemic has underscored the importance of leveraging advanced data analytics and data mining algorithms for predicting disease spread and enhancing disease prevention strategies. The study by Cortés-Martínez et al. presents a thorough examination of various data mining algorithms and their integration with epidemiological models to forecast the evolution of the COVID-19 pandemic. This report aims to encapsulate the methodologies, findings, and significance of leveraging these technological advancements to combat the ongoing and future health crises effectively.

The outbreak of COVID-19 brought about an unprecedented challenge to global health systems, emphasizing the need for innovative solutions to predict and manage viral diseases. Traditional epidemiological models, while useful, often fall short in handling the vast and complex datasets generated during a pandemic. The integration of data mining algorithms with these models offers a promising avenue to enhance the accuracy and efficiency of disease predictions and prevention strategies.

The research conducted a comprehensive analysis of 35 studies, incorporating a wide array of data mining algorithms applied alongside epidemiological prediction models. The study focused on algorithms used for disease classification, clustering of disease incidences, anomaly detection, and association rule learning, assessing their advantages, drawbacks, and suitability for specific scenarios in disease prediction and prevention.

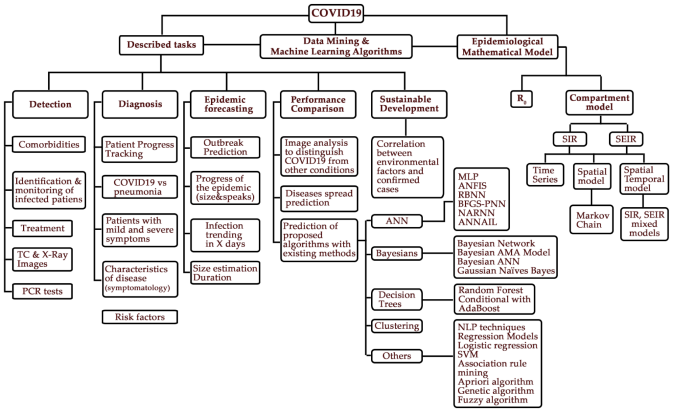

Fig. 5. Tree diagram with main concepts found during the literature review.

Data Mining Algorithms and Their Applications

Classification Algorithms: Essential for diagnosing diseases based on symptoms and test results, offering a systematic approach to identifying infected individuals.

Clustering Algorithms: Useful in identifying patterns and high-risk areas, aiding early detection of disease outbreaks.

Anomaly Detection: Plays a critical role in recognizing unusual patterns that may indicate disease emergence, facilitating timely interventions.

Association Rule Learning: Uncovers relationships between various factors and disease risks, enhancing understanding and prevention strategies.

Integration with Epidemiological Models

The study highlights the integration of data mining algorithms with epidemiological models, showcasing successful applications in predicting disease spread, assessing public health interventions, and optimizing disease management strategies. This synergy has proven instrumental in providing actionable insights for healthcare decision-makers.

Challenges and Recommendations

Despite the advancements, the study identifies several challenges, including data privacy concerns, the need for standardized data formats, and the integration of predictive analytics into existing healthcare systems. The report recommends fostering interdisciplinary collaborations, developing robust data governance frameworks, and encouraging continuous innovation in algorithm development to address these challenges effectively.

The convergence of data mining algorithms and epidemiological models represents a significant leap forward in disease prediction and prevention. By harnessing the power of big data and advanced analytics, healthcare providers can enhance patient care, optimize resource allocation, and implement effective intervention strategies. As the healthcare landscape continues to evolve, the continued exploration and adoption of these technologies will be paramount in managing current and future pandemics.

E. Using Data Mining Techniques to Fight and Control Epidemics: A Scoping Review [2]

This paper's main goal was to compile and analyze published articles to identify the most favored data mining methods and gaps in knowledge, particularly in the context of pandemics. This focus is a result of growing public health concerns about pandemic threats and the need for data mining techniques to successfully counter these threats.

A systematic search of the Web of Science, Scopus, and Pubmed databases from 2010 up to 16 Oct 2020 was done, using "data mining", "prediction model", "data mining techniques", "data mining methods", "pandemics", "pandemic", "COVID-19", "SARS-CoV-2", and "coronavirus disease" as keywords. Articles were included for review if they focused on pandemic diseases such as COVID-19, utilized data mining or knowledge discovery techniques as identified in the study by Patel and Patel [13], and were published in English. Exclusions applied to articles unrelated to pandemics or COVID-19, book chapters, letters, briefs, commentaries, reviews, non-English articles, those employing image processing methods, and any without accessible full text. Efforts were made to retrieve full texts from non-open access articles to minimize bias. See Figure 6 for a PRISMA diagram showcasing the process of how the studies were

determined as eligible. The study first produced 335 citations, which were then narrowed down to 50 eligible through a scoping review process. These databases were the primary sources for identifying relevant literature on the application of data mining techniques in this scenario.

The primary challenge discussed is the dynamic and complex nature of pandemic-related data, which often comes from sources such as social media platforms, hospital records, and public health databases. With new studies being published daily, the literature review conducted up to 16 Oct 2020 may not cover all relevant studies. Another challenge is the limitation of the literature search to only three journal databases. This approach overlooks research published in other databases or journals, having a potential bias on the review's outcomes and insights. Also, the study points out the exclusion of non-English papers from the review process. This leads to missing insights and findings published in other languages. Future research could benefit from incorporating non-English studies using automatic translation tools, which could provide a more global perspective on the subject.

Furthermore, the paper discusses the predominance of supervised learning techniques (90%) in the reviewed studies, highlighting their utility in predictive modelling and trend analysis of pandemic data. Natural Language Processing (NLP) was the most used technique (22%), showing its key

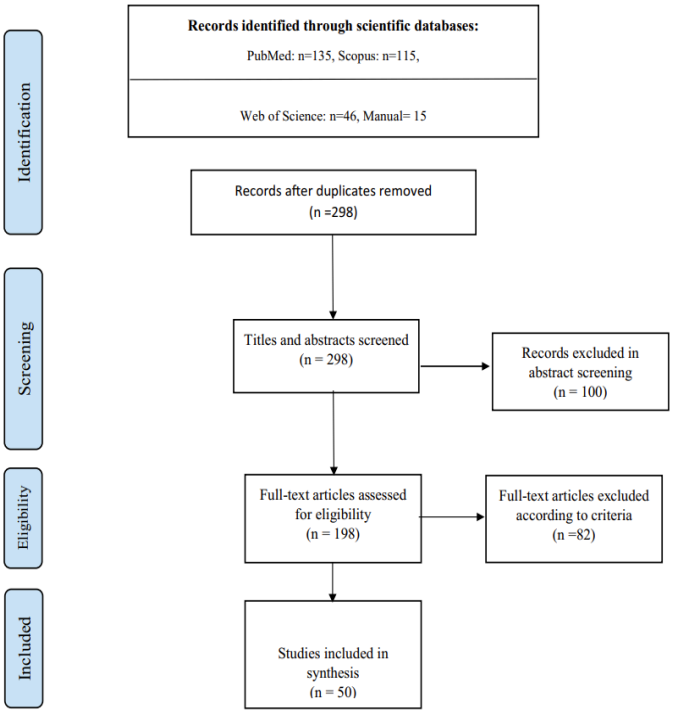

Fig. 6. PRISMA diagram for the identification, screening and eligibility of studies

and role in extracting information from unstructured data, such as social media posts and clinical notes (See Figure 7). These techniques, along with others like logistic regression and time series analysis, random forest, and Artificial Neural Networks (ANN) were highlighted for their potential in addressing aspects of a pandemic. These methods were chosen for their ability to efficiently process and analyze complex, large-scale data sets, to offer insights on disease characteristics, transmission patterns, and patient outcomes. The use of supervised learning techniques shows a preference for models that can be trained on labelled datasets to predict future pandemic trends and outcomes accurately.

The diversity of data sources and geographic focus of the studies reviewed, particularly on China, highlights the localized nature of initial pandemic research efforts. This geographic concentration may influence the generalizability of the findings, as most pandemics, including COVID-19, often begin or have impact points in specific regions. One crucial advancement is the emergence of social media as a unique data source that provides real-time monitoring of public sentiment and disease spread through text mining methods. The transition to digital data sources represents a significant evolution in pandemic research methodologies, providing a more dynamic and instantaneous viewpoint on public health emergencies.

DM techniques	Frequency		Studies
NLP techniques	11	22.00%	[20–30]
Logistic regression	10	20.00%	[31–40]
Time series	7	14.00%	[20, 41–46]
Random forest	7	14.00%	[47, 45, 48, 49, 42, 50, 51]
Regression models	7	12.00%	[52, 53, 40, 49, 54, 55, 39]
Decision tree	6	12.00%	[51, 48, 56–58,39]
ANN	5	10.00%	[52, 59, 60, 21, 61]
Naive Bayes	3	6.00%	[62–64]
SVM	2	4.00%	[49, 51]
Association rule mining	2	4.00%	[66, 58, 67]
Clustering	2	4.00%	[34, 30]
Apriori algorithm	1	2.00%	[65]
Genetic algorithm	1	2.00%	[55]
Fuzzy algorithm	1	2.00%	[41]

Fig. 7. Frequency of DM techniques in reviewed studies

F. Predicting the Incidence of COVID-19 using Data Mining [1]

The objective of this paper focuses on the urgent need for accurate predictions of COVID-19 incidence worldwide to assist health professionals in making decisions. The study utilized COVID-19 epidemiological data compiled by the Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE). The data, which is updated daily, contains information on COVID-19 cases in 252 geographic regions worldwide since January 22, 2020. More specifically, the study analyzed data up to March 29, 2020, comprising 17,136 records with variables including latitude, longitude, date, and case records. This data set was later used in the study to develop a model to predict COVID-19 incidence rates across different regions in the world.

The study encountered several challenges, including the need for a model that could accurately handle the variability in incidence rates across different regions and the requirement

for predictions incorporating regional spread dynamics. The proposed model, employing a group of regression learners, aimed to mitigate these issues by classifying regions into three groups based on the incidence rates and optimizing predictions for each.

The proposed solution involved developing a predictive model using a Least-Square Boosting Classification algorithm. This model aimed to predict the incidence of COVID-19 within a forthcoming two-week period analyzing patterns in the data from each geographic region and its neighboring areas collected two weeks prior. The model classified regions into three groups based on the incidence rate (less than 200, between 200 and 1000, and above 1000) to tailor predictions to different scales of the outbreak. The results of this model were auspicious with a high degree of accuracy. Strengthening the performance in forecasting was also the mean absolute error (MAE) for the model's predictions which were remarkably low. Specifically, the model had an accuracy rate of 98.45% when comparing its predictions with the actual case numbers between March 30 and April 12, 2020. The MAE ranged from 4.71% to 8.54% across different incidence rate groups, predicting a significant global increase in cases over two weeks. With this approach, the model was able to predict a global incidence of 1,134,018 new cases by April 12, 2020, with the highest increases in Europe, North America, and Asia (See Figure 8).

Date	*Continents*						*Total number of confirmed cases*
	Africa	**Asia**	**Australian**	**Europe**	**North America**	**South America**	
22 Jan ~ 29 Mar	4995	161,986	4522	385,097	150,877	11,740	719,217
30-Mar	635	7720	802	37,853	19,269	1906	68,185
31-Mar	820	7227	722	37,433	16,890	2000	65,092
1-Apr	472	7533	338	38,512	19,625	1508	67,988
2-Apr	1046	6438	981	44,047	18,435	1955	72,902
3-Apr	1047	6790	780	53,087	19,802	2359	83,865
4-Apr	1015	9739	872	51,954	19,302	2258	85,140
5-Apr	1014	10,563	1226	47,352	19,579	2490	82,224
6-Apr	1447	6867	1015	48,562	19,060	2530	79,481
7-Apr	1636	8027	1057	51,192	20,191	2768	84,871
8-Apr	2087	6786	1444	56,826	19,546	2550	89,239
9-Apr	2157	7749	1270	55,316	20,475	2685	89,652
10-Apr	1976	5818	1430	54,377	20,819	2573	86,993
11-Apr	1849	8962	1390	56,284	19,627	2351	90,463
12-Apr	1930	6781	1199	54,870	20,337	2806	87,923
Total	19,131	107,000	14,526	687,665	272,957	32,739	1,134,018
Prevalence growth rate	**283.00**	−33.94	221.23	78.57	80.91	178.87	57.67

Fig. 8. Model's prediction of the COVID-19 new cases for the next 2 weeks

G. K-Means Clustering Identifies Diverse Clinical Phenotypes in COVID-19 Patients: Implications for Mortality Risks and Remdesivir [6]

The effectiveness of remdesivir in reducing mortality among COVID-19 hospitalized patients remains a subject of debate. The focus of this paper is to pinpoint specific clinical profiles of COVID-19 patients who stand to gain the most from remdesivir treatment and to validate those findings using an independent data set.

Exploring the effect of antiviral usage, or lack thereof, on mortality across various phenotypes of hospitalized COVID-19 patients represents a critical challenge. Remdesivir stands as the initial antiviral treatment sanctioned for addressing this infection.

The data comes from consecutive COVID-19 patients hospitalized between February 2020 and February 2021. The initial group, or derivation cohort, consisted of individuals admitted to the Hospital Clinic of Barcelona.The derivation cohort consisted of subjects admitted to the Hospital Clinic of Barcelona, while the validation cohort comprised patients from Hospital Universitari Mutua de Terrassa (Terrassa) and Hospital Universitari La Fe (Valencia), both tertiary centers in Spain. K-means clustering was employed to group patients based on reverse transcription polymerase chain reaction (rRT-PCR) cycle threshold (Ct) values, lymphocyte counts at diagnosis, and pre-test symptom duration. The impact of remdesivir on 60-day mortality within each cluster was assessed.

In the study period, 1160 consecutive adults in the derivation cohort were assessed. Figure 6 shows the main epidemiological and clinical characteristics of these patients. 59.4 % of the cohort were male and the median age was 66 years.The median duration of symptoms before testing was 6 days, with a median lymphocyte count of $0.8 \times 10^{\wedge}9/L$ at onset. Overall 60-day mortality was 14.22% and the percent of people admitted into intensive care was 21 %. It was observed that there was higher mortality in Patients with the shortest duration from symptom onset to testing, as well as those with lower lymphocyte counts and Ct values of rRT-PCR at COVID-19 diagnosis exhibited higher mortality rates.

	Patients (*N* = 1160)
Patient characteristics	
Age–median (IQR), in years	66 (55–78)
Sex male, *n* (%)	688 (59.4)
Comorbidities, *n* (%)	
Hypertension	517 (44.6)
Chronic heart disease	291 (25.1)
Chronic lung disease	276 (23.8)
Diabetes mellitus	221 (19.0)
Solid neoplasm	178 (15.3)
Hematological malignancies	78 (6.7)
Chronic liver diseases	71 (6.1)
Vital signs at admission, median (IQR)	
Temperature (°C)	37.3 (36.6–38)
Respiratory rate (rpm)	20 (18–24)
Oxygen saturation (%)	95 (93–97)
Laboratory values at admission, median (IQR)	
Ferritin (ng/mL)	580 (274–1088)
CRP (mg/dL)	7.9 (3.8–14.2)
D-dimer (ng/mL)	700 (400–1300)
LDH (U/L)	315 (251–400)
Lymphocyte count ($\times 10^9$/L)	0.8 (0.6–1.1)
Median (IQR) cycle threshold (Ct) at COVID-19 diagnosis	26.5 (22.0–30.1)
Ct ≤ 20, *n* (%)	173 (14.9)
Ct between 21 and 25, *n* (%)	297 (25.6)
Ct > 25, *n* (%)	690 (59.5)
Median (IQR) days of pre–test duration of symptoms	6 (3–8)
Intensive care unit admission, *n* (%)	244 (21.0)
60–day mortality, *n* (%)	165 (14.2)

Fig. 9. Main epidemiologic and clinical characteristics of patients in the derivation cohort.

This is a great approach to analyzing COVID-19 data. Everyone has had experience with modern medicine and almost no one reacts the same. This can be said for more than just COVID-19. To break it up into clusters based on statistical categories is a really good idea and makes perfect sense.

H. Artificial intelligence-based decision support model for new drug development planning [7]

Developing a new drug is a tricky business, high-risk, high return. While there is potential to make huge profits, the overall success rate is incredibly low. Companies are always looking for new ways to increase the odds of developing a successful drug. One rapidly growing solution is that of big data and AI.

The focus of this paper is to develop a decision-support model to recommend the most developable drugs for pharmaceutical companies. The model is based on the fact that "users" information in recommender systems can correspond to "pharmaceutical companies, while information about "purchased/rated items" can be likened to "successfully developed drugs." To go along with recommender systems, three other approaches were applied: association rule learning, collaborative filtering, and content-based filtering. The results from all these approaches were then combined to create a hybrid model.

There have been reported problems with content-based filtering approaches such as cold-start, sparsity, and scalability of data. To deal with this many hybrid recommender systems have been created to incorporate different approaches to create high-quality recommendations.

The data on the drugs was pulled from IQVIA™ Pipeline Intelligence. From here COVID-19 drug data was obtained for the periods: April 2020 - November 2020.

The degree of advancement in the clinical trial phase was compared to the prediction scores obtained from the hybrid model. The more advanced the clinical trial phase the higher the score predicted by the hybrid model. Significantly, the two companies that have achieved success in vaccine development thus far, Pfizer and Moderna, held the first (0.92) and fourth (0.79) prediction scores. Figure 10 shows the hybrid model predictions for companies trying to develop a COVID-19 vaccine.

Phase advanced	Prediction score (mean)	Number of companies
0	0.029	41
1	0.159	23
2	0.188	14
3	0.411	8

Fig. 10. Hybrid model prediction for vaccine development

Creating a model to help determine the most developable drug is a game-breaking idea. This allows companies to spend less time and resources on a product that might never make it to the market. A model like this could be game-breaking for pharmaceutical companies.

III. Discussion and Conclusion

The application of NLP, supervised learning, clustering algorithms, and association rules has demonstrated high potential in enhancing our understanding and management of the pandemic. For instance, NLP's role in misinformation tracking and sentiment analysis has been essential in public health communication strategies. Similarly, supervised learning algorithms have offered valuable insights into infection trends, patient outcomes, and identifying high-risk demographics, allowing for more targeted interventions. Clustering algorithms, particularly K-means, have provided a framework for stratifying patient groups, leading to more personalized treatment approaches, as evidenced in the study of diverse clinical phenotypes among COVID-19 patients.

Despite these advancements, the integration of DM techniques in healthcare, especially during a global emergency, is filled with challenges and limitations. Data quality and accessibility issues have been recurrent, with the heterogeneity of data sources complicating analysis efforts. Also, ethical considerations around data privacy and the potential for bias in data-driven decisions have highlighted the need for transparent methodologies and strict guidelines. Additionally, specific techniques such as K-Means clustering have faced criticism for occasionally misinterpreting data, which shows the limitations of certain algorithms in handling complex datasets.

Looking forward, the knowledge gained from this investigation highlights the necessity of ongoing innovation in DM techniques. There is immense opportunity to create more reliable models that can manage the complexity of pandemic-related data.This includes tackling the issues of accessibility and data quality, improving the ethical framework that regulates data use, and investigating how to incorporate emerging technologies such as artificial intelligence and machine learning in predictive modelling.

To conclude, by leveraging these DM techniques, researchers and healthcare professionals have been able to gain valuable insights into the virus's spread, its impacts, and effective response strategies. As the pandemic evolves, so too must our approaches to data analysis and interpretation. The challenges encountered and the lessons learned provide a foundation for future research, emphasizing the need for innovative solutions, ethical considerations and global collaboration in the quest to harness the full potential of data mining in healthcare and beyond.

IV. Biography

Alexandros Ioannou is a fourth-year undergraduate student, currently finishing up a Bachelor of Science in Computer Science and Data Analytics at Wilfrid Laurier University. He is interested and passionate about technology, programming, AI, ML, video game design, web development, cybersecurity, and other fields regarding IT.

Cameron Anderson is a fourth-year undergraduate student pursuing a Bachelor of Science in Computer Science at

Wilfrid Laurier University and will be graduating in June 2024. He is interested in computer graphics, game design and the ever-growing use of AI and machine learning and plans on pursuing a role in one of these fields shortly.

Luke Aikman is a fourth-year undergraduate student pursuing a Bachelor of Science in Computer Science and Math at Wilfrid Laurier University and will be graduating in June 2024. He is interested in using data to drive decisions and hopes to one day be able to put this to use in the world of sports.

Lubna Al Rifaie is a fourth-year undergraduate student pursuing a Bachelor of Science in Honours Computer Science at Wilfrid Laurier University and will be graduating in December 2024. She specializes in Big Data Systems and software development, minoring in mathematics. She is deeply passionate about technology, artificial intelligence, and machine learning, demonstrating a strong interest in leveraging these fields within game design, and IT, and further advancing AI innovations.

References

[1] Ahouz, F., & Golabpour, A. (2021). Predicting the incidence of COVID-19 using data mining. BMC Public Health. https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-021-11058-3

[2] Safdari, R., Rezayi, S., Saeedi, S., Tanhapour, M., & Gholamzadeh, M. (2021). Using data mining techniques to fight and control epidemics: A scoping review. Health and Technology. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8102070/

[3] Allmuttar, A. Y. O., & Alkhafaji, S. K. D. (2023, June). Using data mining techniques deep analysis and theoretical investigation of covid-19 pandemic. Measurement. Sensors. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10017173/#bib30

[4] K. Dotzlaw, R. Dotzlaw, C. K. Leung, A. G. M. Pazdor, S. Szturm and D. Tan, "Mining Big Healthcare Data to Predict Long COVID Cases," 2023 IEEE International Conference on Industrial Technology (ICIT), Orlando, FL, USA, 2023, pp. 1-6, doi: https://ieeexplore.ieee.org/document/10143145

[5] Razzak, M. I., Imran, M., & Xu, G. (2020). Big Data Analytics for Preventive Medicine. Neural computing & applications. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7088441/

[6] Garcia-Vidal, C., Teijón-Lumbreras, C., Aiello, T.F. et al. K-Means Clustering Identifies Diverse Clinical Phenotypes in COVID-19 Patients: Implications for Mortality Risks and Remdesivir Impact. Infect Dis Ther (2024). https://doi.org/10.1007/s40121-024-00938-x

[7] Jung YL, Yoo HS, Hwang J. Artificial intelligence-based decision support model for new drug development planning. Expert Syst Appl. 2022 Jul 15;198:116825. doi: 10.1016/j.eswa.2022.116825. Epub 2022 Mar 8. PMID: 35283560; PMCID: PMC8902892.

[8] Y. Tan, "An Improved KNN Text Classification Algorithm Based on K-Medoids and Rough Set," 2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, China, 2018, pp. 109-113, doi: https://ieeexplore.ieee.org/abstract/document/8530289

[9] Kim, E. H.-J., Jeong, Y. K., Kim, Y., Kang, K. Y., & Song, M. (2016). Topic-based content and sentiment analysis of the Ebola virus on Twitter and in the news. Journal of Information Science, 42(6), 763-781. https://doi.org/10.1177/0165551515608733

[10] *Home*. Humanitarian Data Exchange. (n.d.). https://data.humdata.org/dataset/long-covidresearchagenda

[11] Household pulse survey public use file (PUF). (n.d.-a). https://www.census.gov/programs-surveys/household-pulse-survey/datasets.html

[12] Cortés-Martínez, K. V., Estrada-Esquivel, H., Martínez-Rebollar, A., Hernández-Pérez, Y., & Ortiz-Hernández, J. (2022, May 23). The state of the art of data mining algorithms for predicting the COVID-19 pandemic. MDPI. https://www.mdpi.com/2075-1680/11/5/242

[13] Patel S, Patel H. Survey of Data Mining Techniques used in Healthcare Domain. International Journal of Information Sciences and Techniques. 2016. https://doi.org/10.5121/ijist.2016.6206