Exploring Data Mining Techniques In COVID-19 Research

Group 5: Lubna Al Rifaie, Alexandros Ioannou, Cameron Anderson, Luke Aikman

Introduction

- COVID-19 pandemic posed unpredictable and difficult challenges.
- Data mining techniques were needed to extract knowledge and information from large datasets.
- Necessary for:
  - Tracking the spread of the virus
  - Predicting outcomes
  - Responding to the virus

Paper Overview

- “Using Data Mining Techniques for COVID-19: A Systematic Review”[1]
  - Application of supervised learning techniques. Natural Language Processing being the most used DM technique.
- “Using data mining techniques to fight and control epidemics: A scoping review”[2]
  - Supervised learning techniques: Logistic Regression and Classification
  - Unsupervised learning techniques: Clustering and deep learning approaches

Methods Used

- The primary goal of these papers was to compile a list of papers on the use of DM techniques in pandemics.
- **Natural Language Processing (NLP):**
  - Most used method (22%) in the literature for COVID-19
  - Disease spread, public sentiment, potential treatments
- **Supervised Learning:**
  - Most used across studies (90%)
  - Used to predict disease outcomes and trends
- **Clustering Algorithms:**
  - Grouping patients based on symptoms or outcomes
- **Association Rules and Frequent-Itemset Mining:**
  - Used for finding relationships between symptoms or coexisting conditions.

Methods Used

Author	Main approaches	Clinical scope	The applied method of data mining	Software (Environment)	Data source
Abd-Alrazaq A et al. [30]	Infoveillance	Social behavior	Text mining	Python	Twitter
Ahamad MM [19]	Disease characteristics	Infectious disease	Decision Tree, Random Forest, gradient boosting Machine, SVM	SPSS	Github repository
Ren X et al.	Treatment	Pharmacology	Association rule mining method, and association knowledge network	R	Traditional Chinese medicine system pharmacology database
Zhang Y et al. [31]	Infoveillance	Psychology	Time series, NLP, and deep learning	Python	Weibo social network
Sudirman ID Nugraha DY [59]	Risk factors	Infectious	Naive Bayes method	Rapid Miner	Ministry of Public Health Thailand
Huang C et al. [20]	Disease characteristics	Infectious disease	Text mining	Python	Sina Weibo social network
Han X et al. [32]	Infoveillance	Infectious disease	Time series, Random forest, Spatial Distribution	Python	Sina Weibo social network
Ibrahim et al. [61]	Tracing transmission	Epidemiology	ANN	not mentioned	CDC
Foieni F et al.[22]	Disease characteristics	Respiratory medicine	Multivariate Regression	SPSS	WHO
Zhao ZR et al. [46]	Patient monitoring and follow- up	Respiratory medicine Regression model	SPSS	COVID-19 PUI registry	Respiratory medicine Regression model
Fan Q et al. [60]	Risk factors	Cardiology	Logistic regression	SPSS	Wuhan Tongji hospital
Lei MT et al. [62]	Tracing transmission	Epidemiology	CART, Linear regression	SPSS	Macao Meteorological and Geo- physical Bureau
Dong YL et al. [42]	Patient monitoring and follow-up	up Infectious disease	Logistic regression	SPSS	Wuhan union hospital
Roland LT et al. [26]	Disease characteristics	Respiratory medicine	Logistic regression	SPSS	San Francisco (USF) institutional
Zhou YW et al.[51]	Early diagnosis	Infectious disease	Logistic regression, Nomograms	R	47 locations in Sichuan province
Li S et al. [54]	Early diagnosis	Psychology	Text mining	SPSS	Weibo posts
Ayyoubzadeh SM et al. [34]	Infoveillance	Epidemiology scope Linear regression and long short	term memory (LSTM) models	Python	Google data
Qiang X et al. [50]	Active case prediction	Infectious disease	Random forest (RF) method	R	China national genomics data center
Liu. Q et al. [27]	Disease characteristics	Infectious disease	Logistic regression	SPSS	Union Hospital, Tongji medical
KostkovaP et al. [41]	Outbreak prediction	Public health	Text mining	Not mentioned	Twitter
Kostoff RN [35]	Infoveillance	Informatics	Text mining	Not mentioned	Medical literature
Szomszo M et al. [36]	Infoveillance	Informatics	Text mining, linked resource	Not mentioned	Twitter

Table 3 Frequency of data mining techniques in reviewed studies

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]

Methods Used

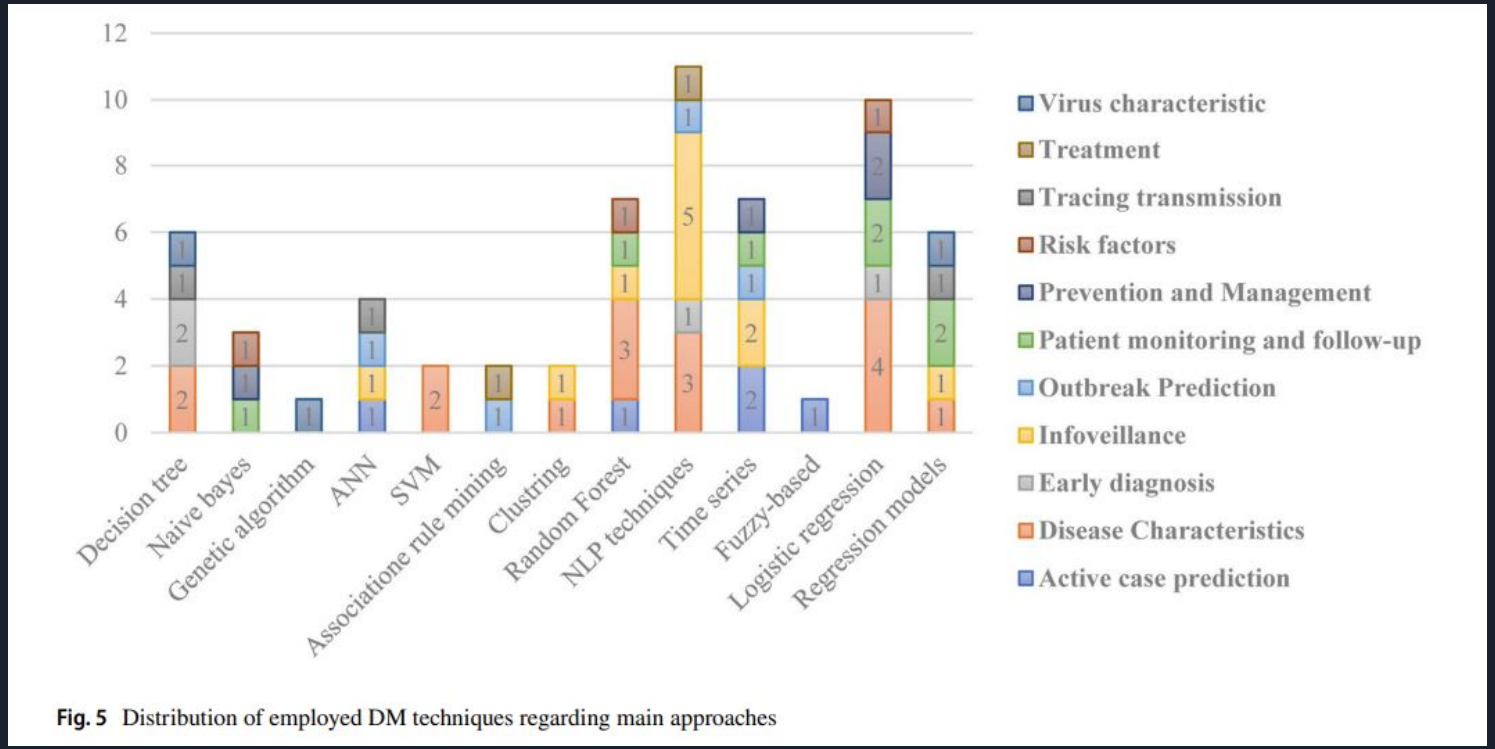


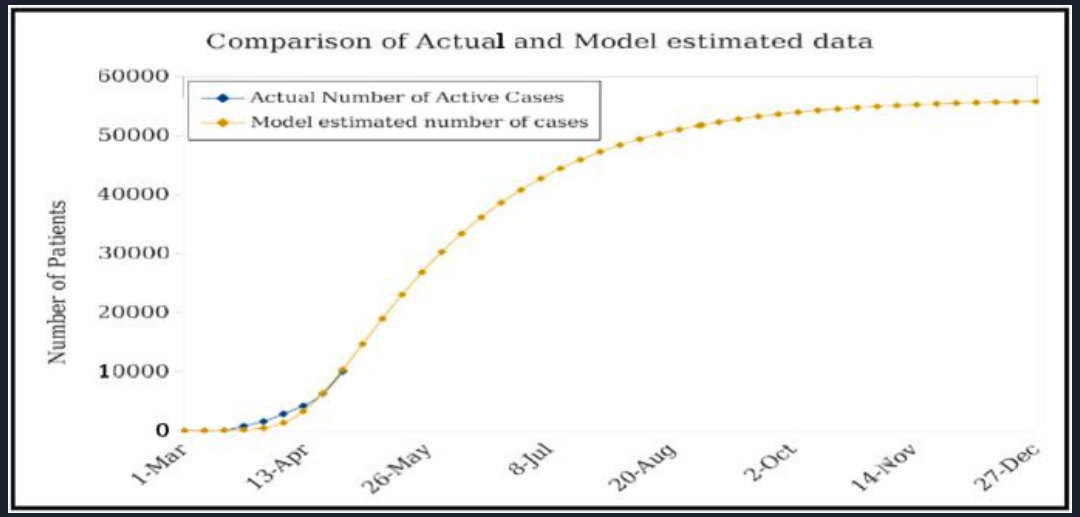
Fig. 5 Distribution of employed DM techniques regarding main approaches

Advantages and Limitations

- **Advantages:**
  - **Diverse Data Handling**
  - **Predictive Power**
  - **Pattern Recognition**
- **Limitations:**
  - **Data Quality**
  - **Data Availability**

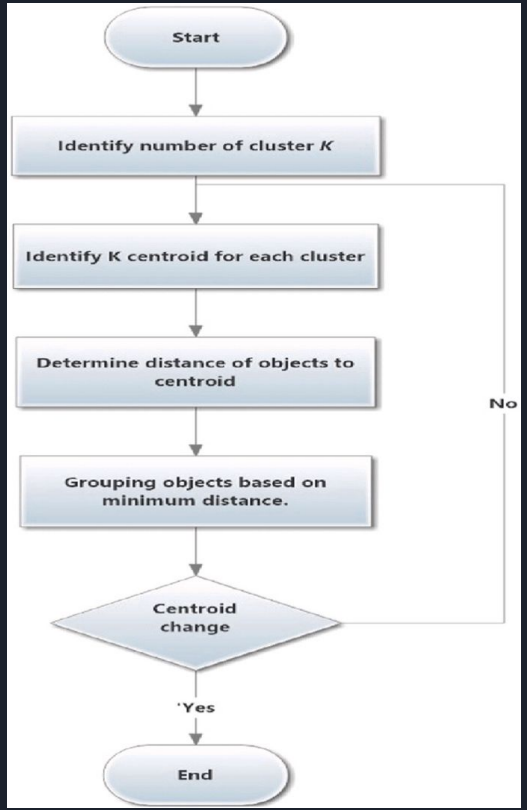
Paper Overview

- “Using data mining techniques deep analysis and theoretical investigation of COVID-19 pandemic”
  - Focused on covid-19 human body data in Iraq
  - Application of K-Means Clustering



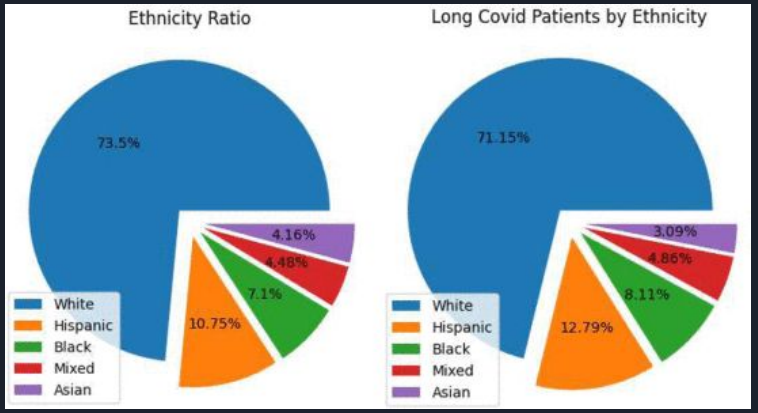
Analysis

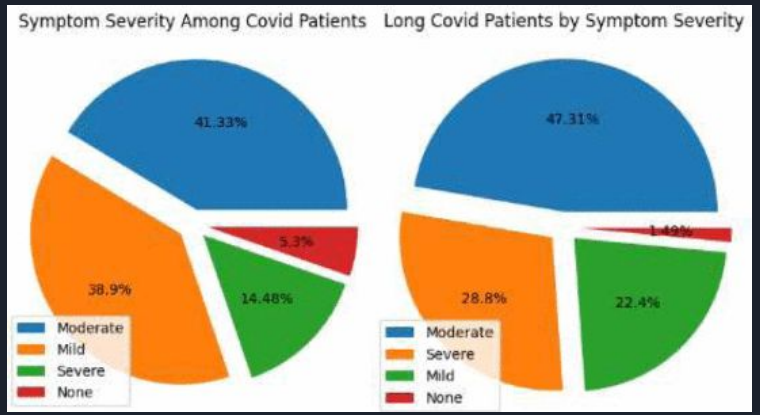
- **Advantages**
  - Lower parameter, less intense model preferred
- **Limitations**
  - Occasionally K-Means will evaluate irrelevant information or ignore essential facts
  - Data accessibility



Paper Overview

- “Mining Big Healthcare Data to Predict Long COVID Cases”
  - Focuses on predictions post COVID
  - Demographic and symptom analysis
  - Frequent Itemsets and Association Rules

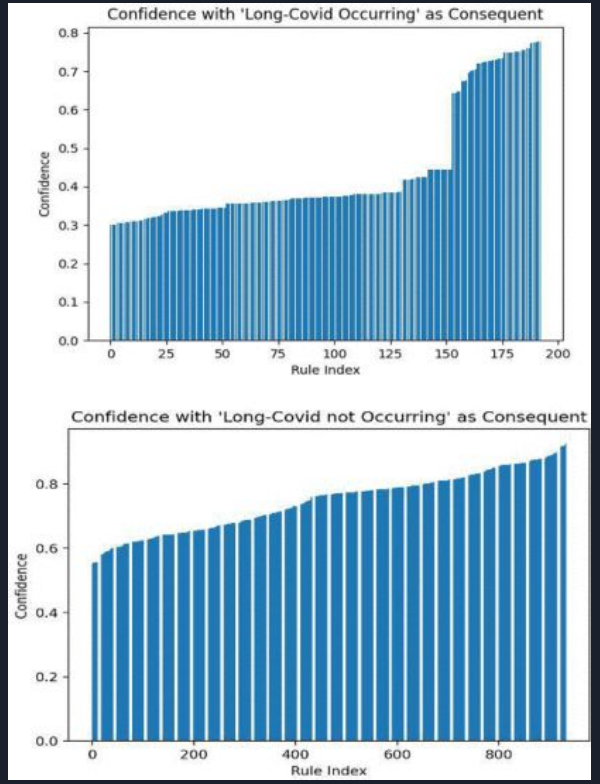




Analysis

- **Advantages**
  - Combination of different classifications of data can provide better results

- **Limitations**
  - Few studies on long COVID and the combination of demographic and symptom data



Paper Overview

- **“K-Means Clustering Identifies Diverse Clinical Phenotypes in COVID-19 Patients”**
  - Identify patients with similar viral stats
  - Mortality associated with use or not of antivirals in hospitalized covid 19 patients
  - Elbow method determined best number of clusters to be 5

Main characteristics of Clusters

K-means cluster	Median Ct (IQR)	Median days of pre-test duration of symptoms (IQR)	Median lymphocyte count (IQR) ($\times 10^9$/L)	60-day mortality (%)	60-day mortality/pts receiving remdesivir (%)	60-day mortality/pts who did not receive remdesivir (%)	*p* value
Cluster 1							
Derivation cohort *n* = 100	26 (23–30)	5 (3–7)	1.7 (1.5–2)	2	0	2.4	0.54
Validation cohort *n* = 167	25 (22–29)	6 (4–7)	1.8 (1.6–2.2)	6.6	0	7.2	0.28
Cluster 2							
Derivation cohort *n* = 273	24 (22–26)	8 (7–9)	0.8 (0.6–1)	11	0	11.3	0.35
Validation cohort *n* = 292	21 (18–25)	8 (7–9)	0.9 (0.7–1.1)	7.2	3.2	7.7	0.37

Paper Overview

- “Artificial Intelligence -based support for model for new drug development planning
  - New drug development success is currently very low
  - The approach taken combines association rules, collaborative filtering and content-based filtering approaches
  - Applied to see the success probability of a company developing a new covid vaccine

Comparison of Degree of Advancement in Clinical Trial Phase with Prediction Score

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

Introduction to Medical Data Analysis

“Big data analytics for preventive medicine”

- Unlocking Insights, Improving Care, and Reducing Costs
- Overview of the complexity and importance of medical data analysis.

Challenges in Healthcare Data Analysis

- Data Volume
- Data Variety
- Data Quality
- Data Privacy and Security
- Interoperability
- Resource Constraints
- Complexity of Healthcare Systems
- Adoption Barriers

Advantages of Using Data Analytics in Disease Prevention:

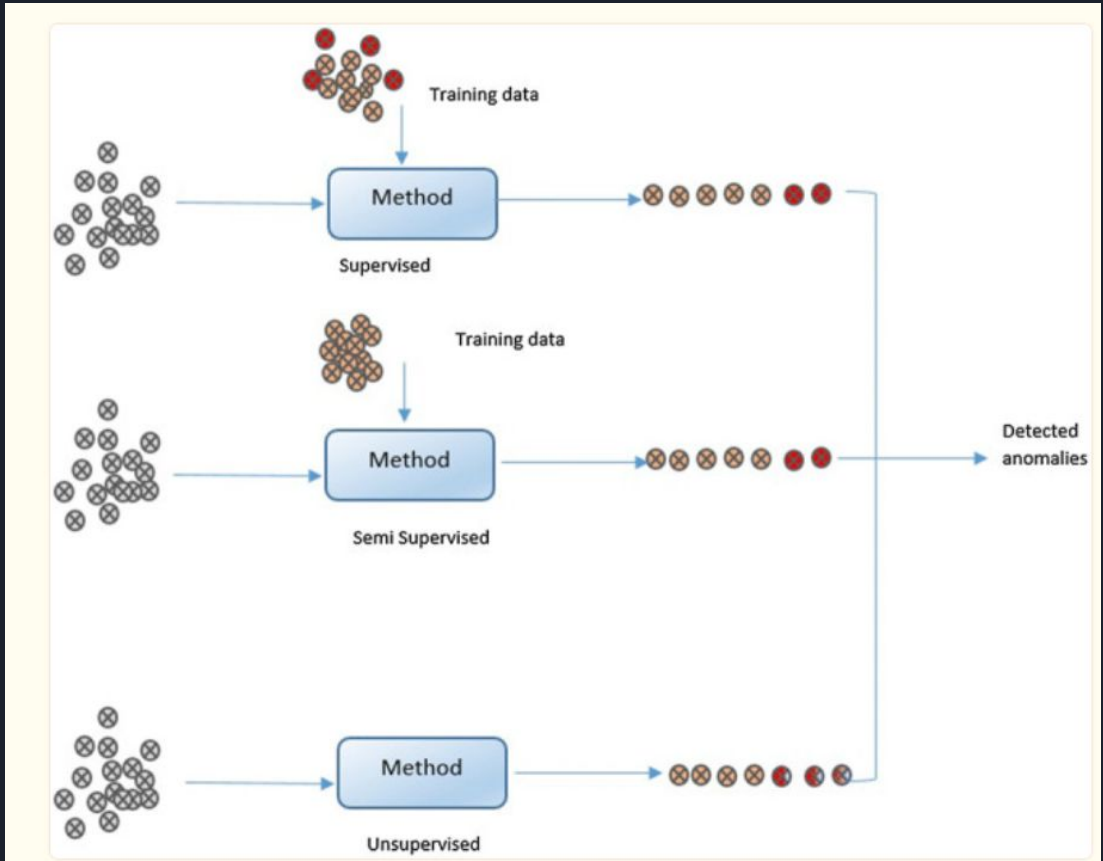
- Early Detection
- Personalized Medicine
- Predictive Analytics
- Evidence-Based Decision Making
- Cost Savings
- Continuous Improvement
- Population Health Management

Role of Data Analytics in Disease Prevention:

- Importance of Disease Prevention
  - Emphasize the significance of disease prevention in healthcare.

- Role of Data Analytics in Disease Prevention
  - Data analytics can contribute to disease prevention by efficiently analyzing large volumes of complex healthcare data.

Anomaly Detection in Data Mining Methods



Thanks for Listening!

Any Questions?

References

1. Ghosh, S., & Das, L. C. (2022). Using Data Mining Techniques for COVID-19: A Systematic Review. International Journal on Data Science and Technology. https://sciencepublishinggroup.com/article/10.11648/j.ijdst.20220802.11
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: 10.1109/ICIT58465.2023.10143145.
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

References

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.