COVID-19 ANALYSIS USING BIGDATA ALGORITHMS

Raja Kishor Duggirala[1*]

Abstract—
In the current era of internet, high volume of data is generated rapidly and made available in data repositories. Big Data supports in understanding, tracking, and managing the spread of the COVID virus, as well as developing strategies to combat its impact. The present paper presents a big data solution for analysing voluminous COVID-19 epidemiological data. The present work also presents the solution that helps users to get a better understanding of information about the confirmed cases of COVID-19. Evaluation of the results show the benefits of this data science solution in discovering useful knowledge from big COVID-19 data.

Keywords—data science, coronavirus disease, COVID-19, big data, big data algorithm, big data application, big data mining and analytics

[1*]Department of CSE, Dr. L. Bullayya College of Engineering, Visakhapatnam, A.P., India

***Corresponding Author:** Raja Kishor Duggirala
*Department of CSE, Dr. L. Bullayya College of Engineering, Visakhapatnam, A.P., India
Email: rajakishor@gmail.com

DOI: - 10.53555/ecb/2022.11.12.194

Introduction:

The COVID-19 pandemic has brought unprecedented challenges, necessitating innovative approaches to disease surveillance, epidemiology, and public health interventions. Big Data Science, which involves the collection, analysis, and interpretation of vast and complex datasets, has played a pivotal role in addressing these challenges.

Big Data Science, characterized by the collection, analysis, and interpretation of vast and complex datasets, has played a critical role in the COVID-19 response. This approach encompasses various data sources, including clinical data, epidemiological data, genomic sequences, mobility data, social media content, and more. It has allowed researchers, healthcare professionals, and policymakers to gain insights into the virus's transmission dynamics, clinical outcomes, and the effectiveness of interventions.

One of the primary applications of Big Data Science in the context of COVID-19 is in epidemiological modelling. Through the integration of diverse data sources, researchers have developed models that predict the disease's spread, identify potential hotspots, and evaluate the impact of public health measures. These models have been invaluable in guiding policy decisions and resource allocation.

Additionally, genomic sequencing data has been a cornerstone of understanding the virus's genetic variability, which is crucial for the development of vaccines and treatments. The analysis of viral genomes using Big Data techniques has enabled the rapid development of effective vaccines and the monitoring of new variants of concern.

Big Data Science has also proven vital in real-time surveillance and monitoring. Mobile phone data, social media sentiment analysis, and other digital data sources have been leveraged to track population movements, public compliance with health measures, and public sentiment, aiding in the rapid response to changing circumstances.

Despite its remarkable contributions, Big Data Science in the context of COVID-19 faces several challenges, including data privacy concerns, data quality issues, and the need for harmonization and standardization of data across regions. Ethical considerations surrounding the use of personal data also pose significant challenges.

Looking forward, the potential for Big Data Science in the fight against COVID-19 is vast. As more data sources become available and as data analysis techniques continue to advance, the ability to predict outbreaks, monitor vaccine distribution, and identify new therapeutic options will be greatly enhanced. However, addressing ethical and privacy concerns and ensuring equitable access to data-driven solutions will be essential as we continue to navigate the pandemic.

In the current era of big data [1-6], high volume of big data can be generated and collected from a wide variety of rich data sources at a rapid rate. Due to differences in level of veracity, some of these big data are precise while some others are imprecise and uncertain. Embedded in these big data are useful information and valuable knowledge that can be discovered by big data science and engineering (BigDataSE) [7-9], which applies techniques from various related areas—such as data mining [10-15], machine learning [16-19], as well as mathematical and statistical modelling [20]—to real-life applications and services and/or for social good. Examples of rich sources of these valuable big data include: images of people or products (e.g., human face, agricultural products) [21, 22]; entertainment or games (e.g., movies, chess) [23, 24]; networks (e.g., co-authorship networks [25], communication networks [6], sensor networks [26], social networks) [27-31]; stock markets [32, 33]; traffic conditions [34-38]; music [39, 40]; as well as healthcare, bio-medical, and/or bio-engineering applications (e.g., disease reports [41, 42], genomic data like genomic data [43, 44], epidemiological data and statistics [45-47]).

Scope of the Present Work:

Big Data Science has been a crucial component of the global response to COVID-19, providing the tools and insights necessary to understand and combat the virus effectively. Its applications in epidemiology, genomics, and real-time monitoring have proven invaluable. As the pandemic continues to evolve, Big Data Science will remain a key asset in this effort to mitigate its impact.

Methodology:

Knowledge discovered from these big data would be valuable. For instance, knowledge discovered from the epidemiological data—such as data related to cases who suffered from viral diseases like (a) severe acute respiratory syndrome (SARS) that broke out in 2002–2004, (b) Middle East respiratory syndrome (MERS) that broke out in 2012–2015, and (c) coronavirus disease 2019 (COVID-19) that broke out in 2019 and became pandemic in 2020—helps researchers, epidemiologists and policy makers to get a better understanding of the disease. This, in turn, may inspire them to come up ways to detect, control and combat the disease.

Partially because of the COVID-19 pandemic, many researchers have explored different aspects of the COVID-19. These include clinical and treatment information [48, 49], as well as drug discovery [50, 51], related on research medical and health sciences. In contrast, we—as computer scientists with expertise in data science and engineering—focus on a data science and engineering aspect of epidemiological data. Epidemiological data are excellent examples in illustrating the common 7V's for characterizing big data: Due to the high number of cumulative COVID-19 cases (e.g., more than 53 million cumulative COVID-19 cases globally [52] as of November 15, 2020), the volume of epidemiological data is huge. With the high number of new COVID-19 cases, new data are generated at a high velocity (e.g., about 594 thousand daily new cases globally [52] on November 15, 2020— which sadly implies more than 400 new COVID-19 cases per minute, or close to 7 new cases per second, globally). These new data are usually reported on a daily basis. These data are usually collected from a wide variety of data sources (e.g., regional health authorities within a province, from which data are integrated and reported at higher levels such as a national level). For instance, in the Canadian province of Manitoba, COVID-19 data can be gathered from Winnipeg Regional Health Authority (WRHA) and four other health authorities1. Moreover, a wide variety of data (e.g., gender, age, symptoms, clinical course and outcomes, transmission methods) are collected too.

Partially due to the fast dissemination of the information and partially due to the privacy-preservation of the individual cases, some details (e.g., transmission methods) of the cases are unstated or unknown. This leads to data of different veracity—some data are precise while some are imprecise and uncertain [53, 54]. To elaborate, it is not unusual to have known values for some of the attributes (e.g., known hospitalization status like "hospitalized and admitted to the intensive care unit (ICU)") but unknown/NULL values for some others (e.g., unstated transmission methods of disease). Moreover, some data are quite detailed (e.g., "on January 23, a 56-year old male presented to Sunnybrook Health Sciences Centre in Toronto with a new onset of fever and non-productive cough following return from Wuhan, China, the day prior" [55]), whereas some other data are more abstract and general (e.g., "on Week 3—i.e., the third full week—of 2020, a male in his 50s—who was transmitted through international travel—in the province of Ontario showed symptoms of fever and cough") to preserving the privacy [56-59] of individual cases.

Partially because of the COVID-19 pandemic, many researchers have explored on different aspects of the COVID-19 disease. These led to numerous works on COVID-19 in different disciplines or areas: For medical and health sciences, there have been (a) systematic reviews on literature about medical research on COVID-19 [61, 62], (b) clinical and treatment information [48, 49], as well as (c) drug discovery and vaccine development [50, 51]. For social sciences, there have been studies on crisis management for the COVID-19 outbreak [63]. For natural sciences and engineering (NSE), there have been works focusing on (a) artificial intelligence (AI)- driven informatics, sensing, imaging for tracking, testing, diagnosis, treatment and prognosis [64]—such as those imaging-based diagnosis of COVID-19 using chest computed tomography (CT) images [65, 66]—and (b) mathematical modelling of the spread of COVID-19 [67].

The current paper is also for NSE by taking on a computational favor. However, our designed and developed data science solution examines textual-based COVID-19 epidemiological data (rather than images). Instead of projecting the spread of the disease, this data science solution discovers common characteristics among COVID-19 cases belonging to a certain gender, age group combination, and compares them with those belonging to other combinations. The discovered knowledge helps users to get a better understanding of information about the confirmed cases of COVID-19. Although this solution is designed for big data science of COVID-19 data, it would be applicable to data science of other big data in many real-life applications and services.

1) Data Collection, Integration and Preprocessing

To evaluate and demonstrate the usefulness of this data science solution, we tested it with different COVID-19
epidemiological data including the Canada cases from Statistics Canada2 [68, 69]. With this dataset, data have been collected and integrated from provincial and territorial public health authorities by the Public Health Agency of Canada (PHAC).

We preprocess data and generalize some attributes to obtain a dataset with the following attributes:
1. A unique privacy-preserving identifier for each case
2. A generalized region/location
3. Episode week (or onset week of symptoms): From Week 3 (i.e., week of January 19-25, 2020) to now
4. Gender (cf. sex at birth, which consists of male and female), including (a) male, (b) female, (c) others including unstated gender and non-binary gender (e.g.,
lesbian, gay, bisexual, transgender, queer/ questioning, two-spirited (LGBTQ2+)).
5. Age group: ≤ 19, 20s, 30s, 40s, 50s, 60s, 70s, and ≥ 80s.
6. Occupation group, including:
a) healthcare worker,
b) school or daycare worker (or attendee),
c) long-term care resident, and
d) other occupation.
7. Asymptomatic: Yes and No
8. Set of 13 symptoms, including cough, fever, chills, sore throat, runny nose, shortness of breath, nausea, headache, weakness, pain, irritability, diarrhea, and other symptoms.
9. Hospital status, including:
a) hospitalized in the ICU,
b) hospitalized but not in the ICU, and
c) not hospitalized.
10. Transmission method, including:
a) community exposures, and
b) travel exposures.
11. Clinical outcome: Recovered and death
12. Recovery week As of November 12, 2020, the dataset has captured
209,811 COVID-19 cases in Canada. Among them, 190,108 cases with stated episode week. Moreover, although the first Canadian case occurred in Week 3, there were not more than two new daily cases for following few weeks. To preserve 17 privacy of these early cases and to cumulate statistically significant mass for analysis, cases from Weeks 3-8 were grouped into (Episode) Week 8 (February 23-29) with 107 cases. From Week 9 onward, the data reflect their reported episode weeks.

Results and Discussions:
Among 12 aforementioned attributes, we examine 16 combinations of gender, age group. In addition, also compared the COVID-19 case distribution with the
corresponding distribution of the estimated Canadian population (for July 2020) [70]. See Fig. 1 for the population distribution.

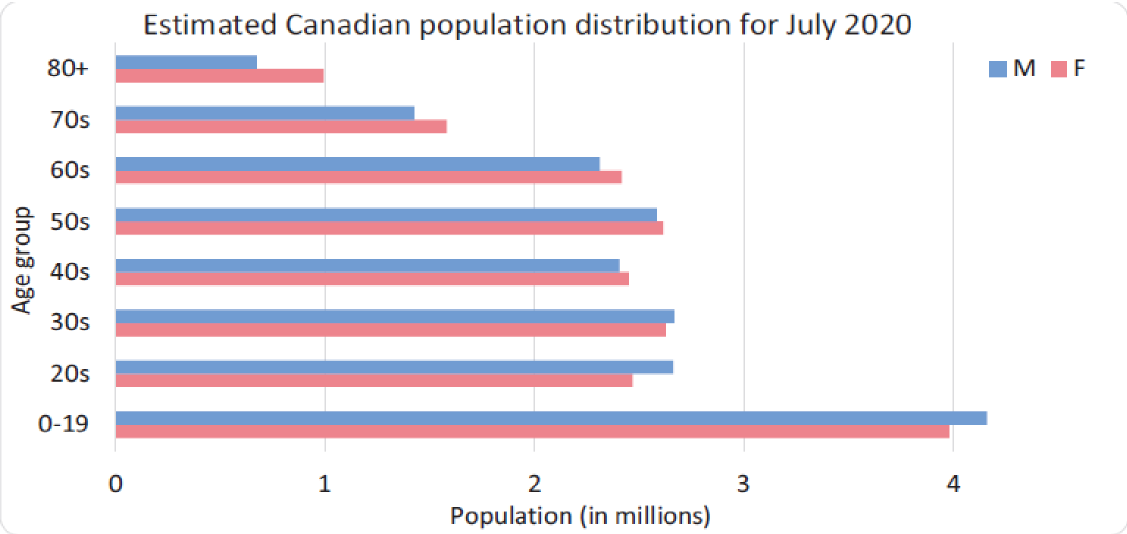

Fig. 1. Distribution of estimated population for July 2020.

2) Big Data Science on Cases:
Once the data are pre-processed, this data science solution first analyses and mines the national data. With 201,341 COVID-19 cases with stated gender and age (out of an estimated Canadian population of 38,005,238), the solution reveals that about 0.53% of the population contacted the disease.

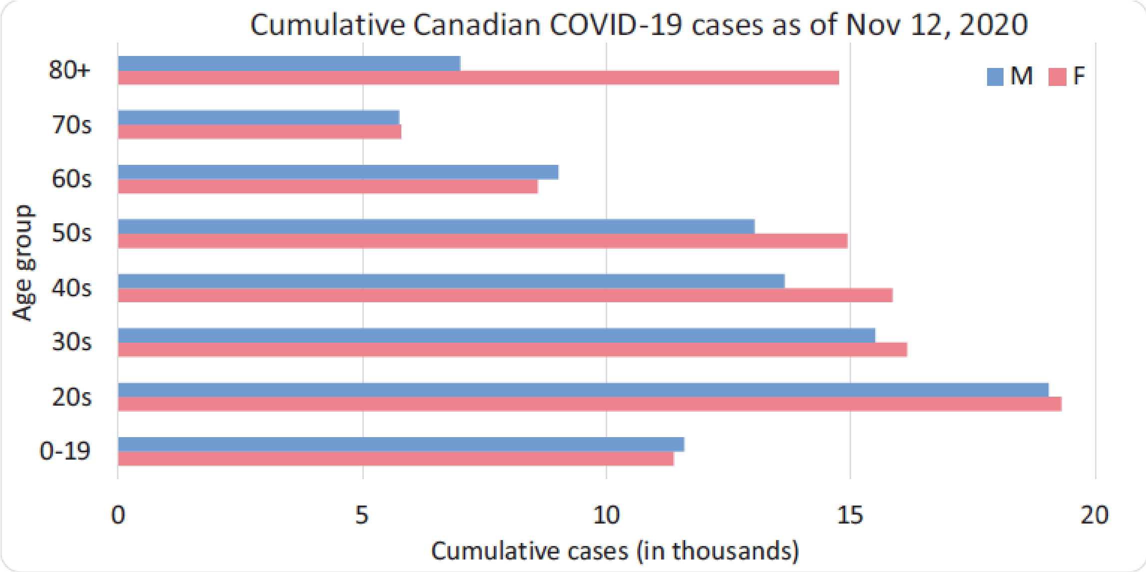

Fig. 2. Distribution of cumulative COVID-19 cases as of Nov 12, 2020.

TABLE I. DISTRIBUTION OF CUMULATIVE COVID-19 CASES (AND PERCENTAGES WITH RESPECT TO POPULATION OF THE CORREPONDING GENDER, AGE GROUP COMBINATION AS OF NOVEMBER 12, 2020

	Male		Female		Age Group
	#cases	*wrt corr. pop'n*	*#cases*	*wrt corr. pop'n*	*wrt corr. pop'n*
0-19	11,594	0.28%	11,374	0.29%	**0.28%**
20s	19,049	0.72%	19,316	0.78%	**0.75%**
30s	15,497	0.58%	16,151	0.62%	**0.60%**
40s	13,651	0.57%	15,851	0.65%	**0.61%**
50s	13,040	0.50%	14,935	0.57%	**0.54%**
60s	9,007	0.39%	8,584	0.36%	**0.37%**
70s	5,743	0.40%	5,790	0.37%	**0.38%**
80+	7,004	1.04%	14,755	1.49%	**1.31%**
Total	**94,585**	**0.50%**	**106,756**	**0.56%**	**0.53%**

Then, this data science solution analyses and mines all 16 gender, age group combinations. The resulting distribution of COVID-19 cases is shown in Fig. 2 and Table I. The bar chart reveals that (a) despite being the most populated age groups, youth of 0-19 does not have the highest number of cases. Instead, (b) youth of 20s have the highest number. In contrast, (c) seniors in their 70s have the lowest number of cases. Moreover, (d) female in their 80s have more cases than their male counterparts.

Table I confirms the above observations. Moreover, it also reveals that (a) age groups 20s-40s and 80+ (as well as female in their 50s) appear to be more vulnerable to the disease as they have higher COVID-19 percentages than the national norm. Here, the percentage is computed by dividing the number of cases in a specific group (i.e., a specific gender, age group combination) by the population of the corresponding combination. For instance, 19,049 cases of male in their 20s correspond to 0.28% of this population group of about 2.6 million male in their 20s. (b) Among all age groups, seniors in 80+ have the highest risk—with a COVID-19 percentage of 1.31% of their corresponding population (cf. the national norm of 0.53% of the national population).

The table also reveals that (c) female appears to be slightly more vulnerable to the disease than their male counterparts. (d) In all age groups from 0-59, percentages of female COVID- 19 cases are slightly higher than their male counterparts. (e) For age groups 60s-70s, the opposite is observed. (f) Among all age groups, female in their 80+ have the highest risk—with a COVID-19 percentage of 1.49% of their corresponding population (cf. 1.04% of male in 80+).

3) Big Data Science on Hospital Status

In addition to examining the cumulative cases, this solution also examines the hospital status among the 16 combinations. Table II reveals that, (a) as the age increases, the absolute number of hospitalized cases also increases. When combined with Table I, we observe that (b) despite the number of cases decreases from age groups 20s to 70s, the number of hospitalization increases. This means that, when young people catches COVID-19, a majority of them do not need to be hospitalized. When people age, their chance of requiring hospitalization once they catch COVID-19 increases. (c) Between the two genders, more male in their 30+ are admitted into the ICU than female.

Cells in Table III shows the percentage of hospitalized cases with respect to COVID-19 patients in their corresponding gender, age group combination. For instance, 38 male COVID-19 patients in their 20s admitted to the ICU (as shown in Table II) account for 0.77% (as shown in Table III) of all 19,049 male COVID-19 patients in their 20s. Table III reveals that, (a) for seniors 60+, the hospitalization percentages among all COVID-19 cases are high—ranging from 14.01% to 25.57% (cf. national norm of 7.05%)—and peak at 70s. In particular, (b) males in their 70s have highest percentages of both ICU admission (8.51% wrt COVID-19 cases for males in 70s) and hospitalization (8.51%+20.02% = 28.53%). In contrast, (c) males in their 80s have the highest percentage of non-ICU hospitalization (23.70%).

Population (In Millions **TABLE II.**

CUMULATIVE NUMBER OF HOSPITALIZATION AS OF NOVEMBER 12, 2020

	Male		Female		Age Group
	ICU admitted	*Non-ICU hospitalized*	*ICU admitted*	*Non-ICU hospitalized*	*Total hospitalized*
0-19	11	79	11	95	**196**
20s	38	147	54	193	**432**
30s	74	288	62	292	**716**
40s	159	438	99	372	**1,068**
50s	421	777	211	557	**1,966**
60s	548	957	269	690	**2,464**
70s	489	1,150	268	1,042	**2,949**
80+	204	1,660	194	2,346	**4,404**
Total	**1,944**	**5,496**	**1,168**	**5,587**	**14,195**

TABLE III. PERCENTAGE OF HOSPITALIZATION WITH RESPECT TO COVID-19 CASES OF THE CORREPONDING GENDER, AGE GROUP- COMBINATION AS OF NOVEMBER 12, 2020

	Male		Female		Age Group
	ICU admitted	*Non-ICU hospitalized*	*ICU admitted*	*Non-ICU hospitalized*	*Total hospitalized*
0-19	0.09%	0.68%	0.10%	0.84%	**0.85%**
20s	0.20%	0.77%	0.28%	1.00%	**1.13%**
30s	0.48%	1.86%	0.38%	1.81%	**2.26%**
40s	1.16%	3.21%	0.62%	2.35%	**3.62%**
50s	3.23%	5.96%	1.41%	3.73%	**7.03%**
60s	6.08%	10.63%	3.13%	8.04%	**14.01%**
70s	8.51%	20.02%	4.63%	18.00%	**25.57%**
80+	2.91%	23.70%	1.31%	15.90%	**20.24%**
Total	**2.06%**	**5.81%**	**1.09%**	**5.23%**	**7.05%**

4) Big Data Science on Occupation Groups

This solution also examines different occupation groups. Table IV shows the number of healthcare workers for some gender, age group-combinations (and their percentages w.r.t. COVID-19 cases in the corresponding combination). It reveals that (a) female healthcare workers in their 30s-50s account for more than a quarter of COVID-19 cases in their respective combinations. For instance, 5,308 (33.49%) of 15,851 COVID- 19 cases for females in their 40s are healthcare workers. (b) In terms of both absolute number (in terms of cases) and relative number (w.r.t. cases in their combinations), female healthcare workers have much higher numbers (about 4x higher) than their male counterparts. For completeness, Table IV also includes the total numbers for all age groups (including 0-19 and 70+) in the bottom row.

TABLE IV. NUMBER OF HEALTHCARE WORKERS (AND THEIR PERCENTAGE WITH RESPECT TO COVID-19 CASES OF THE CORREPONDING GENDER, AGE GROUP COMBINATION) AS OF NOVEMBER 12, 2020

	Male		Female		Age Group
	healthcare workers	*wrt cases*	*healthcare workers*	*wrt cases*	*wrt cases*
20s	893	4.69%	3,751	19.42%	12.10%
30s	1,206	7.78%	4,497	27.84%	18.02%
40s	1,300	9.52%	5,308	33.49%	22.40%
50s	1,178	9.03%	4,605	30.83%	20.67%
60s	389	4.32%	1,493	17.39%	10.70%
All ages	**5,075**	**5.37%**	**19,937**	**18.68%**	**12.42%**

5) Frequent and Contrast Pattern Mining

In addition to conducting big data analytics on attributes, This solution also mine frequent and contrast patterns for each combination. For instance, we observe the following from males in their 20s: (a) Frequent singleton pattern {community exposures}:14524 reveals that 14,524 males in their 20s exposed to COVID-19 from the community (i.e., domestic acquisition), which account for 76.2% of all 19,049 male COVID-19 cases in their 20s (including known and unstated transmission methods). (b) Similarly, pattern {not hospitalized}:12175 reveals that 12,175 (i.e., 63.9%) males COVID-19 cases in their 20s do not need hospitalization, which account for 63.9% of all 19,049 male COVID-19 cases in their 20s (including known and unstated hospital status).

As this solution provides users with flexibility of ignoring NULL values (e.g., unstated transmission methods), (a) the aforementioned 14,524 males in their 20s who exposed to COVID-19 from the community account for 97.6% of all 14,876 male COVID-19 cases in their 20s with known transmission methods. Similarly, (b) the aforementioned 12,175 males in their 20s who do not need hospitalization account for 98.5% of 12,360 male COVID-19 cases in their 20s with known hospital status.

Frequent non-singleton pattern {community exposures, not hospitalized}: 11435 reveals that, among 19,049 males COVID- 19 cases in their 20s, 11,435 (60.0%) exposed via the community but do not need hospitalization. These account for 96.5% of 11,853 male COVID-19 cases in their 20s with known transmission methods and hospital status.

As users have flexibility to express their interest or preference (say, finding frequent pattern consisting of only symptoms), This solution then incorporates user preference into mining frequent patterns satisfying the user preference. For instance, it finds the following patterns from males in their 20s: (a) Frequent pattern {cough}:1528 reveals that 1,528 male COVID-19 cases in their 20s show cough as a symptom.

(b) Similarly, frequent patterns {headache}:1409, {sore throat}: 1142, {chills}:964 and {fever}:910 show the numbers of male COVID-19 cases in their 20s show cough these symptoms.

(c) Frequent non-singleton {cough, headache}:771 reveals that 771 male COVID-19 cases in their 20s show both cough and headache. Similarly, {cough, sore throat}:643 reveals that 643 male COVID-19 cases in their 20s show both cough and sore throat. (d) However, {cough, headache, sore throat}:353 reveals that, while cough commonly occurred with headache or sore throat, but not frequently occurred with both headache and sore throat, among male COVID-19 cases in their 20s.

This data science solution applies a similar procedure to other gender, age group-combinations for discovery of frequent patterns from each combination and comparison among patterns discovered from these combinations. From the comparisons and contrasts, it is observed that following: Between the two genders, (a) more males tend to have fever (e.g., 7.85% of male cases in 20s vs. 6.68% of female counterparts), but (b) more females tend to have soar throat (e.g., 12.38% of female cases in 20s vs. 9.85% of male counterparts) and runny nose (e.g., 9.03% of female cases in 20s vs. 7.25% of male counterparts). Moreover, among different age groups, a commonality is that (a) cough is the most common symptom. In terms of differences, (b) while cases in most age groups experienced headache, seniors in 80+ have lower percentages of this symptom (e.g.,
0.59% cases in 80+ vs. 13.29% of cases in 20s). (c) Similar 19 comments apply to chills (e.g., 0.70% cases in 80+ vs. 8.22% of cases in 20s).

B. Functionality Check with Related Works

After demonstrating the features and usefulness of This data science solution in analyzing real-life COVID-19 data, let us evaluate its functionality when compared with related works. First, most of the related works are observed to report mostly the numbers of cases and deaths. They do not provide privacypreserving details and epidemiological characteristics of those COVID-19 cases, which are provided by this solution. Second, This solution also provides details for each gender, age group combination, which are unavailable in the related works.

Conclusion:

In this paper, it is presented a data science solution for conducting data science on big COVID-19 epidemological data. The solution generalizes some attributes (e.g., age into age groups) for effective analysis. Instead of ignoring unstated/ NULL values of some attributes, the solution provides users with flexibility of including or excluding these values. It also provides users with flexibility to express their preference (e.g., "must include symptoms") in mining of frequent patterns. It discovers frequent patterns from each of the 16 gender, age group combinations. Moreover, it compares and contrasts the discovered frequent patterns among these combinations. Taking into account differences in population and/or the number of cases in each of the 16 combinations, This solution computes relative frequency (with respect to population and/or the number of cases in the respective combination) in addition to showing the absolute frequency of the attributes and/or frequent patterns. Evaluation results show the practicality of This solution in providing rich knowledge about characteristics of COVID-19 cases. This helps researchers, epidemiologists and policy makers to get a better understanding of the disease, which may inspire them to come up ways to detect, control and combat the disease. As ongoing and future work, transfer knowledge learned from the current work to data science on other big data in many real-life applications and services.

REFERENCES

1. A. Alsaig, et al., "A critical analysis of the V-model of big data," in IEEE TrustCom/BigDataSE 2018, pp. 1809-1813.
2. A. Kobusinska, et al., "Emerging trends, issues and challenges in Internet of Things, big data and cloud computing," FGCS 87, 2018, pp. 416-419.
3. K. Kritikos, "Towards dynamic and optimal big data placement," in IEEE TrustCom/BigDataSE 2018, pp. 1730-1737.
4. C.K. Leung, "Big data analysis and mining," in Encyclopaedia of Information Science and Technology, 4e, 2018, pp. 338-348.
5. C.K. Leung, "Big data computing and mining in a smart world," in Big Data Analyses, Services, and Smart Data, 2021, pp. 15-27.
6. B. Yin, et al., "A cooperative edge computing scheme for reducing the cost of transferring big data in 5G networks," in IEEE TrustCom/BigDataSE 2019, pp. 700-706.
7. K.E. Dierckens, et al., "A data science and engineering solution for fast k-means clustering of big data," in IEEE TrustCom/BigDataSE/ICESS 2017, pp. 925-932.
8. C.K. Leung, "Data science for big data applications and services: data lake management, data analytics and visualization," in Big Data Analyses, Services, and Smart Data, 2021, pp. 28-44.
9. C.K. Leung, F. Jiang, "A data science solution for mining interesting patterns from uncertain big data," in IEEE BDCloud 2014, pp. 235-242.
10. A.K. Chanda, et al., "A new framework for mining weighted periodic patterns in time series databases," ESWA 79, 2017, pp. 207-224.
11. A. Fariha, et al., "Mining frequent patterns from human interactions in meetings using directed acyclic graphs," in PAKDD 2013, Part I, pp. 38- 49.

12. C.K. Leung, C.L. Carmichael, “FpVAT: A visual analytic tool for supporting frequent pattern mining,” ACM SIGKDD Explorations 11(2), 2009, pp. 39-48.
13. C.K. Leung, et al., “Fast algorithms for frequent itemset mining from uncertain data,” in IEEE ICDM 2014, pp. 893-898.
14. C.K. Leung, et al., “Parallel social network mining for interesting 'following' patterns,” CCPE 28(15), 2016, pp. 3994-4012.
15. R. Mo, et al., “A differential privacy-based protecting data preprocessing method for big data mining,” in IEEE Trust Com/Big DataSE 2019, pp. 693-699.
16. S. Ahn, et al., “A fuzzy logic based machine learning tool for supporting big data business analytics in complex artificial intelligence environments,” in FUZZ-IEEE 2019, pp. 1259-1264.
17. A. Bari, G. Saatcioglu, “Emotion artificial intelligence derived from ensemble learning,” in IEEE Trust Com/Big Data SE 2018, pp. 1763-1770.
18. J. Gonzalez-Lopez, et al., “Large-scale multi-label ensemble learning on Spark,” in IEEE TrustCom/BigDataSE/ICESS 2017, pp. 893-900.
19. C.K. Leung, et al., “An innovative fuzzy logic-based machine learning algorithm for supporting predictive analytics on big transportation data,” in FUZZ-IEEE 2020. doi: 10.1109/FUZZ48607.2020.9177823
20. C.K. Leung, “Mathematical model for propagation of influence in a social network,” in Encyclopedia of Social Network Analysis and Mining, 2e, 2018, pp. 1261-1269.
21. L. Ale, et al., “Lightweight deep learning model for facial expression recognition,” in IEEE TrustCom/ BigDataSE 2019, pp. 707-712.
22. B. Min, et al., “Image classification for agricultural products using transfer learning,” in BigDAS 2020.
23. [23] J.A. Brown, et al., “A machine learning system for supporting advanced knowledge discovery from chess game data,” in IEEE ICMLA 2017, pp. 649-654.
24. D. Choudhery, C.K. Leung, “Social media mining: prediction of box office revenue,” in IDEAS 2017, pp. 20-29.
25. C.K. Leung, C.L. Carmichael, “FpViz: A visualizer for frequent pattern mining,” in KDD-VAKD 2009, pp. 30-39.
26. H. Ren, et al., “Prediction algorithm based on weather forecast for energy harvesting wireless sensor networks,” in IEEE TrustCom/BigDataSE 2018, pp. 1785-1790.
27. F. Jiang, et al., “Finding popular friends in social networks,” in CGC 2012, pp. 501-508.
28. C.K. Leung, C.L. Carmichael, “Exploring social networks: a frequent pattern visualization approach,” in IEEE SocialCom 2010, pp. 419-424.
29. C.K. Leung, et al., “Big data analytics of social network data: who cares most about you on Facebook?” in Highlighting the Importance of Big Data Management and Analysis for Various, 2018, pp. 1-15.
30. C.K. Leung, F. Jiang, “Big data analytics of social networks for the discovery of "following" patterns,” in DaWaK 2015, pp. 123-135.
31. Y. Yang, M.O. Shafiq, “Identifying high value users in twitter based on text mining approaches,” in IEEE TrustCom/BigDataSE 2019, pp. 634-641.
32. C.K. Leung, et al., “A machine learning approach for stock price prediction,” in IDEAS 2014, pp. 274-277.
33. K.J. Morris, et al, “Token-based adaptive time-series prediction by ensembling linear and non-linear estimators: a machine learning approach 20 for predictive analytics on big stock data,” in IEEE ICMLA 2018,pp. 1486-1491.
34. A.A. Audu, et al., “An intelligent predictive analytics system for transportation analytics on open data towards the development of a smart city,” in CISIS 2019, pp. 224-236.
35. P.P.F. Balbin, et al., “Predictive analytics on open big data for supporting smart transportation services,” Procedia Computer Science 176, 2020, pp. 3009-3018.
36. Y. Huang, et al., “Diffusion convolutional recurrent neural network with rank influence learning for traffic forecasting,” in IEEE TrustCom/BigDataSE 2019, pp. 678-685.
37. C.K. Leung, et al., “Effective classification of ground transportation modes for urban data mining in smart cities,” in DaWaK 2018, pp. 83-97.
38. C.K. Leung, et al., “Urban analytics of big transportation data for supporting smart cities,” in DaWaK 2019, pp. 24-33.
39. K.E. Barkwell, et al., “Big data visualisation and visual analytics for music data mining,” in IV 2018, pp. 235-240.
40. C. Fan, et al., “Social network mining for recommendation of friends based on music interests,” in IEEE/ACM ASONAM 2018, pp. 833-840.
41. C.K. Leung, et al., “Data science for healthcare predictive analytics,” in IDEAS 2020, pp. 8:1-8:10.

42. J. Souza, et al., "An innovative big data predictive analytics framework over hybrid big data sources with an application for disease analytics," in AINA 2020, pp. 669-680.
43. J. De Guia, et al., "DeepGx: deep learning using gene expression for cancer classification," in IEEE/ACM ASONAM 2019, pp. 913-920.
44. C.K. Leung, et al., "Predictive analytics on genomic data with highperformance computing," in IEEE BIBM 2020, pp. 2187-2194.
45. Y. Chen, et al., "Temporal data analytics on COVID-19 data with ubiquitous computing," in IEEE ISPA-BD Cloud-SocialCom-SustainCom 2020, pp. 958-965. doi:. 10.1109/ISPA-BDCloud-Social Com-Sustain Com 51426.2020.00146
46. P. Gupta, et al., "Vertical data mining from relational data and its application to COVID-19 data," in Big Data Analyses, Services, and Smart Data, 2021, pp. 106-116.
47. Q. Liu, et al., "A two-dimensional sparse matrix profile DenseNet for COVID-19 diagnosis using chest CT images," IEEE Access 8, 2020, pp. 213718-213728.
48. A.A. Ardakani, et al., "Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: results of 10 convolutional neural networks," Comp. Bio. Med. 121, 2020, pp. 103795:1-103795:9.
49. D. Barh, et al.,, "Multi-omics-based identification of SARS-CoV-2 infection biology and candidate drugs against COVID-19," Comput. Biol. Medicine 126, 2020, pp. 104051:1-104051:13.
50. M.B. Jamshidi, et al., "Artificial intelligence and COVID-19: deep learning approaches for diagnosis and treatment," IEEE Access 8, 2020, pp. 109581-109595.
51. B. Robson, "COVID-19 coronavirus spike protein analysis for synthetic vaccines, a peptidomimetic antagonist, and therapeutic drugs, and analysis of a proposed achilles' heel conserved region to minimize probability of escape mutations and drug resistance," Comp. Bio. Med. 121, 2020, pp. 103749:1-103749:28.
52. World Health Organization, WHO coronavirus disease (COVID-19) dashboard. https://covid19.who.int/
53. F. Jiang, C.K. Leung, "A data analytic algorithm for managing, querying, and processing uncertain big data in cloud environments," Algorithms 8(4), 2015, pp. 1175-1194.
54. C.K. Leung, "Uncertain frequent pattern mining," in Frequent Pattern Mining, 2014, pp. 417-453.
55. X. Marchand-Senécal, et al., "Diagnosis and management of first case of COVID-19 in Canada: lessons applied from SARS-CoV-1," Clinical Infectious Diseases, 2020. doi:10.1093/cid/ciaa227
56. C.S. Eom, et al., "Effective privacy preserving data publishing by vectorization," Information Sciences 527, 2020, pp. 311-328.
57. C. Luo, et al., "Efficient privacy-preserving outsourcing of large-scale QR factorization," in IEEE Trust Com/Big Data SE/ICESS 2017, pp. 917-924.
58. A.M. Olawoyin, et al., "Privacy-preserving spatio-temporal patient data publishing," in DEXA 2020, Part II, pp. 407-416.
59. B.H. Wodi, et al., "Fast privacy-preserving keyword search on encrypted outsourced data," in IEEE BigData 2019, pp. 6266-6275. doi: 10.1109/ BigData47090.2019.9046058
60. C.K. Leung, et al., "Visual analytics of social networks: mining and visualizing co-authorship networks," in HCII-FAC 2011, pp. 335-345.
61. W.T. Li, et al., "Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis," BMC Medical Informatics Decis. Mak. 20(1), 2020, pp. 247:1-247:13.
62. A.S. Albahri, et al., "Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID- 19): a systematic review," J. Medical Syst. 44(7), 2020, pp. 122:1-122:11.
63. W. Kuo, J. He, "Guest editorial: crisis management - from nuclear accidents to outbreaks of COVID-19 and infectious diseases," IEEE Trans. Reliab. 69(3), 2020, pp. 846-850.
64. A.A. Amini, et al., "Editorial special issue on "AI-driven informatics, sensing, imaging and big data analytics for fighting the COVID-19 pandemic". IEEE JBHI 24(10), 2020, pp. 2731-2732.
65.] D. Shen, et al., "Guest editorial: special issue on imaging-based diagnosis of COVID-19," IEEE TMI 39(8), 2020, pp. 2569-2571.
66. Y. Zhang, et al., "A five-layer deep convolutional neural network with stochastic pooling for chest CT-based COVID-19 diagnosis," Mach. Vis. Appl. 32(1), 2021, pp. 14:1-14:13.

67. A. Viguerie, et al., "Simulating the spread of COVID-19 via a spatiallyresolved susceptible-exposed-infected-recovered-deceased (SEIRD) model with heterogeneous diffusion," Appl. Math. Lett. 111, 2021, pp. 106617:1-106617:9.
68. Public Health Agency of Canada, "Detailed preliminary information on confirmed cases of COVID-19 (revised)," Statistics Canada Table 13-10- 0781-01. doi:10.25318/1310078101-eng
69. Public Health Agency of Canada, "Preliminary dataset on confirmed cases of COVID-19," Statistics Canada Table 13-26-0003. https://www150. statcan.gc.ca/n1/ en/ catalogue/ 13260003
70. Statistics Canada, "Population estimates on July 1st, by age and sex,"Table 17-10-0005-01. doi: 10.25318/1710000501-eng