2020 IEEE International Conference on Big Data (Big Data)

# Machine Learning and OLAP on Big COVID-19 Data

Carson K. Leung*, Yubo Chen, Calvin S.H. Hoi, Siyuan Shang
*Department of Computer Science, University of Manitoba*
Winnipeg, MB, Canada
*Email: kleung@cs.umanitoba.ca

Alfredo Cuzzocrea
*Big Data Engineering and Analytics Lab, University of Calabria*
Rende, Italy

2020 IEEE International Conference on Big Data (Big Data) | 978-1-7281-6251-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/BigData50022.2020.9378407

***Abstract*—In the current technological era, huge amounts of big data are generated and collected from a wide variety of rich data sources. These big data can be of different levels of veracity in the sense that some of them are precise while some others are imprecise and uncertain. Embedded in these big data are useful information and valuable knowledge to be discovered. An example of these big data is healthcare and epidemiological data such as data related to patients who suffered from epidemic diseases like the coronavirus disease 2019 (COVID-19). Knowledge discovered from these epidemiological data—via data science techniques such as machine learning, data mining, and online analytical processing (OLAP)—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. In this paper, we present a machine learning and big data analytic tool for processing and analyzing COVID-19 epidemiological data. Specifically, the tool makes good use of taxonomy and OLAP to generalize some specific attributes into some generalized attributes for effective big data analytics. Instead of ignoring unknown or unstated values of some attributes, the tool provides users with flexibility of including or excluding these values, depending on their preference and applications. Moreover, the tool discovers frequent patterns and their related patterns, which help reveal some useful knowledge such as absolute and relative frequency of the patterns. Furthermore, the tool learns from the patterns discovered from historical data and predicts useful information such as clinical outcomes for future data. As such, the tool helps users to get a better understanding of information about the confirmed cases of COVID-19. Although this tool is designed for machine learning and analytics of big epidemiological data, it would be applicable to machine learning and analytics of big data in many other real-life applications and services.**

***Keywords—big data, machine learning, online analytical processing, OLAP, data science, data analytics, data mining, coronavirus disease, COVID-19, epidemiological data***

## I. Introduction

In the current technological era, big data are everywhere. To elaborate, huge amounts of data have been easily generated and collected from a wide variety of rich data sources at a rapid rate. These big data can be of different levels of veracity (e.g., precise data, imprecise and uncertain data [1-3]). Examples of big data include:

- network (e.g., social network) data [4-10],
- financial time series [11-13],
- transportation data [14-17],
- omic data (e.g., genomic data) [18, 19],
- disease reports [20-22], as well as
- epidemiological data and statistics.

Useful information and valuable knowledge is usually embedded in these big data. This calls for data science [23], which aims to discover knowledge from these big data via data mining algorithms [24-26], machine learning tools [27-29], online analytical processing (OLAP) techniques [30-32], mathematical and statistical models [33, 34], data analytics, and visual analytics. The discovered knowledge is useful. For instance, knowledge discovered from these epidemiological data helps researchers, epidemiologists and policy makers to get a better understanding of the disease, which may inspire them to come up ways to detect, prevent, and/or control diseases such as viral diseases. Examples of viral diseases include:

- severe acute respiratory syndrome (SARS), with outbreak in 2002–2004;
- Middle East respiratory syndrome (MERS), with outbreak in 2012–2015; and
- coronavirus disease 2019 (COVID-19), with outbreak started in 2019 and became pandemic in 2020

Due to the COVID-19 pandemic, many researchers have focused on different aspects of the COVID-19 disease. These include clinical and treatment information [35, 36], as well as drug discovery [18, 37], related on research medical and health sciences. In contrast, as computer scientists, we focus on other aspects of COVID-19 data—namely, epidemiological data.

Many existing works on the COVID-19 epidemiological data focused on showing the numbers of confirmed cases and mortality spatially and/or temporally. In other words, they show:

- spatial differences among different continents, countries, regions, or sovereignties; and/or
- temporal differences among weeks or days along the timeline—e.g., to show the effects of public health strategies and mitigation techniques such as social/

physical distancing, stay-at-home orders, and lockdowns in “flattening the (epidemic) curve”.

As the numbers of inhabitants and tests both play roles in the data and their analyses, they help in the computation of figures like (a) the numbers of confirmed cases and mortality per thousand/million inhabitants and (b) the number of tests per thousand inhabitants.

While the numbers of confirmed cases and mortality are important in showing the severity of the disease in a certain location at a specific time or time interval, there are other important knowledge that can be discovered from the epidemiological data for revealing additional information associated with the disease. For instance, knowing that more confirmed cases and mortality reported today when compared with yesterday indicates the severity of the COVID-19 situations in Canada. However, these numbers do not reveal information such as:

- Which age groups tend to be more vulnerable to the disease (i.e., who is most at risk for COVID-19)?
- Which age groups tend to be less vulnerable to the disease?
- What is likelihood of recovery for COVID-19 survivors who were admitted to the intensive care units (ICU)?

In this paper, we present a machine learning and big data analytic tool to discover this additional information associated with the disease from the epidemiological data. The tool collects a wide variety of data—such as (a) administrative information, (b) case details, (c) symptoms, (d) clinical course and outcomes, (e) exposures, etc.—from a different data sources. With the increasing number of cases in Canada (and around the world), these data are big and updated frequently. Due to the nature of the data, it is not unusual to have different levels of veracity—i.e., with known values for some of the attributes (e.g., known hospitalization status like “hospitalized and ICU admitted”) and 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” [38]). 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”), for preserving the privacy [39-42] of the individuals.

It becomes logical to have taxonomy to perform OLAP such as generalizing some very specific details into their generalized or aggregated forms to give an overview (i.e., a big picture) of data for data analytics. With the taxonomy, one can drill down if detailed information is needed. As a side-benefit, aggregated counts for attributes reduce the dimensions of the data and the search space for machine learning on big data. Aggregated counts for many of these attributes are expected to be sufficiently frequent to be qualified as frequent patterns. The discovered frequent patterns can then be used in training a supervised learning model for associative classification to predict the clinical course and outcomes for new data.

Our *key contributions* of this paper include our design and development of a machine learning tool for big COVID-19 epidemiological data. Our tool incorporates:

- OLAP techniques, with taxonomy for summarizing specific details of COVID-19 cases by their more generalized forms for purposes like preserving privacy of COVID-19 cases and preparing for data analytics of big data;
- handling of NULL values, which allows users to include or exclude NULL values in the analysis;
- data mining algorithms for the discovery of frequent patterns; and
- machine learning procedures for conducting supervised learning such that the resulting associative classifier—which was trained on historical data—can predict the clinical course and outcomes for new data.

Our tool helps users (e.g., researchers, epidemiologists and policy makers) to get a better understanding of information about the confirmed cases of COVID-19. This, in turns, may inspire them to come up ways to detect, control and combat the disease. Moreover, despite that this tool is designed for machine learning and analytics of big epidemiological data, it is applicable to machine learning and analytics of big data in many other real-life applications and services.

The remainder of this paper is organized as follows. Next section discusses some background and related work. Section III presents our machine learning tool. Section IV shows evaluation results, and Section V draws the conclusions.

## II. Background and Related Works

### A. COVID-19 Research

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. Examples include:

- systematic reviews on literature about medical and health science research on COVID-19 [43, 44]
- clinical and treatment information [35, 36], as well as drug discovery and vaccine development [18, 37], which focus more on the medical and health science aspects
- crisis management for the COVID-19 outbreak [45], which focuses more on the social science aspects
- artificial intelligence (AI)-driven informatics, sensing, imaging for tracking, testing, diagnosis, treatment and prognosis [46] such as those imaging-based diagnosis of COVID-19 using chest computed tomography (CT) images [47, 48]
- mathematical modelling of the spread of COVID-19 [49]

In contrast, the current paper focuses more on natural sciences and engineering aspects—especially, takes on a more computational favor. Moreover, our designed and developed machine learning tool examines textual-based COVID-19 epidemiological data (rather than images). Instead of projecting the spread of the disease, our tool predicts the clinical outcomes—e.g., whether the case recovered or died from the disease. Furthermore, our tool conducts machine learning on big data, and it helps users to get a better understanding of information about the confirmed cases of COVID-19. Although this tool is designed for machine learning and analytics of big epidemiological data, it would be applicable to machine learning and analytics of big data in many other real-life applications and services.

## *B. Confirmed Cases and Mortality*

Many existing works on the COVID-19 epidemiological data focused on reporting the numbers of confirmed cases and mortality spatially, which highlight spatial differences among different continents, countries, regions, or sovereignties. Examples of these works include data and dashboards reported by organizations like:

- World Health Organization (WHO) [50];
- Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)[1];
- European Center for Disease Prevention and Control (ECDC)[2];
- governments (e.g., Government of Canada[3]); as well as
- major news channels/media/networks (e.g., newspaper, TV[4]) and Wikipedia[5].

See Tables I, II, III and IV for some examples showing top-10 countries with new (or cumulative) cases (or deaths) based on the WHO data [50]. Specifically:

- Table I lists the top-10 countries with the highest daily number of new COVID-19 cases, as well as the global daily number of new cases, on November 15, 2020.
- Table II lists the top-10 countries with the highest daily number of new COVID-19 deaths, as well as the global daily number of new cases, on November 15, 2020.
- Tables III lists the top-10 countries with the highest cumulative number of COVID-19 cases, as well as the global cumulative number of cases, as of November 15, 2020.
- Tables IV lists the top-10 countries with the highest cumulative number of COVID-19 deaths, as well as the global cumulative number of deaths, as of November 15, 2020.

Observed from these tables, several countries—such as Brazil, France, India, UK, and USA—have been hit hard by COVID-19 as they appear on all four tables.

TABLE I. COUNTRIES WITH THE TOP-10 NUMBER OF NEW COVID-19 CASES ON NOVEMBER 15, 2020

| **Rank** | **Country** | **New Cases** |
|---|---|---|
| | **Global** | **594,000** |
| 1 | USA | 181,066 |
| 2 | India | 41,100 |
| 3 | Italy | 37,249 |
| 4 | France | 32,059 |
| 5 | Brazil | 29,070 |
| 6 | UK | 26,860 |
| 7 | Poland | 25,571 |
| 8 | Russia | 22,571 |
| 9 | Germany | 16,947 |
| 10 | Argentina | 11,859 |

TABLE II. COUNTRIES WITH THE TOP-10 NUMBER OF NEW DEATHS FROM COVID-19 ON NOVEMBER 15, 2020

| **Rank** | **Country** | **New Deaths** |
|---|---|---|
| | **Global** | **8,212** |
| 1 | USA | 1,356 |
| 2 | Mexico | 568 |
| 3 | Poland | 546 |
| 4 | Italy | 544 |
| 5 | UK | 462 |
| 6 | Brazil | 456 |
| 7 | Iran | 452 |
| 8 | India | 447 |
| 9 | France | 354 |
| 10 | Russia | 352 |

TABLE III. COUNTRIES WITH THE TOP-10 CUMULATIVE TOTAL NUMBER OF COVID-19 CASES AS OF NOVEMBER 15, 2020

| **Rank** | **Country** | **Cumulative Cases** |
|---|---|---|
| | **Global** | **53,766,728** |
| 1 | USA | 10,641,431 |
| 2 | India | 8,814,579 |
| 3 | Brazil | 5,810,652 |
| 4 | Russia | 1,925,825 |
| 5 | France | 1,918,345 |
| 6 | Spain | 1,458,591 |
| 7 | UK | 1,344,360 |
| 8 | Argentina | 1,296,378 |
| 9 | Colombia | 1,182,697 |
| 10 | Italy | 1,144,552 |

TABLE IV. COUNTRIES WITH THE TOP-10 CUMULATIVE TOTAL NUMBER OF DEATHS FROM COVID-19 AS OF NOVEMBER 15, 2020

| **Rank** | **Country** | **Cumulative Deaths** |
|---|---|---|
| | **Global** | **1,308,975** |
| 1 | USA | 242,542 |
| 2 | Brazil | 164,737 |
| 3 | India | 129,635 |
| 4 | Mexico | 97,624 |
| 5 | UK | 51,766 |
| 6 | Italy | 44,683 |
| 7 | France | 43,913 |
| 8 | Iran | 41,034 |
| 9 | Spain | 40,769 |
| 10 | Peru | 35,106 |

[1] https://coronavirus.jhu.edu/map.html

[2] https://qap.ecdc.europa.eu/public/extensions/COVID-19/COVID-19.html

[3] https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection.html

[4] https://newsinteractives.cbc.ca/coronavirustracker/

[5] https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Canada, https://en.wikipedia.org/wiki/Template:COVID-19_pandemic_data/Canada_medical_cases

Given the numbers of inhabitants and tests both play roles in the data and analyses, the numbers of confirmed cases and mortality are sometimes represented in terms of per million inhabitants. See Tables V and VI, which reveal COVID-19 situations in terms of infection *rate* and death *rate* (cf. Tables I to IV show the absolute numbers of infection and death) based on the WHO data [15]. Observed from these tables, some geographically small or sparsely populated countries have been hit hard by COVID-19. For example, in Andorra with population around 77,000, the infection rate is at a worrisome level of around a case per 13.5 inhabitants, and the death rate is around a death per 1,030 inhabitants (cf. around a case per 171.8 inhabitants and a death per 1,350 inhabitants in the USA).

TABLE V. Countries with the Top-k Cumulative Total Number of COVID-19 Cases Per Million Inhabitants as of Nov 15, 2020

| Rank | Country (or Region) | Cum Cases Per 1M Pop'n |
|---|---|---|
| 1 | Andorra | 74,095.6 |
| 2 | Bahrain | 49,673.4 |
| 3 | Qatar | 47,055.7 |
| 4 | Belgium | 45,832.7 |
| 5 | Aruba | 43,450.2 |
| 6 | Montenegro | 42,810.8 |
| 7 | Czechia | 42,789.2 |
| 8 | French Polynesia | 41,672.0 |
| 9 | Luxembourg | 41,424.8 |
| 10 | Armenia | 39,597.5 |
|  | **Global** | **6,887.6** |

TABLE VI. Countries with the Top-k Cumulative Total Number of Deaths Per Million Inhabitants from COVID-19 as of November 15, 2020

| Rank | Country (or Region) | Cum Deaths Per 1M Popn |
|---|---|---|
| 1 | San Marino | 1,237.6 |
| 2 | Belgium | 1,234.1 |
| 3 | Peru | 1,064.7 |
| 4 | Andorra | 970.7 |
| 5 | Spain | 872.0 |
| 6 | Argentina | 775.4 |
| 7 | Brazil | 775.0 |
| 8 | Chile | 773.0 |
| 9 | UK | 762.5 |
| 10 | Mexico | 757.2 |
|  | **Global** | **167.7** |

As "a picture is worth a thousand words", the numbers of cases and mortality are sometimes represented in graphical forms by using *bubble maps*. In a bubble map, the number of cases (or deaths) for each country is indicated by the radius of the bubble representing the country. The larger the bubble representing a country, the more severity is its COVID-19 situation. On the one hand, the users can then easily spot those countries with severe COVID-19 situations due to their large bubble sizes. On the other hand, bubbles may overlap. As such, the overlapping and/or containment of bubbles can make it difficult to users to visualize the severity of the disease in dense regions such as Eastern Caribbean and Southeastern Europe.

Alternatively, the numbers of cases and mortality are sometimes represented by *choropleth maps*. These maps use different shading, coloring, or placing of symbols within predefined areas to indicate the number of cases (or deaths) for each country. The darker the shading of a country, the more severity is its COVID-19 situation. On the one hand, the users can then easily spot those countries with severe COVID-19 situations due to their shading. On the other hand, small countries in terms of geographic areas or sizes (e.g., Andorra, Monaco, San Marino) may not be easily visible on the map, let alone visualizing their shading.

These numbers of confirmed cases and mortality are important in showing the severity of the disease in a certain location at a specific time or time interval. However, it is equally important to explore and discover other useful knowledge from the epidemiological data because the discovered knowledge can reveal useful information (e.g., some characteristics of COVID-19 cases) associated with the disease. This, in turn, helps users to get a better understanding on characteristics of the confirmed cases of COVID-19 (rather than just the numbers of cases).

## III. Our Machine Learning Tool

In this section, we describe our machine learning tool for big data analytics of COVID-19 epidemiological data.

### A. *Collection and Integration of Data*

Big COVID-19 epidemiological data can be of a wide variety (e.g., different types of data). They are usually generated and collected from various data sources.

As a concrete example, in Canada, health care is a responsibility of provincial governments. So, Canadian COVID-19 epidemiological data are gathered from each province (or territory), and provincial data are obtained from *health regions* (which are also known as *health authorities*) within the province. For instance, in the province of Manitoba, COVID-19 data can be gathered from Winnipeg Regional Health Authority (WRHA) and four other health authorities[6]. Similarly, data for the province of British Columbia (BC) can be gathered from five health authorities such as Vancouver Coastal Health (VCH), which obtained data from 14 local health areas (LHA) within the three health service delivery areas (HSDA) in the VCH. In BC, there are 88 HSDA within the 16 LHA among the five health authorities[7]. As a third example, data from the province of Ontario can be gathered from public health units within the provincial 14 local health integration networks (LHIN)[8].

In terms of data types, COVID-19 epidemiological data usually include:

- administrative information, which includes:
  - an unique privacy-preserving identifier for each case,
  - its location, and
  - episode day (i.e., symptom onset day or its closest day).

[6] https://www.gov.mb.ca/health/rha/

[7] https://www2.gov.bc.ca/gov/content/data/geographic-data-services/land-use/administrative-boundaries/health-boundaries

[8] http://www.lhins.on.ca/

- case details, which include:
    - gender,
    - age, and
    - specific occupation of the cases.
- symptom-related data, which include additional information for the case who is not asymptomatic (i.e., symptomatic case) such as:
    - onset day of symptoms, and
    - a collection of symptoms (e.g., cough, fever, chills, sore throat, runny nose, shortness of breath, nausea, headache, weakness, pain, irritability, diarrhea, and other symptoms).
- clinical course and outcomes, which include:
    - hospital status (e.g., hospitalized in the intensive care unit (ICU), non-ICU hospitalized, not hospitalized).
    - For recovered case, it also includes additional information such as the recovery day.
    - For the case who has not recovered, it indicates that the case died while infected by COVID-19.
- exposures, which include transmission methods.

### B. *Handling of NULL values*

After collecting and integrating data from heterogeneous sources, we observe that there are some missing, unstated or unknown information (i.e., NULL values). Given the nature of these COVID-19 cases, it is not unusual to have NULL values because values may not be available or recorded. For some other attributes related to case details (e.g., personal information like gender, age), patients may prefer not to report it due the privacy concerns.

To elaborate, since data are collected from administrative health regions, their locations (or generalized regions within a country) are known. For other attributes, their values can be NULL to indicate that they are unknown or not stated. Although NULL values are usually ignored in many other real-life applications, our tool captures and counts NULL values instead. The rational is that, due to nature of the COVID-19 cases (e.g., for timely reporting of cases, privacy-preservation of the identity of cases), it is not too surprising to observe a significant number of NULL values. Ignoring these many NULL values may lead to inaccurate or incomplete analysis of the data. Hence, for each of these nullable attribute, in additional to those stated values, our tool captures and counts NULL values.

### C. *Preprocessing of Data with Taxonomy and OLAP*

In addition to observing NULL values in the data, we also observe that values for some attributes are too specific (e.g., reported symptom onset day, when may be inaccurate, partially due delays in testing). As another example, due to numerous values for some attributes (e.g., age, occupation), it would be logical to group similar values into a mega-value (say, ages can be binned into age groups). Hence, our tool generalizes some attributes by exploiting taxonomy and OLAP. In other words, data can be stored in a data cube so that users can (a) drill down to find more details and (b) drill/roll up to get aggregate values (e.g., usually count or sum of values). To elaborate, our tool generalizes data by:

- applying taxonomy to case locations to group them into local health regions, which then generalize to become provinces, and then to regions within a country (e.g., Prairies, Atlantic region);
- applying temporal hierarchy to group days into week (e.g., episode week, onset week of symptoms, recovery week);
- grouping ages to age groups (e.g., ≤ 19 years old, 20-29 years old, ..., 70-79 years old, ≥ 80 years old);
- generalizing specific occupation of the cases to some generalized key occupation groups—say, (a) health care workers, (b) school or daycare workers, (c) long-term care residents, and (d) others;
- generalizing specific transmission methods to some generalized key transmission methods—say, (a) community exposures, (b) travel exposures, and (c) others; as well as
- transforming any set of $m$ symptoms (with potential set size of 1 to $m$) into $m$ Boolean attributes, each indicates whether a symptom is reported or not.

Note that generalization of some data helps preserve privacy of some COVID-19 cases. Another side-benefit is to increase the frequency of some attributes in preparation of frequent pattern mining.

### D. *Mining of Frequent Patterns*

After preprocessing and generalizing data, our tool conduct big data analytics on the resulting COVID-19 data. With at least 11 attributes and $m$ symptoms (e.g., $m$ = 13 symptoms listed above), there can be a total of $(11+m)$ dimensions in a data cube when data are stored in the cube. For each dimension, there can be $n_i$ stated values for the attribute/dimension. Then, with the NULL value and ALL value, there can be $(n_i + 2)$ values for the attribute. The total number of cells in the cube can be the product of the number of values (i.e., $n_i + 2$) in each dimension over at least $(11+m)$ dimensions. Thus, the search space can be large.

Our tool first provides users with insights about each dimension $D$. It can do by setting all other dimensions to ALL and enumerating all values of $D$. It repeats the same procedure for each dimension. The dimension with highest frequency is the *most frequent singleton pattern*. On the other end of the spectrum, the dimension with lowest frequency is the *rarest singleton pattern*.

To a further extent, by setting all but $k$ dimensions to ALL and enumerating all values for the $k$ dimensions, cells with high frequency give frequent non-singleton patterns. Conversely, cells with low frequency give rare non-singleton patterns.

Given the large search space, finding frequent or rare patterns with the aforementioned procedure can be time consuming. Our tool also provides an alternative by applying traditional frequent pattern mining algorithms to find frequent and rare patterns. Benefits of using these algorithms include the constant pruning of search space provided by the exploitation of the property that a super-pattern is infrequent if any of its sub-pattern is infrequent.

In addition to finding frequent patterns, our tool also provides users the flexibility to find patterns related to (or complementary to) the mined frequent patterns. This gives insights about relative importance of the mined frequent patterns. Specifically, our tool provides the relative percentages of the frequent patterns when compared with their related patterns (with NULL values included or excluded). The tool first finds frequent patterns by applying traditional frequent pattern mining algorithms and then looks up frequency of related patterns by enumerating values for the attributes in the frequent patterns. The related patterns can be looked up from the data cube or the mined patterns.

### *E. Prediction of Outcomes by Supervised Learning*

Once the frequent patterns are mined, they can be used for associative classification, which is a supervised learning technique. By training our tool with different combinations of attribute-values, it can make predictions. A useful prediction is to predict the likelihood of clinical outcomes (e.g., recovered or deceased).

## IV. Evaluation

### *A. A Case Study on Real-Life COVID-19 Data*

#### *1) Collection and Integration of Data*

To evaluate and demonstrate the usefulness of our machine learning tool, we tested it with different COVID-19 epidemiological data including the Canada cases from Statistics Canada [51]. With this dataset, data have been collected and integrated from provincial and territorial public health authorities by the Public Health Agency of Canada (PHAC). Our tool preprocesses the data and incorporates NULL values in 9 of the 11 attributes:

1. A mandatory attribute for the unique privacy-preserving identifier for each case
2. Another mandatory attribute for the generalized region covering
    a) British Columbia & Yukon,
    b) Prairies (i.e., Alberta, Saskatchewan, Manitoba) & the Northwest Territories,
    c) Ontario & Nunavut,
    d) Quebec, as well as
    e) Atlantic (i.e., New Brunswick, Nova Scotia, Prince Edward Island, Newfoundland and Labrador)
3. Episode week (or onset week of symptoms): From Week 3 (i.e.., week of January 12-18, 2020) to now, and NULL
4. Gender, including NULL
5. Age group: ≤ 19, 20s, 30s, 40s, 50s, 60s, 70s, ≥ 80s, and NULL (e.g., unknown, prefer not to declared)
6. Occupation group, including:
    a) health care worker,
    b) school or daycare worker (or attendee),
    c) long-term care resident,
    d) other occupation, and
    e) NULL
7. Asymptomatic: Yes, No, and NULL
8. Hospital status, including:
    a) hospitalized in the ICU,
    b) hospitalized but not in the ICU
    c) not hospitalized, and
    d) NULL
9. Transmission method, including:
    a) community exposures,
    b) travel exposures, and
    c) NULL
10. Clinical outcome: Recovered, death, and NULL
11. Recovery week, including NULL

As of November 12, the dataset has captured 209,811 COVID-19 cases in Canada. When conducting data analytics, our tool ignores the unique identifier and focuses on the remaining 10 attributes. There are 5 generalized regions, 53 weeks in year 2020 (for episode week), 2 stated gender values, 8 stated age groups, 4 stated occupation groups, 2 Boolean values (i.e., yes or no) indicating whether the case is asymptomatic or not, 3 stated hospital status, 2 stated transmission methods, 2 stated values for clinical outcomes (i.e., recovery or death), and 53 weeks in year 2020 (for recovery week). Thus, there can be 21,573,120 possible combinations of stated values for these 10 attributes.

#### *2) Handling of NULL values*

Knowing that, with exception of the five generalized regions, the remaining nine attributes are nullable. This adds to each attribute another possible value (namely, NULL to indicate that the value for the attribute is unstated or unknown). This increases the number of possible combinations of (stated and unstated) values for the 10 attributes to 212,576,400.

#### *3) Preprocessing of Data with Taxonomy and OLAP*

On the one hand, capturing COVID-19 cases in the dataset in a data cube provides users flexibility to apply OLAP operations such as drill downs to the details of specific cases, roll ups to some aggregated counts. On the other hand, having an ALL value capturing aggregated counts for each dimension further increases the number of combinations. This is also the total number of cells in the data cube—namely, 1,161,600,000 cells, i.e., around 1.2 billion cells.

With this setting, frequent patterns can be found by setting some attribute values to ALL. Specifically, setting values of all attributes—except the transmission method—to ALL, i.e., the cell ⟨ALL, ..., ALL, transmission method = "community exposures", ALL, ALL⟩, gives frequency of a singleton pattern {community exposures}. This reveals that 164,280 COVID-19 cases in Canada were transmitted through community exposures. Replacing the value of transmission method gives two related singleton patterns {travel exposures} and {unstated exposures}. The two corresponding cells in the data cube reveal that (a) 4,810 cases were transmitted through travel exposures and (b) transmissions of other 40,721 cases were unstated.

With these frequency values for related attribute-values about transmission methods, our tool provides users with relative frequency information. Specifically, it reports to the users that, among 209,811 cases, 78.3% of cases were transmitted through community exposures, 2.3% of cases were transmitted through travel exposures and transmissions of the remaining 19.4% were unstated. See Table VII.

TABLE VII. DISTRIBUTION OF TRANSMISSION METHOD

| **Transmission Method** | **Absolute Frequency (i.e., #Cases)** | **Relative Frequency** |
|---|---|---|
| Community exposures | 164,280 | 78.3% |
| Travel exposures | 4,810 | 2.3% |
| Unstated transmission method | 40,721 | 19.4% |
| **Total** | **209,811** | **100%** |

Moreover, in addition to showing the relative frequencies with NULL category for the attribute transmission methods, our tool also provides users with flexible of ignoring the NULL category for the attribute and focusing only on stated/known values. Specifically, it reports to the users that, among 169,090 cases with known transmission methods, 97.2% of cases were transmitted through community exposures and the remaining 2.8% were transmitted through travel exposures. See Table VIII.

TABLE VIII. DISTRIBUTION OF STATED TRANSMISSION METHOD

| ***Stated* Transmission Method** | **Absolute Frequency (i.e., #Cases)** | **Relative Frequency** |
|---|---|---|
| Community exposures | 164,280 | 97.2% |
| Travel exposures | 4,810 | 2.8% |
| **Total** | **169,090** | **100%** |

Similarly, our tool provides data distributions for some other attributes. See Tables IX to XII, which reveal knowledge like:

- A majority of cases covered.
- More than half of cases were not hospitalized.
- There is no significant difference between the genders, though slightly more female cases than males.
- There is also no significant difference among most age groups, though slightly more young cases (especially, those in their 20s) than the elderly.

TABLE IX. DISTRIBUTION OF CLINICAL OUTCOME

| **Clinical Outcome** | **Frequency** | | |
|---|---|---|---|
| | ***Absolute*** | ***Relative*** | |
| Recovered | 158,528 | 94.3% | 75.6% |
| Deceased | 9,541 | 5.7% | 4.5% |
| **#cases w/ *stated* clinical outcome** | **168,069** | **100%** | **80.1%** |
| Unstated clinical outcomes | 41,742 | | 19.9% |
| **Total #cases** | **209,811** | | **100%** |

TABLE X. DISTRIBUTION OF HOSPITAL STATUS

| **Hospital Status** | **Frequency** | | |
|---|---|---|---|
| | ***Absolute*** | ***Relative*** | |
| Not hospitalized | 122,174 | 89.5% | 58.23% |
| Hospitalized but not ICU admitted | 11,140 | 8.2% | 5.31% |
| ICU admitted | 3,203 | 2.3% | 1.53% |
| **#cases w/ *stated* hospital status** | **136,517** | **100%** | **65.07%** |
| Unstated hospital status | 73,294 | | 34.93% |
| **Total #cases** | **209,811** | | **100%** |

TABLE XI. DISTRIBUTION OF GENDER

| **Gender** | **Frequency** | | |
|---|---|---|---|
| | ***Absolute*** | ***Relative*** | |
| Female | 106,878 | 53.0% | 50.9% |
| Male | 94,736 | 47.0% | 45.2% |
| **#cases w/ *stated* gender** | **201,614** | **100%** | **96.1%** |
| Unstated gender | 8,197 | | 3.9% |
| **Total #cases** | **209,811** | | **100%** |

TABLE XII. DISTRIBUTION OF AGE GROUP

| **Age Group** | **Frequency** | | |
|---|---|---|---|
| | ***Absolute*** | ***Relative*** | |
| 20s | 40,029 | 19.12% | 19.08% |
| 30s | 32,904 | 15.72% | 15.68% |
| 40s | 30,637 | 14.63% | 14.60% |
| 50s | 28,960 | 13.83% | 13.80% |
| ≤ 19 years old | 24,487 | 11.70% | 11.67% |
| ≥ 80s | 22,241 | 10.62% | 10.60% |
| 60s | 18,206 | 8.70% | 8.68% |
| 70s | 11,890 | 5.68% | 5.67% |
| **#cases w/ *stated* age group** | **209,354** | **100%** | **99.78%** |
| Unstated age group | 457 | | 0.22% |
| **Total #cases** | **209,811** | | **100%** |

### *4) Mining of Frequent Patterns*

While our tool makes good use of the data cube in providing users with insight about distributions of different attributes (i.e., singleton patterns and their related patterns), searching through numerous cells in a data cube can be time consuming. Hence, our tool provides users with an alternative by using traditional frequent pattern mining algorithms to find frequent patterns. See Table XIII for top-10 frequent patterns, Table XIV for top-5 frequent singleton patterns (i.e., patterns involving only one attribute), and Table XV for top-10 frequent non-singleton patterns (i.e., patterns involving more than one attributes). They reveal knowledge like:

- A majority of cases were (a) transmitted via community exposures and (b) recovered.
- A majority of those community-exposed cases were (a) recovered and (a) not hospitalized.
- More than half of the cases were not hospitalized.

- Slightly more than half of the cases were females.
- Many cases that were not hospitalized and recovered.

TABLE XIII. TOP-10 FREQUENT PATTERNS

| Frequent Pattern | Frequency | |
|---|---|---|
| | *Absolute* | *Relative* |
| {community exposures} | 164,280 | 78.3% |
| {recovered} | 158,528 | 75.6% |
| {community exposures, recovered} | 130,291 | 62.1% |
| {not hospitalized} | 122,174 | 58.2% |
| {community exposures, not hospitalized} | 115,448 | 55.0% |
| {female} | 106,878 | 50.9% |
| {not hospitalized, recovered} | 97,422 | 46.4% |
| {male} | 94,736 | 45.2% |
| {community exposures, not hospitalized, recovered} | 92,559 | 44.1% |
| {female, community exposures} | 88,480 | 42.2% |
| **All Canadian COVID-19 cases** | **209,811** | **100%** |

TABLE XIV. TOP-5 FREQUENT SINGLETON PATTERNS

| Frequent Singleton Pattern | Frequency | |
|---|---|---|
| | *Absolute* | *Relative* |
| {community exposures} | 164,280 | 78.3% |
| {recovered} | 158,528 | 75.6% |
| {not hospitalized} | 122,174 | 58.2% |
| {female} | 106,878 | 50.9% |
| {male} | 94,736 | 45.2% |
| **All Canadian COVID-19 cases** | **209,811** | **100%** |

TABLE XV. TOP-10 FREQUENT NON-SINGLETON PATTERNS

| Frequent Non-singleton Pattern | Frequency | |
|---|---|---|
| | *Absolute* | *Relative* |
| {community exposures, recovered} | 130,291 | 62.1% |
| {community exposures, not hospitalized} | 115,448 | 55.0% |
| {not hospitalized, recovered} | 97,422 | 46.4% |
| {community exposures, not hospitalized, recovered} | 92,559 | 44.1% |
| {female, community exposures} | 88,480 | 42.2% |
| {female, recovered} | 84,371 | 40.2% |
| {male, community exposures} | 75,294 | 35.9% |
| {male, recovered} | 72,786 | 34.7% |
| {female, community exposures, recovered} | 71,138 | 33.9% |
| {female, not hospitalized} | 65,204 | 31.1% |
| **All Canadian COVID-19 cases** | **209,811** | **100%** |

Based on the discovered frequent patterns, our tool allows users to further explore and expand the discovered patterns. To elaborate, after finding a frequent singleton pattern {community exposures}, our tool allows users to expand the pattern to explore the hospital status. As shown in Table XVI, the expanded pattern {community exposures, not hospitalized}—which is a frequent non-singleton pattern—reveals that a majority of cases transmitted via community exposures did not need to be hospitalized. Along this direction, the users can further explore the expanded pattern to find patterns like {community exposures, not hospitalized, recovered}, which reveals that a majority of cases transmitted via community exposures were not hospitalized but recovered.

In addition to showing these frequent patterns, our tool also returns other patterns (which may be not so frequent) related to the frequent patterns. As a side-benefit, these related patterns—as shown in Tables XVI to XVIII—provide additional information such as relative frequency of the frequent patterns with respect to all cases and/or groups of cases. For instance, Table XVI reveals that most (i.e., 78.3%) of the cases were exposed through the community (rather than other transmission methods).

TABLE XVI. SOME FREQUENT PATTERNS AND THEIR RELATED PATTERNS

| Pattern | | | Frequency | |
|---|---|---|---|---|
| | | | *Absolute* | *Relative* |
| {community exposures} | | | 164,280 | 78.3% |
| {community exposures, | not hospitalized} | | 115,448 | 55.0% |
| | not hospitalized, | recovered} | 92,559 | 44.1% |
| | | unstated clinical outcome} | 20,328 | 9.7% |
| | | deceased} | 2,561 | 1.2% |
| | unstated hospital status} | | 35,869 | 17.1% |
| | non-ICU hospitalized} | | 10,190 | 4.9% |
| | ICU hospitalized} | | 2,773 | 1.3% |
| {unstated transmission method} | | | 40,721 | 19.4% |
| {travel exposures} | | | 4,810 | 2.3% |
| **All Canadian COVID-19 cases** | | | **209,811** | **100%** |

TABLE XVII. SOME FREQUENT PATTERNS ABOUT COMMUNITY EXPOSURES AND THEIR RELATED PATTERNS

| Pattern with Community Exposures & Hospital Status | | Frequency | | |
|---|---|---|---|---|
| | | *Absolute* | *Relative* | |
| {community exposures, | not hospitalized} | 115,448 | 89.9% | 70.3% |
| | non-ICU hospitalized} | 10,190 | 7.9% | 6.2% |
| | ICU hospitalized} | 2,773 | 2.2% | 1.7% |
| **Total for all *stated* hospital status assoc with {community exposures}** | | **128,411** | **100%** | **78.2%** |
| {community exposures, | unstated hospital status} | 35,869 | | 21.8% |
| **Total for all hospital status assoc with {community exposures}** | | **164,280** | | **100%** |

TABLE XVIII. SOME FREQUENT PATTERNS ABOUT COMMUNITY EXPOSURES & NON-HOSPITALIZATION, AND THEIR RELATED PATTERNS

| Pattern with Community Exposures, Non-hospitalization & Clinical Outcome | | Frequency | | |
|---|---|---|---|---|
| | | *Absolute* | *Relative* | |
| {com. exp., not hosp., | recovered} | 92,559 | 97.3% | 80.2% |
| | deceased} | 2,561 | 2.7% | 2.2% |
| **Total for all *stated* clinical outcomes assoc w/ {com. exp., not hospitalized}** | | **95,120** | **100%** | **82.4%** |
| {com. exp., not hosp., | unstated clinical outcome} | 20,328 | | 17.6% |
| **Total for all clinical outcomes assoc w/ {community exp., not hospitalized}** | | **115,448** | | **100%** |

Among these 164,280 community-exposed cases, 115,448 (i.e., 70.3% as indicated in Table XVII) of them—which is 55.0% of all COVID-19 cases—were not hospitalized. Only 6.2% were hospitalized (non-ICU or ICU admitted). Table XVII also reveals that, when considering only 128,411 community-exposed cases with stated clinical outcomes (by ignoring those 35,869 cases with unstated clinical outcomes, which account for 21.8% of community-exposed cases or 17.1% of all cases), 89.9% of them did not need to be hospitalized.

To a further extent, among the 115,448 non-hospitalized community-exposed cases, 92,559 (i.e., 80.2% as indicated in Table XVIII) of them—which account for 70.3% of community-exposed cases and 44.1% of all COVID-19 cases) were recovered. The table also reveals that, when considering only 95,120 non-hospitalized community-exposed cases with stated clinical outcomes (by ignoring those 20,328 non-hospitalized community-exposed cases with unstated clinical outcomes, which account for 17.6% of non-hospitalized community-exposed cases, 12.4% of community-exposed cases, or 9.7% of all cases), 97.3% of them were recovered.

*5) Prediction of Outcomes by Supervised Learning*

Once frequent patterns (especially, frequent non-singleton patterns) are discovered, our tool makes good use of them in forming association rules. These rules are then used in associative classification—i.e., associative supervised learning—for predicting the clinical outcomes.

For example, based on frequent patterns {community exposures, not hospitalized, recovered} and {community exposures, not hospitalized} with respective frequencies of 92,559 and 115,448, our tool infers an associative classification rule:

{community exposures, not hospitalized} → recovered,

which is supported by 92,559 COVID-19 cases and with 80% confidence. Similarly, based on frequent patterns {community exposures, recovered} and {community exposures} with respective frequencies of 130,291 and 164,280, our tool infers another rule:

{community exposures} → recovered,

which is supported by more cases (i.e., 130,291 cases) and with 79% confidence. Some additional samples of associative classification rules for the prediction of clinical outcomes are shown in Table XIX.

TABLE XIX. SAMPLE RULES FOR CLINICAL OUTCOME PREDICTION

| Associative Classifier Prediction | Support | Confidence |
|---|---|---|
| {travel exp., not hospitalized} → recovered | 3,003 | 97.4% |
| {male, travel exp. not hospitalized}→ recov'd | 1,608 | 97.5% |
| {40s, community exposures} → recovered | 20,688 | 85.4% |
| {50s, community exposures} → recovered | 19,270 | 84.5% |
| {40s, not hospitalized} → recovered | 16,683 | 83.6% |
| {50s, not hospitalized} → recovered | 14,821 | 83.5% |
| {30s, community exposures} → recovered | 21,343 | 83.1% |
| {20s, community exposures} → recovered | 24,444 | 81.9% |
| {60s, community exposures} → recovered | 11,394 | 80.9% |

### *B. Functionality Check with Related Works*

After demonstrating the features and usefulness of our machine learning tool 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 privacy-preserving details and epidemiological characteristics of those COVID-19 cases, which are provided by our tool. Second, for those related works that provide overall data distribution of cases, they are mostly confined to single dimensions/attributes. In contrast, our tool provides multi-dimensional information such as relationships among attributes in the form of frequent patterns (and their related patterns) and associative classification rules. Third, for related works focused on prediction, they mostly predict the trends (e.g., number of new cases) instead of clinical outcomes. In contrast, our tool makes good use of the discovered frequent patterns discovered from historical data to predict clinical outcomes for future data.

## V. CONCLUSIONS

In this paper, we presented a machine learning tool for big analytics on big COVID-19 epidemological data. The tool makes good use of taxonomy and OLAP to generalize some attributes for effective analysis. Instead of ignoring unstated values of some attributes, the tool provides users with flexibility of including or excluding these values. Moreover, the tool also discovers frequent patterns and their related patterns, which help reveal some useful knowledge such as absolute and relative frequency of the patterns. Our tool trains a supervised learning model based on the frequent patterns discovered from historical data, and predicts clinical outcomes (e.g., recovered or deceased from COVID-19) for future data. Evaluation results show the practicality of our tool 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, we transfer knowledge learned from the current work to machine learning and analytics of big data in many other real-life applications and services. Moreover, we explore the incorporation of our machine learning tool with a COVID-19 visualizer [52] such that the machine learning serves as a back-end engine for big data analytics and the visualizer serves as a front-end interface for information visualization and visual analytics of big COVID-19 epidemological data.

## ACKNOWLEDGMENT

This work is partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), as well as the University of Manitoba.

## REFERENCES

[1] A. Alim, X. Zhao, J. Cho, F. Chen, “Uncertainty-aware opinion inference under adversarial attacks,” in IEEE BigData 2019, pp. 6-15.

[2] 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.

[3] C.K. Leung, et al., “Fast algorithms for frequent itemset mining from uncertain data,” in IEEE ICDM 2014, pp. 893-898.

[4] G. Chatzimilioudis, et al., “A novel distributed framework for optimizing query routing trees in wireless sensor networks via optimal operator placement,” JCSS 79(3), 2013, pp. 349-368.

[5] A. Cuzzocrea, “Combining multidimensional user models and knowledge representation and management techniques for making web services knowledge-aware,” WIAS 4(3), 2006, pp. 289-312.

[6] C. He, S. Sun, B. Li, X. Tu, D. Yu, “Finding mutual X at WeChat-scale social network in ten minitues,” in IEEE BigData 2019, pp. 288-297.

[7] F. Jiang, C.K. Leung, S.K. Tanbeer, “Finding popular friends in social networks,” in CGC 2012, pp. 501-508 .

[8] C.K. Leung, C.L. Carmichael, “Exploring social networks: a frequent pattern visualization approach,” in IEEE SocialCom 2010, pp. 419-424.

[9] C.K. Leung, A. Cuzzocrea, J.J. Mai, D. Deng, F. Jiang, “Personalized DeepInf: enhanced social influence prediction with deep learning and transfer learning,” in IEEE BigData 2019, pp. 2871-2880.

[10] C.K. Leung, F. Jiang, “Big data analytics of social networks for the discovery of "following" patterns,” in DaWaK 2015, pp. 123-135.

[11] A.K. Chanda, et al. “A new framework for mining weighted periodic patterns in time series databases,” ESWA 79, 2017, pp. 207-224.

[12] C.K. Leung, R.K. MacKinnon, Y. Wang, “A machine learning approach for stock price prediction,” in IDEAS 2014, pp. 274-277.

[13] R. Sharma, A. Mateush, J. Übi, “Tale of three states: analysis of large person-to-person online financial transactions in three Baltic countries,” in IEEE BigData 2019, pp. 1497-1505.

[14] 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.

[15] 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

[16] C.K. Leung, et al., “Data mining on open public transit data for transportation analytics during pre-COVID-19 era and COVID-19 era,” in INCoS 2020, pp. 133-144.

[17] C.K. Leung, et al., “Urban analytics of big transportation data for supporting smart cities,” in DaWaK 2019, pp. 24-33.

[18] 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.

[19] O.A. Sarumi, C.K. Leung, “Exploiting anti-monotonic constraints for mining palindromic motifs from big genomic data,” in IEEE BigData 2019, pp. 4864-4873.

[20] P. Gupta, et al., “Vertical data mining from relational data and its application to COVID-19 data,” Big Data Analyses, Services, and Smart Data, 2021, pp. 106-116.

[21] 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.

[22] S. Tsumoto, et al., “Estimation of disease code from electronic patient records, in IEEE BigData 2019, pp. 2698-2707.

[23] C.K. Leung, F. Jiang, “A data science solution for mining interesting patterns from uncertain big data,” in IEEE BDCloud 2014, pp. 235-242.

[24] A. Fariha, et al., “Mining frequent patterns from human interactions in meetings using directed acyclic graphs,” in PAKDD 2013 (I), pp. 38-49.

[25] C.K. Leung, “Uncertain frequent pattern mining,” Frequent Pattern Mining, 2014, pp. 417-453.

[26] A.Y. Shahir, et al., “Mining vessel trajectories for illegal fishing detection, in IEEE BigData 2019, pp. 1917-1927.

[27] 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.

[28] 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.

[29] K.J. Morris, et al., “Token-based adaptive time-series prediction by ensembling linear and non-linear estimators: a machine learning approach for predictive analytics on big stock data,” in IEEE ICMLA 2018, pp. 1486-1491.

[30] A. Cuzzocrea, et al., “OLAP analysis of multidimensional tweet streams for supporting advanced analytics,” in ACM SAC 2016, pp. 992-999.

[31] A. Cuzzocrea, C.K. Leung, “Efficiently compressing OLAP data cubes via R-tree based recursive partitions,” in ISMIS 2012, pp. 455-465.

[32] A. Cuzzocrea, I. Song, “Big graph analytics: the state of the art and future research agenda,” in DOLAP 2014, pp. 99-101.

[33] S. Hirai, K. Yamanishi, “Detecting model changes and their early warning signals using MDL change statistics,” in IEEE BigData 2019, pp. 84-93.

[34] C.K. Leung, “Mathematical model for propagation of influence in a social network,” Encyclopedia of Social Network Analysis and Mining, 2e, 2018, pp. 1261-1269

[35] 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.

[36] M.B. Jamshidi, et al., “Artificial intelligence and COVID-19: deep learning approaches for diagnosis and treatment,” IEEE Access 8, 2020, pp. 109581-109595.

[37] 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.

[38] 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

[39] C.S. Eom, et al., “Effective privacy preserving data publishing by vectorization,” Information Sciences 527, 2020, pp. 311-328.

[40] C.K. Leung, et al., “Privacy-preserving frequent pattern mining from big uncertain data,” in IEEE BigData 2018, pp. 5101-5110.

[41] A.M. Olawoyin, et al., “Privacy-preserving spatio-temporal patient data publishing,” in DEXA 2020 (II), pp. 407-416.

[42] 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

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] D. Shen, et al., “Guest editorial: special issue on imaging-based diagnosis of COVID-19,” IEEE TMI 39(8), 2020, pp. 2569-2571.

[48] 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.

[49] A. Viguerie, et al., “Simulating the spread of COVID-19 via a spatially-resolved susceptible-exposed-infected-recovered-deceased (SEIRD) model with heterogeneous diffusion,” Appl. Math. Lett. 111, 2021, pp. 106617:1-106617:9.

[50] World Health Organization, WHO coronavirus disease (COVID-19) dashboard. https://covid19.who.int/

[51] 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

[52] C.K. Leung, et al, “Big data visualization and visual analytics of COVID-19 data,” in IV 2020, pp. 387-392. doi:10.1109/IV51561.2020.00073