International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com

Volume 10, Issue 7, May-June-2024 | Published Online : 20th June 2024

Exploring Data Science for COVID-19 Epidemiological Analysis: A Machine Learning Approach

R. Suganya[1], Sreejith S[2], Ankitha [3], Sai Moneesh[4], Surya [5]

[1]Associate Professor, Dept. of CSE (Data Science),

2, 3, 4,5New Horizon College of Engineering

ABSTRACT

In the contemporary era of big data, vast quantities of data can be generated and gathered from diverse, rich sources at an unprecedented pace. Within this massive volume of data lies valuable information and insights. For instance, healthcare and epidemiological data, such as information pertaining to patients affected by viral diseases like COVID-19, offer a wealth of knowledge. Extracting knowledge from these epidemiological datasets through data science enables researchers, epidemiologists, and policymakers to gain deeper insights into the disease. This understanding can inspire the development of strategies for detection, control, and mitigation of the disease. In this paper, we introduce a data science solution designed to analyse extensive COVID-19 epidemiological data. This solution aids users in comprehending detailed information regarding confirmed COVID-19 cases. Our evaluation results highlight the advantages of our data science solution in uncovering significant knowledge from substantial COVID-19 datasets.

Keywords : Big data, Epidemiological data, COVID-19, Data science Healthcare, Knowledge discovery, Disease control

I. INTRODUCTION

Big data represents an emerging paradigm applicable to data whose sheer volume surpasses the capabilities of commonly used software tools to capture, manage, and process within a reasonable timeframe. This data, which can be in various formats and types, is often generated or collected at high velocities from a diverse array of rich sources. Examples include:

- Audio and video such as music data
- Biodiversity data
- Biomedical and healthcare data, along with disease reports
- Census data
- Meteorological data
- Patent records
- Social media and social network data
- Time series
- Transportation and urban data for smart cities
- Web content and web logs

Furthermore, these big data can possess high value and vary in veracity, impacting their reliability for business decision- making. Consequently, massively parallel processing databases, scalable storage systems, cloud computing platforms, and high-performance computing techniques (such as MapReduce, edge

computing, fog computing, and dew computing) are essential for managing big data.

II. RELATED WORK

Due to its global impacts, numerous studies have been conducted related to COVID-19. For example, in the social sciences, researchers have investigated crisis management in response to the outbreak. In the medical and health sciences, studies have focused on

Beyond handling, big data science is crucial for uncovering implicit, previously unknown, and useful information embedded within big data. Generally, data science leverages data analytics [high- performance computing [visual analytics techniques and data mining and machine learning. Analysing and mining big data can be beneficial for social good. For example, analysing healthcare data and disease reports enhances our understanding of diseases.

Over the past century, several notable diseases have emerged, including the 1918 "Spanish flu" pandemic (1918–1920), the 1957–1958 "Asian flu" pandemic, the 1968 "Hong Kong flu" pandemic (1968–1970), the 2009 "Swine flu" pandemic (2009–2010), and the coronavirus disease 2019 (COVID-19). The latter emerged in 2019, became a pandemic in 2020, and continued to prevail into 2021.[1] managing clinical and treatment information and the development of vaccines. From the perspective of natural sciences and engineering, researchers have employed techniques such as data mining, data science, machine learning, mathematics, and statistics to contribute to

COVID-19 research. For instance, some have used artificial intelligence (AI) to track, test, diagnose, and treat COVID-19, including the detection of cases through AI-driven analyses of chest computed tomography (CT) images.

In contrast, our focus is on analysing large volumes of alphanumeric COVID-19 epidemiological data. Many related works on COVID-19 epidemiological data, especially notable dashboards, primarily aim to report the numbers of new cases and deaths. These numbers are often visualized graphically to aid comprehension for the general public. However, visualizing this information using bubble maps can lead to overlapping bubbles, making it difficult to discern individual data points. Similarly, choropleth maps use shading or colouring to represent numbers, where darker shades indicate higher numbers, but small locations may not be easily visible.

Beyond reporting the number of cases and deaths, it is crucial to examine common characteristics associated with COVID-19 cases. For instance, mining COVID-19 epidemiological data can reveal valuable insights such as transmission methods, symptomatic versus asymptomatic cases, and hospitalization requirements (e.g., ICU, regular wards, no hospitalization). Understanding transmission methods can help decision-makers and health officers implement measures to break transmission links. Information on symptomatic cases aids healthcare providers in detecting COVID-19, while data on asymptomatic cases is vital for preventing disease spread through measures like self-quarantine. Knowing hospitalization requirements assists in planning for potential patient demand.

Consequently, several studies have analysed and visualized these characteristics of COVID-19 cases to provide a deeper understanding of the pandemic [2].

III. FINDINGS AND ANALYSIS

Over the past decade, data mining, in conjunction with statistics and big data, has evolved into the field of Data Science.

Simultaneously, the field of machine learning has seen significant growth, bringing Artificial Intelligence (AI) into prominence. While we use the terms Data Science and Machine Learning interchangeably, AI extends beyond machine learning to encompass planning, reasoning, and other advanced functions.

Data plays a crucial role in epidemiology and the study of infectious diseases. It is essential not only for detecting and preventing the spread of infectious

diseases but also for developing treatments and vaccines. For COVID-19, data is collected on infected individuals, including their personal, work, and travel details, contacts, and social activities. This data can be analysed using various data science techniques to understand and prevent the disease's spread.

For example, contact tracing graphs can be analysed using link analysis techniques to identify individuals likely to contract COVID-19 from current infections. Clustering techniques can identify potential clusters of COVID-19 cases, and decision trees can classify individuals based on susceptibility factors, such as ethnicity or location. These techniques can also help identify asymptomatic individuals who might be potential spreaders of the virus. By studying the behavior patterns and genetic makeup of asymptomatic individuals who test positive, researchers can understand why they might be asymptomatic and use this information to test and warn their contacts.

Prevention is critical, given the significant impact of the disease on survivors' lives. By gathering data from countries with few COVID-19 cases, researchers can identify trends that might help prevent the disease elsewhere. Data Science is also instrumental in developing treatments and vaccines. Data on the virus's DNA, individuals' genetic makeup, and information from databases on related infectious diseases like the Spanish Flu and SARS can be integrated and analysed to develop effective treatments and vaccines for COVID-19.

Comprehensive data collection, storage, sharing, and analysis are vital for developing solutions to the pandemic. These solutions can be medical, such as treatments and vaccines, or behavioural and social, such as wearing protective equipment. Experiments, like comparing infection rates between groups wearing different protective gear or studying outcomes of varying social gathering sizes, can provide valuable insights. Accurate and reliable data is essential for these analyses, and thorough risk analysis under various scenarios is crucial to inform the public about potential dangers effectively.[4]

IV. CONCLUSION

In this paper, we introduce a machine learning tool designed for comprehensive analysis of extensive COVID-19 epidemiological data. This tool effectively utilizes taxonomy and OLAP to generalize certain attributes, facilitating a more efficient analysis. Rather than disregarding unstated attribute values, it offers users the flexibility to include or exclude these values as needed. Additionally, the tool identifies frequent patterns and their related patterns, thereby uncovering valuable knowledge such as the absolute and relative frequencies of these patterns.

Our tool employs a supervised learning model trained on frequent patterns derived from historical data to predict clinical outcomes (e.g., recovery or death from COVID-19) for future cases. The evaluation results demonstrate the tool's practicality in providing detailed insights into the characteristics of COVID-19 cases, aiding researchers, epidemiologists, and policymakers in gaining a deeper understanding of the disease. This enhanced understanding can inspire strategies for detection, control, and combatting the disease.

For ongoing and future work, we aim to transfer the knowledge gained from this study to the machine learning and analytics of big data in various other real-life applications and services. Additionally, we are exploring the integration of our machine learning tool with a COVID-19 visualizer, where the machine learning component acts as a back-end engine for big data analytics and the visualizer functions as a front-end interface for information visualization and visual analytics of extensive COVID-19 epidemiological data.[3]

II. REFERENCES

[1]. Carson K Leung, yubo Chen, Big Data Science on COVID 19 Data, Publisher: IEEE, Year:2022

[2]. Carson K. Leung; Chenru Zhao, Big Data Intelligence Solution for Health Analytics of

COVID-19 Data with Spatial Hierarchy,Publisher: IEEE, Year: 2021

[3]. Carson K. Leung, Yubo Chen, Calvin S.H Machine Learning and OLAP on Big COVID-19 Data Publisher: IEEE, Year: 2020

[4]. Bhavani Thuraisingham, Data Science, COVID-19 Pandemic, Privacy and Civil Libertie,Publisher: IEEE, Year: 2020