Analysis of COVID-19 blockchain data using social signals for contact tracing

by

Tadepalli Sarada Kiranmayee

A thesis submitted to
The Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements
of the degree of

Master of Science

Department of Computer Science
The University of Manitoba
Winnipeg, Manitoba, Canada
December 2021

Thesis advisor

Dr. Ruppa K. Thulasiram

Author

Tadepalli Sarada Kiranmayee

Analysis of COVID-19 blockchain data using social signals for contact tracing

Abstract

Communities around the world are adapting to the fast-changing global situation due to the unprecedented effect of COVID-19. The data related to COVID-19 spread is being analyzed for identifying outbreaks and for trying to predict their future movement across geographies with the help of advanced machine learning models. The main challenge faced by researchers was using the centralized data sources, aggregating relevant data, and standardizing them, at a global level.

In my thesis, I have first studied the difference between the centralized and blockchain-backed data provided by Mipasa (powered by the IBM Blockchain Platform and the IBM Cloud) and have developed a knowledge graph for COVID-19 cases in USA and Japan. The creation of a knowledge graph helped in predicting the regions which could witness the formation of new clusters. This observation helped in isolating regions and prevented the further spread of the virus. This led to the development of Decentralized Applications.

The first decentralized application for COVID-19 symptoms tracking using Blockchain is developed to enhance reliable data collection for training Machine Learning (ML) models. The Blockchain integration in this application helped patients to provide

COVID-19 symptoms data with trust. In addition to this, the data was first verified by an entity of the decentralized network (e.g. a COVID-19 testing lab). Then, with the consent of the patient, this data was provided to the centralized system for re-training the ML model. This re-training performed with verified data, updates the ML model and provides accurate results.

The data collected from different platforms helped in identifying outbreaks, contact tracing, and the creation of a machine learning model for predicting the future movements of the outbreak to minimize the spread. However, there was delay in taking measures which cost many lives, and many local businesses were shut down around the world. As a solution to this problem, I have developed a Decentralized architecture where the Blockchain Oracle smart contract could access the data outside the Blockchain, the second application. The REST API provided the daily aggregated data from three platforms (Twitter, Google Mobility Data, and COVID-19 cases) for Toronto, Canada. This aggregated data helped participants of the Blockchain ecosystem to initiate the smart contracts early decisions like implementing lockdown, deploying officials for contact tracing, and providing financial support for businesses and individuals affected due to lockdown.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to begin by thanking God, my advisor, committee, family, friends and everyone who have supported me along the way.

This thesis is dedicated to my Mother, Husband and Daughter.

Chapter 1

Introduction

Coronavirus (COVID-19), a new virus outbreak that is believed to have started in early December 2019 in Wuhan, China. It rapidly spread to different countries around the world [1] and was declared as a pandemic by World Health Organization (WHO) on the 11th of March 2020. By the end of March 2020 around 166 countries were affected [2] and has been affecting countries with its mutations (delta, omicron etc.) even today after almost 2 years since its initial days. A unified platform, where governments, local authorities and hospitals can share COVID-19 data with trust is absolutely essential. This data will help in building applications that will provide accurate results for preventing further spread. The applications which are developed using trusted and verifiable data also guarantees the data privacy of the patients [3]. The required properties for data analysis, i.e., trust and verifiable data - can be provided by Blockchain technology. It maintains an immutable ledger of records in a decentralized way has attracted various researchers and businesses. This concept was initially used for with cryptocurrencies. However, it is constantly explored in

different areas as a source of trust or data validation.

Many data analysts around the world are helping in providing best insights from the COVID-19 data (confirmed cases, deaths and recovered cases). This helps governments of different countries to create best policies or measures for their citizens. The COVID-19 data was provided by many agencies like World Health Organization (WHO), European Centre for Disease Prevention and Control (ECDC) and John Hopkins University etc. The COVID-19 Situation Report by WHO provided these statistical data. However, on the 18th of March 2020, for Situation Report 58 they shifted the cutoff time that compromised the data (the data from midnight to 9 am (CET) overlapped) [4]. Integrating worldwide COVID-19 related data and providing it to others for creation of predictive models. **Mipasa** - "is a multi-party, multi-source verifiable data processing, analytics and sharing platform" [5] developed by HACERA corporation. It provides solution for this problem by standardizing and normalizing information from all contributing data streams (for example, date and time of information using a unified timezone or using ISO 3166 for country codes). It uses blockchain technology called **Hyperledger fabric** that helps in tamper-proof data consumption. Mipasa was announced on 27th of March 2020 by WHO in partnership with technology giants, government agencies and international health organizations. It is a blockchain based control and communications system that aims for rapid and precise detection of COVID-19 carriers and hot spots. It also promises to share health and location data among the levels of organizations with privacy as its main consideration [6]. The data provided by the promising blockchain backed technology helps the data analyst to perform productive models. This study my thesis on blockchain

data analysis for COVID-19 helped in creating two architectures for decentralized applications.

The first decentralized application, eliminates the current problem of interoperability of electronic medical information between organizations where the data is centralized. Consider the COVID-19 data collection where organizations like WHO requires a platform for smooth data interoperability. Early symptoms diagnosis will help in reducing the COVID-19 fatality rate. Numerous applications [7][8] are developed for self - diagnosis, collection and storage of data in a centralized system to train machine learning (ML) models. Currently, patients report the symptoms which are stored in the centralized storage and then they are verified by a lab. Meanwhile, the ML models are trained with this early data resulting in incorrect predictions. So, the COVID-19 data should be verified by a medically trusted authority to prevent incorrect conclusions. Addressing this issue, a platform is developed where governments, local authorities, and hospitals can share COVID-19 data with trust. This trust-worthy data collection will help in building reliable COVID-19 applications in future, which will provide accurate results and help in preventing further spread [3].

The second Decentralized architecture is developed with the Blockchain Oracle smart contract that can access the data outside the Blockchain. The REST API provided the daily aggregated data from three platforms (Twitter, Google Mobility Data, and COVID-19 cases) for Canada. This aggregated data helps the participants (Government, Non-Essential and Essential services) of the Blockchain ecosystem to initiate the smart contracts when a decision by the government like lockdown, travel restrictions, gathering restrictions and closure of the non-essential businesses etc. [9]

to reduce the COVID-19 spread so that the government can support the businesses and the unemployed personnel with relief funds instantly [10] during the COVID-19 outbreak. When implementing lockdown officials can be deployed by predicting more crowded areas by analysing the Google Mobility and Twitter data for contact tracing.

Our main contributions include:

The first decentralized application (DApp) is developed using Blockchain for COVID-19 symptoms tracking to enhance reliable data collection and training Machine Learning (ML) models. The Blockchain integration helps patients to provide COVID-19 symptoms data with trust. Moreover, the data is verified by an entity of the decentralized network (e.g. a COVID-19 testing lab) and then it is provided to the centralized system for retraining the ML model with patient's permission.

* In this study, we propose a COVID-19 early symptoms data collection using Blockchain-based architecture.

* A DApp is designed with a ML model deployed on the client-side that instantly predicts COVID-19 positive or negative based on the symptoms entered by the patient. The model predicts with 90% accuracy using SGDClassifier [11].

* Smart contracts are triggered for sending the symptoms data to different accounts in Blockchain ecosystem.

The second decentralized application analyses the social signals from different APIs like Twitter, COVID-19 Community Mobility Data by Google and data from Canada COVID-19 cases for the government to decide on lockdown. It will also help in depositing the relief fund to the non-essential business accounts in the Ethereum

private network when there is lockdown. Moreover, this architecture will update the health officials with the high mobility areas and also the essential services with the store capacity.

* In this study, we propose an Oracle based decentralized architecture with the entities like Government, essential business, non-essential business along with health officials as participates in the private Blockchain ecosystem.

* The REST API is developed to provide an interface between the processed data from different sources (i.e. Twitter, COVID-19 Community Mobility Data by Google and data from Canada COVID-19 cases) and the blockchain oracle smart contract with 'Monitor' event.

* Once the smart contract is triggered with the 'Lockdown', 'Restrictions' or 'No Lockdwon' decision will trigger the smart contracts associated to the decision taken. Ethereum accounts like health officials will get the details about mobility of the people, non-essential business will be credited with the relief funds and the essential businesses will be prompted the store capacity instantly.

The rest of the thesis is structured as follows: In Chapter 2, related work provides the recent research work on Blockchain data analysis, COVID-19 data analysis and decentralized applications. In Chapter 3, methodology is provided. In Chapter 4, showcases implementation for the Decentralized Application development followed by results and in Chapter 5, I conclude this study with several observations.

Chapter 2

Related Work

In this chapter, the related work for Blockchain and COVID-19 is discussed.

2.1 Blockchain

Christidis and Devetsikiotis [12] defined Blockchain as a distributed append-only time stamped data structure. Blockchain is a distributed peer-to-peer network among non-trusting members whose interactions are verified without the intervention of a central authority. Permissionless (public) and permissioned (private and federated) - are two different categories [13] of a Blockchain network. A new user or miner can join anytime in permissionless blockchain in the network to perform transactions. However, only the allowed users with predefined characteristics and have permissions participate in the network operations is called permissioned blockchains. In this paper, blockchain cluster analysis is performed for both permissionless platform (Bitcoin and Etherum) and permissioned platform (Hyperledger fabric).

2.1.1 Cryptocurrencies

In 2008, Bitcoin and blockchain concepts were first conceive by a pseudonym - Satoshi Nakamoto, an unknown person or a group [14]. It is a combination of cryptographic hash functions and open distributed ledger for digital currency applications. Many countries avoid Bitcoin since it violates many government policies and also it has high price volatility in the market. However, Blockchain - its underlying technology, has attracted the attention of researchers. Distributed ledger, decentralized, tamper-proof construction, information transparency and openness are few of the advantages of the Blockchain. The applications of blockchain technology is not only restricted to digital currency but also used for supply chain management, finance, monitoring market, health care system, copyright protection and smart energy etc. [15].

Ethereum framework is available to run the Decentralized Applications [16]. It is a Turing-complete programming language (Solidity) which helps the users can create "smart contracts" . It is defined as an execution of actions is specified by a set of rules. Blockchain with a complete programming language inclusion was a breakthrough that provided full freedom to developers to create applications. The Ethereum and Bitcoin process is similar. In maintaining a constant frequency of block creation in every 12-15 seconds algorithms help in regulating Proof-of-Work (PoW) puzzles difficulty. The miner is rewarded 5 ethers that has the advantages, firstly, circulation of ethers without a central authority and secondly transactions are validated in a decentralized manner. The miner hacking is significantly reduced due to the sheer computing power. However, the Ethereum is now using Proof-of-Stake

consensus algorithm.

Recently, Blockchain-based cryptocurrencies have created a new trend in the world of financial system [17]. Permission-less Blockchains store massive amounts of data. There is huge business value in performing analysis of this massive Blockchain data. These authors have collected and processed on-chain data from Ethereum. They named this dataset as XBlock-ETH, which consisted of Blockchain transactions, smart contracts, and cryptocurrencies (i.e., tokens). They presented basic statistical and exploration data analysis.

2.1.2 Blockchain Data Analysis

Ron and Shamir [18] discussed about the bitcoin - a digital payment system. Here transactions are performed anonymously by people. They collected the full history of Bitcoin, and many statistical properties were analyzed and associated with the transaction graph. This paper provided insights on various questions related to the users, for example, the behavior of the users, movement of bitcoins between accounts to maintain their privacy and how they spend and acquire bitcoins and also balance in the accounts. This analysis helped in additional discovery of strange long chains in the transaction graph and a large transaction occurred in November 2010 that the users tried to hide with fork-merge structures.

Zhou et al. [19] proposed - a ledger data query platform - Ledgerdata Refiner. It provides an interface to retrieve data such as transactions and blocks with the help of ledger data analysis. Historical data of any state can be tracked. Ledger state schemas are clustered after analysis which helps the users to query the ledger data.

Spagnuolo et al. presented a modular framework called BitIodine [20]. It parsed the blockchain network and the addresses were clustered that are likely belongs to a same user or group of users. It was implemented by classifying such users and adding labels to them. Finally, extracting complex information from the Bitcoin network and visualizing them. The labeled users are scraped from website where the Bitcoin address information is provided. It helped in manual exploration of the paths to or from an addresses or users. This framework was evaluated on many real-world scenarios like links were found to Silk Road Wallets and CryptoLocker ransomware where the details of the victims and the ransom paid by them were extracted and this prototype was used for Bitcoin forensic analysis tools.

Oggier et al. [21] created a graph mining tool, BiVA. It was used for analyzing and visualizing the Bitcoin network. Bitcoin data was used to explore and also visual the subgraphs. The algorithms include clustering algorithms and address aggregation mechanisms. The interface supports basic visualization algorithms. It was adaptable and new algorithms can be integrated. Ashley Madison data hack was provided as a case study [22].

Battista et al. [23] provided a visual analysis depicting the Bitcoin flow mixes with other currency flows in a transaction graph. It tracks the transactions and stores at multiple UTXOs (unspent transaction outputs) graphically. Some applications were accumulation, distribution, and mixing for the analysts to reveal Bitcoin flow patterns.

Chen et al. [24] collected all Ethereum transaction data. They constructed three graphs - smart contract creation graph (CCG), money flow graph (MFG) and smart

contract innovation graph (CIG). To characterize the major activities helped to draw insights from these graphs. Attack Forensics and Anomaly detection were the two new security issues that were addressed using A new approach called cross-graph analysis.

Dillenberger et al. [25] discussed about providing convenient predictive models by connecting the analytic engines to blockchains for configurable dashboards, provenance histories and compliance checking. They also showcased the combination of the blockchain data with external data sources for private analytic which is secure. This can enable creation of artificial intelligence model over geographically dispersed data, and enable provenance and lineage tracking for trusted artificial intelligence models.

In [26] the authors have tried to integrate the Blockchain as a datastore into large systems. They also discussed new challenges and the data management techniques. Firstly, the architectural layer is designed using the Blockchain as the data store. Next, they analyzed the decentralized application positioning in this architecture. Further, they explored distributed data store i.e. data administration aspect. This architecture provided trustworthy data and in turn enabled reliable data analytics. Lastly, privacy and quality assurance issues are discussed for the data governance in the Blockchain prospective.

Machine Learning models are mostly centralized. The large datasets used for analysis are often proprietary. If the data is not updated frequently by adding recent data, then the machine learning models published will be out-dated. Harris and Waggoner [27] proposed a framework to solve this problem with smart contracts to host a continuously updated model. This model is on a Blockchain so that it can be

free to use on interface. Both non-financial (gamified) and financial incentive structures for providing reliable data is proposed for maintaining the model's accuracy with respect to some test set proposed. The advantages like public and decentralized ML model, good data for model updation, and transparency and trust along with some disadvantages like overwhelming the third-party intervention, Blockchain network and non-familiarity for the collaborative and decentralized AI of their model were identified.

2.1.3 Crowdfunding

Zhu and Zhou [28] discusses a new method of raising money for startups is Equity crowdfunding via the Internet. Low barriers to entry, low cost, and high speed and it encourages innovation. This study involves current problems for equity crowdfunding in China. Based on the characteristics of blockchain technology, this study explores its practical applications in equity crowdfunding. 1) It used for the registration of stocks and shares of a firm financed by crowdfunding; 2) It simplifies the transaction and transfer of crowdfunding equities; 3) It enables peer to peer transactions between investors and entrepreneurs, and solves the problem of regulatory compliance and security of fund management; 4) It helped in regulating the market conditions, and supports regulatory activities such as managing investors and fighting money laundering.

The COVID-Crypto Relief Fund (Crypto Relief) [29] was announced on 24 April 2021 during the second wave of the COVID-19 pandemic in India. It helped in providing donations for oxygen concentrators and oxygen cylinders.

2.2 Health

2.2.1 COVID-19

Dey et al. [30] discovered the recent outbreak of pneumonia in Wuhan, China. It was caused by the SARS-CoV-2 and they prioritized the analysis of the epidemiological data of this novel virus with focus on predicting large number of people getting infected all around the globe. The study provided a way to gather epidemiological outbreak data for analyzing COVID-19. World Health Organization, Johns Hopkins University, National Health Commission, Chinese Center for Disease Control and Prevention along with an online community for physicians and health care professionals in China called DXY provided few open datasets for 2019-nCoV. An exploratory data analysis with visualizations is inevitable to understand the number of different cases reported - confirmed, recovered and death in China and around the world. Moreover, these evaluation techniques are inevitable to identify the risks and commence the containment activities at the earliest.

2.2.2 COVID-19 data analysis

Hamzah et al. [31] developed Corona Tracker - an online platform, provides latest and reliable news updates along with statistical analysis on COVID-19. Corona-Tracker community conducted this research aiming to predict and forecast COVID-19 cases, deaths, and recoveries through predictive modelling. The main significance of this model is interpreting public sentiment patterns on healthcare information, assessing political and economic influence towards the spread of the virus.

Zoabi et al. [8] suggested that the burden on healthcare system can be reduced with quick and efficient diagnosis of COVID-19. Efficient screening can help in this process by using prediction models. This paper proposed a ML approach that was trained on records from 51,831 tested individuals using only eight binary features: age $\geq$ 60 years, sex, known contact and five preliminary conditions like fever, soar throat, cough, headache and cold. Israeli Ministry of Health published a nationwide dataset that helped in developing a machine learning model to detects COVID-19 cases with a questionnaire.

Ahamad et al. [32] developed a supervised machine mearning model that helped in identifying the features and predicting COVID-19 disease with very high accuracy. Features examined provided the details of the individuals. They are - age, gender, history of travel, observation of fever, and clinical details such as the severity of cough and incidence of lung infection. Several machine learning algorithms were implemented on the collected data. XGBoost algorithm [33] performed with highest accuracy i.e., ($>$ 85%). Fever (41%), cough (30.3%), lung infection (13.1%) and runny nose (8.43%) were few of the highly predictive statistical features. However, the percentage of people who did not develop any symptoms 54.4%, that data was used for the diagnosis.

Laguarta et al. [34] trained the MIT Open Voice model by building a data collection pipeline of COVID-19 cough recordings through the website (opensigma.mit.edu) between April and May 2020 and created the largest audio COVID-19 cough with 5,320 subjects. The method used was an AI speech processing framework, which leverages acoustic biomarker feature extractors that pre-screens the COVID-19 from

cough recordings. Cough recordings are transformed with Mel Frequency Cepstral Coefficient and are inputted into a Convolutional Neural Network (CNN). The model has been trained on 4256 subjects and tested on the remaining 1064 subjects of our dataset. It achieved COVID-19 sensitivity of 98.5% with a specificity of 94.2% (AUC: 0.97) and for the asymptomatic it achieved sensitivity of 100% with a specificity of 83.2%.

Gupta et al. [35] presented a large Twitter dataset for researchers of public conversation on COVID-19 pandemic. Using natural language processing techniques and machine learning based algorithms they annotated seventeen latent semantic attributes for each public tweet. The latent semantic attributes included ten attributes for ten detected topics, five quantitative attributes for unpleasantness/pleasantness and primary emotions like fear, anger, sadness and joy, and two qualitative attributes for dominant emotion category.

Pham et al. [36] discussed the importance of big data to prevent the COVID-19 outbreak and to prevent the COVID-19 pandemic severe effects. They represented the overview of AI and big data, then the applications were identified to fight against COVID-19, challenges and issues with these new solutions were identified, and concluded with the recommendations to control the COVID-19 situation.

Sheng et al. [37] presented the importance of harnessing the big data for decision making. A review of the big data analytics and the methodological details were discussed in the management corner. They examined contemporary organizational issues and suggested further research in these areas. They provided insights on big data methods for descriptive/diagnostic, predictive and prescriptive analytics, and

leveraged them to study 'black swan' events such as the COVID-19-related global crisis and its aftermath's implications for managers and policymakers.

Wellenius et al. [38] suggested that social distancing helped to mitigate COVID-19 pandemic in the United States. From Google location history of users, the anonymous and aggregated mobility data was used to estimate how the people are responding to the social distancing campaigns in United States. By declaring a state-of-emergency, there was a 10% reduction in the people spending time away from their residence. An additional 25% reduction in movement was noticed with an addition of one or more prevention policies. The subsequent stay-at-home policy provided a supplemented 29% reduction. Their finding suggested that the state wide policies helped in promoting social distancing measures.

Milano and Cannataro [39] analyzed the Italy's COVID-19 data for the period February 24 to March 29, 2020 at the zonal levels. Using statistical testing they segregated groups of similar or dissimilar regions with respect to the ten different types of Italy's COVID-19 data. They also built several similarity matrices which were mapped into networks where nodes represented Italian regions and edges represented similarity relationships. This, network-based analysis uncovered communities of regions that indicated similar behaviour. Community detection algorithms were used for the analysis, to show similar behaviour. The temporal results showcased how regions formed communities and changed with respect to the different available data.

Samuel et al. [40] collected Coronavirus specific Tweets and identified public sentiment associated with the pandemic using sentiment analysis packages in R language. They concluded that an advancement of fear-sentiment over time over COVID-19

cases were approaching peak levels in the United States. The classification of Tweets were based on two types of implementations. Firstly, they used Naive Bayes method on short tweets and observed a strong accuracy of 91%. Secondly, they used classification method-logistic regression on shorter tweets which provided a decent accuracy of 74%. However, on longer tweets both methods displayed relatively weak performance.

Alrazaq [41] discussed the issues related to COVID-19 pandemic, which were communicated among individuals, organizations, and governments through social media. He also concluded that analysing the topics discussed on social media over COVID-19 will help policy makers and health care organizations to determine the demands of their stakeholders and address them.

Rangone and Adalberto [42] discussed about the need for new solutions that can support non-profit organizations around the world. During COVID-19 the philanthropic organisations had a negative impact on fundraising and it suffered a substantial decrease. The Blockchain can play a pivotal role to re-establish pre-pandemic standards and enhance the development of global philanthropy. This paper demonstrates the Blockchain impact on the development of charity 4.0. In this paper, Charity Wall - an Italian social marketplace is studied where the important business associations for its innovative solutions in the charity 4.0 sector provides support to NPOs during their traditional function as well as against COVID-19 in Italy. This is a benchmark analysis, highlighting the Charity Wall solutions and comparing it with charity 4.0 systems on the market.

Chapter 3

Methodology

The difference between centralized and decentralized COVID-19 data, analysis of the blockchain - backed COVID-19 data and the advantages of using Blockchain - backed data is discussed in Chapter 3.1 followed by the methodology for creating two architectures for decentralized applications described in Chapter 3.2 and 3.3.

3.1 Analysis of COVID-19 blockchain-backed data

In this chapter, the structure of the centralized datasets (The COVID Tracking Project [43], COVID-19 dataset in Japan [44] and ECDC COVID-19 [45]) are compared with the Blockchain - backed data Mipasa Blockchain backed dataset. This helps in showcasing the need and importance for normalized data. The blockchain - backed data performs easy and quick data analysis for unprecedented situation like COVID-19. Later, a knowledge tree is created for analysing the Mipasa COVID-19 datasets comparing USA and Japan's response to COVID-19 confirmed cases.

3.1.1 Data Collection

The datasets with the descriptions are described as follows:

3.1.1.1 ECDC Covid-19 data

ECDC Covid-19 data is collected from the website [45]. The dataset features are captured in Table 3.1.

Columns	Description
dateRep	Date in the dd/mm/yyyy format
Day	dd extracted from the date
month year	mm extracted from date yyyy extracted from the date
cases	Total number of COVID19 confirmed cases
deaths	Total number of COVID19 related deaths
countriesAndTerritories	Name of the country or Union Territory
geoId	Geographical ID is the Alpha-2 country code
countryterritoryCode	Alpha-3 country code
popData2019	total population of the country in 2019
continentExp	Name of the continent
Cumulativenumberfor14daysof COVID-19casesper100000	The 14-day notification rate of newCOVID-19 cases, along with the 14-day death rate are the main indicators displayed

Table 3.1: COVID-19 ECDC data description

3.1.1.2 Mipasa ECDC Covid-19 Blockchain data

The ECDC COVID-19 blockchain data is the main dataset which is collected from Mipasa Website [44]. The 'source.csv' is the dataset from the original source i.e., ECDC data. Mipasa standardized and normalized the dataset to present it in two files Output_ECDC_Cases.csv and Output_ECDC_Deaths.csv. The dataset from the Mipasa and its description is in Table 3.2.

Columns	Description
dataId	its a 32 alphanumeric value representation like Wallet ID
countryCode2	Country Codes Alpha-2
cases	number of positive cases
New_cases	Number of new cases
Cumulativecases	Number of cumulative cases

Table 3.2: Mipasa's COVID-19 ECDC Cases description

3.1.1.3 The COVID-19 tracking project

This dataset is provided by the COVID tracking website [43] and the data description is presented in Table 3.3.

Columns	Description
date	date with yyyyddmm format
state	Alpha-2 state code
positive	total number of positive cases
negative	total number of negative cases
pending hospitalizedCurrently	Total number of viral tests that have not been completed Individuals who are currently hospitalized with COVID-19
hospitalizedCumulative	Total number of individuals who have ever been hospitalized with COVID-19
inIcuCurrently	Individuals who are currently hospitalized in the Intensive Care Unit with COVID-19
inIcuCumulative	Total number of individuals who have ever been hospitalized in the Intensive Care Unit with COVID-19
onVentilatorCurrently	Individuals who are currently hospitalized under advanced ventilation with COVID-19. Defini-tions vary by state / territory
onVentilatorCumulative Total number of people that are identified as re-covered from COVID-19.	Total number of individuals who have ever been hospitalized under advanced ventilation with COVID-19 recovered
dataQualityGrade	The COVID Tracking Project grade of the completeness of the data reporting by a state
lastUpdateEt	The COVID Tracking Project compiles data once each day
dateModified	Modified date
checkTimeEt	check time
death hospitalized	Total fatalities with confirmed OR probable COVID-19 case diagnosis Total number of people hospitalised
dateChecked	date when checked
totalTestsViral	Total number of completed viral tests (or spec-imens tested)
positiveTestsViral	Total number of completed viral tests (or spec-imens tested) that return positive
negativeTestsViral	Total number of completed viral tests (or spec-imens tested) that return negative
positiveCasesViral	Total number of people with a completed viral test that return positive as reported by the state or territory
deathConfirmed	Total fatalities with confirmed COVID-19 case diagnosis
deathProbable fips	Total fatalities with probable COVID-19 case diagnosis Federal Information Processing Standards (FIPS) code for the state or territory.
grade	The COVID Tracking Project grade of the completeness of the data reporting by a state

Table 3.3: The COVID tracking project data description

3.1.1.4 The Mipasa's COVID tracking project

This dataset is collected for analysis from the Mipasa and the blockchain data is described in Table 3.4.

Columns	Description
dataId	its a 32 alphanumeric value representation like Wallet ID
date	Date when data collected (ISO 8601 time representation
stateId	US state Alpha-2
positive	number of positive cases

Table 3.4: Mipasa's The COVID tracking project data description

3.1.1.5 Mipasa's COVID-19 in Japan

Output_States.csv dataset is collected from the Mipasa and dataset is described in Table 3.5.

Columns	Description
dataId	its a 32 alphanumeric value representation like Wallet ID
date	Date when data collected (ISO 8601 time repre-sentation
Tests	Number of Testes
Confirmed	number of confirmed cases
Deaths	number of death cases
Recovered	number of recovered cases
Hospitalized	number of hospitalised cases
Severe	number of severe cases
Population	population in the state
Administrative_area_level	administrative division
administrative_area_level_2	the state name
JIS_Code	Japan State codes

Table 3.5: Mipasa's COVID-19 in Japan

3.1.1.6 COVID-19 Dataset In Japan From Kaggle

covid _jpn_prefecture.csv dataset is collected from Kaggle [46] which is processed data. The data was manually collecting from pdf files reports published by Japanese Ministry of Health and converted to excel,and this data is described in Table 3.6 .

Columns	Description
Date	YYYY-MM-DD
Prefecture	Prefecture name. Tokyo, Osaka etc.
Positive	PCR tested positive cases
Tested	PCR tested cases
Discharged	Discharged cases
Fatal	Fatal cases
Hosp_require	Requiring hospitalization
Hosp_severe	Positive and with severe symptoms

Table 3.6: COVID-19 in Japan

Note, the positive cases is reported when a doctor confirms the tests of the patients for COVID-19 at local level. However, a presumptive positive test is passed to a national level body for confirmation. Once it is confirmed by national body then it is considered as confirmed case [46]. Similarly, ECDC dataset shows the confirmed cases where as the COVID Tracking Project provides positive cases for USA.

3.1.2 Analysis

The Mipasa's COVID tracking project and Mipasa ECDC COVID-19 datasets are already normalized which makes the analysis of the data convenient because the preprocessing is minimal. On the other hand, the centralized ECDC and COVID Tracking Project datasets presented in Table 3.1 and Table 3.3 are different datasets

and the data should be universally normalized. Here we can see the advantage of using Mipasa, i.e., if the datasets are different sources and the platform provided standardized and normalized then the data analyst can provide good analysis in less time.

Lastly, collapsible tree is constructed to showcase the ten countries, the effected provinces in Japan and States in USA which were adversely affected by COVID-19. The implementation and results are discussed in Chapter 4.1.1 and Chapter 4.2.1 respectively.

3.2 COVID-19 early symptom prediction using Blockchain and machine learning

The study of blockchain - backed data helped in creating an architecture for the COVID-19 early symptom prediction and our proposed architecture is presented in Figure 3.1. The architecture is divided into three phases:

ML Model Phase, where the Israel early symptom dataset [47] is explored using python libraries for training ML model. The SGDClassifier [11] is used as the incremental ML algorithm for the DApp.

Data Transaction Phase, where the Decentralized Application (DApp) helps in the early symptom tracking. The DApp is developed using HTML, CSS and Web3Py. In this phase, the Ethereum Blockchain test network is used. The smart contracts are deployed to track the transfer between Ethereum accounts of the patients, labs and the centralized storage.

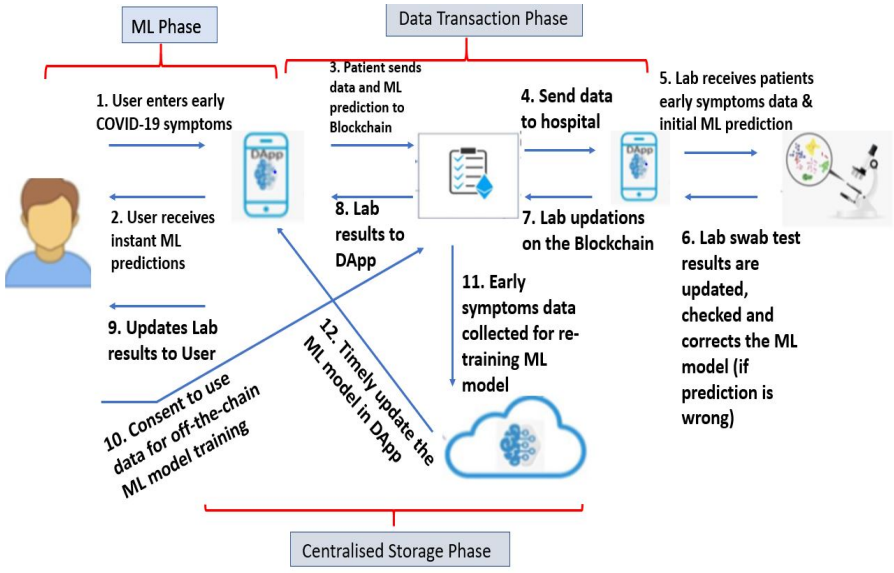

Figure 3.1: Architecture of COVID-19 Early Symptoms Prediction

Centralized Storage Phase, where with the permission of the patient, the verified data is collected in the centralized storage for retraining ML model.

The implementation is discussed in Chapter 4.1.2 and the data transactions on the decentralized network can be completed by spending Gas on the Ethereum network which is discussed in Chapter 4.2.2.1

3.3 Blockchain Oracles for COVID-19 Social Signals

Another Decentralized Application is developed to understand the COVID-19 situation by collecting the data from different public APIs like Twitter, COVID-19 Community Mobility Data by Google and Canada COVID-19 cases. Then, the data

is cleaned and processed. This clean data is stored in the REST API, which provides response to the query from the Oracle smart contract from the Blockchain ecosystem. The proposed architecture is shown in the Figure 3.2. The sequence diagram is shown in the Figure 3.3 and the sequence of steps are as follows:

- The Oracle smart contract generates an event called 'MONITOR' to query the REST API requesting for the external data (Twitter, Mobility and the COVID-19 cases data).
- This is sent to the REST API as a query.
- REST API provides response, i.e., Maximum tweets related to mobility category, highest mobility rate and rate of COVID-19 confirmed cases.
- This response is then provided to Blockchain ecosystem through the Oracle smart contract.
- COVID-19 active cases helps Oracle to generates an event 'Lockdown', 'Restrictions' or 'No Lockdown'.
- The Oracle also provides additional information required by other accounts in the Blockchain ecosystem like 'Store Capacity' to the Essential services, 'Relief Fund' to non-essential services, maximum 'Twitter' and 'Mobility' details to heath officials.

The implementation is elaborated in Chapter 4.1.3 and the Gas spent on the Main Ethereum network is discussed in Chapter 4.2.2.2.

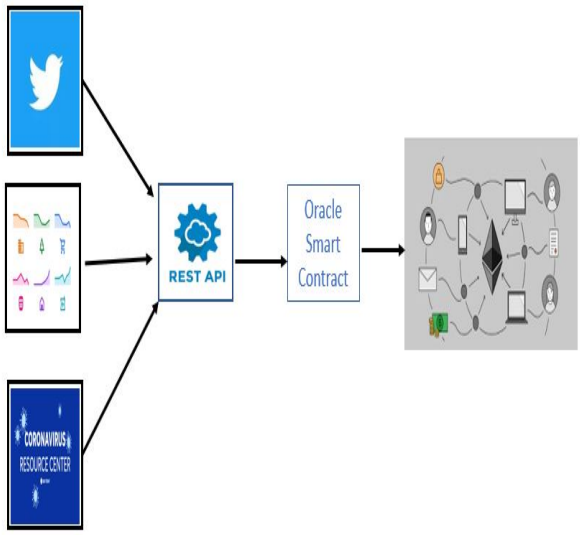

Figure 3.2: Architecture

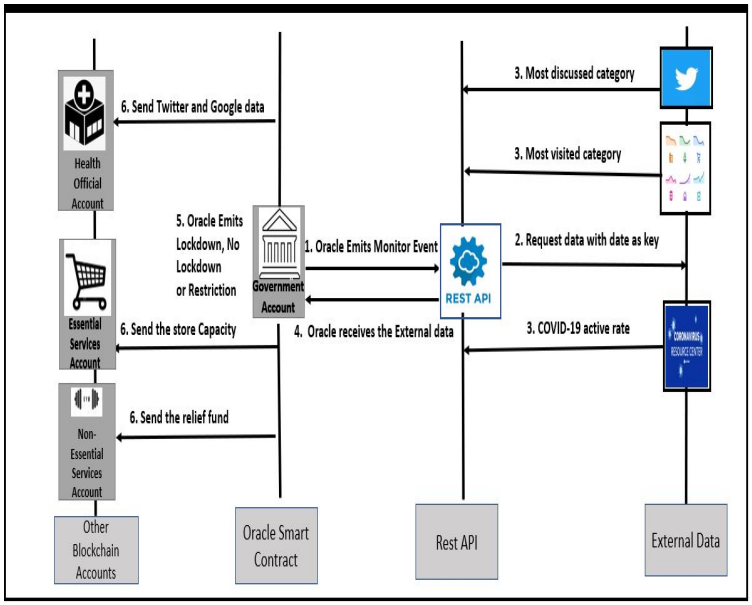

Figure 3.3: Sequence Diagram

Chapter 4

Implementation and Results

In this chapter, the implementation is discussed and the results are showcased.

4.1 Implementation

The implementation details for the COVID-19 blockchain - backed COVID-19 data in Chapter 4.1.1 followed by the implementation for decentralized applications described in Chapter 4.1.2 and 4.1.3.

4.1.1 COVID-19 blockchain-backed data

The ECDC and COVID Tracking Project which are blockchain backed data feed are used to find the countries with COVID-19 confirmed patients and in US states respectively for all the months from January 2020 to July 2020. This analysis of the data is implemented using Python and visualization is done in R programming. The Mipasa's ECDC confirmed cases dataset i.e. Output_ECDC_Cases.csv was down-

loaded. The data description of the dataset is given in Table 3.2. The aim of the analysis is to find the top 10 countries suffering from a greater number of COVID cases every month from beginning of January 2020 to July 2020. This aim can be achieved by firstly extracting the month from the ISO 8601 date format and defining the respective month. Then, finding the maximum confirmed cases each month. The countrycode2 is linked to "Unbounded Taxonomy Representation (UTR)" to avoid inconsistencies country codes in the dataset [48]. To attain the corresponding country name the country code has to be linked to the country UTR [48]. The blockchain-backed data is collected and the collapsible tree is developed using R package. The results are in Chapter 4.2.1.

4.1.2 COVID-19 early symptom prediction using Blockchain and machine learning

4.1.2.1 ML Model Phase

4.1.2.1.1 Data Exploration: The Israeli Ministry of Health released data of individuals who were tested for SARS-CoV-2 via Real-Time polymerase chain reaction (RT-PCR) using nasopharyngeal swab [47]. The dataset contains results of all the residents who were tested for COVID-19 nationwide. In addition to the test date and result there is a lot of information available, including clinical symptoms, gender and a binary indication as to whether the tested individual is aged 60 years or above. The dataset has over 1 million records from the period August 9th, 2020 through October 10th, 2020. The data is pre-processed by translating Hebrew to English for the exploratory data analysis and the description of the dataset is given in Table 4.1

and Table 4.2:

Sex (Male/Female)
Age $\geq$ 60 years (True/False)
Known contact with an individual confirmed to have COVID-19 (True/False)

Table 4.1: Patient Information

Fever (True/False)
Cough (True/False)
Sore throat (True/False)
Shortness of breath (True/False)
Headache (True/False)

Table 4.2: Symptoms

The exploratory data analysis is performed and the Figure 4.1 showcases the insights.

Figure 4.1(a) shows that in the dataset 80% of the cases reported were negative and fewer than 20% were positive and the rest were inconclusive. Figure 4.1(b) categorizes the COVID-19 patients' symptoms based on age. The analysis clearly shows that most of the people who reported COVID-19 symptoms were not older than 60 years. Figure 4.1(c) shows fever, cough and headache as the common symptom of all the COVID-19 symptoms reported in the majority of the cases. The data in Figure 4.1(d) shows that male and female patients, who reported COVID-19 symptoms, were observed in equal proportions.

The"Known contact with an individual confirmed to have COVID-19 (True/False)" in Table 4.1 is not considered since data provided was insufficient. The

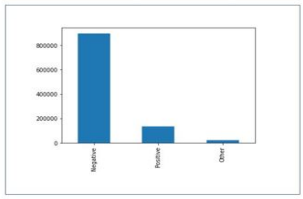

Fig. 2(a) COVID-19 Early Symptoms- "Positive Vs Negative Vs Others".

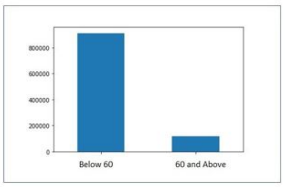

Fig. 2(b) COVID-19 Early Symptoms- Age.

Row Labels	Count of corona_result
Female	**532331**
Negative	458380
Other	9836
Positive	64115
Male	**512436**
Negative	434651
Other	9583
Positive	68202
(blank)	**3808**
Negative	3265
Other	131
Positive	412
Grand Total	**1048575**

Fig. 2(c) Gender based COVID-19 cases.

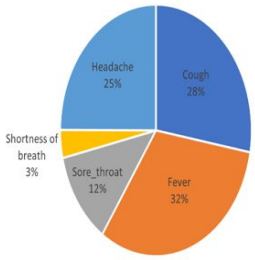

Fig. 2(d) Symptoms for COVID-19 cases.

Figure 4.1: Exploratory Data Analysis

understanding of the dataset is important for creating reliable ML model.

4.1.2.1.2 ML Model Selection: Predictions are made by the ML model deployed on the DApp, when a person enters the symptoms for preliminary testing of COVID-19. The ML model makes predictions on-the-fly and also gets retrained with data provided to it. Two incremental ML models considered are Perceptron model and SGDClassifier as the partial_fit() helps in retraining the model with new batch of data. The advantage of using such ML models is that it will keep a track on the evolving symptoms with different COVID-19 variants.

A. Perceptron Model: Perceptron is the most simple neural network model. This is a ML model, which is a binary or two-class classification. The hyperplane helps in learning for a decision boundary that separates two classes using a line in

the feature space. This model will be appropriate for the problem where the classes (COVID-19 positive or negative) can be separated well by a line or linear model or linearly separable function [49].

B. SGDClassifier Model: Stochastic Gradient Descent (SGD) [11] is an optimization algorithm that minimizes a cost function by finding the values of parameters or coefficients of functions. It is used for discriminative learning of linear classifiers under convex loss functions such as SVM and Logistic regression. The update to the coefficients is performed during each training instance, rather than at the end because of which it is used for incremental learning of COVID-19 symptoms. It is a classifier that implements a SGD learning routine supporting various loss functions and penalties for classification. Scikit-learn provides SGDClassifier module to implement SGD classification.

4.1.2.1.3 Evaluation For selecting the best ML model for DApp, different metrics [50] are considered as shown in Table 4.3 and 4.4.

Accuracy is a ratio of correctly predicted observations to the total observations. A model is best when it has the high accuracy. Accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Perceptron model is approximately 88% accurate where as SGDClassifier has 90% accuracy as evaluation results shown in Table 4.3 and 4.4.

We use the following metrics for evaluation of the results.

Precision is the ratio of predicted positive observations accurately to the total predicted positive observations.

Recall is the ratio of accurately predicted positive observations to the all obser-

vations in actual class.

F1-Score is the weighted average of Precision and Recall. In other words, $F1-Score = 2*(Recall*Precision)/(Recall+Precision)$. F1 is usually more useful than accuracy, especially if you have an uneven class distribution, i.e., if the cost of false positives and false negatives are very different. Perceptron model has 94% F1-score where as SGDClassifier has 95% F1-score.

	Precision	**Recall**	**F1-Score**	**Support**
Positive	0.88	1.00	0.94	268906
Negative	0.84	0.09	0.16	39802
Accuracy	-	-	0.88	308708
Macro Average	0.86	0.54	0.55	308708
Weighted Average	0.88	0.88	0.84	308708

Table 4.3: Evaluation Results For Perceptron Model

	Precision	**Recall**	**F1-Score**	**Support**
Positive	0.91	0.99	0.95	268906
Negative	0.81	0.33	0.47	39802
Accuracy	-	-	0.90	308708
Macro Average	0.86	0.66	0.71	308708
Weighted Average	0.90	0.90	0.88	308708

Table 4.4: Evaluation Results For SGDClassifier Model

Considering these metrics for ML models, Perceptron model showed 88% accuracy and F1-Score 94% where as SGDClassifier shows 90% accuracy and F1-score 95% as shown in Table 4.3 and 4.4.

The SGDClassifer outperforms Perceptron model except for the Recall metric

for positive COVID cases. Hence, SGDClassifier model is selected for the proposed architecture.

4.1.2.2 Data Transaction Phase

4.1.2.2.1 COVID-19 symptoms data transaction and predictions: The DApp shown in Figure 3.1, consists of a questionnaire for the patient to enter the early symptoms data **D** in accordance with Tables 4.1 and 4.2. This data **D** is provided and the ML model deployed on the client machine for the initial predictions. The prediction **P** will help the patient with instant COVID-19 results. Next, the DApp provides the list of the nearest registered labs along with addresses in the network where the user can get a swab test. The early symptoms and ML prediction result is sent as a transaction **T(U, D, P)** on the Blockchain to the lab selected by the user address **U**.

4.1.2.2.2 Blockchain: The Blockchain used here is the Ethereum test network called Ganache. The smart contracts developed, using Solidity, save and send the patient's symptoms data as a transaction to the lab account **L**. After the lab performs the swab-tests and when the results for the patient are ready, the prediction **P** made by the ML algorithm are checked by the lab technician. If it is wrong, it is negated and the test results **Q** are updated otherwise the result is retained as **Q**. This correction is sent as a transaction **T(L, Q)** on the Blockchain and the result **Q** is also updated to the patient.

Additionally, the smart contracts are called by the centralized storage account **C**, then the aggregated data **F(U, L, D, P, Q)** consisting of the User and Lab addresses,

patients symptoms data, ML model predictions and the verified ML results by the hospital respectively, is appended to the centralized dataset as a record. This record **F** helps in contributing reliable dataset for good ML model predictions in the future.

4.1.2.3 Centralized Storage Phase

The ML model in this architecture is trained off-chain in the centralized storage. Larger the amount of reliable data collected, the better will be the retraining of the model. The ML model is updated on the client side machines on day-to-day basis.

Chapter 4.2.2.1 has the details related to the gas spent by the smart contracts.

4.1.3 Blockchain Oracles for COVID-19 Social Signals

This work was implemented using Flask, Web3Py and Ganache. The steps in methodology are explained in detail.

4.1.3.1 Data Collection and Processing:

4.1.3.1.1 Twitter data The Twitter historical data from 08-Mar-2020 to 20-Mar-2020 from Toronto, Canada was collected. This period is crucial because the first lockdown was declared on 14-Mar-2020 in Canada [9]. Totally, 836,759 COVID-19 related tweets were collected and processed. The data was scraped using the Python library called snscrape. Snscrape [51] is a social networking services (SNS), which scrapes social network like Twitter and provided the data related to users, user profiles, hashtags, searches, threads, and list posts. Table 4.5 shows the fields related to the tweets collected for the hashtags with the aim to derive insights about

people's concerns regarding the categories like 'Retail', 'Parks', 'Supermarket', 'Public Transport', 'Residence' and 'Work Place'. The user details are not required for the application development so the privacy of the users is out of scope. The data is processed by converting the 'datetime' field to only 'date' and also tweets by removing unnecessary characters like stopwords, puctuations etc. Further, Lemmatization [52] is performed, i.e, a process of converting a word into its root word. Later, it helped in creating a word cloud. Figure 4.2 shows the concerns of the people during the period. According to the word cloud, the main concern was 'panic buying'. These cleaned tweets provided the insights to predict people mobility details which may be related to the topics discussed. Figure 4.3 showcases that the total number of tweets per day related to the above mentioned categories. It is very clear that the tweets related to 'Residental' category was the most discussed followed by 'Retail' and 'Supermarkets'.

Field	**Description**
Datetime	Date and Time when the tweet was created
Tweet Id	Tweet ID number
Text	Tweet text
Username	Username who created the tweet

Table 4.5: Twitter Data Collection

4.1.3.1.2 Google Mobility Data COVID-19 Community mobility data by Google as shown in Table 4.6 was collected from 08-Mar-2020 to 20-Mar-2020 for Toronto sub-region only [53]. The data was cleaned and processed. The mobility data categories are 'Parks', 'Supermarket', 'Public Transport', 'Residence', 'Workplace', 'Retail'. According to Figure 4.4, overall movement of people declined over the period

Figure 4.2: Twitter Word Cloud

of time. People preferred staying at home as we can see that the 'residential' mobility was maximum by the end of the period, followed by the most visited places were 'supermarkets' and 'parks'.

Field	Description
date	Date of creation
country_region	Country name
sub_region_1	Province name
sub_region_2	City name
retail	Retail mobility trends
grocery	Grocery mobility trends
park	Park mobility trends
transit	Transit mobility trends
workplaces	Workplace mobility trends
residential	Residential mobility trends

Table 4.6: Community Mobility Data by Google

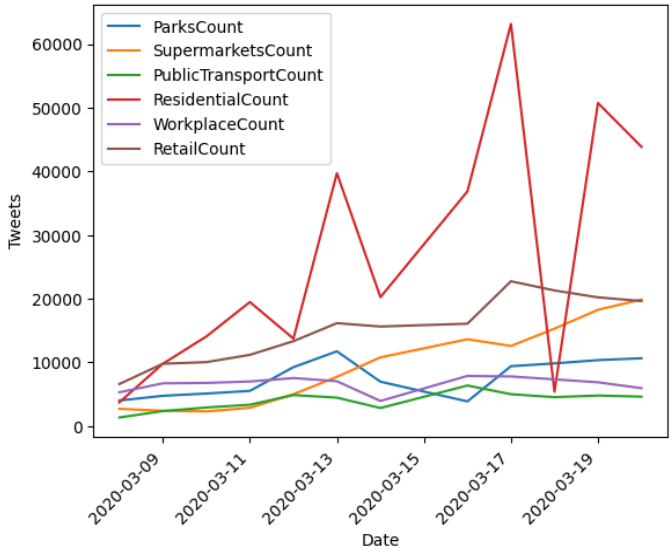

Figure 4.3: Tweets count

4.1.3.1.3 COVID-19 Cases Ontario COVID-19 cases as shown in Table 4.7 were collected from COVID-19 daily epidemiology update [54] from 8-Mar-2020 to 20-Mar-2020. The Figure 4.5 shows that the number of active cases increased over this period of the time.

Later, these results are stored in JSON files and depending on the query from Oracle smart contract, the data is retrieved and provided to Oracle seamlessly.

4.1.3.1.4 REST API REST API ingests the JSON files and when queried by Oracle smart contract it responds with the maximum category for that date, fetch the social signals required.

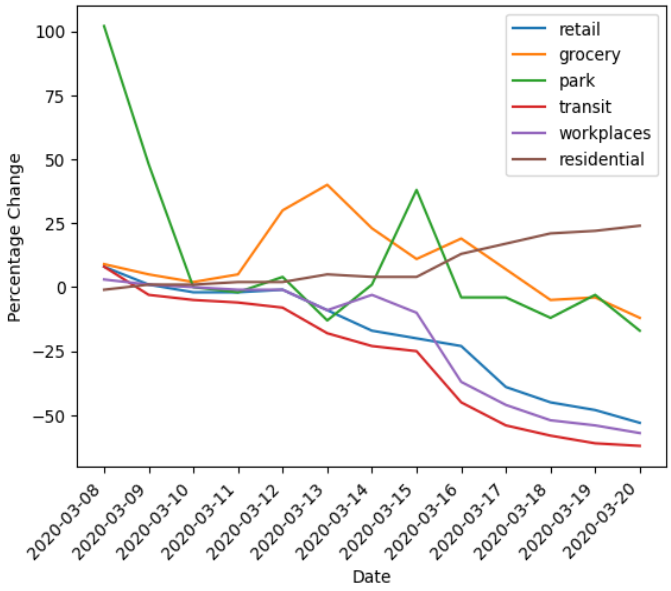

Figure 4.4: Google Mobility count -Toronto

Field	**Description**
pruid	Province ID
prname	Province Name
prnameFR	Province name in French
date	Date
numconf	Number of Confirmed Cases
numprob	Number of Probable cases
numdeaths	Number of Deaths
numtotal	Total number of Cases
numtested	Total Number of tested Cases
rateactive	Rate of Active Cases
raterecovered	Rate of Recovered Cases

Table 4.7: Canada COVID-19 Cases Data

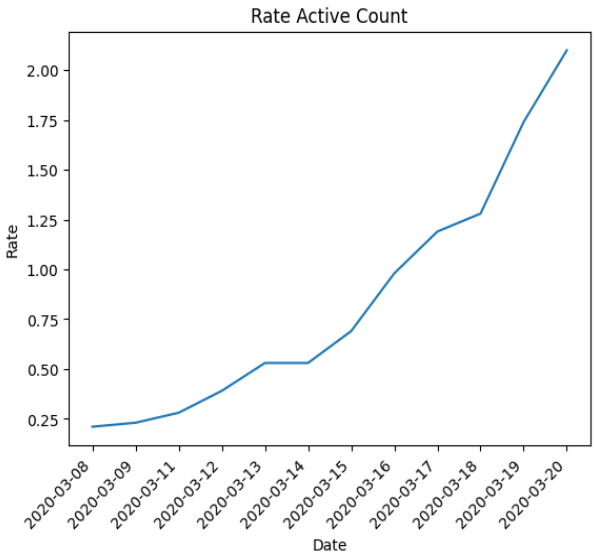

Figure 4.5: Active Rate COVID-19 Cases - Toronto

4.1.3.2 Blockchain Oracle, Defi and REST API

The Figure 3.3 shows the process where the big data and the blockchain work together. This process is explained below:

4.1.3.2.1 Ganache For the study, a private blockchain Ganache [55] is used with four accounts - Government (G), Non-Essential Business (N), Essential Business (E) and Contact Tracing (H) are considered.

4.1.3.2.2 Blockchain Oracle Smart Contract Now the goal is to provide the Blockchain with data from external sources. Smart contracts cannot interact with the external data directly. As they are codes that are automatically executed when a partial or a full agreement is met and stored on the distributed ledger [56]. This limitation can be overcome by using Blockchain Oracles. They have the ability to

access the external data [57]. Openzepplin - "provides security products to build, automate, and operate decentralized applications" [58] has a module called Ownable.sol [59]. This helps the smart contract to interact with the outside world. Once a smart contract inherits this module, it becomes an Oracle smart contract. Here, account G has the Oracle smart contract, so it interacts with the external data. The Oracle smart contract provides the following functions:

* G generates 'MONITOR' event (E_m) which is a request to REST API for the external data.

* The external data, consisting of max number of tweets (T), community maximum mobility (M) and COVID-19 cases (C) the active COVID-19 cases is provided to G.

* The Oracle smart contract will get a response (R_m) with the T, M and C. Depending on C, the decision for lockdown, restrictions or no lockdown is taken by G. It also provides store capacity (α) and relief funds (β) inputs.

* This will trigger the smart contracts related to the decision taken to provide the data T and M to H, α to E and β to N.

The gas spent by the smart contracts for the transaction of the data is explained in Chapter 4.2.2.2.

4.2 Results

The results for the COVID-19 blockchain - backed COVID-19 data in Chapter 4.2.1 followed results for the decentralized applications, i.e., the Gas spent on Ethereum network for performing the transactions are described in Chapter 4.2.2.1 and 4.2.2.2 respectively.

4.2.1 The analysis of COVID-19 blockchain-backed data

In Chapter 3.1 the analysis for the Blockchain - backed data was discussed. Here, results for the Knowledge tree created using Collapsible Tree [60] with R package the Figure 4.6 shows the 10 countries with maximum number of COVID-19 confirmed cases for months from January 2020 to July 2020. We can easily comprehend that Japan was having the maximum number of confirmed cases in January and February, but the government took important measures like effective contact tracing for early detection and early response to clusters [61] for containment of COVID-19 cases along with required measures like “social distance”, “wearing mask” and “hand hygiene for example hand washing”. We can see the efforts paid off as the number of cases decreased and the first wave ended early in Japan. However, USA persistently had maximum number of cases. Hence, we can understand that February marked the end of the first wave in Japan but unfortunately, it was not the same with USA. Lastly, the results from the above analysis provides 10 most affected US states with maximum number of COVID-19 confirmed cases each month shown in Figure 4.7. The Figure 4.7 shows number of cases in a particular month for Japan and USA. Figure 4.7 shows the top countries for the month of February with maximum number of COVID-19

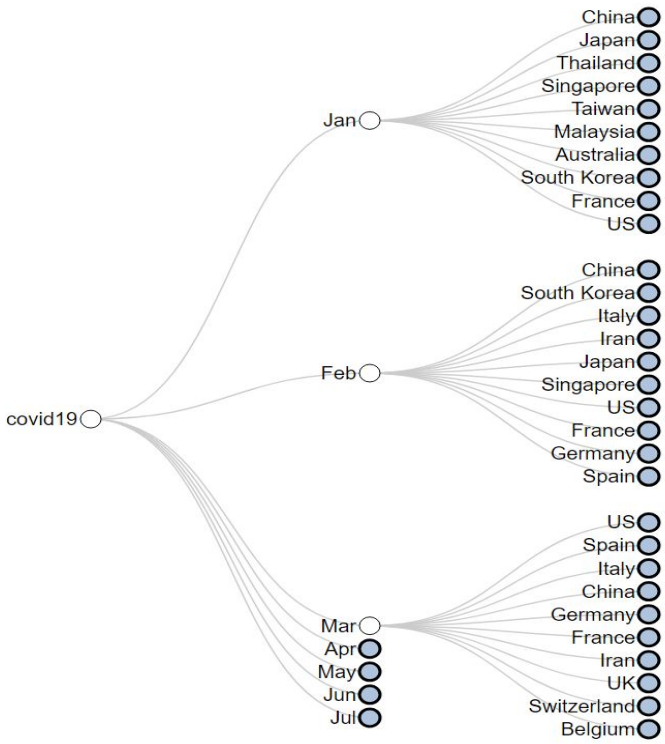

Figure 4.6: 10 most affected countries with COVID-19 patients January-March

confirmed case numbers on the edges. Now, we would like to know within USA and Japan which of the 10 states or provinces is having maximum COVID-19 patients. Figure 4.8 shows the hierarchy tree structure for the cases in Japanese provinces. It became possible to create this tree hierarchy by joining the COVID-19 Tracking Data for USA in Figure 4.9 and COVID-19 in Japan datasets in Figure 4.8 respectively. We can see that COVID-19 Tracking data did not have January COVID-19 data whereas ECDC data shows six COVID-19 cases. This hierarchical tree structure in Figure 4.8 shows the provinces are suffering the most and in turn the country is suffering of the pandemic. Moreover, there are more provinces suffering in Japan whereas US had only Washington with COVID-19 cases in Figure 4.9.

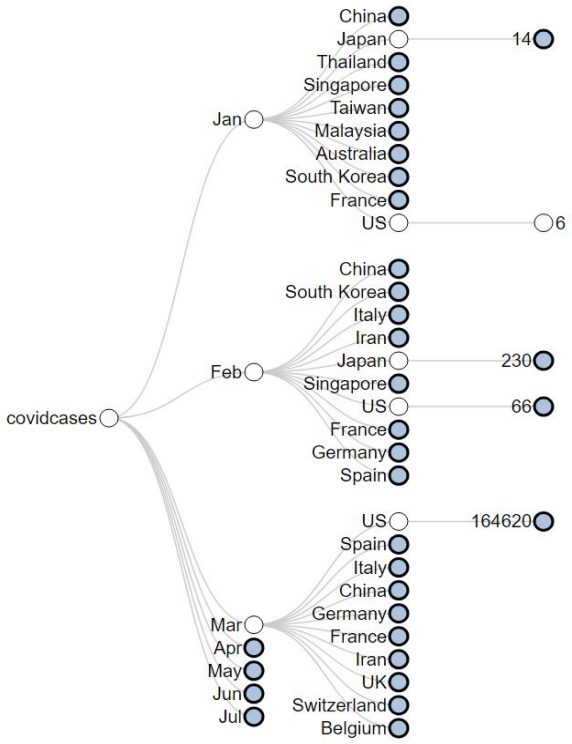

Figure 4.7: USA cases and Japan COVID-19 cases

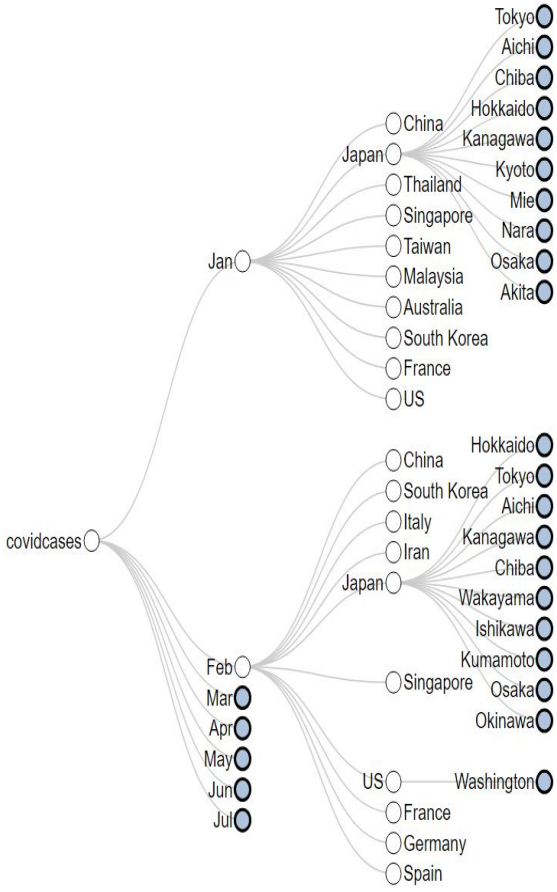

Figure 4.8: COVID-19 affected Provinces in Japan and States in USA for month of January and February

We can also analyse the data corresponding to a particular state in USA. Figure 4.9 and Figure 4.8 shows the monthly positive cases in New York and monthly confirmed cases in Tokyo respectively. We can understand that the government has taken measure to reduce the cases in Tokyo since January 2020 but on the other hand the containment measures of COVID-19 in New York showed effect only after April 2020.

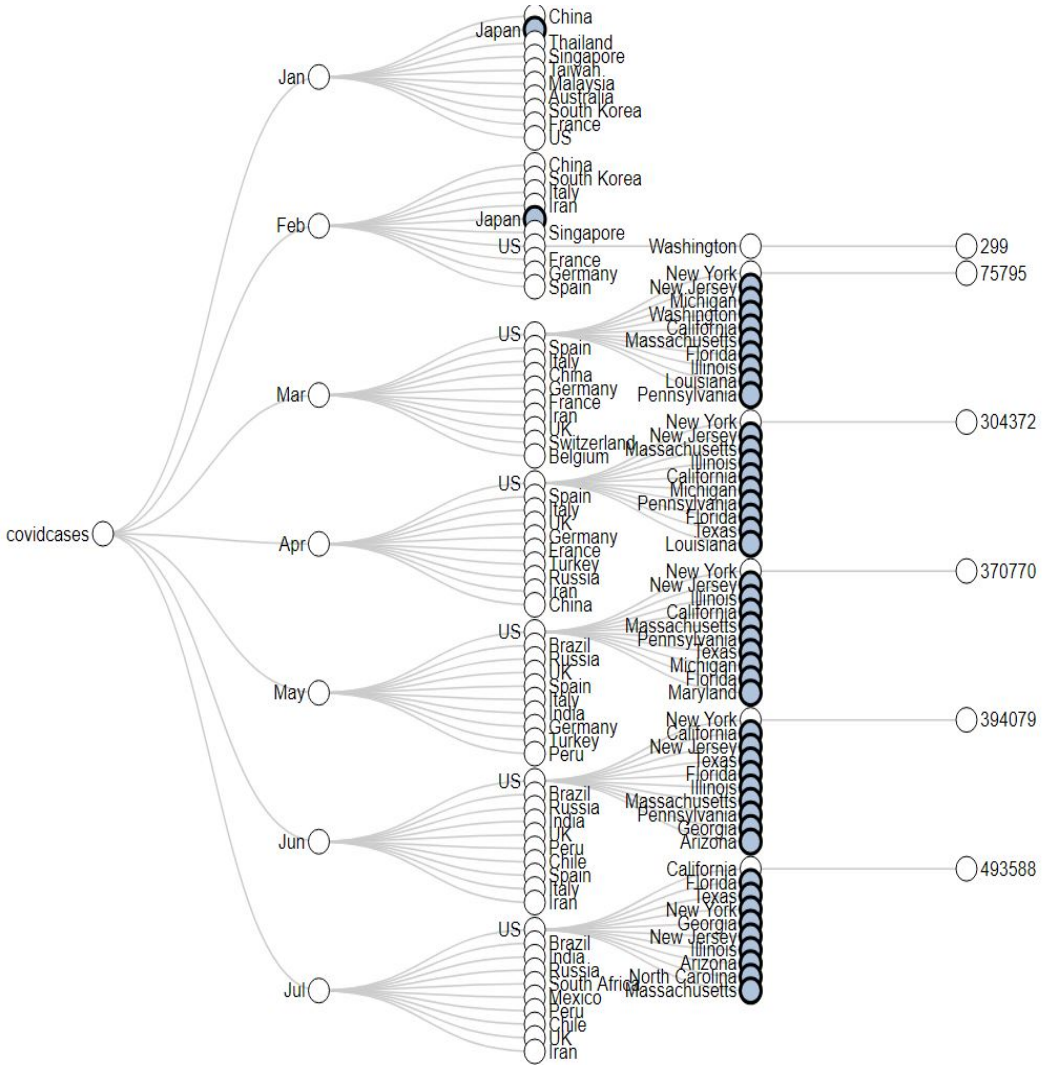

Figure 4.9: COVID-19 epicenter drift in USA

4.2.2 Ethereum Gas Cost

Executing a smart contract or conducting a transaction on the Ethereum Blockchain Network requires a price to be paid called Gas. It is commonly referred as **gwei**. The gas is the price paid to allocate resources for the Ethereum virtual machine (EVM) which helps to self-execute the smart contracts in a decentralized applications in a secured but decentralized fashion. The gas price varies as it is determined by supply and demand between the network's miners [62].

In Chapter 4.2.2.1 and Chapter 4.2.2.1 provides the details related to the Ethereum Gas Cost for both the Decentralized Applications.

4.2.2.1 COVID-19 early symptom prediction using Blockchain and machine learning

In Chapter 3.2 the architecture and the implementation for COVID-19 early symptom prediction using Blockchain and machine learning was discussed. Table 4.8 shows details related to the Ethereum gas spent when this proposed Decentralized Application is deployed on the Main network. Here, 1 Gwei = 0.00000170 USD [63]. Every time, when data is added either by a lab or by a patient, there is a gas cost for a transaction. For the smart contract deployment on the Main Ethereum network, the gas spent is 2019063.

4.2.2.2 Blockchain Oracles for COVID-19 Social Signals

In the Chapter 3.3 the architecture and the implementation for Blockchain Oracles for COVID-19 Social Signals was discussed. Here, 1 Gwei = 0.0.00000328 USD [63].

Description	Gas Spent	USD
Smart contract deployment	2019063	3.4255 USD
User Symptoms data	217378	0.3688 USD
Hospital Verification	70671	0.1199 USD

Table 4.8: Ethereum gas spent for Smart Contracts in COVID-19 early symptom prediction using Blockchain and machine learning

When there is data transfer in the Blockchain the gas is spent and Table 4.9 provides the details of Gas spent when Decentralized Application is deployed on the Main Network. For the smart contract deployment on the main Ethereum network, the gas spent is 1556763.

Description	Gas Spent	USD
Smart contract deployment	1556763	5.1061 USD
Set Essential data	68975	0.226271 USD
Set Non-Essential	68985	02262708 USD
set Mobility Twitter	129843	0.42589 USD

Table 4.9: Ethereum gas spent for Smart Contracts in Blockchain Oracles for COVID-19 Social Signals

4.2.3 Results Summary

The study of the data formats of the centralized and blockchain data as discussed in Chapter 3.1 and Chapter 4.1.1 helped in understanding the advantages of using Blockchain technology for data storage. This insight is used for the creation of the COVID-19 decentralized application (DApp) leveraging big data analysis and blockchain smart contracts for recording all the transactions on a Distributed ledger technology (DLT) as seen in Chapter 3 and Chapter 4.1.

The first DApp helps in predicting the COVID-19 positive or negative from the data of the symptoms provided by the patients. The predictions are sent to an entity in the Blockchain ecosystem (a testing lab) where the predictions are compared with the actual results. This helps to improve the accuracy of the ML model. The architecture is discussed in Chapter 3.2 and development steps are presented in Chapter 4.1.2.

The second DApp is developed according to the architecture described in Chapter 3.3 and implementation details are presented in Chapter 4.1.3. Big data analysis which helps in aggregating the data from three platforms - Twitter, Canada COVID-19 cases, and Google Mobility data. The REST API is stored with the processed data in JSON format. The REST API responds with the required data when there is a request from the Blockchain ecosystem through Oracle smart contract. Once the data is reviewed by the Government officials and one of the decisions is taken - Lockdown, No lockdown, or Restrictions. The smart contract initiates the transactions, i.e., it will deposit relief funds in the closed non-essential services account to support them during this period. Similarly, the essential services account should limit the number of people entering the business space (store or office) to slow down the spread of the virus. The smart contract will provide the 'Store capacity' details to all the essential services. The movements of the people can be predicted from the social signals - Twitter and Google Mobility data, this will help health officials manage services.

Chapter 5

Conclusion and Contribution

In this chapter, the conclusion with a glimpse of the future work is provided along with our contribution.

5.1 Conclusion

This research presented analysis on the COVID-19 blockchain data. It also showcases the importance of normalized data collected from different platforms - provided by Mipasa. Data preprocessing was minimum for Mipasa data when compared to the data provided by the centralised system. Blockchain data is analysed and insights related to first wave of COVID-19 in USA and Japan were discussed by visualizing the data as a network (collapsible tree).

The first decentralized application (DApp) created for symptom tracking will help in the direct transfer of data to/from patient and lab. The predictions of machine learning used in DApp are constantly checked by the laboratory with the test results.

This reliable data on Blockchain will in turn help retraining the ML model and provide more accurate predictions. Moreover, this reliable data can also be used instantly for the purpose of research and development by the government for prompt decision-making.

The second DApp used the Blockchain oracle smart contract for collecting social signals worked as an interface between the external data and the blockchain. The advantage of this system is that the events will be notified to the participants instantly and the details will be recorded on the distributed ledger. This study can be used in other scenarios caused by natural disasters, financial catastrophic events etc.

This work can be further extended by understanding the timeline of the government measure which helped in containment of COVID-19.

The COVID-19 results can be stored in the InterPlanetary File System (IPFS), which is a peer-to-peer network for storing and sharing data in a distributed file system. This will allow extension of the DApp to utilize some Deep Learning algorithms to study numerous other applications, like COVID-19 predictions by Chest X-Ray Images [64].

5.2 Contributions

1. Sarada Tadepalli and Ruppa K. Thulasiram (2021), A Chapter in the Springer Nature book titled *Assessing COVID-19 and Other Pandemics and Epidemics using Computational Modelling and Data Analysis*; article title: *Analysis of Blockchain Backed Covid19 Data*, (Ed. Subhendu K. Pani, Sujata Dash, Wellington PD Santos, Syed Ahmad Chan Bukhari, Francesco Flammini), pp.285-299,

2021.

2. ICBA Paper: Sarada K. Tadepalli, Ruppa K. Thulasiram, *COVID-19 Early Symptom Prediction Using Blockchain and Machine Learning*, 3rd International Congress on Blockchain and Applications, Salamanca, Spain, Springer Lecture Notes in Networks and Systems, Vol.320, pp.243-251, Oct. 2021.

Bibliography

[1] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, Z. Cheng, T. Yu, J. Xia, Y. Wei, W. Wu, X. Xie, W. Yin, H. Li, M. Liu, and B. Cao, "Clinical features of patients infected with 2019 novel coronavirus in wuhan, china," *The Lancet*, vol. 395, 01 2020.

[2] M. Khafaie and F. Rahim, "Article history: Cross-country comparison of case fatality rates of covid-19/sars-cov-2," *Osong public health and research perspectives*, vol. 11, pp. 74–80, 04 2020.

[3] M. A. Hamade and A. Holdings, "Covid-19: How to fight disease outbreaks with data." [Online]. Available: https://www.weforum.org/agenda/2020/03/covid-19-how-to-fight-disease-outbreaks-with-data/

[4] M. R. Hannah Ritchie, Esteban Ortiz-Ospina and J. Hasell, "Covid-19 deaths and cases: how do sources compare?" (Accessed: 2020-06-12). [Online]. Available: https://ourworldindata.org/covid-sources-comparison

[5] G. Singh and J. Levi. Mipasa project and ibm blockchain team on open data platform to support covid-19 response. (Accessed:

2020-06-11). [Online]. Available: ibm.com/blogs/blockchain/2020/03/mipasa-project-and-ibm-blockchain-team-on-open-data-platform-to-support-covid-19-response/

[6] V. Chamola, V. Hassija, V. Gupta, and M. Guizani, "A comprehensive review of the covid-19 pandemic and the role of iot, drones, ai, blockchain and 5g in managing its impact," *IEEE Access*, vol. PP, 05 2020.

[7] C.Menni, V. AM, F. MB, S. CH, N. LH, D. DA, V. T. Ganesh S, C. MJ, E.-S. M. JS, V. A, H. P, B. RCE, M. M, F. M, W. J, O. S, C. AT, S. CJ, and S. TD, "Real-time tracking of self-reported symptoms to predict potential covid-19," *Nat Med*, pp. 1037–1040, 07 2020.

[8] Z. Yazeed, D.-R. Shira, and N. Shomron, "Machine learning-based prediction of covid-19 diagnosis based on symptoms," *npj Digital Medicine*, vol. 4, 01 2021.

[9] T. Hale, N. Angrist, R. Goldszmidt, B. Kira, A. Petherick, T. Phillips, S. Webster, E. Cameron-Blake, L. Hallas, S. Majumdar, and H. Tatlow, "A global panel database of pandemic policies (oxford covid-19 government response tracker)," *Nature Human Behaviour*, vol. 5, 04 2021.

[10] Tracking economic relief plans around the world during the coronavirus outbreak. Accessed on(2021-10-22)). [Online]. Available: https://taxfoundation.org/coronavirus-country-by-country-responses/

[11] S. Ruder, "An overview of gradient descent optimization algorithms," *CoRR*, vol. abs/1609.04747, 2016. [Online]. Available: http://arxiv.org/abs/1609.04747

[12] K. Christidis and M. Devetsikiotis, "Blockchains and smart contracts for the internet of things," *IEEE Access*, vol. 4, pp. 2292–2303, 2016.

[13] M. Xu, X. Chen, and G. Kou, "A systematic review of blockchain," *Financial Innovation*, vol. 5, pp. 1–14, 12 2019.

[14] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," *Cryptography Mailing list at https://metzdowd.com*, 03, 2009.

[15] F. Casino, T. Dasaklis, and C. Patsakis, "A systematic literature review of blockchain-based applications: Current status, classification and open issues," *Telematics Informatics*, vol. 36, pp. 55–81, 2019.

[16] V. Buterin, "Ethereum white paper: A next generation smart contract & decentralized application platform," 2013. [Online]. Available: https://github.com/ethereum/wiki/wiki/White-Paper

[17] P. Zheng, Z. Zheng, J. Wu, and H.-N. Dai, "Xblock-eth: Extracting and exploring blockchain data from ethereum," *IEEE Open Journal of the Computer Society*, vol. 1, pp. 95–106, 2020.

[18] D. Ron and A. Shamir, "Quantitative analysis of the full bitcoin transaction graph," in *Financial Cryptography and Data Security*, A.-R. Sadeghi, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 6–24.

[19] E. Zhou, H. Sun, B. Pi, J. Sun, K. Yamashita, and Y. Nomura, "Ledgerdata refiner: A powerful ledger data query platform for hyperledger fabric," in *2019*

Sixth International Conference on Internet of Things: Systems, Management and Security (IOTSMS), 2019, pp. 433–440.

[20] M. Spagnuolo, F. Maggi, and S. Zanero, "Bitiodine: Extracting intelligence from the bitcoin network," vol. 8437, 03 2014, pp. 457–468.

[21] F. Oggier, S. Phetsouvanh, and A. Datta, "Biva: Bitcoin network visualization analysis," in *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2018, pp. 1469–1474.

[22] S. Thomsen. Online cheating site ashley madison hacked. (Accessed: 21 July 2015). [Online]. Available: https://www.businessinsider.com/cheating-affair-website-ashley-madison-hacked-user-data-leaked-2015-7

[23] G. D. Battista, V. D. Donato, M. Patrignani, M. Pizzonia, V. Roselli, and R. Tamassia, "Bitconeview: visualization of flows in the bitcoin transaction graph," in *2015 IEEE Symposium on Visualization for Cyber Security (VizSec)*, 2015, pp. 1–8.

[24] T. Chen, Y. Zhu, Z. Li, J. Chen, X. Li, X. Luo, X. Lin, and X. Zhange, "Understanding ethereum via graph analysis," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, 2018, pp. 1484–1492.

[25] D. Dillenberger, P. Novotny, Q. Zhang, P. Jayachandran, H. Gupta, S. Mehta, S. Hans, S. Chakraborty, M. Walli, J. Thomas, R. Vaculín, K. Sarpatwar, and D. Verma, "Blockchain analytics and artificial intelligence," *IBM Journal of Research and Development*, vol. PP, pp. 1–1, 02 2019.

[26] H.-Y. Paik, X. Xu, H. M. N. D. Bandara, S. U. Lee, and S. K. Lo, "Analysis of data management in blockchain-based systems: From architecture to governance," *IEEE Access*, vol. 7, pp. 186 091–186 107, 2019.

[27] J. D. Harris and B. Waggoner, "Decentralized & collaborative AI on blockchain," *CoRR*, vol. abs/1907.07247, 2019. [Online]. Available: http://arxiv.org/abs/1907.07247

[28] H. Zhu and Z. Z. Zhou, "Analysis and outlook of applications of blockchain technology to equity crowdfunding in China," *Financial Innovation*, vol. 2, no. 1, pp. 1–11, December 2016.

[29] Wikipedia contributors, "Covid-crypto relief fund — Wikipedia, the free encyclopedia," 2021, [Online; accessed 23-October-2021]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=COVID-Crypto_Relief_Fund&oldid=1051394786

[30] S. Dey, U. Siddiqi, and A. Howlader, "Analyzing the epidemiological outbreak of covid-19: A visual exploratory data analysis (eda) approach," *Journal of Medical Virology*, vol. 92, 03 2020.

[31] F. A. Binti Hamzah, C. Hau, H. Nazri, D. Ligot, G. Lee, M. Shaib, U. Zaidon, A. Abdullah, M. Chung, C. Ong, and P. Chew, "Coronatracker: World-wide covid-19 outbreak data analysis and prediction," 03 2020.

[32] M. M. Ahamad, S. Aktar, M. Rashed-Al-Mahfuz, M. S. Uddin, P. Lio', H. Xu, M. A. Summers, J. Quinn, and M. Moni, "A machine learning model to iden-

tify early stage symptoms of sars-cov-2 infected patients," *Expert Systems with Applications*, vol. 160, pp. 113 661 – 113 661, 2020.

[33] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *CoRR*, vol. abs/1603.02754, 2016. [Online]. Available: http://arxiv.org/abs/1603.02754

[34] J. Laguarta, F. Hueto, and B. Subirana, "Covid-19 artificial intelligence diagnosis using only cough recordings," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, pp. 275–281, 2020.

[35] R. K. Gupta, A. Vishwanath, and Y. Yang, "COVID-19 twitter dataset with latent topics, sentiments and emotions attributes," *CoRR*, vol. abs/2007.06954, 2020. [Online]. Available: https://arxiv.org/abs/2007.06954

[36] Q.-V. Pham, D. C. Nguyen, T. Huynh-The, W.-J. Hwang, and P. N. Pathirana, "Artificial intelligence (ai) and big data for coronavirus (covid-19) pandemic: A survey on the state-of-the-arts," *IEEE Access*, vol. 8, pp. 130 820–130 839, 2020.

[37] J. Sheng, J. Amankwah-Amoah, Z. Khan, and X. Wang, "Covid-19 pandemic in the new era of big data analytics: Methodological innovations and future research directions," *British Journal of Management*, vol. 32, no. 4, pp. 1164–1183, Oct. 2021.

[38] G. A. Wellenius, S. Vispute, V. Espinosa, A. Fabrikant, T. C. Tsai, J. Hennessy, A. Dai, B. Williams, K. Gadepalli, A. Boulanger, and et al., "Impacts of social distancing policies on mobility and covid-19 case growth in the

us," *Nature Communications*, vol. 12, no. 1, May 2021. [Online]. Available: http://dx.doi.org/10.1038/s41467-021-23404-5

[39] M. Milano and M. Cannataro, "Statistical and network-based analysis of italian covid-19 data: Communities detection and temporal evolution," *International Journal of Environmental Research and Public Health*, vol. 17, no. 12, 2020. [Online]. Available: https://www.mdpi.com/1660-4601/17/12/4182

[40] J. Samuel, G. G. M. N. Ali, M. M. Rahman, E. Esawi, and Y. Samuel, "Covid-19 public sentiment insights and machine learning for tweets classification," *Information*, vol. 11, no. 6, p. 314, Jun 2020. [Online]. Available: http://dx.doi.org/10.3390/info11060314

[41] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, and Z. Shah, "Top concerns of tweeters during the covid-19 pandemic: Infoveillance study," *J Med Internet Res*, vol. 22, no. 4, p. e19016, Apr 2020. [Online]. Available: http://www.jmir.org/2020/4/e19016/

[42] Rangone and Adalberto, "Managing charity 4.0 with blockchain: a case study at the time of covid-19," *International Review on Public and Nonprofit Marketing*, vol. 1, pp. 1–31, 03, 2021.

[43] The covid19 tracking project. (Accessed: 2020-07-31). [Online]. Available: https://covidtracking.com/

[44] Mipasa-because i care. (Accessed: 2020-07-31). [Online]. Available: https://app.mipasa.org/

[45] "European centre for disease prevention and control," (Accessed: 2020-07-31). [Online]. Available: https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide

[46] How japan has responded to covid-19 pandemic. (Accessed: 2020-07-31). [Online]. Available: https://valdaiclub.com/a/highlights/how-japan-has-responded-to-covid-19-pandemic/

[47] Covid-19-isreal government data. (Accessed: 2021-03-09). [Online]. Available: https://data.gov.il/dataset/covid-19

[48] Mipasa-finding, viewing using datasets. (Accessed: 2020-07-31). [Online]. Available: https://mipasa.org/documentation/finding-viewing-using-datasets/

[49] Perceptron model. (Accessed: 2021-03-23). [Online]. Available: https://machinelearningmastery.com/perceptron-algorithm-for-classification-in-python

[50] L. Yangguang, Z. Yangming, W. Shiting, and T. Chaogang, "A strategy on selecting performance metrics for classifier evaluation," *International Journal of Mobile Computing and Multimedia Communications*, vol. 6, pp. 20–35, 10 2014.

[51] JustAnotherArchivist, "snscrape: A social networking service scraper in python," 2021, [Online; accessed 23-October-2021]. [Online]. Available: https://github.com/JustAnotherArchivist/snscrape

[52] Wikipedia contributors, "Lemmatisation — Wikipedia, the free encyclopedia," 2021, [Online; accessed 25-October-2021]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Lemmatisation&oldid=1022854248

[53] "Covid-19 community mobility data," (Accessed: 2021-08-31). [Online]. Available: https://www.google.com/covid19/mobility/

[54] "Covid-19 daily epidemiology update," (Accessed: 2021-08-31). [Online]. Available: https://health-infobase.canada.ca/covid-19/epidemiological-summary-covid-19-cases.html

[55] "Ganache," (Accessed: 2021-08-31). [Online]. Available: https://www.trufflesuite.com/docs/ganache/overview

[56] M. Alharby, A. Aldweesh, and A. v. Moorsel, "Blockchain-based smart contracts: A systematic mapping study of academic research (2018)," in *2018 International Conference on Cloud Computing, Big Data and Blockchain (ICCBB)*, 2018, pp. 1–6.

[57] A. Beniiche, "A study of blockchain oracles," *CoRR*, vol. abs/2004.07140, 2020. [Online]. Available: https://arxiv.org/abs/2004.07140

[58] "Openzeppelin docs," (Accessed: 2021-08-31). [Online]. Available: https://docs.openzeppelin.com/contracts/2.x/access-control

[59] J. Chittoda, *Mastering Blockchain Programming with Solidity*. Packt Publishing, 2019. [Online]. Available: https://books.google.ca/books?id=WB84yAEACAAJ

[60] Collapsible tree. (Accessed: 2020-08-31). [Online]. Available: https://github.com/AdeelK93/collapsibleTree

[61] Covid-19 dataset in japan. (Accessed: 2020-08-31). [Online]. Available: https://www.kaggle.com/lisphilar/covid19-dataset-in-japan

[62] J. Frankenfield, "Gas (ethereum)," (Accessed: 2020-11-30). [Online]. Available: hhttps://www.investopedia.com/terms/g/gas-ethereum.asp

[63] "Gwei to usd convertor," (Accessed: 2021-03-23). [Online]. Available: https://www.curvert.com/en/eth-usd/

[64] S. Tabik, A. Gómez-Ríos, J. L. Martín-Rodríguez, I. Sevillano-García, M. Rey-Area, D. Charte, E. Guirado, J. L. Suárez, J. Luengo, M. A. Valero-González, P. García-Villanova, E. Olmedo-Sánchez, and F. Herrera, "Covidgr dataset and covid-sdnet methodology for predicting covid-19 based on chest x-ray images," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 12, pp. 3595–3605, 2020.