Exploring relationship between COVID-19 cases and eating habits using data of London boroughs

Abdulhadi Algbear
College of Computer Science and Engineering
University of Jeddah
Jeddah, Saudi Arabia
Abdulhadi.IT@hotmail.com

Mohammed Ali Alqarni
College of Computer Science and Engineering
University of Jeddah
Jeddah, Saudi Arabia
alqarni@uj.edu.sa

Muhammad Usman Ilyas
College of Computer Science and Engineering
University of Jeddah
Jeddah, Saudi Arabia
milyas@uj.edu.sa

Muhammad Murtaza Khan
College of Computer Science and Engineering
University of Jeddah
Jeddah, Saudi Arabia
mkhan@uj.edu.sa

***Abstract*—COVID-19 has affected everyone in the world in one way or another. At the time of this writing, there are approximately 110.9 million reported cases with approximately 2.4 million deaths across the world this makes the ratio of deaths to total infections a little over 2%. To better understand the reasons for COVID-19 related infections and deaths, efforts are underway to uncover relationships between them and existing health conditions. Some studies have focused on causes of infection and use of preventive equipment for protection, while others have focused on identifying relationships between deaths and existing diabetes, heart condition or hyper-tension. Research has established that pre-existing health conditions can be associated to eating habits of people. Therefore, we have tried to determine if there is any relationship between eating habits of people and COVID-19 infections. This has been done by making use of data related to purchases made by residents from Tesco supermarket, for London Boroughs. The data related to pre-existing health conditions, for same regions, was obtained from the London Datastore. Our study indicates that for the London Boroughs' data, food products containing alcohol, carbohydrates and fats are weakly correlated with the number of COVID-19 cases. We believe that these results warrant a more detailed investigation of causality.**

Keywords—COVID-19, correlation, mutual information, regression, food groups

I. INTRODUCTION

Modern data aggregation methods have made large, diverse data sets available that can be used to determine and establish relationships between different facets of life. Thus, data about jobs, economy, housing, health, environment, purchases is available and can be used to determine direct relationships at scales at which it was not previously possible. Availability of large amounts of data has ushered in a new era in data analytics. Thus, when Coronavirus disease 2019 (COVID-19), also known as Severe acute respiratory syndrome Coronavirus 2 (SARS-COV-2), began spreading, data collection along with its availability and analysis became important, albeit less than finding a treatment or developing a vaccine, but still important for tracking the spread of the disease and identifying super spreaders [1][2]. Considering the fact that approximately 110.9 million people have been infected with the virus [3], monitoring the spread of COVID-19 is still an important area of research.

A secondary area of focus for researchers has been to understand if there is any relationship between pre-existing health conditions and COVID-19 infections or deaths. The myths and conspiracies surrounding infection of COVID-19 due to stress were addressed by Georgiou et al. in [5]. The authors clarified the myth that people with stress are not more likely to be affected by COVID-19 compared to others. In [6], Jordan et al. observed that different studies based on data collected from Wuhan, Italy and UK citing increased risk of COVID-19 related deaths for people suffering from pre-existing health conditions. However, they highlighted that these studies comprised of a small population ranging from 100 to 40,000 participants with data that is not readily available and, in some cases, incomplete. Therefore, there is a need for improved data acquisition for analysis and, hence, reaching better conclusions. It was highlighted in [7], based on a study by Chinese Center for Disease Control (CDC) of approximately 44,000 lab-tested positive cases, that advanced age, heart conditions, cancer, hyper-tension, chronic respiratory diseases, diabetes increase the risk of fatality in case of a Coronavirus infection. Data collected from patients in China suggested that smoking and obesity were linked with higher risk of severe infection and death [8]. In another study, Stefan et al. [9] identified that patients with obesity are at increased risk for severe COVID-19 symptoms.

In this work we try to identify if eating habits have a direct relationship with the number of COVID-19 cases in a particular region. This is based on the assumption that eating habits generally effect the health of an individual, since pre-existing conditions seem to have a relationship with COVID-19. Therefore, it will be interesting to see if any food group has any relationship with the number of COVID-19 cases in a geographic region. To conduct this analysis, data for COVID-19 cases, along with the data of pre-existing health conditions and data related to eating habits of people is required for a particular region. All of this data was not available at the same spatial resolution and for the same temporal window. However, we were able to compile data from different sources to obtain data at the resolution of Boroughs for London region.

The rest of the paper is organized as follows. Section II introduces the sources and type of data used in this study. Section III presents a correlation-based analysis between

COVID-19 cases, deaths and pre-existing health conditions and a few food groups. Section IV presents regression analysis between prediction of COVID-19 cases and eating habits. Finally, Section V presents our conclusions and future perspectives.

II. Literature review

Predicting the causes of an illness or factors that may increase chances of being affected by it are important for controlling the disease. Massive impact of COVID-19 on people's lives and economies of countries has led many researchers to investigate the reasons for its causes and factors that may protect certain individuals while make other more vulnerable.

Since patient numbers continue to grow, healthcare systems are faced with the challenge of reporting and treating cases. Therefore, the authors in [10] focused on the correct number of COVID-19 cases as they were being under-reported due to various factors. They proposed to use the data from one country, with adequate confirmed cases and reported deaths, to correct the number of reported cases for another country. They proposed vulnerability factor for adjusting the dissimilarity in the population demographics of the target country and the benchmark country and to predict the average death rate.

In [11], Atkins et al. established a relation between pre-existing health conditions and chances of being infected by Covid-19. Using data from UK Biobank baseline assessment 2006–2010, hospital discharge records and death records, they used logistic regression to determine the relationship between preexisting health conditions and hospitalized lab-confirmed COVID-19 cases. They concluded that age group, sex, ethnicity, education, preexisting diagnoses of dementia, type 2 diabetes, chronic obstructive pulmonary disease, pneumonia, depression, atrial fibrillation, and hypertension emerged as independent factors that may affect COVID-19 hospitalization cases.

Authors in [12] tested Logistic, Bertalanffy and Gompertz models for predicting the number of COVID-19 cases. Betalanffy and Gompertz models examine factors that oversee and affect growth and extinction law of populations. The authors used SARS data for development of algorithm and demonstrated that it could be extended for analysis of COVID-19 data. Wang et al. used COVID-19 epidemiological data collected at John Hopkins University and the SARS data collected in 2003 from China to predict the movement of epidemic using logistic growth model [15].

Gupta et al., in [16], proposed the use of linear and polynomial regression to predict the number of COVID-19 cases by using the Susceptible, Exposed, Infected, and Recovered (SEIR) model. They tested their algorithm on data collected locally between January and May 2020 and predicted 175,000 to 200,000 cases for the three weeks test period. In [20], authors propose the use of polynomial regression and support vector regression to estimate the outbreak of COIVD-19.

Rustam et al. tested linear regression, least absolute shrinkage, selection operator (LASSO), support vector machines (SVM) and exponential smoothing (ES) to predict the number of infections, deaths and recoveries. Their results suggested that ES, which is type of a time series analysis, provided the best prediction while SVM results were least accurate [17].

Tian et al. [18], compared the prediction accuracies obtained for COVID-19 data, made available by John Hopkins University, by using long-short-term-memory model (LSTM), hierarchical Bayes model, and hidden Markov model. For data spanning 84 days for six countries, they observed that hierarchical Bayes model correctly estimated the plateau in the growth curve of incidents while LSTM resulted in the least accuracy in daily predictions. Alzahrani et al. proposed the use of autoregressive integrated moving average (ARIMA) model for predicting the number of COVID-19 cases. ARIMA is a time series model and provided the best results among models tested by them. They predicted that the number of cases will rise to 127,000 by the end of May 2020 [21]. They performed quantitative comparison of the proposed method with other method using root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE) and R^2.

Istaiteh et al. compared the efficacy of ARIMA, LSTM, and Convolutional Neural Networks (CNN) for forecasting the number of COVID-19 cases [23]. Using the data made available by John Hopkins University they trained on data from 189 countries to predict the number of cases for one week. Their results indicated that the CNN architecture as superior to LSTM and ARIMA architecture. However, they concluded that a hybrid machine learning model seemed to be the best approach.

Since COVID-19 is a type of influenza, Kumar et al. in [13] proposed to use Chest X-rays to diagnose COVID-19 using Synthetic Minority Oversampling Technique (SMOTE) to balance intra-class variability for binary classification, pneumonia or normal. They reported an accuracy of 97.3% for Random Forest and 97.7% for XGBoost predictive classifiers. Other classifiers such as Logistic Regression (LR), k-Nearest Neighbor (kNN), Decision Tree (DT), Adaboost and Naive Bayes (NB) resulted in an accuracy of 96.6%, 94.7%, 93.1%, 97.3%, 92.1% and 88.9%, respectively.

Yan et al., in [14], predicted the survival rate of patients using XGBoost classifier by analyzing lactic dehydrogenase (LDH), lymphocyte and high-sensitivity C-reactive protein (hs-CRP) biomarkers in blood samples of patients. These biomarkers help measure the breakdown of tissue in various diseases including pneumonia. Similar to the work of Yan et al., Booth et al. [22] proposed the use of biomarkers for predicting the unfortunate death cases in case of a COVID-19 infection. They identified C-reactive protein, blood urea nitrogen, serum calcium, serum albumin and lactic acid as the biomarkers that can be used as data for training support vector machine (SVM).

In [19], Khanday et al. proposed the use of different machine learning algorithms for differentiating between COVID-19, severe acute respiratory syndrome (SARS) and middle east respiratory syndrome (MERS) cases by using the text of clinical reports. The data used in the research was obtained from Github and comprised of only 212 reports. They tested inverse document frequency, bag of words, report lengths among other features for classification. Among logistic regression, decision tree, random forest, Adaboost, Naïve Bayes and SVMs they obtained the best results with Naïve Bayes.

Gao et al. proposed the use of mortality risk prediction model for COVID-19 (MRPMC) in [24]. Making use of clinical records in the form of electronic health records, they attempted to categorize patients by mortality risk at the time of admission. The predictions were made for the next twenty days at time of admission. The data was collected from Tongji Medical College (Sino-French New City Campus and Optical Valley Campus) of Tongji Hospital and central hospital of Wuhan. They used random forest, SVMs, logistic regression, neural network, k-nearest neighbor and gradient boost methods for training and prediction. The proposed MRPMC model comprised of logistic regression, random forest, neural network and gradient boosting-based models.

III. DATA

Generally, residential regions across England are divided hierarchically into lower layer super output area (LSOA), medium layer super output area (MSOA) and Boroughs. These regions range from an average population of 1,500 to 10,000 to 100,000. The government of England has developed a comprehensive data collection system which provides data at different scales ranging from Boroughs to LSOA [25]. This data is made publicly available at London Datastore and provides public access to different indicators, collected at different times, available at different scales. For example, the data related to health or pre-existing health conditions, i.e., asthma, diabetes, heart conditions, hypertension and obesity, is available at MSOA scale and was collected in 2019. However, it has been indicated that the data may not be very accurate at LSOA and MSOA level and, hence, may present a better representation at larger scale. Data related to child poverty and deprivation is also available at MSOA scale and was collected in 2019. Data related to the number of COVID-19 cases is available at Borough scale, however, the data related to the number of deaths is available at MSOA scale.

Population data (i.e., total population, density per square kilometer, number of people below 18 years of age, number of people between 18 and 65 and number of people above 65 years of age) was compiled from a recent study by Aeillo et al. [4]. In their work the authors have shared the data for purchase of food items, from Tesco food mart, at LSOA, MSOA and Borough levels, during the year 2015. The data was collected from 411 Tesco stores for 1.6 million customers. Since the data collected did not cover all the people living in that area, the authors have calculated normalized representation, i.e., a measure of the number of customers as per the number of total residents of that community. The other interesting data provided by them is the caloric content of purchased items including, but not limited to, eggs, oils, red meat, poultry, fruits and vegetables, sweets, wines, water, spirits, beer, fish, grains, dairy. The authors later aggregated these individual food items into larger food groups, i.e., fats, saturated fats, sugar, proteins, carbohydrates, fibers and alcohol. The authors utilized this data to determine the extent of correlation between obesity and diabetes with respect to food groups. In another study they highlighted the extent of correlation between hypertension and heart disease with respect to the above-mentioned food groups [26].

In their work, Aiello et al. demonstrated that diseases such as hyper-tension, diabetes, cardiovascular diseases and obesity are correlated to food purchases, i.e., eating habits using Tesco food products related dataset. Since authors have identified a link between pre-existing health conditions and COVID-19 cases / deaths. Aeillo et al. demonstrated that there was a relationship between food consumption and health disorders like asthma, cardiovascular issues. Inspired by this we propose to identify if there exists any relationship between food consumption and number of COVID-19 cases for the regions of London Boroughs. In this regard, first we have tried to identify if a linear relationship exists between food consumption and COVID-19 cases or deaths.

IV. PROPOSED METHDOLOGY AND RESULTS

Correlation is a measure of the strength of a linear relationship between two quantities. It indicates the relative change in variation of the two quantities and helps establish the basis of any linear relationship between them. This is the same measure employed by Aeillo et al. [4] to highlight the relationship between food consumption and health of residents in London region. It was established in [4][26] with food data from 2016 and health data from 2018 that there is a strong correlation between food consumption and health issues like diabetes, obesity, hypertension and cholesterol. In [9], Stefan et al. demonstrated that people with high blood glucose levels and obesity are at higher risk in case of COVID-19 infection. Therefore, we attempted to explore the relationship between food consumption in a particular region and COVID-19 infections.

To determine if a linear relationship exists between COVID-19 cases and the purchase of food items in a particular Borough of London, we calculated correlation between them. Instead of presenting the result for all 33 boroughs we have presented below the result for six boroughs in Table II. These boroughs were randomly selected. For reference, sample data used for calculating the correlation of the first column, i.e., Barnet borough, is shown in Table I.

In Table I, each row represents the data for a month and the columns represent the food items purchased or the number of reported cases. The COVID-19 data is from March 2020 to December 2020 while the food items data is available from January 2015 till December 2015. We have only selected the data from March till December 2015 for this analysis, assuming that the food item consumption does not vary significantly across years and follows a general trend depending on weather and population demographics.

TABLE I. AVERAGE CALORIC CONTENT OF PURCHASES BROKEN UP BY FOOD GROUPS AND MONTHS FOR BARNET BOROUGH.

Month	Cases	Fat (kcal)	Salt (kcal)	Sugar (kcal)	Protein (kcal)	Carbs (kcal)	Fiber (kcal)	Alcohol (kcal)
March	553	8.676	0.573	9.903	5.167	17.323	1.621	0.210
April	746	8.482	0.551	9.287	5.066	16.726	1.625	0.239
May	199	8.471	0.549	9.267	5.100	16.703	1.672	0.219
June	37	8.259	0.539	9.292	5.028	16.443	1.609	0.219
July	705	8.248	0.541	9.146	5.036	16.539	1.594	0.230
Aug.	211	8.471	0.561	9.166	5.182	16.614	1.605	0.228
Sept.	549	8.562	0.571	9.491	5.129	17.071	1.653	0.214
Oct.	2105	8.756	0.580	9.827	5.200	17.335	1.651	0.226
Nov.	2669	8.923	0.587	9.887	5.165	17.331	1.675	0.239
Dec.	10118	9.530	0.590	9.957	5.111	17.401	1.611	0.346

TABLE II. CORRELATION COEFFICIENT BETWEEN NUMBER OF CASES AND FOOD GROUPS FOR BARNET, CROYDON, EALING, HACKNEY, LAMBETH, TOWER HAMLETS

	Barnet	Croydon	Ealing	Hackney	Lambeth	T.Hamlets
Fat	0.924	0.855	0.896	0.889	0.912	0.826
Salt	0.653	0.733	0.621	0.675	0.669	0.591
Sugar	0.615	0.566	0.626	0.586	0.644	0.584
Proteins	0.100	-0.155	0.217	-0.041	-0.248	0.172
Carbs	0.584	0.695	0.764	0.647	0.671	0.622
Fiber	-0.097	-0.185	0.154	-0.191	-0.248	-0.034
Alcohol	0.966	0.958	0.943	0.943	0.961	0.959

The correlation results for the data of Table I are presented in the first column of Table II. For the region of Barnet, highest correlation is between the number of cases and alcohol (0.966, 0.958, 0.943, 0.943, 0.961, and 0.959) and these values highlighted in green color in Table II. This is followed closely by fat, highlighted in orange color (0.924, 0.855, 0.896, 0.889, 0.912, and 0.826). After these three food types, the correlation coefficient values drop significantly and salt and carbohydrates share the next highest correlation (0.653, 0.733, 0.764, 0.675, 0.671, and 0.622) which is highlighted in blue. The same trend of alcohol having the highest correlation with the number of cases and fat having the second highest correlation was observed for all the boroughs. The data in Table I represents the number of cases and the nutritional energy of average product in kcals [4]. This means it is the average number of calories purchased by residents of an area across all purchased products for a particular food type [4].

Another comparison that was made was between the cumulative data of each borough. To do so the number of cases was summed up for the duration of ten months. This resulted in a cumulative number of cases for each region. Similarly, the food group purchased for a particular region for the duration of 10 months was aggregated and divided by the total number of months to get an average. We have only presented this data for a few boroughs because of constraints of available space. The actual data comprises of 33 rows, one for each borough and 8 columns. Some data samples are shared in Table III while the correlation results are shared in Table IV.

TABLE III. AVERAGE CALORIC VALUES OF FOOD GROUPS PER PURCHASE OVER 10 MONTH TIME PERIOD AND AGGREGATE REPORTED COVID-19 CASES FOR BOROUGHS.

Borough	Fat (kcal)	Salt (kcal)	Sugar (kcal)	Protein (kcal)	Carb (kcal)	Fiber (kcal)	Alcohol (kcal)	Cases
Barnet	8.594	0.567	9.530	5.130	17.026	1.639	0.228	17411
Bexley	9.119	0.591	10.941	5.304	19.997	1.65r7	0.201	13976
Brent	8.962	0.570	10.149	5.133	18.726	1.586	0.207	14555
Sutton	9.358	0.552	11.629	5.207	19.756	1.701	0.223	9058
T. Hamlets	9.004	0.585	10.225	5.273	18.506	1.602	0.199	18087
Waltham Forest	9.075	0.613	9.795	5.406	18.293	1.574	0.237	15003

The values given in Table III represent the average kcals of purchased items over a period of 10 months and the total number of reported cases for those months. The results of correlation in Table IV indicate that carbohydrates have the highest positive correlation with the number of COVID-19 cases while alcohol has the highest negative correlation with the number of COVID-19 cases. This seems to indicate that the regions where more carbohydrates were used those regions had more reported cases of COVID-19 while regions which had higher sales of alcohol had negative correlation with number of COVID-19 cases. The magnitude of correlation values for these two food groups is close to 0.6 which is significant as the rest of the food groups have correlation values around 0.35.

TABLE IV. CORRELATION COEFFICIENT BETWEEN REPORTED COVID-19 CASES AND FOOD ITEMS BETWEEN BOROUGHS OVER 10 MONTH PERIOD.

	Fat	Salt	Sugar	Protein	Carbs	Fiber	Alcohol
Cases	0.274	0.360	0.367	-0.289	0.557	-0.354	-0.583

Finally, we tested different regression algorithms for predicting the number of cases using the food item with the maximum correlation with the number of cases. In this regard we tested linear regression, support vector regression, polynomial regression, decision tree and random forest regression for predicting the number of cases. Implementations of these algorithms available in the scikit-learn library were used for this purpose. We attempted to predict the total number of cases for a region based on the number of cases of other regions and certain food groups. To measure the performance of the regression algorithm we use the coefficient of determination or R^2 score, which measures the amount of variation in the output variable depending upon the independent input variable(s). The ideal value for R^2 score is 1.0 while negative values indicate poor performance.

TABLE V. R^2 SCORE FOR PREDICTING NUMBER OF CASES IN BOROUGHS USING FOOD ITEM DATA FROM

	LR	SVR	PR	DT	RF
Alcohol	0.621	0.353	0.314	-0.113	-0.063
Carbs	0.605	0.209	-1.495	0.664	0.584

From Table V, the best prediction accuracy was obtained using linear regression where for carbohydrates and alcohol the R^2 score was approximately 0.6. Carbohydrates had the highest correlation score when the aggregated data was used as shown in Table IV and for carbohydrates the decision trees provide the best R^2 score of 66.4%. Since alcohol is negatively correlated to the number of cases it provides a high R^2 score for linear regression but for decision tree and random forest regression methods provide poor R^2 scores.

V. CONCLUSION

Recent research has highlighted that people with pre-existing health conditions like cholesterol, hypertension, high blood glucose, and obesity are more susceptible to suffer from adverse effects of COVID-19. Researchers have also established a link between food purchases and adverse health conditions of people living in boroughs of London. Inspired by these studies, we have investigated if correlation exists between number of COVID-19 cases in London boroughs and food purchases in those regions. The data being considered is not for the same year, however, we have

noticed relatively high correlation with alcohol, carbohydrates and fats under different evaluation scenarios. This is a preliminary work and the insights obtained by us suggest that a more comprehensive study should be done to explore the relationship in more detail. We have noticed relatively high correlation of alcohol and fat with COVID-19 cases by considering the data of each borough over a period of 10 months. We have also noticed relatively high correlation between alcohol, carbohydrates purchased over a period of 10 months and the number of cases during that period for all the boroughs. Therefore, in our future work we propose to explore moving average prediction algorithms like ARIMA to incorporate time series analysis for all boroughs together. We believe that this could improve prediction results for the number of cases based upon time series analysis of boroughs monthly data.

REFERENCES

[1] E. Emanuel, G. Persad, R. Upshur, B. Thome, M. Parker, A. Glickman, C. Zhang, C. Boyle, M. Smith, J. Phillips, "Fair Allocation of Scarce Medical Resources in the Time of Covid-19", The New England Journal of Medicine, no.382, pp. 2049-2055, 2020.

[2] J. Grein, N. Ohmagari, D. Shin, G. Diaz, E. Asperges, A. Castagna, T. Feldt, G. Green, M. Green, F. Lescure, E. Nicastri, R. Oda,"Compassionate Use of Remdesivir for Patients with Severe Covid-19", The New England Journal of Medicine, no.382, pp. 2327-2336, 2020.

[3] "COVID-19 Coronavirus Pandemic", Coronavirus Update (Live): 110,951,913 Cases and 2,454,722 Deaths from COVID-19 Virus Pandemic - Worldometer (worldometers.info). *(Last accessed 20th Feb 2021)*

[4] L. Aiello, D. Quercial, R. Schifanella, L. Prete, "Tesco grocery 1.0, a large-scale dataset of grocery purchases in London", Scientific Data, vol.7, no.1, pp. 1-11, 2020

[5] N. Georgiou, P. Delfabbro, R. Balzan, "COVID-19-related conspiracy beliefs and their relationship with perceived stress and pre-existing conspiracy beliefs", Personality and Individual Differences, vol. 166, no. 1, 2020.

[6] R. Jordan, K. Cheng, "Covid-19: risk factors for severe disease and death", BMJ 2020; 368

[7] Z. Wu Z, J. M. McGoogan, "Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72314 cases from the Chinese Center for Disease Control and Prevention", Journal of American Medical Association (JAMA), 2020. doi:10.1001/jama.2020.2648 pmid:32091533

[8] D. Wang, B. Hu, C. Hu, "Clinical characteristics of 138 hospitalized patients with 2019 Novel Coronavirus–Infected pneumonia in Wuhan, China.", Journal of American Medical Association (JAMA), 2020;323:1061-9. doi:10.1001/jama.2020.1585 pmid:32031570

[9] N. Stefan, A. Birkenfeld, M. Schulze, D. Ludwig, "Obesity and impaired metabolic health in patients with Covid-19", Nature Reviews Endocrinology, pp.1-2, 2020.

[10] A. Lachmann, "Correcting under-reported COVID-19 case numbers." MedRxiv (2020).

[11] J. Atkins, J. Masoli, J. Delgado, L. Pilling, C. Kuo, G. Kuchal, D. Melzer, "Preexisting comorbidities predicting COVID-19 and mortality in the UK biobank community cohort." The Journals of Gerontology: Series A 75.11 (2020): 2224-2230.

[12] J. Lin, L. Kewen, Y. Jiang, G. Xin, T. Zhao, "Prediction and Analysis of Cornovirus Disease 2019", arXiv:2003.05447, 2020.

[13] R. Kumar, R. Arora, V. Bansal, V. Sahayasheela, H. Buckchash, J. Imran, N. Narayanan, G. Pandian, B. Raman, "Accurate prediction of COVID-19 using chest x-ray images through deep feature learning model with smote and machine learning classifiers." MedRxiv (2020).

[14] L. Yan, H. Zhang, J. Goncalves, Y. Xiao, M. Wang, G. Yuqi, C. Sun, "A machine learning-based model for survival prediction in patients with severe COVID-19 infection." MedRxiv (2020).

[15] P. Wang, X. Zhang, L. Jiayang, Z. Bangren, "Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics." Chaos, Solitons & Fractals, no. 139, p.110058, 2020.

[16] R. Gupta, G. Panday, P. Chaudhary, K. Saibal, "Machine learning models for government to predict COVID-19 outbreak." Digital Government: Research and Practice, vol.1, no.4, pp.1-6, 2020.

[17] F. Rustam, A. Reshi, A. Mehmood, S. Ullah, B. On, W. Aslam, G. Choi, "COVID-19 future forecasting using supervised machine learning models." IEEE access, no.8, pp. 101489-101499, 2020.

[18] Y. Tian, I. Luthra, X. Zhang, "Forecasting COVID-19 cases using machine learning models." MedRxiv, 2020.

[19] A. Khanday, S. Rabani, Q. Khan, N. Rouf, M. Mohiuddin, "Machine learning based approaches for detecting COVID-19 using clinical text data", International Journal of Information Technology, vol.12, pp. 731-739, 2020.

[20] R. Tamhane, S. Mulge. "Prediction of COVID-19 outbreak using machine learning." International Research Journal of Engineering and Technology (IRJET), vol.7, no.5, 2020.

[21] S. Alzahrani, I. Aljamaan, E. Al-Fakih. "Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions." Journal of infection and public health, vol.13, no.7, pp.914-919, 2020.

[22] A. Booth, A. Elizabeth, P. McCaffrey. "Development of a prognostic model for mortality in COVID-19 infection using machine learning." Modern Pathology, pp.1-10, 2020.

[23] O. Istaiteh, T. Owais, N. Al-Madi, S. Abu-Soud, "Machine Learning Approaches for COVID-19 Forecasting." 2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA). IEEE, 2020.

[24] Y. Gao, G. Cai, W. Fang, H. Li, S. Wang, "Machine learning based early warning system enables accurate mortality risk prediction for COVID-19." Nature communications, vol.11, no.1, pp.1-10, 2020.

[25] "London Data Store", \url{https://data.london.gov.uk/}. Last accessed 2nd August 2020.

[26] L. Aiello, D. Quercial, R. Schifanella, L. Prete, "Large-scale and high-resolution analysis of food purchases and health outcomes", EPJ Data Science, vol.8, no.1, pp. 1-14, 2019.