Review Of The Relationship Between Weather And The Rate Curve Of COVID-19 Disease In The Tabuk Region By Machine Learning Algorithms

Latifah Albalawi1, Mona Alasmari2, Rabda Alenzy3, Manal Alswat4, Majed Aborokbah5, Mohammad Hijji6

1,2,3,4,5,6Faculty Of Computers And Information Technology, University Of Tabuk, Saudi Arabia.

Abstract—In light of this global pandemic, scientists have sought to decipher the COVID-19 virus to save humanity from this deadly disease, determine the active role of data scientists in providing important research on this disease and search the reasons for its decline or spread, as much research has been conducted. In this review paper, we want to study the relationship between different weather conditions around the world, particularly in the Tabuk region and the rate of increase and decrease in the number of deaths of COVID-19 patients to predict the rate of infection over the one year. A dataset was created to collect cases for the region for a full year of seasons from 31 March 2020 until 31 March 2021. The cases were divided into daily recoveries and deaths and classified according to the seasons of the year. We then applied a set of machine-learning algorithms to help us understand this relationship. To achieve this goal, three neural network, k-mean and decision tree classifier machine learning models were built, trained and tested. The experimental results showed that the neural network model outperformed the other two models, as it achieved 80% prediction accuracy.

Keywords—Machine Learning, Prediction, Weather Seasons, COVID-19, Tabuk Region, Saudi Arabia

I. INTRODUCTION

Considering the outbreaks of many different diseases during the past years and the spread of many chronic respiratory diseases, the efforts of the World Health Organization have focused on controlling them before their concentration and the emergence of more resistant strains the next day. On 31 December 2019, specifically in Wuhan, China, a new map of the SARS-CoV-2 virus was drawn [1], [2].

It mutated to become the most dangerous COVID-19 in the history of respiratory disease [3]. After discovering this rapidly spreading mutation and the severity of its symptoms, the World Health Organization declared a state of emergency with strict and severe international efforts [4]. However, this disease was not fully controlled and was stimulated with preventive measures daily. On 5 April 2021, the number of infections worldwide reached 131 million, including 2,850,000 deaths [5], which is a clear indication of the rapid spread and transmission of COVID-19. Saudi Arabia detected the first cases of infection with the new coronavirus, COVID-19, on Monday, 2 March 2020 and the Ministry of Health announced the emergence of laboratory results confirming the registration of the first case [6]. The infection occurred in a citizen coming from Iran
through the Kingdom of Bahrain in the city of Riyadh, and his presence was not disclosed in Saudi Arabia port [7].

Smith and Coşkun et al. [8], [9] suggest a direct relationship between human population density, weather and transmission speed of the virus. Dalziel [10] stated that climate predictions are the main key to predicting the curve of the occurrence of the virus.

Similarly, Epstein [11] reported that poor weather due to long-term climate change indefinitely contributed to the spread of the West Nile virus (WNV) in the United States and Europe.

Researchers are still publishing a significant amount of research to determine the relationship between the spread of the virus and weather; however, there has been limited and scarce research examining the positive relationship between them, especially since most of them focused on temperatures while neglecting other factors, such as the quality of aerodynamics on COVID-19 [12], [13].

Previous studies by Xie, Jingui and Yongjian Zhu [14] focused on the relationship between temperature and humidity retroactively changing with COVID-19 and found it linked through linear regression. Many previous researchers have asked whether high temperatures during the summer could slow the spread of COVID-19, as is the case with other seasonal viruses [15], [16].

Machine learning algorithms and artificial intelligence (AI) have contributed to data analysis and the fight against COVID-19. Data analysis in Python provides safer, more accurate and more effective solutions. Modern applications supported by AI mainly include assessment and diagnosis of COVID-19. As well as groundbreaking basic and clinical research, commercial products have been developed that successfully integrate AI to combat COVID-19 and clearly demonstrate the potential of AI in medical imaging and more, especially during the ongoing pandemic. [17]

This study aims to discuss the role of machine learning algorithms in predicting the relationship of weather to COVID-19 and its widespread prevalence. This is especially true with data analysis and AI in the fight against COVID-19. This will inspire the future of the importance of data science, applications and systematic research. We present, first, the most related research that dealt with the study of the relationship of climate with viruses and seasonal epidemics, and second, the most important research that dealt with the study of the relationship of climate with COVID-19.

A. Studying the relationship of climate with viruses and seasonal epidemics

Previous studies have shown that decreases in temperature and humidity increase the risk of influenza attacks in cold climates [9]. Other studies show that the spread of aerosol influenza virus depends on surroundings, relative humidity and temperature. Twenty experiments were conducted at relative humidity ranging from 20% to 80% and 5°C. This indicated that both cold and dry conditions favoured the transition. Also, inhaling dry air can cause dryness of the nasal mucosa, which in turn makes humans more susceptible to infection with respiratory viruses [18].

Influenza and respiratory syncytial virus (RSV) activity consistently peaked during the winter months in temperate regions, while diversity was present in tropical regions. Many temperate regions experienced semi-
annual influenza activity, with peaks occurring in winter and summer. Bi-annual activity was relatively common in the tropics of Southeast Asia for both viruses.

The series model applied to influenza data from 85 countries confirmed the existence, in general, of latitudinal gradients in the timing and duration, seasonal breadth and interannual variability of epidemics. Eighty percent of tropical locations experienced distinct RSV seasons lasting six months or less, while the proportion was 50% for influenza [19]. Although there is evidence that climate is related to changes in WNV epidemiology, there are other vectors of diseases that contribute to their spread [20].

A relationship between the outbreak of severe acute respiratory syndrome (SARS) and the temperature of the environment seven days before the start, an interval that corresponds to the known incubation period known for SARS. The optimum environmental temperature associated with SARS cases, which encourages the growth of the virus, is between 16°C and 28°C [21].

A sharp rise or decrease in the temperature of the environment related to cold has led to an increase in the incidence of SARS due to the possible effect of the weather on the human immune system [22].

**B. Studying the relationship of climate to COVID-19**

The question that was asked with the spread of Covid-19, will the Corona virus disappear when the temperature rises? Based on the fact that some viruses infection decreases during high temperature, but this does not confirm the extent of the effect of high temperature on the Covid-19 virus.[62]

Studies in China indicate that doubling time is positively correlated with temperature and humidity reversal, which indicates a decrease in the rate of development of COVID-19 with the advent of spring and summer in the northern hemisphere (20 °C). The increase is expected to delay the doubling time by 108 days. These variables explain 18% of the variance in the doubling time of the disease; the remaining 82% may be related to other factors [23], [24].

These results are consistent with other studies that have indicated that aerosols and the influenza virus spread depending on relative humidity and temperature. Other authors suggest that some diseases spread faster at higher humidity levels [25]. They also found that an increase in confirmed cases occurred in the provinces of China in humid conditions ranging from cold and dry (Jilin or Helu Longjiang) to tropical (Guangxi or Singapore), indicating that changes in weather, as expected with the advent of spring and summer, will not necessarily lead to a decrease in the pandemic [26], [27].

On the other hand, Jin Bu, et al. [28] concluded that constant warmth and dry weather are conducive to the survival of COVID-19 and speculated so. Conditions such as temperatures ranging from 13°C to 19°C and humidity between 50% and 80% are suitable for the survival and transmission of this new coronavirus.

However, COVID-19 data does not cover all counties since the beginning of the outbreak, as some southern Chinese cities (such as Sanya, Haikou and Danzhou) had average daily maximum temperatures of 30°C [29]. This means that ambient temperature does not have a significant effect on the transmission of the virus.

In a study conducted in Finland, weather conditions, such as temperature and relative humidity, were unrelated to the occurrence of COVID-19 during the first wave in the arctic and sub-arctic winter and spring. Inference depends on a relatively small number of cases within a specific period [30].
<table>
<thead>
<tr>
<th>Literature</th>
<th>Country</th>
<th>Study time (seasons)</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>B Oliveira et al [29]</td>
<td>China</td>
<td>1 January to 23 March 2020 (Winter)</td>
<td>There is no correlation between temperature and covid-19</td>
</tr>
<tr>
<td>W. J. Tabachnick et al [28]</td>
<td>United State</td>
<td>12 December 2019 to 22 April 2020 (Winter–Spring)</td>
<td>The weather variables are more relevant in predicting the mortality rate when compared to the other census variables such as population, age and urbanization.</td>
</tr>
<tr>
<td>K. Ghosh et al [44,45]</td>
<td>India</td>
<td>30th January to 31st May 2020 (Winter–Spring)</td>
<td>The huge disparity in the growth rate and total cases of COVID-19 is not only due to climate but also other factors.</td>
</tr>
<tr>
<td>B. Heibati et al [31]</td>
<td>Finland</td>
<td>January 1 to May 31 (Winter–Spring)</td>
<td>Weather conditions such as temperature and humidity have nothing to do with the occurrence of COVID-19 during the first wave in the Arctic and Subarctic.</td>
</tr>
<tr>
<td>M. Alkhowailed et al [33]</td>
<td>Saudi Arabia</td>
<td>March 5 to May 20, 2020 (Spring)</td>
<td>The number of positive cases of COVID-19 increases due to the decrease in temperature and humidity, and the average decrease in wind speed is associated with a higher number of cases.</td>
</tr>
</tbody>
</table>
For a study published by Şahin, Mehmet in April 2020 showing the impact of weather on the spread of infection depending on human activity at work and social gatherings, where the results proved that temperature negatively affects the index of cases that an increase in temperature by 1 degree Celsius reduces cases Amount 0.0006.[61]

In another study, the new data revealed interesting facts about the spread of COVID-19 in Saudi Arabia [31], showing that the number of positive cases of COVID-19 increased due to a decrease in temperature or humidity, while it was also found that the average decrease in wind speed was related to an increase in the positive number of cases.

However, this study was not comprehensive for all seasons of the year; it was during a short period and the study was only on major cities with dense populations. Therefore, we conducted a study over a full year, beginning with the emergence of COVID-19 in the Tabuk region until now.

Considering our goal in this research, we focus on the Tabuk region of Saudi Arabia and study the relationship between different seasons and climate change and its relationship to the number of deaths according to climate through a sequence like that of a neural network. In TABLE I presents a summary of previous studies on the relationship between and COVID-19.
II. STUDY AREA

In this section, we discuss the area in which we worked on the study. The Tabuk region in Saudi Arabia is in the northwest and surrounded by the Red Sea to the west. It is located on the seacoast of the Dhaba province, field governorate, face and innovations. It is considered that the Tabuk region plays a role in civilization and growth [32]. TABLE II shows a related dataset.

<table>
<thead>
<tr>
<th>seasons</th>
<th>death</th>
<th>recovered</th>
<th>infected</th>
<th>date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spring</td>
<td>0</td>
<td>10</td>
<td>6</td>
<td>09/04/21</td>
</tr>
<tr>
<td>Spring</td>
<td>0</td>
<td>11</td>
<td>12</td>
<td>08/04/21</td>
</tr>
<tr>
<td>Spring</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td>07/04/21</td>
</tr>
<tr>
<td>Spring</td>
<td>0</td>
<td>4</td>
<td>11</td>
<td>06/04/21</td>
</tr>
<tr>
<td>Spring</td>
<td>0</td>
<td>3</td>
<td>12</td>
<td>05/04/21</td>
</tr>
<tr>
<td>Spring</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>20/04/01</td>
</tr>
<tr>
<td>Spring</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>20/03/30</td>
</tr>
<tr>
<td>Spring</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>20/03/29</td>
</tr>
<tr>
<td>Spring</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>20/03/28</td>
</tr>
<tr>
<td>Spring</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>20/03/20</td>
</tr>
</tbody>
</table>

Based on methods of data collection and analysis in the selection of Tabuk region, a list of data was made for eight governorates in addition to the administrative capital of the region. Then the infection curve index was calculated and divided into confirmed cases, critical cases, recovered cases and deaths. the extent of the impact of weather variables on the spread of the virus throughout the year.
III. DATASET AND METHODOLOGY USED

Urban areas that are most vulnerable to the COVID-19 pandemic can be identified using a variety of methods and assessment criteria [33]. This process is complex and requires a wide range of environmental, natural, meteorological and health data in addition to quantitative and qualitative criteria to make the final decision. Data integration and data mining technology are suitable approaches to supporting decision-making and planning procedures. These technologies have been used to guide the sustainable development of the Tabuk region and the interrelationships of COVID-19 disease, in particular, to reduce and avoid the pandemic. This study demonstrates the application of a proposed model to predict the COVID-19 epidemic with the help of Python to determine the areas of the Tabuk region most vulnerable to the pandemic and their impact on urban areas in the region. In TABLE II To achieve this, different data sets were used to understand the different confirmed case units and characteristic types of factors affecting disease prevalence in the study area, using a carefully proposed model a data set containing 5 columns and 367 rows was used where data were collected on daily on infected, recovered and death cases based on the season by used for pandemic prediction modelling.

A. Preparing and pre-processing procedures

In general, data scientists spend most of their time processing data. This procedure consists of selecting appropriate features and cleaning them and preparing them to become inputs or independent variables for a machine learning model. From data pretreatment to data mining postprocessing, this process consists of a succession of transformation phases.

At the beginning of the data collection stage, a detailed plan of action was drawn up based on scientific foundations in data science in terms of working on steps:

Data Preprocessing, Data Mining, Postprocessing

a. Data cleaning

The most critical stages in this section are estimating unavailable data in the database, removing noise in data, eliminating outliers and unrelated data and eliminating data anomalies. When the data set was completed, there were missing days of estimation of deaths or injuries, or sometimes there were unrecorded days, we followed the categorical data method. This method is intended for handling categorical data. Missing values have been replaced by mode value (the most frequently occurring value)

b. Data integration
In most cases, the data is kept in different files and resources. It is necessary to integrate data before applying data mining techniques. This stage includes removing missing values, outliers, and redundant data.

Data were collected from multiple sources, depending on the relative convergence between the data ratios and the relative variance in the data rate. Work was done to build a data set with a structured framework, where each category was collected in a separate framework.

**c. Data reduction**

All data are not always demanded in data mining, and only part of the data needs to be processed. We dropped a set of data that does not serve the research direction, such as cases in areas that are not administratively affiliated to the Tabuk region, the type or gender of the victim or his name, to reduce the volume of data that is not used.

**d. Data transformation**

Since data are provided through sources that generate or keep data regardless of data mining procedures, it is necessary to prepare data considering the condition and the given problem for data mining algorithm insertion. To prepare the data, we transformed them from their initial form to a suitable form for the algorithm.

**B. Feature selection**

The establishment of big datasets and their requirements for machine learning techniques is a significant challenge. To address this issue, novel approaches are in demand. Feature selection in machine learning refers to selecting the best features in our data to provide for our model. Where data were collected on daily on infected, recovered and death cases based on the season by used for pandemic prediction modelling. (see Fig. 1) [33].

**C. Classification in data mining**

The classification procedure attempts to find a model that discriminates and describes classes and data concepts. This is a data analysis task.

**a. Classifier models**

This paper uses binary classifiers and basic and hybrid algorithms to improve the prediction’s performance and more accurate diagnosis of COVID-19. In the following section, we briefly explain each algorithm.

**b. Hybrid classification models**

This study proposed a combination of random forest and basic algorithms, such as decision trees, k-means, and neural networks algorithms. The COVID-19 patients’ dataset was divided into training and test datasets utilizing 80% training and 20% testing [34].
D. Review the proposed method

This study proposed a combination of Random Forest and basic algorithms such as Decision Tree, K-means, and Neural Networks algorithm. COVID-19 patients' dataset is divided into training and test dataset utilizing 80% training 20% testing. [34]

We then evaluated the accuracy using Python. We experimented with different combinations of basic and hybrid algorithms (see Fig. 2).

E. The Used analysis tools in Python

Python lets one choose among a huge number of modules that implement the main steps necessary to develop a classification/data mining application (e.g. frameworks, classifiers, clustering algorithms, plotting and visualization analysis of results) [35].

G. Proposed method

In this model, we combined random forest in the stacking part and basic algorithms such as DT, k-means, and neural networks in the stacking algorithm’s fundamental part.
IV. RESULTS

There is a wide range of classification algorithms, each of which has its own strengths and weaknesses. In fact, none of the learning algorithms has the best performance, considering the current supervised learning issues [36]. We evaluated the model to find an optimal solution from various classification models generated through a complicated and repetitive process. Machine learning model evaluation can be intricate. Generally, we divided the dataset into two categories: (i) training dataset (ii) and test dataset. The training dataset was used to train the model, and the test dataset was employed to test the model. We then evaluated the model performance based on error criteria to determine the model accuracy. In Fig. 3, we showed the model and the evaluated results.

<table>
<thead>
<tr>
<th>Class Name</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neural network models</td>
<td>0.79729</td>
</tr>
<tr>
<td>K-mean</td>
<td>0.77027</td>
</tr>
<tr>
<td>Decision Tree Classifier</td>
<td>0.66216</td>
</tr>
</tbody>
</table>

Fig. 3. Proposed model’s evaluated results

B. A comparison of the three algorithms

The party of means k is one of the most straightforward and commonly used unsupervised machine learning algorithms. ‘The purpose of K-means is to find basic trends by grouping related data points together. To arrive at this end, it is an aim to find k classes’ [37], [38].

To refine the locations of the beginning points in data mining, the k-means algorithm begins with a collection of randomly chosen centroids and iterates (repetitions) calculations to optimise the positions of the centroids [39].

A decision tree has a root node, roots and leaf nodes. Each internal node corresponds to an attribute test, each branch corresponds to the product of a test, and each leaf node corresponds to a class name. The root node is the first node in the tree [40].

Neural networks are based on simulations of the way neurons work in the brain that respond to specific signals from dendrites. Each entry is multiplied by its weight, and the multiplying values are added to create the expected sum. The weighted sum is then applied to the activation function, producing the perceptron’s output [37], [38].
In short, decision tree deduction is a way of classifying inference from a training example laid out on a tree structure, rather than the simple assumption on which it is constructed when the algorithm is applied manually, as we did in Fig. 4.

Hence, Decision Tree has lower accuracy than K-means then Neural Networks, therefore Neural Network algorithm that gives better predictive product.

TABLE III. COMPARISON BETWEEN MACHINE LEARNING MODELS

<table>
<thead>
<tr>
<th>Method</th>
<th>Task</th>
<th>Algorithm type</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision tree</td>
<td>Prediction</td>
<td>Supervised</td>
<td>66.21%</td>
</tr>
<tr>
<td>(TD)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K-means</td>
<td>Prediction</td>
<td>Unsupervised</td>
<td>77.02%</td>
</tr>
<tr>
<td>NN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neural networks</td>
<td>Prediction</td>
<td>Supervised</td>
<td>79.72%</td>
</tr>
<tr>
<td>(NN)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
V. DISCUSSION

This study resulted in new findings about the relationship between the COVID-19 virus and the mortality curve regarding weather factors in the Tabuk region, where it was noted that the death rate in summer was higher than in the other seasons, as shown in Fig. 6. However, this does not mean that the weather was the sole reason for this. There may also be other factors, such as noncompliance with the quarantine restrictions at that time [28]. There was a generalisation of study periods on a quarterly basis, and the results taken from the United States in our study contradict these observations around the world because we cannot correlate nearby weather conditions with the number of new cases. In these countries, generalisations by weather factors alone may lead to poor results.

As for the study of India and other countries, the predictions in the results cannot correlate the approximate temperatures of the seasons with the number of new cases in this study. This could be due to the limited number of reported cases [44–46].

Fig. 4. Decision Tree is a structure that includes a root node
These studies are consistent with the results of our study, as we did not clearly see any correlation between weather indicators and conditions in the Tabuk region, perhaps due to the lack of available data and a decrease in research motivation due to imposed restrictions and social distancing, as many scientists had to stop their research during the pandemic.

Also, in this study, some limitations should be acknowledged. Our analysis can be subject to caveats. While the number of incidents was specifically taken from WHO reports, many countries do not report cases and have implemented many public health strategies. In the United States and several other countries, including the United Kingdom and Japan, Korea has conducted comprehensive tests to classify potentially positive COVID-19 subjects, including asymptomatic subjects, for example, and has chosen to test people with symptoms or who have had contact with COVID-19-positive cases. In fact, this may contribute to unrecognized cases; thus, while a community can have a greater proportion of positive cases of COVID-19, they may remain undetected before they are transmitted to the more vulnerable community. Additional climatic conditions, such as sun exposure and cloud cover, were not considered in our study, which may also have played an important role in the spread of COVID-19. Comparative trends can be predicted from these climatic variables because solar and cloud cover correspond to normal temperatures. Moreover, there is still little knowledge about the rates of external versus internal transmission and direct versus indirect transmission. Also, environmental effects are important for outward transmissions. Recently, cases have increased in warm and humid regions due to other reasons, such as social distancing, quarantine steps and cultural traditions, such as demographic differences, healthcare and infrastructure and social policies such as the lockdown process.

Future studies will aim to infer the combined effects of these variables to fully understand transmission dynamics. Where recommend that many countries have implemented a ‘flattening-the-curve’ policy to immediately reduce healthcare stresses, disperse the number of individuals seeking treatment over time, adequately handle patients’ medical needs, and delays the spread of the epidemic throughout the country. Although appropriate quarantine steps contribute to the ‘spreading of the curve’, we agree that warm humid environments in most of Europe and North America will not see a benefit in the coming days, provided that the atmosphere plays a major role in the spread of COVID-19 [47].

We strongly emphasize the need to take appropriate quarantine steps, including in cold and humid areas where the incidence is lower, so that COVID-19 can be easily minimized and defended against spread. Warm weather alone, as in hot and humid countries, will not be enough to avoid the spread of COVID-19 if adequate protective precautions are not taken because we already have evidence that positive cases of COVID-19 are rising in hot and humid areas as well. Besides temperature, the number of infected cases in each region will play a role, including population density, public health policy, political and social institutions, access to healthcare, health intervention and global interdependence. These considerations can be used in future work,
and the association between weather and the transmission of COVID-19 is further investigated using a statistical and epidemiological model.

VI. CONCLUSION

In conclusion, this study provides important information on the impact of weather on the spread of COVID-19. This information can lead to a better understanding of weather parameters and the spread of the virus in the Tabuk region. The results can also help us predict areas at risk for weather-based virus spread in already affected countries and countries with high population density, such as India. Recently, the spread and development of the virus have increased rapidly.

Many studies have been conducted in Saudi Arabia regarding COVID-19, as it ranked first in the Middle East and fourteenth globally. Saudi Arabia was also first in the world in response by the government and entrepreneurs to the coronavirus pandemic [48]. In this study, we used and pre-processed datasets and uploaded them to kaggle [49] and used machine learning algorithms such as decision tree, k-means and neural networks in a stacking classifier. The results showed that neural network algorithms diagnose patients with 80% accuracy. We can improve the performance of the models by increasing the data in the standard format of the COVID-19 dataset and studying and diagnosing cases in general for all regions of Saudi Arabia. The research team recently came up with a quick diagnostic tool for coronavirus, which may help us to study cases faster and more accurately in the future.

VII. REFERENCES


[22] K. Bloom-Feshbach et al., “The George Washington University Health Sciences Research Commons Global Health Faculty Publications Global Health.”.


[32] TABUK City Profile.


[34] World’s Most Vulnerable Countries Lack the Capacity to Respond to a Global Pandemic Credit: MFD/Elyas Alwazir | Office of the High Representative for the Least Developed Countries, Landlocked Developing Countries and Small Island Developing States.


[36] Types of Machine Learning Algorithms You Should Know | by David Fumo | Towards Data Science


[39] [Data Mining]-Clustering algorithm: K-means algorithm understanding and SparkCore implementation - Programmer All.

[40] Data Mining - Decision Tree Induction (tutorialspoint.com).


“The Kingdom is the first in the world in the response of the government and entrepreneurs to the Corona pandemic.”


