Analysis of COVID-19 Database for Defining the Most Important Symptoms

Mohanad A. Salih[1*], Murtadha M. Hamad[2], Wesam M. Jasim[3]

[1*,2,3]*Department of Computer Science, College of Computer Science and Information Technology, University of Anbar, Iraq*

Email: [2]dr.mortadha61@gmail.com, [3]co.wesam.jasim@uoanbar.edu.iq
Corresponding Email: [1]mohanad.a.alhadethy@gmail.com*

Abstract: In Wuhan, China, COVID-19 was found at the end of 2019. The virus quickly spread to every country in the world. In order to offer enough treatment at the appropriate time, implementing detection strategies for patient status (Negative, Positive) is one of the top issues for managers and decision-makers in modern hospitals of all sorts.. Enhancing the standard of medical treatment may help to stop a COVID-19 pandemic. A summary of the health care provided to comparable individuals may be obtained by grouping patients with similar characteristics and symptoms. In most cases, algorithms are employed in medical machine learning. This article creates an electronic questionnaire. It covers every nation in the world and is promoted online utilizing Twitter and other networking sites. In the course of getting ready to prepare COVID-19 data, the raw COVID-19 data including non-applicable occurrences is handled after the feedback interpretation and analysis process is complete, after the questionnaires were distributed, and after all the findings have been collated. The patient state is predicted using these traits and symptoms, either positively or negatively.

Keywords: Covid-19 Data, Symptoms, Analysis, Diagnostic.

1. INTRODUCTION

COVID-19 is a hurricane to the majority of the nations in the world. The globe is working together to combat this pandemic. Sneezing, coughing, chatting, and touch all contribute to the transmission of COVID-19. Therefore, it poses a risk for COVID-19 to spread quickly. A nation's capacity to fight the epidemic could be enhanced by higher-quality healthcare. The situation is made worse by the fact that COVID-19 may live for up to 72 hours in environments that seem normal (J.FX. Susanto Soekiman, et al., 2021)[1].
The best way to increase the COVID-19 survival rate is early discovery, and the work currently

being done can benefit from the use of medical data mining technologies. Medical machine learning is frequently used in the medical field. If patterns of patient data are acquired, it is generally accepted that medical machine learning may address the urgent issues related to COVID-19 (Shailja Gupta, et al., 2020)[2]. Clinicians are assisted in the development of trustworthy diagnostic methods and processes by medical machine learning. Both medical supervised and unsupervised learning are frequently employed in the fields of data science and medical machine learning to address a variety of real-world issues.

(Marco Polignano, et al., 2020)[3]. Clustering patients with comparable COVID-19 symptoms aids in gaining insight into the standard of care given to a range of individuals.

Masaki Kato and colleagues 2020) [4]. Unsupervised machine learning techniques are being used for clustering. Unsupervised medical machine learning algorithms frequently distinguish between data created using vision and unlabeled data contained in the data collection.

(Mohamed Alloghani, et al., 2020; Zfania Tom Korach, et al., 2019). [5] [6] This work provides a real dataset based on patient data to assist researchers in predicting cases of COVID-19 possibilities (negative, positive) using unsupervised medical machine learning techniques.

I.   Related Works

Only two recent studies include the COVID-19 dataset, which was developed to forecast and examine COVID-19 instances.

G. Manimannan, & Co., A. Pompaavai, 2019, Up till March 24, 2020, using secondary sources of data from the Indian Health and Family Welfare Organization, it is attempted in this study to determine the Indian states and Union Territories that have been impacted by COVID-19 [7].According to J. Hutagalung et al. (2020): The categorization of COVID-19 cases and fatalities in Southeast Asia is discussed in this essay. The data was then organized into several clusters using Rapid Miner tools. Data from the WHO for 2020 fatalities in April, national statistics, and COVID-19 confirmed cases have all been utilized [8].

2. METHODOLOGY

In this research a questionnaire electronic is created It includes all countries of the world and Promote it via the Internet using social networking sites and Twitter. According to table 2, the symptoms and traits that were questioned for the training dataset were restricted to personal characteristics and several laboratory test variables. The patient state is predicted using these traits and symptoms, either positively or negatively. In order to get comprehensive patterns about them, the questionnaire was filled out by 3000 patients from various countries. Iraq's Al-Anbar University website hosted the database. You may get it online by clicking this link. https://github.com/mohanadalhadethy/covid-19-database.

Once all of the data was acquired and the questionnaires were delivered, the return analysis and interpretation procedure was complete. The method used to prepare COVID-19 data involves processing the raw COVID-19 data, which includes events that aren't relevant. The pre-cleaning and data processing approaches were used to clean the data set. This was brought on by contradictions and mistakes that needed to be fixed. Data from the COVID-19 was imported into Excel or a csv file for analysis and alterations. During the pre-processing and cleaning of the data, Outlier and missing values are removed from the dataset. Data

generalization is therefore considered to be one of the methods for minimizing data.

Table 1: lists the patients' characteristics and symptoms questionnaire.

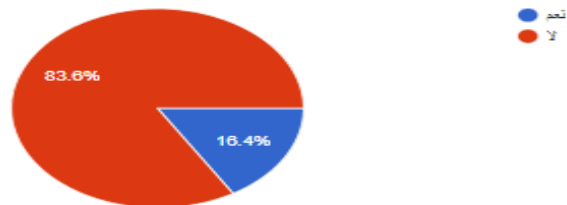| Question NO.	Description	Question NO.	Description
Q1	the patient's age	Q12	Does the Patient Experience Fatigue? (Yes, No)
Q2	Patient's Sexuality	Q13	Does the Patient Have Sputum? (Yes, No)
Q3	Administration of the Sick	Q14	Does the Patient Experience Chest Pain? (Yes, No)
Q4	Nationality of the Patient	Q15	Does the Patient Have Diarrhea? (Yes, No)
Q5	Body temperature of the patient (High, Medium, Low)	Q16	Does the Patient Have Hypertension? (Yes, No)
Q6	The patient may be coughing. (Yes, No)	Q17	Does the Patient Have Expectoration? (Yes, No)
Q7	Is the patient experiencing a headache? (Yes, No)	Q18	Does the Patient Experience Myalgia? (Yes, No)
Q8	Does the Patient Experience Loss of Smell? (Yes, No)	Q19	Does the Patient Experience Gasping? (Yes, No)
Q9	Does the Patient Experience a Loss of Taste? (Yes, No)	Q20	Does the Patient Experience Drowsiness? (Yes, No)
Q10	Does the patient have any recurring illnesses? (Yes, No)	Q21	Anorexia: Does the Patient Have It? (Yes, No)
Q11	Does the Patient Suffer from Vertigo? (Yes, No)	Q22	Patient COVID-19 Status (Positive, Negative) Target Class

3. RESULTS AND DISCUSSION

The results of the questionnaire showed that a number of features are very similar to the symptoms of seasonal flu. In addition, the symptoms differ from one person to another (for example, the temperature of many people with covid-19 did not feel a rise in temperature, unlike other patients who had a high temperature) as well. The results of the questionnaire proved that only 4 out of 22 features included in the questionnaire were the main symptoms by which the patient could be diagnosed, is the result positive or negative, namely (headache by 51%, sense of smell by 52%, fatigue by 76%, loss of appetite by a percentage 56%,). The most susceptible ages are those between 25-35 years, due to the frequent mixing of these ages with
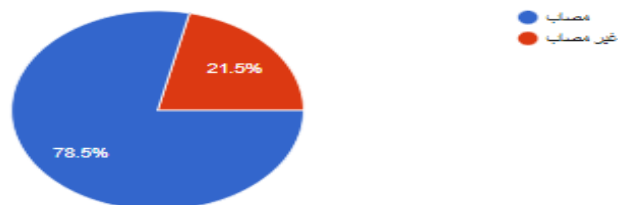
others through work. The questionnaire also proved that the death rate from covid-19 disease was 9%, especially for those who suffer from diseases Chronic (heart, kidney, diabetes) As shown in figure 1.

هل عانى المريض من تورم العين ؟
2,574 ردًا

- نعم
- لا

83.6%

16.4%

ماهي حالة المريض بعد اجراء الفحوصات المختبرية ؟
2,536 ردًا

- مصاب
- غير مصاب

21.5%

78.5%

هل عانى المريض من التهاب الحلق؟
2,567 ردًا

- نعم
- لا

52.2%

47.8%
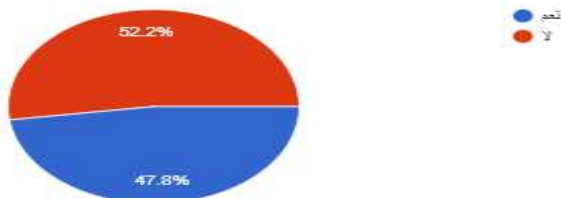
هل عانى المريض من القيء؟
2,573 ردًا

- نعم
- لا

77.3%

22.7%
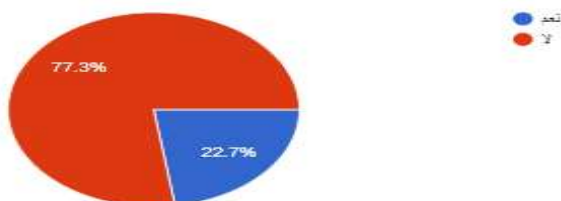
هل عانى المريض من الضيق في التنفس ؟ اذا كان ليس طول فترة احتضان المرض فيرجى الاجابة ب ريما
2,577 ردًا
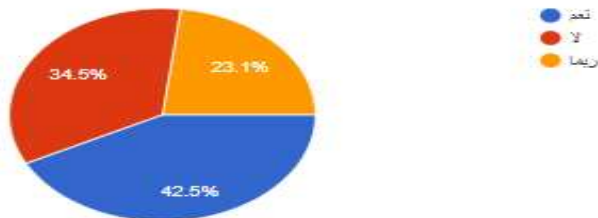
هل عانى المريض من رشح الانف ؟ اذا كان ليس طول فترة احتضان المرض فيرجى الاجابة ب ريما
2,576 ردًا
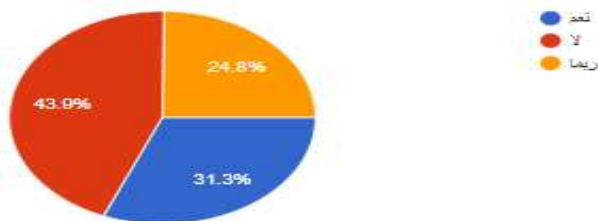
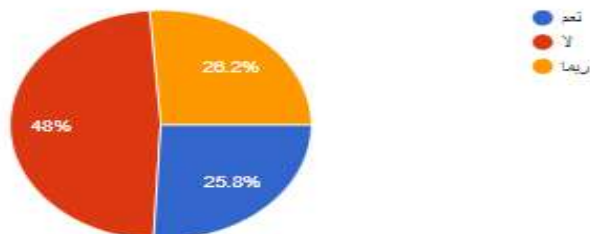هل عانى المريض من القشعريرة؟ اذا كان ليس طول فترة احتضان المرض فيرجى الاجابة ب ريما
2,573 ردًا
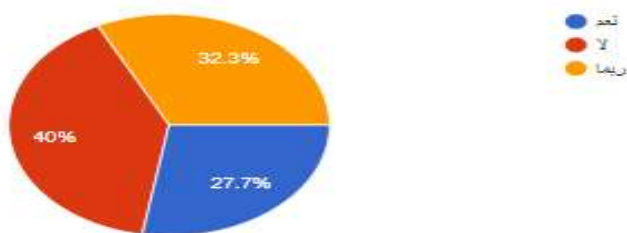
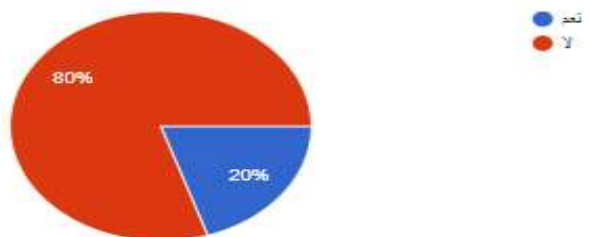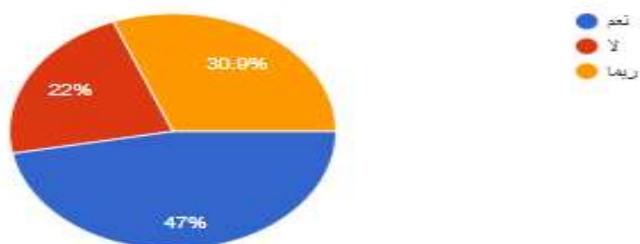هل عانى المريض من البرد ؟اذا كان ليس طول فترة احتضان المرض فيرجى الاجابة ب ريما
2,580 ردًا
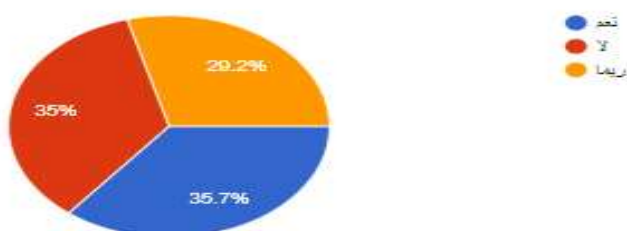
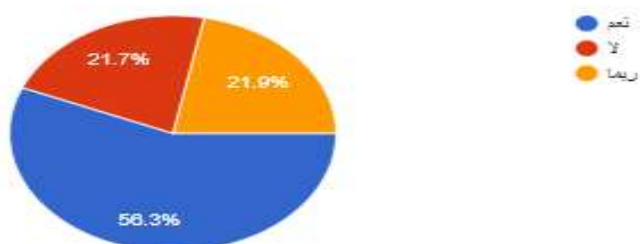هل عانى المريض من الإسهال؟اذا كان ليس طول فترة احتضان المرض فيرجى الاجابة ب ربما

2,585 ردًا

- تعم
- لا
- ربما

26.2%

48%

25.8%

خلال فترة احتضان المرض هل عانى المريض من ارتفاع ضغط الدم؟

2,571 ردًا

- تعم
- لا

80%

20%

هل عانى المريض من نعاس؟اذا كان ليس طول فترة احتضان المرض فيرجى الاجابة ب ربما

2,576 ردًا

- تعم
- لا
- ربما

30.9%

22%

47%

هل يعاني المريض من فقدان الشهية؟اذا كان ليس طول فترة احتضان المرض فيرجى الاجابة ب ربما

2,581 ردًا

- تعم
- لا
- ربما

21.7%

21.9%

56.3%

هل عانى المريض من دوار؟اذا كان ليس طول فترة احتضان المرض فيرجى الاجابة ب ربما

2,588 ردًا

- تعم
- لا
- ربما

36.7%
28.1%
35.2%
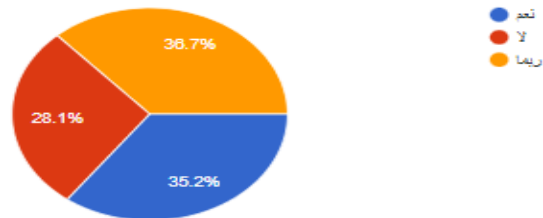
هل عانى المريض من التعب؟اذا كان ليس طول فترة احتضان المرض فيرجى الاجابة ب ربما

2,590 ردًا

- تعم
- لا
- ربما

16.8%
76.2%
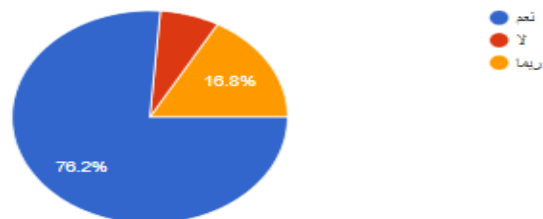
هل عانى المريض من الصداع؟اذا كان ليس طول فترة احتضان المرض فيرجى الاجابة ب ربما

2,589 ردًا

- تعم
- لا
- ربما

32.6%
15.8%
51.6%
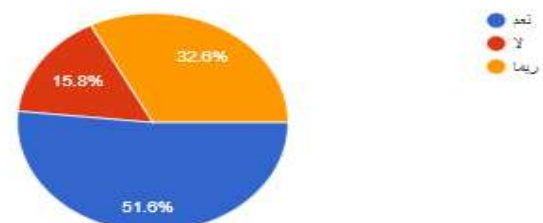
هل عانى المريض من فقدان حاسة الشم؟ اذا كان ليس طول فترة احتضان المرض فيرجى الاجابة ب ربما

2,591 ردًا

- تعم
- لا
- ربما

30.5%
16.7%
52.8%

Figure 1: Statistical percentages of symptoms for covid-19 disease
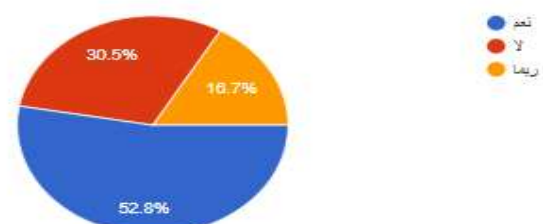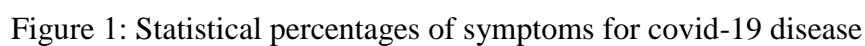
4. CONCLUSION

In this paper, a questionnaire electronic is created It includes all countries of the world and Promote it via the Internet using social networking sites and Twitter. Once the feedback interpretation and analysis process has been finished once the questionnaires have been sent out and all the results have been compiled, the raw COVID-19 data containing non-applicable occurrences is processed as part of the COVID-19 data processing method. The results of the questionnaire proved that only 4 out of 22 features included in the questionnaire were the main symptoms by which the patient could be diagnosed, is the result positive or negative, namely (headache by 51%, sense of smell by 52%, fatigue by 76%, loss of appetite by a percentage 56%,). The patient state is predicted using these traits and symptoms, either positively or negatively.

5. REFERENCES

1. Soekiman, J. F. S., Putranto, T. D., Susilo, D., & Garcia, E. M. A. (2021). Economic Sector during the COVID-19 Pandemic: Indonesian Instagram Users Behaviour. Webology, 18(1).
2. Gupta, S., Kaur, M., Lakra, S., & Dixit, Y. (2020). A Comparative Theoretical and Empirical Analysis of Machine Learning Algorithms. Webology, 17(1).
3. Polignano, M., Suriano, V., Lops, P., de Gemmis, M., & Semeraro, G. (2020). A study of Machine Learning models for Clinical Coding of Medical Reports at CodiEsp 2020. In Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings.
4. Kato, M., Asami, Y., Wajsbrot, D. B., Wang, X., Boucher, M., Prieto, R., & Pappadopulos, E. (2020). Clustering patients by depression symptoms to predict venlafaxine ER antidepressant efficacy: Individual patient data analysis. Journal of Psychiatric Research, 129, 160-167.
5. Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. Supervised and Unsupervised Learning for Data Science, 3-21.
6. Korach, Z. T., Cato, K. D., Collins, S. A., Kang, M. J., Knaplund, C., Dykes, P. C., & Zhou, L. (2019). Unsupervised machine learning of topics documented by nurses about hospitalized patients prior to a rapid-response event, 952-963.
7. Poompaavai, A., & Manimannan, G. (2019). Clustering study of Indian states and union territories affected by coronavirus (COVID-19) using k-means algorithm. International Journal of Data Mining And Emerging Technologies, 9(2), 43-51.
8. Hutagalung, J., Ginantra, N. L. W. S. R., Bhawika, G. W., Parwita, W. G. S., Wanto, A., & Panjaitan, P. D. (2021, February). COVID-19 Cases and Deaths in Southeast Asia Clustering using K-Mean Algorithm. In Journal of Physics: Conference Series (Vol. 1783, No. 1, p. 012027). IOP Publishing.