Predicting the Possibility of COVID-19 Infection using a Fuzzy Logic System

Shadab Hafiz Choudhury	Tanbin Akter Mitaly	Azmary Zannat Aurin	Sabbir Mollah	Abdullah Al Rafi
1631335042	1712438642	1621233042	1712490642	1731371642

Abstract *– Diagnosing a disease is a difficult task for medical professionals. With such a high infection rate, the COVID-19 pandemic requires a fast way of diagnosing patients. As test kits are not always available, often preliminary diagnosis has to be carried out purely based on the patient's pre-existing conditions and the symptoms they are currently experiencing. A Type-2 Fuzzy Logic system has been designed here that will help provide a preliminary diagnosis on how likely the symptoms a patient is suffering from are due to a COVID-19 infection. Both statistical data and Rule Base information were collected from publicly available databases and datasets. While the Fuzzy Logic system is fully functional, it is not accurate enough for use in medical contexts. Rather, it is aimed at the general populace who can use it as an aid in determining whether or not to seek out proper medical care based on the likelihood of being infected.*

1. Introduction

The Coronavirus (COVID-19) first originated in Wuhan in the Hubei Province of China. Currently, it is a worldwide pandemic with more than 6 million confirmed cases around the world. As it is a highly contagious virus with both airborne and contract transmission vectors, detection of COVID-19 is paramount in order to combat this threat.

Artificial Intelligence has a long history of applications in medical diagnosis. Diagnosing medical diseases is a difficult process that requires one to consider a staggering variety of parameters, ranging from the patient' symptoms to their medical history and the presence of various factors in the body. It is possible to develop Artificial Intelligence systems that can account for all the relevant factors. One approach for creating such a system is using Fuzzy Logic.

A Fuzzy Logic System (FLS) is a form of AI based on the theory of Fuzzy Sets developed by L. A. Zadeh in 1965 [1]. It focuses on the relationship between language and numbers. The core concept of Fuzzy Logic revolves around Fuzzy sets, which assumes that all entities have a degree of membership in a particular universe. The degree of membership is a continuous range between 0 and 1, where 0 means it is not a member of the universe and 1 means it is fully a member. Intermediate values indicate partial membership.

There is generally a large degree of vagueness present in medical diagnosis. In many cases, only partial information is available. In other cases, the available information is vague and generally unusable for most forms of AI. However, Fuzzy Logic is highly suitable for working with this kind of information. For instance, if a patient reports they have a 'Slight Cough', an FLS-based diagnosis system can use 'Slight' as a linguistic variable rather than requiring a numeric variable. An FLS uses a set of IF-THEN rules to evaluate inputs and produce an output. The set of IF-THEN rules is referred to as a Rule Base.

One of the shortcomings of a basic FLS is that it becomes inaccurate if there is any uncertainty in the Membership functions. In order to account for cases where the exact membership functions are not known, Type-2 Fuzzy Logic Systems was further developed by Zadeh in 1975 [12]. In a Type-2 FLS, membership functions are considered three dimensional. They can be represented as a two-dimensional membership curve, with an Upper Membership Function and a Lower Membership Function [11]. For an example of a Type-2 Membership Function, refer below to Figure 4.3.

Using a Type-2 Fuzzy Logic System allows us to represent the uncertainties in the Membership Functions themselves. This is particularly useful for making this system usable by general people rather than medical professionals. FLSs have been proven to be effective. The goal of this project was to develop an FLS that could be used by anyone to get an idea of how likely it is that they are affected by COVID-19 and so should seek proper medical care.

The project was developed entirely in Python, using the PyIT2FLS package to build the Type-2 Fuzzy Logic System. [1]

2. Related Work

A large amount of work has been done so far in the application of Fuzzy Logic in medical diagnosis. Type-1 Fuzzy Logic Systems have been used in the diagnosis of Heart Disease [3], for example. In this paper, a large amount of crisp data was available, so a Type-1 Fuzzy Logic System was used where the inputs included various factors that contribute to heart disease. Symptoms like high blood pressure or high cholesterol were taken into account by the Rule Base. This paper is an extremely straightforward implementation of Fuzzy Logic in medical diagnosis and highlights the effectiveness of this technique.

Other examples of the successes of using Fuzzy Logic in medical diagnosis include the diagnosis of Cervical Cancer by Awotunde and Matiluko in 2014 [4], where a significant number of patients were accurately diagnosed. Fuzzy Logic has also been used to diagnose other forms of cancer, such as Lung Cancer [5].

However, the techniques used in those papers cannot be directly replicated in the context of COVID-19, due to a lack of data. A closer approach can be seen in the use of Type-2 Fuzzy Logic Systems to diagnose most common diseases by Erin and Abiyev in 2019 [6]. This paper targets a much broader spectrum of diseases and their diagnosis, using Type-2 FLS to order to account for the increased uncertainty in the membership functions. They were able to diagnose the possibility of a Common Cold and of Flu by using a total of five contributing factors or symptoms as input.

For COVID-19 related research, most studies have focused on the epidemiological aspects, studying the spread of the virus. Other research has focused on isolating the genetic characteristics of the virus and the effectiveness of drugs and treatment procedures. Not much research has been done on the use of artificial intelligence to analyze and diagnose the visible symptoms of COVID-19.

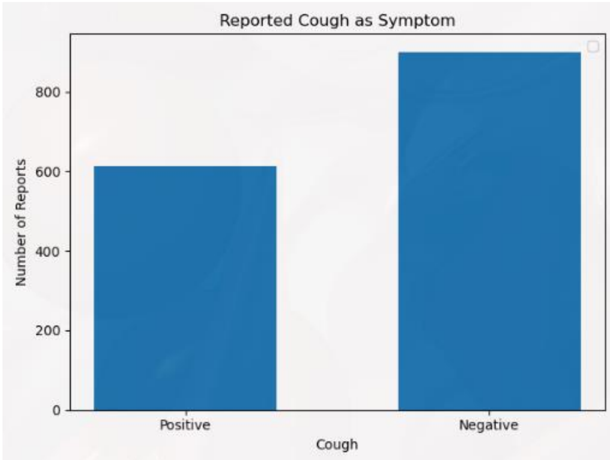
3. Data Collection and Preprocessing

As this project is based on a situation that is currently developing and evolving, a wide breadth of data was available. Most of the data was focused on epidemiology, however, and relatively little detail was available on the exact symptoms experienced by the patients themselves.

The primary goal during data collection was to find data that correlated the onset of symptoms and the severity of those symptoms with the possibility of an individual contracting the highly infectious COVID-19. However, these efforts were stymied by the lack of information on severity. Information such as the exact level of fever or the respiratory rate was kept private by the hospitals, due to it being confidential medical information belonging to their patients.

However, there were many medical research papers published throughout the last few months, which were collated into the COVID-19 Open Research Dataset Challenge [7]. These papers contained data about the spread of the epidemic, the symptoms reported by the patients, the severity of those symptoms, factors contributing to severe cases [8], and the efficacy of various medical treatments and measures. These papers were essential for building the Rule Base of our FLS.

In order to build our membership functions, we instead looked at datasets containing data about known infections and the general symptoms reported. We found two such data sets [15, 16], containing information such as date of admission, geographical location, age and reported symptoms. By preprocessing and charting the data, we were able to reach conclusions regarding the most common symptoms reported. After combining the two datasets and removing duplicate tuples and unnecessary tuples (the symptoms reported by those who tested negative for COVID-19), we had a sample size of 1514 patients. These 1514 patients all reported positive for COVID-19. While data regarding those who showed symptoms but were tested to be negative for COVID-19 would have been extremely useful, such data was not available.



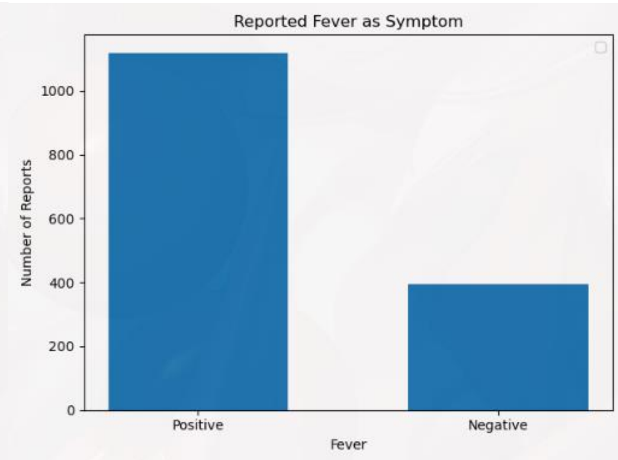
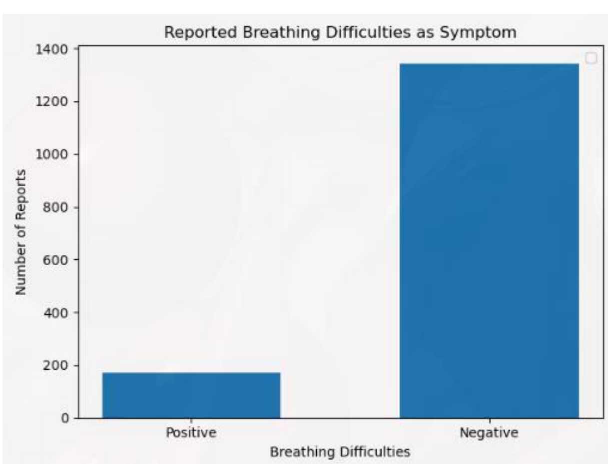


Fig: 3.1: Cough as Symptom

Fig. 3.2: Fever as Symptom



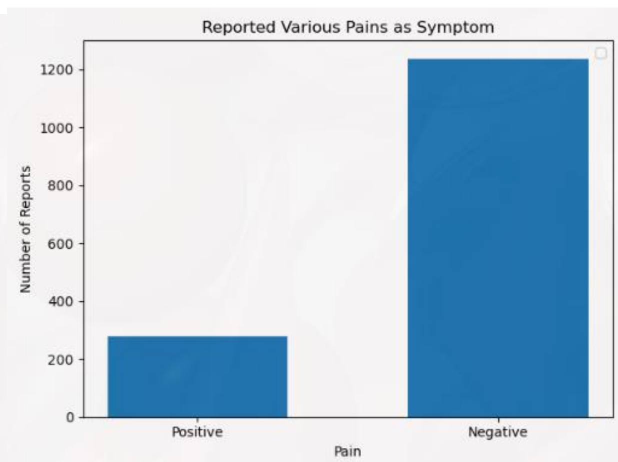


Fig: 3.3: Breathing Difficulties

Fig. 3.4: Muscle, Chest and other Pain

4. Design of Fuzzy Logic System

There were three main stages in the design of our FLS: Fuzzification, or determining the inputs and the membership function for each of the antecedents; Creation of a Rule Base; Defuzzification, or evaluating the outputs of the system.

4.1. Determining Membership Functions

A wide variety of symptoms were reported by individuals suffering from COVID-19. We combined all the symptoms and contributing risk factors into four general categories: Cough, Fever, Breathing Difficulty and Additional Risks.

Lacking concrete information about the severity of each symptom, we were forced to use a roundabout approach making the primary assumption that the patients reported a specific symptom only if it was severe enough. In general, our membership functions were constructed by extrapolating information from both other papers and datasets.

All of the membership functions are within the domain [0, 10]. This domain represents the input from a sliding scale the user can enter. The table given below shows the approximate domains of each of our linguistic variables. Note that these domains are separated into the Upper and Lower Membership Functions. Additionally, the standard deviations of the gaussian functions are given here. With this information the mathematical information of the actual curves can be determined.

Inputs	**Linguistic Variable**	**UMF Range**	**LMF Range**	**MF Shape**	**Standard Deviation**
Coughing	Cough is Negative	[0, 7]	[0, 2]	Trapezoidal	NA
	Cough is Positive	[5, 10)	[8.5, 10]	Trapezoidal	NA
Fever	Low Fever	[0, 4]	[0, 2]	Gaussian	1.0
	Moderate Fever	[1, 9]	[3, 7]	Gaussian	1.0
	High Fever	[6, 10]	[8, 10]	Gaussian	1.0
Breathing Difficulties	Low Difficulty	[0, 4]	[0, 2]	Gaussian	1.0
	Moderate Difficulty	[1, 9]	[3, 7]	Gaussian	1.0
	High Difficulty	[6, 10]	[8, 10]	Gaussian	1.0
Additional Risks	Low Additional Risk	[0, 8]	[0, 4]	Gaussian	2.0
	High Additional Risk	[2, 10]	[6, 10]	Gaussian	2.0

Output					
Overall (Likelihood of being infected with COVID-19	Unlikely	[0, 4]	[0, 1.5]	Gaussian	1.0
	Likely	[2, 10]	[4, 8]	Gaussian	1.0
	Extremely Likely	[6, 10]	[8, 10]	Gaussian	1.0

As evident from the table, this is one of the strengths of using a Type-2 Fuzzy Logic System. Inaccuracies caused due to incomplete data can be accounted for better.

As there are no exact figures available to us, we used Gaussian membership functions, which are generally suitable for working with approximate values. Each of the membership functions for our antecedents and consequents is given here.

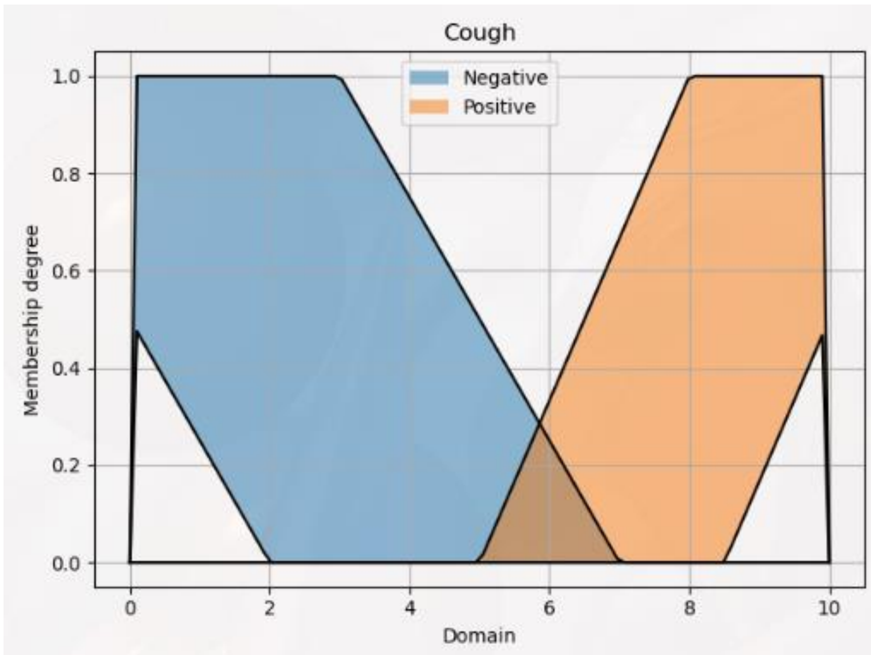


Fig: 4.1.1: Membership Functions for the Input Coughing

For Cough, our data tells us that about 40% of it reported it as a symptom, and 60% did not. By taking the assumption we made earlier, we determined that by our assumption, the upper 40th percentile of Cough severity would account for the patients who reported coughing as a symptom. This allowed us to build Trapezoidal membership functions for both the linguistic variables for this input.

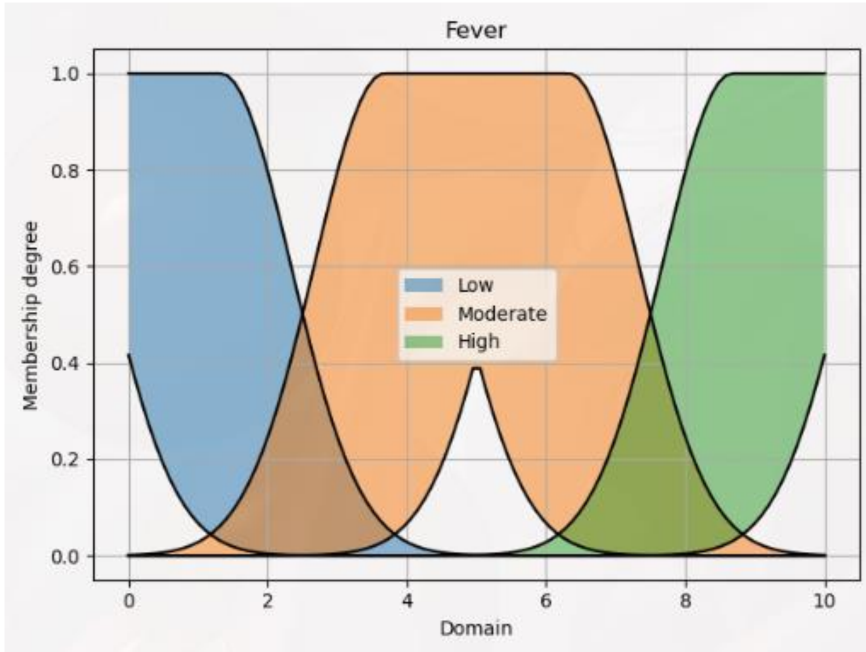


Fig. 4.1.2: Membership Function for the Input Fever

For Fever, one thing to note is that unlike coughing, we can actually map the linguistic variables to approximate values. We could note that a body temperature between 98.6 °F and 99.5 °F is a low Fever, up to 102 °F is a moderate fever and above that is a strong fever. However, we did not implement this because we wanted to keep our input domain consistent. We leave it up to the user of this FLS to carry this out.

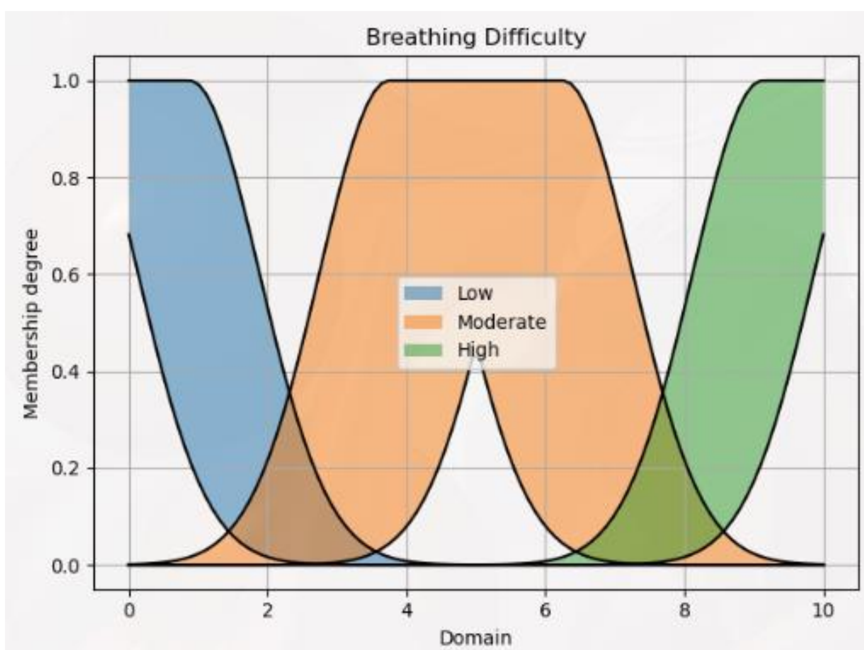


Fig. 4.1.3: Membership Functions for the input Breathing Difficulty

As breathing difficulty has been reported as one of the major symptoms of COVID-19, we made sure to account for that in our membership function. As can be seen, the areas of the Low and High ends are relatively narrow compared to the area for Moderate breathing difficulty. The reason is that the average user will likely be able to distinguish between experiencing very little or no breathing difficulty, or extreme difficulty, but it's harder to say an exact value for 'Moderate difficulty'.

Finally, Additional Risks were one of the major difficulties. The risk and danger of COVID-19 infection is magnified by various comorbidities such as cardiovascular issues or respiratory issues. We picked out the most common of these comorbidities: Hypertension [9], Diabetes, other Cardiovascular Issues, existing Respiratory Issues, and Immunological Issues. Taking into account Age and Environment as well, which have been explored in some medical papers [10], we built a secondary function that takes the input of these conditions and generates a value for Additional Risks.

Under 10 years of age	0.5	Has Hypertension	1.5
Over 50 years of age	0.5	Has Diabetes	1.5
Over 60 years of age	1.0	Has prior cardiac/heart issues	1.0
Over 80 years of age	1.5	Has prior respiratory issues.	3.0
Lives in polluted environment	1.0	Has prior immunological issues.	4.0

It should be noted that this calculation is relatively inaccurate, and further research is required to narrow down the values into what can be considered clinically acceptable.

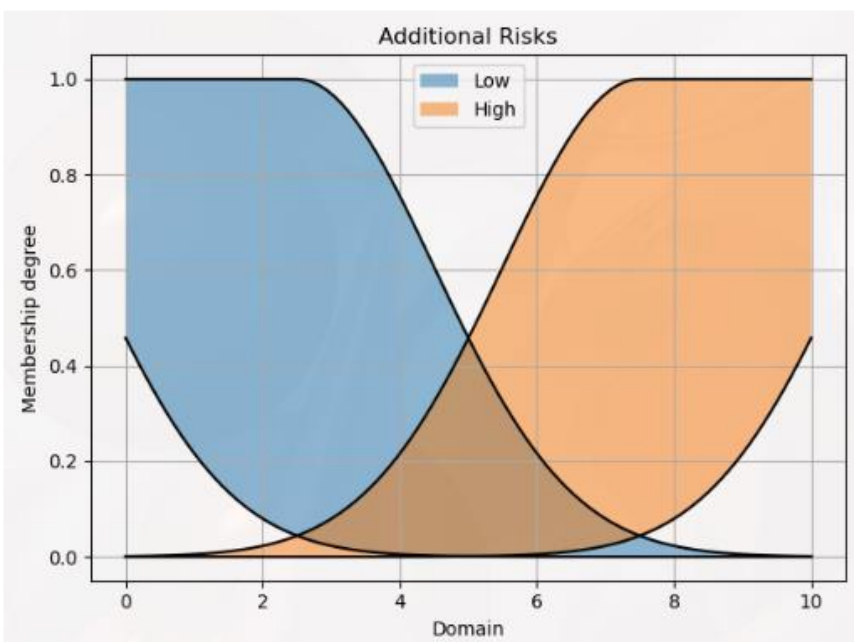


Fig. 4.1.4: Membership Functions for the input Additional Risks

Since our input of Additional Risks is an approximate calculation, the Membership function reflects that by having very large range curves.

Finally, we have our output Membership function. Note that the domain of the variable 'Likely' is biased to the right - this is in order to account for the fact that someone who is affected is more likely in general to show a symptom than someone who is not.

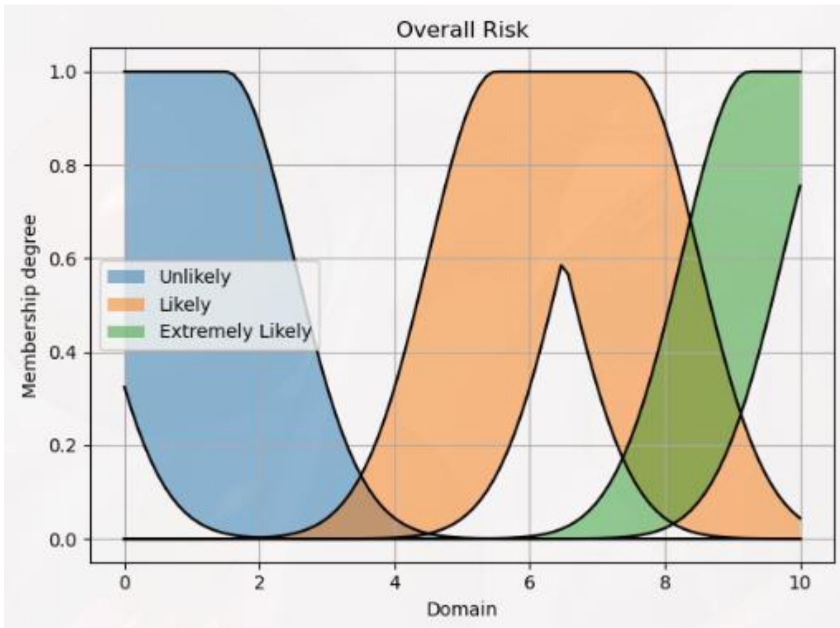


Fig. 4.1.5: Membership Functions for the Output Overall Risk

4.2. Rule Base

For the rule base, we built a total of ten rules governing the Fuzzy Logic System. While the system would have been much more accurate with a larger rule base, we did not have enough Inputs or Linguistic Variables to generate a larger rule base. The rule base is given below as a table:

Inputs				Output
Cough	**Fever**	**Breathing Diff.**	**Additional**	**Risk**
Negative	Low	Low	Low	Unlikely
Positive	Moderate	Low	Low	Unlikely
Negative	High	Low	Low	Unlikely
Negative	High	Low	High	Unlikely
Negative	Low	High	Low	Likely

Negative	High	Moderate	Low	Likely
Positive	Moderate	Moderate	High	Very Likely
Positive	Low	High	High	Very Likely
Positive	Moderate	Moderate	High	Very Likely
Positive	High	High	High	Very Likely

4.3. Defuzzification

Defuzzification of the output set is carried out by the Centroid Method using an Enhanced Karnik-Mendel Algorithm [13]. The Centroid Method [14] evaluates an output membership function by taking the α-cuts of the final output membership form. It evaluates the centroids of each of the planes created by the α-cuts and calculates the overall centroid by taking the unions of all these centroids.

Initially defuzzification converts the Type-2 output Membership Function into a Type-1 Function. An example generated of the Type-1 Membership function for the linguistic variable “Very Likely” is given here.

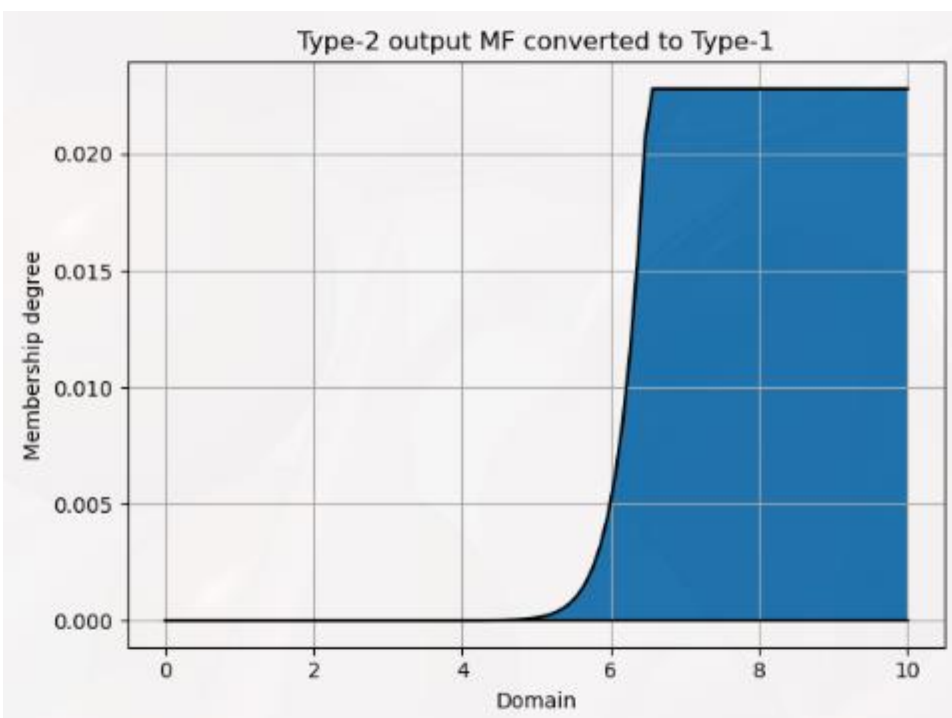


Fig. 4.3.1: Type-1 MF converted from the output Type-2 MF

From the Type-1 MF, we can calculate a crisp value for the output by using the centroid method to defuzzify this. The crisp value here is 8.

5. Evaluation

Due to the issues outlined earlier in Section 3, it should be noted that this FLS is not particularly accurate, especially towards the middle of the range. We categorized the results in three different ways: accurate, vague and inaccurate.

Accurate output is what is expected. For instance, entering the symptoms of someone who is confirmed to be suffering from COVID-19 should return a high value as the output, above 60%. Outputs Between 40% and 60% are considered vague, as the system appears to lack enough information to clearly give an answer. Below 40% is considered accurate for those who are not infected, but inaccurate for those who are (false negatives).

As our data contains only discrete values for a symptom rather than precise continuous values, the evaluation was carried out by randomizing values about the discrete linguistic variable. For instance, each instance of 'Pneumonia' or 'Acute Respiratory Distress' was assigned a random value between 7 and 10 for 'Breathing Difficulty', falling under the linguistic variable 'High'. 'Shortness of Breath' was assigned values between 1 and 9, falling under 'Moderate' for Breathing Difficulty. Patients who did not report any symptoms related to breathing difficulty were given random values between 0 and 3, signifying a very little or negligible amount.

We took a sample of 125 patients from our datasets who had relatively complete data, consisting of symptoms, preexisting conditions, and age. Due to the randomization, the evaluation process was run 20 times in order to get an average value. The evaluation was carried out using a separate python script than the main program.

Finally, in order to compare and justify our use of a Type-2 FLS instead of a Type-1 FLS, a Type-1 FLS system was developed using similar antecedents and consequents, membership functions and rules. The Type-1 system was put through the same evaluation process as the Type-2 FLS.

FLS	Accurate	Vague	Inaccurate
Type-1	0.10%	98.0%	0.10%
Type-2	14.4%	80.0%	5.60%

While both systems are fairly inaccurate, the Type-1 FLS was unable to consistently generate any kind of acceptable results in comparison to the Type-2 system.

The current system was implemented into a basic front-end. The front-end application, in its current form, is a very simple desktop application built in Tkinter that asks the user to enter their symptoms through sliding scales for intensity of coughing, fever or breathing difficulty, and checkboxes for preexisting conditions. This kind of layout is intuitive and easily achievable on any system.

The UI is intentionally kept simple so that people with no medical knowledge can make full use of it, and the antecedents of the FLS were designed with this in mind.

The calculation is processed quickly, returning the results with negligible delay.

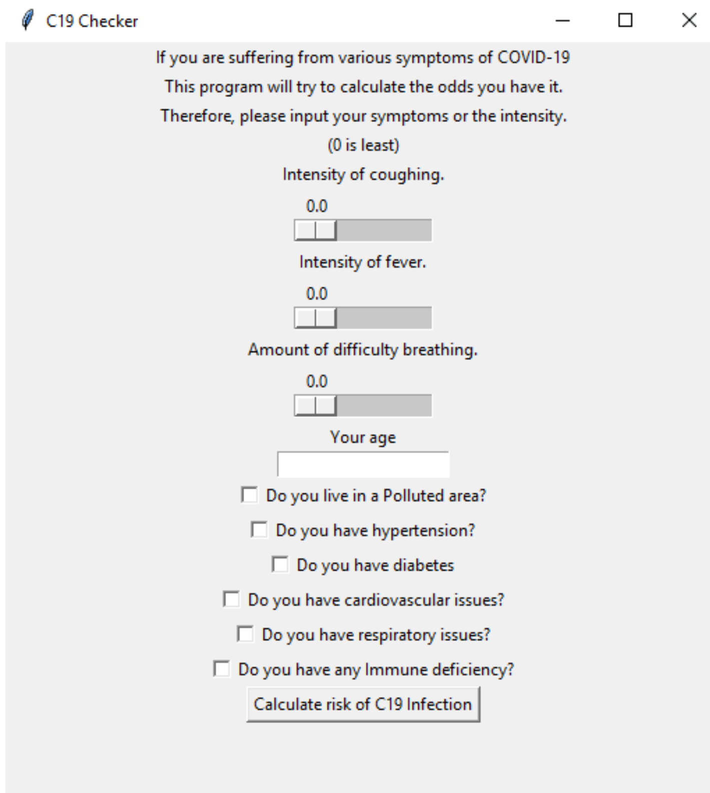


Fig. 4.5: User Interface for using the FLS.

6. Conclusions and Future Work

Overall, the project was successful in developing the FLS for its purposes. However, the main concerns remain about the accuracy of the results of this Fuzzy Logic System, as it was developed using incomplete and relatively inaccurate data. With higher quality data a more effective system with a larger rule base and more accurate membership functions should be possible.

In terms of its basic goals, this FLS is capable of taking a user's input and evaluating whether the user is likely or unlikely to be affected by COVID-19. However, it is inaccurate, and will generally not be of huge use to the general public at this stage.

As shown by the evaluation process, the system returns an accurate result less than 1/5 of the time, and give an unhelpful value when it does return them. Further work is also needed to refine the rule base so that it can provide accurate values.

In the future, once the FLS has been made more accurate, the front-end could also be developed further and built into a website or mobile app for ease of use and distribution.

6. References

[1] A. A. Haghrah and S. GhaemiPyIT2FLS: A New Python Toolkit for Interval Type 2 Fuzzy Logic Systems, 2019, arXiv:1909.10051

[2] L. A. Zadeh. Fuzzy Sets, Information and Control, Vol. 8 (1965), 338-353

[3] Hasan Kahtan, Kamal Z. Zamli, Wan Nor Ashikin Wan Ahmad Fatthi, Azma Abdullah, Mansoor Abdulleteef, and Noor Shahaiyusniezam Kamarulzaman. 2018. Heart Disease Diagnosis System Using Fuzzy Logic. In Proceedings of the 2018 7th International Conference on Software and Computer Applications (ICSCA 2018). Association for Computing Machinery, New York, NY, USA, 297–301. DOI: https://doi.org/10.1145/3185089.3185118

[4] Awotunde J.B., Matiluko O.E. and Fatai O.W.2014. Medical Diagnosis System Using Fuzzy Logic, African Journal of Computing & ICT, 7(2), 99–106

[5] Besime Erin and Rahib H. Abiyev. 2019. Diagnosis of Common Diseases Using Type-2 Fuzzy System. In Proceedings of the 3rd International Conference on Machine Learning and Soft Computing (ICMLSC 2019). Association for Computing Machinery, New York, NY, USA, 239–243. DOI: https://doi.org/10.1145/3310986.3311028

[6] K. Lavanya, M. S. Durai, and N. C. S. N. Iyengar, "Fuzzy rule based inference system for detection and diagnosis of lung cancer," International Journal of Latest Trends in Computing (EISSN: 2045-5364) Volume, vol. 2, 2011

[7] COVID-19 Open Research Dataset Challenge (CORD-19), Allen Institute for AI, https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

[8] Incidence, clinical characteristics and prognostic factor of patients with COVID-19: a systematic review and meta-analysis Xianxian Zhao, Bili Zhang, Pan Li, Chaoqun Ma, Jiawei Gu, Pan Hou, Zhifu Guo, Hong Wu, Yuan Bai medRxiv 2020.03.17.20037572; doi: https://doi.org/10.1101/2020.03.17.20037572

[9] Hypertension in patients hospitalized with COVID-19 in Wuhan, China: A single-center retrospective observational study. Zhenhua Zeng, Tong Sha, Yuan Zhang, Feng Wu, Hongbin Hu, Haijun Li, Jiafa Han, Wenhong Song, Qiaobing Huang, Zhongqing Chen. medRxiv 2020.04.06.20054825; doi: https://doi.org/10.1101/2020.04.06.20054825

[10] Coronavirus Disease-19: The First 7,755 Cases in the Republic of Korea. Osong Public Health Res Perspect. 2020;11(2):85-90. Published online April 30, 2020. DOI: https://doi.org/10.24171/j.phrp.2020.11.2.05

[11] O. Castillo, P. Melin, J. Kacprzyk and W. Pedrycz, "Type-2 Fuzzy Logic: Theory and Applications," 2007 IEEE International Conference on Granular Computing (GRC 2007), Fremont, CA, 2007, pp. 145-145, doi: 10.1109/GrC.2007.118.

[12] L. A. Zadeh, "The Concept of a Linguistic Variable and Its Application to Approximate Reasoning–1," Information Sciences, vol. 8, pp. 199–249, 1975

[13] Wu, Dongrui & Mendel, Jerry. (2007). Enhanced Karnik-Mendel Algorithms for Interval Type-2 Fuzzy Sets and Systems. IEEE T. Fuzzy Systems. 17. 184 - 189. 10.1109/NAFIPS.2007.383834.

[14] N. N. Karnik and J. M. Mendel, "Centroid of a type-2 fuzzy set," Inf. Sci., vol. 132, pp. 195–220, 2001.

[15] Novel Coronavirus 2019 Dataset, Kaggle. https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=covid_19_data.csv

[16] Wolfram Research, "Patient Medical Data for Novel Coronavirus COVID-19" from the Wolfram Data Repository (2020). https://doi.org/10.24097/wolfram.11224.data