Covidnator

Nakul Gupta(nakul.gupta@sjsu.edu)
Huang-Kai Hsu(huang-kai.hsu@sjsu.edu)
Hyesung Ko(hyesung.ko@sjsu.edu)
Huyen Nguyen(huyenng1220@@gmail.com)

Charles Davidson College of Engineering @San Jose State University One Washington Square, San Jose, California 95192-0080, USA

Abstract — The proposal of this project is to analyze the COVID-19 survey dataset that is collected by Carnegie Mellon University (CMU) and University of Maryland (UMD). The main target feature that we would like to analyze is how COVID-19 positive cases relate to the contact people make. More specifically, if people have minimal and essential only contacts such as for work and grocery, how likely they will get COVID-19 compared to people that had contacts with positive cases and social gathering etc. Moreover, we would like to see how tobacco impacts differently on the group of ages for male and female with COVID-19. Based on the analysis of level of contacts, we will provide insights on what businesses are more vulnerable than others.

Index Terms— *Covid-19, Tobacco, Machine Learning, IoT, Data Analytics,*

INTRODUCTION

Since the Covid-19 started spreading around the world, the spreading became faster. According to the Center for Disease Control and Prevention(CDC), between 100K and 200K people got new positive results from the test and between 500 and 2500 people have died due to the Covid-19 everyday[1]. Since the vaccine for Covid-19 is not available at this moment, what people can do is prevention of Covid-19 spreading. For example, people wear masks, keep the distance among others, or stay home. On the other hand, in addition to developing the vaccine, many researchers have been researching the factors which can lead to prevention of spreading of Covid-19. For example, Zhang et al, identified that airborn is the main route of the spreading of the Covid-19[2], which is why wearing a mask is the most effective way to stop the spread so far. Our study focuses on how smoking cigarettes and age affect the Covid-19 symptoms and detection of wearing masks with IoT devices.

TECHNOLOGY

A. Technology use in the Covid-19 analysis
    a. Jupyter Notebook
    b. Scikit Learn Library
    c. Pandas Library
    d. Google Data Studio
    e. Seaborn Library
B. Machine Learning Models
    a. XGBoost
    b. Random Forest Tree
    c. Logistic regression
C. Covidnator System Future implementation needs approximately million-dollar funding in order to develop a mature software system to prevent future flu. Such as COVID The software system contains machine learning, deep learning, Internet of Things, and Artificial Intelligence. Thankfully, the open-sources have made it possible for scholars and research engineers to create prototype systems and structures. In our object detection system, we are using technologies like Tensorflow, Python, OpenCV, and Tensorflow Models. Tensorflow is an open-source library to do machine learning, and it is maintained by Google. This library is enriched, and it is used in many enterprise companies to do research. Such as Facebook, NVIDIA, and more. In order to write efficient code, we are choosing Python3.7 or above because it is the most current and stable version according to the machine learning community. The sample code that is provided on the Tensorflow document is in Python. Moreover, many Python libraries like numpy, matplotlib, and pyvisa made the data analysis process sample to do. OpenCV is a library of programming functions mainly aimed at real-time computer vision, so it is used in the prototype. Lastly, the Tensorflow Models is an open-source community for small to train their system.

DATASET

A. Percentage of population of smoking cigarette in US by states
    a. This dataset provides the percentage of population of each state. Through this study research, we analyze if smoking cigarettes is related to Covid-19 or it causes severe symptoms compared to the non-smoking population.
B. US Covid-19 positive cases and death
    a. This dataset provides the general information of Covid-19 cases, such as positive cases and death cases of each state in the US.
C. Covid-19 hospitalization and pre-condition dataset
    a. This dataset provides the information of the patients who have been hospitalized related to the Covid-19. The dataset include pre-condition of the patients, such as obesity, asthma, tobacco usage, hypertension, sex, and age, date of hospital entry end died.
D. Medium age of each states in US
    a. This dataset includes the medium age of each state in the US. With the medium age, we can provide the relationship between medium age and death rate as well as the relationship between medium age and percentage of smokers.

VISUALIZATIONS

The data exploration has been carried out using Google Data Studio. The datasets provided by the Carnegie Mellon University captured from the Facebook survey and the tobacco information supplemented with COVID-19 details were loaded into the Data Studio. Below are some correlations that were established based on the data provided.

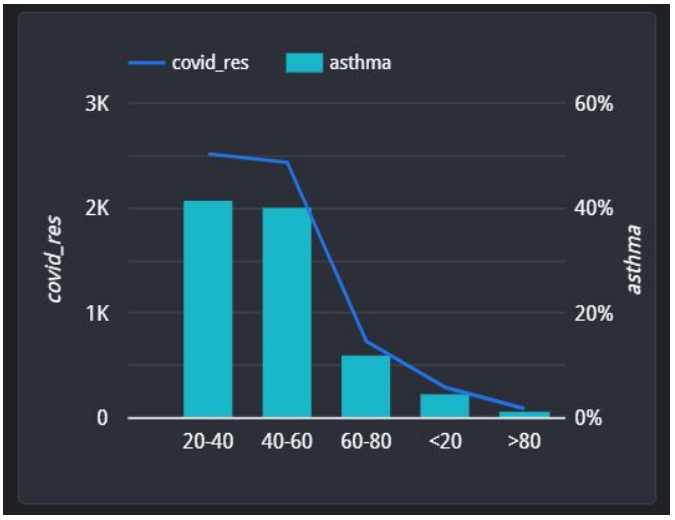


Fig 1. Correlation of asthma to COVID-19 for different age groups

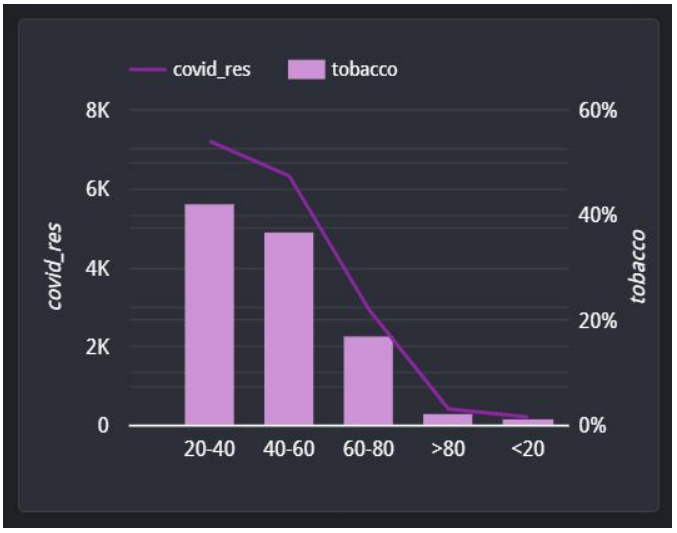


Fig 2. Correlation of tobacco to COVID-19 for different age groups

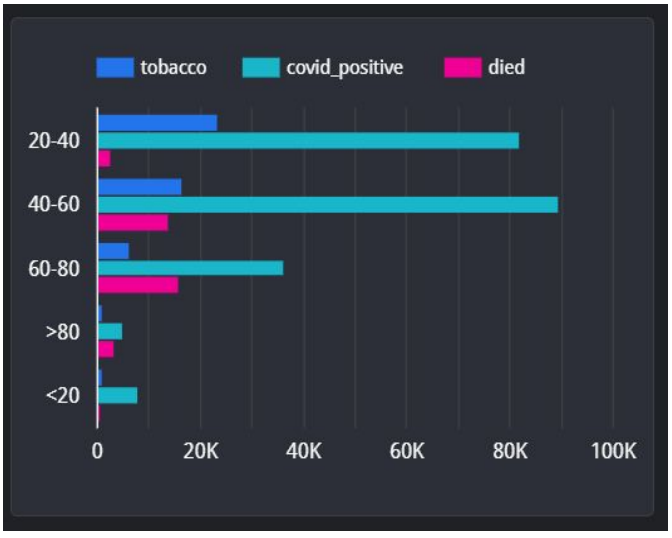


Fig 3. Percentage of tobacco consumer having COVID-19 that died

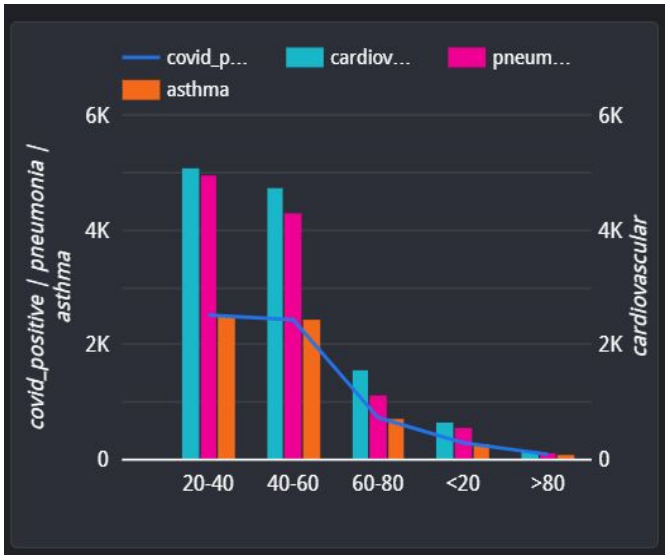


Fig 4. Correlation of other diseases with COVID-19

The initial analysis of data was able to establish that the people consuming tobacco are prone to get COVID-19. This becomes more prominent in the age group of 20-60 and especially in males.

ANALYSIS

In this study, we tried to analyze the impact of tobacco with Covid19 patients. We used that dataset given from the Government of Mexico, which provides patients' history. For patients that use tobacco, the percentage of all tobacco user patients got hospitalized is 33% while the patients that are not tobacco users is 31%. The percentage of tobacco user patients die is 14% while the non-smoking patients die percentage is 12%. The percentage of tobacco user patients that were admitted to ICU is 2.7% while it's 2.6 for non-smoking patients. We also compared the effects of tobacco vs patients that have diabetes as pre-condition. The result was that diabetes has a lot more impact on the rate of hospitalization, death and ICU vs tobacco.

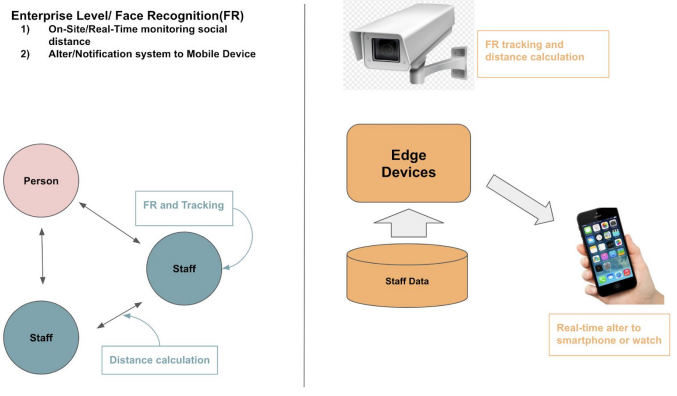
	Non-Smoking Patients	Tobacco User Patients	Patients With Diabetes	Tobacco User with Diabetes Patients
Die Percentage	12.1%	13.8% • Female: 2.2% • Male: 11.5%	27.6%	29.5%
Hospitalized Percentage	30.7%	32.7% • Female: 6.1% • Male: 26.6%	57.9%	60.1%
ICU Percentage	2.6%	2.7% • Female: 0.3% • Male: 2.4 %	5.3%	5.6%

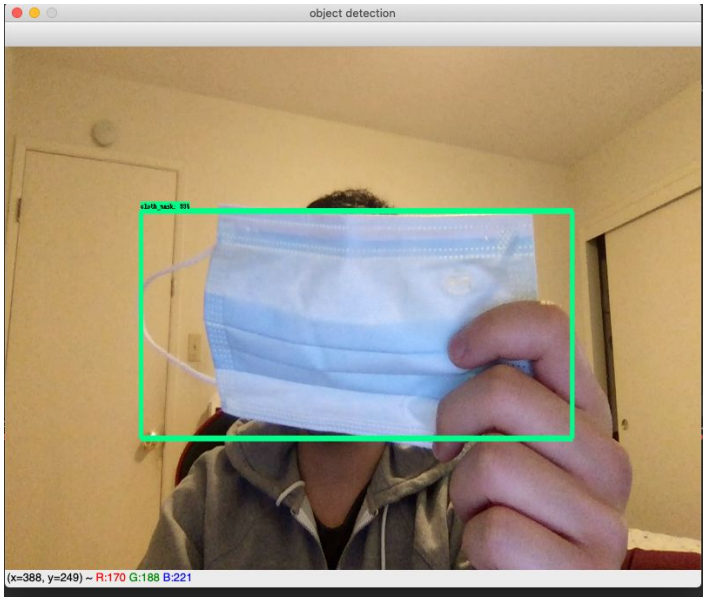
Based on the above comparison, it shows that overall tobacco user patients have slightly higher percentages on non-smoking patients. However, if a patient has diabetes as a pre-condition, they will have a lot higher percentages. With diabetes and being a tobacco user, patients will likely get hospitalized.

Overall, being a tobacco user can give the patient a little more severe when compared with non-smoking patients. However, other pre-condition is much worse than just using tobacco,

VI. FUTURE IMPLEMENTATION

Through our research, the overall covid cases are increasing. The in-depth data analysis is a small subset of the covid operation. Many dashboards and statistical analyses are published on the internet. However, the information and expert' advice is not effective to reduce the number of COVID cast. In fact, the mumps vaccine is the quickest to have ever been developed, according to *National Geographic*. Currently, the COVID vaccine has some promises by the US Federal Governor Officials. However, the question will be how fast the operation can the government deliver the vaccine to everyone. The process is complicated. Therefore, the post-pandemic opening is necessary for the citizen to go back to work. In this case, we design a software system that is combining Machine Learning (ML), Deep Learning (DL), the Internet of Things (IoTs), and Artificial Intelligence (AI). The system is Covidnator, and the idea is to implement cameras in public to gather facial data and use ML to train distance calculation. Therefore, we can have businesses install the camera and the app to regulate and remind their people on social distance. The cameras have facial recognition, mask detection, and algorithms for calculating social distance. The implementation uses open-sources like Tensorflow, OpenCV, Python, and open-sources Tensorflow models. As a result, businesses can be open with better protection and sustain the economy.





VII. CONCLUSION

Since the Covid19 is the first pandemic that our generations have experienced, we didn't know anything to prepare enough. By doing this research, we hope that it can provide more useful information for people. Moreover, it can help us to know what to prepare for the next pandemic, if we have one again.

Thanks to Professor Rakesh Ranjan for his devoted guidance. The clarity of his comments helped us to know what to do correctly so that it can be used by the general public.

REFERENCES

[1] https://covid.cdc.gov/covid-data-tracker, 2020

[2] Identifying airborne transmission as the dominant route for the spread of COVID-19 Renyi Zhang, Yixin Li, Annie L. Zhang, Yuan Wang, Mario J. Molina Proceedings of the National Academy of Sciences Jun 2020

[3] https://research.fb.com/prophet-forecasting-at-scale/

[4] https://www.gob.mx/salud/documentos/datos-abiertos-152127