A Machine Learning (ML) Modelling Approach in Monitoring and Controlling the Viral Pandemic-COVID 19

Mr. Susheel George Joseph MCA, MTech, MPhil, UGC-NET, (PhD)
(Associate Professor, Department of Computer Applications, Kristu Jyoti College of Management and Technology, Changanassery, Kottayam, Kerala, India.)

Abstract. As the world grapples with COVID-19, every ounce of technological innovation and ingenuity harnessed to fight this pandemic brings us one step closer to overcoming it. Artificial intelligence (AI) and machine learning are playing a key role in better understanding and addressing the COVID-19 crisis. Machine learning technology enables computers to mimic human intelligence and ingest large volumes of data to quickly identify patterns and insights. In the fight against COVID-19, organizations have been quick to apply their machine learning expertise in several areas: scaling customer communications, understanding how COVID-19 spreads, and speeding up research and treatment.

Keywords. COVID 19, Machine Learning, Quarantine, Pandemic, Classification, Clustering, Time Series, Regression, Sentiment Analysis.

1. Introduction

Viral pandemics are a serious threat. COVID-19 is not the first, and it won't be the last. A pandemic is an epidemic of an infectious disease that has spread across a large region, for instance multiple continents or worldwide, affecting a substantial number of people. A widespread endemic disease with a stable number of infected people is not a pandemic. Widespread endemic diseases with a stable number of infected people such as recurrences of seasonal influenza are generally excluded as they occur simultaneously in large regions of the globe rather than being spread worldwide.

Throughout human history, there have been a number of pandemics of diseases- smallpox and tuberculosis. The most fatal pandemic in recorded history was the Black Death (also known as The Plague), which killed an estimated 75–200 million people in the 14th century. Other notable pandemics include the 1918 influenza pandemic (Spanish flu). Current pandemics include COVID-19 and HIV/AIDS.

Etiology, the modern branch of science that deals with the causes of infectious disease, recognizes five major modes of disease transmission: airborne, waterborne, bloodborne, by direct contact, and through vector (insects or other creatures that carry germs from one species to another). As humans began traveling overseas and across lands which were previously isolated, diseases have been spread by all five transmission modes.

But, like never before, we are collecting and sharing what we learn about the virus. Hundreds of research teams around the world are combining their efforts to collect data and develop solutions.

How machine learning is helping us to:

- ✓ Choose who all are most at risk,
- ✓ Diagnose patients,
- ✓ Research and develop drugs faster,
- ✓ Find a pool of existing drugs that can help
- ✓ Predict the spread of the disease,
- ✓ Closely observe and understand viruses better,
- ✓ Trace where viruses come from, and
- ✓ Understanding how COVID-19 spreads
- ✓ Researchers and practitioners analyse large volumes of data to forecast the spread of COVID-19, in order to act as an early warning system for future pandemics and to identify vulnerable populations.
- ✓ Leaders make more informed decisions in the face of COVID-19.
- ✓ Examining ways to limit the spread of COVID-19, particularly among vulnerable populations.

✓ Predict the next pandemic.

Let's fight this pandemic – and prepare ourselves better.

2. Machine Learning (ML) Approach for monitoring

Since there is a direct impact visible on the depth and length of disruption and shape of the recovery curves, a machine learning (ML) approach could help in reasoning-based monitoring of indicators for the aforesaid questions.

2.1. Factors to be observed to obtain the Depth of Disruption

Time to implement social distancing after community transmission is confirmed– Since this pandemic is spreading through community and local transmission, it is very crucial to monitor the time taken to implement social distancing which can be accomplished by using time series analysis.

The purpose of *Time Series Forecasting* is generally twofold- to understand or model the stochastic mechanisms giving rise to an observed series and to predict or forecast the future values of a series based on the history of that series.

Number of cases- absolute: *Classification algorithms* can be used for monitoring the number of absolute active COVID19 cases.

Geographic distribution of cases relative to economic contribution- *Clustering algorithms* can help in monitoring this indicator as it allows grouping a set of objects.

Cuts in spending on durable goods- Due to reduced supply and shortage of components, promotional offers and discounts are also being cut on finished products which is leading to cuts in spending on durable goods such as refrigerators, air conditioners, LCDs etc. This indicator can be monitored with the help of *Regression algorithms* where the "outcome variable" of downfall in demand can be analysed based on the "input features" of cutting promotional offers and discounts.

Extent of behaviour shift- *Sentiment analysis* is used to understand the change in behaviour and is increasingly being used for social media monitoring, brand monitoring, the voice of the customer (VoC), customer service, and market research. In terms of COVID19 lockdown, it is very crucial to analyse the post COVID19 period on the behaviour shift in spending on socializing such as eating out at restaurants, entertainment etc.

Extent of travel reduction- Post COVID19 situation needs to be analysed in terms of the extent of travel reduction with the help of t*ime series analysis* and *deep learning* models. It will impact both tourist and business travels due to one or more independent variables such as employment stability, travel alternatives, urgency of travel and travel cost.

2.2. Factors to be observed to obtain the Length of Disruption

Rate of change of cases- Various factors such as lack of community and local transmission, self-quarantine and self-isolation will allow the chain to break. A *time-series analysis* will help in understanding the rate of change of COVID19 cases.

Evidence of virus seasonality– T*ime Series Analysis* can provide predictions for COVID19 seasonality in a *linear or nonlinear pattern* that repeats at regular or irregular intervals. It is also stated that with the increase in temperature, the impact and spread of this virus will decrease but it is still unproven. A timeseries analysis will provide the seasonality data and identify the patterns, if observed.

% of cases treated at home- This indicator will comprise of *structured data classification* and can be handled by classification algorithms.

% utilization of hospital beds- Utilization forecasting uses *linear regression models* to extrapolate and make predictions based on existing data. This help in flattening the curve to let the active cases remain below the threshold capacity of hospitals to treat the infected people.

Availability of therapies- Availability of therapies based on infection severity and spread can be dealt by the simple *binary classification algorithms* and shall allow prediction of future cases with such medical diagnosis details.

Case fatality ratio Vs. other countries- Case fatality rate is the proportion of deaths from a certain disease compared to the total number of people diagnosed with the disease for a certain period of time. *Time-series and Logistic regression algorithms* can be used for monitoring this indicator as the ICMR mentioned that "Till we see a significant number of cases to indicate community transmission, let us not overinterpret things".

Late payment/credit defaults- The best performing model for detection of defaulting credit card customers has been *naive Bayes model.*

Stock market & volatility indexes- Since stock market and volatility indexes require hypotheses, hence *k-nearest neighbors algorithm (k-NN) algorithm* is used for both *classification and regression*. It is a useful technique which can assign weights to the contributions of the neighbors so that the nearer neighbors contribute more to the average than the more distant ones.

Initial claims for unemployment- The government of India has announced a 1.7 lakh crore relief package. The time taken for disbursement of package will decide the length of disruption that is going to last. *Predictive ML time series models* will help in faster claims fulfilment.

2.3. Economic factors to be observed to obtain the Shape of Recovery

Effective integration of public health measures with economic activity- It is necessary to effectively integrate public health measures with that economic activity to conduct it. For example, regular sanitization of workplace and disease prevention steps in a manufacturing unit may need a certification from the authorized body before starting the business operations which may require sustaining economic models. *Correlation and regression models* will help in analysing this indicator.

Potential for different disease characteristics over time- A *predictive ML model* for finding potential threat for different disease characteristics can be analysed with *time series forecasting* which will help in risk mitigation and making business continuity plans. A *random forest model* can help in this type of monitoring as it may require constructing a multitude of decision trees.

Bounce-back in economic activity- It is again a *predictive model* indicator based on *time-series analysis* for uplifting the slowed down sectors which will depend on the depth and length of disruption and economic policies.

Various epidemiological and economic indicators- As per the report of the WHO-China Joint Mission on COVID-19, there has been a relentless focus on improving key performance indicators, for example, constantly enhancing the speed of case detection, isolation and early treatment. It is very important to understand the epidemiological indicators based on areas without active cases, areas with sporadic cases (*non-linear regression*), areas with community clusters (*clustering*) and areas with community transmissions (*classification*).

3. Identify the risk individuals from COVID-19

ML has proven to be invaluable in predicting risks in many spheres. With medical risk specifically, ML is potentially interesting in three key ways.

Infection risk: What is the risk of a specific individual or group getting COVID-19?

Severity risk: What is the risk of a specific individual or group developing severe COVID-19 symptoms or complications that would require hospitalization or intensive care?

Outcome risk: What is the risk that a specific treatment will be ineffective for a certain individual or group, and how likely are they to die?

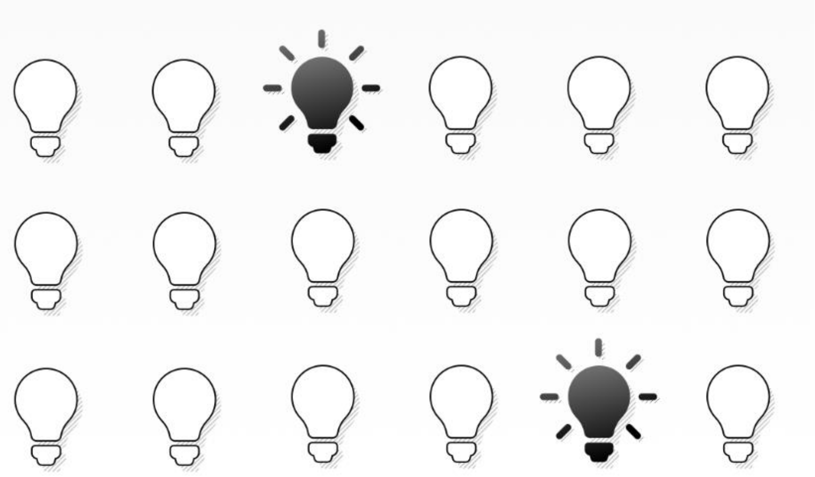

Figure 1. Identification of persons who are risk from COVID-19

ML can potentially help predict all three risks. Although it's still too early for much COVID-19-specific ML research to have been conducted and published, early experiments are promising. Furthermore, here this paper focus on how ML is used in related areas and imagine how it could help with risk prediction for COVID-19.

3.1 Predicting the risk of infection

Early statistics show that important risk factors that determine how likely an individual is to contract COVID-19 include: *Age, Pre-existing conditions, General hygiene habits, Social habits, Number of human interactions, Frequency of interactions, Location and climate, Socio-economic status.*

Risk research for the current pandemic is still in the early stages. For eg., DeCapprio et al. have used ML to build an initial Vulnerability Index for COVID-19. Measures such as wearing masks, washing hands, and social distancing are all likely to influence overall risk as well.

3.2 Predicting who is at risk of developing a severe case

Once a person or group has become infected, we need to predict the risk of that person or group developing complications or requiring advanced medical care. Many people experience only mild symptoms, while others develop severe lung disease or acute respiratory distress syndrome (ARDS), which is potentially deadly. It's not possible to treat and closely monitor everyone with mild symptoms, but it's far better to start treatment early if more severe symptoms are likely to develop.

Recent researchers published an article showing that ML could potentially predict the likelihood of a patient developing ARDS as well as the risk of mortality, just by looking at the initial symptoms. The researchers acknowledge the limitations of this research: *"A clear limitation of this study is the size of the dataset; 53 patients with some incomplete data as well as a limited spectrum of severity."* But the study lays important groundwork for applying machine learning once more data becomes available.

3.3 Predicting treatment outcomes

An extension of severity prediction is predicting the treatment's outcome, which is often literally a matter of predicting life and death. Clearly, it would be useful to know how likely a patient is to survive, given certain symptoms. But on top of this, it's important to keep in mind that not all patients are treated in the same way. Given a specific patient or group, how effective is a specific treatment likely to be?

If we can predict the outcomes of specific treatment methods, then doctors can treat patients more effectively. Using ML to personalize treatment plans is not specific to COVID-19, and ML has previously been used to predict treatment outcomes for patients with epilepsy as well as responses to cancer immunotherapy.

Because treatment options for COVID-19 are still evolving, it will likely be some time before we see ML applied to predicting outcomes for specific treatments. But outcome prediction remains an important part of risk assessment.

4. Observe patients and diagnose COVID-19

When a new pandemic hits, diagnosing individuals is challenging. Testing on a large scale is difficult and tests are likely to be expensive, especially in the beginning. Anyone who has any symptoms of COVID-19 is likely to be very concerned that they have contracted the disease, even if the same symptoms are indicative of many other, potentially milder diseases too.

Instead of taking medical samples from each patient and waiting for slow, expensive lab reports to come back, a simpler, faster, and cheaper test (even if it's less accurate) would be useful in gathering data on a larger scale. This data could be used for further research, as well as for observing, filtering and triaging patients.

When it comes to using ML to help diagnose COVID-19, applicable research areas are mainly:

- Using face scans to identify symptoms, such as whether or not the patient has a fever,
- Using wearable technology such as smart watches to look for tell-tale patterns in a patient's resting heart rate,
- Using ML-powered chatbots to screen patients based on self-reported symptoms.

4.1 Diagnosing patients using face scans

Although there are few precise details available; Upon entering the hospital, patients are given an automatic face scan, which uses ML to detect whether they have a fever.

On its own, this data is probably not extremely helpful, but when dealing with hundreds or even thousands of patients, every piece of data is important in helping triage them effectively.

4.2 Using wearable technology to screen for resting heart rate

Apple Watch can be used to detect common heart issues with the help of ML. But patterns in resting heart rate can be indicative of more specific problems too, and some preliminary research using Fitbit data indicates that changes in resting heart rate can help identify "ILI" or "influenza-like illness" patients.

Similarly, a sleep and activity tracking ring, uses body temperature, heart rate, and breathing rate from OURA to try to "identify patterns of onset, progression, recovery for COVID-19."

4.3 Using chatbots for screening and diagnosis

If doctors spend too much time answering worried patients' basic questions, they have less time to focus on treating patients who need them more. Develop a "self-triage" systems, where patients complete a questionnaire about their symptoms and medical history before being advised whether to stay home, call a doctor, or visit a hospital. Companies Like Microsoft, have released chatbots that help people self-identify their best course of action, given their specific symptoms.

ML is currently more suited to helping screen COVID-19 patients rather than reliably diagnosing them. Doing real diagnostics is challenging, partly because any diagnostic algorithm also has to be robust to mutations.

Pardis Sabeti discusses some related challenges in a Ted Talk: "We also could see that, as the virus was moving between humans, it was mutating. And each of those mutations are so important, because the diagnostics, the vaccines, the therapies that we're using, are all based on that genome sequence – fundamentally, that's what drives it."

If we do all of this work for a specific virus and then that virus mutates, a lot of work is potentially wasted and has to be redone. If we do find a ML algorithm that can quickly and accurately detect COVID-19, it will need to be robust enough to handle mutations.

5 Effective management and development of drugs

5.1. Speeding up drug development

In response to a new pandemic, it's critical to come up with a vaccine, a reliable diagnostic method, and a drug for treatment – fast. Current methods involve a lot of trial and error, and can take months to isolate even one viable vaccine candidate.

ML can speed up this process significantly without sacrificing quality control. When researchers were trying to find small molecule inhibitors of the Ebola virus, they discovered that training *Bayesian ML models* with viral pseudotype entry assay and the Ebola virus replication assay data helped speed up the scoring process. This accelerated process quickly identified three potential molecules for testing. Similarly, researchers working on H7N9 discovered that ML-assisted virtual screening and scoring led to substantial improvements in the accuracy of the scores. Using the *random forest algorithm /a classification algorithm made up of lots of decision trees* provided the best results with H7N9.

In the COVID-19 pandemic, where a virus is spreading rapidly, getting more accurate scores faster is critical to speeding up drug development.

5.2. Identifying effective existing drugs

Repurpose drugs that have already been tested and used to treat other diseases.

But there are thousands of drug candidates, and we don't have time to test them all – so how do we find the right one?

ML can help us prioritize drug candidates much faster by automatically:

- Building knowledge graphs and
- Predicting interactions between drugs and viral proteins

5.2.1. Building biomedical knowledge graphs

A lot of what we know about drugs, viruses, and their mechanism is spread across a huge number of research articles. We can use natural language processing (ML applied to text) to read and interpret a large number of scientific articles and build biomedical knowledge graphs, which are structured networks that connect different entities, such as drugs and proteins).

Specifically, scientists have customized an ML-built knowledge graph and applied it to COVID-19 to find a connection between the virus and the potential drug candidate Baricitinib.

COVID-19 most likely uses the protein ACE2 to enter our lung cells. This process – known as endocytosis – is regulated by AAK1 (another protein). Baricitinib inhibits AAK1, and could potentially also prevent COVID-19's entry into our lung cells.

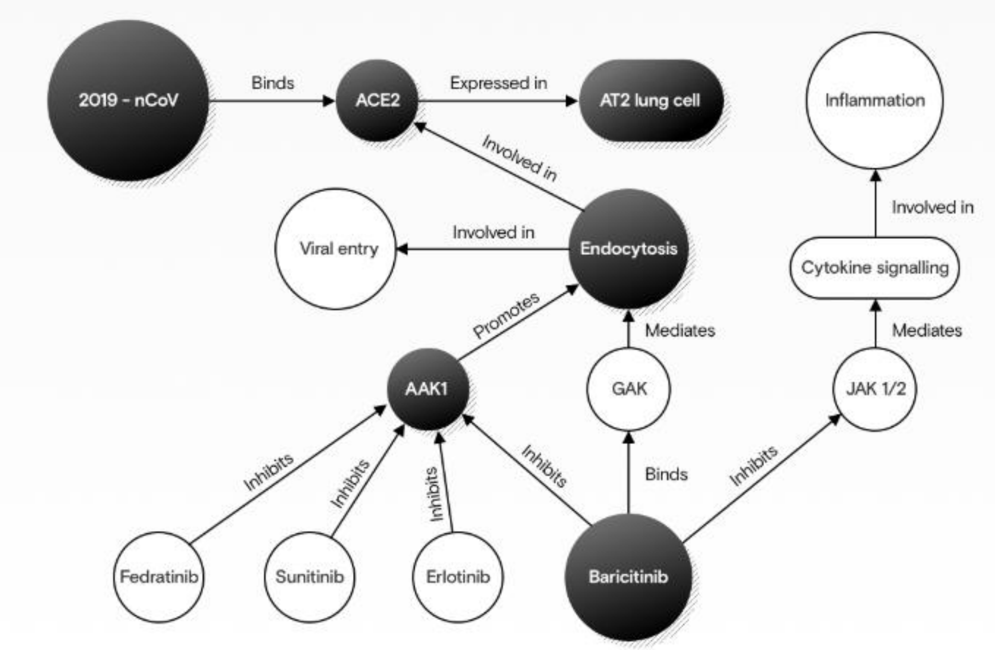

Figure 2. Bio-medical Knowledge Graph

5.2.2 Predicting drug-target interactions

ML can be used to identify drug candidates by predicting drug-target interactions (DTIs) between the virus's proteins and existing drugs.

These interactions are highly complex, so researchers mostly choose neural networks to identify them (1, 2, 3). These networks are trained on large DTI databases to generate lists of particular drug candidates that are most likely to bind to and inhibit the virus's proteins.

One research team has developed an end-to-end framework for using neural networks to process knowledge graphs. The model is then trained to interpret the knowledge graph and can be used to accurately predict DTIs.

6. Prediction of disease using social networks

Develop strategies to actively work against the pandemic.

Obtain the following inferences-

- Where we are?
- How many people are infected? and
- Where are these people?

Unfortunately, pandemics – especially those caused by viruses – are difficult and expensive to keep track of. Usually the government answers these questions, together with the health system. Every day/week the responsible agency counts and publicizes the number of new patients diagnosed with the disease. But one of the problems here is that there might be a big time and space gap between contracting the disease, developing the first symptoms, and testing positive.

In this digital world, a candidate who is starting to develop covid 19 symptoms might live in a small town with no nearby hospitals capable of performing the test. The candidate be able to access social networks and immediately input his/her health details and the spread of the disease. Shortly, only a ML model can learn to process at scale.

By interpreting the content of public interactions on social media, a ML model assesses the likelihood of novel virus contamination. The model might not be able to classify people on an individual level, but it can use all of this data to estimate the spread of the pandemic in real time and to forecast the spread in the upcoming weeks. The value of this information in decision-making processes in the midst of a rapidly evolving pandemic cannot be overstated.

7. Understand and attack virus

To understand COVID-19 is to understand its proteins – whether and how a person get sick depends entirely on how these proteins interact with their bodies. But interpreting them is no easy task. ML can help improve our understanding of viruses by analysing their proteins.

7.1. Predicting viral-host protein-protein interactions

Protein-protein interactions (PPIs) between viruses and human body cells determine body's reactions to pathogens. The virus-host interactome is the entire map of interactions between a virus's and a host's proteins. This interactome can be seen as a blueprint of how the virus infects our bodies and replicates in our cells.ML models trained with protein data have been successfully used to predict the most likely virus-host PPIs for HIV and H1N1 – greatly reducing the effort required to map the whole virus-host interactome.

Understanding how a virus interacts with our bodies is extremely important in the development of new treatments and the discovery of new drugs.

7.2. Predicting protein folding

A protein's structure is linked to its function – and once this structure is understood, we can guess its role in the cell, and scientists can develop drugs that work with the protein's unique shape. But defining a protein's 3D structure is no easy task – the range of possible structures for a single protein is astronomical: a protein composed of 100 amino acids has 3100 possible conformations. Moreover, there are one billion above known protein sequences, but we have only been able to identify the structures of less than 0.1% of them.

Using artificial neural networks (ANN), models can be built that can predict protein structures, finally making it feasible to identify protein structures using computational methods.

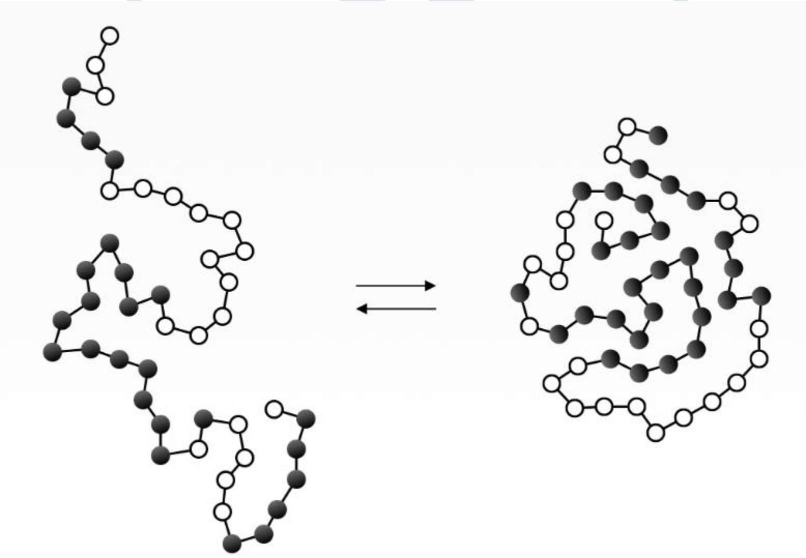

Figure 3. Folded Vs unfolded protein

7.3. How to attack the virus

Epitopes are clusters of amino acids found on the outside of a virus. Antibodies bind to epitopes, which is how our immune system recognizes and eliminates the virus. So, finding and classifying epitopes is essential in determining which part of a molecule to target when we develop vaccines.

Compared to traditional vaccines, which contain inactivated pathogens, epitope-based vaccines are safer – they prevent disease without the risk of potentially deadly side effects. Locating the correct epitope can be a time-consuming, expensive process. With a new pandemic, such as COVID-19, locating epitopes faster speeds up the process of developing effective vaccines.

This is where ML can help. Support vector machines (SVM), hidden Markov Models, and artificial neural networks (specifically deep learning) have all proven to be faster and more accurate at identifying epitopes than human.

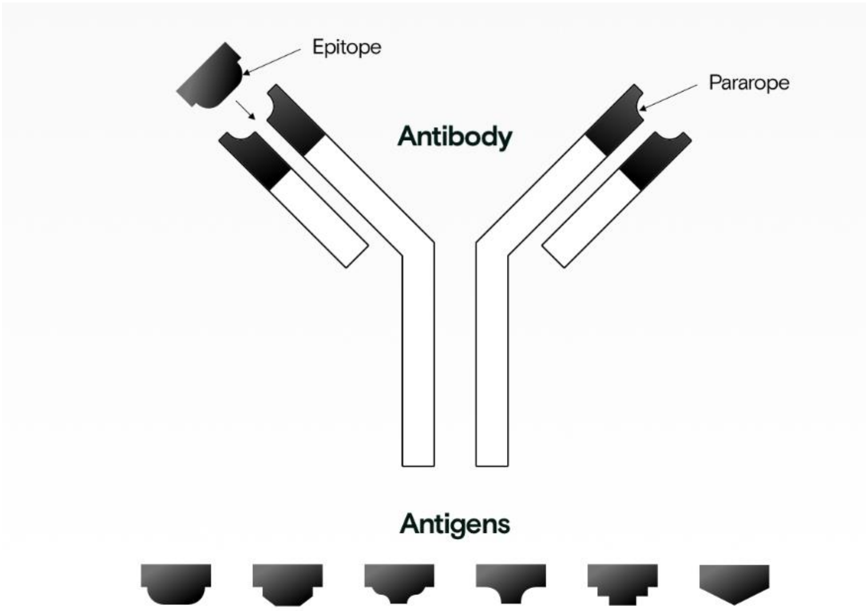

Figure 4. Figuring out the attack of virus

8. Host identification and Risk prediction in the real world

A zoonotic pandemic – like the one we are experiencing with the novel coronavirus – is a pandemic caused by an infectious disease that originates in a different species (such as bats) and spreads to humans. Viruses such as Ebola, HIV, NIPPA or COVID-19 can survive unnoticed in the natural world for a long time, waiting for the next mutation and the next opportunity to infect us. They hide in animals – called reservoir hosts – that are unaffected by the illness. Knowing who these reservoir hosts are is vital in fighting a pandemic – once we've found them, we can develop strategies to control the spread of the disease and prevent more outbreaks from happening.

Accurately predicting whether a strain of influenza is going to make a zoonotic leap-jumping from one species to another - can help doctors and medical professionals anticipate potential pandemics and prepare accordingly. With machine learning the researchers were then able to identify potentially zoonotic strains of influenza with high levels of accuracy. More work needs to be done to establish prediction models for direct transmission, but knowing which strains of influenza are likely to make a leap is an important first step in preparing for the next pandemic.

Conclusion

A situation like COVID-19 demands prompt and informed action. And that's where citizens play the biggest role. By holding the government accountable, because lives are at stake. A country needs solid, responsible, high-quality policy decisions now more than ever before.

ML is an important tool in fighting the pandemic both current and upcoming. If data is collected effectively, pool the knowledge, and combine the skills, many lives will be saved – both now and in the future.

References

[1] "The Elements of Statistical Learning", Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie, Springer Science & Business Media,

[2] "Pattern Recognition and Machine Learning", Christopher Bishop

[3] "Artificial Intelligence: A Modern Approach", Peter Norvig and Stuart J. Russell, Prentice Hall, 4th Edition, 2020

[4] "Machine Learning: The Art and Science of Algorithms that Make Sense of Data", Peter Flach

[5] "Applied Predictive Modeling", Kjell Johnson and Max Kuhn

[6] "Neural Networks and Deep Learning: A Textbook", Charu C. Aggarwal

[7] "Foundations of Statistical Natural Language Processing", Christopher D. Manning and Hinrich Schütze

[8] "Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics", Anirban Das Gupta

[9] "Plague and Pestilence: A History of Infectious Disease", Enslow. ISBN 978-0-89490-957-3.

Web references

- https://www.nature.com/articles/d41586-020-01393-7
- https://www.weforum.org/agenda/2020/05/how-ai-and-machine-learning-are-helping-to-fight-covid-19/
- https://towardsdatascience.com/covid-19-machine-learning-4f064df53c43
- https://www.medrxiv.org/content/10.1101/2020.04.08.20057679v2
- https://www.kaggle.com/tags/covid19
- https://analyticsindiamag.com/a-machine-learning-approach-for-monitoring-covid19-indicators/Altman, Linda Jacobs (1998).