International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 5, May 2020||

A Machine Learning Model for Detecting COVID-19 Virus Infections – A Case of Kenya

Kipkebut Andrew[1] , Lang'at Jonah [2]

P.G. Student, Department of Computer Science and I.T, Kabarak University–Nakuru, Kenya[1]

P.G. Student, Department of Computer Science and I.T, Kabarak University – Nakuru, Kenya[2]

ABSTRACT: COVID-19 is one of the major public health problems in the world today .Early prediction of a COVID 19 spread is the key for controlling the disease .Knowing the extent of transmission of the virus in a population can help a government policymakers, health care providers, medical officers to better target medical resources to areas affected. In this paper a model is developed to detect corona virus infections using machine learning technique, this model can be used as a proactive tool to identify potential outbreaks of Corona virus in individuals. In this study three popular classification algorithms are considered:Support Vector Machine (SVM), Artificial Neural Network (ANN) and J48 .The Data set is collected from Kenya ministry of health (MOH)from March to May 2020. Feature symptoms used include fever, sore throat, headaches, chest pain among others. Machine learning and data mining experiments are conducted using WEKA tool. The results and discussions are presented in form of descriptive statistics and detection metrics. The study observed that SVM combine with chi – square feature selection performed better.

KEYWORDS: Covid 19, Support vector machines, Artificial neural networks, Machine learning, J48

I. INTRODUCTION

The COVID-19 pandemic, also known as the coronavirus pandemic, is an ongoing pandemic of coronavirus disease 2019 (COVID 19), caused by severe acute respiratory syndrome coronavirus 2 (SARS CoV 2). The outbreak was first identified in Wuhan China in December 2019 (Fauci, Lane and Redfield, 2020). According to the World Health Organization COVID-19 has been declared an outbreak and a public health emergency of international concern on 30 January 2020, and a pandemic on 11th March 2020. As of 27th May 2020, more than 5.64 million cases of COVID-19 had been reported in more than 188 countries and territories, resulting in more than 352,000 deaths. More than 2.3 million people have recovered from the virus according to John Hopkins University (2020).

According (Covid and Team, 2020) this virus is primarily spread between people during close contact most often via small droplets produced by coughing, sneezing and talking. The droplets usually fall to the ground or onto surfaces rather than travelling through air over long distances. Less commonly, people may become infected by touching a contaminated surface and then touching their face. It is most contagious during the first three days after the onset of symptoms, although spread is possible before symptoms appear, and also from people who do not show symptoms at all (Covid- 19). Common symptoms include fever, cough, fatigue, shortness of breath, and loss of sense of smell (WHO, 2020) Complications may include pneumonia and acute respiratory distress syndrome. The time lapse from exposure to onset of symptoms is typically around five days but may range from two to fourteen days (Zhaoet al. ,2020) .There is no known vaccine or specific antiviral treatment (WHO, 2020) . Recommended preventive measures include hand washing, covering one's mouth when coughing, maintaining social distance fall the time , wearing a face mask in social gathering, self-isolation in case of infection (Cascella et al., 2020). Authorities worldwide have responded by implementing travel restrictions, lockdowns, cessation and facility closures such as schools , bars, churches, mosques, synagogues and borders. This pandemic has caused great global social and economic disruption including the largest global recession since the great depression. In Kenya for example the government is working to increase citizens testing capacity and contact tracing.

Paper is organized as follows. Section II describes related work in infectious disease detection. Section III presents methodology, Section IV experimental analysis and section V is the research conclusion.

II.RELATED WORK

The panic of infectious disease transmission has led governments to set up progressions to detect individuals at risk. As such as in airport terminals, temperature checks are performed using a thermal camera to identify people with high temperature. This check is one step of the multiple tasks taken forward to stop spread of infections. Recent approaches

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 5, May 2020||

using mathematical modelling are improving this type of surveillance. A similar system was developed to detect infected patients by classification using vital signs (sun et al.,2015) Hence, respiration rate, heart rate, and facial temperature among others were used to successfully classify individuals at higher risk for influenza using neural network and fuzzy clustering method. Fuzzy clustering methods differ from the *k*-means clustering because of the addition of the membership values (degree of belief) and the fuzzifier. In that each point can belong to multiple groups also known as clusters. This shows the ability to develop effective methods for identifying populations at risk. This method is necessary and part of the process, even in the case of emergent infectious diseases, where efforts have to be prioritized. The use of machine learning methods can be used in more sophisticated contexts. For instance, a combination of support vector machine (SVM) learning algorithm, Matlab and cross-validation method have in many years proven reliable though still lack in accuracy and precision. COVID 19 is a new pandemic and few research has been conducted on this area.

Mathematical model based systems, developed using artificial intelligence that is based on machine learning techniques are useful in predicting and diagnosing many diseases (Ghazaly et al, 2020). A Well-defined covid symptoms is sufficient enough to be used by Machine Learning (ML) techniques to effectively and efficiently predict infections .This methods through proper data pre-processing and Feature selection are key in accurately predicting both positive and negative cases. Common machine learning techniques that can be used especially in medical diagnosis expert systems include , Naïve bayes, KNN, Support Vector Machine (SVM), Naïve Bayes, Decision Tree and Artificial Neural Network (ANN).

"Modeling the current COVID-19 outbreak is much more challenging, simply because researchers know very little about the disease" (Fauci et al., 2020)

III.METHODOLOGY

Data collection

On 13 March, the first case in Kenya was recorded, a 27-year-old Kenyan woman who travelled from the US via Londonwas confirmed (MOH, 2020), during this period the Kenyan government identified and isolated a number of people who had come into contact with the first case. In this study the covid 19 data was collected from Kenya local and national health care repositories. As of May 27, 2020 the statistics are shown in fig 1.

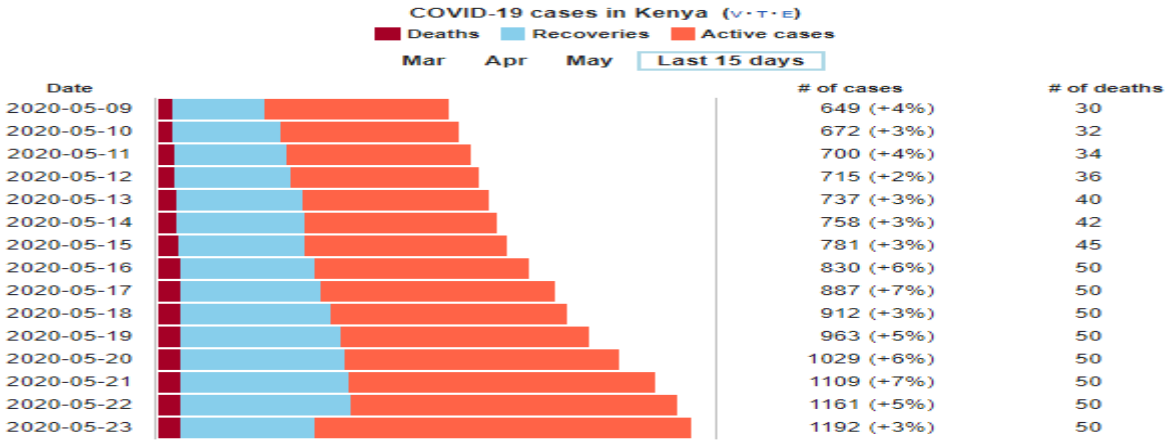

Fig. 1 Kenya covid statistics as of May 23, 2020 (source, MOH Kenya)

In this study data was collected as per the following criteria of symptoms/features present in a patient i.e Age, fever, Tiredness, Sore throat(sthroat),Difficulty in breathing(Dbreathing),chestpain, cough,headache,traveled,Class)

This feature are of type nominal (Yes, No) ,numeric, predictive class (positive or negative) .The final output is either a patient test positive or negative for Covid -19

Table 1: Sample data

Patient ID	Gender	Age	Cough	fever	Tired	Sthroat	Dbreathing	chestpain	headache	traveled	Class
1	Male	2	Yes	Yes	No	No	Yes	No	Yes	No	Negative
2	Female	20	No	No	No	No	No	No	No	No	Positive
3	Male	52	Yes	No	No	No	No	No	Yes	No	Negative
4	Female	70	Yes	yes	Yes	Yes	Yes	yes	Yes	yes	Positive
5	Male	76	No	yes	Yes	Yes	Yes	yes	No	yes	Positive
6	Female	60	Yes	No	Yes	No	Yes	No	Yes	No	Negative

From Table 1 we can note that some patients exhibit none of the symptoms of Covid 19 but tested positive , this was noted in young people below the age of 35. Majority of patients above 40 yrs.of age with the symptom tested positive especially if they had travel from COVID-19 hotspots.The methods adopted by this study includedatapre-processing , Feature selection and Classification as shown in fig 4.

Data Pre-processing

Data pre-processing whichinvolves preparing raw data from the database for automatic analysis it includes. The pre-processing step consists of creating a feature vector template and then building the feature vectors for the training data all the data was converted to nominal (NumericToNominal as per WEKA with attribute indices set to R first-last, this is represented in arff format named Covid.arff.

Feature selection methods

These methods select features that are highly correlated with a given class (positive or negative) and are uncorrelated with each other. Most redundant features areremoved because ofhigh correlation with the residual feature set. It involves deleting irrelevant and less important features especially for large data set. In this study Correlation feature selection (CFS) and Chi square (CS) methodwere considered for this purpose.

Correlation feature selection (CFS)

CFS is a simple filter algorithm that ranks feature subsets according to a correlation based heuristic evaluation function. The bias of the evaluation function is toward subsets that contain features that are highly correlated with the class (positive or negative) and uncorrelated with each other. Irrelevant features should be ignored because they will have low correlation with the class. Redundant features should be screened out as they will be highly correlated with one or more of the remaining features (wahba et al., 2015). The acceptance of a feature will depend on the extent to which it predicts classes in areas of the instance space not already predicted by other features. $r_{cf1+}r_{cf2...}$ r_{cf} variables are referred to as correlations.
The CFS criterion is defined in equation1 (eq1).

$$\mathrm{CFS} = \max_{S_k}\left[\frac{r_{cf_1} + r_{cf_2} + \cdots + r_{cf_k}}{\sqrt{k + 2(r_{f_1f_2} + \cdots + r_{f_if_j} + \cdots + r_{f_kf_{k-1}})}}\right]. \quad \text{eq(1)}$$

Chi –Square feature selection method (CS)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 5, May 2020||

Chi -square feature selection method (yang et al 2016) is very effective in this context when proper training and testing methods are used. In the Chi-square formula eq(2) the best terms t_k for the class c_i (positive ,negative) are the ones distributed most differently in the sets of positive and negative examples of class c_i.

$$\text{Chi} - \text{square}(t_k, c_i) = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \quad \text{eq(2)}$$

In the equation 2 the variables are described as follows.

N = Total number of patients cases,

A = Number of patients in class c_i that contain the symptom t_k;

B = Number of patients that contain the symptoms tk in other classes;

C = Number of patients in class c_i that do not contain the symptom t_k;

D = Number of patientsthat do not contain the symptom t_k in other classes.

Each feature is assigned a score in each class, all these scores are combined with a single final score max (Chi-square(t_k, c_i)) (Bahassine, S., et al. F, 2018).

Model Design

The experiments were conducted using WEKA. WEKA (Waikato Environment for knowledge analysis), an Open source data mining tool written in Java programming that can be used for data pre-processing, classification, clustering and visualization. It contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces (Hall, Frank, Holmes, Pfahringer, Reutemann and Witten, 2009). Weka has an extensive collection of different machine learning and data mining algorithms and its proven a very helpful data mining tool for developing prediction model by classify the accuracy on the basis of datasets. Weka Explorer has several panel like pre-process, classify, cluster, associate, select attribute and visualize. In this study only pre-processing and classification is used to develop the model .Each classifier is associated with some unique parameters.

Support Vector Machine (SVM)

Support vector machines (SVM), also support vector networks are supervised learning models with associated learning algorithms that analyses data used for classification and regression analysis (Cortes and Vapnik, 1995). Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a framework that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM framework is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible (fig 2). New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall, in addition to performing linear classification. SVM extracts a best possible hyper-plane that segregate the two classes as shown in the diagram below.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 5, May 2020||

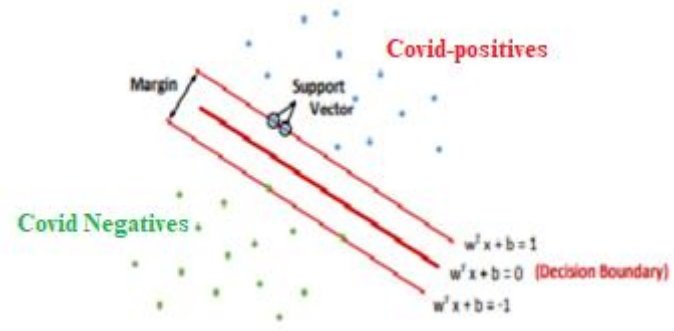

Fig 2 SVM boundary decision scheme

Artificial neural network (ANN)

Artificial neural network consists of a collection of processing elements that are highly interconnected and transform a set of inputs to a set of desired outputs. The result of the transformation is determined by the characteristics of the elements and the weights associated with the interconnections among them. A neural network conducts an analysis of the information and provides a probability estimate that it matches with the data it has been trained to recognize. The neural network gains the experience initially by training the system with both the input and output of the desired problem. In Neural networks the first stage is the feed forward where each input (xi) receives an input signal then send it to the hidden units (z1,z2,z3,…) , then computes its activation signal and then sends that signal to the output unit(y_i). This is illustrated in fig 3.

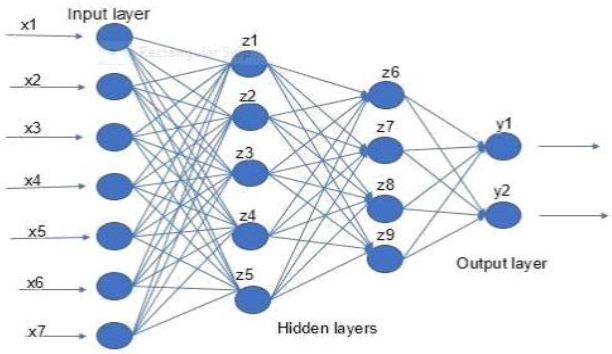

Fig 3: Artificial Neural Network

J48 Decision tree based algorithm

J48 classifier is a simple C4.5 decision tree for classification, which is an extension of Quinlan's earlier ID3 (Iterative Dichotomiser 3) algorithm. It is based on binary tree. The decision tree approach is most useful in classification of a problem. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier(Margaret and Sridhar, 2006).

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 5, May 2020||

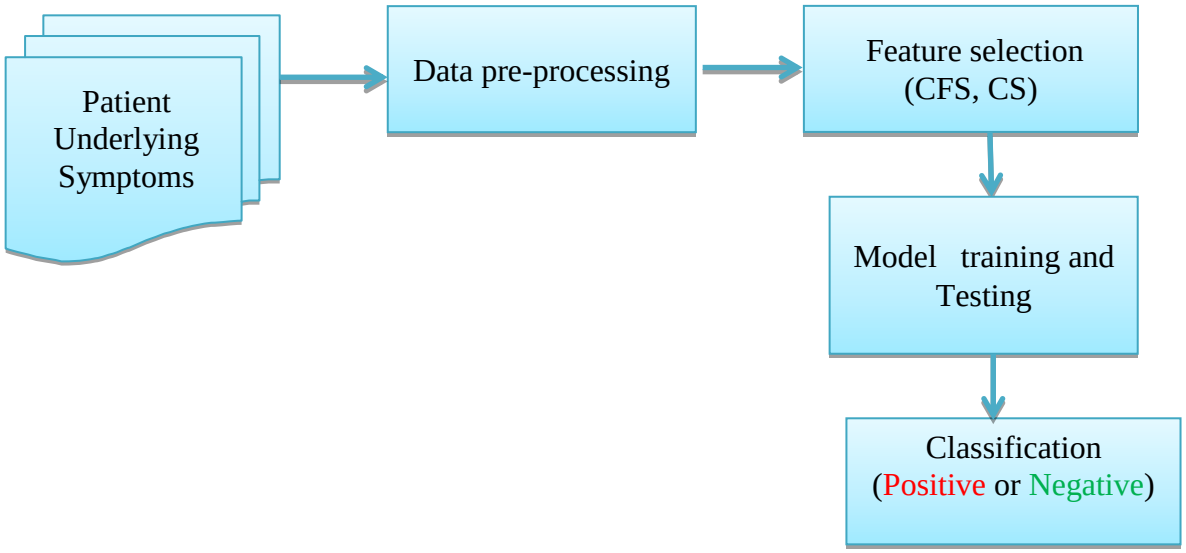

Fig.4 Model design

The model design in figure 4, contains the following steps: getting patients symptoms, data preprocessing , Feature selection ,Model training , testing and finally classification.

IV.EXPERIMENTAL ANALYSIS

When Correlation feature selection (CFS) with best first search was used for feature selection , the attributes was reduced to five, the presence or absence of difficulty in breathing , chest pain ,fever ,sore throat and age are key features in determining positive or negative cases in patients.
In the case of Chi square feature selection with ranker search the following features were selected and ranked.

Table 2: Sample Chi- square attributes selection

Rank	Attributes
66.563	10 difficulty breathing
19.645	3 Fever
19.645	8 chest pain
19.645	7 Age
18.545	5 soar throat
0.957	9 Gender

The ranked decimal column represents how important an attribute is to the class based on the feature selection method. The attribute column gives the average position of the attribute in respect all the selected attributes table 2. Average merit in table 3 also tells us how important thatattribute is (i.e. the higher the number the better), this is averaged over the folds (10) of the cross validation.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 5, May 2020||

Table 3: Chi-square 10 folds cross validation on the data set.

Average merit	average rank	Attribute
55.745 +- 3.425	1 +- 0	10 Dbreathing
15.796 +- 1.712	2.2 +- 0.4	3 fever
15.796 +- 1.712	3.4 +- 0.8	8 chestpain
15.796 +- 1.712	3.6 +- 0.8	7 Age
14.894 +- 1.281	4.8 +- 0.4	5 sorethroat
0.831 +- 0.534	6 +- 0	9 Gender

The experiments were conducted on the three classifier and their accuracies, true positive and false negative and root means square error (RMSE) are recorded.

J48 algorithm was able to correctly classify 54 (62.069 %) and incorrectly classified 33instances.(37.931 %). TP of 0.621, FP 0.483 and RMSE of 0.4611 which is relatively good was also recorded.

Neural network correctly classified 60 (68.9655 %), incorrectly classified 27 (31.0345 %) TP of 0.690, FP 0.466 and RMSE of 0.5371 which is an improvement 7 % as compared to J48 on the true positive but dropped in false positive 0.49%. A small RMSE is always recommended.According Ian ,witten and frank (2005) a good model should have Kappa closer to 1 , MAEs and RMSE closer to zero.

SVM correctly classified Instances 69 (79.3103 %) and incorrectly classified instances 18 (20.6897 %) instances. TP of 0.793 FP of 0.362 and RMSE of 0.4549 SVM recorded a great improvement on accuracy, TP, FP and RMSE by almost 10%.

V.CONCLUSION

As per these experiments, we can conclude that SVM when combined with Chi- square feature selection recorded the optimal detection criteria. However since COVID-19 is an ongoing pandemic. There is little data available for research, especially on the virus symptoms and manifestation. More academic research is required on this areain order to further the study. It would be also interesting to see if the use of ensemble hybrid learning will improve on accuracy when data is scaledup.

REFERENCES

[1] Ai, T.; Yang, Z.; Hou, H.; Zhan, C.; Chen, C.; Lv, W.; Tao, Q.;Sun, Z.; Xia, L. Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease 2020 (COVID-19) in China: A Report of 1014 Cases. Radiology 2020, 2020, 200642.

[2] Bai, Y., Yao, L., Wei, T., Tian, F., Jin, D. Y., Chen, L., & Wang, M. (2020). Presumed asymptomatic carrier transmission of COVID-19. *Jama*, *323*(14), 1406-1407.

[3] Bahassine, S., et al. Feature selection using an improved Chi-square for Arabic text classification. Journal of King Saud University – Computer and Information Sciences (2018), ttps://doi.org/10.1016/j.jksuci.2018.05.010.

[4] COVID, C., & Team, R. (2020). Severe outcomes among patients with coronavirus disease 2019 (COVID-19)—United States, February 12–March 16, 2020. *MMWR Morb Mortal Wkly Rep*, *69*(12), 343-346.

[5] COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)". ArcGIS. Johns Hopkins University. Retrieved 27 May 2020.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 5, May 2020||

[6] Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C., & Di Napoli, R. (2020). Features, evaluation and treatment coronavirus (COVID-19). In *Statpearls [internet]*. StatPearls Publishing.
[7] Ghazaly, N. M., Abdel-Fattah, M. A., & Abd El-Aziz, A. A. (2020). Novel Coronavirus Forecasting Model using Nonlinear Autoregressive Artificial Neural Network. *Journal of advanced science*.
[8] Fauci, A. S., Lane, H. C., & Redfield, R. R. (2020). Covid-19—navigating the uncharted
[9] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explolations newsletter, 11(1), 10-18.
[10] Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. (February 2020). "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China". *Lancet*. 395 (10223):
[11] https://www.health.go.ke/covid-19/
[12] https://en.wikipedia.org/wiki/Coronavirus_disease_2019
[13] https://www.who.int/emergencies/diseases/novel-coronavirus-2019?gclid=CjwKCAjwiMj2BRBFEiwAYfTbCtwtfnur4Ua98Sak0mcGeZnxo_tFOjz9RwdWYLXNz0cbAaW-PVjJoxoCYy0QAvD_BwE
[14] Kong, J., Wang, S., & Wahba, G. (2015). Using distance covariance for improved variable selection with application to learning genetic risk models. *Statistics in medicine*, *34*(10), 1708-1720.
[15] Margaret H. Danham,S. Sridhar(2006), " Data mining,Introductory and Advanced Topics", Person education , 1st ed., 2006
[16] McCallum, Andrew. "Graphical Models, Lecture2: Bayesian Network Represention" (PDF). Retrieved 22 October 2019.
[17] Sun G., Matsui T., Hakozaki Y., Abe S. An infectious disease/fever screening radar system which stratifies higher-risk patients within ten seconds using a neural network and the fuzzy grouping method. J. Infect. 2015;70(3):230–236.
[18] Zhou, P.; Yang, X.-L.; Wang, X.-G.; Hu, B.; Zhang, L.; Zhang,W.; Si, H.-R.; Zhu, Y.; Li, B.; Huang, C.-L.; et al. A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin. Nature 2020, 579, 270.