MML Classification Techniques for the Pathogen Based on Pnuemonia-nCOVID-19 and the Detection of Closely Related Lung Diseases Using Efficacious Learning Algorithms

M. Kannan and C. Priya

Abstract The main purpose of this topic is to provide an excellent classification method for predicting the disease based on the key aspect of the disease. Here, we used a multiclass variable database for the prediction; also the methods, random forest and linear SVC, are used for the classification. Furthermore, based on the confusion matrix, we can know the outcome of the prediction model. In this, all the results are discussed using the confusion matrix. Infectious diseases such as nCOVID-19 cause serious damage to the human body's immune system. It recently emerged from China and affects neighbors' country and flu-like symptoms initially manifest in 89.9%. The disease spreads faster than SARS-CoV and MERS-CoV, and soon, the disease begins to spread from one person to another, with high fever (101.4 F), inhalation or dyspnea, sore throat, sneezing and coughing. In India, as of January 31, 2020, the number of cases was one, and on March 28, 2020, the outrage began to rise to 909. In addition, COVID is also caused by pneumonia-related illnesses. So far, such epidemics have been studied and diagnosed by reverse transcriptase polymerase chain reaction (RT-PCR) and serology laboratory testing. Chest X-ray or computed tomography helps identify damaged and white cells in the affected body, identifying pathogens, and the presence of abundant metagenomic sequence in RNA is a major clinical challenge. Since the vaccine has not yet been announced, the current treatment is supplemental care. In this study, we compared machine learning classification methods such as NN, SVM, MLP, RF and KNN, which are widely used in the healthcare sector to diagnose disease by X-ray. Doctors often prescribe chest radiography to diagnose and/or predict infections, since we have read numerous articles on coronavirus. Further, in the clinical perspective, machine learning plays a

M. Kannan (✉)
PhD Research Scholar, Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, India
e-mail: kannanmuthusamy.research@gmail.com

C. Priya
Associate Professor, Department of Information Technology, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, India
e-mail: drcpriya.research@gmail.com

C. Chakraborty et al. (eds.), *Internet of Medical Things for Smart Healthcare*, Studies in Big Data 80, https://doi.org/10.1007/978-981-15-8097-0_3

vital role in solving the problem of prognosis and, thanks to treatment monitoring, there are effective mechanisms. In the presence of airborne diseases, we need an effective tool such as machine learning to investigate this, because nCOVID is transmitted by sneezing or coughing and/or other pulmonary syndrome. Therefore, this review summarizes the current outbreaks of coronavirus and its closely related lung viruses such as influenza and pneumonia, medical-based machine learning (MML) techniques and comparative analysis of MML for infectious diseases.

Keywords nCOVID-19 · Chest X-ray · Pulmonary diseases · Pneumonia · Medical machine learning · Classification models · Outbreaks · World Health Organization

1 Introduction

The latest nCOVID-19 public health epidemic problem is one of the infectious viral diseases that first erupted in Wuhan, Hubei Territory, China (2019). This is followed by the announcement of a blockade of the world's countries. Day after day, the disease will increase the infection among people. The serious symptoms of COVID are breathing difficulties, in which it occurs within the incubation period of 1–14 days. Furthermore, in a few decades, the World Health Organization-WHO has to continually face many viral diseases. For example [1]: Avian influenza in 1997, SARS-CoV in 2002–2003, H1N1 influenza in 2009, MERS-CoV in 2012 and EZV in 2014. These diseases cause severe acute respiratory syndrome and various lung diseases in the human brain and increase mortality with these new types of viruses; the mortality rate increases, making WHO the top priority in controlling the spread of COVID. Compared to other infectious diseases, nCOVID causes a serious outbreak worldwide. Currently, there are no current vaccines or other supplements against nCOVID due to the lack of scientific and clinical studies. As of April 4, 2020, a total of 205 countries and territories have been reported and about 11,17,860 positive cases have been identified in China and India has 2,902 confirmed cases. In addition, there is a huge increase in people's mortality every day. The disease also has close contact with the flu virus and viral pneumonia, in which the flu causes viral cough in which pneumonia also causes severe respiratory problems. For this reason, in the medical industry, computed tomography [2] is highly recommended for checking and diagnosing disease such as pneumonia. In today's world, the human race is increasingly affected by combinations of pandemic and epidemic diseases. Most of these viruses are infected with low resistance levels or the main airways of the human body. According to a report published by pathologists [3] so far, COV is known to cause more germs in the respiratory tract and intestinal tract, as well as other bacterial diseases such as influenza. Thus, the WHO named the "corona" from the family Coronoviridae and is divided into four categories: α, β, γ, δ by various pathologists. The human body contains a multitude of RNA-filled blood cell proteins and DNA culture. Any viral disease can easily occur through the blood immunity cell. In the

medical industry, blood cell count is also important test and tremendous. Basically, the hemocytometer and some chemical compounds will count the blood cells. Therefore, it is a tedious task to count all infected and low immunity blood cells. For this reason, Mahmudhul Alam and Taqual Isam proposed a "YOLO" ("only seen once") method for automatic blood cell counting in 2019 [4] (Tables 1 and 2).

The Centers for Disease Control and Prevention (CDC) have reported that the contagious respiratory swine flu virus or H1N1 is one of the most serious public health problems and that humans can easily become infected with fever, cough and chills. As a result, the type-A influenza virus is influenced by pigs. It is classified into avain, pig, pandemic, seasonal. The human-mammalian (avian) transaction scenario is rare, but compared to other flu types, the avian influenza virus and COV have a similar structure along with the infectious stages, https://www.cdc.gov/flu/pdf/avianflu/avian-flu-transmission.pdf. The human gene is constructed as a multiple cell sequence, which is sensitive and can be easily influenced by bacteria and other pandemic viruses, such as influenza disease, if the gene contains low immunity. Machine learning methods are very effective in the multidisciplinary area, including

Table 1 Total number of suspected cases [31]—2020

Country	1st May	2nd May	3rd May	4th May	5th May	6th May	7th May	8th May
Afghanistan	2171	2335	2469	2704	2894	3224	3392	3563
Algeria	4006	4154	4295	4474	4648	4838	4997	5182
Australia	6762	6767	6783	6801	6825	6849	6875	6896
Austria	15,424	15,458	15,470	15,538	15,569	15,586	15,651	15,673
Bangladesh	7667	8238	8790	9455	10,143	10929	11719	12425
Brazil	85,380	91,589	96,559	101,147	107,780	114,715	125,218	135,106
Canada	53,236	55,061	56,714	59,474	60,772	62,046	63,496	64,922
China	83,956	83,959	83,961	83,964	83,966	83,968	83,970	83,976
France	**129,581**	**130,185**	**130,979**	**131,287**	**131,863**	**132,967**	**137,150**	**137,779**
Germany	**159,119**	**161,703**	**162,496**	**163,175**	**163,860**	**164,897**	**166,091**	**167,300**
Ghana	2074	2074	2169	2169	2719	2719	3091	3091
India	35,043	37,336	39,980	42,533	46,433	49,391	52,952	56,342
Indonesia	10,118	10,551	10,843	11,192	11,587	12,071	12,438	12,776
Iran	94,640	95,646	96,448	97,424	98,647	99,970	101,650	103,135
Ireland	20,612	20,833	21,176	21,506	21,722	21,983	22,248	22,385
Israel	15,946	16,101	16,185	16,208	16,246	16,289	16,310	16,381
Japan	14,281	14,544	14,839	15,057	15,231	15,354	15,463	15,547
Mexico	19,224	20,739	22,088	23,471	24,905	26,025	27,634	29,616
Netherlands	39,316	39,791	40,236	40,571	40,770	41,087	41,319	41,774
Pakistan	16,817	18,114	19,103	20,186	21,501	22,550	24,073	25,837
Paraguay	266	333	370	396	415	431	440	462
Peru	36,976	40,459	42,534	45,928	47,372	51,189	54,817	58,526
Philippines	8488	8772	8928	9223	9485	9684	10,004	10,343
Poland	12,877	13,105	13,375	13,693	14,006	14,431	14,740	15,047
Portugal	24,987	25,351	25,190	25,524	25,524	25,702	26,182	26,715
Puerto Rico	1539	1575	1757	1808	1843	1924	1968	2031

(continued)

Table 1 (continued)

Country	1st May	2nd May	3rd May	4th May	5th May	6th May	7th May	8th May
Qatar	13,409	14,096	14,872	15,551	16,191	17,142	17,972	18,890
Romania	12,240	12,567	12,732	13,163	13,512	13,837	14,107	14,499
Russia	**106,498**	**114,431**	**124,054**	**134,687**	**145,268**	**155,370**	**165,929**	**177,160**
Saudi Arabia	22,753	24,097	25,459	27,011	28,656	30,251	31,938	33,731
Singapore	16,169	17,101	17,548	18,205	18,778	19,410	20,198	20,939
South Africa	5647	5951	6336	6783	7220	7572	7808	8232
Spain	**215,216**	**216,582**	**217,466**	**218,011**	**219,329**	**220,325**		**221,447**
Turkey	**120,204**	**122,392**	**124,375**	**126,045**	**127,659**	**129,491**	**131,744**	**133,721**
United Arab Emirates	12,481	13,038	13,599	14,163	14,730	15,192	15,738	16,240
United Kingdom	**171,253**	**177,454**	**182,260**	**186,599**	**190,584**	**194,990**	**201,201**	**206,715**
United States	**1,069,826**	**1,103,781**	**1,133,069**	**1,158,041**	**1,180,634**	**1,204,475**	**1,228,603**	**1,256,972**

immunology, virology, microbiology and other health testing laboratories, the ability to split big data into multiple test data sets and training for the prediction model of illnesses. Computed tomography (CT) and X-rays tests are crucial medical features [5] that are used to identify pneumonia. With the spread of infectious diseases such as COVID fever and pneumonia, it is very difficult to diagnose their true disease, and doctors are analyzing the true impact of the disease with a chest X-ray.

Today, with the large amount of data, descriptive or statistical analysis is sometimes confusing and also creates an arduous task that includes understanding and extracting knowledge. One of the efficient applications of the machine learning methodology and its techniques helps many other industries; even the clinical industry uses it widely and quickly. Machine learning, a subfield of AI [6], allows the system to read data for multiple uses. The collected data set is divided into training and test data sets for future forecasting. Therefore, ML techniques are mainly used for classification and prediction with three different learning methods: supervised, unsupervised and reinforcement. Each of them has unique diagnostic benefits. Furthermore, in the medical field, early disease prediction is a rather difficult task, the disease and related data can only be viewed by experienced doctors. Sometimes, it even confuses the experts. But machine learning has the ability to build a prediction model together with the previously available original dataset. They are increasingly used in the health industry, such as EEG, ECG and radiology (Table 3 and Fig. 1).

2 Recent Pandemic

In recent decades, China has faced several infectious diseases, such as human-zoonotic viral infection, including severe acute respiratory syndrome and Middle East respiratory syndrome. Coronavirus-19 is a family of Coroviridae/RNA virus. Virulent diseases such as SARS and MERS [7] identified with zoonotic animals; therefore,

Table 2 Death cases [31]—2020

Country	1st May	2nd May	3rd May	4th May	5th May	6th May	7th May	8th May
Afghanistan	64	68	72	85	90	95	104	106
Algeria	450	453	459	463	465	470	476	483
Australia	92	93	93	95	95	96	97	97
Bangladesh	168	170	175	177	182	183	186	199
Brazil	5901	6329	6750	7025	7321	7921	8536	9146
Canada	3184	3391	3566	3682	3854	4043	4232	4408
China	4637	4637	4637	4637	4637	4637	4637	4637
France	**24,376**	**24,594**	**24,760**	**24,895**	**25,201**	**25,531**	**25,809**	**25,987**
Germany	6288	6575	6649	6692	6831	6996	7119	7266
Ghana	17	17	18	18	18	18	18	18
India	1147	1218	1301	1373	1568	1694	1783	1886
Indonesia	792	800	831	845	864	872	895	930
Iran	6028	6091	6156	6203	6277	6340	6418	6486
Ireland	1232	1265	1265	1303	1319	1339	1375	1403
Israel	222	225	229	232	235	238	239	240
Italy	**27,967**	**28,236**	**28,710**	**28,884**	**29,079**	**29,315**	**29,684**	**29,958**
Mexico	1859	1972	2061	2154	2271	2507	2704	2961
Netherlands	4795	4893	4987	5056	5082	5168	5204	5288
Pakistan	385	417	440	462	486	526	564	594
Paraguay	10	10	10	10	10	10	10	10
Peru	1051	1124	1200	1286	1344	1444	1533	1627
Philippines	568	579	603	607	623	637	658	685
Poland	644	651	664	678	698	716	733	755
Portugal	1007	1007	1023	1063	1063	1074	1089	1105
Puerto Rico	92	94	95	97	97	99	99	102
Qatar	10	12	12	12	12	12	12	12
Romania	717	744	771	780	803	827	858	876
Russia	1073	1169	1222	1280	1356	1451	1537	1625
Saudi Arabia	162	169	176	184	191	200	209	219
Singapore	15	16	17	18	18	18	20	20
South Africa	103	116	123	131	138	148	153	161
Spain	**24,824**	**25,100**	**25,264**	**25,428**	**25,613**	**25,857**		**26,070**
Turkey	3174	3258	3336	3397	3461	3520	3584	3641
United Arab Emirates	105	111	119	126	137	146	157	165
United Kingdom	**26,771**	**27,510**	**28,131**	**28,446**	**28,734**	**29,427**	**30,076**	**30,615**
United States	63,006	65,068	66,385	67,682	68,934	71,078	73,431	75,670

Table 3 The strong correlation of the two previous epidemic diseases: Information on these data has collected from the World Health Organization

Influenza	COVID
It causes respiratory disease, asymptomatic	It causes basic, respiratory disease and asymptomatic
Transmitted through droplets, cough	Also transmitted through droplets, sneeze, direct contact
The interval range of this disease is 3 days	The incubation period of this disease is 1–14 days
It shows mild symptoms only	It will also show mild symptoms
Mortality ranges is (0.1%) lower	The mortality ratio between 3–4%
Children are highly affected	Adult cases are more
Antivirals and vaccine are available	No vaccine or therapeutic

According to this disease, the coronavirus and the flu virus have a certain similarity, which will be represented in the above table

they cause low morbidity and transmission between people. In contrast, nCOVID-19 has wreaked havoc with humans. The result of this high mortality varies with the short term. Therefore, the zoonotic virus, SARS-CoV, was transmitted through bats and civet cats and originated in 2002. Probably in 2012, MERS-CoV was identified, which is transmitted through bats to the camel. The clinical manifestation of both viruses is fever, chills, dyspnoea, myalgia, respiratory problems, diarrhea, general malaise, dilemma and pneumonia. Compared to other zoonotic diseases, nCOVID-19 has tremendous vigor. The 0th death case of MERS-CoV [8] is reported in Saudi Arabia (Jeddah); this virus is belonging to the lineage C of the *betacoronavirus* in which it causes severe respiratory illness to the people. In July 2013, total, 91 cases were identified from the Arab peoples and the fatality range is 50%. Simultaneously,

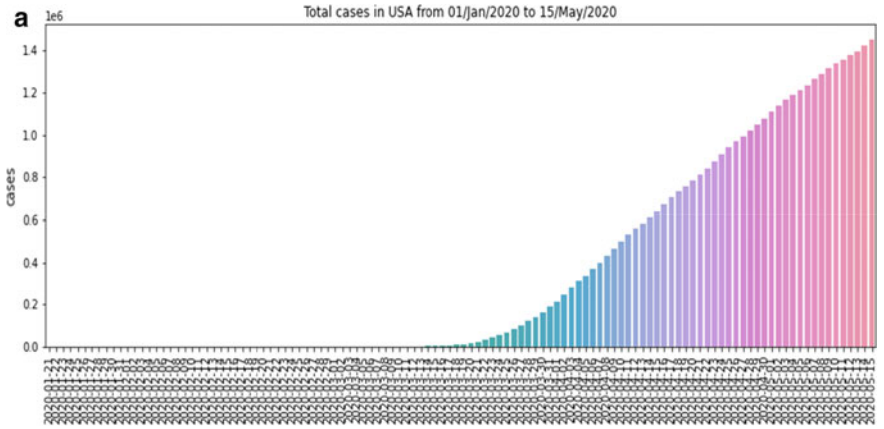

Fig. 1 **a**. USA COVID + cases **b**. USA death cases

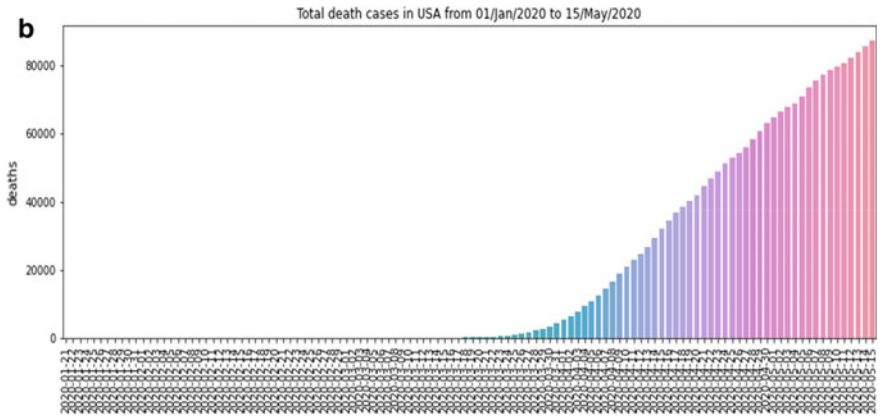

Fig. 1 (continued)

another 27 International countries including UK, France, Italy, Germany also have found and reported to the World Health Organization.

The MERS-CoV infected person is strongly confirmed by the laboratory test; for this in the medical industry, a sample of sputum [9] is collected from the infected person and then the samples are tested by using real-time RT-PCR laboratory method. In 2009, influenza family viruses [10], termed airborne diseases, threatened people. And these have made its connection to the population through droplets that usually come from sneezing. Along with respiratory viruses, H1N1, H5N1 and H5N7 are also included. And these are related in SARS and MERS. We obtained the data through several online sources. Their data list and associated mortality rates are explained in further sections (Table 4).

Table 4 Specimen ranges. https://apps.who.int/iris/bitstream/handle/10665/331329/WHO-COVID-19-laboratory-2020.4-eng.pdf

Type of the specimens collected from the infected/suspected cases	Temperature range to store the collected specimens
Nasopharyngeal/Oropharyngeal	2–8°C if $\leq$ 5 days–70°C (dry ice) if > 5 days
Sputum	2–8°C if $\leq$ 2 days–70°C (dry ice) if > 2 days
Serum	2–8°C if $\leq$ 5 days–70°C (dry ice) if > 5 days
Blood	2–8°C if $\leq$ 5 days–70°C (dry ice) if > 5 days
Urine	2–8°C if $\leq$ 5 days–70°C (dry ice) if > 5 days
Stool	2–8°C if $\leq$ 5 days–70°C (dry ice) if > 5 days

3 Pneumonia Infection

Pneumonia, which is caused by some fungi or negative bacteria in the human body, is attacked by various soft parts of the body such as the throat, lungs and blood vessels. They cause a variety of diseases. These include the following.

- Breathless
- Nausea and vomiting
- Kidney damage
- Mental disorder
- Coughing
- Color changes in many organs of the body
- Chest pain
- Cancer (Fig. 2).

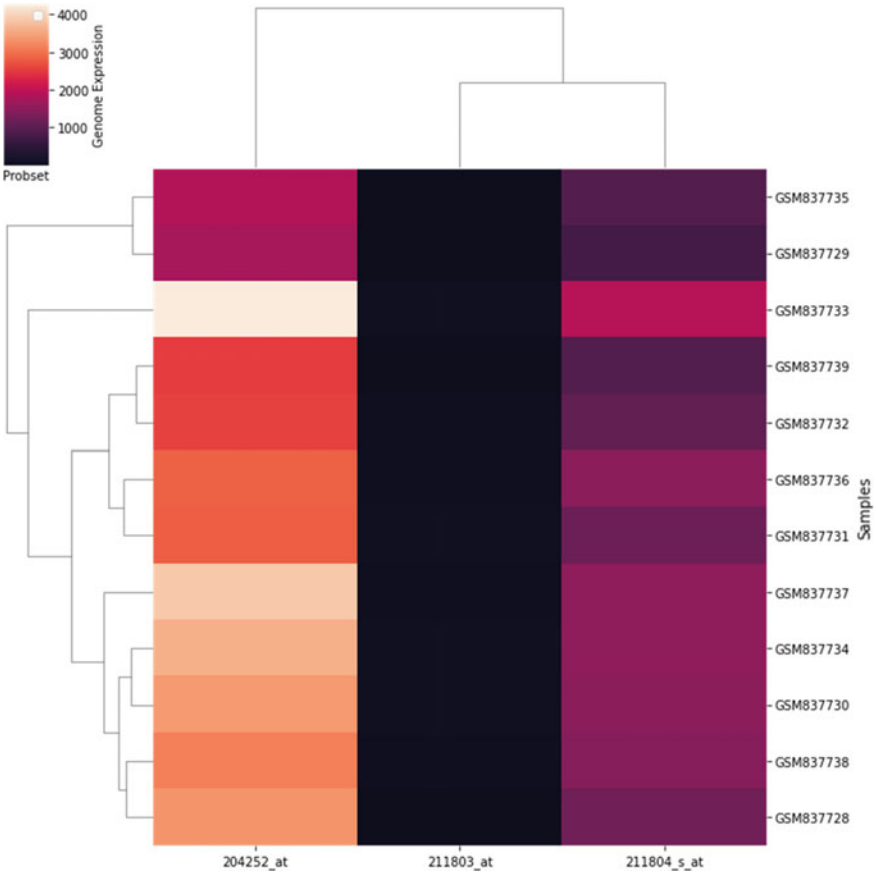

Fig. 2 The probable sequence of genomes for lung cancer [29, 30]

Pneumonia is also one of the routes to travel the infectious disease (Coronavirus) from one person to another through the droplets, simultaneously. Clinicians prefer chest X-ray test to diagnose the active pneumonia from the respirator. As the proof of, a study published by Sufang Tian et al. in Journal of Thoracic Oncology (IASLC Special Report) found that two people had coronavirus from lung cancer.

4 Pneumonia with COVID-19

Initially, the two cases were discovered from the province of china. At first, both patients are admitted for the lung cancer treatment [11]. First, an 84-year-old woman was admitted to Wuhan, China, for treatment of lung cancer due to increased stress, and doctors prescribed a CT-scan for the patient. Pneumonia infection in the lung was detected by scan report. Because of this, the patient was transferred to the special ward and the medical test was carried out with the swap spaceman. At the end of the report, it was revealed that the patient was infected with coronavirus.

Next, a 73-year-old man was admitted to lung cancer surgery. He was slightly healed and discharged in a few weeks. Shortly thereafter, coughs with fever were frequent, and she was taken back to the hospital for examination, and CT scans and nucleic acid tests revealed that she and her pneumonia had been affected. In the end, both patients were diagnosed with hypertension at an early age, comparing the outcome. The below Fig. 3a the normal and COVID-19 patient lung X-ray image, which are collected from, https://github.com/ieee8023/covid-chestxray-dataset/tree/master/images [31].

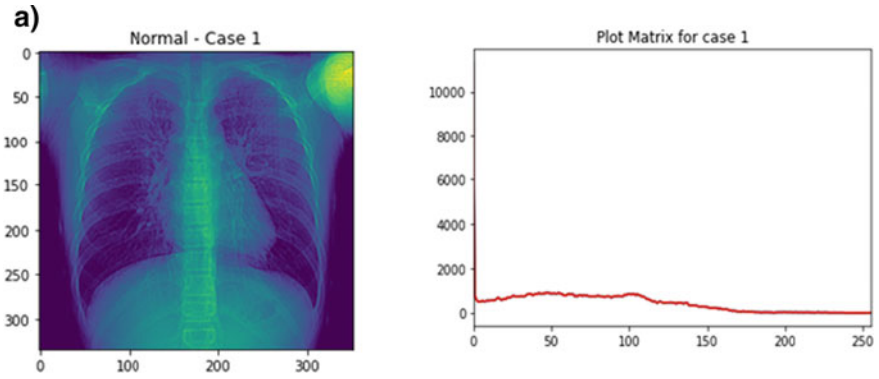

Fig. 3 **a**, **b**, **c**. Graphical representation of chest X-ray test image with plot matrix

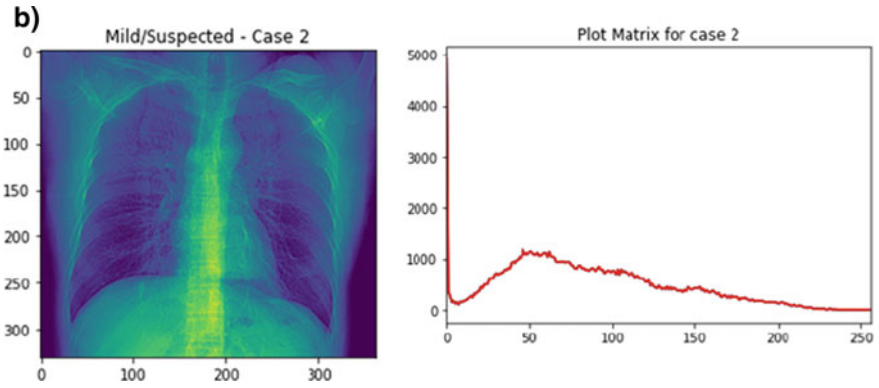

Fig. 3 (continued)

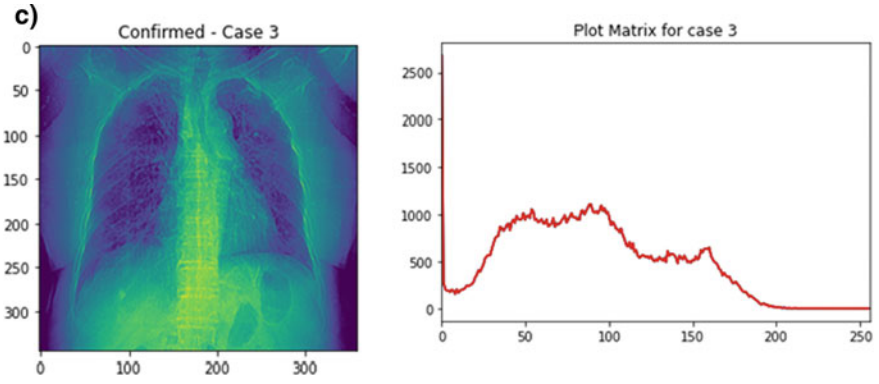

Fig. 3 (continued)

5 Related Literature in ML with Healthcare

For COVID-19 in viral infections, most medical examiners prefer X-ray testing to detect pathogens or pneumonia. Day by day, infectious disease is becoming more and more unpredictable in humans. In the medical field, chest X-ray is an important clinical trial that can be very useful for pathogen identification. Machine learning and deep learning enable examiners to identify the disease in its early stages using the confident accuracy of scan reports. Here, we reviewed the literature on the CAD-based predictive model. The literature helps to understand the reality of ML and DL prognostic models in the medical field (Fig. 4).

Anuja Kumar and Rajalakshmi [12] have proposed an automated detection model to identify pneumonia from chest X-ray imaging using deep Siamese NN. In chest X-ray, all viral pneumonia is captured as a white substance. The pathogen can spread to the left lung or to the right side. The convolution neural network has come into

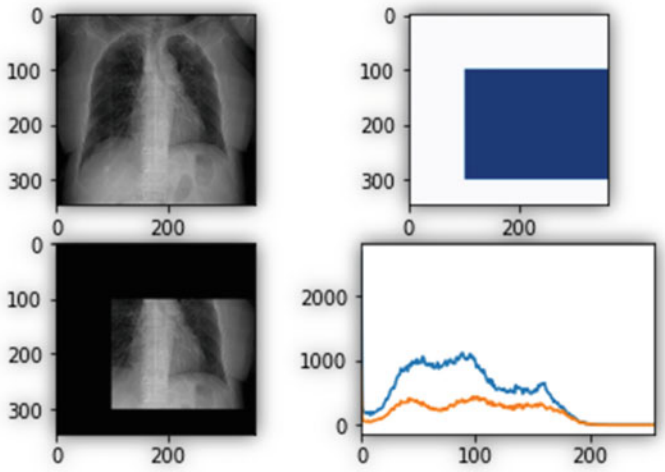

Fig. 4 COVID + lung X-ray with histo plot

the picture to extract image pixels. Initially, the CXR image is retrieved from the original image (mm) pixel size. After that, the transformed image is divided into two sections, like the left and the right. The Siamese network model has been used to calculate the end of the segment. To increase the efficiency of the proposed model, the author approaches the K-fold CV technique to divide the model into k-values. Furthermore, AUROC is used to identify the overall performance of the proposed model. Furthermore, the authors divide the problem into three categories, namely normal image, viral pneumonia picture and bacterial pneumonia.

Sebastian Gundel et al. [13] have proposed an abnormal classification based on deep learning using scan image. The model operates on a different task, such as division, abnormality and regional classification.

In the modern and e-technological world, wearable devices are the most common and most useful electronic devices for humans. These devices automatically detect human health situations such as headache, fever, BMI, heart rate, blood pressure, etc., based on the monitoring sensors. Paulo Resque et al. [14] investigate five main ML algorithms, such as: support vector machine, random forest, naive Bayes, KNN and neural network for the health problem of the epileptic seizure problem. Analysis-based work is performed using the patient's EEG dataset. For the implementation of the model, the R language was used to calculate the model. In the result of the evaluation of the model is calculated based on accuracy. Here, kappa statistics have also been used for comparison of results. The kappa statistics will be calculated using the formula [14] [Paulo et al.].

$$\text{KAPPA} = \frac{\text{Observed Accuracy} - \text{Expected Accuracy}}{1 - \text{Expected Accuracy}}.$$

Finally, the SVM model produces 97.31% accuracy with computational complexity O (n^3).

Ibrahim Almubark et al. [15] discussed the prediction of Alzheimer's disease using neuropathological patient data. Alzheimer's disease is a disease that primarily attacks the human neurological system. To improve the accuracy of the classification, the neurophysiological data set was used. The datasets are grouped into three parts: (i) common neurological test data, (ii) mild cognitive data indicating the reaction and response time, and (iii) combination of both. The authors choose four classification algorithms, namely SVM, RF, GB and AB are used to classify the data. For these three types of data, a total of three experiments are performed and fivefold cross-validations and leave-one-out are used to evaluate the model.

In the healthcare sector [16], computed tomography and mammography are the most used for making decisions, as if the patient was in normal or abnormal conditions. SVM and image processing are the techniques in which image processing plays a key role in the medical science to detect lung cancer through the acquired image. In today's life, most people admitted because of the cancer-oriented problem, the mortality rate also increases. Up to now, anticipation is much more important. The author explains "How are computed tomography images used to detect the defect?". Image processing techniques are used here to clean the X-ray image such as acquisition, feature extraction, segmentation, noise reduction, filter, etc. SVM is one of the supervised techniques; it helps to identify and classify the + ve and -ve ratio of cancer patients. For this application, the Java framework and JSP were used to build the application model.

We can say that chronic diseases are lifelong illnesses. Decision trees, random forest and SVM [17] [Swetha Ganikar] have been used to test whether or not the patient is infected. Chronic diseases such as diabetes, liver and heart disease are collected from an open database. All data were used for each method. The RF generates 98% accuracy in the benchmarking phase.

Thirunavukarasu et al. [18] discussed the prognosis of liver disease. Because of the length of health records, the author uses machine learning classification techniques to find the hidden area for the best predictive model. Supervised learning methods, such as KNN, LR, and SVM, were used to predict the disease. Therefore, the performance of the proposed models is calculated using the confusion matrix of the proposed model.

Muhammad Imran Faisal et al. [19] used machine learning classifiers and ensemble classification techniques to detect lung cancer based on its symptoms. The UCI reference dataset is used for this analysis. 10-fold cross validation of training and testing data after the initial data preprocessing phase. The data are applied to each ML method and set the classifier to choose the best classifier for the problem.

Israeli AI-Tauraika [20] and others have discussed MERS-CoV, a family of Coronaviridae, which mainly affects animals. The authors use data mining techniques to develop a predictive model of MERS virus. Approximately, 1082 records of patient data were collected to create the model. The NB classifier and J48 methods were used to validate the model. After processing the data, the collected data will be applied to the selected sample. J48 and NB is a well-known classifiers and supervised learning

methods for creating tree-like image after classification. Using the WEKA software tool, the predictive model has been successfully built, and with the use of 10-fold cross-validation techniques, both the stability model and the recovery model yield the most accurate result. The stability model provides 55.69% accuracy for the J48 method compared to NB and J48. The patient recovery model provides 71.58% accuracy for the NB classification.

6 Roles of Machine Learning

Artificial intelligence is a computer-based technology, also known as "machine intelligence." Artificial intelligence is more powerful for making multitrack decisions and can learn the data set itself to predict the future. Today, AI technologies are used in multiple industries. Artificial intelligence extends the technique to machine learning. Hence, ML is also called as a subset of AI. It also further expands another subset called "deep learning." Artificial intelligence aims to develop and build the computer, and efficiently learn large amounts of data. The technology helps machine learning to extract the relevant functionality of the large synthetic data set. The term "machine learning" was coined by the computer scientist "Arthur Samuel" in 1959, who was the developer of the artificial intelligence model. The main role of ML is to learn any type of data. Machine learning has the ability to predict the future, so it plays on multiple disciplines. It is also possible to perform automatic operations such as data extraction, data grouping, data classification, etc.

Artificial intelligence and machine learning are an emerging technology, currently used by the era all over the world, especially in the medical field with the name of medical image analysis, including magnetic resonance imaging, X-rays and/or computed tomography. We know that in the hospital sector, computer scanners are used to diagnose the level of disease along with the patient's clinical symptoms. Sometimes, the result can lead to low accuracy due to the medical report. This is the problem that will normally occur in the medical area.

The era of machine learning technology offers numerous classification and prediction algorithms to determine the most accurate result of severely infected cases. Recently, coronavirus has caused major outbreaks on earth. Many people have been seriously affected, so the graphical representation of the report will be shown in Fig. 1a, b. In the pandemic situation, research-oriented results will be the most common need. The main strength of this document is to discuss the role of the machine learning techniques most used in bioinformatics (Figs. 5 and 6).

As mentioned earlier, ML is a subset of AI, taking the initiative to learn the data itself. It plays a key role in medical forecasting using the statistical and probabilistic classification technique. For example, in the medical field, early diagnosis of the disease is a difficult task. But in machine learning, it offers many classification and prediction techniques to predict disease at an early stage. In addition, in the recent pandemic, viral disease is a COVID-19 that spreads through the flu drops and is

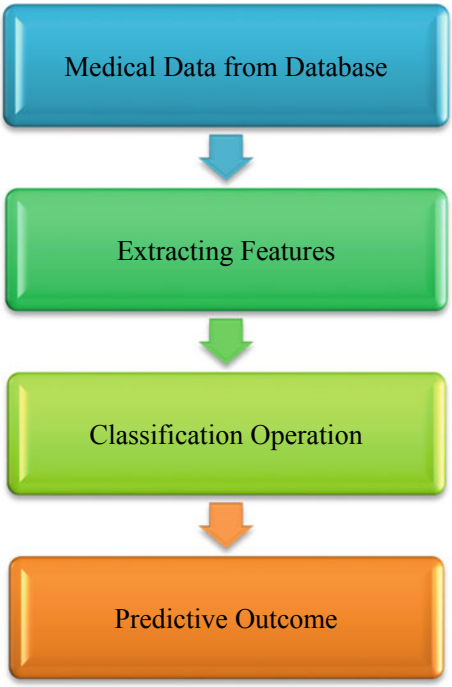

Fig. 5 Disease prediction using machine learning

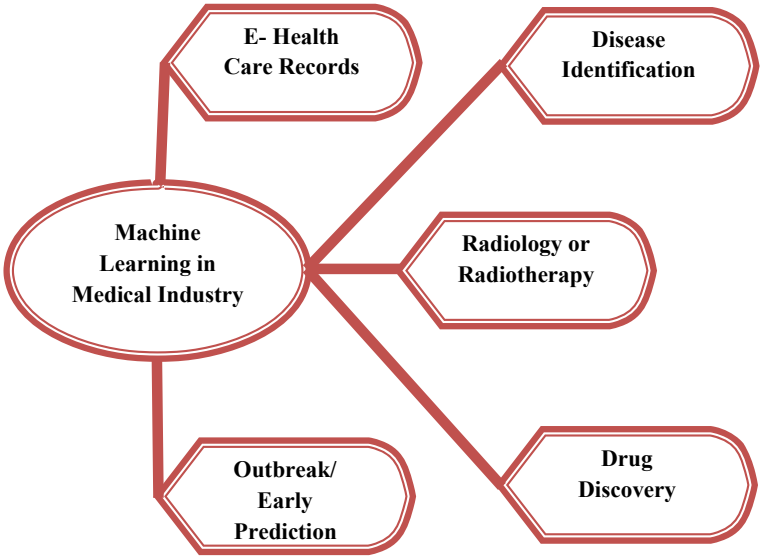

Fig. 6 ML in medical

mainly infected in the chest and lung areas. For this, clinical examiners suggest tests and radiographic results to diagnose the infection.

Although due to the lack of small medical imaging problems, we sometimes get low precision results. For this, machine learning helps to avoid this problem during the disease detection phase. Machine learning uses several methods to process data, including [21], supervised, unsupervised, semi-supervised, strengthening, multi-tasking, ensembles, neural networks and instance-based learning. Primarily, it can be divided into two parts, (i) supervised learning and (ii) unsupervised learning. These two methods are the most used in medical diagnosis.

6.1 Supervised Learning

Supervised learning is one of the machine learning methods used to perform some statistical operations based on classification and regression. Structured data called "*labeled data*" was used to train the model. In the supervised learning model, the data structure is already known. The data will be separated according to the known characteristics. For example, consider the DNA sequence [22]. We know that the human cell contains numerous proteins. Each of them has a unique sequence or type. The chromosome is unique and varies from one human model to another. Based on the sequence module, the algorithm will classify the genome. The learning model is further divided into two subparts.

- Regression
- Classification.

Work flow:

- Data collection/Preprocessing and data preparation
- Extract the data
- Data split up
- Apply the model and evaluate the performance
- Train the model again and again for better accuracy.

Example:

- Data classification—To classify, suspected versus infected case.

6.2 Unsupervised Learning

It is one of the learning models. When we are entering into unsupervised learning method, the label is not mandatory of the data. Basically, unsupervised learning does not know about the data. Based on the data correlation, the data will be segregated. Major operations are,

- Clustering
- Association.

In other words, the cluster-based analysis method is majorly used to get the several types of data and it makes as a group. The supervised learning is working under the specific rules, which is clearly defined. Unlike, unsupervised learning methods are working under the condition-based rules, or in other words, it observes the information from the unlabeled data.

Work flow:

- Data preparation
- To learn the information from the preprocessed data
- Set the centroid point
- Make the similar data as a group
- Assign the cluster of data to each centroid.

Example:

- Doctors and ward boys—the method will work based on the similar feature.

6.3 Semi-supervised Learning

Semi-supervised learning is a one kind of machine learning approach in which the method will handle both labeled and unlabeled data, for example, like in the ration of 40:60 approximately. In other words, it contains a small number of labeled data and huge amount of unlabeled data. So that this learning method is referred as a combination of supervised and unsupervised learning. The fruit-full applications like, speech analysis, web content classification and cell protein sequence classification are working under the semi-supervised learning methods. Multiple literature [22] reviewed, semi-supervised methods are most useful for the findings and also provide better accuracy due to have the capability to learn combined data, than the supervised model.

6.4 Reinforcement Learning

Machine learning distributes a unique category of learning methods called reinforcement learning, is an automated decision-making process. The model will learn the data by using the past experience, environment and it works under the reward based system, unlike supervised and unsupervised. In the artificial intelligence, it is a type of dynamic programming, has the capability to train the model, while in the case of data is absent. Suppose the result is not satisfied after the training phase, then the model can take up the punishment/or the reward to train the model one again. Until,

the process will continue, till the model gets the correct result based on the reward value. For this, the model is simply referred as an agent-reward-based model.

6.5 Neural Networks Learning

In general, the neural network is known as ANN, which is one of the learning processes. The main feature of ANN is the processing of the input elements, which automatically read the input data based on the characteristics. It is a revolutionary neural network that contains multiple layers of nodes, which will be treated like a neuron, used for pattern recognition. These are divided into three levels, namely: input layer (receive input), hidden layer (calculation) and the output layer (check the signal result). In general, the perceptron network is developed for the huge data set consisting of numerous attribute classes. This model contains the weight values of each neuron. The model will be trained until no error occurs. MLP levels are interconnected to synaptic links called edges which have some weight values for the calculation (Fig. 7).

The hidden layer receives input signals from the input layer through the communication link. For the purpose of processing values, each link will be assigned some weight values. Here, weight values are input information. After this process, the layer will calculate the net input value for the future process. It will be sent to the release layer called output layer. If the network is unable to give the desired output, the model will be re-trained with different weight values.

7 Methodology

7.1 Naïve Bayes Classifier

Based on the Bayes theorem, the naive Bayes algorithm will evaluate the probability of the class. The model is helpful, even if the attribute of the data belongs to some other attribute. Therefore, in machine learning, this model is also called "probability classifier." Finally, it will return the result based on the predicted class.

$$\text{Bayes theorem, } \frac{\mathrm{P(H|E)} = \mathrm{P(E|H).P(H)}}{\mathrm{P(E)}}$$

The NB classification is categorized into three parts: GaussianNB, MultinomialNB and Bernoulli model.

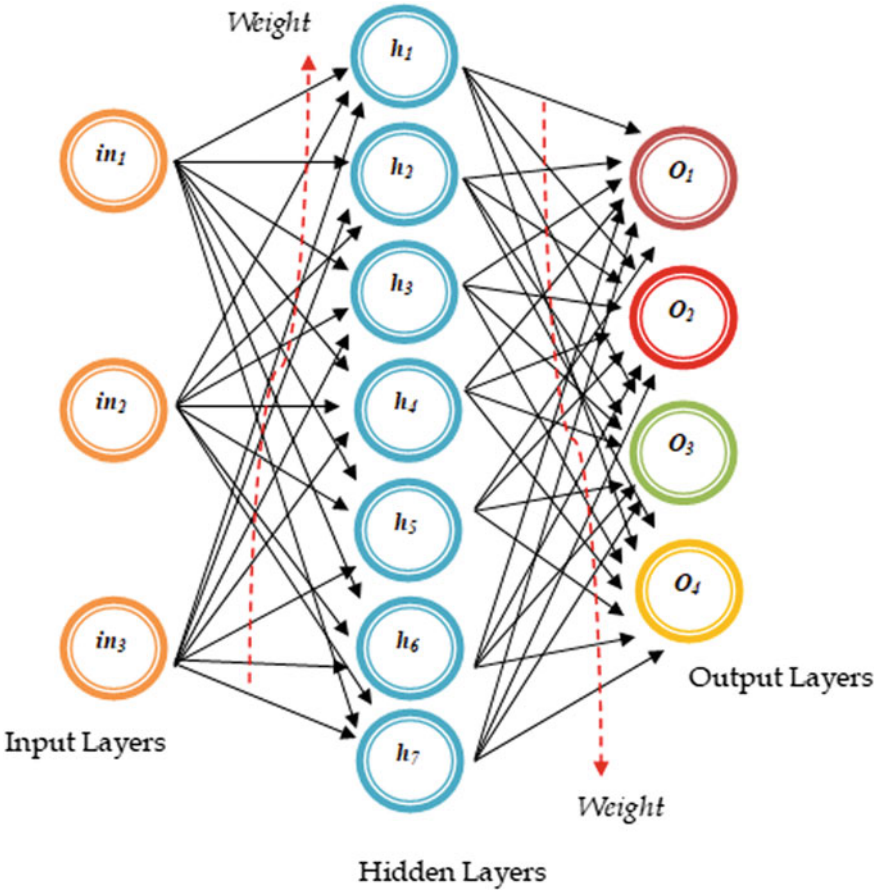

Fig. 7 Simple N-network architecture

Example:

Sweating	Vomiting	Dehydration	Fever	Fever (Multiclass)
Y	Y	Y	Y (1)	High
Y	N	Y	Y (1)	Medium
N	N	Y	N (0)	Normal
Y	N	N	N (0)	Normal
N	Y	Y	Y (1)	Prob
Y	Y	N	Y (1)	Prob
N	N	Y	N (0)	Normal

- ***Gaussian Model***: Typically, the model is used when the data are in numeric format
- ***Multinomial Model***: It deals with multiclass variables, which is why this model is called multinomial, which is suitable for text-based data.
- ***Bernoulli Model:*** This model is used, while the data are in vector-based (binary values) format.

7.2 Random Forest

It is an ensemble classification model, one of the supervised techniques. It can create multiple decision trees to classify attributes using "bagging." The model also separates the relevant features for classification. Once the model is trained, the output of each tree is integrated into a single group. The model predicts the final result based on the high number of votes cast on the tree. This is an ensemble classifier, which means that the properties are more deeply divided than a normal decision tree. This random forest classification does not allow for excessive matching (overfitting) of the data, which is a major advantage compared to other classifiers.

7.3 Linear SVC

Support vector clustering is a method that is commonly used for the hierarchical clustering problem. It is a non-parametric classification model, which automatically calculates the value with its method, regardless of the format of the data. Like the support vector machine, the kernel function is used to compile data. This method works by using a decision boundary or a hyperplane to separate the data points.

7.4 Feature Selection

The selection of characteristics is an important step in the problem of classification and forecasting. The feature selection model allows the machine to classify the variable for training. Manually, specific attributes cannot be extracted during big data. In other words, we do not know which class is associated with which class. Assuming that the decision is made by man, the model will have errors during the training phase. In addition, selecting an important class is also a complicated task. Therefore, machine learning provides a feature selection method (such as PCA) to extract relevant variables/attributes from large amounts of data. For example: Data classification, if the data set contains multiple variables, the machine needs preliminary knowledge to extract the feature to increase the degree of precision. The method is designed to,

(1) Reduce the processing time
(2) Simplify the problem
(3) Reach best accuracy score
(4) Reduce the data length
(5) Choose the best fit and data correlation.

8 Experimental Setup

For this experimental analysis, we collected primary tumor [32] data from the open dataset repository, which is a UCI repository in CSV format. The dataset contains 18 attributes (including age and gender) and 339 instances. In addition, the data set consists of a multiclass variable. So we import the MultinomialNB classifier from the sklearn. Based on the dataset we have, the most common type of cancer known as "adeno" is known to be the most pathogenic when looking at which histological type is most harmful (Figs. 8 and 9).

Python and Sklearn have been used for testing. Machine learning provides several useful guidelines for clinical prognosis. Sometimes, based on the data, the model will provide less accuracy during the training phase. To avoid such a problem, we initially conducted a general performance analysis of ML models (RF, Linear SVC, LR, and NP) with our clinical dataset. All models achieved full accuracy score (Fig. 10). To ensure this, we select two classifier models such as SVM and RF for the training phase.

As we have noted above, there are many types of histologic variants in the primary tumor database. Therefore, we have the responsibility of extracting the feature from

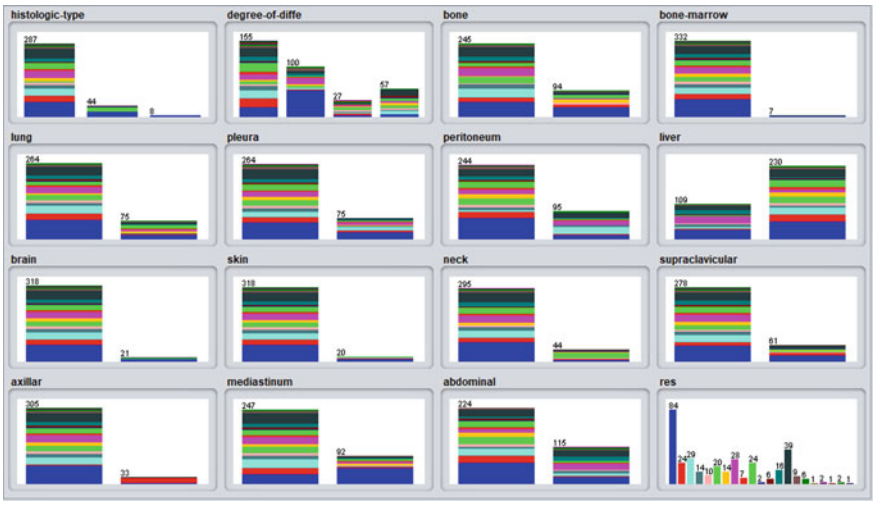

Fig. 8 Visualization of data attributes

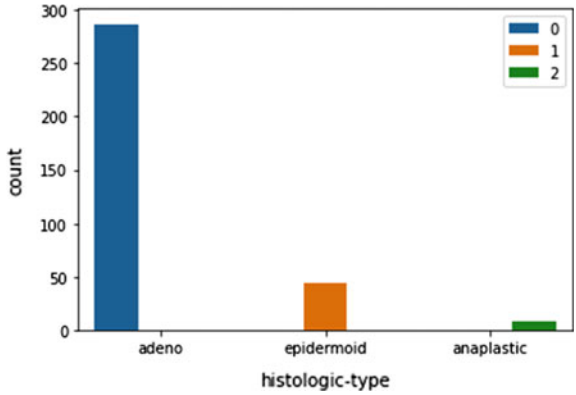

Fig. 9 Total count of histologic-type

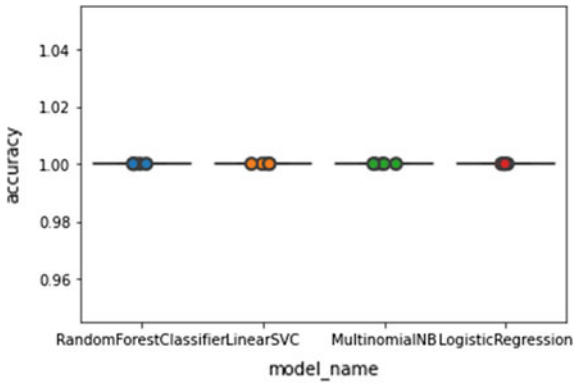

Fig. 10 Total accuracy of the selected model

the histologic (cancer type) type. Here, TfidfVectorizer and CountVectorizer are used to convert text data into a matrix format [33]. In this unique performance, we choose the random forest and linear support vector machine classifier to demonstrate model accuracy with the primary tumor dataset. After the two models work equally well, we obtain the same confusion matrix of the RF and linear SVC classifier (Figs. 11, 12).

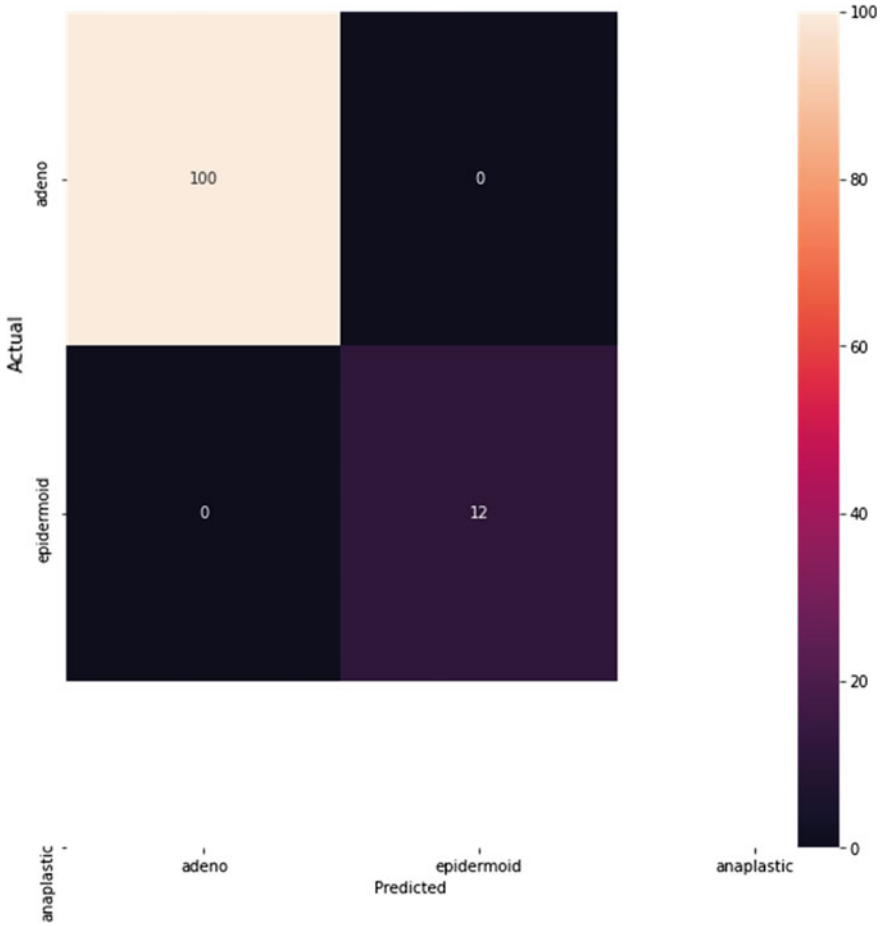

Fig. 11 Confusion matrix–linear SVC

Fig. 12 Confusion matrix–random forest

Predicted Species	Lungs	Class
Actual Species		
Lungs	**12**	**0**
Class	**0**	**100**

9 Publications Relevant to ML and DL for Medical Imaging

Author(s)	Aim	Model description	Methodology	Performance evaluator	Result and accuracy
Jie Ren, Kai Song et al. [23]	Discovering virus of metageonomic sequence data	Proposed "reference-free" and "alignment-free" method to disassemble the virus. DeepVirFinder has used to find the sequence of the viral genome (DNA)	Machine learning and deep learning-convolutional neural networks	AUROC	1. Viral seq (500): 95% 2. Viral seq (1000): 97% 3. Viral seq (3000): 98%
Zhenyu Tang, Wei Zhao et al. [24]	The assessment model to detect the COVID patients complication	Based on the quantitative features of the Coronavirus, the data have trained the model. Using lung CT images and the main features is helping the RF model to detect, if the infected patient is in severe or normal condition.	Random forest	AUC curve	1. Total accuracy: 87% 2. Accuracy in AUC: 91%
Ghanshyam Verma et al. [25]	Infected gene classification for viral respiratory infection	Early stage detection of viral infection using top most viral genes	KNN, linear SVM, RF and SVM with RBF	10-fold, hold-out	1. Overall accuracy in 10-fold: SVM with RBF 2. Overall accuracy in hold-out: Random forest
Okeke Stephen et al. [26]	Classification and detection of pneumonia infection in chest X-ray	The Convloutional neural network method has approached to analyze the pneumonia in the X-ray image using several network layers	Deep learning	Epochs for training and loss	1. Average accuracy of training set: 95% 2. Average accuracy of validation set: 93%
Dhiraj Dahiwade et al. [27]	General disease prediction with symptoms	The work was carried on the Java platform. The general disease patient dataset is used for the prediction. Based on the ML preprocessing, the training dataset is created for the problem analysis	KNN and CNN	–	1. Best algorithm: KNN 2. Best Time Complexity: CNN
Amani Yahyaoui et al. [28]	Diabetes prediction	Based on the decision support system (DSS), the ML and DL model has approached to predict the diabetes. The performance analysis of ML and DL has also conducted	SVM, RF and CNN	Kappa Co-efficient	1. Class Diabetic-82% (RF) 2. Class Non-diabetic-86.7% (SVM)
Shuaijing Xu, Hao Wu and Rongfang Bie [34]	Anomaly detection on chest X-ray	CXNet-m1 is a proposed network structure, was used to train the model and Softmax cross_entropy has approached to classify the X-ray image due to the format of binary	Deep neural network	F1 Score and AUC	1. Accuracy rate in old data: 67.6% 2. Accuracy rate in new data: 84.4% 3. Accuracy rate in OpenI: 93.6%

10 Conclusion and Future Work

COVID-19 is a serious viral problem, and early detection of this virus is a complex task because it is directly or indirectly linked to other viral genome proteins. Since December 2019, many countries have been enslaved by the disease and have killed countless human species. We have found many positive literature to solve such a genetic classification problem. Computer-assisted diagnosis (CAD) is one of the medical techniques that play a major role in the problem prediction area. The medical key findings of this disease are the human lungs that are the major influences. So, clinicians suggest X-ray or CT scans for medical examination to find the cause of the infection. Nowadays, many high quality technology and medical disciplines are helping to identify drugs and vaccines for this. However, systematic research on this is being carried out and implemented.

Through the daily COVID-19 statistics, we know that many people infected with the virus have reduced and fully recovered, despite the deaths caused by this quarantine or infection. Immunity is the first of all parts of the human body. Such diseases attack the body through immune deficiency. Early detection of viral immunity may prevent infection. For this purpose, in the pharmaceutical industry, the method of "immunotherapy" is being manipulated. In some cases, these methods may not be effective due to the high viral component. Therefore, in the near future, we will be working with machine learning and deep learning to make these computational models more important to a clinician in order to meet their difficulties and increase their accuracy.

References

1. Li, J., et al.: Machine learning methods for predicting human-adaptive influenza a viruses based on viral nucleotide compositions. Mol. Biol. Evol. **37**(4), 1224–1236 (2020). https://doi.org/10.1093/molbev/msz276
2. "Title Page," J. Sex. Med. **14**(5), e205 (2017), https://doi.org/10.1016/s1743-6095(17)31143-8
3. Madjid, M., Safavi-Naeini, P., Solomon, S.D., Vardeny, O.: Potential effects of coronaviruses on the cardiovascular system: a review. JAMA Cardiol. **10**, 1–10 (2020). https://doi.org/10.1001/jamacardio.2020.1286
4. Alam, M.M., Islam, M.T.: Machine learning approach of automatic identification and counting of blood cells. Healthc. Technol. Lett. **6**(4), 103–108 (2019). https://doi.org/10.1049/htl.2018.5098
5. Abbas, A., Abdelsamea, M.M., Gaber, M.M.: Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network, 2020, [Online]. Available: http://arxiv.org/abs/2003.13815
6. Luo, Y., et al.: Machine learning for the prediction of severe pneumonia during posttransplant hospitalization in recipients of a deceased-donor kidney transplant. Ann. Transl. Med. **8**(4), 82–82 (2020). https://doi.org/10.21037/atm.2020.01.09
7. Meo, S.A., et al.: Novel coronavirus 2019-nCoV: prevalence, biological and clinical characteristics comparison with SARS-CoV and MERS-CoV. Eur. Rev. Med. Pharmacol. Sci. **24**(4), 2012–2019 (2020). https://doi.org/10.26355/eurrev_202002_20379

8. Salamatbakhsh, M., Mobaraki, K., Sadeghimohammadi, S., Ahmadzadeh, J.: The global burden of premature mortality due to the middle east respiratory syndrome (MERS) using standard expected years of life lost, 2012 to 2019. BMC Publ. Health **19**(1), 1–7 (2019). https://doi.org/10.1186/s12889-019-7899-2
9. Cho, S.Y., et al.: MERS-CoV outbreak following a single patient exposure in an emergency room in South Korea: an epidemiological outbreak study. Lancet **388**(10048), 994–1001 (2016). https://doi.org/10.1016/S0140-6736(16)30623-7
10. Ramanathan, K., et al.: Transmission of SARS and MERS coronaviruses and influenza virus in healthcare settings: the possible role of dry surface contamination. J. Hosp. Infect. **92**(January), 235–250 (2020)
11. Tian, S., Hu, W., Niu, L., Liu, H., Xu, H., Xiao, S.Y.: Pulmonary pathology of early-phase 2019 novel coronavirus (COVID-19) pneumonia in two patients with lung cancer. J. Thorac. Oncol. **15**(5), 700–704 (2020). https://doi.org/10.1016/j.jtho.2020.02.010
12. Acharya, A.K., Satapathy, R.: A deep learning based approach towards the automatic diagnosis of pneumonia from chest radio-graphs. Biomed. Pharmacol. J. **13**(1), 449–455 (2020). https://doi.org/10.13005/bpj/1905
13. S. Guendel et al.: Multi-task learning for chest X-ray abnormality classification on noisy labels, pp. 1–10 (2019), [Online]. Available: http://arxiv.org/abs/1905.06362
14. Resque, P., Barros, A., Rosario, D., Cerqueira, E.: An investigation of different machine learning approaches for epileptic seizure detection. 15th Int. Wirel. Commun. Mob. Comput. Conf. IWCMC **2019**, 301–306 (2019). https://doi.org/10.1109/IWCMC.2019.8766652
15. Almubark, I., Chang, L.C., Nguyen, T., Turner, R.S., Jiang, X.: Early detection of alzheimer's disease using patient neuropsychological and cognitive data and machine learning techniques. Proc.–2019 IEEE Int. Conf. Big Data, Big Data 2019. **2**(Mci), 5971–5973 (2019), https://doi.org/10.1109/bigdata47090.2019.9006583
16. Rahane, W., Dalvi, H., Magar, Y., Kalane, A., Jondhale, S.: Lung cancer detection using image processing and machine learning healthcare. Proc. 2018 Int. Conf. Curr. Trends Towar. Converging Technol. ICCTCT 2018, pp. 1–5 (2018). https://doi.org/10.1109/icctct.2018.8551008
17. Ganiger, S., Rajashekharaiah, K.M.M.: Chronic diseases diagnosis using machine learning. 2018 Int. Conf. Circuits Syst. Digit. Enterp. Technol. ICCSDET 2018. pp. 1–6 (2018). https://doi.org/10.1109/iccsdet.2018.8821235
18. Thirunavukkarasu, K., Singh, A.S., Irfan, M., Chowdhury, A.: Prediction of liver disease using classification algorithms. 2018 4th Int. Conf. Comput. Commun. Autom. ICCCA 2018. **6**(9), 1–3 (2018). https://doi.org/10.1109/ccaa.2018.8777655
19. Faisal, M.I., Bashir, S., Khan, Z.S., Hassan Khan, F.: An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer. 2018 3rd Int. Conf. Emerg. Trends Eng. Sci. Technol. ICEEST 2018. pp. 1–4 (2019). https://doi.org/10.1109/iceest.2018.8643311
20. Al-Turaiki, I., Alshahrani, M., Almutairi, T.: Building predictive models for MERS-CoV infections using data mining techniques. J. Infect. Public Health **9**(6), 744–748 (2016). https://doi.org/10.1016/j.jiph.2016.09.007
21. Dey, A.: Machine learning algorithms: a review. Int. J. Comput. Sci. Inf. Technol. **7**(3), 1174–1179 (2016). [Online]. Available: www.ijcsit.com
22. Libbrecht, M.W., Noble, W.S.: Machine learning applications in genetics and genomics. Nat. Rev. Genet. **16**(6), 321–332 (2015). https://doi.org/10.1038/nrg3920
23. Ren, J., et al.: Identifying viruses from metagenomic data using deep learning. Quant. Biol. **8**(1), 64–77 (2020). https://doi.org/10.1007/s40484-019-0187-4
24. Tang, Z. et al.: Severity assessment of coronavirus disease 2019 (COVID-19) using quantitative features from chest CT images. **2019**, pp. 1–18 (2020). [Online]. Available: http://arxiv.org/abs/2003.11988
25. Verma, G., Jha, A., Rebholz-Schuhmann, D., Madden, M.G.: Using machine learning to distinguish infected from non-infected subjects at an early stage based on viral inoculation. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). **11371 LNBI**(November), 105–121 (2019). https://doi.org/10.1007/978-3-030-06016-9_11

26. Stephen, O., Sain, M., Maduh, U.J., Jeong, D.U.: An efficient deep learning approach to pneumonia classification in healthcare. J. Healthc. Eng. **2019**, (2019). https://doi.org/10.1155/2019/4180949
27. Dahiwade, D, Patle, G., Meshram, E.: Designing disease prediction model using machine learning approach. Proc. 3rd Int. Conf. Comput. Methodol. Commun. ICCMC 2019. Iccmc, 1211–1215 (2019). https://doi.org/10.1109/iccmc.2019.8819782
28. Yahyaoui, A., Jamil, A., Rasheed, J., Yesiltepe, M.: A decision support system for diabetes prediction using machine learning and deep learning techniques. 1st Int. Informatics Softw. Eng. Conf. Innov. Technol. Digit. Transform. IISEC 2019—Proc. 2, 1–4 (2019). https://doi.org/10.1109/ubmyk48245.2019.8965556
29. https://www.kaggle.com/paultimothymooney/coronavirus-genome-sequence
30. Human lung cancer genomes, http://biogps.org/dataset/tag/lung%20cancer/
31. Total death and confirmed cases-COVID-19, Github, https://github.com/datasets/covid-19/tree/master/data
32. Dataset-Primary Tumor, https://datahub.io/machine-learning/primary-tumor
33. Scikit-learn.org
34. Xu, S., Wu, H., Bie, R.: CXNet-m1: anomaly detection on chest X-Rays with image-based deep learning. In: IEEE Access, vol. 7, pp. 4466–4477 (2019). https://doi.org/10.1109/ACCESS.2018.2885997