Accurate computation: COVID-19 rRT-PCR positive test dataset using stages classification through textual big data mining with machine learning

Shalini Ramanathan¹ · Mohan Ramasundaram¹

Accepted: 16 December 2020 / Published online: 4 January 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC part of Springer Nature 2021

Abstract

In every field of life, advanced technology has become a rapid outcome, particularly in the medical field. The recent epidemic of the coronavirus disease 2019 (COVID-19) has promptly become outbreaks to identify early action from suspected cases at the primary stage over the risk prediction. It is overbearing to progress a control system that will locate the coronavirus. At present, the confirmation of COVID-19 infection by the ideal standard test of reverse transcription–polymerase chain reaction (rRT-PCR) by the extension of RNA viral, although it presents identified from deficiencies of long reversal time to generate results in 2–4 h of corona with a necessity of certified laboratories. In this proposed system, a machine learning (ML) algorithm is used to classify the textual clinical report into four classes by using the textual data mining method. The algorithm of the ensemble ML classifier has performed feature extraction using the advanced techniques of term frequency–inverse document frequency (TF/IDF) which is an effective information retrieval technique from the corona dataset. Humans get infected by coronaviruses in three ways: first, mild respiratory disease which is globally pandemic, and human coronaviruses are caused by HCoV-NL63, HCoV-OC43, HCoV-HKU1, and HCoV-229E; second, the zoonotic Middle East respiratory syndrome coronavirus (MERS-CoV); and finally, higher case casualty rate defined as severe acute respiratory syndrome coronavirus (SARS-CoV). By using the machine learning techniques, the three-way COVID-19 stages are classified by the extraction of the feature using the data retrieval process. The TF/IDF is used to measure and evaluate statistically the text data mining of COVID-19 patient's record list for classification and prediction of the coronavirus. This study established the feasibility of techniques to analyze blood tests and machine learning as an alternative to rRT-PCR for detecting the category of COVID-19-positive patients.

Keywords COVID-19 · RT-PCR test · Machine learning · Text data mining · TF-IDF · Feature extraction · Classification

Extended author information available on the last page of the article

1 Introduction

The epidemic disease caused by COVID-19 requires an extraordinary response of intensity. There are more than 150 states around the world affected by corona. To handle the spread of the COVID-19 infection, worldwide governments and millions of residents have taken extreme measures, such as quarantine. Symptomatology of COVID-19 showed a large number of patients who were infected by corona, but some of the patients were also affected by corona asymptotically. These efforts are differentiating between the corona test positive and negative with the limited problems individually. Thus, the stages of identifying the SARS-CoV-2 virus have been believed to be crucial to recognize positive cases and thus control the pandemic. Therefore, the current trial of choice is the RT-PCR based on respiratory specimens examination performed in the laboratory. The automatic, reliable classification algorithms are helpful for training COVID-19 cases by considering the number of patients. The high demand generally is known for nasopharyngeal swab tests named as rRT-PCR due to the extension of worldwide virus that highlights the type of diagnosis limitations on a large scale, such as the expensive equipment, trained personnel, reagents for demanding things that can easily overcome supply, and at the turnaround time, the need of laboratories' certificate. For instance, the shortage of specialized laboratories and reagents forced the government to limit the testing of swab those who showed clearly the symptoms of SARS-CoV, thus leading to several virus-infected people and infection rates that were underestimated largely.

The laboratory medicine useful for easy analysis of coronavirus by using a simple blood test might aid to recognize the positivity/negativity of COVID-19 through rRT-PCR tests. This work consideration motivated us strongly to apply an advanced method of machine learning to routine and to evaluate the stages of COVID-19 infection for the feasibility of a predictive model. The proposed research classifies the stages using various techniques; the positive case records are available in the UCI repository of an original raw dataset in the proposed text data mining process to classify the stages into three types. A useful and more accessible, accurate, less expensive, and faster COVID-19 classification was proposed in this research.

2 Literature survey

Due to the spread of COVID-19, several territories and countries have been experiencing an increasing number of infected cases and deaths which remain a real threat to the public health sectors (Jamshidi et al. [1]). The research extracts a response to the struggle of the virus through AI and some deep learning (DL) techniques which have been demonstrated to reach the goal, including extreme learning machine (ELM) and generative adversarial networks (GANs). A user-friendly platform describes a combination of a bioinformatics approach with

different aspects from structured and unstructured data sources that are randomly put together for researchers and physicians. The recent COVID-19 publications and the medical reports were examined to choose both inputs and targets that might simplify to reach a consistent artificial neural network (ANN)-based tool for experiments associated with COVID-19. Research and diagnostics capable of deep learning on chest radiographs image classifier are based on COVID-Net, which were obtainable to classify chest X-ray images (Wang et al. [2]). This survey model aims to transfer knowledge for organizing and integrating images of chest X-ray according to three labels: regular, COVID-19, and viral pneumonia. Depending upon the accuracy of loss values, the models of ResNet-101 and ResNet-152 with the better effect of fusions improved dynamically by their ratio weights during their training process. This improved technology has produced higher sensitivity than radiologists in the diagnosis and screening of lung nodules. 96.1% accuracy was achieved by analyzing corona and classifying the type of chest image on the rest set.

Diagnosis of COVID in a timely manner through tomography is essential for both patient care and disease control (Li et al. [3]). Computer tomography (CT) is analyzed as a useful tool for corona diagnosis, yet the disease outbreaks have placed tremendous pressures on reading radiologists and potentially lead to fatigue-related misdiagnosis. In this work, we propose a novel approach for effective and efficient COVID-19 classification networks training using a small number of COVID-19 CT examinations and an archive of negative samples. Experimental results showed that the research is achieved as superior performance consuming about half of the negative sample cases, extensively reducing a model of training time. Several laboratories have confirmed that corona cases have been identified in an alarming rate with reportedly confirmed more than 2.2 million cases as of April 20, 2020 (Chamola et al. [4]). Numerous false reports, unsolicited fears, and misinformation regarding this virus were regularly circulated since the outbreak of the corona. In this survey, the use of technologies such as artificial intelligence (AI), 5G, Internet of things (IoT), blockchain, and unmanned aerial vehicles (UAVs), among others, was explored to mitigate the impact of the COVID-19 outbreak.

The platform of COVID-19 provides a quick diagnostic through serology testing, and molecular testing is also the important method to control the epidemic corona outbreak (Charizadeh et al. [5]). COVID-19 life cycle manages various stages: the preparedness phase, preventive phase, recovery phase, and response phase. The viral distribution of spatial and temporal RNA, antibiotics, and antigens at the time of

Fig. 1 Proposed block diagram

Fig. 2 COVID-19 rRT-PCR molecular test

corona infection to humans has shortened an immortal biological treatment for accurate analysis of COVID-19 diseases. The training provides the advanced encouragement of COVID-19 pandemic improvements in our global public health sector to realize a better struggle against outbreaks in the future (Figs. 1, 2).

3 Proposed methodology

3.1 Data collection

WHO declared the COVID-19 epidemic a health emergency. The researchers and hospitals have been giving open access regarding the corona pandemic data. The record has been collected from the open-source data repository from UCI, in which several corona-positive patient data are stored, as shown in different stages presented in Fig. 3. The original raw dataset of COVID-19 information is collected through the repository from medical data. Each attribute was collected from sample data of swab testing rRT-PCR. The proposed method using the COVID-19 data record is analyzed using advanced tools of machine learning techniques. The doctors will diagnose the pandemic coronavirus disease by taking a specimen swab test for the person affected. The data consist of several attributes, namely patient id, sex, offset, age, survival, needed supplemental O2,

Fig. 3 Overall proposed methodology

temperature, intubation, leukocyte count, lymphocyte count, neutrophil count, view, folder, date, file name, modality, location, DOI, and URL [6–8].

Since the dataset is a work of text, data mining can easily extract clinical notes and data findings. Clinical notes of COVID-19-positive cases' sample text record consist of text as the attribute finding is a label of the corresponding query text. Our dataset has three classes: mild, moderate, and severe, which consist

of clinical text of corona stages being categorized and the corresponding report length.

3.2 Machine learning

The novel coronavirus 2019, which has been termed as pandemic by the World Health Organization (WHO), has placed the world's numerous governments in a risky position. The outbreak of COVID-19, whose impacts were previously witnessed by the China citizens alone, has become a concern of every country virtually throughout the world [9–15] (Table 1).

3.2.1 Data preprocessing

The text data are unstructured, which need to be advanced such that machine learning techniques can be done. Various steps are being followed in this phase. The text is being scrubbed by removing the excessive text. The dataset consists of original raw data of the proposed system, with some noise present in it, so that the data preprocessing is used to filter the noisy and irrelevant data.

3.2.2 TF-IDF techniques

The machine learning techniques used term frequency–inverse document frequency (TF/IDF) for the text data mining process. The proposed system defines the use of text data retrieval from a huge amount of corona-positive data, which are distributed through a text and stored in a search engine using TF-IDF techniques which were used as retrieval schemes from search engine for classifying complete search text record. The results show that the accurate prediction of COVID-19 stages classification was expressively improved by exploiting features by text data retrieval. The next stage considers overturned lists according to those searching query words and finally sorts the target file from the record of searching index lists.

3.2.2.1 Feature extraction Term frequency–inverse document frequency (TF-IDF) is common in which a weighted statistically and broadly used in text analysis and text data retrieval. TF-IDF obtains one word that has a high frequency in one record of the file; if this word appears often, then it can be conserved as the main keyword to differentiate this file from one another. Term frequency (TF) is a time word performing in this record; fundamentally, a searching name with high reality is correlated with this file [16–20].

TF is defined as:

$$TF_{ij} = \frac{e_{ij}}{\sum_k e_{k,j} + 1}. \quad (1)$$

Inverse document frequency (IDF) is defined as:

Table 1 Proposed specimen type with temperature

Type of specimen	Materials collection	Storage temperature Until testing takes place in country laboratory	Recommended temperature for shipment according to expected shipment time
Nasopharyngeal and oropharyngeal swab	Dacron or polyester flocked swabs	2–8 °C	2–8 °C if ≤ 5 days –70 °C (dry ice) if > 5 days

$$IDF_i = \log \left(\frac{C}{K_i + 1} \right). \quad (2)$$

In Eq. (1), ‘ e ’ is the epoch word, $e_{k,j}$ is the sum of all the searching words in the file, and 1 is added in the denominator to avoid it from becoming zero.

In the IDF equation, ‘ C ’ as w_i , mentions the size of the word and similarly 1 is added in the denominator to avoid it becoming equivalent to zero, and ‘ k_i ’ is the integer of word file collection. Combining TF with IDF is essentially using TF to modify, which specifies the weight of the word W_i infiled j .

$$W_{i,j} = TF_i \times IDF_i \quad (3)$$

Figure 3 shows the overall proposed methodology of COVID-19 stages classification by using the improved machine learning techniques such as TF-IDF which gives a full data text retrieval method.

Algorithm 1: Machine Learning-Based Text Retrieval Process

Input:

C : the complete documents of the training set

U : the unique terms in all documents

Output:

Statistically Weight Matrix

Procedure:

- 1 for each term $t_i \in U$ do
- 2 for each document $d_j \in C$ do
- 3 w_{ij} = number of occurrences of term t_i in document d_j .
- 4 $W_{i,j} = TF_i \times IDF_i$
- 5 End for of complete document
- 6 End for of term

Features extraction of the testing report of COVID-19 was analyzed by various methods of sample testing for confirming a corona disease. The index value

was matched with the query values for analyzing in which stages the patients are affected mostly, which will be helpful for further decision-making schema. There are many methods for swab testing, and finally storing the data from the dataset of a repository with the original data is used to predict the classification stage of COVID-19.

3.3 COVID-19 stages classification

Classification of coronavirus stages has become practically a field in the proposed research due to the increased key procedures used for establishing the feasibility by indeed assigning a set of forms into predefined groupings based on their entire content, which contains a similarity matching model, word count model, word tagging model, machine learning methods, and so on. And m_i can be defined as a vector with word having statistical weights of unstructured entire text data of corona-positive record. It is measured as shown in Fig. 4.

$$m_j = \langle W_{1,i}, W_{2,i}, W_{3,i}, \dots, W_{n,i} \rangle. \quad (4)$$

Using machine learning techniques, positive corona cases were identified using several types of corona stages and were classified under the three stages of mild, moderate, and severe. The proposed research has been applied to advanced algorithms to predict the locations having most patients affected by the COVID-19. These techniques can predict the patients until they reached the severe stage; this research classifies the COVID-19 stages accurately.

Fig. 4 COVID-19 stages classification

Table 2 COVID-19 testing from rRT-PCR dataset for feature extraction

Feature data type	Data type
Gender categorical	Categorical
Age numerical (discrete)	Numerical (discrete)
Leukocytes (WBC) numerical (continuous)	Numerical (continuous)
C-reactive protein (CRP) numerical (continuous)	Numerical (continuous)
Platelets numerical (continuous)	Numerical (continuous)
Transaminases (ALT) numerical (continuous)	Numerical (continuous)
Transaminases (AST) numerical (continuous)	Numerical (continuous)
Gamma-glutamyltransferase (GGT) numerical (continuous)	Numerical (continuous)
Lactate dehydrogenase (LDH) numerical (continuous)	Numerical (continuous)
Monocytes numerical (continuous)	Numerical (continuous)
Lymphocytes numerical (continuous)	Numerical (continuous)
Neutrophils numerical (continuous)	Numerical (continuous)
Basophils numerical (continuous)	Numerical (continuous)
Eosinophils numerical (continuous)	Numerical (continuous)
Swab categorical	Categorical

4 Results and discussion

In this section, the evaluation of the proposed method is enhanced with the feature extraction dataset of COVID-19. The proposed system is compared with the existing system in terms of sensitivity, specificity, accuracy, corona classification accuracy, time complexity, and prediction methods processed as shown in Table 2.

4.1 Sensitivity, specificity, and accuracy

Here, the evaluation of the proposed enhanced machine learning and text data mining method has been compared with the existing techniques, and the presented TF-IDF techniques are used to classify the stages of COVID-19 by similarity matching and are compared with the current classification of SVM and AI classifier in terms of sensitivity, specificity, and accuracy of the COVID-19 stages of infected patients, and they have been calculated by the following equations:

The statistical measures that can be considered are sensitivity, specificity, and accuracy

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} * 100 \quad (5)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} * 100 \quad (6)$$

Table 3 Performance analysis of the proposed and existing machine learning algorithms

Parameters	TF-IDF (%)	SVM (%)	Artificial intelligence (%)
Sensitivity	93	73	81.5
Specificity	90	78.8	74.9
Accuracy	98.4	63.4	74

Fig. 5 Comparison of statistical parameters

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} * 100. \quad (7)$$

A true positive and true negative accurate classification of corona stages is labeled by the proposed classifier techniques. The true positive indicates a proper classification of corona stages; if this label has an inappropriate classifier, then it indicates the false positive of the records, where

TP specifies the true positive,

FP denotes the false positive,

TN indicates the true negative,

FN represents the false negative.

The proposed TF-IDF method is used to classify the stages of the coronavirus accurately, which has been shown in the experimental result of Table 3, and the chart shown in Fig. 5 demonstrated the comparison.

The comparison tables for the existing ML algorithms with our developed techniques are illustrated in Table 3. From the comparison table, the proposed method has provided a 93% sensitivity level, 90% specificity level, an accuracy level of 98.4% compared with the existing techniques such as SVM and AI classifier (Fig. 5).

Similarly, the classification accuracy of the given test dataset is represented by the overall percentage of test data records that are correctly classified by the classifier techniques. The specificity and sensitivity are substitutes to the measure of accuracy that are used to evaluate the classifier's performance.

Fig. 6 The accuracy of the training model

Fig. 7 The loss of the training model

4.2 Accurate classification of COVID-19 Stages

The prediction accuracy of the proposed and existing methods can be analyzed through how the stages classify corona as mild, moderate, or severe through text classification from the dataset machine learning techniques (Figs. 6, 7).

As shown in Fig. 7, with the progress in training, the accuracy rate has been high during the comparison of previous verifications. The loss value was unable

Fig. 8 Classification of COVID-19 stages

to predict throughout the entire training process because only the change in the weight value of two models has occurred dynamically. After training, the model has achieved 92.74% classification accuracy of the COVID-19 stage on the test set.

The efficiency of each method is evaluated using the accuracy level of the analyzing process. The accurate stages classification of the COVID-19 has been demonstrated by comparing the proposed and existing methods, as shown in Fig. 8. This shows that the proposed method has given high accuracy for COVID-19 stages classification when compared with the existing methods such as SVM, KNN, and Corona Kit. Thus, the existing algorithm compared with the proposed method has provided good performance with a minimum time of complexity.

5 Conclusion

The COVID-19 first case was found in the Wuhan region, which is located in China. COVID-19 is a widespread disease and threatens the worldwide health system and economy. COVID-19 virus behaves correspondingly to other epidemic viruses. This makes it problematic to identify COVID-19 cases quickly. Therefore, COVID-19 is an important for a global epidemic, and it has confused the worldwide healthcare sector due to the non-availability of drugs or vaccines. Various researchers are working to conquer this deadly virus. The test of nasopharyngeal and an oropharyngeal swab of rRT-PCR testing is taken, and all positive case data are maintained as a record of a dataset. The machine learning techniques are used to classify the patients, who are tested positive for corona, into three different classes of mild, moderate, and severe, from the clinical report of dataset. The TF-IDF technique is used to classify the stages by similarity matching of query searching from the features presented in the test cases report. The probability has been analyzed from the feature set to detect the stages of COVID-19-infected patients. The experimental results show the high accuracy for classifying the stages of COVID-19 with a minimum number of times and good results.

References

1. Jamshidi M, Lalbakhsh A, Talla J, Peroutka Z et al (2020) Artificial intelligence and covid-19: deep learning approaches for diagnosis and treatment. *IEEE Access* 8:109581–109595. <https://doi.org/10.1109/Access.2020.3001973>
2. Wang N, Liu H and Xu C (2020) Deep Learning for the Detection of COVID-19 Using Transfer Learning and Model Integration. In: 10th International conference on electronics information and emergency communication (ICEIEC), p 281–284. IEEE
3. Li Y, Wei D, Chen J, Cao S, Zhou H, Zhu Y, Wu J, Lan L, Sun W, Qian T, Ma K (2020) Efficient and effective training of COVID-19 classification networks with self-supervised dual-track learning to rank. *IEEE J Biomed Health Inf* 24(10):2787–2797
4. Chamola V, Hassija V, Gupta V, Guizani M (2020) A comprehensive review of the COVID-19 pandemic and the role of IoT, drones, AI, blockchain, and 5G in managing its impact. *IEEE Access* 8:90225–90265. <https://doi.org/10.1109/ACCESS.2020.2992341>
5. Zhang J, Gharizadeh B, Lu D, Yue J, Yu M, Liu Y, Zhou M (2020) Navigating the pandemic response life cycle: molecular diagnostics and immunoassays in the context of COVID-19 management. *IEEE Rev Biomed Eng*. <https://doi.org/10.1109/RBME.2020.2991444>
6. Shen D, Wu G, Suk H-I (2017) Deep learning in medical image analysis. *Annu Rev Biomed Eng* 19:221–248
7. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian A, Van Der Laak JA, Van Ginneken B, Sánchez CI (2017) A survey on deep learning in medical image analysis. *Med Imag Anal* 42:60–88
8. Parthasarathy P, Vivekanandan S (2020) Internet of things (IoT) in healthcare-smart health and surveillance, architectures, security analysis and data transfer: a review. *Int J Softw Innov* 7(2):21–40
9. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, Bai J, Lu Y, Fan Z, Song Q, Cao K (2020) Artificial intelligence distinguishes COVID-19 from community-acquired pneumonia on chest CT. *Radiology*. <https://doi.org/10.1148/2Fradiol.2020200905>
10. Vijayarajeswari R, Parthasarathy P, Vivekanandan S, Basha AA (2019) Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform. *Measurement* 146:800–805
11. Lin TY, Goyal P, Girshick R, He K, Dollar P (2017) Focal Loss for Dense Object Detection. In: International conference on computer vision, p. 2980–2988. IEEE
12. Basha AA, Vivekanandan S, Parthasarathy P (2019) Blood glucose regulation for post-operative patients with diabetics and hypertension continuum: a cascade control-based approach. *J Med Syst* 43(4):95
13. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C, Shpanskaya K, Lungren MP (2017) Chestnet: radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint arXiv:1711.05225*
14. Rajaraman S, Candelier S, Kim I, Thoma G, Antani S (2018) Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs. *Appl Sci* 8(10):1715
15. Parthasarathy P, Vivekanandan S (2018) Urate crystal deposition, prevention and various diagnosis techniques of GOUT arthritis disease: a comprehensive review. *Health Info Sci Syst* 6(1):19
16. Stephens O, Sain M, Maduh UJ, Jeong DU (2019) An efficient deep learning approach to pneumonia classification in healthcare. *J Healthc Eng* 2019:1–7. <https://doi.org/10.1155/2019/4180949>
17. Zheng X, Kulhare S, Mehanian C, Chen Z, Wilson B (2018) Feature detection and pneumonia diagnosis based on clinical lung ultrasound imagery using deep learning. *J Acoust Soc Am* 144(3):1668–1668
18. Parthasarathy P, Vivekanandan S (2018) Investigation on uric acid biosensor model for enzyme layer thickness for the application of arthritis disease diagnosis. *Health Inf Sci Syst* 6(1):5
19. Xu X, Jiang X, Ma C, Du P, Li X, Lv S, Yu L, Ni Q, Chen Y, Su J, Lang G (2020) Deep learning system to screen novel coronavirus disease 2019 pneumonia. *arXiv preprint arXiv:2002.09334*
20. Gozes O, Frid-Adar M, Greenspan H, Browning PD, Zhang H, Ji W, Bernheim A, Siegel E (2020) Rapid AI development cycle for the coronavirus (COVID-19) pandemic: Initial results for automated detection & patient monitoring using deep learning CT image analysis. *arXiv preprint arXiv:2003.05037*

Affiliations

Shalini Ramanathan¹ · Mohan Ramasundaram¹

✉ Shalini Ramanathan
406916001@nitt.edu

Mohan Ramasundaram
rmohan@nitt.edu

¹ Department of Computer Science and Engineering, National Institute of Technology,
Tiruchirappalli, Tamil Nadu, India

RETRACTED ARTICLE