SN Computer Science (2023) 4:218
https://doi.org/10.1007/s42979-023-01703-6

ORIGINAL RESEARCH

SARS-CoV-2 Diagnosis Using Transcriptome Data: A Machine Learning Approach

Pratheeba Jeyananthan[1]

Received: 19 April 2022 / Accepted: 24 January 2023

Abstract

SARS-CoV-2 pandemic is the big issue of the whole world right now. The health community is struggling to rescue the public and countries from this spread, which revives time to time with different waves. Even the vaccination seems to be not prevents this spread. Accurate identification of infected people on time is essential these days to control the spread. So far, Polymerase chain reaction (PCR) and rapid antigen tests are widely used in this identification, accepting their own drawbacks. False negative cases are the menaces in this scenario. To avoid these problems, this study uses machine learning techniques to build a classification model with higher accuracy to filter the COVID-19 cases from the non-COVID individuals. Transcriptome data of the SARS-CoV-2 patients along with the control are used in this stratification using three different feature selection algorithms and seven classification models. Differently expressed genes also studied between these two groups of people and used in this classification. Results shows that mutual information (or DEGs) along with naïve Bayes (or SVM) gives the best accuracy (0.98 ± 0.04) among these methods.

Keywords COVID-19 diagnosis · Feature selection · Transcriptome data · Machine learning models · Differently expressed genes · GO analysis

Introduction

SARS-CoV-2 is the current health issue of the globe. Even though considerable people get vaccinated, it is not under the control and evolving with new variants. As it is a contagious disease, it is crucial to identify the infected patients as soon as possible. Delays in the identification of the patients or misinterpretation of the results will even worse the situation. Polymerase chain reaction (PCR) test is widely used in the identification of the infected people. Rapid antigen tests also used in this context with very low accuracy. Literature shows that these methods have considerable drawbacks and their reliability depends on so many factors [1, 2].

There are studies in the literature which use machine learning algorithms in the identification of SARS-CoV-2 patients with significant accuracy. These studies vary in many directions including classification between COVID-19 positive and negative cases [3–5], separating COVID-19 cough from normal cough [6], COVID-19 detection, prognosis and diagnosis [7–12], whether a person is having the risk of COVID-19 or not [13] and differentiating SARS-CoV-2 from other viruses [14–17]. These studies used different types of data as the input for their model and different set of machine learning algorithms were utilized in these predictions.

Type of the input data used in these models plays a crucial role in the accuracy of the prediction despite the machine learning algorithms. In the literature, varieties of data were used such as genome sequences [15], transcriptome data [16], recorded voices [6], symptoms, clinical and morphological features of the patients [3–5, 7–9, 12, 13, 17] and X-ray images [10].

Studying the literature shows that molecular data are rarely used in these machine learning related diagnosis of SARS-CoV-2. However, transcriptome data are the most widely used data in the investigation of diseases in molecular

This article is part of the topical collection "Computer Aided Methods to Combat COVID-19 Pandemic" guest edited by David Clifton, Matthew Brown, Yuan-Ting Zhang and Tapabrata Chakraborty.

✉ Pratheeba Jeyananthan
pratheeba@eng.jfn.ac.lk

[1] Faculty of Engineering, University of Jaffna, Jaffna, Sri Lanka

Published online: 17 February 2023