Which Features Are Useful in Machine Learning-Based COVID-19 Prognostication? A Meta-Analysis

Minhaz Mahmud
Department of Electrical and Electronic Engineering
Bangladesh Army International University of Science & Technology
Cumilla, Bangladesh
ORCID: 0009-0005-4699-1940

Md Rakibul Hasan
Optus Centre for Artificial Intelligence
Curtin University
Perth, Australia
ORCID: 0000-0003-2565-5321

Reabal Najjar
School of Computing
Australian National University
Canberra, Australia
Canberra Health Services
Canberra Hospital
Canberra, Australia
ORCID: 0000-0001-7169-7077

Md Zakir Hossain
Optus Centre for Artificial Intelligence
Curtin University
Perth, Australia
School of Computing
Australian National University
Canberra, Australia
ORCID: 0000-0003-1892-831X

***Abstract*—COVID-19 brings devastating impacts on society and shows how fragile our current healthcare system is. To aid the healthcare system, machine learning (ML) models have shown strong potential in combating COVID-19 across several areas, including diagnosis, complication prediction, mortality risk prediction, and severity assessment. Feature selection is an essential step in developing efficient machine-learning models. In this research, we screened 143 COVID-19 research articles, where 26 articles had enough data to calculate the effect size. We used Cohen's d to calculate the effect size of the features, which reveals several highly significant features yet explored in only a few studies and several less significant features yet explored in many studies. Lymphocyte (%) and hypersensitive c-reactive protein are comparatively less used features but showed a very large effect size. On the other hand, lymphocyte count, c-reactive protein, D-dimer, and creatinine are utilized in several studies but are not necessarily significant. Considering these high-effect-sized features and discarding the low-effect-sized features would benefit future ML model developments for combating COVID-19.**

***Index Terms*—Effect size, Cohen's d, COVID-19, Prognostication, Meta-analysis**

I. Introduction

In December 2019, the novel coronavirus disease (COVID-19) was discovered in Wuhan, China [1]. Since then, it has quickly spread throughout the entire globe. The World Health Organisation (WHO) classified this outbreak as a pandemic in January 2020 [2]. The extremely contagious viral illness, COVID-19, has so far spread rapidly across the world and has raised major issues about global health. Globally, 6.89 million deaths and 761.4 million confirmed cases of COVID-19 have been reported as of March 31, 2023, [3]. The fast spread of COVID-19 has left front-line healthcare professionals exhausted and in serious need of medical supplies.

Many COVID-19 patients experience fast symptom escalation following a period of relatively moderate symptoms, underscoring the need for sophisticated risk classification models. Predictive modeling can identify individuals who have a higher mortality risk and help reduce deaths as quickly as possible. Therefore, it is essential to accurately estimate the disease prognosis and triage critically sick patients in order to lessen the load on the healthcare system and offer the best care for patients. Additionally, physicians and health policymakers frequently employed and relied on predictions provided by various computational and statistical models because of the significant hesitancy surrounding its concluding influence.

Feature selection is a crucial part of developing any predictive model as it heavily affects the performance of such models. It refers to selecting a subset of relevant features to train predictive models, such as machine learning (ML). By removing unnecessary data, feature selection provides a straightforward yet efficient solution to the problem of finding appropriate features for a particular real-life application [4]. Removal of irrelevant data helps increase accuracy and decrease computation time, with a better comprehension of the learning model or data. Mostly, not all the variables in the dataset are relevant when creating a real-world ML model. The addition of redundant variables reduces the model's capacity for generalization and may also reduce a classifier's overall precision. A model also becomes more complex if additional unnecessary variables are added to it.

Effect size is one of the feature selection methods. While

2575-2634/23/$31.00 ©2023 IEEE

statistical significance demonstrates the existence of an effect in a study, practical significance demonstrates that the effect is significant enough to have real-world implications [5]. P values are used to indicate statistical significance, whereas statistical effect sizes are used to indicate practical significance. Statistical significance is impacted by the sample size, and so it alone might be deceptive. No matter how small the effect is in reality, increasing the sample size always increases the chance of finding a statistically significant effect. Contrarily, effect sizes are unaffected by the number of samples [6]. Sullivan and Feinn [6] argued that the p value is not enough and emphasized reporting the effect size in the abstract and results of papers. Through a systematic literature review, Ge *et al.* [7] used effect size to evaluate features used in diagnosing Parkinson's disease.

Researchers have reported many studies on COVID-19 on several targets, for example, diagnosis [8], [9], complication prediction [10], mortality risk prediction [11], and severity assessment [12]. A literature review article helps us know about the summary trends of a certain research domain. Several literature review articles are published for COVID-19 domains; for example, Lalmuanawma *et al.* [13] reviewed ML models, data types, performance, etc. reported by various COVID-19 studies. Abd-Alrazaq *et al.* [14] reported the most prominent research hotspots, publication frequency, origin, article types, and top authors. They mentioned 733 systematic literature review articles on COVID-19.

Most literature reviews lack further analysis to find out which features or models are actually better, given that different authors had different choices, using which they reported different performances. Yaacob *et al.* [15] used effect size on the performance of ML models in predicting COVID-19 confirmed cases. However, ML has been used not only in predicting confirmed cases but also in other aspects, such as severity assessment and mortality risk prediction. In this article, we, therefore, present the calculated effect size of features used in ML-based COVID-19 studies in four aspects.

II. Method

A. Data Curation

Based on our primary target, we set the search keywords as *covid*, artificial intelligence, machine learning, and prognos** (* refers to any keywords that start with the preceding word fragment). We searched five well-known databases on 19 November 2021: Google Scholar, Web of Science, PubMed, Scopus, and MEDLINE. The search revealed 2,403 articles, where 737 articles were removed being duplicates. Further, 217 articles were discarded because of irrelevant titles that do not match our scope. In the remaining 1,449 articles, we applied our inclusion/exclusion criteria.

We included an article if it: (1) specifically addressed COVID-19, (2) applied artificial intelligence or ML algorithm on COVID-19 datasets, (3) was peer-reviewed, (4) was published in English, (5) was published after 2017, and (6) was not a case-study or review article. The above criteria made 1,306 articles excluded. We inspected the remaining 143 articles to calculate the effect size.

Among the 143 articles, only 26 articles have mentioned required data on features, using which we calculate the effect size. The selected articles are broken down into four categories: diagnosis, complication prediction, mortality risk prediction, and severity assessment (Table I). Among the prognostication categories, diagnosis refers to COVID-19 diagnosis using chest x-ray images [8], omics data [9], etc. Studies were categorized as 'complication prediction' if they used ML in predicting the course or complications of COVID in the future (i.e. if a patient ends up in ICU or develops long-term respiratory symptoms) [10]. This is in contrast to 'severity assessment', which is based on stratifying how bad the disease currently is (present) [12].

TABLE I
Distribution of Articles into Four Categories

Category	Screened articles (%)	Articles having required data (%)
Diagnosis	23 (16.1%)	3 (11.5%)
Complication prediction	46 (32.2%)	11 (42.3%)
Mortality risk prediction	35 (24.5%)	5 (19.2%)
Severity assessment	39 (27.3%)	7 (26.9%)

B. Effect Size

Effect size is a statistical concept that uses a quantitative scale to quantify how strongly two variables are related. If two data have comparable qualities and one of them has a higher average than the other, the difference between them is referred to as the effect size. The effect size increases when the difference gets bigger. We can determine whether a difference is real or the result of a change in factors by looking at the statistical effect size.

There are different ways to calculate the effect size, such as standardized mean difference, Cohen's d, Glass' Δ, and Hedges' g. Among these, Cohen's d is the most acclaimed [5], and so used in this study.

To calculate Cohen's d, a pooled standard deviation (s) defined by Jacob Cohen is calculated at first, shown in (1).

$$s = \frac{\sqrt{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}}{n_1 + n_2 - 2} \tag{1}$$

where, n_1 and n_2 are the sample sizes, and s_1 and s_2 are the standard deviations of the two variables.

Cohen's d is then the difference between two means divided by the pooled standard deviation shown in (2).

$$d = \frac{\overline{x_1} - \overline{x_2}}{s} \tag{2}$$

Effect size is classified into different categories based on the magnitude of d: very small (0.01), small (0.20), medium (0.50), large (0.80), very large (1.20), and huge (2.00) [16], [17].

We calculated the effect size of all features in all 26 articles separately. As we calculated it independently based on the data in respective articles, the variations of effect size might skew

the arithmetic mean, and so we considered the median, instead of the mean, to compare the effect sizes among different features.

III. Results and Discussion

A. Frequency of Features

We extracted data on 142 features from 26 COVID-19 studies. Among these, only 24 features are used in at least four studies and are listed in Table II. The most frequently used feature is the age of the participants, which appears in 84.6% of studies. The second most used feature is lymphocyte count, which appears in 50% of the studies.

TABLE II
Frequency of Features Used in 26 Studies

Feature name	Number of studies (%)
Age	22 (84.6%)
Lymphocyte count	13 (50%)
White blood cell count	12 (46.2%)
Lactate dehydrogenase	11 (42.3%)
C-reactive protein	10 (38.5%)
Neutrophil	8 (30.8%)
D-dimer	8 (30.8%)
Lymphocyte (%)	7 (26.9%)
Platelet count	7 (26.9%)
Temperature (c)	7 (26.9%)
Creatinine	6 (23.1%)
Neutrophils (%)	5 (19.2%)
Oxygen saturation	5 (19.2%)
Body mass index	5 (19.2%)
Creatine kinase	5 (19.2%)
Aspartate aminotransferase	4 (15.4%)
Monocyte (%)	4 (15.4%)
Respiratory rate	4 (15.4%)
Ground glass opacity	4 (15.4%)
Diastolic blood pressure	4 (15.4%)
Systolic blood pressure	4 (15.4%)
Eosinophils (%)	4 (15.4%)
Hemoglobin	4 (15.4%)
Consolidation	4 (15.4%)

B. Top Features in Diagnosis

Fig. 1 displays huge and very large effect-sized features in COVID-19 diagnosis studies. Age and lymphocyte (%) have the highest effect size of 1.33, followed by lymphocyte count and neutrophils (%).

C. Top Features in Complication Prediction

Fig. 2 presents huge and very large effect-sized features in the complication prediction category. Oxygen saturation, red blood cell, and hematocrit have huge effect sizes, even though only one research in the complication prediction category employed these features. Lymphocyte (%) has a very large effect size, and five studies in this area exploited this feature.

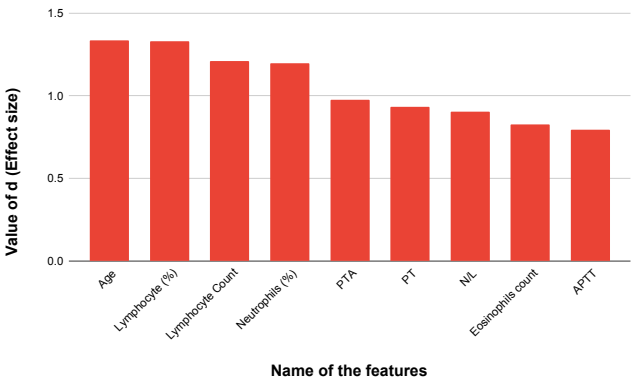

Fig. 1. Huge and very large effect-sized features in COVID-19 diagnosis articles. Acronyms: PTA – Prothrombin time activity; PT – Prothrombin time; N/L – Neutrophil to lymphocyte ratio; APTT – Activation of partial thromboplastin time.

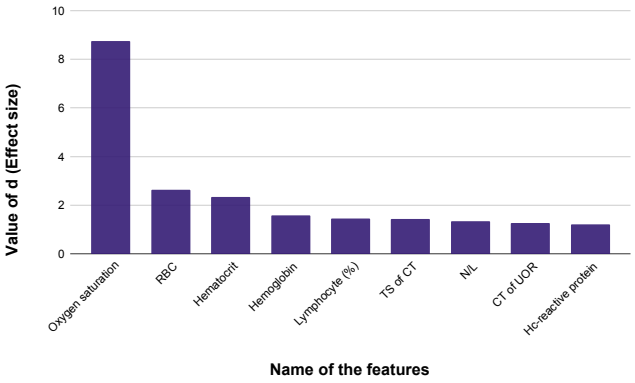

Fig. 2. Huge and very large effect-sized features in COVID-19 complication prediction articles. Acronyms: RBC – Red blood cell; TS of CT – Total score of computed tomography; N/L – Neutrophil to lymphocyte ratio; CT of UOR – Computed tomography scan of Upper lobe of the right lung; Hc-reactive protein – Hypersensitive c-reactive protein.

D. Top Features in Mortality Risk Prediction

Huge and very large effect-sized features in mortality studies are presented in Fig. 3. Although lymphocyte (%) is used in one paper, it has the largest effect size in this category. However, age also has a large effect size and was used in six articles.

E. Top Features in Severity Assessment

Fig. 4 presents huge and very large effect-sized features in severity assessment studies. The highest effect-sized feature, with around 1.46 and used in two research, is temperature. Apart from temperature and age, other features have only been observed in one severity assessment study.

F. Overall Top Features

The top 10 features out of 142 features in terms of effect size are shown in Fig. 5. Although they lie in the very large effect size category, the majority of them are found in just one study.

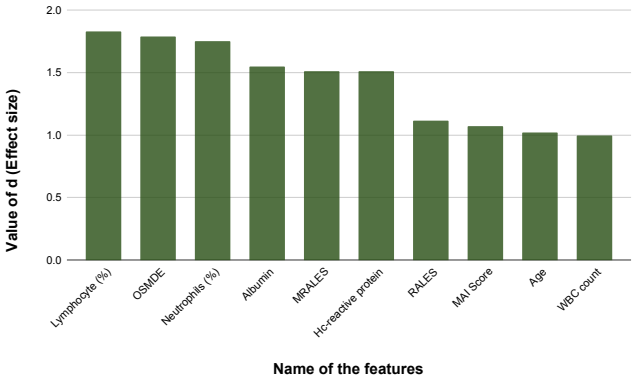

Fig. 3. Huge and very large effect-sized features in COVID-19 mortality risk prediction articles. Acronyms: OSMDE – Oxygen saturation minimum during the encounter; MRALES – Maximum radiographic assessment of lung edema score; Hc-reactive protein – Hypersensitive c-reactive protein; RALES – Radiographic assessment of lung edema score change; MAI Score – Maximum artificial intelligence score; WBC count – White blood cell count.

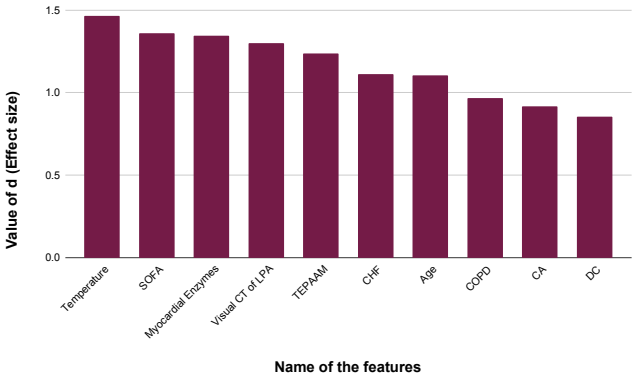

Fig. 4. Huge and very large effect-sized features in COVID-19 severity assessment articles. Acronyms: SOFA – Sequential organ failure assessment; Visual CT of LPA – Visual computed tomography score of lung parenchymal abnormalities; TEPAAM – Total extent of parenchymal abnormalities by automatic measurement; CHF – Congestive heart failure; COPD – Chronic obstructive pulmonary disease; CA – Cardiac arrhythmias; DC – Diabetes with complications.

For example, 'Oxygen saturation minimum during encounter' has the highest effect size among all the 142 features, but it is used in only one study. On the other hand, age was not even among the top 10 features, but it is the most used feature. Lymphocyte (%) is used in seven studies with an effect size $= 1.45$.

Several features – for example, lymphocyte count, c-reactive protein, D-dimer, and creatinine – are used in many studies but are not effective based on their effect size. Yet again, features that have a large effect size but have only appeared in one or two studies are less reliable since they could be biased. We, therefore, present the features that are used in three or more separate studies (reliable) and have a high effect size (effective) in Table III.

In the data preprocessing stage, before developing ML

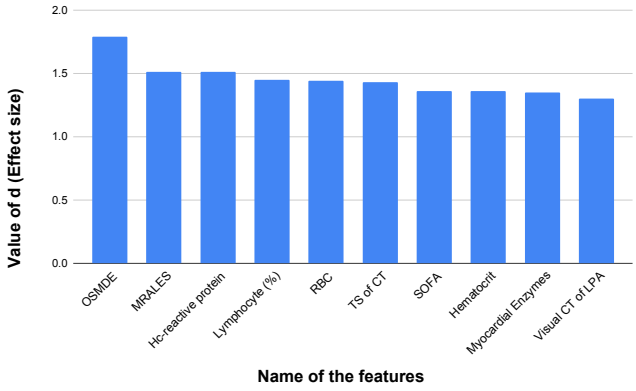

Fig. 5. Top 10 features in terms of effect size in all COVID-19 articles. Acronyms: OSMDE – Oxygen saturation minimum during encounter; MRALES – Maximum radiographic assessment of lung edema score; Hc-reactive protein – Hypersensitive c-reactive protein; RBC – Red blood cell; TS of CT – Total score of computed tomography; SOFA – Sequential organ failure assessment; Visual CT of LPA – Visual computed tomography score of lung parenchymal abnormalities.

TABLE III
RELIABLE AND EFFECTIVE FEATURES

Feature Name	Effect size	Number of studies
Age	1.018	22
Lymphocyte (%)	1.447	7
Neutrophils (%)	1.197	5
Aspartate aminotransferase	0.864	4
Monocyte (%)	0.844	4
Hc-reactive protein	1.510	3
Arterial oxygen partial pressure (PaO2 in mmHg) to fractional inspired oxygen ratio	1.149	3
Neutrophil to lymphocyte ratio	1.060	3

models, future research should consider selecting these reliable and statistically effective features. It would enhance the performance of ML models in terms of accuracy and speed.

IV. CONCLUSION

The traditional way of combating COVID-19 is resource intensive. In order to reduce medical resources and efficiently use them, ML models can be a prospective solution. While developing a model to combat COVID-19, appropriate feature selection is important because there are many features to choose from. We systematically curated 143 articles and calculated the effect size (Cohen's d) of all reported features. Analysis revealed several effective features but seldomly used (e.g., lymphocyte (%) and hypersensitive c-reactive protein), and several less effective features but highly used (e.g., lymphocyte count, c-reactive protein, D-dimer, and creatinine). In addition to Cohen's d, we will include AI techniques in the future, for example, multilayer perceptron and convolutional neural network models for variable selection, feature importance, and significance analysis [18]. They provide useful information to identify features or variables that have the most impact on predicting a particular outcome. Additionally, we plan to expand our meta-analysis to include more recent studies and a larger number of articles to increase the robustness of our

findings. The reliable and effective features reported in this article would help future research to select features, which would improve the model performance, helping humanity combat COVID-19.

REFERENCES

[1] H. Kazemi-Arpanahi, K. Moulaei, and M. Shanbehzadeh, "Design and development of a web-based registry for coronavirus (covid-19) disease," *Medical journal of the Islamic Republic of Iran*, vol. 34, p. 68, 2020. DOI: 10.34171/mjiri.34.68.

[2] H. Shi, X. Han, N. Jiang, *et al.*, "Radiological findings from 81 patients with covid-19 pneumonia in wuhan, china: A descriptive study," *The Lancet infectious diseases*, vol. 20, no. 4, pp. 425–434, 2020. DOI: 10.1016/S1473-3099(20)30086-4.

[3] E. Mathieu, H. Ritchie, L. Rodés-Guirao, *et al.*, "Coronavirus pandemic (covid-19)," *Our World in Data*, 2020, https://ourworldindata.org/coronavirus.

[4] M. Z. Hossain, M. M. Kabir, and M. Shahjahan, "A robust feature selection system with colin's cca network," *Neurocomputing*, vol. 173, pp. 855–863, 2016.

[5] L. D. Kyu, "Alternatives to p value: Confidence interval and effect size," *kja*, vol. 69, no. 6, pp. 555–562, 2016. DOI: 10.4097/kjae.2016.69.6.555.

[6] G. M. Sullivan and R. Feinn, "Using effect size—or why the p value is not enough," *Journal of graduate medical education*, vol. 4, no. 3, pp. 279–282, 2012. DOI: 10.4300/JGME-D-12-00156.1.

[7] W. Ge, C. J. Lueck, D. Apthorp, and H. Suominen, "Which features of postural sway are effective in distinguishing parkinson's disease from controls? a systematic review," *Brain and Behavior*, vol. 11, no. 1, e01929, 2021. DOI: 10.1002/brb3.1929.

[8] J. Manokaran, F. Zabihollahy, A. Hamilton-Wright, and E. Ukwatta, "Detection of COVID-19 from chest x-ray images using transfer learning," *Journal of Medical Imaging*, vol. 8, no. S1, p. 017 503, 2021. DOI: 10.1117/1.JMI.8.S1.017503.

[9] X. Liu, M. R. Hasan, K. A. Ahmed, and M. Z. Hossain, "Machine learning to analyse omic-data for covid-19 diagnosis and prognosis," *BMC bioinformatics*, vol. 24, no. 1, pp. 1–20, 2023. DOI: 10.1186/s12859-022-05127-6.

[10] G. Wu, S. Zhou, Y. Wang, *et al.*, "A prediction model of outcome of sars-cov-2 pneumonia based on laboratory findings," *Scientific Reports*, vol. 10, no. 1, pp. 1–9, 2020. DOI: 10.1038/s41598-020-71114-7.

[11] A. Karthikeyan, A. Garg, P. K. Vinod, and U. D. Priyakumar, "Machine learning based clinical decision support system for early covid-19 mortality prediction," *Frontiers in Public Health*, vol. 9, 2021. DOI: 10.3389/fpubh.2021.626697.

[12] O. Kocadagli, A. Baygul, N. Gokmen, S. Incir, and C. Aktan, "Clinical prognosis evaluation of covid-19 patients: An interpretable hybrid machine learning approach," *Current Research in Translational Medicine*, vol. 70, no. 1, p. 103 319, 2022. DOI: 10.1016/j.retram.2021.103319.

[13] S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, "Applications of machine learning and artificial intelligence for covid-19 (sars-cov-2) pandemic: A review," *Chaos, Solitons & Fractals*, vol. 139, p. 110 059, 2020. DOI: 10.1016/j.chaos.2020.110059.

[14] A. Abd-Alrazaq, J. Schneider, B. Mifsud, *et al.*, "A comprehensive overview of the covid-19 literature: Machine learning–based bibliometric analysis," *J Med Internet Res*, vol. 23, no. 3, e23703, Mar. 2021. DOI: 10.2196/23703.

[15] W. F. W. Yaacob, N. M. Sobri, S. A. M. Nasir, N. I. Nordin, W. F. W. Yaacob, and U. Mukhaiyar, "Machine learning models for COVID-19 confirmed cases prediction: A meta-analysis approach," *Journal of Physics: Conference Series*, vol. 2084, no. 1, p. 012 013, Nov. 2021. DOI: 10.1088/1742-6596/2084/1/012013.

[16] J. Cohen, *Statistical power analysis for the behavioral sciences*. Routledge, 1988. DOI: 10.4324/9780203771587.

[17] S. S. Sawilowsky, "New effect size rules of thumb," *Journal of modern applied statistical methods*, vol. 8, no. 2, p. 26, 2009. DOI: 10.22237/jmasm/1257035100.

[18] Y. Zhang, Y. Yao, Z. Hossain, S. Rahman, and T. Gedeon, "Eeg feature significance analysis," in *Neural Information Processing*, T. Mantoro, M. Lee, M. A. Ayu, K. W. Wong, and A. N. Hidayanto, Eds., Cham: Springer International Publishing, 2021, pp. 212–220, ISBN: 978-3-030-92310-5.