ANNALES ACADEMIAE MEDICAE SILESIENSIS

ACCEPTED MANUSCRIPT

Ann. Acad. Med. Siles. (Online) 2026; DOI: 10.18794/aams/222516

Review

Applications of machine learning methods in long COVID phenotyping: Review

Martyna Szlenk[1], Karol Żmudka[1], Aleksandra Spychał[1], Rafał Kołodziej[2], Patrycja Piłat[1], Jerzy Jaroszewicz[1]

[1]Department of Infectious Diseases and Hepatology, Faculty of Medical Sciences in Zabrze, Medical University of Silesia, Katowice, Poland
[2]Students' Scientific Association, Department of Immunology, Faculty of Medical Science, University of Rzeszów, Poland

Address for correspondence:
Martyna Szlenk
Katedra i Klinika Chorób Zakaźnych i Hepatologii
Górnośląskie Centrum Medyczne im. prof. Leszka Gieca ŚUM
ul. Ziołowa 45/47, 40-635 Katowice
e-mail: s81920@365.sum.edu.pl

Received: 22.02.2026, Revised: 27.04.2026, Accepted: 26.05.2026, Published: June 2026

Publisher: Medical University of Silesia, Katowice, Poland

ABSTRACT

Introduction: Long COVID (LC) is one of the major challenges of modern medicine. The lack of standardized diagnostic criteria, nonspecific symptoms, their temporal variability and chronic nature significantly hinder diagnosis and classification. This creates the need for advanced technological tools to better understand the pathophysiology and identify individual disease phenotypes. Artificial intelligence, particularly machine learning (ML), is a promising approach. This review synthesises current evidence on ML applications in LC phenotyping and outlines core ML concepts relevant to this research context.

Material and methods: A search of the PubMed database identified 47 articles published through July 9, 2024. After manual screening of titles and abstracts, 17 studies published in English were included in the final analysis.

Results: We present an overview of ML applications in LC research, focusing on the identification of phenotypes, subphenotypes and risk factors associated with LC occurrence and severity. The identified phenotypes depict LC as a multisystem condition, in which symptom clusters often involve multiple organ systems rather than a single organ.

Conclusions: ML represents a useful tool for identifying LC patients and classifying them into distinct subphenotypes characterized by different clinical manifestations. This approach improves the precision of diagnostic pathways and supports individualized therapeutic strategies. ML approaches may also improve clinical trial design by facilitating analysis of large, heterogeneous datasets – a particular advantage given LC's variable clinical course.

KEYWORDS

post-acute COVID-19 syndrome, COVID-19, SARS-CoV-2, post-COVID conditions, machine learning, long COVID phenotypes

INTRODUCTION

After the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic, the long-term effects of the disease in convalescents became a major health concern. Long COVID (LC), also referred to as post-acute sequelae of COVID-19 (PASC), is a principal challenge for global healthcare nowadays. Its diagnosis and characterization are complicated by factors such as a wide range of symptoms, varying onset times and underreporting. During and after the COVID-19 pandemic machine learning (ML) methods were used among other applications in the development of diagnostic models, treatment strategies, epidemiological predictions or LC identification. ML methods have advanced predictive modelling, supported patient stratification, and helped identify complex interaction patterns among LC sequelae [1]. Extensive research has

documented a wide spectrum of symptoms associated with the sequelae of coronavirus disease 2019 (COVID-19), including gene expression, proteomics, blood tests, comorbidities, radiological signs, and more. ML methods are thus well-suited to analysing such heterogeneous, high-dimensional data.

The scope of LC phenotyping can be divided into identifying patients at higher risk of developing LC symptoms and subphenotyping those at risk for different sets of symptoms, which may indicate various courses of the condition. Lastly, some studies aim to investigate and characterize the various physiological and pathological changes associated with LC. Moreover, LC can have different trajectories in distinct subpopulations. Most analyzed studies focused on adults; however, some included pediatric populations, while others aimed to describe LC using animal models.

However, ML methods in phenotyping present numerous challenges. First, ML developed phenotyping results could be difficult to reproduce due to data sharing restrictions or data involving large number of diverse features. Furthermore, training data may not be extendable to different populations, meaning that phenotyping developed with data from earlier waves of COVID-19 could be poorly applicable in the future populations.

Despite the fact that the presented publications often cover multiple categories of topics, we have divided them into subsections based on the main aspects they address. Below, we cover different ways of using ML in the phenotyping of LC.

MATERIAL AND METHODS

Search methods

We searched the PubMed database; the search was limited to a single database, which constitutes a limitation of this review. The search strategy was conducted for the general string of “Long COVID, phenotyping, machine learning,” with the exact search string: ((((((Long COVID) OR (Long Haul COVID-19)) OR (Long-Haul COVID)) OR (Post Acute COVID-19 Syndrome)) OR (Post-Acute Sequelae of SARS-CoV-2 Infection)) OR (Post-COVID Conditions)) AND ((Machine Learning) OR (Artificial Intelligence)) AND ((Phenotype) OR (Phenotyping) OR (Phenotypes)). As of July 9, 2024, we found 57 results. Finally, after manually reviewing titles and abstracts we included 17 studies in the review.

Objectives

The narrative review aims to gather current information on potential applications of ML methods in phenotyping COVID-19 patients (Figure 1).

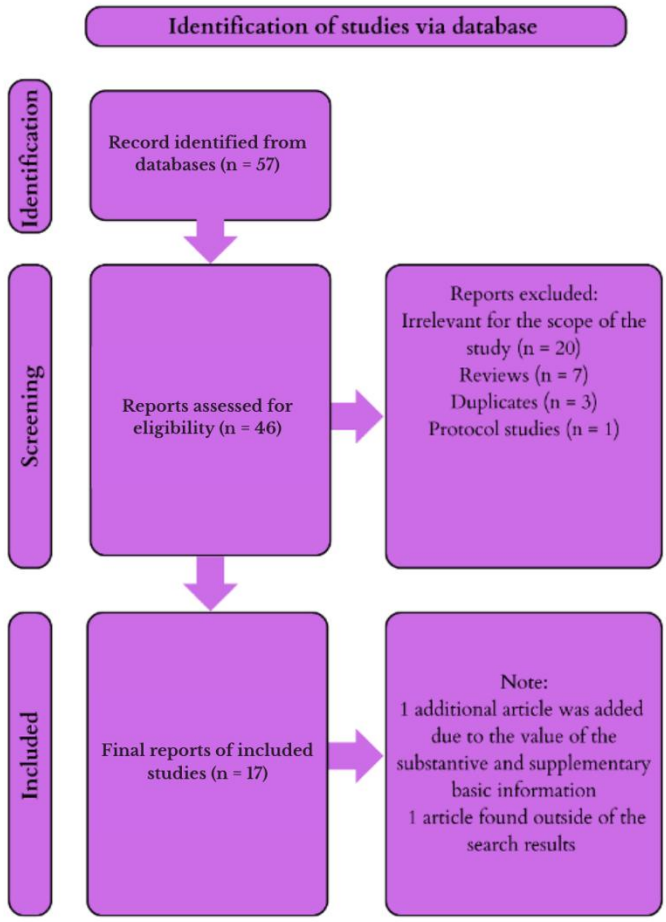

Fig. 1. Diagram of the screening process and article selection

Selection criteria

We included observational studies, cohort studies, clinical trials, and other study designs employing ML methods for LC phenotyping, including studies on pediatric populations and animal models. The considered studies were published in English. We excluded studies that did not clearly demonstrate the application of ML methods to COVID-19 or post-COVID phenotyping. Several included sources are preprints (medRxiv, bioRxiv), which have not undergone formal peer review; this should be considered a limitation when interpreting the corresponding findings.

INTRODUCTION TO MACHINE LEARNING MECHANISMS

The following section provides a concise overview of the key ML concepts relevant to the studies included in this review. Among the methods described, unsupervised learning — particularly clustering algorithms — and supervised learning were most widely applied in the identified LC phenotyping literature. Deep learning (DL) methods were employed only in specific analytical contexts.

Structure of artificial intelligence and learning

Artificial intelligence (AI) is a hard-to-define umbrella term for the use of, among other things: computational algorithms, to achieve computerized intelligence that mimics human intelligence. One of its subordinate functions gaining prominence is ML, which enables a machine to adapt to new data without introducing additional software [2,3]. The main idea of ML is to design algorithms and statistical models, allowing processing and integration of huge amounts of information. Based on the pre-entered data, the system predicts the results and categorizes the observations pertaining to them, while modifying and improving its proceedings based on the new data acquired, with which it calibrates its own statistical models [3,4]. The performance of ML processes depends on a wide range of statistical models, among others: simple decision trees, reinforcement learning, classification or regression analysis, as well as layered neural networks, mainly used in DL of much higher complexity [5,6]. The selection of the appropriate algorithm depends on the structure of the input data. In the context of LC phenotyping research, the most relevant paradigms are supervised learning and unsupervised learning, the latter being predominant in phenotype discovery tasks. Two further paradigms — semi-supervised learning and reinforcement learning — exist but were not employed in the studies reviewed here [4]. Supervised learning is based on training the model to predict specific target variables, through the use of a "training set" consisting of input data, for which specific output data are assigned, which are the results or values of the so-called: "labels." The nature of the output variable determines the task of the algorithm: classification or regression. Classification is based on the analysis of data with the assignment of an example to a particular category (e.g., a patient to a certain subtype of cancer based on medical data). The regression task, on the other hand, aims to predict a specific outcome value (e.g., biomarker values in blood) [4,5,7,8].

Unsupervised learning is distinguished by identifying patterns in unlabeled data sets devoid of "labels," yielding better results in data analysis, stratification or reduction than prediction. In simpler terms, the model focuses on dimensionality reduction and clustering, i.e. finding naturally occurring patterns, by identifying underlying relationships and features by which it selects data into specific groups based on their similarity. For example, the model can find previously unseen connections between different individuals, when grouping them based on genes, environment and medical history [2,7,9].

Steps in developing a machine learning model

Data engineering

We can divide the data used in the creation of ML models into structured data, unstructured data, semi-structured data and metadata, depending on their defined format or organization, and their

source can be, among others: electronic medical records, or medical images acquired from imaging studies [2,5]. The obtained data then undergoes processes of cleaning, transformation and reduction, leading successively to increasing data consistency and removing errors, transforming it to the format required for a specific ML model, and reducing its volume, allowing faster and easier analysis [10,11]. The next step is data extraction (feature extraction), which is a process that identifies and generates the optimal and most appropriate features from the available raw data sets, eliminating redundancy and reducing dimensionality. The resulting set serves as the input necessary for the creation of a specific ML algorithm. The range of extraction methods includes principal component analysis (PCA), linear discriminant analysis (LDA) and independent component analysis (ICA), while depending on the type of data being analyzed. For text data, more complex methods such as Word2Vector are used [7,12,13].

Model training

The data obtained in the processes described above are used to create a master training set, which is a portion of the database that is selected at random. This set is designated as the master set, serving as the first in the iteration cycle. The set is then transformed, balanced and validated allowing for the highest quality of analyzed data in subsequent iteration loop cycles [14]. Depending on the "data labeling", the appropriate type of supervised, unsupervised, semi-supervised or reinforcement learning is introduced. In turn, the type of data, such as tabular data, image data or text data determine the choice of ML algorithm and techniques, the characteristics of which are beyond the scope of this work. Currently, we distinguish between simpler, easier to interpret classical models, operating on tabular or textual data (Table I), and neural networks, which perform better in complex data [4,5,8].

Table I. Summary of examples of the most used classical statistical algorithms in certain types of machine learning (compiled by the authors based on [4,5,7,8])

Type ML	SUPERVISED LEARNING	UNSUPERVISED LEARNING	LEARNING BY REINFORCEMENT
	SEMI-SUPERVISED LEARNING		
Used algorithms	Naïve Bayes	K-means clustering	Monte Carlo Methods
	Decision tree	Hidden Markov model	Q-learning
	Support vector machine	Gaussian mixture model	Deep Q-learning (DQN)
	Linear regression/logistic	Affinity propagation	
	K-nearest neighbors	Hierarchical clustering	

Neural networks are the most widely used ML model, usually consisting of three layers imitating a biological neural network, where each layer has nodes, connected to all the nodes of the other layers [15,16]. Their mechanism of operation is an iterative process, consisting of an error function

that evaluates the difference between the model's predictions and the actual results, a search function that determines the factors that minimize the error function, and an update function based on learning the model by changing the decision-making power of a specific neuron to the end result of the predictions [8,17]. There are many variants of neural networks among others: recurrent networks better suited for processing linguistic data, and spline networks preferred for image data analysis. In addition, the extension of the neural network by increasing the number of layers, allows the processing of much larger and more complex databases. The resulting neural networks are referred to as DL, characterized by much higher process efficiency and the ability to interpret complex data. Each neuronal layer in DL is responsible for transforming raw input data into more complex and general feature representations, which enables the model to process more complex data, e.g., image data [5,16,18].

Evaluation and implementation of the model

The final stages of shaping a ML model, once the appropriate training sets are in place, are based on regularization, selection of hyperparameters, metric evaluation and model comparison. Regularization is a process aimed at balancing model accuracy against generalization. Reducing accuracy eliminates the phenomenon of "overfitting," i.e., over-fitting the model's hypotheses against test data, with reduced analysis performance against new data. Reducing the generalization of the model, in turn, prevents overlooking of underlying feature patterns [4,7,19]. Tuning hyperparameters, responsible for the configuration of the model learning processes, is also an important procedure for model performance and accuracy. They are not optimized during model training, so it is necessary to tune them independently. Nowadays we distinguish numerous tuning methods like grid search, random search, Bayesian optimization, evolutionary methods or heuristic algorithms [20,21].

A further procedure involves evaluating the ML model's metrics, which depend on the task the model performs, such as regression, classification or data clustering. Depending on the type of task performed by the model, appropriate statistical tests are selected to calculate the sensitivity, specificity and accuracy of the model [7,22]. During training, the input data determines a specific type of ML, e.g., supervised, however, further selection of the most appropriate statistical algorithm, depends on the evaluation of model metrics. The result should be based on a multi-criteria analysis including cross-validation, model complexity assessment, ROC curves or statistical tests such as the McNemar test for classifier model comparison or the Wilcoxon test for regression models [20,21,22]. In addition, external validation with a completely independent data set is recommended before implementing the model, allowing validation of model performance [4]. The model created in this way is ready for implementation in the healthcare system, where, on a

closed-loop data feedback basis, it is gradually refined, allowing it to be optimized for clinical conditions [19].

RESULTS

Table II. Summary of included studies

Authors, Year (Ref.)	Sample / Data Source	ML Method	Key Findings / Phenotypes
Zhang et al. (2023) [23]	20,881 and 13,724 patients; EHR (2 US cohorts)	Topic modelling + K-means clustering	4 subphenotypes: cardiac/renal; respiratory; musculoskeletal/neurological; digestive
Estiri et al. (2021) [24]	96,025 non-hospitalized patients; EHR	PheWAS + unsupervised ML	33 phenotypes positively associated with prior SARS-CoV-2 infection (anosmia, alopecia, CFS, T2DM)
Gentilotti et al. (2023) [25]	Multinational cohort (ORCHESTRA); clinical/biochemical/QoL data	Hierarchical clustering	4 phenotypes: chronic fatigue-like (CFs), respiratory (REs), chronic pain (CPs), neurosensorial (NSs)
Canas et al. (2023) [26]	9,804 adults; 1,513 post-COVID; self-reported app data	K-means clustering	3 clusters: cardiorespiratory; central neurological; multi-organ systemic inflammatory
Kisiel et al. (2023) [27]	584 adults (non-hospitalised, hospitalised, outpatient clinic)	K-means clustering	3 phenotypes: none/mild, moderate, severe; severity correlated with general health and work ability
Socia et al. (2023) [28]	Large N3C cohort; EHR	Random forest + ML prediction	LC risk prediction model; plant hardiness zone identified as predictor in mild disease model
Reese et al. (2023) [29]	Large N3C/RECOVER cohort; EHR	Unsupervised clustering	6 generalizable PASC clusters: pulmonary, neuropsychiatric, cardiovascular, severe (high mortality), and 2 additional
Epsi et al. (2024) [30]	1,988 SARS-CoV-2-positive adults; clinical + proteomic	Unsupervised clustering	3 clusters: sensory (elevated ICAM-1); fatigue/cognitive (elevated D-dimer, IL-1RA); breathing/exercise intolerance (obesity)
Epsi et al. (2023) [31]	1,273 SARS-CoV-2-positive adults; clinical	Unsupervised clustering	3 early-symptom clusters: nasal; sensory; respiratory/systemic
Klein et al. (2023) [32]	275 individuals with/without LC; blood immune markers	ML + immune phenotyping	Immune features distinguishing LC: altered myeloid/lymphocyte populations, elevated humoral SARS-CoV-2 responses
Ma et al. (2023) [33]	20 ACE2 transgenic mice; Omicron BA.1 model	ML + behavioural/immunological assessment	BA.1 convalescent mice: spontaneous behavioural changes; vaccination protective against pulmonary immune perturbations
Wang et al. (2023) [34]	117 COVID-19 patients + 28 healthy controls; multi-omics	Unsupervised clustering (cytokines, metabolomics, proteomics)	3 phenotypes: Cluster A (molecular deviation); Cluster B (triglyceride/organic acid); Cluster C (heterogeneous; female-predominant, higher symptom burden)
Prabhakaran et al. (2023) [35]	205 COVID-19 survivors; neuropsychological profiles	Unsupervised clustering	3 neurophenotypes: Dysexecutive Function; Memory-Speed Impaired; Normal Cognition

Authors, Year (Ref.)	Sample / Data Source	ML Method	Key Findings / Phenotypes
Lorman et al. (2023) [36]	<21 years; EHR (RECOVER programme)	XGBoost	Algorithm distinguishing PASC, MIS-C, and non-MIS-C variants in paediatric/young adult population
Sonnweber et al. (2022) [37]	145 patients ≥19 years; CT + clinical + lab (days 60, 100, 180 post-onset)	Multiparameter clustering	Pulmonary LC phenotypes; prolonged systemic inflammation associated with structural/functional lung abnormalities
Ayadi et al. (2023) [38]	27,216 Reddit posts (July 2020–July 2022); social media	NLP + clustering	>78% of posts reported ≥1 LC symptom; leading: fatigue, pain, anxiety; heterogeneity of LC profiles characterised
Pfaff et al. (2023) [39]	N3C + All of Us data repositories; EHR	ML phenotype transfer / computable phenotyping	Cross-site reproducibility of N3C-RECOVER LC model demonstrated; open-source phenotyping workflow validated

HER – electronic health record; LC – long COVID; ML – machine learning; NLP – natural language processing; PASC – post-acute sequelae of SARS-CoV-2; MIS-C – multisystem inflammatory syndrome in children

Characterization of different LC phenotypes

Zhang et al. [23] characterized subphenotypes of LC based on 20,881 and 13,724 patients cohorts using ML methods. After identifying potential PASC topics, the authors derived four subphenotypes by clustering patient clinical data based on similar symptoms by focusing on the largest patient cohorts. The following subphenotypes were presented:

- Subphenotype 1 was identified as cardiac and renal. Compared to other subphenotypes, patients were older, had the highest male proportion, and exhibited greater acute COVID-19 severity. It also had the highest rate of SARS-CoV-2 infections during the pandemic's first wave.
- Subphenotype 2 consisted of patients who developed more respiratory comorbidities, sleeping disorders, anxiety and symptoms such as chest pain or headache. Patients were associated with higher prescription rates for anti-asthma, anti-allergy, and anti-inflammatory medications.
- Subphenotype 3 included patients with musculoskeletal and nervous system implications, such as headaches, sleep-wake disorders and musculoskeletal pain. It showed higher baseline comorbidities related to autoimmune, allergic, musculoskeletal, and nervous system conditions, and was associated with increased prescriptions for pain medications.
- Subphenotype 4 included patients with mainly digestive system and respiratory conditions. Patients were identified with higher prescription rates of digestive system medications.

Another study included 96,025 non-hospitalized patients in the final study cohort. Among those 22,475 patients were positive for the SARS-CoV-2 virus [24]. The authors used clinical data at 3–6 months and 6–9 months since the acute phase of the disease to identify phenotypes that positively associate with a past positive reverse transcription-polymerase chain reaction (RT-PCR) test for COVID-19. The analysis resulted in the recognition of 33 phenotypes that were positively associated with previous SARS-CoV-2 infection. Among these phenotypes, new diagnoses of anosmia and dysgeusia, alopecia, chest pain, chronic fatigue syndrome, shortness of breath, pneumonia, and type 2 diabetes mellitus are significant indicators of a past COVID-19 infection. Moreover, the authors emphasised the importance of vaccination to reduce the risk of COVID-19 sequelae.

The study by Gentilotti et al. [25] used clinical and biochemical features, antibody (Ab) response, Variant of Concern (VoC), and physical and mental quality of life (QoL) data to establish clinical phenotypes of post-COVID-19 syndrome. Using ML methods authors described 4 phenotypes: chronic fatigue-like syndrome (CFs), respiratory syndrome (REs), chronic pain syndrome (CPs) and neurosensorial syndrome (NSs). Moreover, the authors presented factors that were associated with those syndromes. Female sex increased risks for CPs, NSs, and CFs, while chronic pulmonary diseases affected more often REs. Early treatment with monoclonal antibodies and vaccination reduced the likelihood of LC. The greatest QoL reduction was in REs and CPs. Female sex, gastrointestinal symptoms, and renal complications raised the risk of severe LC, while vaccination and early treatment mitigated it.

The study by Canas et al. [26] included adult patients who self-reported symptoms through the mobile app. The data were acquired between 2020 and 2021. Authors identified distinct profiles for vaccinated and unvaccinated patients with post-COVID-19 conditions using data about symptom prevalence, duration, demography, and previous comorbidities. From a cohort of 9,804 patients, 1,513 developed post-COVID-19 condition. The authors identified three clusters of symptoms: a cardiorespiratory cluster, a central neurological cluster, and a multi-organ systemic inflammatory cluster. Additionally, different variants were established for cohorts based on vaccination status and infection variant.

Moreover, association with general health status and working ability with LC phenotypes was assessed by Kisiel et al. [27]. The study included 584 non-hospitalised, hospitalised and outpatient clinic adults. The cluster analysis revealed three phenotypes: none/mild, moderate and severe. General health and work ability were self-reported by appropriate scales. Researchers concluded that the most severe phenotype had the greatest effect on overall health and work capability.

National COVID Cohort Collaborative (N3C) Data Enclave Repository

The N3C is one of the largest collections of clinical data in the United States for COVID-19 research. Here, we present two articles utilizing the database, with a third paper to follow later. Socia et al. [28] analysed data from the United States National Institutes of Health (NIH) National COVID Cohort Collaborative (N3C) Data Enclave Repository in order to develop a prediction tool for LC using ML methods. The researchers separated patients into two cohorts: LC after severe or mild disease. Among many factors contributing to the patients' condition the plant hardiness zone is one especially worth mentioning. The plant hardiness zone is a geographic region characterized by a specific average annual minimum temperature and it was established as a significant factor in the Long COVID after Mild Disease Model, indicating a potential influence of climate and sunlight on the progression of LC.

In another study utilizing N3C Data Enclave Repository authors identified six clusters of PASC patients. The identified clusters were: pulmonary, neuropsychiatric, cardiovascular, and a cluster with severe manifestations and increased mortality [29]. Clusters were significantly associated with pre-existing conditions and acute COVID-19 severity. By assigning new patients from other healthcare centres to clusters previously determined authors showed that the clusters were generalizable across different hospital systems.

Proteomics analysis

The study by Epsi et al. [30] included 1,988 SARS-CoV-2 positive individuals aged over 18 years with quantitative post-COVID symptom scores. The authors aimed to phenotype patients based on their symptoms and identify proteomic predictors of LC. The three clusters highlighted by the authors were: a sensory cluster associated with loss of smell and/or taste, a fatigue/difficulty thinking cluster, and a difficulty breathing/exercise intolerance cluster. The sensory cluster was linked to elevated ICAM-1 concentration, the fatigue/difficulty thinking cluster was associated with elevated D-dimer and IL-1RA levels, while the difficulty breathing/exercise intolerance cluster was characterized by obesity and a higher risk of COVID-19 hospitalization. Moreover, Epsi et al. [31] presented in a previous analysis of 1,273 SARS-CoV-2 positive adults three distinct clusters of patients which included Nasal – associated with reporting runny/stuffy nose and sneezing, Sensory – which highly correlated with loss of smell or taste and Respiratory/Systemic cluster associated with among many, cough, trouble breathing, body aches, chills.

Immune phenotyping

Mount Sinai–Yale long COVID (MY-LC) study group aimed to distinguish immune profiling features of LC. It included individuals older than 18 years old [32]. 275 individuals with or without LC were enrolled in a study. The authors used immune phenotyping and ML methods to identify biological

features associated with LC. The cohort of LC patients experienced more intensely reported symptoms and their QoL has deteriorated significantly. The authors described differences in circulating myeloid and lymphocyte populations relative to the matched controls, as well as evidence of exaggerated humoral responses directed against SARS-CoV-2 among participants with LC. The authors described many changes in levels of soluble immune mediators and hormones concluding that integration of immune phenotyping with ML methods provide the key features mostly associated with LC status.

Another study using immune profiling of LC patients used angiotensin-converting enzyme 2 (ACE2) transgenic mice recovered from Omicron (BA.1) infection to test for pulmonary and behavioral post-acute sequelae [33]. The study aimed to characterize the behavioral and immunological effects of Omicron infection and previous vaccination. The study evaluated sera, tissue and behavioral changes among 20 mice, of which 5 were control group, 5 were vaccinated, 5 were Omicron infected and 5 after previous vaccination were also Omicron infected. The authors demonstrated that BA.1 convalescent mice exhibited spontaneous behavioral changes, emotional alterations, and cognitive-related deficits. Moreover, previous vaccination had a protective effect against post-acute pulmonary immune perturbations.

Multiomics phenotyping

A study by Wang et al. [34] used ML methods to identify pathways differentially altered during acute SARS-CoV-2 infection and convalescence. The study included 117 patients above the age of 18 who underwent acute COVID-19 disease and 28 healthy individuals who served as the control group. Infected patients' blood was analyzed during the acute phase of the infection and 6 months after the acute phase. The authors evaluated cytokines metabolomics and proteomics changes and described 20 cytokines and metabolites, which predicted adverse outcomes after discharge. By using unsupervised clustering three distinct disease phenotypes were established. Most individuals were captured by cluster A, which was associated with significant deviation in molecular profile and had the least number of established PASC risk factors. Cluster B was characterized by a predominant triglyceride and organic acid signature molecules. Cluster C exhibited a more heterogeneous composition of cytokines, proteins, and metabolites. What is especially significant cluster C had a higher proportion of women and more frequently reported symptoms such as insomnia, palpitation, shortness of breath, general weakness, and fatigue. These findings may help explain the sex-based differences in PASC prevalence and symptom burden reported in epidemiological studies.

Neurophenotyping

Given that neurocognitive and behavioral changes are major features of LC, studies focusing on these aspects of the disease are particularly important. While the previously mentioned papers include neurocognitive and behavioral abnormalities in their analyses, the paper by Prabhakaran et al. [35] specifically focuses on these changes. The authors examined neuropsychological profiles of 205 individuals aged above 18 years old, who suffered from COVID-19 disease. With the use of unsupervised ML cluster analysis patients were divided into 3 distinct groups: Dysexecutive Function, Memory-Speed Impaired, and Normal Cognition. The first group was characterized by impaired cognitive flexibility and complex attention. Moreover, those patients reported mild to moderate symptoms of anxiety, attention, memory, fatigue, and pain impairment. Second group was characterized by memory deficits, slowed processing speed, and fatigue. Those symptoms were also associated with reporting of anosmia and more severe COVID-19 infection. In the last group cognitive functions were within normal limits with mild impairment. Additionally, this group was associated with previous vaccination. The study highlights the potential use of ML-developed subphenotypes in LC in different therapeutic approaches among selected subphenotypes.

Pediatric population

Although most of the research is focused on adult population the LC complications are also prevalent in pediatric and young adults' population. Lorman et al. [36] developed and validated a ML algorithm which aimed to distinguish patients with PASC, Multisystem Inflammatory Syndrome in Children (MIS-C) and non-MIS-C variants using XGBoost model. The study included individuals younger than 21 years old. The study analyzed 114 condition features, 181 diagnostic test features, 167 procedure features, and 189 medication features. The model developed by the authors could potentially help determine how specific clinical data can help appropriately classify patients and provide a better understanding of PASC in the young adults' cohort.

Lung imaging studies

As acute COVID-19 is primarily associated with pulmonary symptoms, the persistence of these symptoms is highly significant. The study by Sonnweber et al. [37] involved 145 patients who were 19 years old or older and recruited in early 2020. The data gathered encompassed pulmonary computed tomography (CT) scans, lung function (LF) readouts, symptom prevalence, and clinical and laboratory parameters during acute COVID-19 and follow-up visits on days 60, 100, and 180 post-onset. Clinical features and participants were classified using multiparameter clustering and ML techniques. The authors found that prolonged systemic inflammation is closely associated with ongoing structural and functional abnormalities in the lungs and suggested that screening of

medical records could help find patients with incomplete pulmonary recovery and therefore LC reporting.

Social media analysis

Lastly, although clinical data provide systematic and professionally controlled information, social media posts can also be a valuable source of insight into patients' perspectives on the course of LC. As mentioned earlier, ML methods enable us to analyze large and diverse datasets. Ayadi et al. [38] collected 27,216 posts shared between July 2020 and July 2022 on Long COVID-related Reddit forums. They found that more than 78% of the posts mentioned at least one LC symptom such as fatigue, pain or anxiety. Additionally, the authors demonstrated the benefits of using large volumes of data from Reddit to characterize the heterogeneity of LC profiles.

Addressing reproducibility

Pfaff et al. [39] addressed the issue of difficulty in the reproducibility of ML computable phenotypes by reproducing the output of N3C's trained model in the different datasets (All of Us data enclave). The authors demonstrated the extensibility of the previous model across various environments. This study on ML-based phenotype reuse highlights how open-source software practices and cross-site collaboration can enhance the transparency of phenotyping algorithms, reduce unnecessary rework, and foster open science in informatics. The presented workflow is expected to apply to other phenotyping scenarios and aims to encourage more research teams to minimize rework and support open science by sharing their phenotyping and data manipulation code in this way.

DISCUSSION

The studies included in this review employed a range of ML methodologies, with unsupervised clustering approaches — including k-means clustering, hierarchical clustering, and topic modelling — predominating among phenotyping studies. Supervised algorithms such as XGBoost and random forest were primarily employed in predictive tasks, specifically to distinguish PASC from non-PASC patients or to forecast LC risk following acute infection. This methodological distinction is clinically meaningful: unsupervised approaches allow discovery of previously unrecognized patient subgroups without prior assumptions, whereas supervised models require labeled training data and are more suited to prediction than to phenotype discovery. Despite this methodological diversity, no study directly compared competing ML approaches within the same dataset, limiting the ability to draw conclusions about which method yields the most clinically informative or reproducible phenotypes.

Comparison of phenotyping results across studies reveals both convergent and divergent findings. The number of identified clusters varied considerably. Nevertheless, it is worth noting that

recurring patterns of symptoms can be observed across all studies: a respiratory or cardiorespiratory phenotype, a neurological or neurocognitive phenotype, a fatigue-dominant phenotype, and a musculoskeletal or pain-predominant phenotype. These patterns suggest that, despite different datasets, time points, and analytical strategies, LC manifests in broadly consistent clinical domains. Nonetheless, direct comparisons are hampered by the lack of harmonized case definitions, variable follow-up intervals and heterogeneous exposure periods all of which introduce substantial between-study heterogeneity.

Unfortunately, there are several methodological limitations in the studies mentioned that should be pointed out. Studies relying on electronic health records (EHR), such as those utilizing the N3C Data Enclave, are subject to inherent biases: EHR-based phenotypes capture only documented diagnoses, potentially underrepresenting patients who did not seek medical care or whose symptoms were not formally coded. Conversely, studies based on self-reported symptom data may reflect recall bias and are difficult to externally validate. The risk of overfitting is a concern in smaller cohorts, particularly in studies with high-dimensional proteomic or immune data relative to sample size. Furthermore, models trained predominantly on data from earlier pandemic waves may not generalize well to later variants, as the clinical phenotype of LC has evolved across different SARS-CoV-2 strains. The study by Pfaff et al. [39] represents an important step toward addressing cross-site reproducibility, demonstrating that ML-derived phenotypes can be transferred across independent data repositories, though such validation efforts remain the exception rather than the rule. Improved transparency through open-source code sharing and adherence to reporting standards would significantly enhance reproducibility in this field.

The clinical applicability of ML-derived phenotypes is a central question for future research. While the identification of distinct subphenotypes offers theoretical advantages for patient stratification and individualized treatment, translation to routine clinical practice remains limited. Most identified phenotypes were derived from retrospective data and have not been prospectively validated. The integration of multi-omics data and clinical features holds particular promise for uncovering biological mechanisms underlying phenotypic diversity, as demonstrated by Wang et al. [34] and Klein et al. [32]; however, such approaches require substantial infrastructure and remain largely confined to specialist research settings. The present review is further limited by the restriction of the search to a single database (PubMed), the inclusion of preprint studies that have not undergone peer review, and the narrative rather than systematic nature of the synthesis. Future reviews should employ multi-database searches and adhere to established reporting standards to improve comprehensiveness and reproducibility.

CONCLUSIONS

The studies reviewed here demonstrate that ML methods enable data-driven identification of LC subphenotypes. ML facilitates the identification of patients with LC and enables their categorization into distinct subphenotypes, each characterized by different disease manifestations. This stratification may guide more targeted treatment decisions and refine diagnostic criteria. Additionally, our review revealed the potential for employing big data analysis, consisting of various medical features. Prospective validation of ML-derived subphenotypes in diverse clinical settings remains a priority for future research.

Authors' contribution

Study design – K. Żmudka, J. Jaroszewicz, A. Spychał

Data collection – M. Szlenk, R. Kołodziej, P. Piłat

Manuscript preparation – M. Szlenk, K. Żmudka, A. Spychał

Literature research – M. Szlenk, R. Kołodziej, P. Piłat

Final approval of the version to be published – J. Jaroszewicz

REFERENCES

1. Sarmiento Varón L, González-Puelma J, Medina-Ortiz D, Aldridge J, Alvarez-Saravia D, Uribe-Paredes R, et al. The role of machine learning in health policies during the COVID-19 pandemic and in long COVID management. Front Public Health. 2023;11:1140353. doi: 10.3389/fpubh.2023.1140353.

2. Habehh H, Gohel S. Machine Learning in Healthcare. Curr Genomics. 2021;22(4):291–300. doi: 10.2174/1389202922666210705124359.

3. Rubinger L, Gazendam A, Ekhtiari S, Bhandari M. Machine learning and artificial intelligence in research and healthcare. Injury. 2023;54 Suppl 3:S69–S73. doi: 10.1016/j.injury.2022.01.046.

4. Chafai N, Bonizzi L, Botti S, Badaoui B. Emerging applications of machine learning in genomic medicine and healthcare. Crit Rev Clin Lab Sci. 2024;61(2):140–163. doi: 10.1080/10408363.2023.2259466.

5. Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput Sci. 2021;2(3):160. doi: 10.1007/s42979-021-00592-x.

6. Mintz Y, Brodie R. Introduction to artificial intelligence in medicine. Minim Invasive Ther Allied Technol. 2019;28(2):73–81. doi: 10.1080/13645706.2019.1575882.

7. An Q, Rahman S, Zhou J, Kang JJ. A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges. Sensors (Basel). 2023;23(9):4178. doi: 10.3390/S23094178.

8. Jayatilake SMDAC, Ganegoda GU. Involvement of Machine Learning Tools in Healthcare Decision Making. J Healthc Eng. 2021;2021:6679512. doi: 10.1155/2021/6679512.

9. Eckhardt CM, Madjarova SJ, Williams RJ, Ollivier M, Karlsson J, Pareek A, et al. Unsupervised machine learning methods and emerging applications in healthcare. Knee Surg Sports Traumatol Arthrosc. 2023;31(2):376–381. doi: 10.1007/s00167-022-07233-7.

10. Njeri R. Data Preparation for Machine Learning Modelling. Int J Comput Appl Technol Res. 2022;11(06):231–235. doi: 10.7753/IJCATR1106.1008.

11. Masmoudi O, Jaoua M, Jaoua A, Yacout S. Data Preparation in Machine Learning for Condition-based Maintenance. J Comput Sci. 2021;17(6):525–538. doi: 10.3844/jcssp.2021.525.538.

12. Pratheesh R, Divya V. Feature Extraction to Evaluate the Quality of Data Using Machine Learning Technique. In: 2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE). IEEE; 2024, p. 01–05. doi: 10.1109/ICDCECE60827.2024.10549606.

13. A Semary N, Ahmed W, Amin K, Pławiak P, Hammad M. Enhancing machine learning-based sentiment analysis through feature extraction techniques. PLoS One. 2024;19(2):e0294968. doi: 10.1371/journal.pone.0294968.

14. Browning NJ, Ramakrishnan R, von Lilienfeld OA, Roethlisberger U. Genetic Optimization of Training Sets for Improved Machine Learning Models of Molecular Properties. J Phys Chem Lett. 2017;8(7):1351–1359. doi: 10.1021/acs.jpclett.7b00038.

15. Kolasa K, Admassu B, Hołownia-Voloskova M, Kędzior KJ, Poirrier JE, Perni S. Systematic reviews of machine learning in healthcare: a literature review. Expert Rev Pharmacoecon Outcomes Res. 2024;24(1):63–115. doi: 10.1080/14737167.2023.2279107.

16. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018;19(6):1236–1246. doi: 10.1093/bib/bbx044.

17. Kriegeskorte N, Golan T. Neural network models and deep learning. Curr Biol. 2019;29(7):R231–R236. doi: 10.1016/j.cub.2019.02.034.

18. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24–29. doi: 10.1038/s41591-018-0316-z.

19. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol. 2019;20(5):e262–e273. doi: 10.1016/S1470-2045(19)30149-4.

20. Piccolo SR, Mecham A, Golightly NP, Johnson JL, Miller DB. The ability to classify patients based on gene-expression data varies by algorithm and performance metric. PLoS Comput Biol. 2022;18(3):e1009926. doi: 10.1371/journal.pcbi.1009926.

21. de Lacy N, Ramshaw MJ, Kutz JN. Integrated Evolutionary Learning: An Artificial Intelligence Approach to Joint Learning of Features and Hyperparameters for Optimized, Explainable Machine Learning. Front Artif Intell. 2022;5:832530. doi: 10.3389/frai.2022.832530.

22. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. Sci Rep. 2024;14(1):6086. doi: 10.1038/s41598-024-56706-x.

23. Zhang H, Zang C, Xu Z, Zhang Y, Xu J, Bian J, et al. Data-driven identification of post-acute SARS-CoV-2 infection subphenotypes. Nat Med. 2023;29(1):226–235. doi: 10.1038/s41591-022-02116-3.

24. Estiri H, Strasser ZH, Brat GA, Semenov YR; Consortium for Characterization of COVID-19 by EHR (4CE); Chirag J Patel, Shawn N Murphy. Evolving Phenotypes of non-hospitalized Patients that Indicate Long Covid. medRxiv [Preprint]. 2021;2021.04.25.21255923. doi: 10.1101/2021.04.25.21255923.

25. Gentilotti E, Górska A, Tami A, Gusinow R, Mirandola M, Rodríguez Baño J, et al. Clinical phenotypes and quality of life to define post-COVID-19 syndrome: a cluster analysis of the multinational, prospective ORCHESTRA cohort. EClinicalMedicine. 2023;62:102107. doi: 10.1016/j.eclinm.2023.102107.

26. Canas LS, Molteni E, Deng J, Sudre CH, Murray B, Kerfoot E, et al. Profiling post-COVID-19 condition across different variants of SARS-CoV-2: a prospective longitudinal study in unvaccinated wild-type, unvaccinated alpha-variant, and vaccinated delta-variant populations. Lancet Digit Health. 2023;5(7):e421–e434. doi: 10.1016/S2589-7500(23)00056-0.

27. Kisiel MA, Lee S, Malmquist S, Rykatkin O, Holgert S, Janols H, et al. Clustering Analysis Identified Three Long COVID Phenotypes and Their Association with General Health Status and Working Ability. J Clin Med. 2023;12(11):3617. doi: 10.3390/jcm12113617.

28. Socia D, Larie D, Feuerwerker S, An G, Cockrell C. Prediction of Long COVID Based on Severity of Initial COVID-19 Infection: Differences in predictive feature sets between hospitalized versus non-hospitalized index infections. medRxiv [Preprint]. 2023;2023.01.16.23284634. doi: 10.1101/2023.01.16.23284634.

29. Reese JT, Blau H, Casiraghi E, Bergquist T, Loomba JJ, Callahan TJ, et al. Generalisable long COVID subtypes: findings from the NIH N3C and RECOVER programmes. EBioMedicine. 2023;87:104413. doi: 10.1016/j.ebiom.2022.104413.

30. Epsi NJ, Chenoweth JG, Blair PW, Lindholm DA, Ganesan A, Lalani T, et al. Precision Symptom Phenotyping Identifies Early Clinical and Proteomic Predictors of Distinct COVID-19 Sequelae. J Infect Dis. 2025;232(1):39–49. doi: 10.1093/infdis/jiae318.

31. Epsi NJ, Powers JH, Lindholm DA, Mende K, Malloy A, Ganesan A, et al. A machine learning approach identifies distinct early-symptom cluster phenotypes which correlate with hospitalization, failure to return to activities, and prolonged COVID-19 symptoms. PLoS One. 2023;18(2):e0281272. doi: 10.1371/journal.pone.0281272.

32. Klein J, Wood J, Jaycox JR, Dhodapkar RM, Lu P, Gehlhausen JR, et al. Distinguishing features of long COVID identified through immune profiling. Nature. 2023;623(7985):139–148. doi: 10.1038/s41586-023-06651-y.

33. Ma T, Suryawanshi RK, Miller SR, Ly KK, Thomas R, Elphick N, et al. Post-acute immunological and behavioral sequelae in mice after Omicron infection. bioRxiv [Preprint]. 2023;2023.06.05.543758. doi: 10.1101/2023.06.05.543758.

34. Wang K, Khoramjoo M, Srinivasan K, Gordon PMK, Mandal R, Jackson D, et al. Sequential multi-omics analysis identifies clinical phenotypes and predictive biomarkers for long COVID. Cell Rep Med. 2023;4(11):101254. doi: 10.1016/j.xcrm.2023.101254.

35. Prabhakaran D, Day GS, Munipalli B, Rush BK, Pudalov L, Niazi SK, et al. Neurophenotypes of COVID-19: Risk factors and recovery outcomes. Brain Behav Immun Health. 2023;30:100648. doi: 10.1016/j.bbih.2023.100648.

36. Lorman V, Razzaghi H, Song X, Morse K, Utidjian L, Allen AJ, et al. A machine learning-based phenotype for long COVID in children: An EHR-based study from the RECOVER program. PLoS One. 2023;18(8):e0289774. doi: 10.1371/journal.pone.0289774.

37. Sonnweber T, Tymoszuk P, Sahanic S, Boehm A, Pizzini A, Luger A, et al. Investigating phenotypes of pulmonary COVID-19 recovery: A longitudinal observational prospective multicenter trial. Elife. 2022;11:e72500. doi: 10.7554/eLife.72500.

38. Ayadi H, Bour C, Fischer A, Ghoniem M, Fagherazzi G. The Long COVID experience from a patient's perspective: a clustering analysis of 27,216 Reddit posts. Front Public Health. 2023;11:1227807. doi: 10.3389/fpubh.2023.1227807.

39. Pfaff ER, Girvin AT, Crosskey M, Gangireddy S, Master H, Wei WQ, et al. De-black-boxing health AI: demonstrating reproducible machine learning computable phenotypes using the N3C-RECOVER Long COVID model in the All of Us data repository. J Am Med Inform Assoc. 2023;30(7):1305–1312. doi: 10.1093/jamia/ocad077.