National Conference on Behavioural Economics and Intelligent Decision Systems for Climate Change and Sustainable Development
International Journal of Healthcare Education & Medical Informatics
Volume 13, Issue 1&2 - 2026

Review Article

COVID-19 Detection Using Machine Learning: A Dataset-Centric Review

Ravneet Kaur[1], Vipul Sharma[2]

[1]Research Scholar, [2]Assistant Professor, Department of Computer Science and Engineering, I.K. Gujral Punjab Technical University, Kapurthala, Punjab, India
DOI: https://doi.org/10.24321/2455.9199.202603

I N F O

Corresponding Author:
Ravneet Kaur, IK Gujral Punjab Technical University Kapurthala
E-mail Id:
reet.kahlon@gmail.com
Orcid Id:
https://orcid.org/0009-0000-4519-9576
How to cite this article:
Kaur R, Sharma V. COVID-19 Detection Using Machine Learning: A Dataset-Centric Review . J. HealthCare Edu. & Med. Inform. 2026;13(1&2):166-171.

Date of Submission: 2025-10-04
Date of Acceptance: 2025-10-28

A B S T R A C T

COVID-19, which first emerged in 2019, quickly escalated into a global pandemic, officially declared by the World Health Organization (WHO) in March 2020. The infection exhibits symptoms such as fever, dry cough, sore throat, headache, and shortness of breath—similar to pneumonia and influenza—making early and accurate detection challenging. Individuals with underlying health conditions are particularly vulnerable, emphasising the importance of efficient diagnosis and disease management. Machine learning (ML) and deep learning (DL) have become essential tools in COVID-19 research, leveraging diverse datasets to enhance diagnostic accuracy and prediction capabilities. This review focuses on the use of various datasets—clinical, imaging, audio, and multimodal—in ML and DL models for COVID-19 detection and analysis. The study consolidates findings from existing research to evaluate model performance, highlight dataset significance, and identify current limitations, providing a structured perspective on data-driven approaches for pandemic response and healthcare innovation.

Keywords: COVID-19 Detection, Machine Learning, Deep Learning, Clinical Data, Medical Imaging, Multimodal Fusion

Introduction

The emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in Wuhan, China, in December 2019 marked the beginning of a global health crisis, later designated as COVID-19 by the WHO. As of November 2024, more than 776 million confirmed infections and millions of deaths have been reported worldwide.[1,2] underscoring the pandemic's profound public health and socioeconomic consequences. The virus's rapid transmission, high mutation rate, and variable clinical presentation have made early detection and accurate diagnosis vital for mitigating its spread and improving patient outcomes.

COVID-19 has reached all regions of the world, affecting 240 countries. The United States has recorded the highest number of cases (110 million), followed by China (105 million), India (47 million), France (42 million), and Germany (40 million) Table 1. Despite extensive preventive measures and global vaccination campaigns, healthcare systems continue to face challenges related to reinfection, long-term complications, and diagnostic uncertainty. These figures highlight the urgent need for reliable, data-driven diagnostic and prognostic solutions that can aid in real-time decision-making and pandemic control.

Table 1. COVID-19 case statistics as of October 2025 in the most affected countries

Country	Confirmed Cases (Millions)
United States of America	110
China	105
India	47
France	42
Germany	40

SARS-CoV-2 shares significant genetic similarities with SARS-CoV-1 but differs in terms of transmissibility, incubation period, and clinical severity.[3] Unlike earlier coronavirus outbreaks, COVID-19 has led to unprecedented global restrictions, including lockdowns, curfews, and travel bans. Most infected individuals exhibit mild to moderate symptoms—such as fever, dry cough, sore throat, and fatigue—while severe cases can result in pneumonia, Acute Respiratory Distress Syndrome (ARDS), or multi-organ failure.[4] The virus's ability to mutate and generate new variants adds complexity to diagnosis and treatment, necessitating adaptive, intelligent healthcare technologies.

Artificial Intelligence (AI), particularly ML and DL, has emerged as a transformative force in addressing these challenges. These computational techniques can extract meaningful patterns from diverse and complex datasets, providing a foundation for rapid, automated, and precise diagnosis.[5] A PubMed search conducted in November 2024 identified more than 456,000 publications focused on AI-driven COVID-19 detection,[6] reflecting the significant global research investment in data-centric healthcare innovation.

The success of ML and DL models in COVID-19 diagnosis relies heavily on the nature and quality of the datasets employed. Clinical datasets—including patient demographics, symptoms, laboratory parameters, and comorbidities—facilitate risk stratification and severity prediction. Imaging datasets, such as chest X-rays, Computed Tomography (CT), and Magnetic Resonance Imaging (MRI), capture pulmonary abnormalities and disease progression. Audio datasets—comprising cough, breath, and speech recordings—offer non-invasive, cost-effective screening alternatives. Multimodal datasets, integrating these varied data sources, enhance model generalisability and predictive robustness by leveraging complementary features across modalities.

COVID-19's overlapping symptoms with other respiratory disorders, such as pneumonia or Chronic Obstructive Pulmonary Disease (COPD), further complicate clinical diagnosis. In such cases, Computer-Aided Diagnosis (CAD) systems powered by ML and DL can assist clinicians by providing rapid image interpretation, risk assessment, and monitoring capabilities.[7] These dataset-driven approaches offer scalable, accurate, and interpretable diagnostic support, making them indispensable for pandemic preparedness and control.

Given this context, the present review systematically examines the role of dataset types—clinical, imaging, audio, and multimodal—in the development of ML and DL-based COVID-19 detection systems. Section 2 presents the research strategy, Sections 3–7 discuss individual dataset categories and their applications, Section 8 explores ongoing challenges and future directions, and Section 9 concludes with perspectives on data-centric AI for infectious disease diagnostics.

Research Strategy

Studies that used deep learning and machine learning methods on various COVID-19 datasets were methodically assessed in this review.

A thorough search of the literature was done using Elsevier, Google Scholar, PubMed, and IEEE Xplore. "COVID-19", "coronavirus", "epidemic", "computer-aided diagnosis", "machine learning", and "deep learning" were among the keywords. Additional keywords like "X-ray", "CT scan", "MRI", "clinical data", and "audio recordings" were used to find research that was unique to a certain modality.

The goal was to assess machine learning performance for early diagnosis, risk prediction, and patient monitoring across clinical, imaging, audio, and multimodal datasets. The discussion in the following parts was informed by insights gained from this approach.

Clinical Data

Structured patient data, such as comorbidities, lab results, and demographics, are available in clinical databases. Predicting illness severity, ICU admission, and mortality risk requires the use of these datasets. Early intervention tactics have been supported by the broad use of machine learning and deep learning models for clinical datasets.

Using laboratory testing for admission, Barough et al. (2023) created a mortality risk prediction model that had an Area Under the Curve (AUC) of 0.88. For the purpose of classifying ICU patients from non-ICU patients, Chieregato et al. (2022) suggested a hybrid model that combines deep learning and traditional machine learning.

High prediction accuracy was attained by several research studies, including Zhang et al. (2025) utilising support vector machines and Aktar et al. (2021) combining blood parameters with machine learning.[5–11]

Key research using clinical datasets for COVID-19 prediction is included in Table 2. These studies show that organised patient data can yield useful prognostic information when examined using cutting-edge machine learning algorithms.

Challenges with clinical datasets include missing values, inconsistent measurement protocols, and insufficient granularity. Integration with imaging and audio datasets can improve predictive accuracy and robustness.

Imaging Data

Imaging datasets, including chest X-rays and CT scans, provide visual evidence of pulmonary involvement. Convolutional

ISSN: 2455-9199
DOI: https://doi.org/10.24321/2455.9199.202603

neural networks have achieved high diagnostic performance using these datasets. Apostolopoulos and Mpesiana (2020) achieved 96.78% accuracy using transfer learning on chest X-rays. Li et al. (2020) developed a CT-based model distinguishing COVID-19 from other pneumonia with an AUC of 0.92.[12–18]

Hybrid methods combining radiomic feature extraction with classical machine learning have also been employed. Yang et al. (2021) used random forests on CT-derived features, achieving 99% classification accuracy. Segmentation-based approaches further improve interpretability by highlighting infected regions. Table 3 presents a summary of studies using imaging datasets for COVID-19 detection.

Audio Data

Audio datasets, including cough, breathing, and speech recordings, enable non-invasive COVID-19 detection. Imran et al. (2020) developed AI4COVID-19 using cough recordings with 95% accuracy. Laguarta et al. (2020) achieved an AUC of 0.97 using cough and breathing sounds. Temporal models such as long short-term memory networks and ensemble methods capture dynamic acoustic features.[19–25] Table 4 summarizes representative studies on audio datasets.

Table 2.Summary of Studies on Clinical Data for COVID-19 Prediction

Study	Data Type	ML/DL Model	Outcome Predicted	Performance
Barough et al. (2023) [5]	Clinical & Laboratory	Various ML models	Mortality Risk	AUC 0.88
Chieregato et al. (2022) [6]	Clinical	Hybrid ML/DL	ICU vs Non-ICU	AUC 0.81
Zhang et al. (2025) [7]	Clinical	SVM	Severity Risk	AUC 0.994
Aktar et al. (2021) [8]	Blood Parameters	ML & Statistical	Severity Prediction	>90% Accuracy
Yan et al. (2020) [9]	Clinical	XGBoost	Mortality Risk	AUC 0.91
Li et al. (2021) [10]	Clinical	Random Forest	ICU Admission	AUC 0.89
A. Suruliandi et al.(2024) [11]	Clinical	SVM	Severity Risk	94.3% Accuracy

Table 3.Summary of Studies on Imaging Data for COVID-19 Detection

Study	Data Types	Integrated Fusion Approach	Outcome Predicted	Performance
Yang et al. (2021) [12]	Chest X-ray	CNN	Severity Prediction	99% Accuracy
Xue et al. (2021) [13]	CT + Chest X-ray	VGG16	Mortality Risk	99% Accuracy
Pahar et al. (2022) [14]	Cough + Clinical	Feature-level Fusion	COVID-19 Detection	95% Accuracy
Barua et al. (2021) [15]	X-rays	SVM	COVID-19 Detection	99.64% Accuracy
Zouch et al. (2021) [16]	X-ray +CT	CNN	COVID-19 Detection	99.35% Accuracy
Singh et al. (2021) [17]	CT	CNN	Mortality Prediction	95.4% Accuracy
Abdulsalam et al. (2022) [18]	X-ray	CNN	COVID-19 Detection	96% Accuracy

ISSN: 2455-9199
DOI: https://doi.org/10.24321/2455.9199.202603

Multimodal Data

Multimodal datasets combine clinical, imaging, and audio data to leverage complementary information. Zhang et al. (2022) fused chest X-rays and clinical records using CNN and multilayer perceptrons, achieving an AUC of 0.96. Pahar et al. (2022) integrated cough recordings with clinical data, improving detection accuracy from 91% to 95%.[26–30] Table 5 summarizes studies using multimodal datasets.

Comparative Summary of Dataset Modalities

Machine learning and deep learning models for COVID-19 diagnosis exhibit varying strengths depending on the type of dataset used. Table 6 summarizes the comparative advantages, diagnostic roles, and key limitations of clinical, imaging, audio, and multimodal data sources.

This comparative summary illustrates that multimodal fusion approaches consistently outperform single-modality systems by combining complementary features. However, clinical and audio data remain vital for rapid and non-invasive screening, particularly in low-resource contexts.

Table 4.Summary of Studies on Audio Data for COVID-19 Detection

Study	Data Type	Model	Performance
Imran et al. (2020) [19]	Cough	CNN	95% Accuracy
Laguarta et al. (2020) [20]	Cough + Breathing	CNN	AUC 0.97
Brown et al. (2021) [21]	Cough	Ensemble (MFCC + GB)	91% Accuracy
Coppock et al. (2021) [22]	Cough + Speech	LSTM + CNN	92% Accuracy
Pahar et al. (2021) [23]	Cough	ResNet50	94.5% Accuracy
Benmalek et al. (2021) [24]	Cough	PCA+ML	93% Accuracy
Usman et al. (2021) [25]	Speech	ML	Recall 0.7892.

Table 5.Summary of Studies on Multimodal Data for COVID-19 Detection

Study	Data Types	Integrated Fusion Approach	Performance
Santosh et al. (2022) [26]	Audio + Chest X-ray	CNN	98.70%
Hussain et al. (2021) [27]	Cough + X-ray	Hybrid DL	99.89 % Accuracy
khan et al. (2022) [28]	X-ray + Clinical	DL	97 % Accuracy
Turr et al. (2021) [29]	Biomarkers + X-Ray	Decision-level Fusion	AUC 0.94
Kumar et al. (2021) [30]	X-ray + Speech	Attention-based Fusion	98.91% Accuracy

Table 6.Comparative analysis of dataset modalities for COVID-19 diagnosis

Dataset Type	Diagnostic Stage	Key Features	Advantages	Limitations
Clinical	Early screening & risk stratification	Vital signs, symptoms, blood tests, comorbidities	Enables early prognosis; suitable for resource-limited settings	Missing values; variability in lab measurements
Imaging (X-ray, CT)	Confirmation & severity assessment	Pulmonary opacity, consolidation patterns	High diagnostic accuracy; visual interpretability	Requires imaging equipment; radiation exposure
Audio (Cough, Breath, Speech)	Preliminary screening	Acoustic biomarkers, frequency shifts	Non-invasive, fast, cost-effective	Environmental noise; limited large-scale datasets
Multimodal (Clinical + Imaging + Audio)	Comprehensive diagnosis & monitoring	Combined data representations	Highest predictive accuracy; improved robustness	Complex data fusion; higher computational cost

ISSN: 2455-9199
DOI: https://doi.org/10.24321/2455.9199.202603

Challenges and Future Directions

Despite advances, several challenges remain. Dataset heterogeneity, missing values, class imbalance, and limited annotated samples reduce model generalisation. Models trained on single-population datasets may fail when applied externally. Interpretability of deep learning models is limited, necessitating explainable AI methods, such as Grad-CAM and feature importance ranking.

Emerging directions include multimodal fusion, self-supervised learning, contrastive learning, attention-based architectures, and federated learning for privacy-preserving, generalised model development. Future research should prioritise large, diverse datasets and advanced fusion strategies to enhance robustness and facilitate deployment in clinical settings.

Conclusion

This review demonstrates that machine learning and deep learning approaches have effectively utilised clinical, imaging, audio, and multimodal datasets for COVID-19 detection. Clinical datasets facilitate early screening and risk prediction, imaging provides confirmatory evidence of infection severity, and audio data enable non-invasive, rapid community-level testing. Integrating these modalities through multimodal fusion significantly enhances predictive accuracy and robustness.

Despite promising progress, challenges such as dataset heterogeneity, limited labelled data, and model interpretability remain critical. Moreover, real-world deployment faces additional barriers, including data privacy, regulatory validation, and cross-population generalisation. Addressing these challenges through attention-based fusion, self-supervised learning, and federated learning will be key to achieving scalable, explainable, and privacy-preserving AI solutions. Overall, this dataset-driven perspective lays a foundation for advancing intelligent, reliable, and ethically aligned diagnostic systems for future pandemics.

References

1. COVID-19 cases | WHO COVID-19 dashboard. https://data.who.int/dashboards/covid19/cases?n=c (2 October 2025)
2. Apostolopoulos, I.D., Mpesiana, T.A., 2020. Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. Physical and Engineering Sciences in Medicine, 43(2), pp.635–640.
3. https://pubmed.ncbi.nlm.nih.gov/?db=PubMed
4. Brown, C., et al., 2021. Exploring audio biomarkers for COVID-19 detection using cough recordings. IEEE Access, 9, pp.50325–50335.
5. Barough, S.S., Safavi-Naini, S.A.A., Siavoshi, F. et al. Generalizable machine learning approach for COVID-19 mortality risk prediction using on-admission clinical and laboratory features. Sci Rep 13, 2399 (2023). https://doi.org/10.1038/s41598-023-28943-z.
6. Chieregato, M., Frangiamore, F., Morassi, M., Baresi, C., Nici, S., Bassetti, C., Bnà, C., & Galelli, M. (2022). A hybrid machine learning/deep learning COVID-19 severity predictive model from CT images and clinical data. Scientific reports, 12(1), 4329. https://doi.org/10.1038/s41598-022-07890-1
7. Zhang, H., Wang, Y., Xie, Y., Wang, C., Ma, Y., & Jin, X. (2025). Prediction models based on machine learning algorithms for COVID-19 severity risk. BMC public health, 25(1), 1748. https://doi.org/10.1186/s12889-025-22976-x
8. Aktar S, Ahamad MM, Rashed-Al-Mahfuz M, Azad A, Uddin S, Kamal A, Alyami SA, Lin PI, Islam SMS, Quinn JM, Eapen V, Moni MA Machine Learning Approach to Predicting COVID-19 Disease Severity Based on Clinical Blood Test Data: Statistical Analysis and Model Development JMIR Med Inform 2021;9(4):e25884 doi: 10.2196/25884
9. Yan, L., Zhang, HT., Goncalves, J. et al. An interpretable mortality prediction model for COVID-19 patients. Nat Mach Intell 2, 283–288 (2020). https://doi.org/10.1038/s42256-020-0180-7
10. Y. Song et al., "Deep Learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) With CT Images," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 18, no. 6, pp. 2775-2780, 1 Nov.-Dec. 2021, doi: 10.1109/TCBB.2021.3065361.
11. A. Suruliandi, R. Ame Rayan, S. P. Raja. Prediction of COVID-19 Using a Clinical Dataset With Machine Learning Approaches, International Journal of Interactive Multimedia and Artificial Intelligence, (2024), http://dx.doi.org/10.9781/ijimai.2025.01.003
12. Yang, D., Martinez, C., Visuña, L., Khandhar, H., Bhatt, C., & Carretero, J. (2021). Detection and analysis of COVID-19 in medical images using deep learning techniques. Scientific reports, 11(1), 19638. https://doi.org/10.1038/s41598-021-99015-3
13. Xue, X., Chinnaperumal, S., Abdulsahib, G. M., Manyam, R. R., Marappan, R., Raju, S. K., & Khalaf, O. I. (2023). Design and Analysis of a Deep Learning Ensemble Framework Model for the Detection of COVID-19 and Pneumonia Using Large-Scale CT Scan and X-ray Image Datasets. Bioengineering (Basel, Switzerland), 10(3), 363. https://doi.org/10.3390/bioengineering10030363
14. Pahar, M., Klopper, M., Warren, R., & Niesler, T. (2022). COVID-19 detection in cough, breath and speech using deep transfer learning and

bottleneck features. Computers in biology and medicine, 141, 105153. https://doi.org/10.1016/j.compbiomed.2021.105153

15. Barua, P. D., Muhammad Gowdh, N. F., Rahmat, K., Ramli, N., Ng, W. L., Chan, W. Y., Kuluozturk, M., Dogan, S., Baygin, M., Yaman, O., Tuncer, T., Wen, T., Cheong, K. H., & Acharya, U. R. (2021). Automatic COVID-19 Detection Using Exemplar Hybrid Deep Features with X-ray Images. International journal of environmental research and public health, 18(15), 8052. https://doi.org/10.3390/ijerph18158052
16. Zouch, W., Sagga, D., Echtioui, A., Khemakhem, R., Ghorbel, M., Mhiri, C., & Hamida, A. B. (2022). Detection of COVID-19 from CT and Chest X-ray Images Using Deep Learning Models. Annals of biomedical engineering, 50(7), 825–835. https://doi.org/10.1007/s10439-022-02958-5
17. Singh, M., Bansal, S., Ahuja, S., Dubey, R. K., Panigrahi, B. K., & Dey, N. (2021). Transfer learning-based ensemble support vector machine model for automated COVID-19 detection using lung computerized tomography scan data. Medical & biological engineering & computing, 59(4), 825–839. https://doi.org/10.1007/s11517-020-02299-2
18. Abdulsalam Hamwi, W., & Almustafa, M. M. (2022). Development and integration of VGG and dense transfer-learning systems supported with diverse lung images for discovery of the Coronavirus identity. Informatics in medicine unlocked, 32, 101004. https://doi.org/10.1016/j.imu.2022.101004
19. Imran, A., Posokhova, I., Qureshi, H. N., Masood, U., Riaz, M. S., Ali, K., John, C. N., Hussain, M. I., & Nabeel, M. (2020). AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. Informatics in medicine unlocked, 20, 100378. https://doi.org/10.1016/j.imu.2020.100378
20. J. Laguarta, F. Hueto and B. Subirana, “COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings,” in IEEE Open Journal of Engineering in Medicine and Biology, vol. 1, pp. 275-281, 2020, doi: 10.1109/OJEMB.2020.3026928
21. Han, J., Xia, T., Spathis, D. et al. Sounds of COVID-19: exploring realistic performance of audio-based digital testing. npj Digit. Med. 5, 16 (2022). https://doi.org/10.1038/s41746-021-00553-x
22. Coppock, Harry et al.The Lancet Digital Health, Volume 3, Issue 9, e537 - e538
23. Pahar, M., Klopper, M., Warren, R., & Niesler, T. (2021). COVID-19 cough classification using machine learning and global smartphone recordings. Computers in biology and medicine, 135, 104572. https://doi.org/10.1016/j.compbiomed.2021.104572
24. Benmalek, E., El Mhamdi, J., Jilbab, A., & Jbari, A. (2022). A cough-based Covid-19 detection system using PCA and machine learning classifiers. Applied Computer Science, 18(4), 96-115. https://doi.org/10.35784/acs-2022-31
25. Usman, M., Gunjan, V. K., Wajid, M., Zubair, M., & Siddiquee, K. N. (2022). Speech as a Biomarker for COVID-19 Detection Using Machine Learning. Computational intelligence and neuroscience, 2022, 6093613. https://doi.org/10.1155/2022/6093613
26. Santosh Kumar, Rishab Nagar, Saumya Bhatnagar, Ramesh Vaddi, Sachin Kumar Gupta, Mamoon Rashid, Ali Kashif Bashir, Tamim Alkhalifah Chest X ray and cough sample based deep learning framework for accurate diagnosis of COVID-19,Computers and Electrical Engineering,Volume 103,2022,108391,ISSN 0045-7906,https://doi.org/10.1016/j.compeleceng.2022.108391.
27. Hussain, Shabir & Amran, Gehad Abdullah & Alabrah, Amerah & Alkhalil, Lubna & AL-Bakhran, Ali. (2024). C19-MLE: A Multi-Layer Ensemble Deep Learning Approach for COVID-19 Detection Using Cough Sounds and X-ray Imaging. IEEE Access. PP. 1-1. 10.1109/ACCESS.2024.3521938.
28. Khan, I. U., Aslam, N., Anwar, T., Alsaif, H. S., Chrouf, S. M. B., Alzahrani, N. A., Alamoudi, F. A., Kamaleldin, M. M. A., & Awary, K. B. (2022). Using a Deep Learning Model to Explore the Impact of Clinical Data on COVID-19 Diagnosis Using Chest X-ray. Sensors (Basel, Switzerland), 22(2), 669. https://doi.org/10.3390/s22020669
29. Tur K. (2024). Multi-Modal Machine Learning Approach for COVID-19 Detection Using Biomarkers and X-Ray Imaging. Diagnostics (Basel, Switzerland), 14(24), 2800. https://doi.org/10.3390/diagnostics14242800
30. Kumar, S., Chaube, M. K., Alsamhi, S. H., Gupta, S. K., Guizani, M., Gravina, R., & Fortino, G. (2022). A novel multimodal fusion framework for early diagnosis and accurate classification of COVID-19 patients using X-ray images and speech signal processing techniques. Computer methods and programs in biomedicine, 226, 107109. https://doi.org/10.1016/j.cmpb.2022.107109