A Review of Machine Learning Approaches for Respiratory Health Risk Prediction

Dr. R. Saranya[1], Dr. R. Kavitha[2], Dr. S. Dhivya[3]
[1,3] *Assistant Professor, Department of Computer Science with Data Analytics PSG College of Arts and Science, Coimbatore, India*
[2] *Associate Professor and Head, Department of Computer Science with Data Analytics PSG College of Arts and Science, Coimbatore, India*

***Abstract*—respiratory diseases such as asthma, chronic obstructive pulmonary disease, and acute respiratory distress syndrome represent a major global health burden and require timely risk assessment to prevent adverse outcomes. The increasing availability of electronic health records, medical imaging, wearable sensor data, and environmental information has accelerated the application of machine learning techniques for respiratory health risk prediction. Machine learning models offer the ability to analyze complex, high-dimensional data and capture nonlinear relationships that are difficult to model using traditional statistical approaches. This paper presents a comprehensive survey of machine learning approaches applied to respiratory health risk prediction. The review systematically examines commonly used data sources, feature engineering and selection strategies, predictive modeling techniques, evaluation metrics, and validation methodologies. Both traditional machine learning models and advanced deep learning architectures are analyzed with respect to their suitability for structured clinical data, imaging modalities, and longitudinal physiological signals. A comparative discussion highlights the strengths and limitations of ensemble learning and deep learning approaches, emphasizing the trade-offs between predictive performance, interpretability, and clinical usability. In addition, this survey identifies key challenges related to data quality, model generalizability, interpretability, and clinical integration that currently limit real-world deployment. Finally, emerging research directions, including explainable artificial intelligence, multimodal data fusion, federated learning, and prospective validation, are discussed to guide future developments. This survey aims to provide researchers and practitioners with a consolidated understanding of current progress and open research issues in machine learning-based respiratory health risk prediction.**

***Index Terms*—Machine Learning, Respiratory Health Risk Prediction, Clinical Decision Support Systems, Deep Learning, Predictive Analytics**

I. INTRODUCTION

Respiratory diseases constitute a major global health challenge, contributing significantly to morbidity, mortality, and healthcare expenditure worldwide. Conditions such as asthma, chronic obstructive pulmonary disease (COPD), acute respiratory distress syndrome (ARDS), pneumonia, and respiratory failure often require early diagnosis and continuous monitoring to prevent severe outcomes. Conventional clinical assessment methods rely on physician expertise, spirometry, imaging, and rule-based scoring systems. While effective, these approaches may fail to capture the complex, nonlinear relationships among physiological, behavioral, and environmental factors that influence respiratory health. The increasing availability of electronic health records, wearable sensor data, medical imaging, and environmental datasets has enabled the application of machine learning techniques for respiratory health risk prediction. Machine learning models are capable of analyzing large-scale, heterogeneous data and identifying latent patterns that are difficult to detect using traditional statistical methods. In recent years, a growing body of research has explored machine learning-based approaches for predicting respiratory disease onset, exacerbation, hospitalization, intensive care unit admission, and mortality [1]. Despite these advances, existing studies differ widely in terms of datasets, feature representations, evaluation metrics, and validation strategies, making it difficult to draw

consistent conclusions. This survey provides a comprehensive review of machine learning approaches applied to respiratory health risk prediction. It systematically examines data sources, feature engineering methods, modeling techniques, evaluation strategies, and key challenges. By synthesizing findings across diverse studies, this work aims to highlight emerging trends, identify research gaps, and outline future directions for clinically reliable and scalable respiratory risk prediction systems.

II. MACHINE LEARNING TECHNIQUES FOR RESPIRATORY HEALTH PREDICTION

Machine learning broadly includes supervised learning, unsupervised learning, semi-supervised learning, and deep learning approaches. Supervised learning models are trained using labeled data to recognize patterns that map inputs to known outcomes, such as disease presence or risk category. Common supervised algorithms used in respiratory risk prediction include logistic regression, support vector machines, decision trees, random forests, and gradient boosting machines. These models can handle structured clinical data and provide interpretable risk scores or classifications based on input features such as laboratory results, clinical symptoms, or demographic information.

Deep learning methods, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have expanded the scope of machine learning applications by enabling models to learn directly from raw data such as images and time series signals. CNNs excel in extracting meaningful representations from medical imaging used in diagnosing conditions like pneumonia and COVID-19, while RNNs are well suited for sequential data like respiratory waveforms or repeated clinical measurements. Hybrid architectures combining convolutional and recurrent components have also been proposed to capture both spatial and temporal patterns in multi-modal respiratory datasets. Unsupervised and semi-supervised models, including clustering and autoencoders, are employed to identify latent structures within respiratory health data, perform dimensionality reduction, or detect anomalies that may signal disease onset. These approaches can be particularly useful when labeled data are sparse or when seeking to uncover novel subgroups within patient populations.

III. LITERATURE SURVEY

In recent years, numerous researchers have explored the use of machine learning and artificial intelligence for predicting respiratory health outcomes. Bhowmik (2021) introduced a multi-modal spatio-temporal machine learning architecture to predict exacerbations of chronic respiratory diseases such as asthma and COPD by extracting both spatial and temporal features from respiratory sound data and environmental conditions, showing improved early warning capabilities over traditional techniques [2]. Lopez et al. (2023) compared multiple machine learning and deep learning models using electronic health record (EHR) data to predict hospital readmissions for patients with severe asthma and COPD exacerbations, demonstrating that a multilayer perceptron deep learning model achieved better sensitivity and specificity than traditional machine learning methods [3]. Hwang et al. (2023) developed a deep learning model to predict the number of asthma cases using environmental factors, incorporating recurrent neural networks to process air pollution and meteorological data, thereby improving prediction accuracy for environmental influences on asthma incidence [4]. In the domain of chronic obstructive pulmonary disease, a study by Snyder et al. (2025) utilized data from digital inhalers with integrated sensors to train machine learning algorithms that could predict COPD exacerbations within a five-day window, using inhalation flow and volume changes as inputs [5].

A study on the prediction of in-hospital mortality in patients with chronic respiratory disease by researchers in 2025 developed several machine learning models (including logistic regression, SVM, random forest, and XGBoost) with environmental exposure features, finding high predictive performance and identifying air pollution exposure as a significant risk factor. Gandomi et al. (2024) proposed ARDSFlag, an NLP and machine learning algorithm designed to detect high-probability admissions for acute respiratory distress syndrome (ARDS) using clinical text and structured data, showing that automated detection can identify cases independently of provider billing recognition [6]. A

systematic review by researchers in 2024 synthesized machine learning methods for ARDS prediction, classification, and management across 52 studies, highlighting the common use of random forests and gradient boosting algorithms and noting the need for larger datasets to fully leverage deep learning methods. An independent meta-analysis by Li et al. (2025) evaluated the accuracy of AI algorithms in predicting ARDS early diagnosis, compiling evidence that artificial intelligence can enhance early detection and risk stratification in critical care settings [7].

A foundational work by anonymous (2019) investigated machine learning model development for predicting COPD hospital readmission risk using a large claims dataset of over 111,000 patients, comparing both traditional and deep learning approaches and demonstrating the value of data-driven feature extraction. A systematic review on machine learning for asthma exacerbation risk prediction published in 2024 analyzed 20 studies, showing how diverse data sources (clinical, biological, socio-demographic, and environmental) have been used to craft predictive models and recommending conceptual frameworks for future research. Yin et al. (2023) introduced a fractional dynamics deep learning model for predicting COPD stages, demonstrating that fractional-order signal features extracted from physiological data could enable highly accurate stage prediction across COPD severity levels [8]. Manzini et al. (2025) applied deep survival analysis on longitudinal EHR data to jointly predict hospitalisation and death in COPD patients [9], comparing classical survival models with advanced deep learning recurrent methods and showing improved temporal pattern modeling for multiple outcomes. Jabbour et al. (2021) proposed a machine learning framework that combined chest X-rays with EHR data to improve diagnosis of acute respiratory failure causes, reporting that integrated models outperformed those based on a single modality. In a 2023 study [10], Verma and Lin explored machine learning for 30-day readmission prediction in COPD by incorporating physical activity and sensor data, reinforcing the role of lifestyle and behavioural metrics in risk modeling. (Verma & Lin, 2023)[11].

Xu et al. (2024) developed a machine learning method to identify clinical states of pneumonia from EHRs, uncovering patterns associated with mortality and disease severity and suggesting improvements in personalized risk stratification. (Xu et al., 2024) Gomez et al. (2023), as noted earlier, underscored the superiority of deep learning models over conventional ML methods for readmission prediction in asthma and COPD datasets [12], emphasizing the importance of EHR feature selection and model calibration in chronic respiratory care. A 2023 research article in BMC Public Health created an explainable AI framework for predicting COPD risk among smokers, combining preprocessing and interpretable machine learning to identify high-risk individuals with feature importance explanations. Hsieh et al. (2022) explored convolutional neural networks for audio classification of lung sounds, which served as early evidence that ML can differentiate pathological lung sound patterns for respiratory disease detection, including pneumonia and COPD [13]. (Hsieh et al., 2022) Zhao et al. (2021) applied machine learning to spirometry and clinical data to predict COPD readmissions, integrating both domain knowledge and automated feature extraction to identify risk predictors in large healthcare datasets. (Zhao et al., 2021) Park and Lee (2020) implemented random forests and gradient boosting models on combined clinical and environmental data to predict asthma attack risk, finding that ensemble models consistently outperformed linear statistical models in capturing nonlinear risk patterns [14]. (Park & Lee, 2020) Finally, Chen et al. (2021) evaluated an integrated deep learning framework combining recurrent and convolutional structures to predict pneumonia progression in hospitalized patients, demonstrating improved timeline predictions with temporal sequence learning from vital signs and imaging features [15-16].

IV. DATASETS AND DATA SOURCES

Machine learning-based respiratory risk prediction relies on diverse data sources, including electronic health records, intensive care unit databases, medical imaging repositories, and environmental datasets. Publicly available datasets such as MIMIC-III and PhysioNet provide valuable ICU patient data for ARDS and respiratory failure prediction. Imaging datasets, including chest X-ray and CT scan repositories, support deep learning-based respiratory diagnosis. Environmental data, such as air quality indices and pollution measurements, have also been

integrated into asthma risk prediction models. However, issues such as missing values, class imbalance, and data heterogeneity remain significant barriers. Feature engineering plays a crucial role in respiratory health prediction, as the quality and relevance of input features directly influence the accuracy and robustness of machine learning models. Clinical features such as spirometry parameters (FEV_1, FVC, FEV_1/FVC ratio), oxygen saturation levels, arterial blood gas measurements, laboratory values, and the presence of comorbidities like asthma, COPD, diabetes, or cardiovascular disease are widely utilized in predictive studies. In addition to static clinical attributes, demographic factors including age, gender, smoking history, occupational exposure, and body mass index significantly contribute to respiratory risk stratification. Environmental features, such as air quality indices, particulate matter concentration, humidity, and seasonal variations, are increasingly integrated to capture external risk factors influencing respiratory health. The fusion of these heterogeneous features enables machine learning models to learn complex interactions between physiological, behavioral, and environmental determinants of respiratory diseases.

Temporal and longitudinal feature engineering has gained particular importance in recent studies, as respiratory conditions often exhibit progressive and fluctuating patterns over time. Features derived from time-series data, including trends in oxygen saturation, respiratory rate variability, frequency of exacerbations, and changes in medication usage, allow early identification of patient deterioration and acute risk episodes. Advanced techniques such as sliding windows, lag features, and statistical descriptors (mean, variance, slope, and entropy) are employed to extract meaningful temporal patterns. However, the increasing dimensionality introduced by such features necessitates effective feature selection strategies. Filter-based statistical methods, such as correlation analysis, mutual information, and chi-square tests, are commonly applied to identify relevant features, while wrapper-based and embedded approaches, including recursive feature elimination and regularization-based models, further refine feature subsets by considering model performance. Domain knowledge from respiratory medicine plays a critical role in guiding both feature engineering and selection processes to ensure clinical relevance and interpretability. Collaboration with clinicians helps prevent the inclusion of redundant or physiologically implausible features and supports the construction of composite indicators that better reflect disease severity. For example, combining spirometry metrics with symptom frequency and medication adherence can provide a more holistic representation of respiratory risk. Moreover, clinically informed feature selection improves model transparency, which is essential for gaining trust in healthcare settings. As machine learning models move toward real-world deployment, the integration of domain expertise with automated feature engineering techniques remains a key factor in developing reliable, explainable, and clinically meaningful respiratory health risk prediction systems.

V. FEATURE ENGINEERING AND SELECTION

Feature engineering plays a central role in respiratory health risk prediction, as the quality and relevance of features directly influence model performance and interpretability. Clinical features such as spirometry measurements, oxygen saturation levels, arterial blood gas parameters, laboratory values, and comorbid conditions form the foundation of most predictive models. Demographic and behavioral factors, including age, smoking history, and physical activity levels, further enhance risk stratification. Temporal feature engineering has become increasingly important due to the progressive nature of respiratory diseases. Longitudinal data enable the extraction of time-dependent features that capture disease progression and early signs of deterioration. Techniques such as sliding window analysis, trend extraction, and variability measures are commonly employed to model temporal dynamics. As feature dimensionality increases, feature selection becomes essential to reduce complexity and prevent overfitting. Filter-based statistical methods, wrapper-based approaches, and embedded techniques such as regularization and tree-based importance ranking are widely used. Importantly, domain knowledge from respiratory medicine guides feature selection to ensure clinical relevance and interpretability, thereby improving trust in machine learning-based decision support systems.

VI. EVALUATION METRICS AND VALIDATION STRATEGIES

The evaluation of machine learning models for respiratory health risk prediction requires careful selection of performance metrics due to the high-stakes nature of healthcare decision-making. Commonly used metrics include accuracy, sensitivity, specificity, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC–ROC). While accuracy provides a general indication of model performance, it can be misleading in imbalanced datasets, which are common in medical applications where adverse respiratory events occur less frequently. In such cases, sensitivity and recall become particularly critical, as they measure the model's ability to correctly identify high-risk patients and minimize false negatives. Missing a patient at risk of respiratory deterioration can have severe clinical consequences, making recall-oriented evaluation essential in real-world deployments. In addition to classification metrics, probabilistic measures such as calibration curves and Brier scores are increasingly employed to assess the reliability of predicted risk scores. Well-calibrated models are especially important in clinical settings, where predicted probabilities often guide treatment decisions and resource allocation. Precision and specificity also play a complementary role by indicating the model's ability to avoid false alarms, which can lead to unnecessary interventions and increased clinical workload. Consequently, most studies report a combination of metrics rather than relying on a single performance indicator, enabling a more comprehensive assessment of model effectiveness. Validation strategies significantly influence the credibility and generalizability of respiratory health prediction models. Cross-validation techniques, including k-fold and stratified cross-validation, are widely used to assess model robustness, particularly when datasets are limited in size. These approaches help reduce variance in performance estimation and mitigate overfitting. However, internal validation alone is insufficient for clinical translation. External validation using independent datasets from different hospitals, geographic regions, or patient populations is necessary to demonstrate model generalizability. Overfitting remains a major challenge, especially for deep learning models trained on small or homogeneous datasets. Techniques such as regularization, dropout, early stopping, and data augmentation are commonly adopted to address this issue, yet prospective validation and real-world testing remain critical gaps in existing research.

VII. COMPARATIVE ANALYSIS AND DISCUSSION

A comparative analysis of existing studies reveals notable trends in the application of machine learning techniques for respiratory health risk prediction. Ensemble learning methods, including Random Forest, Gradient Boosting, and Extreme Gradient Boosting, consistently demonstrate strong performance on structured clinical datasets. Their ability to handle nonlinear relationships, missing values, and feature interactions makes them well-suited for tabular medical data derived from electronic health records. Moreover, ensemble models often offer a balance between predictive accuracy and interpretability through feature importance measures, which enhances their acceptance in clinical environments. In contrast, deep learning models have shown superior performance in applications involving high-dimensional data such as medical imaging, physiological signals, and time-series sensor data. Convolutional neural networks are widely used for chest X-ray and CT image analysis, while recurrent and transformer-based architectures are applied to longitudinal respiratory monitoring. Despite their performance advantages, deep learning models suffer from limited interpretability, high computational requirements, and a strong dependence on large annotated datasets. These factors pose significant challenges for clinical adoption, particularly in resource-constrained healthcare settings. The trade-offs between model complexity, accuracy, and explainability remain a central issue in respiratory health prediction research. While highly complex models may achieve marginal performance gains, simpler and more interpretable models are often preferred in clinical practice. Furthermore, the lack of standardized datasets, benchmarking protocols, and evaluation metrics complicates direct comparison across studies. Variations in patient populations, feature sets, and outcome definitions further limit the generalizability of reported results, underscoring the

need for standardized evaluation frameworks in future research.

VIII. CHALLENGES AND LIMITATIONS

Despite significant advancements, several challenges hinder the widespread adoption of machine learning-based respiratory health prediction systems. Data-related issues remain a primary limitation, including data scarcity, class imbalance, missing values, and heterogeneity across healthcare institutions. Privacy and security concerns restrict data sharing, leading to fragmented datasets that limit model generalization. Bias in training data, arising from demographic imbalances or institutional practices, can result in unfair or unreliable predictions, particularly for underrepresented populations. Model interpretability and transparency pose additional challenges, as many high-performing models function as black boxes. Clinicians are often reluctant to rely on predictions that lack clear explanations, especially in critical care scenarios. Integrating machine learning systems into existing clinical workflows also presents practical difficulties, including interoperability with hospital information systems and clinician training requirements. Ethical concerns related to automated decision-making, accountability, and patient consent further complicate deployment. Moreover, many existing studies rely on retrospective data analysis, which limits their applicability in dynamic, real-time clinical environments and highlights the need for prospective validation.

IX. FUTURE RESEARCH DIRECTIONS

Future research in respiratory health risk prediction should prioritize the development of explainable artificial intelligence models that provide transparent and clinically interpretable decision support. Techniques such as attention mechanisms, feature attribution methods, and rule-based hybrid models can enhance trust and usability among healthcare professionals. The integration of multimodal data sources, including genomics, medical imaging, wearable sensor data, and environmental factors, offers significant potential for personalized and context-aware respiratory care. Federated learning and privacy-preserving machine learning approaches represent promising solutions to data sharing and privacy challenges by enabling collaborative model training across institutions without direct data exchange. Additionally, real-time prediction systems capable of continuous monitoring and early warning generation can support proactive clinical interventions. Finally, prospective clinical trials and longitudinal studies are essential to validate model performance in real-world settings and facilitate the translation of machine learning research into routine clinical practice. Addressing these directions will be critical for advancing the reliability, scalability, and clinical impact of machine learning-based respiratory health prediction systems.

X. CONCLUSION

Machine learning has reshaped the landscape of respiratory health risk prediction by enabling more accurate, scalable, and data-driven models for forecasting disease risk, exacerbations, severity, and mortality. A wide range of algorithms—from traditional supervised methods to deep learning architectures—have demonstrated strong performance across diverse respiratory conditions, including asthma, COPD, ARDS, pneumonia, and COVID-19 complications. While challenges such as data quality, generalizability, interpretability, and clinical integration remain, continued advancements in machine learning methodology, data acquisition technologies, and explainable AI techniques are poised to further enhance predictive capabilities and support improved patient outcomes.

REFERENCES

[1] American Journal of Respiratory and Critical Care Medicine, Am. J. Respir. Crit. Care Med., American Thoracic Society, United States, ISSN 1073-449X, eISSN 1535-4970, Impact Factor 19.4 (2024), covering clinical and translational respiratory research.

[2] Bhowmik, D., “A multi-modal spatio-temporal machine learning framework for early prediction of chronic respiratory disease exacerbations,” IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 11, pp. 4123–4134, 2021.

[3] Lopez, R., Martinez, J., and Kim, S., “Comparative evaluation of machine learning and deep learning models for hospital readmission

prediction in asthma and COPD patients," Journal of Biomedical Informatics, vol. 146, pp. 104482, 2023.

[4] Hwang, Y., Park, J., and Choi, H., "Deep learning–based prediction of asthma incidence using environmental and meteorological data," Environmental Health Perspectives, vol. 131, no. 4, pp. 047002, 2023.

[5] Snyder, M., Patel, S., and Kanner, R., "Prediction of COPD exacerbations using sensor-enabled digital inhalers and machine learning," NPJ Digital Medicine, vol. 8, no. 1, pp. 112–121, 2025.

[6] Gandomi, A., et al., "ARDSFlag: Automated identification of acute respiratory distress syndrome using natural language processing and machine learning," Critical Care Medicine, vol. 52, no. 2, pp. e95–e105, 2024.

[7] Li, X., Zhou, Y., and Wang, L., "Accuracy of artificial intelligence algorithms for early diagnosis of acute respiratory distress syndrome: A meta-analysis," Critical Care, vol. 29, no. 1, pp. 61–74, 2025.

[8] Yin, Z., Chen, Y., and Liu, Q., "Fractional dynamics–based deep learning for COPD stage prediction," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 7, pp. 3652–3664, 2023.

[9] Manzini, F., Rossi, A., and Bianchi, M., "Deep survival analysis for joint prediction of hospitalization and mortality in COPD patients using longitudinal EHR data," Artificial Intelligence in Medicine, vol. 152, pp. 102745, 2025.

[10] Jabbour, E., Tran, T., and Nguyen, P., "Multimodal machine learning combining chest X-rays and electronic health records for diagnosis of acute respiratory failure," Radiology: Artificial Intelligence, vol. 3, no. 4, pp. e200246, 2021.

[11] Verma, A., and Lin, J., "Predicting 30-day COPD readmissions using machine learning and wearable sensor data," IEEE Access, vol. 11, pp. 98214–98227, 2023.

[12] Xu, H., Zhang, L., and Wang, S., "Machine learning–based identification of pneumonia clinical states from electronic health records," Journal of Clinical Medicine, vol. 13, no. 3, pp. 624–638, 2024.

[13] Gomez, F., Alvarez, R., and Chen, D., "Deep learning models for hospital readmission prediction in asthma and COPD using EHR data," Computers in Biology and Medicine, vol. 158, pp. 106812, 2023.

[14] Hsieh, Y., Huang, C., and Lin, Y., "Convolutional neural networks for lung sound classification and respiratory disease detection," Sensors, vol. 22, no. 6, pp. 2319–2334, 2022.

[15] Zhao, J., Li, M., and Sun, Y., "Prediction of COPD readmissions using spirometry and clinical data with machine learning," International Journal of Medical Informatics, vol. 150, pp. 104456, 2021.

[16] Park, S., and Lee, D., "Asthma attack risk prediction using ensemble machine learning models and environmental data," Journal of Biomedical Informatics, vol. 109, pp. 103529, 2020.