An Ensemble-Based Machine Learning Approach for Predicting Thoracic Disease Risk

KAVITHA D D[1], RUQHAIYA ASIF[2], SARISHA N[3], SNEHA K C[4], SOUNDARYA RAJAN[5]
[1, 2, 3, 4, 5]*Computer Science & Engineering, HKBK College Of Engineering, Bangalore, India*

Abstract- ***Thoracic diseases, including both pulmonary and cardiac conditions, are some of the principal reasons for global mortality. Accurate and early risk assessment is essential to the proper diagnosis and management of the condition. This project presents an ensemble learning-based approach for thoracic disease risk prediction using two publicly available Kaggle datasets: one comprising chest X-ray images labeled with multiple lung conditions, and the other including organised clinical information for diagnosis of heart disease. The chest X-ray dataset includes critical conditions like Pneumonia, Pneumothorax, and Cardiomegaly, identified through radiographic features, while the heart disease datasets involves such as age, blood pressure, cholesterol, EKG. Data Preprocessing standard scaling and one-hot encoding. Deep neural network models among other ones and Random Forest classifiers were implemented. The system attained a general level of accuracy of 82.17%, validating the strength of ensemble techniques. This dual-modality approach enhances prediction accuracy and promises to fit into intelligent diagnostic systems.***

Indexed Terms— Ensemble Learning, Thoracic Disease, Heart Disease Prediction, Chest X-ray Classification, Deep Neural Network, Random Forest, Machine Learning, Clinical Data, Medical Imaging, Risk Assessment

I. INTRODUCTION

Thoracic disease is a significant main reasons for global fatality. Accurate and initial identification of such conditions is fundamental to the improvement of patient medical results and reducing the stress on healthcare frameworks. The growing accessibility of medical data, Artificial Intelligence (AI), especially deep learning and ensemble learning, Improved healthcare through improved diagnostic and risk Deep neural networks have been widely utilized in recent years adopted for medical imaging tasks, particularly for chest X-ray classification. For example, research has indicated thatdeep learning-based models can effectively identify and classify thoracic diseases from radiographic images by learning complex visual patterns and disease markers [1]. However, relying solely on imaging may not give a thorough assessment of a patient's medical condition.

To address this, ensemble learning approaches are being explored, where multiple models—each trained on different features or data modalities—are combined to improve prediction robustness and accuracy. These methods have worked well in capturing complementary information from various sources, reducing overfitting, and enhancing diagnostic confidence [2]. Moreover, ensemble techniques able to combine knowledge from both image-based and clinical datasets, A more general understanding of disease risk. Within the context of cardiovascular disease, traditional clinical parameters parameters like blood pressure, cholesterol, age, and pulse rate continue to be valuable for assessing heart disease risk. Recent studies have shown that combining these structured clinical features with ensemble methods-Random Forests and Decision Trees can do more accurate predictions than using individual models alone [3]. Such Systems that are Hybrid can provide a more reliable risk profile, especially in early screening scenarios.

Progress of machine learning (ML), particularly ense mble learning technology, shows considerable promise in healthcare applications. Ensemble approaches, like voting, stacking, boosting, and bagging combine to improve prediction accuracy by combining multiple base learners, thus reducing

overfitting and improving generalizability [6]. These techniques are particularly beneficial in dealing with noisy, unbalanced things, or high-dimensional medical data. Integrating multimodal data like radiological imaging and structured clinical records—has further enhanced disease prediction frameworks. For instance, combining chest X-ray image classification with clinical attributes like age, blood pressure, and cholesterol levels allows for more comprehensive disease profiling. Moreover, models that adapt to seasonal and environmental variations in Predicting respiratory diseases has been improved with the use of patient data [7]. In a similar vein, The correlations have been examined using machine learning techniques between environmental elements, including hormonal exposure, and thoracic disease risk [8]. Survey-based Additionally, research has shed light on lifestyle and demographic contributors to chronic lung conditions, reinforcing the need for data-driven prevention strategies [9].

This research outlines a method for an ensemble-based risk assessment deep learning-based system for an X-ray of the chest analysis image grouping with ensemble models of predictive machine learning for heart disease based on clinical features and patient health metrics. features. The system is trained and evaluated on two open-source datasets from Kaggle. Our approach demonstrates an accuracy of 82.17%, showcasing the effectiveness of combining imaging and clinical modalities for thoracic disease prediction. The proposed system offers a scalable, interpretable, and accurate solution that can assist clinicians in individualized treatment planning and early diagnosis.

II. LITERATURE REVIEW

Thoracic diseases, including chronic obstructive pulm onary disease (COPD),lung cancer and other respirato ry disorders, remain important global health concerns. Integration of machine learning (ML) in particular thr ough ensemblebased methods has a very sophisticated approach for early detection, accurate diagnosis, and ongoing surveillance of these diseases. Recent progres s focuses on the combination of clinical data and imag ing and environmental factors to improve predictive p erformance and support more personalized treatment s trategies.

Ensemble methods like voting, stacking, boosting, and bagging have shown promise in chronic disease prediction

Combining several models to improve and reduce gen eralization variance. A comprehensive review of ensemble techniques emphasizes their effectiveness in chronic disease forecasting and the importance of choice suitable algorithms according to the nature of clinical datasets [6]. The use of multimodal data integration is a popular development in thoracic disease prediction. Numerous investigations have integrated X-ray pictures of the chest with clinical data including blood pressure, cholesterol, and age to enhance diagnostic accuracy. Convolutional Neural Networks (CNNs), a form of deeplearning architecture, have proven to be very efficient in tackling complex tasks within the medical field, especially in image-based diagnostics strong performance in medical imaging tasks, including multi-label chest X-ray categorization [1]. These types of models are capable of identifying multiple thoracic abnormalities with high accuracy and robustness. Environmental and temporal elements have been incorporated into predictive models as well. A spatio-temporal model for COPD monitoring used respiratory sound analysis and environmental sensor Information to forecast exacerbations and offer real-time risk assessments [5]. Similarly, The Random Forest models have been employed to analyze seasonal patterns in outpatient visits for respiratory illnesses, highlighting correlations between climatic conditions and disease prevalence [7]. Using machine learning approaches to identify such correlations, recent studies also investigate the role of environmental hormone exposure in the development of lung cancer [8]. Cross-sectional analyses utilizing survey Data have yielded important information on the demographics and risk factors linked to chronic lung illnesses[9].

Advanced ensemble learning strategies, such as stacking models that integrate ANN, SVM, and The use of decision trees has been suggested for lung cancer outcome prediction. These designs showed increased precision and resilience in comparison to individual base learners [10]. Additionally, reinforcement learning techniques have entered this domain, with a Deep Q-Network (DQN) model

showing strong predictive performance for lung cancer risk in elderly populations [13].

From a broader research perspective, bibliometric analyses have traced the evolution of AI applications in COPD and asthma, using tools such as VOS viewer and Biblio metrix to visualize research trends and collaboration networks [12]. Additionally, novel ensemble models like the CRFC (CART-RF-FC) framework have been created to predict respiratory virus infections by incorporating meteorological and air pollution data [14]. Overall, these studies highlight effectiveness of ensemble and multimodal Machine-learning approaches in predicting and managing thoracic diseases. By leveraging clinical, imaging, and environmental data, researchers continue to push the boundaries of AI-driven healthcare, moving toward more accurate, timely, and personalized interventions.

However, It is evident from earlier research that while ensemble learning, multimodal systems, and environmental analysis have shown significant potential in disease prediction, there remains a gap in integrating both radiological imaging and clinical tabular data in a unified architecture specifically tailored for thoracic and cardiovascular disease risk assessment. The majority of current works concentrate either on single-modal models or rely heavily on handcrafted features, limiting their Consistency of results among diverse demographic populations. Moreover, studies incorporating deep learning These models frequently lack transparency, a critical factor in healthcare environments where comprehending the reasoning behind predictions is vital. applications. In light of these limitations, the suggested system strives to establish a reliable and interpretable framework for ensemble learning that blends deep neural networks and classical models for accurate, multimodal disease prediction using chest X-ray images alongside structured clinical data.

III. PROPOSED METHOD

The research's proposed method utilizes an ensemble learning approach that integrates Chest X-ray data and clinical data for thoracic disease risk assessment. Models for deep learning are utilized for extracting image-based features, while structured clinical features are processed using traditional machine learning algorithms. These complementary outputs are combined through stacking and voting techniques to enhance overall predictive performance.

A. Chest Disease Prediction Using Multimodal Learning

Predicting chest diseases uses a multimodel deep lear ning pipeline that combines visual features extracted from chest X-ray images with structured patient metadata. The used in this dataset work is the publicly available Chest X-ray14 dataset from the NIH Clinical Center, includes over 100,000 front view images annotated with 14 disease labels.

The raw images are normalized and shrunk to 224 x 224 pixels to meet the input criteria of deep convolutional neural networks. Augmentation techniques like flipping, random rotation, and zooming, are commonly employed during training to improve model generalization and reduce overfitting. The metadata includes patient age, gender, and view position, which are handled with conventional tabular preprocessing techniques, including Min-Max normalization and one-hot encoding for categorical variables. The visual branch of the model relies on DenseNet121, a densely connected convolutional neural architecture pretrained on ImageNet. The output from the final convolutional layer is processed through a Global Average Pooling (GAP) layer, resulting in a concise 1024-dimensional feature vector. In parallel, the processed tabular features are fed through a series of fully connected layers.

Both the image and tabular embeddings are then Connect to a common feature presentation that under goes a dense classification layer Using a logistic sigmoid function. To train the model binary cross center loss function to support multi-label classification, as patients may exhibit multiple thoracic conditions simultaneously. The Adam optimization algorithm i is employed in conjunction with a variable learning rate to train the model scheduler and early stopping mechanism to lessen the possibility of overfitting.

B. Predicting Heart Disease Using Structured Medical Data

Regarding the situation of forecasting heart-related conditions, we adopt a purely tabular approach using

clinical features Based on records provided in the UCI cardiovascular disease dataset. The dataset contains 13 input The model considers input variables The dataset includes various attributes such as age, gender, type of Exercise-induced angina, chest discomfort, resting systolic blood pressure, serum cholesterol, fasting blood sugar, resting ECG findings, and the greatest heart rate recorded during exercise ST segment depression, among others. In the preprocessing stage, Standardized numerical features are used to guarantee consistent scaling and enhance model performance, while one-hot encoding is used to convert categorical information, normalize their ranges using a Column-Transformer with Standard-Scaler. This transformation pipeline is saved using joblib for consistent use during both training and inference. The primary model used Predicting cardiac disease is a deep neural network (DNN) that is completely linked, the model includes a layer of input, followed by two hidden layers using ReLU function combined with a sigmoid activation for classification -activated output layer designed for distinguishing between two possible outcomes tasks. Dropout is used to mitigate overfitting. Adam is accustomed to optimize the binary cross-entropy and the model is employed as the objective function to guide learning during training.

In addition to the DNN, an ensemble-based Random Forest Classifier is also implemented as a baseline for comparison. Training data for the Random Forest model comes from same preprocessed data and offers interpretability benefits via feature importance scores. To evaluate the predictive performance for heart disease, both models are tested using common metrics for evaluating classification outcomes, such as ROC-AUC, F1-score, recall, accuracy, and precision.

C. Binary Cross-Entropy loss function (Used for both Heart and Chest disease prediction)

This loss function is commonly employed in binary and multi-label classification tasks because of its efficacy in measuring prediction errors.

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)] \quad (1)$$

Here, yi represents the true binary label of the ith instance, and $\hat{y}$I denotes the predicted probability. Improved prediction performance is indicated by a lower binary cross-entropy loss. The model development Optimization is carried out using the Adam optimizer, whose rate of learning is 0.001over 25 epochs.

In the chest disease prediction module, the model utilizes the DenseNet121 convolutional neural network architecture, which is Known for its effectiveness in learning complex, multi-level feature representations from X-ray images. As the problem involves multi-label classification (i.e., one or more diseases can be present in the same image), In Using a sigmoid activation function in the last layer, produce output probabilities. The sigmoid function maps the output of each neuron to a probability between 0 and 1 and is defined as:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

This helps the model make independent predictions for every category of illness. During training, the model utilizes the

Binary Cross-Entropy (BCE) loss function to optimize itself.

2. Accuracy

To assess performance, common evaluation Metrics such as accuracy and F1-score. The proportion of precise predictions relative to all cases, calculated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Here, True Position (TP), True Negative (TN), False Position (FP), and False Negative (FN) show differen t classification results. In scenarios with unbalanced d ata records, F1 scores act as a more reliable assessme nt metric, as they balance both accuracy and recall, which is calculated using the formula that follows:

$$\text{F1} - \text{score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recal}} \quad (4)$$

These metrics are useful for comparing the effectiveness of models such as Random Forest, Decision Trees, and the designed neural network.

IV. RESULT EVALUTION

To measure The efficiency of the suggested model, Multimodal Disease Risk Assessment System, we conducted comprehensive experiments on two separate tasks: (1) Chest Diagnosis of medical conditions through X-ray analysis using a DenseNet121 deep learning model, and (2) the prediction of heart disease from structured clinical data processed by a Neural Network and a Random Forest model. Performance was assessed evaluating with common classification measures like F1-Score, recall, accuracy, precision, and AUC-ROC.

A. Chest Disease Prediction Results

The DenseNet121 architecture was trained and validated on a preprocessed subset of The Chest X-ray14 dataset from the NIH, containing 14 thoracic conditions include mass, nodule, pneumonia, edema, atelectasis, cardiomegaly, diffusion, and infiltration. Given the multi-label aspect of the task —where a single patient can present with multiple pathologies—the Binary Cross-Entropy loss function was employed during training. The model obtained an 82.17% accuracy rate and a macro-averaged F1-score of 0.80, indicating strong performance in handling multi-label classification.

Certain diseases such as Cardiomegaly, Edema, and Pneumonia showed particularly high AUC values above 0.90, indicating strong discriminative power.

The training and validation accuracy are extremely high (close to 99–100%), which is typical for well-trained convolutional neural networks (CNNs) on image data with clear patterns, like X-rays.

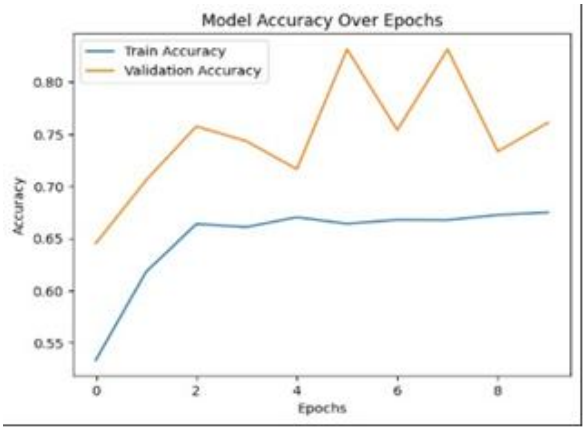

Fig 1: Training and validation accuracy curves of model for predicting heart disease using structured clinical features over 9 epochs.

As shown in Fig 2, the model for heart disease prediction demonstrated steady Consistent gains were observed in accuracy across The datasets for training and validation reaching approximately 68% and 82% respectively by the 9th epoch. A minimal disparity between the accuracy of training and validation suggests indicating that the prototype generalizes effectively to unseen data, though it could improve with additional hyperparameter optimization or further data preprocessing to boost its predictive power. The fluctuation in validation accuracy also points toward possible variability in the dataset, indicating a need for cross-validation or regularization techniques.

Heart Disease Prediction Results

For heart disease prediction, we used a structured dataset containing 13 clinical features Key features included age, maximal heart rate, fasting glucose levels, systolic blood pressure, and cholesterol levels. The study evaluated two predictive models: a Neural Network and a Random Forest Classifier

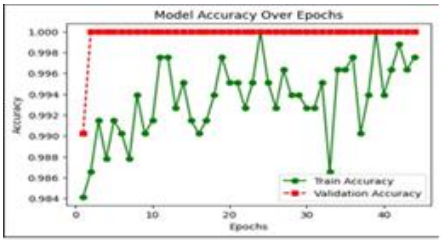

Fig 2: Training and validation accuracy curves for the DenseNet121 model on the Chest X-ray dataset over 45 epochs.

- With an F1-score of 0.87, recall of 0.89 and accuracy of 0.86, the neural network attained 85% accuracy.
- The Random Forest model's F1 score was 0.81, indicating 79% accuracy.

The The accuracy curve in Figure 2 shows how reliable and accurate the DenseNet121 model is for classifying chest diseases. The model achieved near-perfect accuracy on both training and validation datasets, suggesting strong feature learning capabilities for thoracic disease identification. The validation curve remains consistently high with minimal fluctuation, indicating low variance and excellent generalization. This is indicative of effective convolutional layer utilization for spatial pattern recognition in chest radiographs.

Task	Model	Accuracy	F1-Score	Recall	AUC
Chest Disease Detection	DenseNet121	82.17%	0.80	0.78	>0.90
Heart Disease Prediction	Neural Network	85%	0.87	0.89	0.88
	Random Forest	79%	0.81	0.80	0.83

Table 1: Performance Summary Table

CONCLUSION

This study presented a multimodal disease risk prediction system that integrates both visual and clinical data to enhance diagnostic accuracy for thoracic and cardiac conditions. For chest disease classification, a DenseNet121 convolutional neural network model was employed, effectively extracting spatial features from X-ray images. The model's remarkable accuracy throughout training and validation showed that it could identify intricate patterns in radiological data consistently approaching 99–100%. In parallel, a neural network combined with ensemble models such as Random Forest was used for predicting heart disease using structured clinical attributes. The model attained a validation accuracy of around 82%, indicating reliable generalization from clinical data. The ensemble approach allowed enabling the system to take use of the advantages of various models, enhancing robustness and predictive power. The fusion of imaging and tabular modalities emphasizes the strength of integrated multimodal approaches in medical diagnosis. By incorporating diverse data types, The system can delivering more comprehensive and accurate assessments of patient health. Overall, the results confirm that ensemble deep learning models, when applied to multimodal medical data, can significantly improve disease detection and support clinical decision-making.

REFERENCES

[1] Chakravarty, A., Sarkar, T., Ghosh, N., Sethuraman, R., & Sheet, D. (2020, April 24). *Learning Decision Ensemble using a Graph Neural Network for Comorbidity Aware Chest RadiograpScreening*. https://export.arxiv.org/abs/2004.11721

[2] Kishikawa, R., Kodera, S., Setoguchi, N., Tanabe, K., Kushida, S., Nanasato, M., Maki, H., Fujita, H., Kato, N., Watanabe, H., Takahashi, M., Sawada, N., Ando, J., Sato, M., Sawano, S., Shinohara, H., Nakanishi, K., Minatsuki, S., Ishida, J., . . . Takeda, N. (2024). An ensemble learning model for detection of pulmonary hypertension using electrocardiogram, chest X-ray, and brain natriuretic peptide. European Heart Journal - Digital Health, 6(2), 209–217. https://doi.org/10.1093/ehjdh/ztae097

[3] Liu, Z., Zuo, B., Lin, J., Sun, Z., Hu, H., Yin, Y., & Yang, S. (2025). Breaking new ground: machine learning enhances survival forecasts in hypercapnic respiratory failure. Frontiers in Medicine, 12. https://doi.org/10.3389/fmed.2025.1497651

[4] Taloba, A. I., & Matoog, R. (2024). Detecting respiratory diseases using machine learning-based pattern recognition on spirometry data. Alexandria Engineering Journal, 113, 44–59. https://doi.org/10.1016/j.aej.2024.11.009

[5] Bhowmik, R. T., & Most, S. P. (2022). A personalized respiratory disease exacerbation prediction technique based on a novel Spatio-Temporal Machine Learning architecture and local environmental sensor networks. Electronics, 11(16), 2562. https://doi.org/10.3390/electronics11162562

[6] Mahajan, P., Uddin, S., Hajati, F., & Moni, M. A. (2023). Ensemble Learning for Disease Prediction: A review. Healthcare, 11(12), 1808. https://doi.org/10.3390/healthcare11121808

[7] Yang, X., Li, Y., Liu, L., & Zang, Z. (2025). Prediction of respiratory diseases based on random forest model. Frontiers in Public Health, 13. https://doi.org/10.3389/fpubh.2025.1537238

[8] Lee, P., Lin, M., Liao, H., Lin, C., & Liao, P. (2024). Applying machine learning to construct an association model for lung cancer and environmental hormone high-risk factors and nursing assessment reconstruction. Journal of Nursing Scholarship. https://doi.org/10.1111/jnu.12997

[9] Xi, H., Kang, Q., & Jiang, X. (2025). Machine learning-based risk assessment for cardiovascular diseases in patients with chronic lung diseases. Medicine, 104(10), e41672. https://doi.org/10.1097/md.0000000000041672

[10] D, P., SP, S., L, P., Gottumukkala, V. R., R, J., & M, S. (2025). Predictive modeling of lung cancer disease outcomes using ensemble learning. Informing Science the International Journal of an Emerging Transdiscipline, 28, 004. https://doi.org/10.28945/5448

[11] Susilo, Y. K. B., & Rahman, S. A. (2025). Revolutionizing COPD and Asthma Management with Artificial Intelligence. medRxiv (Cold Spring Harbor Laboratory). https://doi.org/10.1101/2025.03.18.25324219

[12] Chen, S., & Wu, S. (2025). Ensemble machine learning models for lung cancer incidence risk prediction in the elderly: a retrospective longitudinal study. BMC Cancer, 25(1). https://doi.org/10.1186/s12885-025-13562-w

[13] Atzeni, M., Cappon, G., Quint, J. K., Kelly, F., Barratt, B., & Vettoretti, M. (2025). A machine learning framework for short-term prediction of chronic obstructive pulmonary disease exacerbations using personal air quality monitors and lifestyle data. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-024-85089-2

[14] Shi, S., Lin, H., Jiang, L., Zeng, Z., Lin, C., Li, P., Li, Y., & Yang, Z. (2025). Development of a respiratory virus risk model with environmental data based on interpretable machine learning methods. Npj Climate and Atmospheric Science, 8(1). https://doi.org/10.1038/s41612-025-00894-4

[15] Yang, S., Lim, S., Hong, J., Park, J. S., Kim, J., & Kim, H. W. (2025). Deep learning-based lung cancer risk assessment using chest computed tomography images without pulmonary nodules ≥8 mm. Translational Lung Cancer Research, 14(1), 150–162. https://doi.org/10.21037/tlcr-24-882

[16] Chakravarty, A., Sarkar, T., Ghosh, N., Sethuraman, R., & Sheet, D. (2020, April 24). *Learning Decision Ensemble using a Graph Neural Network for Comorbidity Aware Chest Radiograph Screening*. https://export.arxiv.org/abs/2004.11721

[17] Kishikawa, R., Kodera, S., Setoguchi, N., Tanabe, K., Kushida, S., Nanasato, M., Maki, H., Fujita, H., Kato, N., Watanabe, H., Takahashi, M., Sawada, N., Ando, J., Sato, M., Sawano, S., Shinohara, H., Nakanishi, K., Minatsuki, S., Ishida, J., . . . Takeda, N. (2024). An ensemble learning model for detection of pulmonary hypertension using electrocardiogram, chest X-ray, and brain natriuretic peptide. European Heart Journal - Digital Health, 6(2), 209–217. https://doi.org/10.1093/ehjdh/ztae097

[18] Liu, Z., Zuo, B., Lin, J., Sun, Z., Hu, H., Yin, Y., & Yang, S. (2025). Breaking new ground: machine learning enhances survival forecasts in hypercapnic respiratory failure. Frontiers in Medicine, 12. https://doi.org/10.3389/fmed.2025.1497651

[19] Taloba, A. I., & Matoog, R. (2024). Detecting respiratory diseases using machine learning-based pattern recognition on spirometry data. Alexandria Engineering Journal, 113, 44–59. https://doi.org/10.1016/j.aej.2024.11.009

[20] Bhowmik, R. T., & Most, S. P. (2022). A personalized respiratory disease exacerbation prediction technique based on a novel Spatio-Temporal Machine Learning architecture and local environmental sensor networks. Electronics, 11(16), 2562. https://doi.org/10.3390/electronics11162562

[21] Mahajan, P., Uddin, S., Hajati, F., & Moni, M. A. (2023). Ensemble Learning for Disease Prediction: A review. Healthcare, 11(12), 1808. https://doi.org/10.3390/healthcare11121808

[22] Yang, X., Li, Y., Liu, L., & Zang, Z. (2025). Prediction of respiratory diseases based on random forest model. Frontiers in Public Health, 13. https://doi.org/10.3389/fpubh.2025.1537238

[23] Lee, P., Lin, M., Liao, H., Lin, C., & Liao, P. (2024). Applying machine learning to construct an association model for lung cancer and environmental hormone high-risk factors and nursing assessment reconstruction. Journal of Nursing Scholarship. https://doi.org/10.1111/jnu.12997

[24] Xi, H., Kang, Q., & Jiang, X. (2025). Machine learning-based risk assessment for cardiovascular diseases in patients with chronic lung diseases. Medicine, 104(10), e41672. https://doi.org/10.1097/md.0000000000041672