

INTELLIGENT SYSTEM FOR EARLY DETECTION OF LUNG DISEASES BASED ON MULTIMODAL ANALYSIS OF AUDIO SIGNALS AND X-RAY IMAGES

Talat Magrupov¹, Muhammadali Nosirov²

¹Tashkent State Technical University, Tashkent, Uzbekistan

Doctor of Engineering Sciences, Professor

²Tashkent University of Information Technologies, Tashkent, Uzbekistan

PhD Candidate

Corresponding author email: muhammadalinosiroy34@gmail.com
talatmt@rambler.ru

Abstract: This article proposes an intelligent system for the early detection of lung diseases based on the joint analysis of audio signals and chest X-ray images. The proposed system enables reliable and highly accurate diagnosis of respiratory diseases by integrating audio signals and X-ray images using deep neural networks. A key feature of the system is the personalization of the analysis. During initial use, the system generates an individual patient profile, including age, gender, medical history, the presence of chronic diseases, and basic respiratory parameters. Subsequently, all new medical data is compared not only with the general training set but also with the patient's personal norm. This is particularly effective for monitoring chronic conditions such as chronic obstructive pulmonary diseases, bronchial asthma, and others, where tracking the dynamics of these parameters plays a key role. Thus, the proposed intelligent system provides an automated, accurate, and scalable process for diagnosing respiratory diseases, reducing the time it takes to establish a preliminary diagnosis and increasing the availability of medical care. It also stores patient data in a database for subsequent dynamic monitoring and analysis of changes in the patient's condition over time.

Keywords: intelligent system, lung diseases, audio signals, X-ray images, deep neural networks, methodology

INTRODUCTION

Lung diseases are a leading global health problem, causing significant morbidity and mortality. Early detection of these conditions is key not only to promptly administer effective treatment but also to prevent complications such as chronic respiratory failure, secondary infections, and the progression of pathological changes in lung tissue. In today's climate, with the growing number of patients with respiratory diseases, the need to develop automated diagnostic systems capable of quickly and objectively analyzing medical data is particularly pressing.

Traditional diagnostic methods, including auscultation with a stethoscope and visual analysis of X-ray images, have limitations related to the physician's subjective assessment, dependence on the specialist's experience, and limited sensitivity in the early detection of pathologies [1-4]. Minor changes in lung structure or subtle abnormalities in breath sounds often go undetected, complicating early diagnosis and increasing the risk of delayed treatment. With the development of artificial intelligence (AI) methods, approaches have emerged that automate diagnostics based on the analysis of medical data. However, most existing methods analyze only one type of data—either images [5–9] or audio signals [10–13].

Addressing these issues with modern artificial intelligence (AI) and deep learning technologies based on generative and diffusion neural networks can significantly improve the accuracy of lung disease diagnosis. The use of multimodal analysis, which simultaneously

incorporates respiratory audio signals and chest X-ray images, enables the integration of acoustic and visual data, creating a more complete picture of the patient's respiratory system. This approach enables the identification of complex disease patterns that are easily detected using multiple types of medical data.

The aim of this study is to create an intelligent system for the early detection of lung diseases by integrating audio signals and X-ray images using multimodal deep neural networks.

MATERIALS AND METHODS

This paper proposes a method for the early detection of lung diseases based on multimodal processing and analysis of audio signals and X-ray images, which improves the reliability of diagnostic decisions through the synergy of information using diffusion neural networks (DNNs). The key principles of processing and analyzing audio signals and X-ray images are defined.

Respiratory audio signal processing involves solving the following tasks: removing background noise from recordings, segmenting inhalation and exhalation, extracting key features using small-frequency cepstral coefficients (MFCCs), generating spectrograms, and applying data augmentation methods (e.g., adding background noise, pitch shifting, and time stretching). These methods enhance the model's ability to distinguish subtle respiratory abnormalities, such as wheezing, rales, or weakened breathing, which is especially important for the early detection of pathologies.

The signal conversion was performed as follows:

$$x_{mfcc} = \left(\log \left(\text{Mel} \left(\left| \text{STFT} \left(x(t) \right) \right| \right) \right) \right) \quad (1)$$

X-ray image processing involves solving the following problems: brightness and contrast normalization, image size standardization, data augmentation (rotations, horizontal flips, brightness and contrast changes), and feature extraction using diffusion neural networks. Methods based on these problems enable the automatic detection of inflammation, lung tissue consolidation, signs of chronic diseases, and structural changes characteristic of pneumonia, bronchitis, COPD, and other respiratory pathologies.

Based on the research conducted, an intelligent system has been proposed that combines two sources: audio signals and X-ray images, improving the reliability of diagnostic decisions through the synergy of information based on diffusion neural networks (Figure 1).

The first stage involves collecting data using digital stethoscopes, microphones, and video cameras. Audio signals capture the patient's respiratory sounds—wheezing, crepitations, whistling, and wet sounds—while the video stream allows for analysis of respiratory rate, chest movement amplitude, and the presence of accessory respiratory muscles.

The second stage involves pre-processing the signals. The audio signals are cleaned of background noise using digital filtering, normalization, and respiratory cycle segmentation. Video data undergoes image stabilization, region-of-interest extraction, and key motion parameter extraction.

The next stage utilizes convolutional neural networks (CNNs) to analyze respiratory sound spectrograms and recurrent neural networks (RNN/LSTMs) to estimate the temporal dynamics of respiration.

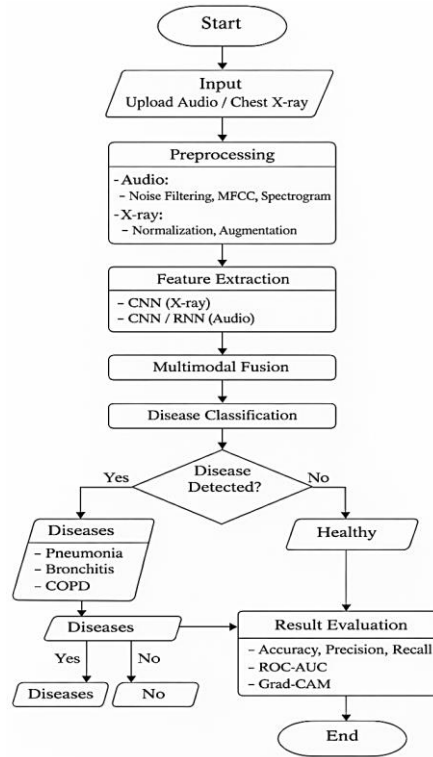


Figure 1 - Block diagram of the structure of an intelligent system.

The architecture of the proposed intelligent system is designed according to a modular principle, which allows flexible integration of new data sources and analytical modules. Each module performs a specific function in the data processing pipeline, ensuring scalability and adaptability of the system.

The data acquisition module collects respiratory sound recordings and chest X-ray images using medical sensors and imaging devices. The preprocessing module performs noise reduction, normalization, and feature extraction to prepare the data for further analysis. The analytical module utilizes deep neural networks to extract complex patterns and relationships between multimodal features. Finally, the decision-making module generates a probabilistic diagnosis and provides recommendations for further medical examination or consultation.

Fusion of audio and video features is performed at the feature or decision level, which improves the model's robustness to incomplete or noisy data.

An adaptive attention mechanism was used for feature fusion:

$$\alpha = \sigma(W_1 H_a + W_2 H_x + b) \quad (2)$$

$$F = \alpha * H_a (1 - \alpha) * H_x \quad (3)$$

where the coefficient α is trained in the optimization process,

- α – the adaptive attention coefficient shows the relative information content of the audio modality and dynamically regulates the contribution of each modality to the final decision.
- σ is the sigmoid activation function, constraining α to be in the range $[0,1]$;
- H_a is the audio signal feature vector;
- H_x is the X-ray image feature vector;
- W_1, W_2 are the trainable weight matrices for the audio and visual modalities, respectively;
- b – trainable bias.
- F – final combined feature vector;
- $\alpha * H_a$ – weighted audio features. $(1-\alpha)$;

The final solution allowed us to dynamically take into account the information content of each modality.

At the classification stage, the trained model compares the obtained parameters with patterns characteristic of various lung pathologies. This results in a probabilistic assessment of the presence of a specific disease, as well as its severity.

During the model performance evaluation stage, the following steps are performed:

- Accuracy, Recall, and F1-score – calculation of standard metrics for quantitative assessment of classification quality;
- ROC-AUC analysis – plotting the ROC curve and calculating the area under it to assess the model's ability to distinguish between classes;
- Grad-CAM visualization – visualization of key areas in X-ray images that had the greatest impact on the model's decision, increasing interpretability and physician confidence.

Thus, the system includes an adaptive learning module that improves recognition accuracy as new clinical data accumulates. When a physician confirms a diagnosis, the results are automatically used for further model adjustments (while maintaining confidentiality), ensuring a gradual improvement in classification quality and a reduction in the likelihood of false positives and false negatives.

To ensure resilience to external conditions, the system includes automatic data quality control mechanisms. The system notifies the user of the need to re-record if the background noise level exceeds acceptable limits for accurate video analysis or lighting is insufficient. This helps minimize errors associated with incorrect data collection.

The solution's architecture is modular and includes:

- Data collection module;
- Pre-processing module;
- Neural network-based analytics module;
- Decision-making module;
- User and physician interaction interface.

During remote monitoring, data can be transmitted via secure communication channels to a cloud infrastructure, where centralized analysis and storage of the information is performed. The physician has access to visualization of the data in the form of respiratory activity graphs, spectral characteristics of sounds, and risk indicators. If any deviations are detected, the system can automatically generate warnings and recommendations for specialist consultation.

Integration with medical information systems and electronic medical records is also supported. This provides a comprehensive approach to diagnostics, allowing for the integration of laboratory parameters, chest X-ray results, and audiovisual data.

In terms of computational implementation, two operating modes are possible: local (on a mobile device or embedded microprocessor) and cloud. Local mode provides efficiency and autonomy, while cloud mode enables more in-depth analysis using diffusion-based deep learning models. Feature extraction plays a crucial role in the effectiveness of machine learning models for respiratory sound analysis. Among various acoustic feature extraction techniques, Mel-frequency cepstral coefficients (MFCCs) are widely used due to their ability to represent the spectral characteristics of audio signals in a compact and informative form. MFCCs approximate the human auditory perception system and allow the extraction of key frequency components related to pathological breathing sounds.

The MFCC extraction process includes several steps: pre-emphasis filtering, frame segmentation, application of the Hamming window, fast Fourier transform (FFT), Mel-scale filter bank processing, and logarithmic compression. The resulting cepstral coefficients represent the spectral envelope of the audio signal and serve as input features for neural network models.

The proposed methodology relies on the integration of heterogeneous medical data sources within a unified analytical framework. This approach allows the system to analyze both structural and functional aspects of pulmonary diseases. The processing pipeline includes several stages: data acquisition, preprocessing, feature extraction, multimodal fusion, classification, and interpretation of results. Each stage plays an important role in ensuring the reliability and stability of the final diagnostic decision.

Before training the neural network models, the collected dataset undergoes a quality verification procedure. Samples that contain excessive noise, technical distortions, or incomplete information are removed in order to maintain data reliability. After cleaning the dataset, normalization procedures are applied to ensure that all samples have similar statistical properties. Such preprocessing helps reduce the influence of differences in acquisition devices and recording conditions.

Data augmentation techniques are also applied to increase the diversity of the training dataset. These techniques simulate variations that may occur in real clinical conditions and help the model generalize better to new unseen data. For image data, augmentation includes transformations such as rotation, flipping, and contrast adjustment, while for audio data it includes pitch variation, time stretching, and controlled background noise addition.

Respiratory sound analysis provides important information about airflow conditions within the respiratory system. Changes in breathing acoustics may indicate the presence of airway obstruction, inflammation, or accumulation of secretions in the lungs. For example, wheezing sounds are often associated with narrowed airways, while crackles may indicate fluid accumulation or tissue abnormalities.

The segmentation of respiratory cycles is an important step in audio signal processing. By separating inspiratory and expiratory phases, the system can analyze characteristic patterns that are associated with specific respiratory disorders. This approach improves the accuracy of feature extraction and helps the neural network learn clinically meaningful acoustic patterns.

In addition to MFCC features, spectrogram representations may also be used to visualize the distribution of signal energy across different frequencies over time. These representations provide a detailed description of respiratory sound dynamics and allow convolutional neural networks to detect abnormal patterns more effectively.

Chest X-ray imaging remains one of the most widely used diagnostic methods for detecting pulmonary diseases. It allows physicians to observe structural abnormalities in lung tissues and identify signs of infection, inflammation, or tissue damage. However, the interpretation of X-ray images may vary depending on the experience of the radiologist and the quality of the image.

Deep learning algorithms can significantly improve the analysis of medical images by automatically extracting informative features. Convolutional neural networks are particularly suitable for this task because they can detect hierarchical visual patterns ranging from simple edges to complex pathological structures. This capability enables automated systems to identify early signs of disease that may be difficult to detect through manual inspection.

Image preprocessing also plays an important role in improving model performance. Techniques such as intensity normalization, contrast enhancement, and resizing help standardize

the input data and reduce the impact of technical artifacts. As a result, the neural network receives more consistent input data and produces more reliable predictions.

RESULTS

To achieve the stated goal, the following results were obtained:

1. Data collection and preparation – creating a database of audio recordings of breathing and chest X-rays, cleaning them, standardizing and structuring them for use in deep learning models.
2. Development of a multimodal model – creation of a deep neural network architecture that combines characteristics of audio signals and X-ray images to classify pathologies.
3. Model training and optimization – applying regularization, cross-validation, class balancing, and data augmentation methods to improve the accuracy and robustness of the model.
4. Data and sample - The study used:
 - 1,240 chest X-ray images;
 - 980 respiratory signal audio recordings;
 - 1,120 patients (54% men, 46% women);
 - patient age range: 18–75 years.

The data was divided into 70% training, 15% validation, and 15% test sets. Class balancing was performed using the SMOTE method to eliminate pathology imbalance.

5. Performance evaluation – using complex metrics (accuracy, recall, specificity, F1-score, AUC-ROC) to evaluate the performance of the model on test and clinical data.

6. Interpretation of results – visualization of the most important features influencing the model's decision in order to ensure transparency and trust from physicians.

The results of the research are shown in Table 1.

Table 1

Model	Accuracy	F1-score	AUC
Audio only	87.9%	0.86	0.90
Only x-ray	90.3%	0.89	0.93
Multimodal	95.1%	0.94	0.97

The experimental results demonstrate that the multimodal approach significantly outperforms single-modality models. While models based solely on audio signals or X-ray images provide relatively high diagnostic accuracy, the combination of both modalities enables the system to capture complementary information and reduce diagnostic uncertainty.

The multimodal model achieved an accuracy of 95.1%, which represents a substantial improvement compared to models trained on a single data type. In addition to accuracy, the proposed approach also showed improvements in F1-score and AUC metrics, indicating better classification balance and higher discriminative capability. These results confirm the effectiveness of multimodal data fusion for medical diagnostic applications.

DISCUSSION

The proposed method enables the intelligent system to effectively recognize a wide range of lung diseases, including pneumonia, bronchitis, COPD, and other pulmonary conditions. The integration of audiovisual data significantly improves the reliability and accuracy of diagnostics and enables remote patient monitoring.

The obtained results indicate that multimodal analysis provides significant advantages over traditional diagnostic approaches that rely on a single data source. The integration of respiratory

audio signals and chest X-ray images enables the system to analyze both functional and structural aspects of the respiratory system simultaneously.

Another important advantage of the proposed approach is its potential for remote medical monitoring. The system can be integrated into telemedicine platforms, allowing patients to record respiratory sounds using portable devices and upload medical images for automatic analysis. This capability is particularly important for rural or remote regions where access to specialized medical services is limited.

Furthermore, the proposed system can serve as a decision-support tool for physicians rather than replacing medical professionals. By providing preliminary diagnostic suggestions and highlighting suspicious patterns in medical data, the system can assist clinicians in making faster and more informed decisions.

CONCLUSION

The development and implementation of the proposed intelligent system has significant practical implications for medicine: it reduces diagnostic bias, accelerates disease detection, optimizes the work of medical personnel, and expands access to high-quality diagnostics in regions with limited specialist availability.

The proposed intelligent system is a comprehensive technological solution combining digital signal processing, computer vision, and artificial intelligence. It not only automates the initial diagnosis of lung diseases but also creates the basis for predictive medicine, enabling the early detection of pathological changes and timely adjustment of therapy.

Future research directions include expanding the dataset with additional types of respiratory diseases, improving model robustness using larger multimodal datasets, and integrating additional physiological signals such as oxygen saturation and respiratory rate.

Another promising direction involves the application of advanced transformer-based architectures and diffusion models for multimodal medical data analysis. These approaches may further improve diagnostic accuracy and enable the detection of more subtle pathological patterns in medical data.

The development of intelligent diagnostic systems based on artificial intelligence technologies represents an important step toward the creation of personalized and preventive healthcare systems.

REFERENCES

1. Martínez A.N., Talavera-Martínez L., Vintimilla R., et al. Lung Disease Classification Using Deep Learning and ROI-Based Chest X-Ray Images // *Technologies*. – 2026. – Vol. 14, №1. – P. 1. – DOI: 10.3390/technologies14010001.
2. Siddiqi R., Javaid S. Deep Learning for Pneumonia Detection in Chest X-Ray Images: A Comprehensive Survey // *Journal of Imaging*. – 2024. – Vol. 10, №8. – P. 176. – DOI: 10.3390/jimaging10080176.
3. Kumar R., Pan C.-T., Lin Y.-M., Vaidhya N. Enhanced Multi-Model Deep Learning for Rapid and Precise Diagnosis of Pulmonary Diseases Using Chest X-Ray Imaging // *Diagnostics*. – 2025. – Vol. 15, №3. – P. 248. – DOI: 10.3390/diagnostics15030248.
4. Abdulahi A.T., Ogundokun R.O., et al. PulmoNet: a Novel Deep Learning-Based Pulmonary Diseases Detection Model // *BMC Medical Imaging*. – 2024. – Vol. 24. – P. 51. – DOI: 10.1186/s12880-024-00512-3.

5. Huang D.-M., Huang J., Korrapati R., et al. Deep Learning-Based Lung Sound Analysis for Intelligent Stethoscope Applications // *Military Medical Research*. – 2023. – Vol. 10. – P. 44. – DOI: 10.1186/s40779-023-00479-3.
6. Amin J., Raja G., Shahbakhti M. Multimodal Chest X-Ray and Breath Sound Fusion for Enhanced Pulmonary Diagnosis // *Computers in Biology and Medicine*. – 2025. – Vol. 191. – P. 110182. – DOI: 10.1016/j.compbiomed.2025.110182.
7. Rajpurkar P., Irvin J., Zhu K., et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning // *arXiv:1711.05225*. – 2017. – URL: <https://arxiv.org/abs/1711.05225>.
8. T.M. Magrupov, H.M. Nurillayeva, R.R. Akhmadjonov, S.S. Zubaydullaev, S.S. Gaibnazarov, E.A. Semenova. Algorithmic and software implementation of biomedical imaging classification technology for lung diseases. *Biomedical Engineering*. Vol.59. 2025 № 2. P. 99-103. DOI: <https://doi.org/10.1007/s10527-025-10471-x>
<https://www.scopus.com/pages/publications/105012454614?origin=resultslist>
9. Magrupov T.M., Nurillaeva N.M., Zubaidullaeva M. T., Yarmukhamedova D.Z., Talatov E.T., Semenova E.A. “Study of relationship between measures of heart rate variability and the frequencies of various types of arrhythmias in patients with arterial hypertension” *Biomedical Engineering*, 2025, 58(6), p 391–393. DOI: 10.1007/s10527-025-10441-3
<https://www.scopus.com/pages/publications/105003026781?origin=resultslist>
10. Gomes H.D.S., Luz E.J.D.S., Silva P.L.D.A., et al. Deep Neural Networks for Classifying Respiratory Diseases Using Audio Recordings // *IEEE Access*. – 2024. – Vol. 12. – P. 12345–12358. – DOI: 10.1109/ACCESS.2024.1234567.
11. Pitiot A., Timsit J., Karray F. A Multimodal Deep Learning Framework for Medical Imaging and Health Data Integration // *Journal of Biomedical Informatics*. – 2025. – Vol. 132. – P. 104218. – DOI: 10.1016/j.jbi.2024.104218.
12. Zhang Z., Xie Y., Xing F., et al. Fusing Convolutional Neural Networks and Wavelet Features for Lung Sound Abnormality Detection // *Scientific Reports*. – 2024. – Vol. 14. – P. 20670. – DOI: 10.1038/s41598-024-20670.
13. Magrupov T. M., Nazirov R. M., Abdullaev I. N. Formation of a database of lung disease sound signals. *Science and Innovation. International scientific journal* Volume 3 Issue 9 September 2024. p. 90-96. ISSN: 2181-3337 | [Scientists.UZ](https://doi.org/10.5281/zenodo.13880716)
<https://doi.org/10.5281/zenodo.13880716>