Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 21, MARCH 2025

DOI: 10.37943/21WPWL2968

Aigerim Aitim
Master of Technical Sciences, Assistant Professor of the Department of Information Systems
a.aitim@iitu.edu.kz, orcid.org/0000-0003-2982-214X
International Information Technology University, Kazakhstan

Zhamilya Abdildanova
Bachelor's student, Department of Information Systems
selena.jamilya@gmail.com, orcid.org/0009-0009-8926-462X
International Information Technology University, Kazakhstan

Symbat Tynystykbayeva
Bachelor's student, Department of Information Systems
tsimka.milka@gmail.com, orcid.org/0009-0003-6843-7575
International Information Technology University, Kazakhstan

Aidana Muratbekova
Bachelor's student, Department of Information Systems
muratbekova1429@gmail.com, orcid.org/0009-0006-9064-3780
International Information Technology University, Kazakhstan

Nurbike Nalhozha
Bachelor's student, Department of Information Systems
nnalkozha006@mail.ru, orcid.org/0009-0007-1826-5209
International Information Technology University, Kazakhstan

ARTIFICIAL INTELLIGENCE-ENHANCED MOBILE DIAGNOSTICS USING DECISION TREES FOR EARLY DETECTION OF RESPIRATORY DISEASES

Abstract: This article is devoted to the identification of early diagnosis of respiratory lung diseases, such as chronic obstructive pulmonary disease and pneumonia, to reduce mortality and prevent complications. One of the most effective methods of structuring data is the Decision Tree model. The research focuses on the development and evaluation of a decision tree model, which is used to obtain data in the form of questionnaires, text files from patients, where they describe in detail the entire process of the disease, describing their symptoms and general condition at different time periods. There are a few criteria that patients must answer for a more accurate diagnosis. The developed methodology will allow processing relevant data with various symptoms to obtain reliable identification of the signs of the disease, as well as the stages of its progression; this can be done without the use of complex and high-tech devices that make diagnosis very accessible and feasible in the shortest possible time, if resources and time are limited. The article describes the model, carefully collected, and processed the necessary data, and then the results will be described in detail, covering many indicators such as accuracy, responsiveness, F1 score and ROC-AUC. The results of this analysis strongly suggest that this model is effective enough to provide a high level of accuracy combined with extensive capabilities, which determines its practical importance for use in real conditions. It is noted that the decision tree model can significantly improve the quality of diagnostics, since it is possible to structure a large amount of data and thus collect many years of human experience.

Received: 22.11.2024 Accepted: 24.03.2025 Published: 30.03.2025

DOI: 10.37943/21WPWL2968

Keywords: artificial intelligence; respiratory diseases; medical diagnostics; Decision Tree; machine learning; early diagnosis; healthcare; mobile-based diagnostics

Introduction

Artificial intelligence technologies are playing an increasingly important role in medical diagnostics, and one of the promising areas of their application is the diagnosis of respiratory diseases. Respiratory diseases such as pneumonia and chronic obstructive pulmonary disease continue to be a serious public health problem and one of the leading causes of death worldwide [1]. Scientists are developing new drugs and vaccines against these diseases. However, in most cases, therapeutic intervention is initiated when the disease has progressed to advanced stages, delaying effective treatment [2]. According to WHO, about 2.5 million people die from pneumonia alone every year. Early diagnosis can significantly reduce mortality, but, unfortunately, traditional diagnostic methods require specialized equipment, which is often unavailable in conditions of limited medical resources [3].

Recent years have seen a considerable focus on the role of artificial intelligence (AI) in medical diagnostics, not least for challenges associated with respiratory care. Recent literature has emphasized the promise AI holds for improved diagnostic accuracy, optimization of clinical workflows, and increased access to health care services [4]. Explainable AI techniques demonstrate transparency during decision-making, which is essential for clinical acceptance [5]. The COVID-19 pandemic has proven that traditional diagnostic methods are inadequate for such urgent needs to detect diseases as early as possible [6]. In the early stages of the pandemic, many people underestimated their symptoms, leading to overwhelmed hospitals and severe complications in untreated patients. This demonstrated the critical need for early detection and diagnosis. Under these conditions, artificial intelligence technologies have opened new opportunities for more accessible and faster diagnostics.

Literature review

Yoo et al. created a decision-tree chest X-ray COVID-19 diagnostic tool. Pandemic diagnosis calls for artificial intelligence as their approach achieved 95% accuracy [7]. While Bian et al. demonstrated the increase of low-dose CT diagnosis, Shen and Liu demonstrated the benefit of radiography and respiratory pattern analysis. Early screening is improved by means of multifactorial analysis in methods [8], [9]. Ohno et al. demonstrated how machine learning enhances interobserver consistency of lung CT texture analysis. This approach lowers error [10] and advances detection of COPD and interstitial lung disease.

More widely available and less expensive are survey-driven, NLP-based diagnoses. Feinstein and colleagues validated a gradient boosting COPD screening model based on a questionnaire. In population studies, data-driven approaches with high case identification sensitivity and a reasonably affordable substitute for spirometry were successful in resource-limited settings [11]. For early illness diagnosis, Dreisbach et al. demonstrated how NLP may identify symptom patterns from patient-authored texts [12]. Georgakopoulou claims that wearable gadgets and artificial intelligence can identify early respiratory tract infections signs. She studies safe and ethical artificial intelligence incorporation into medical systems [13].

These technologies show the mHealth AI transformational power. Still, massive, annotated datasets, algorithm openness, data privacy remain concerns. Al-Anazi et al. recommended ethical oversight for data privacy [14] as well as algorithmic fairness in artificial intelligence applications. Isangula and Haule suggest localized, user-friendly designs that go beyond pragmatic restrictions of tool distribution in low-resource environments. Hussein and Kim claim that imaginative design and robust preprocessing will help to overcome these challenges. For precise feature picking, Sangwan recommended random forests and regression models [15].

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 21, MARCH 2025

Machine Learning Made Possible Early Lung Cancer Biomarkers Discovery Xie & colleagues. Six metabolites used collectively improved early identification with 98% sensitivity and 100% specificity [16].

The Decision Tree was chosen as one of the methods of classification because it effectively works with little and medium data and provides high interpretability of the model decisions, which is important in medical applications. Machine learning is applied in many fields of activity and early diagnosis of various diseases. However, most models are based on the analysis of images, such as lung X-rays, or physiological data, such as spirometry readings [17]. Works related to textual data analysis are relatively rare, although text questionnaires are commonly used for collecting information about a patient's condition. Early research suggests textual questionnaire data may be useful for diagnosis, but its accuracy largely depends on the model and quality of the data. The results of this work present a new approach to diagnosing respiratory diseases using text data, which is suitable for conditions with limited access to medical technologies [18].

This model provides a simplified version of a decision tree based on text questionnaires, making diagnosis easier with text data already available, wherein patients describe their symptoms, thus making diagnosis more accessible to a wide range of users. The result of this work is to develop a model capable of early diagnosis of respiratory system ailments, using only textual data obtained from questionnaires in which patients describe their symptoms and general condition in detail.

The novelty of this study lies in developing a model focused on textual information analysis and testing it on real data obtained from patient questionnaires. This allows researchers to assess the efficiency of the Decision Tree in diagnosing respiratory diseases [19]. Text data provided by patients undergo preprocessing steps such as tokenization, stop word removal, and vectorization to convert it into an analyzable form [20].

Methods and Materials

The dataset includes 870 patient questionnaires collected from medical facilities and online health platforms. The age of patients ranges from 18 to 75 years, with a balanced gender distribution (52% female, 48% male). Each questionnaire contains an average of 14-18 symptoms described in textual form, resulting in an average text length of 120-150 words per response. Accurate data is necessary for this topic due to its relevance. In order to develop an accurate respiratory disease diagnostic model, a comprehensive data set has been gathered and prepared. The dataset comprises text questionnaires completed by patients who have consented to provide a description of their symptoms. The questionnaires provide a comprehensive overview of the progression of diseases at various phases, as they include detailed descriptions of symptoms and general health. The participants provided a Comprehensive account of their symptoms and their emotional responses. All descriptions of the patients' well-being at various phases of the disease were documented, providing us with a more comprehensive understanding of the disease and the emotions that an individual experiences during each stage. This data was employed to develop a model that is capable of predicting a diagnosis based on the symptoms inputted by the user. The entire data collection and processing procedure was conducted in compliance with ethical and privacy standards.

In the present tense, the patient provided a detailed account of his symptoms and well-being using questionnaires in the study. A questionnaire is completed by each patient, which includes both textual and structured data, including their age and gender. These data are essential for a more comprehensive examination of the disease's progression across various age groups and genders. In order to establish a training sample, all these questionnaires are collected and transmitted after undergoing specific phases. The diagnostic system analyzes

DOI: 10.37943/21WPWL2968

the symptoms and provides the result, as illustrated in figure 1. The user inputs his symptoms, which are subsequently evaluated by the profile and subsequently stored in the database. The data is subsequently transmitted through the data analysis system and subsequently analyzed using a decision tree. The output is then saved and sent to the user after the processing. This diagram shows the flow and elements:

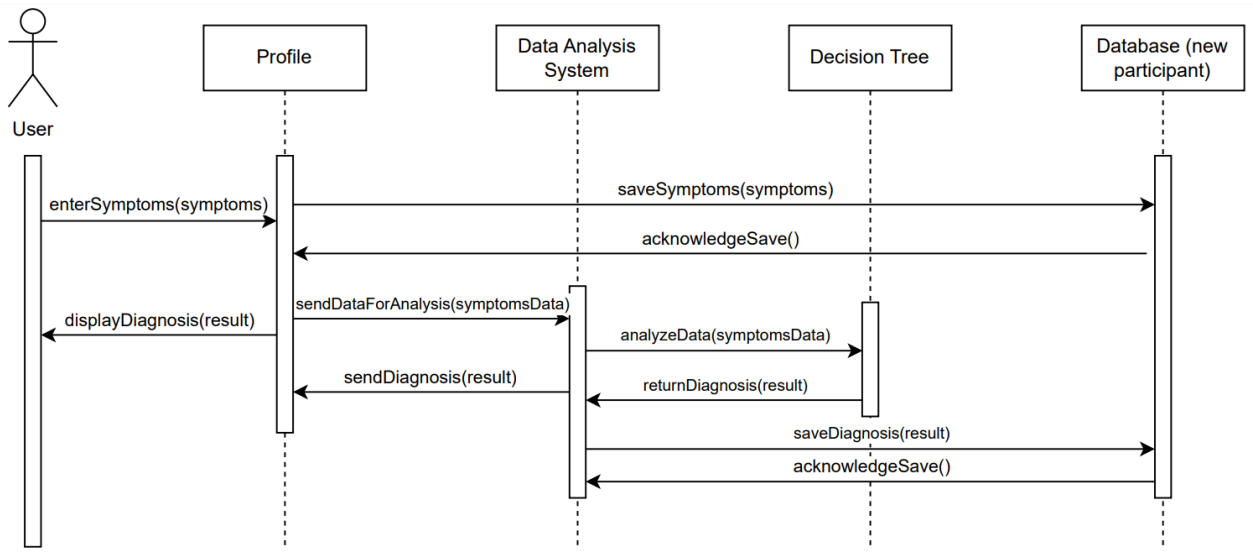

Figure 1. Diagram of the sequence of user interaction with the system

The diagram shows the stepwise interaction between parts of the system to diagnose user symptoms based on data entered. Integration with the data analysis system and database ensures reliable storage and analysis, which in turn creates an efficient system.

The questions should be clear and comprehensible, and the surveys should be structured and unambiguously distinguishable to allow the patient to provide an in-depth explanation of their emotions. The training and test samples were subsequently transmitted to the preprocessing and analysis stage after the surveys were structured and systematized. These characteristics are indispensable for the creation of comprehensive training and testing datasets, which are subsequently employed in the development and analysis phases of models, including Decision Tree modeling. The structure and content of these attributes are illustrated in the following detailed table:

Table 1. Structured Questionnaire Attributes for Respiratory Disease Diagnosis

№	Attribute Category	Sub-Category / Details
1	Basic Information	Application Form ID, Patient ID, Age, Gender
2	Symptoms and Characteristics	Main Symptoms Description, Duration, Intensity, Triggers, Frequency of Manifestations
3	Medical History	Chronic Diseases, Allergies, Smoking History, Frequency of Doctor Visits (Respiratory)
4	Vital Factors	Physical Activity Level, Working Conditions, Environmental Factors
5	Physiological Parameters	Body Temperature, Saturation Level, Respiratory Rate, Pulse
6	Assessment of Subjective Factors	Stress and Anxiety Levels, Subjective Condition Description, Quality of Life Assessment
7	Family History	Respiratory Diseases in Close Relatives (Predisposition)

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 21, MARCH 2025

A patient's respiratory health is fully assessed using standardized questionnaires, as shown in the table above. Each category collects essential data for diagnosis and prediction. Basic demographic and identity information is provided, while symptoms are detailed to show their impact. Medical history examines diseases that may cause or worsen respiratory issues. Vital variables evaluate lifestyle and environmental factors. Physiological measures provide quantitative health markers, while subjective evaluations capture the patient's emotional and stress levels. Finally, family history checks for respiratory illness genetics. This organized data provides a reliable training sample for prediction models like the Decision Tree, which analyzes and diagnoses respiratory diseases. This dataset comprises several factors used to better study and diagnose respiratory disorders. The patient profile, including age, gender, and medical history; symptoms, intensity, triggers, and frequency; external factors, including environmental conditions and work environment; and other parameters, like temperature, stress level, and subjective assessment, make up the questionnaire. A Decision Tree model training sample will be based on this data. A data structure with classes for the patient's profile, symptoms, the Decision Tree model, and the diagnosis is created using questionnaires and open-ended questions that require detailed data and will be repeated depending on the patient's illness. Next, pre-process data to train the model. This allows the machine learning model to use questionnaire text data efficiently. Data must be numeric for preprocessing. Data preprocessing organizes and prepares data for analysis. Without this stage, text input will be excessively varied and hard to handle, affecting Decision Tree model accuracy.

Successful data preparation necessitates the execution of the following steps:

1. Tokenization. The initial stage is tokenizing the data. This technique disaggregates text into discrete words or, more accurately, tokens. This facilitates the disaggregation of lengthy textual descriptions into their constituent elements, which may then be evaluated and processed independently. Tokens facilitate the identification of distinct symptom indicators for further study. This should function as dividing an array of words into distinct components. For instance, the phrase "I have a strong cough and chest pain" can be transformed into the tokens ["strong", "cough", "pain", "chest"].

```
import nltk
nltk.download('punkt')
text = " I have a bad cough and chest pain "
tokens = nltk.word_tokenize(text)
print(tokens)
```

Figure 2. Tokenization Process Using NLTK

2. Eliminate stop words. The subsequent phase in preprocessing involves identifying and removing stop words. Stop words are terms that lack significant informational value for analysis (e.g., "and," "in," "on"). Eliminating such terms decreases data volume and enhances model efficiency. Stop words like "and" and "in" will be eliminated from the text "I have a severe cough and chest pain", resulting in "severe cough chest pain".

```
stop_words = set(nltk.corpus.stopwords.words('english'))
tokens = [word for word in tokens if word not in stop_words]
print(tokens)
```

Figure 3. Stop Words Removal Using NLTK

3. Lemmatization and stemming reduce words to their fundamental form (for instance, "cough" remains "cough"). This streamlines data analysis and diminishes the variety of unique

DOI: 10.37943/21WPWL2968

terms, hence enhancing processing efficiency. For instance, "coughing," "coughing," and "coughing" will be consolidated to "coughing."

```
from nltk.stem import WordNetLemmatizer
    lemmatizer = WordNetLemmatizer()
    lemmatized_tokens = [lemmatizer.lemmatize(token) for token in tokens]
    print(lemmatized_tokens)
```

Figure 4. Lemmatization Process with WordNet Lemmatizer

4. Vectorization transforms textual input into a numerical representation suitable for model analysis. This study used the TF-IDF (Term Frequency-Inverse Document Frequency) technique, which allocates a weight to each word according to its frequency within the text. For instance, post-vectorization, the phrase "severe cough" can be denoted as a vector [0.3, 0.5, 0.9 ...], with the values indicating the importance of the phrases.

```
from sklearn.feature_extraction.text import TfidfVectorizer
    vectorizer = TfidfVectorizer()
    vector = vectorizer.fit_transform(["severe cough chest pain "])
    print(vector.toarray())
```

Figure 5. Text Vectorization Using TF-IDF.

The Decision Tree model was used to categorize symptom data because it can analyze sign hierarchies, which is crucial for illness diagnosis. The Decision Tree evaluates one sign at each level (cough, patient age). The model can readily identify diagnostically important symptoms. The Decision Tree makes it easier to comprehend data and analyze crucial signals since it shows which symptoms and phases had the most influence on diagnosis. The Decision Tree model works well with small and medium-sized samples, making it suitable for our questionnaire dataset. Decision Tree uses fewer resources than neural networks, which is useful for mobile apps to conserve computing resources. Initialize Decision Tree models to generate them. Entropy is used to initialize the model and choose features with maximal information for each node. This strategy helps the model locate the optimal separations, improving diagnostic accuracy. In decision trees, entropy measures how well a feature classifies data. Entropy helps choose a Decision Tree model by identifying the most informative characteristics for categorization. Let it be explored why entropy is the best criteria and how it helps the model make better decisions at each level. Entropy is employed in decision trees to select the optimal feature for classifying data. Entropy measures data uncertainty or "chaos" and helps choose a feature to segregate data at the decision tree node. Calculating entropy:

$$Entropy = -\sum_{i=1}^{n} p_i log_2(p_i) \quad (1)$$

where p_i is the probability of each class within the node and n represents the total number of classes. A lower entropy value indicates greater homogeneity within the group post-separation, signifying a "cleaner" node. The model calculates the entropy of each node according to the class distribution to assess the extent to which the feature differentiates data within the node. It subsequently identifies the feature with the lowest entropy, which optimally divides the data, and employs it on the subsequent node. The model selects a feature that minimizes entropy, hence optimizing node cleanliness for the most accurate categorization possible. This entropy-based criteria enables the model to identify the most relevant characteristics, hence decreasing uncertainty and enhancing classification accuracy at each stage of decision tree construction.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 21, MARCH 2025

During the training of a Decision Tree, it identifies the appropriate features for partitioning according to the entropy criteria. The model requires high-quality data with effective characteristics that distinctly differentiate the classes. Therefore, for the entropy to approach zero, each node must mostly include a single class. The Decision Tree model will autonomously discover splits; however, high-quality, and pertinent characteristics in the data will enhance the tree's ability to accurately discern attributes that will generate nodes with low entropy.

The Decision Tree identifies the feature with the minimal entropy for each split, as this yields the most useful data segmentation. Consequently, the model's accuracy improves as it consistently selects the feature that most effectively diminishes uncertainty and categorizes the data into distinct groups. The Decision Tree model is trained on preprocessed questionnaire data with a training dataset. At each level, the model identifies the indicators that most effectively categorize the symptoms and constructs a tree that partitions the data according to the likelihood of association with certain illness categories. The learning and evaluation of the Decision Tree model involve sequential data processing and the construction of a decision hierarchy, wherein the model identifies the most pertinent attributes for data differentiation at each level. Examine this procedure in further detail:

Table 2. Process of Learning and Testing the Decision Tree Model

№	Phase	Description
1	Model Training	Uses preprocessed data with features (e.g., age, symptoms). Selects features based on criteria like entropy or Gini index for optimal data splits. Creates nodes to separate data, continuing until stopping criteria (e.g., depth, node purity) are met.
2	Classification & Tree Construction	Constructs a hierarchical structure with nodes representing features. Traverses top-down, evaluating features to determine final classifications.
3	Model Testing	Tests on a separate dataset to assess generalization. Traverses' decision tree to make predictions for each test case. Evaluates performance using metrics (Accuracy, Recall, F1-score).

The table summarizes the key phases of the Decision Tree model's learning and testing process. During Model Training, preprocessed data is used to select features based on criteria like entropy, creating nodes to optimally separate data. Classification and Tree Construction results in a hierarchical structure, with nodes representing features and leaves providing final predictions. In the Model Testing phase, the model's performance is evaluated on new data to assess its predictive accuracy. Traversing the tree ensures feature-based predictions, and performance metrics like Accuracy, Recall, and F1-score gauge its real-world applicability. This structured approach enhances transparency and interpretability in diagnosis.

To illustrate how the Decision Tree model functions, consider a dataset with features such as "cough," "chest pain," and "fever." When presented with new input data, the model begins at the root node – evaluating whether a cough is present. Based on the presence or absence of this symptom, the model moves down the appropriate branch to the next decision point, such as checking for chest pain. This process continues until the model reaches a final diagnosis, such as "pneumonia" or "viral infection." This stepwise traversal and evaluation of features ensure a structured and interpretable approach to medical diagnosis.

The stages of data preprocessing and training of the Decision Tree model are essential for creating an accurate system for diagnosing respiratory diseases. The preprocessed data ensures high quality and consistency of the input data, and the Decision Tree model makes it possible to effectively use this data for symptom-based diagnosis.

DOI: 10.37943/21WPWL2968

Further, to fully evaluate the method, it is necessary to evaluate in detail the accuracy and performance of the model using metrics such as Accuracy, Recall, F1-score, and ROC-AUC, which will allow us to analyze the results of predictions and evaluate the effectiveness of the proposed diagnostic method.

The Decision Tree model is evaluated using key metrics such as Accuracy, Recall, F1-Score and ROC-AUC. These metrics allow you to assess how well the model recognizes symptoms and predicts respiratory diseases.

The accuracy, which shows the proportion of correct predictions among all predictions of the model. This metric shows how well the model classifies the data, is determined using this formula:

$$Accuracy = \frac{\text{The number of correct predictions}}{\text{Total number of predictions}} = \frac{520}{600} \approx 0.87 \quad (2)$$

An accuracy of 87% shows that the model effectively classifies most of the data.

The next metric is recall, which measures the model's ability to identify positive cases. In our case, the positive case is sick patients. Recall evaluates how well the model identifies real cases of the disease among all positive examples. This metric is especially important for medical diagnostics, as it shows the ability of the model to find cases of diseases:

$$Recall = \frac{\text{Correct positive predictions}}{\text{The total number of truly positive cases}} = \frac{280}{300} \approx 0.93 \quad (3)$$

Using this formula, recall can be found, and in the study, recall is 93 percent, which indicates the high ability of the model to correctly identify sick patients.

The F1-Score is the harmonic mean of accuracy and completeness, which is useful when the balance between these metrics is important. F1-score evaluated by this formula:

$$F1 = 2 * \frac{Accuracy * Recall}{Accuracy + Recall} = 2 * \frac{0.87 * 0.93}{0.87 + 0.93} = 2 * \frac{0.8091}{1.8} = 0.9\ (90\%) \quad (4)$$

An F1-Score of 90% only proves it is balanced between the correct classification of positive and negative cases.

The next parameter that needs to be changed is Receiver Operating Characteristic - Area Under Curve next (ROC-AUC). The ROC-AUC measures how the model could discern between the "sick" and "healthy" classes. The ROC curve – the Receiver Operating Characteristic – is based on the parameters True Positive Rate (TPR) and False Positive Rate (FPR). AUC indicates how good the model is at distinguishing positive cases from negative ones. It can be calculated with this code:

```
from sklearn.metrics import roc_auc_score, roc_curve
import matplotlib.pyplot as plt
y_true = [1, 0, 1, 1, 0, 1, 0, 0, 1, 0]
y_scores = [0.85, 0.1, 0.78, 0.92, 0.2, 0.88, 0.05, 0.3, 0.9, 0.15]
auc = roc_auc_score(y_true, y_scores)
fpr, tpr, thresholds = roc_curve(y_true, y_scores)
plt.figure()
plt.plot(fpr, tpr, color='blue', label=f'ROC curve (AUC = {auc:.2f})')
plt.plot([0, 1], [0, 1], color='gray', linestyle='--', label='Random Guessing')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve')
plt.legend(loc="lower right")
plt.show()
auc
```

Figure 6. ROC-AUC Curve Representation for Model Evaluation

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 21, MARCH 2025

The ROC-AUC equal to 0.96. This indicates that the model has a high ability to distinguish between the "sick" and "healthy" classes, since the ROC-AUC value is close to 1.

Diverse types of errors may arise within the diagnostic system for respiratory disorders, ranging from inaccurate data entry by the patient to errors in predictions generated by the model. These flaws must be addressed appropriately to reduce the probability of erroneous diagnosis, as this is a medical application where inaccuracies may affect patient health. Data entry errors may arise when the patient provides missing or inaccurate information, such as omitting age or inputting symptoms in an inappropriate manner. The system issues a notification for the user to elucidate or amend the data. For example, if the age is input using letters instead of numbers, the system will request a correct re-entry.

Errors may occur during preprocessing when data do not correctly navigate the phases of tokenization, stop word removal, or vectorization. If an error is detected, the system re-evaluates the data and, if required, does further preprocessing. The method further alerts the developer in instances of persistent issues for further examination.

During the model analysis phase, if the data exceeds the model's training or if the model is inadequately trained to comprehend all symptoms, it will assess the confidence level in its predictions at such times. Should the confidence level drop beneath the designated threshold – such as 70% – the forecast is deemed untrustworthy, prompting the algorithm to recommend further evaluation by a professional. Implementing error handling at each level contributes to the development of a resilient diagnostic system, hence augmenting patient safety and refining the model's accuracy. Figure 7 depicts the primary phases of the error handling procedure. Upon data entry, the system verifies their accuracy and, if required, issues a correction request to the user.

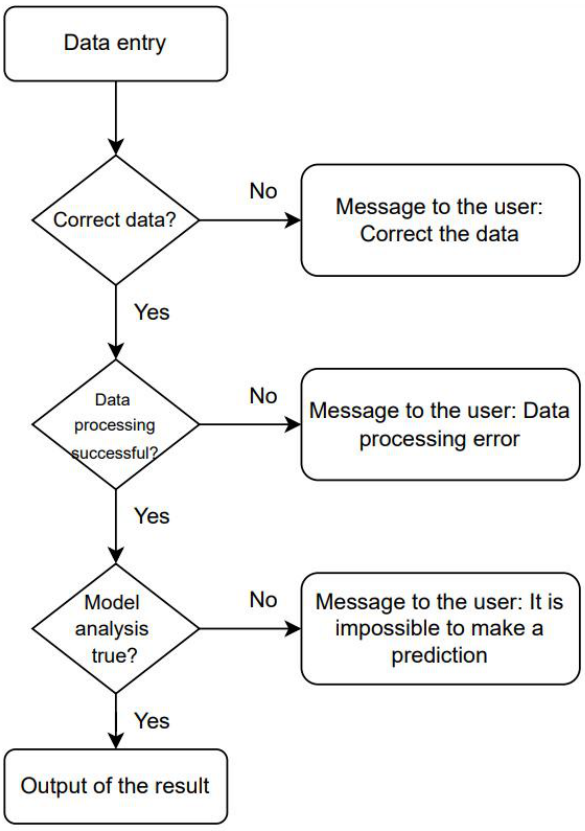

Figure 7. Error handling diagram in the diagnostic system

DOI: 10.37943/21WPWL2968

The error flow diagram shows potential errors at each stage of the process – from data entry to receiving the result, illustrating how the system processes them: sends a message to the user, corrects the data, or performs a recheck. This helps developers anticipate possible failures in advance, optimize error handling, and improve the user experience, making the system more reliable and user-friendly.

High-quality control and monitoring mechanisms are presented to ensure the quality of the respiratory disease diagnostic model. One of the quality controls is the confidence threshold but allows the assessment of the reliability of the predictions made. If the model's confidence in the prediction falls below 70%, the system flags it as unreliable and further suggests reevaluation or retraining of the model.

To keep the model stable, cross-validation is applied; it allows you to test the model on different datasets. This approach allows you to judge the average performance level of your model and how stable it would be on new data. Using, for example, 5x cross validation helps to understand potential deviations in the accuracy and predictive power of the model.

This also includes the quality control of regular assessment of some performance measures like accuracy, recall, and F1-score. In such cases, if any measure falls below a set level, the possible problems with predictions are analyzed for improvement of the model. These safeguards assure the reliability and accuracy of the model over the long term and reduce the likelihood of errors; hence, the predictive power for medical applications is very high.

Results

The evaluation of the developed decision tree model for the diagnosis of respiratory diseases based on the data that the patient entered in the form of text illustrates the high level of effectiveness for various types of key metrics. This section provides an idea of the model's performance and practical significance using metrics such as accuracy, F1-Score, recall and ROC-AUC.

The Recall model has reached a value of 93%, which in turn measures its ability to identify all positive cases of sick patients. A high Recall value is important in medicine because it ensures that the model can identify almost all patients who have respiratory diseases and thereby reduce the risk of missing critical cases. As an example, if a patient shows symptoms that indicate pneumonia, then a high Recall level indicates a high probability of correctly identifying a patient who needs help.

The F1-Score model which represents the average accuracy and recall was 90%. This indicator is important when you need to balance accuracy and recall avoiding false cases. The high values of this model demonstrate the effectiveness of the model in providing balanced predictions, which is critical in medical applications since both types of errors can have huge consequences.

The accuracy model was identified at 87%, which indicates that it correctly identified 87% of all submitted data. This parameter confirms the model's ability to correctly diagnose patients based on the symptoms they have introduced and makes it a useful tool for primary screening and early diagnosis of respiratory diseases in resource-limited settings. The high accuracy of the model minimizes the likelihood of erroneous classification and ensures that patients receive the correct result.

The ROC-AUC model reached a value of 0.96 at launch, which in turn confirms its ability to distinguish between "sick" and "healthy" patients. A high ROC-AUC value, which is close to 1, makes it clear that the model effectively separates positive and negative classes. Figure 4 shows the ROC curve of our model for the diagnosis of respiratory diseases. The result interprets that the model predicts the diseases with high accuracy and practically does not allow

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 21, MARCH 2025

classifying errors. The curve moves towards the upper-left corner of the graph, which indicates high sensitivity True Positive Rate with a minimal level of false positives False Positive Rate.

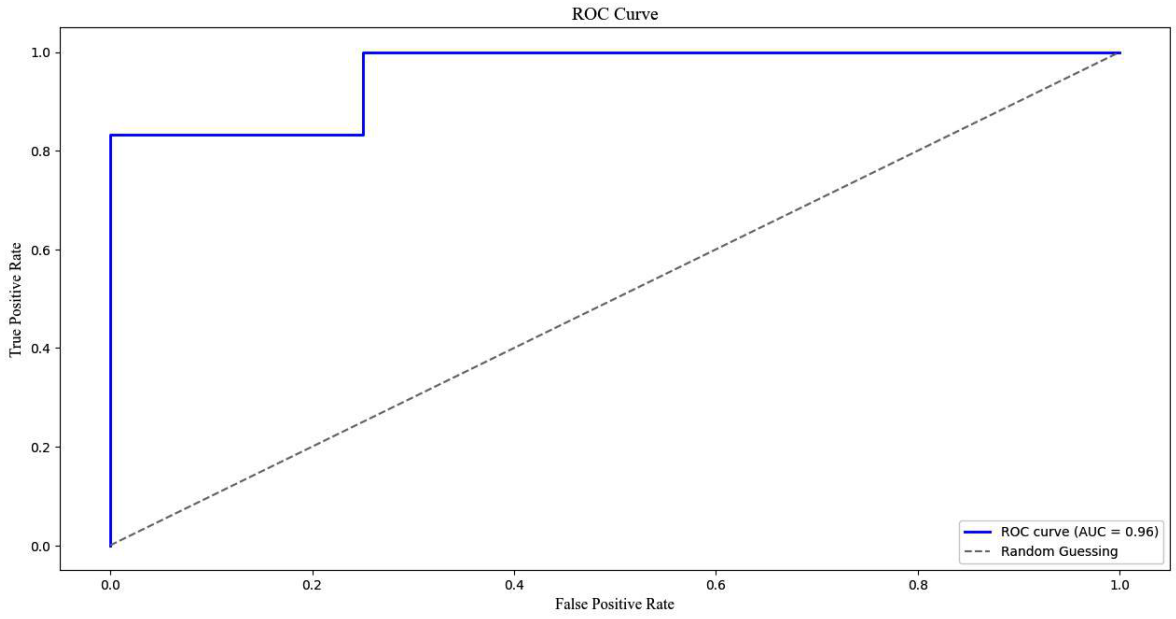

Figure 8. ROC curve for evaluation of the classification model

When analyzing the key signs, it shows that “persistent cough”, “chest pain” and “duration of temperature” are among the most critical symptoms for accurate diagnosis. These signs were recognized by the model because they are primarily associated with respiratory diseases. “Age and gender also matter, but with less weight, which shows their importance in certain cases. Understanding the importance of each feature helps the model analyze and interpret the decision-making process, which in turn increases its reliability.

To get a full picture of how reliable the model is, especially when working with uneven data, this study calculated Matthews Correlation Coefficient (MCC) and Cohen’s Kappa in addition to standard metrics like accuracy, recall, F1-score, and ROC-AUC. These metrics assess both the predictive power and the agreement between the model’s predictions and actual diagnoses. The all results of metrics are shown in the table 3.

Table 3. The results of metrics

№	Metric	Value
1	Accuracy	87%
2	Recall	93%
3	F1-score	90%
4	ROC-AUC	0.96
5	MCC	0.74
6	Cohen’s Kappa	0.76

To give the proposed Decision Tree model a full evaluation, other performance measures were found, such as Cohen’s Kappa and Matthews Correlation Coefficient (MCC). These metrics

DOI: 10.37943/21WPWL2968

tell us a lot about how well the model can deal with uneven data and how well the predictions match up with the real labels. The numbers we got – MCC = 0.74 and Cohen's Kappa = 0.76 – show that the model is still very good at making predictions.

During training and assessment, a 5-fold cross-validation process was used to guarantee the suggested model's robustness and generalizability. Using this method, the dataset is divided into five equal parts. The model is trained using four of the components, and testing is done with the fifth. Five times, this procedure is carried out. The average performance over all folds can be used to accurately measure the model's effectiveness.

To objectively assess the performance of the proposed Decision Tree model, a comparison experiment was conducted utilizing many other machine learning methods, including Random Forest, gradient boosting (XGBoost), and a multilayer neural network (MLP). All models were trained and evaluated on the same dataset including text surveys of patients with symptoms of respiratory illnesses. This technique assures the validity and comparability of the outcomes. The table 4 shows the parameters of the models that were used for training.

Table 4. Parameters of the models

№	Model	Main parameters
1	Decision Tree	Entropy criterion, Max depth = 10, Min Samples Split = 2
2	Random Forest	100 trees, Max depth = 10, Criterion = "gini"
3	XGBoost	Max depth = 6, Learning rate = 0.1, Subsample = 0.8, Number of Estimators = 100
4	MLP	2 hidden layers of 128 neurons each, Activation = ReLU, Output Activation = Softmax, Optimizer = Adam, Learning Rate = 0.001, Batch Size = 32, Epochs = 50

All machine learning models were trained and evaluated using the same preprocessed dataset of 870 patient questionnaires to provide an impartial and equitable comparison. To assess each model's stability, five-fold cross-validation was performed. The Decision Tree model was set up with entropy as the splitting criterion and a maximum depth of 10. For Random Forest, 100 trees with a maximum depth of 10 were used. A learning rate of 0.1 and a maximum depth of 6 were used for training XGBoost. Finally, the neural network (MLP), which contains two hidden layers with 128 neurons each, uses the ReLU activation function. To balance accuracy and computing efficiency and ensure practical application in mobile contexts, these values were chosen based on early testing. The comparison results are presented in Table 5.

Table 5. The results of the comparative analysis

№	Model	Accuracy	Recall	F1-score	ROC-AUC	MCC	Cohen's Kappa
1	Decision Tree	87%	93%	90%	0.96	0.74	0.76
2	Random Forest	90%	91%	90%	0.97	0.79	0.80
3	XGBoost	89%	92%	91%	0.97	0.77	0.78
4	MLP	85%	89%	86%	0.95	0.70	0.72

The results of the tests on the models are shown in table 5 for both conventional and advanced metrics. The Matthews Correlation Coefficient (MCC) and Cohen's Kappa were calculated, along with the standard accuracy, recall, F1-score, and ROC-AUC, to see how balanced and reliable the predictions were. This was especially important when there was a chance of class

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 21, MARCH 2025

imbalance. The findings indicate that when compared to a single decision tree, the ensemble models (Random Forest and XGBoost) exhibit somewhat higher accuracy and F1-measure. The main benefit of Decision Trees, however, is their high interpretability and simplicity of use in mobile applications. This is particularly crucial when creating diagnostic systems for patients and medical professionals that lack specialized knowledge.

The achieved results and their visualization confirm the effectiveness of the Decision Tree model in the diagnosis of colds based on the concept of textual information from patient questionnaires. The low ROC-AUC value and reliance on clinically important features make it a practical tool for real-world supplements, especially in resource-limited settings where traditional diagnostic methods may be less inaccessible. Further self-improvement of feature selection and data preprocessing can reduce the predictive power of the model and guarantee even greater diagnostic reliability.

In summary, obtained results show high accuracy and stability of Decision Tree model in the aspect of respiratory disease diagnosis from text data. Its strong performance on key metrics and focus on clinically relevant indicators lend themselves to real-world implementation, particularly where resources are scarce. Expanded refinement of the model with respect to features selection and data preprocessing will maximize its predictive performance and give more reliable diagnostics.

Discussion

The model, which is based on the decision tree technique for the diagnosis of respiratory disorders using textual data of symptoms presented by patients, produced impressive results, stressing the technological potential and limits of this approach. This section highlights the key findings, their practical implications, and future research goals.

The primary purpose of the project was to develop a model capable of properly diagnosing respiratory disorders based on symptoms gathered from patient surveys. The model had a high recall rate of 93%, indicating its ability to identify nearly all positive cases. This is especially important in the medical field, where early detection of diseases can save lives. The accuracy model was 87%, and the F1 measure was 90%; these signs show that the model operates in a balanced manner and minimizes false positive and false negative findings. The ROC-AUC value of 0.96 verifies the model's capacity to differentiate between ill and healthy individuals and increases its chances for future usage in clinical practice.

The suggested method is a less expensive and more effective alternative to standard diagnostic procedures that need specialized equipment and skilled staff. The use of patient questionnaires to aid in early diagnosis, particularly in constrained contexts such as distant places or underserved regions. Patient involvement can improve compliance with treatment suggestions, allowing patients to become active participants in the diagnostic process.

The comparative analysis demonstrates that while imaging-based approaches offer slightly higher accuracy, survey-driven and text-based methods provide a more accessible, cost-effective, and patient-friendly solution, particularly for early screening in resource-limited settings. The proposed decision tree model fits into this niche, offering a balance between diagnostic accuracy (87%), recall (93%), and operational simplicity, ensuring its applicability in real-world, mobile-based healthcare scenarios.

Keep in mind that the model has major limitations. One of the primary issues was the reliance on the quality of the supplied data. The model's success is directly dependent on the clarity and completeness of patients' replies, which may contain errors, language disparities, medical literacy competence, and a desire to adequately disclose symptoms. Tokenization, stop word elimination, and lemmatization are examples of preprocessing processes that increase complexity when working with diverse datasets.

DOI: 10.37943/21WPWL2968

The incorporation of multimodal data, such as physiological measures, appears to be a viable approach for model development. Combining the supplied text data with objective significant indicators will offer a more complete picture of the patient's state, increasing clinical value. The use of real-time data processing and the ability to monitor the diagnosis based on changes in symptoms increase the practicality. Furthermore, tailoring suggestions based on medical history and past findings can increase therapy accuracy and efficacy.

Because they work with patients' sensory data, confidential and ethical elements of otitis media therapy should be considered. Building confidence and adoption of the model among users and healthcare professionals requires ensuring dependable data protection, patient confidentiality assurances, and adherence to ethical norms.

In conclusion, the created model, which is based on decision trees and text data supplied by patients, has tremendous promise for diagnosing respiratory disorders. Despite the limitations connected with data quality and information generalizability, the model's accessibility, openness, and high efficiency make it a desirable instructional tool, particularly in times of limited resources. Further advances, such as the incorporation of multimodal data, data resolution, and enhanced error handling systems, have the potential to significantly enhance its diagnostic capabilities. This study is a big step toward applying AI to democratize medical diagnoses and enhance health systems.

Conclusion

The development of a mobile application for the early diagnosis of lung diseases using a decision tree model, that uses people's text data, is very promising. There has been recent interest in artificial intelligence and machine learning. To this end, it is targeted to develop an intelligence that replicates the cognitive capabilities of human beings while simultaneously having the capability to store data and experiences way larger than that of human beings. Because this text-based approach uses artificial intelligence and machine learning to provide affordable and cost-effective diagnostic solutions.

For accurate diagnosis of patients using only questionnaires, it is necessary to carefully structure the responses to convert them into analyzable data. It also provides diagnosis of respiratory diseases without the use of complex and expensive medical equipment.

The results confirm the effectiveness of the decision tree model, which provides high memorability, accuracy, and ROC-AUC indicators, which proves its ability to distinguish healthy people from sick people solely since textual data. The decision tree provides transparency and interpretability of structures, which further enhances its suitability for medical applications. Because the model can clearly trace and provides an understanding of the entire decision-making process, which is crucial in the context of healthcare.

However, the diagnosis of patients will become more effective when using the decision tree in combination with other methods that use X-rays and sound data for diagnosis. Since the decision tree has limitations. But to further optimize them, it is necessary to eliminate them. The accuracy of the model largely depends on the quality and completeness of the patient data entered, while deviations are possible due to differences in language, descriptions, and medical literacy. Inaccurate or incomplete data may affect the reliability of the model's forecasts. In addition, although the decision tree model efficiently handles small and medium-sized datasets, the ability to generalize more broadly may be limited when applied to different populations or patients with different characteristics.

It is necessary to emphasize the improvements of the model in the future. Future improvements may include data set extensions and the inclusion of multimodal data, such as physiological measurements or imaging, for more accurate diagnosis and coverage of a wider range of data. Efforts to improve data preprocessing techniques, including reliable handling of text

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 21, MARCH 2025

data variability, are needed to improve consistency and reduce errors. The integration of real-time data processing and personalized diagnostic recommendations based on the patient's medical history can further enhance the usefulness and accuracy of the model.

In summary, this study highlights the potential of the artificial intelligence-based decision tree model, offering accessible, accurate and effective diagnostic tools. This decision tree model has demonstrated the possibility of using textual data for early diagnosis and, therefore, has made a significant contribution to the diagnosis of diseases in the field of medicine, with great prospects in areas where all the possibilities of medicine are not available. Further innovations and improvements in this approach open huge prospects for improving the diagnosis of respiratory diseases and improving patient outcomes worldwide.

References

[1] Chronic obstructive pulmonary disease (COPD). Available: https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)

[2] Peng, J., Chen, C., Zhou, M., Xie, X., Zhou, Y., & Luo, C.-H. (2020). A Machine-learning Approach to Forecast Aggravation Risk in Patients with Acute Exacerbation of Chronic Obstructive Pulmonary Disease with Clinical Indicators. *Scientific Reports, 10(1)*, 3118. https://doi.org/10.1038/s41598-020-60042-1

[3] Hroub, N.A., Alsannaa, A.N., Alowaifeer, M., Alfarraj, M., & Okafor, E. (2024). Explainable deep learning diagnostic system for prediction of lung disease from medical images. *Computers in Biology and Medicine, 170,* 108012. https://doi.org/10.1016/j.compbiomed.2024.108012

[4] Choi, Y., Choi, H., Lee, H., Lee, S., & Lee, H. (2022). Lightweight skip connections with efficient feature stacking for respiratory sound classification. *IEEE Access, 10*, 53027–53042. https://doi.org/10.1109/ACCESS.2022.3174678

[5] Gairola, S., Tom, F., Kwatra, N., & Jain, M. (2020). RespireNet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting. *IEEE Transactions on Biomedical Engineering, 67*(12), 3435-3446. https://doi.org/10.1109/EMBC46164.2021.9630091

[6] Stas, T., Lauwers, E., Ides, K., Verhulst, S., Delputte, P., & Steckel, J. (2024). Convolutional neural network for the detection of respiratory crackles. *IEEE Access, 12*, 3472839. https://doi.org/10.1109/ACCESS.2024.3472839

[7] Yoo, S.H., Geng, H., Chiu, T.L., Yu, S.K., Cho, D.C., Heo, J., Choi, M.S., Choi, I.H., Cong, N.V., Min, B. J., & Lee, H. (2020). Deep learning-based decision-tree classifier for COVID-19 diagnosis from chest X-ray imaging. Frontiers in Medicine, 7, 427. https://doi.org/10.3389/fmed.2020.00427

[8] Bian, H., Zhu, S., Zhang, Y., Fei, Q., Peng, X., Jin, Z., Zhou, T., & Zhao, H. (2024). Artificial Intelligence in Chronic Obstructive Pulmonary Disease: Research Status, Trends, and Future Directions – A Bibliometric Analysis from 2009 to 2023. International Journal of Chronic Obstructive Pulmonary Disease, 19, 1849-1864. https://doi.org/10.2147/COPD.S474402

[9] Shen, X., & Liu, H. (2024). Using machine learning for early detection of chronic obstructive pulmonary disease: a narrative review. Respiratory Research, 25(1), 336. https://doi.org/10.1186/s12931-024-02960-6

[10] Ohno, Y., Aoyagi, K., Takenaka, D., Yoshikawa, T., Ikezaki, A., Fujisawa, Y., Murayama, K., Hattori, H., & Toyama, H. (2021). Machine learning for lung CT texture analysis: Improvement of inter-observer agreement for radiological finding classification in patients with pulmonary diseases. European Journal of Radiology, 134, 109410. https://doi.org/10.1016/j.ejrad.2020.109410

[11] Feinstein, L., Wilkerson, J., Salo, P.M., MacNell, N., Bridge, M.F., Fessler, M.B., Thorne, P.S., Mendy, A., Cohn, R.D., Curry, M.D., & Zeldin, D.C. (2020). Validation of Questionnaire-based Case Definitions for Chronic Obstructive Pulmonary Disease. Epidemiology, 31(3), 459-466. https://doi.org/10.1097/EDE.0000000000001176

[12] Dreisbach, C., Koleck, T. A., Bourne, P. E., & Bakken, S. (2019). A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. International Journal of Medical Informatics, 125, 37-46. https://doi.org/10.1016/j.ijmedinf.2019.02.008

[13] Georgakopoulou, V.E. (2024). The Role of Artificial Intelligence in Combatting Respiratory Tract Infections. Cureus, 16(7), e63635. https://doi.org/10.7759/cureus.63635
[14] Al-Anazi, S., Al-Omari, A., Alanazi, S., Marar, A., Asad, M., Alawaji, F., & Alwateid, S. (2024). Artificial intelligence in respiratory care: Current scenario and future perspective. Annals of Thoracic Medicine, 19(2), 117-130. https://doi.org/10.4103/atm.atm_192_23
[15] Sangwan, P. (2022). Prediction of lung disease using machine and deep learning techniques: A review. International Journal of Health Sciences, 6(S2), 7583–7601. https://doi.org/10.53730/ijhs.v6nS2.6833
[16] Xie, Y., Meng, W., Li, R., Wang, Y., Qian, X., Chan, C., Yu, Z., Fan, X., Pan, H., Xie, C., Wu, Q., Yan, P., Liu, L., Tang, Y., Yao, X., Wang, M., & Leung, E. (2021). Early lung cancer diagnostic biomarker discovery by machine learning methods. Translational Oncology, 14(1), 100907. https://doi.org/10.1016/j.tranon.2020.100907
[17] Perna, D., & Tagarelli, A. (2019). Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks. *IEEE Journal of Biomedical and Health Informatics, 23*(6), 2644-2653. https://doi.org/10.1109/CBMS.2019.00020
[18] Infante, T., Cavaliere, C., Punzo, B., Grimaldi, V., Salvatore, M., & Napoli, C. (2021). Radiogenomics and artificial intelligence approaches applied to cardiac computed tomography angiography and cardiac magnetic resonance for precision medicine in coronary heart disease: A systematic review. *Circulation: Cardiovascular Imaging, 14*(12), e013939. https://doi.org/10.1161/CIRCIMAGING.121.013025
[19] Ibrahim, D.M., Elshennawy, N.M., & Sarhan, A.M. (2021). Deep-chest: multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases. *Computers in Biology and Medicine, 132*, 104348. https://doi.org/10.1016/j.compbiomed.2021.104348
[20] Chen, W., Sun, Q., Chen, X., Xie, G., Wu, H., & Xu, C. (2021). Deep learning methods for heart sounds classification: A systematic review. *Entropy, 23*(5), 667. https://doi.org/10.3390/e23060667