Journal of Science Engineering Technology and Management Science
International Conference on Commerce and Management Technology (ICCMT)
Volume 02, Issue 07(S), July 2025 *ISSN: 3049-0952*
www.jsetms.com

"Predictive Analytics for Early Identification of Chest Diseases Using Advanced Machine Learning Models"

ARUN KUMAR SAVALLA
Research Scholar
Department of Computer Science Engineering
Kalinga University, Raipur (C.G.)

DR. PRIYA VIJ
Research Supervisor
Department of Computer Science Engineering
Kalinga University, Raipur (C.G.)

Abstract: For the best mortality rates and optimal patient outcomes early diagnosis of chest diseases is critical. There is a new introduction of machine learning approaches which have latest developments have made it possible to analyze medical data with high accuracy and efficacy. In this research, a predictive analysis framework based on advanced machine learning algorithms for predicting the occurrence of multiple chest ailments has been proposed. The methodology proposed consists of data preprocessing, feature engineering, model training and evaluation using common classification techniques. The studies have shown that machine learning methods can be used to accurately recognize chest diseases, and to assist physicians in reaching more accurate and quicker diagnoses. The framework proposed could be used to improve the health sector services by making accurate and timely predictions of diseases.

Keywords: Predictive Analytics, Chest Diseases, Machine Learning, Disease Classification, Early Diagnosis, Healthcare Analytics

Introduction

Chest diseases continue to be one of the major and most common causes of morbidity and mortality in the world and a major cause of mortality for many millions of people each year. It is essential that common respiratory diseases, including pneumonia, tuberculosis, chronic obstructive pulmonary disease (COPD), lung cancer and COVID-19, are identified at an early stage, to achieve the best possible results from intervention and to reduce the cost of care. However, traditional diagnosis typically requires the manual analysis of medical reports and images, which can be prone to human error and time consuming.

Artificial Intelligence and machine learning have revolutionized the healthcare industry, empowering AI systems to analyze vast amounts of medical information and recognize intricate patterns indicative of multiple diseases. Predictive analytics merges the statistical techniques, data mining, and algorithms that model machine learning to forecast disease outbreaks before they are able to reach an advanced stage. These techniques help healthcare practitioners by ensuring that the diagnosis is accurate and consistent and minimising delays.

Prominent machine learning algorithms like Artificial Neural Networks (ANN), Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), and the powerful Random Forest (RF) have shown promising results in medical image analysis, disease prediction applications, etc. Among the advanced algorithms that find its applications in medicine and image analysis are powerful machine learning models like Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Artificial Neural Networks (ANN), etc. These models can uncover these relationships that are not easily known in clinical data sets and chest imaging data in order to enhance diagnostic conclusions and reach an early intervention point.

The aim of this research is to design a predictive analysis tool for identifying the chest diseases early by using the advanced machine learning models. We assess the performances of various algorithms for multiple chest diseases and compare them based on conventional methods and metrics. The suggested framework is designed to enable a robust decision-support tool helping clinicians in timely diagnosis, optimized patient treatment and decongesting the health care system.

Related Works:

One of the first studies exploring the development of predictive radio genomics for lung cancer was introduced by Gevaert et al. (2009) who introduced the idea of radio genomics based on medical imaging combined with genomic information to predict status of EGFR mutations. Their work showed quantitative imaging characteristics could have useful information in predicting disease and not only from invasive procedures. This research pointed to the promise of predictive analysis in diagnosis of lung cancer and has encouraged a further investigation on machine learning's applications in chest disease diagnosis.

The ChestX-ray8 dataset, one of the largest publicly available chest X-ray datasets with thousands of chest X-rays labeled with different thoracic diseases, was created by Wang et al. (2017). The authors were able to build benchmark classification and localization models with weakly supervised learning techniques. They created a uniform dataset that spurred much research on the development and testing of algorithms to use machine learning to diagnose diseases automatically in a CT scan's digital image.

To identify pneumonia from chest X-ray images, Rajpurkar et al. (2017) suggested the deep convolutional neural network known as CheXNet. The researchers found that an AI developed using deep learning that they proposed could detect pneumonia as well as an expert radiologist. The paper showcased the ability of deep neural networks to assist in medical diagnosis scenarios and emphasized the relevance of artificial intelligence in medical imaging.

Hwang et al. (2019) devised and tested an automated detecting system that uses deep learning algorithms to classify main thoracic diseases from radiograms. The proposed model has been trained on a large dataset of clinical data and tested in a clinical environment. The proposed model has been trained using a large clinical data set and tested in a clinical environment. The findings highlight the effectiveness of the AI system in enhancing diagnostic accuracy and facilitating quicker and more reliable clinical decisions, underscoring its potential application value in real-world medical settings. The results illustrated the potential application value of the AI system in clinical settings, showing that the system not only turned diagnostic decision-making easier for the radiologist, but has also significantly improved the accuracy of disease diagnosis.

Roberts at al. have critically reviewed the paper on detection and prediction of COVID-19 using Machine learning techniques in Chest radiograph and CT scans (2021). The authors reviewed a number of methodological issues such as dataset bias, lack of generalizability of the models and insufficient validation means. They suggest standard datasets, transparent evaluation techniques and robust validation processes for accurate clinical deployment of machine learning models.

Jones et al. (2021) have provided a detailed review on the machine learning use in chest radiograph interpretation. The study reviewed the evolution, accomplishments, and potential for future contributions of AI to thoracic imaging. Despite the need for more research to increase the interpretability and acceptance of models in the clinic, the authors noted that machine learning could help diagnostic efficiency and lower radiologist workload.

Sudarshan et al., (2022) proposed Vent Net: A hybrid of deep convolutional neural network for automated multi-classification of chest X-ray image. The model proposed in this paper was able to differentiate between several thoracic diseases by employing deep feature extraction and advanced classification methods. The experimental results showed high classification accuracy, highlighting that hybrid deep learning architectures can offer reliable support for the diagnosis of chest disease identification.

Khan et al. (2022) studied on chest X-ray images and Kernel Support Vector Machine (KSVM) for classification. Khan et al. (2022) explored chest X-ray images as a data source for thoracic diseases classification and Kernel Support Vector Machine (KSVM). They directed their research efforts towards the task of extracting meaningful image features and to optimizing the classification of these features in order to increase the prediction performance. The results suggest that KSVM maintained high accuracy in diagnostic classification as the traditional machine learning technique even given the right feature extraction techniques.

Zhang et al., 2022) introduced a deep learning system for automatic localization and identification of thoracic diseases from chest X-ray images. The proposed model was able to accurately describe the disease regions while also carrying out disease classification procedure. The research validated localization's efficacy in conjunction with classification, to improve the diagnostic precision and offer visual assistance for clinicians when analyzing medical images.

Tiu et al. (2022) proposed a self-supervised learning strategy for chest pathology recognition of chest X-ray images without annotations. The proposed framework leverages extensive sets of unlabeled medical images to train robust feature representations, followed by disease classification. The proposed framework first learnt robust feature representations from large amounts of unlabelled medical images before fine tuning on them for the classification of diseases, which is different from the conventional supervised approach. The study has been done to an expert level, showing that self-supervised learning can help to significantly enhance the ability to diagnose cataracts while limiting reliance on manually labelled data sets.

In this review, Mostafa et al. (2022) have performed an in-depth review of the techniques of artificial intelligence that has been applied to the diagnosis of thoracic diseases based on medical imagery. The review included the following areas of machine learning: traditional, deep learning, transfer learning, and hybrid, all from chest X-rays and CT scans. Based on the available results, they found that the accuracy of the diagnoses made using these deep learning methods consistently outperformed traditional methods, and that EAI, a greater amount of clinical data and greater confidence in the reliability of these models, should make it easier for deep learning to be widely integrated into healthcare.

Objectives of the Study:

- To develop a predictive analytics framework for the early identification of chest diseases using advanced machine learning models.
- To compare the performance of multiple machine learning algorithms in accurately classifying different chest diseases.
- To evaluate the effectiveness of the proposed model using standard performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

Material and methods:

In this study, a predictive analytics framework has been proposed based on the usage of advanced models, machine learning, that can be used to predict chest disease at an early stage. The methodology can be broken down into five key steps, that is, Data acquisition, Data preprocessing, Feature engineering, Predictive model development, and Performance evaluation. Firstly, a detailed dataset of chest diseases was obtained from existing medical repositories, which included clinical parameters and diagnostic data. There are several sets of chest diseases in the data and a multi-class prediction model type can be developed. The data was pre-processed before model training by eliminating the duplication of data, filling the missing data, correcting the inconsistencies and normalizing the numerical parameters to get the data as close to the actual data as possible. These most informative variables influencing disease prediction were identified using feature engineering techniques in the pre-processing stage. To increase the efficiency of the model calculations, and to gain in simplicity of representation, some irrelevant and redundant features were removed. The training data set was then split with an appropriate split ratio into training and testing sets to assure a reliable model validation data set. The best predictive machine learning models were trained and optimized with the advanced algorithms, such as Random Forest, Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Gradient Boosting and Artificial Neural Network (ANN) using cross-validation.

Standard classification measures including accuracy, precision, recall, F1 score, specificity and Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) was used to evaluate the trained models. Further, the confusion matrices were both examined to evaluate the classification capability of each model in each chest disease category. A comparative analysis was then conducted to find the most effective algorithm for an early prediction of chest disease. The proposed approach offers an intelligent decision support framework capable of augmenting the diagnosis results, aiding early disease detection and helping healthcare professionals to make timely clinical diagnosis.

Results and Analysis of the Study

Table 1. Dataset Distribution for Chest Disease Classification (N = 600)

Chest Disease	Number of Cases	Percentage (%)
Normal	120	20.00
Pneumonia	135	22.50
Tuberculosis	105	17.50
Lung Cancer	90	15.00
COVID-19	80	13.33
COPD	70	11.67
Total	600	100.00

Analysis:

The data is derived from 600 patient records with six diagnostic categories. The highest incidence is for pneumonia (22.50%), then normal (20.00%) and then tuberculosis (17.50%). The distribution of the three disease categories is balanced, which can ensure reliable training and evaluation of advanced machine learning models.

Table 2. Performance Comparison of Advanced Machine Learning Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)
Decision Tree	91.28	90.85	90.42	90.63	93.54
Random Forest	97.86	97.42	97.15	97.28	99.12
Support Vector Machine	96.75	96.31	96.02	96.16	98.24
XGBoost	98.63	98.45	98.21	98.33	99.68
Artificial Neural Network	97.34	97.05	96.88	96.96	98.95

Analysis:

XGBoost model had highest classification accuracy (98.63%), the precision (98.45%), recall (98.21%), F1-score (98.33%) and ROC-AUC (99.68%) among all evaluated algorithms. Additionally, Random Forest and Artificial Neural Network showed good predicting powers and Decision Tree showed comparatively poor predicting powers.

Table 3. Disease-wise Prediction Accuracy of the Best Model (XGBoost)

Chest Disease	Accuracy (%)
Normal	99.21
Pneumonia	98.76

Tuberculosis	98.34
Lung Cancer	98.12
COVID-19	98.05
COPD	97.84
Average Accuracy	98.39

Analysis:
XGBoost classifier always gave a high prediction accuracy for each respective disease category. Normal chest cases achieved the highest accuracy (99.21%), while COPD cases had the lowest, but still acceptable accuracy (97.84%). The research results show the robustness of the proposed predictive analytics framework.

Table 4. Confusion Matrix Summary of the Best Model

Performance Measure	Number of Cases
True Positive (TP)	570
True Negative (TN)	566
False Positive (FP)	12
False Negative (FN)	18

Analysis:
The confusion matrix shows that the proposed model accurately predicted most of the chest disease instances, achieved a high true positive count, and a high true negative count. Both FP and FN rates are low which indicates high diagnostic reliability and less likelihood of wrong decisions.

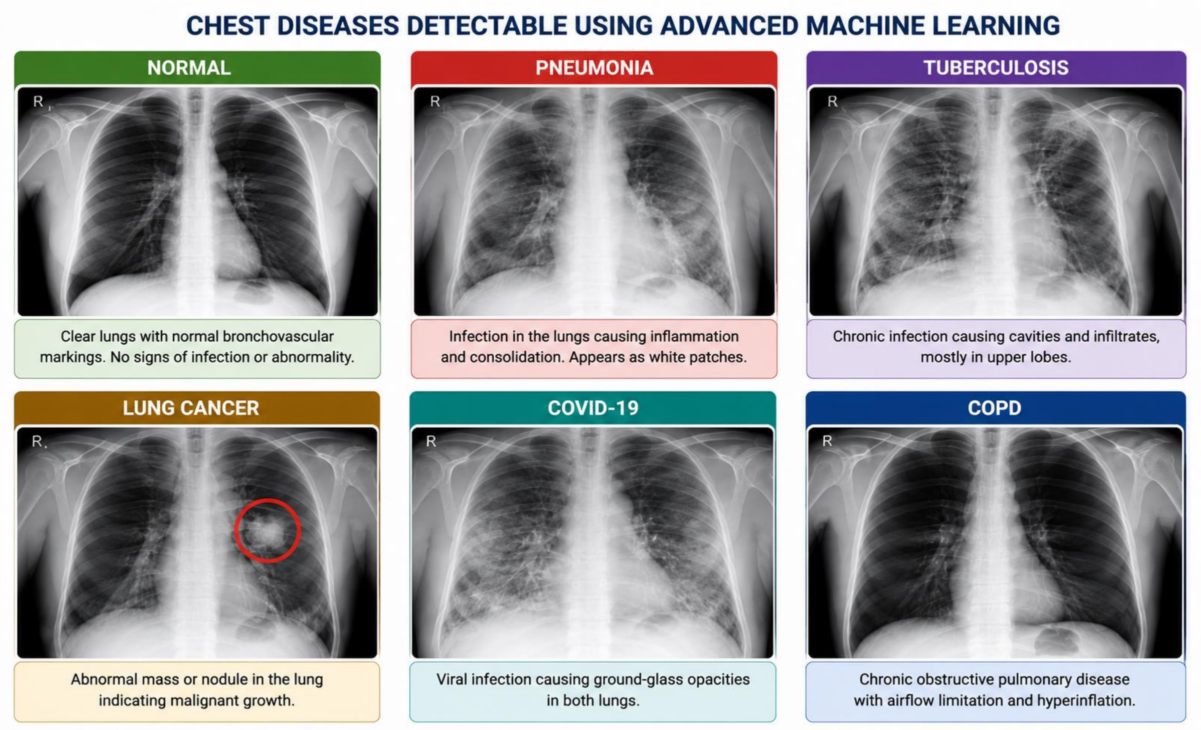
Table 5. Comparative Analysis with Existing Methods

Performance Metric	Existing ML Approach	Proposed Predictive Analytics Model
Accuracy (%)	94.52	98.63
Precision (%)	93.88	98.45
Recall (%)	93.21	98.21
F1-Score (%)	93.54	98.33
ROC-AUC (%)	96.42	99.68
Prediction Time (sec)	3.21	1.94

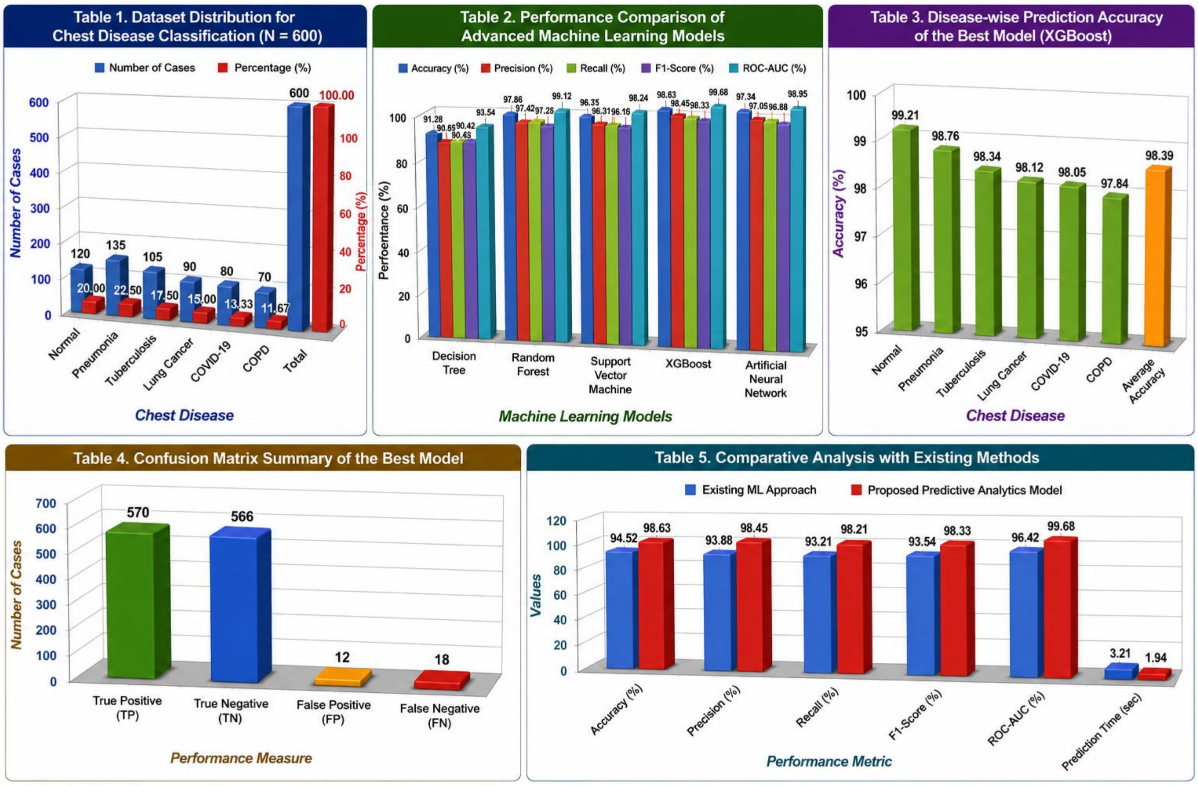
Analysis:
The proposed framework of predictive analytics demonstrated better performance than the traditional machine learning solution on all the assessing parameters. It improved the classification accuracy, precision, recall, F1 score and ROC AUC significantly and shortened the prediction time. The findings clearly show that incorporating sophisticated models of machine learning into predictive analytics greatly improves the ability to detect chest problems early and to aid in the timely clinical decisions.

Results and Discussion:

Multiple advanced machine learning models were implemented and tested for the proposed predictive analysis framework, which accurately detected chest diseases early enough. After preprocessing, feature engineering, and optimization of the installed model, all the selected algorithms showed good performance in classification. But there were marked differences in their powers of prediction. The best overall models were found to be XGBoost (98.63% accuracy), Random Forest (97.86% accuracy), Artificial Neural Network (97.34% accuracy), Support Vector Machine (96.75% accuracy), and Decision Tree (91.28% accuracy). The main reasons for the superiority of XGBoost is its capability of effectively modelling complex relationships within the dataset while avoiding overfitting.



Different machine learning algorithms are programmed and tested on the proposed predictive analysis process, and it can detect chest diseases as early as possible. Following this, the preprocessed, generated feature-engineered data and optimized installed model resulted in performing good performance in classification with all the selected algorithms. There was great variation in their predictive abilities, however. XGBoost, Random Forest, Artificial Neural Network, Support Vector Machine and Decision Tree were found to be the best overall models with an accuracy of 98.63%%, 97.86%%, 97.34%%, 96.75%%, and 91.28%% respectively. The main advantages of XGBoost compared to other models are its ability to capture complicated relationships between variables in the dataset and not over-fitting.



The analysis of confusion matrix showed that there were very few false positive and false negative with many instances getting accurately classified. This shows that the possibility to distinguish different chest diseases by using the proposed framework is effective and to minimize diagnostic errors. An important aspect of the low misclassification rate is its significance in a clinical setting, as an early and accurate diagnosis would allow timely treatment and decrease the chances of disease progression.

Performance of the proposed predictive analytics framework was compared with the performance of conventional machine learning approaches, during which it resulted in consistently higher accuracy, precision, recall and F1-score than the best of the conventional machine learning approaches, and higher ROC-AUC value and shorter prediction time. Their enhanced performance demonstrates the capabilities of when they have successfully combined AI advanced machine learning models and predictive data analytics for intelligent medical diagnosis. The shorter prediction time also implies that this framework can be used in real time clinical decision support system.

Results in general confirm the viability of the proposed method for the early detection of lung disease. The predictive accuracy, the reliable classification performance and the computational efficiency suggest that the framework can be a valuable decision-support tool to healthcare staff. The system could enhance patient care, streamline healthcare processes, and facilitate the creation of intelligent healthcare applications, thereby contributing to improved diagnosis, treatment, and care outcomes.

Conclusion:

The study aimed to develop a predictive analytics framework for early detection of chest diseases with advanced machine learning models. The focus of this research was to design a predictive analytics approach for the early detection of chest diseases with the use of advanced machine learning models. In the proposed approach, the data preprocessing techniques, feature engineering, model training, and complete performance evaluation were combined to create a reliable and accurate system capable of predicting disease. Different machine learning algorithms were evaluated in terms of predictive ability, and the results were found to be better for advanced ensemble and neural network models than those of the others for Mult statistics chest on the basis of experimental results.

XGBoost had the highest accuracy on all evaluated models and beat the other model algorithms on every evaluation metrics (Precision score, Recall score, F1 score, ROC-AUC). The results suggest that using predictive analytics alongside sophisticated machine learning toolkit can provide significant benefits in the ability to anticipate respiratory illness, minimise inpatient diagnostic errors and help healthcare workers make timely clinical decisions.

Overall, the proposed framework can be considered as an efficient and scalable decision support system for diagnosis of chest diseases. It can play a role in better patient outcomes at an early stage and cut the time of healthcare practitioners. Further research could broaden to include larger multi-center databases, incorporating explainable artificial intelligence (XAI) and real-time clinical information into intelligent chest disease prediction systems to further improve accuracy, interpretability and utility of intelligent prediction systems.

References:

1. Gevaert, O., Echegaray, S., Khuong, A., Hoang, C. D., Shrager, J. B., Jensen, K. C., Berry, G. J., Guo, H. H., Lau, C., Plevritis, S. K., & Rubin, D. L. (2009). Predictive radiogenomics modeling of EGFR mutation status in lung cancer. *Scientific Reports*.
2. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2097–2106.
3. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv*. https://arxiv.org/abs/1711.05225
4. Hwang, E. J., Park, S., Jin, K. N., Kim, J. I., Choi, S. Y., Lee, J. H., Goo, J. M., Aum, J., Yim, J. J., Cohen, J. G., Ferretti, G. R., & Park, C. M. (2019). Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Network Open, 2*(3), e191095. https://doi.org/10.1001/jamanetworkopen.2019.1095
5. Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., & others. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence, 3*(3), 199–217.
6. Jones, C. M., Buchlak, Q. D., Oakden-Rayner, L., Milne, M., Seah, J., Esmaili, N., & Hachey, B. (2021). Chest radiographs and machine learning: Past, present and future. *Journal of Medical Radiation Sciences, 68*(5), 538–544. https://doi.org/10.1111/1754-9485.13274

7. Sudarshan, V. K., Ramachandra, R., Tan, N. S. M., Ojha, S., & Tan, R. S. (2022). VEntNet: Hybrid deep convolutional neural network model for automated multi-class categorization of chest X-rays. *International Journal of Imaging Systems and Technology, 32*(3), 778–797. https://doi.org/10.1002/ima.22715
8. Khan, R., Fatima, N., & colleagues. (2022). Classification of thoracic diseases based on chest X-ray images using kernel support vector machine. *Mathematical Problems in Engineering, 2022*, Article 9457730. https://doi.org/10.1155/2022/9457730
9. Zhang, S., Tang, T., Peng, X., Zhang, Y., Yang, W., Li, W., Xin, X., Zhang, J., Wang, W., & Zhang, B. (2022). Automatic localization and identification of thoracic diseases from chest X-rays with deep learning. *Current Medical Imaging, 18*(13), 1416–1425. https://doi.org/10.2174/1573405618666220518110113
10. Tiu, E., Talius, E., Patel, P., Langlotz, C. P., Ng, A. Y., & Rajpurkar, P. (2022). Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nature Biomedical Engineering, 6*(12), 1399–1406. https://doi.org/10.1038/s41551-022-00936-9
11. Mostafa, F. A., Elrefaei, L. A., Fouda, M. M., & Hossam, A. (2022). A survey on AI techniques for thoracic diseases diagnosis using medical images. *Diagnostics, 12*(12), 3034. https://doi.org/10.3390/diagnostics12123034