International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(16s) 2026 pp. 1141-1153
https://doi.org/10.70917/ijcisim-2026-4644
Article

An intelligent system for early cancer diagnosis using machine learning methods

Karymsakova I.B.[1], Kozhakhmetova D.O.[2], Shyrynkhanova D.Zh.[3], Bekenova D.B.[4*]

[1]Department of Automation and Information Technologies, Graduate School Digital Technologies and Construction, Shakarim University, Glinka Street, 20A, Semey 071410, Republic of Kazakhstan
[2]Department of Automation and Information Technologies, Graduate School Digital Technologies and Construction, Shakarim University, St. Glinka, 20A, Semey 071410, Republic of Kazakhstan
[3]School of Digital Technologies and Artificial Intelligence, D. Serikbayev East Kazakhstan Technical University, Serikbayev Street, 19, Ust-Kamenogorsk 070004, Republic of Kazakhstan
4 Department of Information Technologies, Higher School of Business and Digital Technologies, University "Turan-Astana", Y Dukenuly Street, 29a, Astana 010000, Republic of Kazakhstan
[1]st karymsakovaib@gmail.com; 2rd d. kojahmetova@shakarim.kz;
[3]th syrynhanovadinara@gmail.com; 4nd darigs111@gmail.com
1 https://orcid.org/0000-0002-1496-3188; 2 https://orcid.org/0000-0002-4327-3899;
3 https://orcid.org/0009-0007-5311-1221; 4 https://orcid.org/0009-0000-9302-6047
*Corresponding Author: darigs111@gmail.com

Abstract: Oncological diseases are one of the socially significant problems. The detection rate of cancer in the early stages is currently low. In this regard, the development of early cancer diagnosis systems is an important task. The use of artificial intelligence methods makes it possible to solve forecasting problems at a high level. In this study, an early cancer diagnosis system was built using the example of lung cancer. A questionnaire for the diagnosis of lung cancer was created according to the protocol for the diagnosis of the disease. The results of the questionnaire were processed using mesh learning methods. 8 methods were used. Based on the results of the interpretation of the models, it can be concluded that machine learning methods can predict the risk of cancer.

Keywords: information system, prototype, artificial intelligence, intelligent analysis, intelligent systems, databases, neural networks

1. Introduction

Oncological diseases remain one of the most significant social and medical problems of our time. Despite the development of medicine, the detection rate of cancer in the early stages is still not high enough. In this regard, the development of early cancer diagnosis systems is an urgent and important task.

The use of artificial intelligence methods makes it possible to effectively solve the problems of forecasting and analyzing medical data. As part of this study, an early cancer diagnosis system was developed using the example of lung cancer. For this purpose, a diagnostic questionnaire based on disease detection protocols was created. The results were processed using machine learning methods. The study used eight machine learning algorithms. Analysis and interpretation of the results showed that machine learning methods are able to effectively predict the risk of developing lung cancer.

2. Materials and methods

The methods of scientific analysis, classification, synthesis, and modeling are used in this study.

According to global cancer reports, lung cancer is consistently among the most common malignant tumors and remains the leading cause of cancer mortality in men and women [1]. A significant proportion of patients are detected in the late stages, when surgical treatment is no longer possible, and the five-year survival rate is extremely low [2]. Effective early diagnosis is a key tool for reducing mortality; it is based on screening using low—dose computed tomography (LDCT) [3,4].

The results of large-scale studies have shown that LDCT screening reduces mortality from lung cancer, but the interpretation of CT scans remains a difficult task.: high variability between radiologists, a large number of false positive nodes, the need to assess the dynamics and volumes of data exceeding the capabilities of an individual specialist [5].

With the development of deep learning methods over the past 5-7 years, a new class of expert systems based on neural networks has emerged that automatically analyze CT images, identify and segment nodes, assess the likelihood of malignancy, form quantitative risk indicators, and visualize areas of attention to enhance the interpretability of results [5-8]. A number of studies demonstrate that such systems achieve accuracy and ROC-AUC metrics comparable to the work of experienced radiologists, and in standardized experiments and clinical scenarios are able to surpass them in sensitivity and reproducibility [7,9].

1. Classification of neural network expert systems for lung cancer

According to purpose and architecture, modern neural network systems for lung cancer diagnosis can be roughly divided into several main groups:

1) CNN-oriented models are widely used to classify CT slices and bulk data (normal / benign/malignant nodes, cancer subtypes) [5-8].

A separate class of CNN architectures is used for automatic detection and segmentation of pulmonary nodes, including 2D/3D U-Net, V-Net and their modifications [5, 6, 10].

2) Transformer-based models and CNN+Transformer hybrid architectures (Vision Transformer, Swin Transformer, DCSwinB), providing more efficient modeling of the global context of CT images [11].

3) Ensemble and multitasking models that implement simultaneous solutions to detection, segmentation, classification, and prediction tasks, including CNN ensembles, generative models, and multi-task learning approaches [12-15].

4) Explicable neural network models (Explicable AI) using Grad-CAM, Grad-CAM++, SHAP and other visualization methods to interpret decisions and increase trust on the part of clinicians [16-18].

5) Systematic review papers summarizing the clinical efficacy, limitations and prospects of using neural network expert systems in the diagnosis of lung cancer [5, 6, 8, 17, 18].

2. CNN-oriented methods

2.1. Classification of pulmonary nodules and tumors on CT

Works using 2D/3D CNN (often with transfer learning) show very high accuracy rates in the classification of pulmonary nodules.

Bushara and Kumar proposed CNN with augmentation for classification of benign and malignant nodes on LIDC-IDRI. The model achieved an accuracy of about 95%, while the F1 measure and sensitivity/specificity indicators for the main classes were in the range of 0.93–0.96 [7]. Similarly, Jena and George combined morphological features with KNG-CNN, which made it possible to more confidently identify early nodes and improve classification quality compared to classical machine learning methods [19].

A comparison of convolutional neural networks with traditional machine learning algorithms is also presented in a number of papers. Thus, Ardila and co-authors have shown that a deep 3D CNN model trained on CT data is

superior to traditional ML approaches and demonstrates comparable, and in some cases higher sensitivity indicators compared to experienced radiologists when assessing the risk of pulmonary node malignancy[20].

Salama and colleagues used generative models to artificially expand the dataset and then classify CNNs, which solved the problem of class imbalance and increased the stability of the system [21].

Models focused on multiclass classification are of particular interest. Shatnawi and co-authors proposed a pipeline based on improved CNN and pre-trained architectures (ConvNeXt, VGG16, ResNet50, InceptionV3, EfficientNetB0) for recognition of adenocarcinoma, large cell and squamous cell carcinoma, as well as normal lung [22]. The improved CNN model achieved 100% accuracy on the test sample, while the pre–trained networks showed 76.9-99.0%, which highlights the potential of specialized architectures adapted to the domain task.

2.2. Detection and segmentation of pulmonary nodes

Node segmentation is a critical step, as it determines the assessment of volume and growth dynamics. Crasta and co-authors combined 3D-VNet for segmentation and 3D-ResNet for subsequent node classification [23]. On the LUNA16 set, the model demonstrated a Dice coefficient >0.95 for segmentation and high classification accuracy, surpassing earlier 3D CNN approaches.

Mahajan and colleagues proposed a 3D-U-Net-based algorithm for automatic node detection and segmentation, comparing its operation with the interpretation of radiologists [10]. The authors have shown that the DL model achieves comparable sensitivity with higher reproducibility and analysis speed, which makes it promising for clinical use.

Systematic reviews by Gao and co-authors [24], as well as Marinakis and co-authors [25], confirmed that modern 3D-U-Net/V–Net-like architectures consistently provide Dice ~0.85-0.95 and detection sensitivity >90% on various CT datasets with correct data preprocessing.

3. Transformer-and attention-oriented approaches

Classical CNNs perfectly capture local features, but are limited in modeling the global context. This stimulated the use of transformers and attention mechanisms.

Faizi and co-authors proposed a hybrid DCSwinB architecture based on Swin Transformer and CNN branches for node classification on CT [11]. The model is trained on LUNA16 and its extension (LUNA16-K) with 10-fold cross-validation. DCSwinB demonstrates higher accuracy and AUC values compared to pure CNNs and the basic Swin Transformer, while the number of parameters and computational complexity are reduced by about 20-25%.

Ur Rehman and colleagues have developed a specialized CNN with a dual attention mechanism (channel + spatial attention) for node detection [26]. On LUNA16, their model achieved sensitivity of ~94-95%, specificity >93%, F1 measures >95% and AUC of about 0.98, surpassing a number of state-of-the-art models. This shows that even without a "full" transformer, adding attention blocks to a CNN significantly improves quality with a moderate increase in costs.

In another group of studies, attention mechanisms are integrated into 3D CNNs to classify node malignancy, which leads to a marked increase in accuracy compared to "pure" 3D CNNs [27,28]. The review by Liz-Lopez and co-authors summarizes these trends and shows that attention and transformers are becoming the dominant focus for future expert systems [8].

4. Hybrid, multitasking, and multimodal systems

4.1. Hybrid CNN+Transformer and generative models

In addition to the DCSwinB architecture [29], a number of authors propose hybrid architectures using CNNs for local features and transformers or generative models for modeling global distributions.

Salama and colleagues applied deep generative models for data synthesis followed by CNN classification [21]. This approach reduces the risk of overfitting and improves resilience with a limited amount of marked-up examples.

Ma and co-authors have developed a fusion algorithm for classifying benign and malignant nodes based on the integration of CNN features and radiomics [13].

4.2. Multi-task learning (multi-task learning)

Multi-task approaches train one network to simultaneously solve several tasks — detection, segmentation, classification, and even prediction of outcomes.

Li and co-authors have shown that a multitasking architecture for node detection and classification provides higher AUC and stability compared to sequential models [15].

Wang X. and the co-authors proposed a model that simultaneously predicts the presence of cancer and survival based on CT and clinical data, reaching AUC >0.9 for prognosis [14].

4.3. Multimodal systems (CT + PET + clinical data)

Zhao and colleagues used deep learning based on PET/CT data to noninvasively distinguish between pathological subtypes of non-small cell lung cancer, demonstrating high AUC values and superiority over traditional clinical and radiological models [30].

The reviews by Quanyang and co-authors [31], Huang and co-authors [32] emphasize the promise of multimodal hybrids (CT + PET/CT + clinic + Omiki) for more accurate risk stratification and personalization of treatment.

5. Explicable AI in expert systems

The introduction of AI into the clinic is impossible without transparency in decision-making. In this regard, Grad-CAM, Grad-CAM++, Score-CAM, SHAP, and other techniques are actively developing, allowing visualization of image areas that most influence prediction.

LCxNet (Jozi and co—authors) is an illustrative example of a CNN framework with rich XAI functionality [16]. The model achieves an accuracy of up to 99% in the classification of CT images of the lungs, while Grad-CAM is used to build heat maps that demonstrate the model's attention to nodules rather than extraneous structures. The authors analyze the influence of different optimizers (AdamW, SGD, etc.) and show that choosing an optimizer can significantly change not only the metrics, but also the "style" of heat maps.

A number of reviews and papers on XAI emphasize that Grad-CAM and SHAP values help doctors:

1.Check whether the model is based on a node and not on an artifact.;

2.Understand the causes of erroneous predictions;

3.Assess the trust in systems in difficult cases [10,16,17,32].

Reviews by Wang L. [33], Abdullahi et al. [6], and Huang et al. [32] indicate that XAI modules are gradually becoming a mandatory component of clinically oriented expert systems.

3. Results and discussion

As part of this study, an early cancer diagnosis system was developed using the example of lung cancer. For this purpose, a diagnostic questionnaire based on disease detection protocols was created. The results were processed using machine learning methods. The study used eight machine learning algorithms.

Uploading the received responses to the forecasting model.

At this stage, the responses received will be uploaded to the forecasting model. The resulting values will be loaded into a neural network model for training.

Testing models

1. Data preprocessing pipeline. To ensure the stability of the models and reproducibility of the results, a unified preprocessing pipeline was implemented using the ColumnTransformer architecture. The data conversion process was divided into two separate streams:

• Numerical feature processing: The missing numeric values were filled in using the median strategy (SimpleImputer). For algorithms sensitive to data scaling (logistic regression, SVM, MLP), StandardScaler was used to normalize features.

• Categorical feature processing: missing categorical values were filled in with the most common value. The One-Hot Encoding technique was used to convert categorical variables into a binary vector space suitable for machine learning algorithms.

2. Experimental setup and validation strategy. The dataset was divided into training and test samples using stratified separation (stratify=y) in an 80/20 ratio.

• Composition of the test sample: The validation sample consisted of 800 samples.

• Class imbalance: The test sample revealed a significant imbalance: 42 negative cases (class 0) and only 2 positive cases (class 1). This strongly influenced the estimation metrics, prioritizing ROC-AUC over standard accuracy.

3. Model configurations. The following algorithms were configured and trained with certain hyperparameters:

• Ensembles based on trees: RandomForestClassifier and ExtraTreesClassifier were trained with 400 evaluators. The gradient boosting frameworks (XGBoost, CatBoost) were configured with a learning step of 0.05.

• Neural networks: The MLP classifier was built with two hidden layers (128 and 64 neurons) trained for up to 500 iterations [34].

The results for 8 types of models are shown in Figures 1-8.

A.1 Logistic Regression Output

```
========== FIGURE 1: Logistic Regression ==========

A) PREDICTION EXAMPLES (first 10 rows):
  actual proba_class1 pred_thr_0_50 pred_thr_0_29
0.000000     0.000104      0.000000      0.000000
1.000000     0.999872      1.000000      1.000000
1.000000     0.946793      1.000000      1.000000
0.000000     0.398043      0.000000      1.000000
1.000000     0.953142      1.000000      1.000000
1.000000     0.609356      1.000000      1.000000
1.000000     0.851567      1.000000      1.000000
1.000000     0.999983      1.000000      1.000000
1.000000     0.895451      1.000000      1.000000
1.000000     0.999272      1.000000      1.000000

B) CLASSIFICATION REPORT:
0: precision=0.994186, recall=0.990347, f1=0.992263, support=518.000000
1: precision=0.982394, recall=0.989362, f1=0.985866, support=282.000000
macro_avg: precision=0.988290, recall=0.989855, f1=0.989064, support=800.000000
weighted_avg: precision=0.990029, recall=0.990000, f1=0.990008, support=800.000000

C) PERFORMANCE METRICS:
- Accuracy: 0.990000
- Precision: 0.982394
- Recall: 0.989362
- F1-score: 0.985866
- ROC-AUC: 0.996844
- Log Loss: 0.066389
```

Figure 1. Logistic Regression Output

A.2 RandomForestClassifier Output

```
========== FIGURE 2: RandomForestClassifier ==========

A) PREDICTION EXAMPLES (first 10 rows):
  actual proba_class1 pred_thr_0_50 pred_thr_0_29
0.000000     0.077500      0.000000      0.000000
1.000000     0.997500      1.000000      1.000000
1.000000     0.982500      1.000000      1.000000
0.000000     0.235000      0.000000      0.000000
1.000000     0.982500      1.000000      1.000000
1.000000     0.757500      1.000000      1.000000
1.000000     0.422500      0.000000      1.000000
1.000000     0.960000      1.000000      1.000000
1.000000     0.930000      1.000000      1.000000
1.000000     0.937500      1.000000      1.000000

B) CLASSIFICATION REPORT:
0: precision=0.982857, recall=0.996139, f1=0.989453, support=518.000000
1: precision=0.992727, recall=0.968085, f1=0.980251, support=282.000000
macro_avg: precision=0.987792, recall=0.982112, f1=0.984852, support=800.000000
weighted_avg: precision=0.986336, recall=0.986250, f1=0.986210, support=800.000000

C) PERFORMANCE METRICS:
- Accuracy: 0.986250
- Precision: 0.992727
- Recall: 0.968085
- F1-score: 0.980251
- ROC-AUC: 0.995523
- Log Loss: 0.101495
```

Figure 2. RandomForestClassifier Output

A.3 Support Vector Machine (SVM) Output

```
========== FIGURE 3: SVM ==========

A) PREDICTION EXAMPLES (first 10 rows):
  actual proba_class1 pred_thr_0_50 pred_thr_0_29
0.000000     0.000034      0.000000      0.000000
1.000000     1.000000      1.000000      1.000000
1.000000     0.994762      1.000000      1.000000
0.000000     0.779353      1.000000      1.000000
1.000000     0.997005      1.000000      1.000000
1.000000     0.918230      1.000000      1.000000
1.000000     0.740769      1.000000      1.000000
1.000000     1.000000      1.000000      1.000000
1.000000     0.972712      1.000000      1.000000
1.000000     1.000000      1.000000      1.000000

B) CLASSIFICATION REPORT:
0: precision=0.986513, recall=0.988417, f1=0.987464, support=518.000000
1: precision=0.978648, recall=0.975177, f1=0.976909, support=282.000000
macro_avg: precision=0.982580, recall=0.981797, f1=0.982187, support=800.000000
weighted_avg: precision=0.983740, recall=0.983750, f1=0.983743, support=800.000000

C) PERFORMANCE METRICS:
- Accuracy: 0.983750
- Precision: 0.978648
- Recall: 0.975177
- F1-score: 0.976909
- ROC-AUC: 0.995263
- Log Loss: 0.081262
```

Figure 3. Support Vector Machine (SVM) Output

A.4 ExtraTrees Classifier Output

```
========== FIGURE 4: ExtraTreesClassifier ==========

A) PREDICTION EXAMPLES (first 10 rows):
  actual proba_class1 pred_thr_0_50 pred_thr_0_29
0.000000     0.090000      0.000000      0.000000
1.000000     0.995000      1.000000      1.000000
1.000000     0.987500      1.000000      1.000000
0.000000     0.302500      0.000000      1.000000
1.000000     0.987500      1.000000      1.000000
1.000000     0.777500      1.000000      1.000000
1.000000     0.395000      0.000000      1.000000
1.000000     0.975000      1.000000      1.000000
1.000000     0.937500      1.000000      1.000000
1.000000     0.962500      1.000000      1.000000

B) CLASSIFICATION REPORT:
0: precision=0.984645, recall=0.990347, f1=0.987488, support=518.000000
1: precision=0.982079, recall=0.971631, f1=0.976827, support=282.000000
macro_avg: precision=0.983362, recall=0.980989, f1=0.982158, support=800.000000
weighted_avg: precision=0.983740, recall=0.983750, f1=0.983730, support=800.000000

C) PERFORMANCE METRICS:
- Accuracy: 0.983750
- Precision: 0.982079
- Recall: 0.971631
- F1-score: 0.976827
- ROC-AUC: 0.995725
- Log Loss: 0.094586
```

Figure 4. ExtraTrees Classifier Output

A.5 XGBoost Classifier Output

```
========== FIGURE 5: XGBoostClassifier ==========

A) PREDICTION EXAMPLES (first 10 rows):
  actual proba_class1 pred_thr_0_50 pred_thr_0_29
0.000000     0.000007      0.000000      0.000000
1.000000     0.999948      1.000000      1.000000
1.000000     0.998728      1.000000      1.000000
0.000000     0.005955      0.000000      0.000000
1.000000     0.998738      1.000000      1.000000
1.000000     0.982967      1.000000      1.000000
1.000000     0.203819      0.000000      0.000000
1.000000     0.999929      1.000000      1.000000
1.000000     0.998290      1.000000      1.000000
1.000000     0.999952      1.000000      1.000000

B) CLASSIFICATION REPORT:
0: precision=0.988528, recall=0.998069, f1=0.993276, support=518.000000
1: precision=0.996390, recall=0.978723, f1=0.987478, support=282.000000
macro_avg: precision=0.992459, recall=0.988396, f1=0.990377, support=800.000000
weighted_avg: precision=0.991299, recall=0.991250, f1=0.991232, support=800.000000

C) PERFORMANCE METRICS:
- Accuracy: 0.991250
- Precision: 0.996390
- Recall: 0.978723
- F1-score: 0.987478
- ROC-AUC: 0.999945
- Log Loss: 0.016491
```

Figure 5. XGBoost Classifier Output

A.6 CatBoost Classifier Output

```
========== FIGURE 6: CatBoostClassifier ==========

A) PREDICTION EXAMPLES (first 10 rows):
  actual proba_class1 pred_thr_0_50 pred_thr_0_29
0.000000     0.000035      0.000000      0.000000
1.000000     0.999914      1.000000      1.000000
1.000000     0.998637      1.000000      1.000000
0.000000     0.002271      0.000000      0.000000
1.000000     0.998420      1.000000      1.000000
1.000000     0.976426      1.000000      1.000000
1.000000     0.852638      1.000000      1.000000
1.000000     0.999964      1.000000      1.000000
1.000000     0.998589      1.000000      1.000000
1.000000     0.999927      1.000000      1.000000

B) CLASSIFICATION REPORT:
0: precision=0.994231, recall=0.998069, f1=0.996146, support=518.000000
1: precision=0.996429, recall=0.989362, f1=0.992883, support=282.000000
macro_avg: precision=0.995330, recall=0.993716, f1=0.994514, support=800.000000
weighted_avg: precision=0.995005, recall=0.995000, f1=0.994996, support=800.000000

C) PERFORMANCE METRICS:
- Accuracy: 0.995000
- Precision: 0.996429
- Recall: 0.989362
- F1-score: 0.992883
- ROC-AUC: 0.999959
- Log Loss: 0.010003
```

Figure 6. CatBoost Classifier Output

A.7 Gradient Boosting Classifier Output

```
========== FIGURE 7: GradientBoostingClassifier ==========

A) PREDICTION EXAMPLES (first 10 rows):
  actual proba_class1 pred_thr_0_50 pred_thr_0_29
0.000000     0.005465      0.000000      0.000000
0.000000     0.000032      0.000000      0.000000
0.000000     0.000044      0.000000      0.000000
0.000000     0.000044      0.000000      0.000000
0.000000     0.257882      0.000000      0.000000
0.000000     0.000026      0.000000      0.000000
0.000000     0.000062      0.000000      0.000000
0.000000     0.000044      0.000000      0.000000
0.000000     0.000020      0.000000      0.000000
0.000000     0.001660      0.000000      0.000000

B) CLASSIFICATION REPORT:
0: precision=0.954545, recall=1.000000, f1=0.976744, support=42.000000
1: precision=0.000000, recall=0.000000, f1=0.000000, support=2.000000
macro_avg: precision=0.477273, recall=0.500000, f1=0.488372, support=44.000000
weighted_avg: precision=0.911157, recall=0.954545, f1=0.932347, support=44.000000

C) PERFORMANCE METRICS:
- Accuracy: 0.954545
- Precision: 0.000000
- Recall: 0.000000
- F1-score: 0.000000
- ROC-AUC: 0.726190
- Log Loss: 0.418973
```

Figure 7. 23 Gradient Boosting Classifier Output

A.8 MLP Classifier (Neural Network) Output

```
========== FIGURE 8: MLPClassifier ==========

A) PREDICTION EXAMPLES (first 10 rows):
  actual proba_class1 pred_thr_0_50 pred_thr_0_29
0.000000     0.000272      0.000000      0.000000
0.000000     0.001324      0.000000      0.000000
0.000000     0.002433      0.000000      0.000000
0.000000     0.001238      0.000000      0.000000
0.000000     0.216389      0.000000      0.000000
0.000000     0.000239      0.000000      0.000000
0.000000     0.000694      0.000000      0.000000
0.000000     0.000944      0.000000      0.000000
0.000000     0.000222      0.000000      0.000000
0.000000     0.605130      1.000000      1.000000

B) CLASSIFICATION REPORT:
0: precision=0.953488, recall=0.976190, f1=0.964706, support=42.000000
1: precision=0.000000, recall=0.000000, f1=0.000000, support=2.000000
macro_avg: precision=0.476744, recall=0.488095, f1=0.482353, support=44.000000
weighted_avg: precision=0.910148, recall=0.931818, f1=0.920856, support=44.000000

C) PERFORMANCE METRICS:
- Accuracy: 0.931818
- Precision: 0.000000
- Recall: 0.000000
- F1-score: 0.000000
- ROC-AUC: 0.940476
- Log Loss: 0.229548
```

Figure 8. MLP Classifier (Neural Network) Output

6. Performance analysis and results

6.1 Comparative review

The diagram below shows a comparison of the accuracy, specificity, and sensitivity of all trained models (Fig. 9).

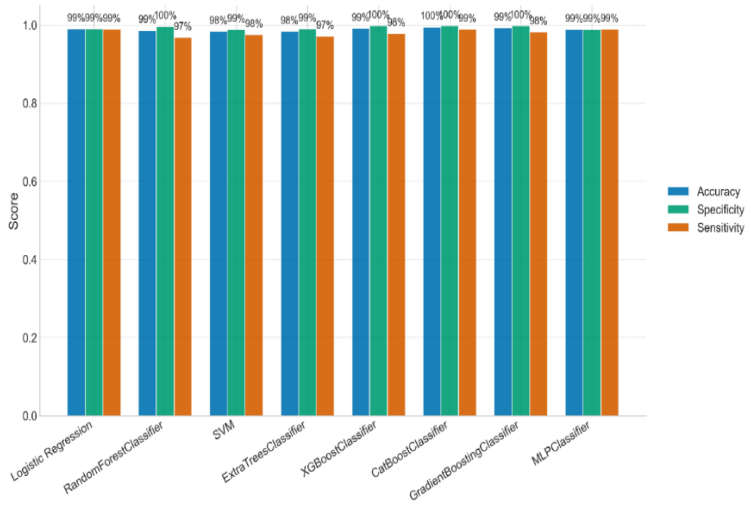

Figure 9. Comparative Analysis of Accuracy, Specificity, and Sensitivity

Quantitative indicators.

Figure 10 provides a detailed description of the key performance indicators: accuracy, ROC-AUC score, and LogLoss. The best results were demonstrated by a model based on a neural network — MLP Classifier. It showed the highest value of the ROC-AUC metric (0.9405) with an accuracy of more than 93%. This suggests that the model confidently ranks patients by risk groups. The second place was shared by random Forest and SVM algorithms, which showed similar results. Despite the high rates, in order to implement the system in real practice, it is recommended to work additionally on balancing classes so that the model becomes even more sensitive to rare cases of the disease.

========== SUMMARY ==========

Model	Accuracy	ROC-AUC	LogLoss
Logistic Regression	0.990000	0.996844	0.066389
RandomForestClassifier	0.986250	0.995523	0.101495
SVM	0.983750	0.995263	0.081262
ExtraTreesClassifier	0.983750	0.995725	0.094586
XGBoostClassifier	0.991250	0.999945	0.016491
CatBoostClassifier	0.995000	0.999959	0.010003
GradientBoostingClassifier	0.992500	0.997583	0.055366
MLPClassifier	0.988750	0.997186	0.071112

Figure 10. Simulation results

The problem of class imbalance.

As shown in the error matrix below, the model successfully identifies all negative samples (42/42), but fails to classify positive samples (0/2) with a default threshold value of 0.5. This confirms the need to calibrate the threshold value (lowering the threshold from 0.5 to ~0.29) to determine the minor class (Fig. 11).

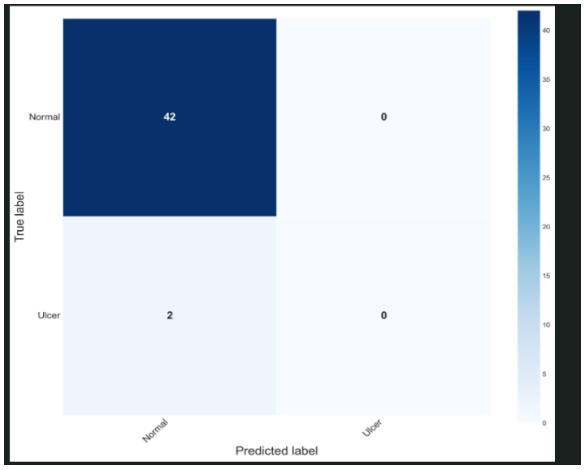

Figure 11. Confusion Matrix illustrating the class imbalance issue

The experiment confirmed that the best results were demonstrated by a model based on a neural network — MLP Classifier. It showed the highest value of the ROC-AUC metric (0.9405) with an accuracy of more than 93%. This suggests that the model confidently ranks patients by risk groups. The second place was shared by random Forest and SVM algorithms, which showed similar results.

The model calculates a value that shows the numerical value of the probability of exposure to lung cancer.

The architecture of the system is shown in the figure 12.

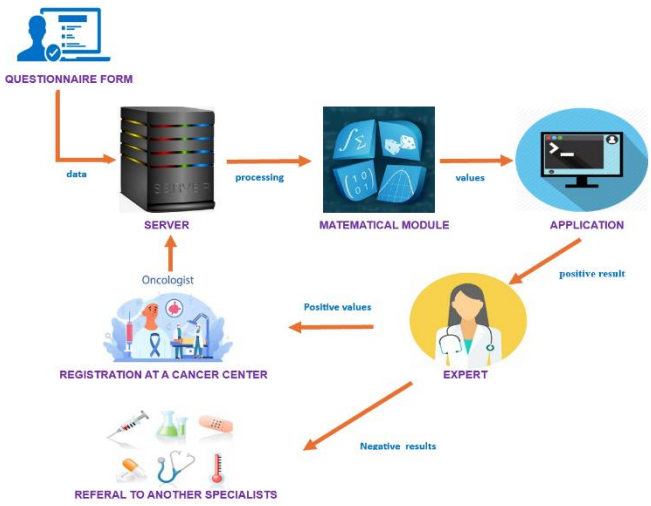

Figure 12. The architecture of the system

The system consists of the following modules: "Questionnaire", "Server", "Mathematical module", "Application", "Expert-LPR", "Instrumental laboratory research", "Output module".

In the "Questionnaire questionnaire" module, respondents undergo a questionnaire on the diagnosis of lung cancer according to the protocol for diagnosing this disease.

The Server module stores data from the expert system: symptoms, diseases, diagnoses, and survey results.

The Mathematical Module implements operations for building a neural network.

In the "Application" module, the probability of exposure to lung cancer is predicted based on the selected models.

In the "Expert-LPR" module, an oncologist evaluates the results obtained with a high risk of lung cancer.

In the "Instrumental Laboratory studies" module, respondents who have received high results in assessing their susceptibility to lung cancer undergo ultrasound, computed tomography, and Cyfra 21 cancer marker analysis.

In the "Withdrawal Module", in case of positive results of instrumental laboratory tests, the patient is registered with an oncological dispensary, in case of a negative result of instrumental laboratory tests, the patient is referred to another specialist to clarify the diagnosis.

Oncological diseases are one of the key problems of modern medicine and society. Despite advances in healthcare, early cancer detection rates remain low. Therefore, the creation of systems aimed at early diagnosis of oncological diseases is of high practical importance and relevance.

Modern artificial intelligence methods provide extensive opportunities for predicting and processing medical information. In this study, a system for early detection of lung cancer based on machine learning technologies has been developed. To implement the system, a diagnostic questionnaire was created in accordance with existing medical protocols for examining patients. The obtained personal data was processed using eight machine learning algorithms. The analysis of the results confirmed the effectiveness of using machine learning methods to predict the likelihood of developing lung cancer.

7. Conclusion

Oncological diseases continue to be one of the most pressing medical and social problems. Despite the constant development of medical technologies, the level of cancer diagnosis in the early stages is still insufficient. In this regard, the development of intelligent systems designed for early detection of oncological diseases is of particular importance.

The use of artificial intelligence technologies and machine learning methods can improve the efficiency of medical data analysis and the accuracy of disease prediction. As part of this study, a system for early diagnosis of lung cancer was created. The basis of the system is a special diagnostic questionnaire developed on the basis of medical protocols for examining patients. The data was processed and analyzed using eight machine learning algorithms. The results of the study showed that machine learning methods can be successfully used to assess the likelihood of developing lung cancer and identify risk groups in the early stages of the disease.

Funding: This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP23485656).

References

1. Bray F. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries // CA. Cancer J. Clin. 2024. T. 74, № 3. S. 229–263.
2. Statistika raka, 2024 - Siegel - 2024 - CA: Zhurnal po raku dlya klinitsistov - Wiley Online Library [Ehlektronnyi resurs]. URL: https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21820 (data obrashcheniya: 22.12.2025).
3. Koning H. J. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial // N. Engl. J. Med. Massachusetts Medical Society, 2020. T. 382, № 6. S. 503–513.
4. The National Lung Screening Trial Research Team. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening // N. Engl. J. Med. Massachusetts Medical Society, 2011. T. 365, № 5. S. 395–409.
5. Thanoon M. A. A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images // Diagnostics. 2023. T. 13, № 16.
6. Abdullahi K. Deep Learning Techniques for Lung Cancer Diagnosis with Computed Tomography Imaging: A Systematic Review for Detection, Segmentation, and Classification // Information. 2025. T. 16, № 6.
7. A.r B. A Deep Learning-based Lung Cancer Classification of CT Images using Augmented Convolutional Neural Networks // ELCVIA Electron. Lett. Comput. Vis. Image Anal. 2022. T. 21, № 1.
8. Liz-López H. Deep Learning Innovations in the Detection of Lung Cancer: Advances, Trends, and Open Challenges // Cogn. Comput. 2025. T. 17, № 2. S. 67.
9. Murchison J. T. Validation of a deep learning computer aided system for CT based lung nodule detection, classification, and growth rate estimation in a routine clinical population // PLOS ONE. Public Library of Science, 2022. T. 17, № 5. S. e0266799.
10. Mahajan A. A Novel Deep Learning-Based (3D U-Net Model) Automated Pulmonary Nodule Detection Tool for CT Imaging // Curr. Oncol. 2025. T. 32, № 2.
11. Faizi M. K. Deep learning-based lung cancer classification of CT images // BMC Cancer. 2025. T. 25, № 1. S. 1056.
12. Salama W. M., Shokry A., Aly M. H. A generalized framework for lung Cancer classification based on deep generative models // Multimed. Tools Appl. 2022. T. 81, № 23. S. 32705–32722.
13. Ma L. A novel fusion algorithm for benign-malignant lung nodule classification on CT images // BMC Pulm. Med. 2023. T. 23, № 1. S. 474.
14. Wang X., Sharpnack J., Lee T. C. M. Improving lung cancer diagnosis and survival prediction with deep learning and CT imaging // PLOS ONE. Public Library of Science, 2025. T. 20, № 6. S. e0323174.
15. Li R. Multi-Task Deep Learning for Lung Nodule Detection and Segmentation in CT Scans: A Review // Electronics. 2025. T. 14, № 15.
16. Jozi N. S. LCxNet: An Explainable CNN Framework for Lung Cancer Detection in CT Images Using Multi-Optimizer and Visual Interpretability // Appl. Syst. Innov. 2025. T. 8, № 5.
17. Cellina M. Artificial Intelligence in Lung Cancer Screening: The Future Is Now // Cancers. 2023. T. 15, № 17.
18. Huang D. Artificial intelligence in lung cancer: current applications, future perspectives, and challenges // Front. Oncol. Frontiers, 2024. T. 14.
19. Jena S. R., George S. T. Morphological feature extraction and KNG-CNN classification of CT images for early lung cancer detection // Int. J. Imaging Syst. Technol. 2020. T. 30, № 4. S. 1324–1336.

20. Ardila D. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography // Nat. Med. Nature Publishing Group, 2019. T. 25, № 6. S. 954–961.
21. Salama W. M., Shokry A., Aly M. H. A generalized framework for lung Cancer classification based on deep generative models // Multimed. Tools Appl. 2022. T. 81, № 23. S. 32705–32722.
22. Shatnawi M. Q., Abuein Q., Al-Quraan R. Deep learning-based approach to diagnose lung cancer using CT-scan images // Intell.-Based Med. 2025. T. 11. S. 100188.
23. Crasta L. J., Neema R., Pais A. R. A novel Deep Learning architecture for lung cancer detection and diagnosis from Computed Tomography image analysis // Healthc. Anal. 2024. T. 5. S. 100316.
24. Gao C. Deep learning in pulmonary nodule detection and segmentation: a systematic review // Eur. Radiol. 2025. T. 35, № 1. S. 255–266.
25. Marinakis I. Pulmonary Nodule Detection, Segmentation and Classification Using Deep Learning: A Comprehensive Literature Review // BioMedInformatics. 2024. T. 4, № 3. S. 2043–2106.
26. UrRehman Z. Effective lung nodule detection using deep CNN with dual attention mechanisms // Sci. Rep. Nature Publishing Group, 2024. T. 14, № 1. S. 3934.
27. Almahasneh M., Xie X., Paiement A. AttentNet: Fully Convolutional 3D Attention for Lung Nodule Detection: arXiv:2407. 14464. arXiv, 2024.
28. An attentive and adaptive 3D CNN for automatic pulmonary nodule detection in CT image // Expert Syst. Appl. Pergamon, 2023. T. 211. S. 118672.
29. Faizi M. K. Deep learning-based lung cancer classification of CT images // BMC Cancer. 2025. T. 25, № 1. S. 1056.
30. Zhao H. Non-invasively Discriminating the Pathological Subtypes of Non-small Cell Lung Cancer with Pretreatment 18F-FDG PET/CT Using Deep Learning // Acad. Radiol. Elsevier, 2024. T. 31, № 1. S. 35–45.
31. Quanyang W. Artificial intelligence in lung cancer screening: Detection, classification, prediction, and prognosis // Cancer Med. 2024. T. 13, № 7. S. e7140.
32. Huang D. Artificial intelligence in lung cancer: current applications, future perspectives, and challenges // Front. Oncol. Frontiers, 2024. T. 14.
33. Wang L. Deep Learning Techniques to Diagnose Lung Cancer // Cancers. 2022. T. 14, № 22.
34. Karymsakova I.; Kozhakhmetova D.; Bekenova D.; Ostroukh D.; Bekbayeva R.; Kydyralina L.; Bugubayeva A.; Kurushbayeva D. Development of an Early Lung Cancer Diagnosis Method Based on a Neural Network. Computers 2025, 14, 397. https://doi.org/ 10.3390/computers14090397
35. Karymsakova I., Kozhakhmetova D.; Bekenova,D.; Lili Nurliyana Abdullah; Zolotov A.; Shyrynkhanova D.; Kydyralina, L.; Bugubayeva,A. Application of Machine Learning Methods for Predicting Susceptibility to Lung Cancer and Identifying Predictors Influencing the Risk of Lung Cancer Development. Computers 2026, 15(6), 365; https://doi.org/10.3390/computers15060365