UDC 004.8
IRSTI 28.23.15

https://doi.org/10.55452/1998-6688-2024-21-4-32-44

[1]Nam D.,
Master of Tech. Sci., PhD Student, ORCID ID:0000-0002-9356-3114,
e-mail: d.nam@kbtu.kz

[1]Kazakh-British Technical University, Almaty, Kazakhstan

CLASSIFICATION OF LUNG CALCIFICATIONS AND CANCER IN LUNGS-RADS SYSTEM BASED ON RADIOLOGICAL FEATURES

Abstract

Lung cancer represents a significant health challenge both in Kazakhstan and globally, standing out as one of the most fatal forms of cancer. Diagnosis of lung cancer is challenging as symptoms often remain undetectable in the early stages. Furthermore, lung cancer shares clinical features with various other pulmonary conditions, complicating its accurate identification. Accurate diagnosis typically involves lung puncture for subsequent biopsy, a highly invasive and painful procedure for the patient. Therefore, it is crucial to distinguish false positive cases in the diagnostic stage of computed tomography scans. We conducted a comparative analysis of five machine learning models (Logistic Regression, Decision Tree, Random Forest, SVM, and Naïve Bayes Algorithms) based on radiological features extracted from annotated computed tomography scans. We opted for classical machine learning methods because their decision-making process is easier to control compared to neural networks. We evaluated the models in terms of binary and multi-class classification to determine whether a given nodule is related to calcifications or cancers, as well as its classification according to Lung-RADS, enabling the management of whether further biopsy or only routine monitoring is necessary. We used Precision to evaluate the number of False Positive predictions in the binary classification task. Precision emerged as a pivotal metric in our assessment, offering insights into the number of false positive predictions specifically in the binary classification task. For the multi-class classification aspect, we turned to Quadratic Kappa, a robust measure that accounts for the ordinal nature of the Lung-RADS classes. Our analysis was underpinned by a combination of local Kazakhstani data and the publicly available LIDC-IDRI dataset, underscoring our commitment to leveraging diverse data sources to bolster diagnostic capabilities.

Key words: lung cancer classification, radiological feature extraction, ordinal data, medical image processing, computer vision, machine learning.

Introduction

Lung cancer stands as the deadliest form of cancer in Kazakhstan and all over the world [1]. The diagnosis of lung cancer is complicated by the fact that it is difficult to identify in its early stages and can resemble other pulmonary diseases. Additionally, symptoms of lung cancer often do not manifest in the early stages, further complicating the diagnostic process. However, the late onset of lung cancer symptoms is not only one reason of is not the only reason for the difficulty of diagnosis. Lung cancer shares similar signs and symptoms with several other conditions. The diagnostics of lung cancer are separated into three main steps. First, the abnormal area is indicated on the X-ray image. The next step is computed tomography. In cases of suspected lung cancer, a biopsy is often recommended. Biopsy procedures are highly invasive and carry a number of negative health consequences. Therefore, reducing the number of false positive results is an important task in medical data processing. In this current article, we explored the potential application of clustering based on the depth and area of nodules for both calcifications and tumors according to the Lung-RADS system.

Literature review

Convolutional neural networks are widely used for lung cancer classification on CT images. The authors of Lung-EffNet [2] proposed a transfer learning framework based on EfficientNet architecture [3] for the classification of lung CT according to normal, adenocarcinoma, large cell carcinoma, and squamous cell carcinoma. The authors of the next observed article proposed DL-CAD [4] for the detection of missed lung cancer after CT screening. The algorithm utilized the DenseNet [5] architecture for image classification into cancer and non-cancer classes. The authors exclusively examined cases that were missed during the previous CT screening to evaluate the DL-CAD. LR3, according to Lung-RADS, was employed as a positive indicator. Another way of lung cancer detection is an application of Generative adversarial neural networks for cancer detection. The authors of MTL-MGAN [6] proposed an application of a modified generative adversarial network (MGAN) and transfer learning. MGAN has been used to create is to create two intermediary domains that act as connectors between the source and target domains which allows an increase in lung cancer classification quality. The authors of the next observed article [7] proposed an application of an Optimized Ensemble of Hybrid RNN-GAN Models for lung nodule classification for cancer and non-cancer cases.

Although deep neural networks demonstrate high performance in medical image processing, their application is complicated by several factors:

1. The need for significant computational resources. Generally, networks with a large number of training parameters yield better results [8].
2. The requirement for a large volume of training data is challenging in medical data processing tasks due to ethical and privacy concerns. Additionally, medical data needs to be pre-annotated by a team of experienced clinician experts.
3. The training process of the model acts as a black box, making it difficult to control the decision-making process of the neural network. For example, in the paper [9], the process is described where a neural network mistakenly identified images with band-aids as skin diseases. Another example illustrating that it is impossible to precisely interpret the decision-making process of the neural network from the expert side is described in papers [10], [11]. The authors of these papers demonstrate that the neural network learned to distinguish the race of a person from an X-ray image.

Due to these limitations, classical machine learning methods are still widely used in computer vision for medical data processing. The authors of [12] proposed an application of Random Forest for lung cancer classification. The authors used an open-source LIDC-IDRI dataset [13] with lung CT images. The authors applied median and Gaussian filters to remove noise from the original image. Then the authors applied a watershed algorithm [14] for nodule segmentation. The nodule was used for the extraction of radiological features, such as area, eccentricity, mean intensity, centroid, and diameter. The following values were used as input for the Random Forest classifier (RF) [15]. The authors of [16] also used an RF for lung nodule classification. The model used the output of the CAD system for lung cancer detection as an input for an improved Random Forest classifier. RF was improved by updating the sampling and feature selection process for better performance with imbalanced data. The authors of [17] applied a support vector machine (SVM) for nodule classification. Otsu thresholding-based algorithm [18] was used for nodule segmentation. Gabor filter [19] and Gray-Level Co-occurrence Matrix (GLCM) [20] were used for feature extractions. Extracted features were fed to SVM for the next classification. Classical machine learning algorithms also were compared with each other in [21]. The authors compared Bayesian Network, Logistic Regression, J48, Random Forest, and Naïve Bayes Algorithms for binary classification on open source Kaggle dataset with 309 observations and 16 attributed.

For the current research, we also applied classical machine-learning approaches for lung nodule classification. We provide a comparative analysis of five classical machine-learning algorithms (Logistic Regression, Decision Tree, Random Forest, SVM, and Naïve Bayes) based on the

radiological features of lesions for multiclass classification according to 5 classes: 4 cancer classes according to the Lung-RADS System and one non-cancer class (calcification).

Main provision

The main provision of the research could be described as:

1. Extraction of radiological features from the original DICOM image based on Pearson [22] and Spearman [23] correlation.
2. Classification between lung cancer and lung calcification instead of classification between cancer and non-cancer areas.
3. The use of the Lung-RADS system for cancer nodule classification.
4. Comparative analysis for binary (cancer, non-cancer) and multi-class (calcification and Lung-RADS classes) classification of five classical ML algorithms.
5. The use of Kazakstani local data allows to take into account economic and environmental specification of the region.

We compared algorithms based on Accuracy, Precision, Recall, and F1 for binary and multi-class classification, and also we compared the number of False Positive and False Negative Predictions. Additionally, we used Quadratic Kappa for the calculation the quality of multi-class classification based on ordinal data.

Methods and Materials

We used the dataset with lung cancer CT images of Kazakhstani patients with corresponding binary masks [24] and supplemented it via calcinate cases. We used a preliminary segmentated area as an input of the pipeline. Then we extracted radiological features, such as mean, mean of positive values, the mean value of the circle described around the centroid, square, and square of positive elements only. We used these features for classification according to 5 classes: four classes according to the Lung-RADS system and calcinate (non-cancer class). We compared Logistic Regression, Decision Tree, Random Forest, SVM, and Naïve Bayes Algorithms for classification. Fig 1. shows the overview of the pipeline for lung lesion classification.

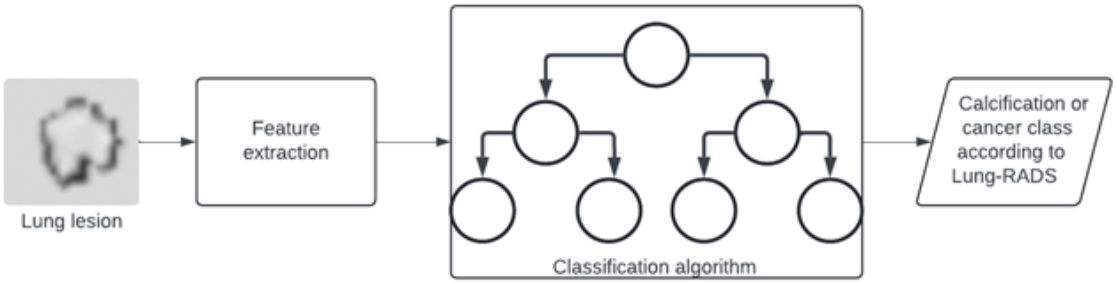

Figure 1 – Pipeline of lung lesion classification

Dataset description

We supplemented the dataset with lung cancer CT images via calcification cases. The total number of images is 1134. The number of images with lung cancer is 972, and the number of images with calcification is 162. The dataset consists of a CT image, corresponding binary mask, class according to Lung-RADS system or calcification, and the binary value of cancer existence (0 for calcinates, 1 for cancer). The distributions between cancer and label classes are shown in Fig 2. We worked with imbalanced data, predominantly consisting of positive cases.

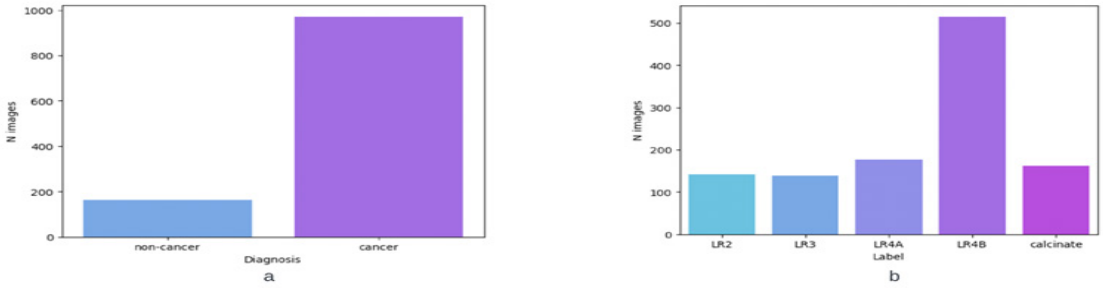

Figure 2 – The distribution between (a) cancer and non-cancer classes (b) classes according to Lung-RADS and calcification

The dataset had two target functions. The first one is the label. It is the class name according to the Lung-RADS system that includes 4 classes and calcification. A brief description of each class has been provided in Table 1.

Table 1 – Dataset label description

Name	N samples	Description
Calcification	162	Calcification refers to the accumulation of calcium salts in tissues or organs, often observed as white spots on medical imaging such as X-rays or CT scans. It can occur in various parts of the body and may indicate underlying conditions such as infections, trauma, or metabolic disorders.
LR2	142	LR2, or Lung-RADS category 2, is a classification system used in lung cancer screening to categorize pulmonary nodules as benign based on specific imaging features. Nodules in this category typically have low suspicion for malignancy and require routine surveillance.
LR3	138	LR3, or Lung-RADS category 3, is a classification used in lung cancer screening to identify nodules with intermediate suspicion for malignancy. These nodules may require additional imaging or follow-up to assess for changes over time.
LR4A	177	LR4A, or Lung-RADS category 4A, indicates nodules with a moderate level of suspicion for malignancy. These nodules often require further evaluation, such as biopsy or PET-CT imaging, to determine if they are cancerous.
LR4B	515	LR4B, or Lung-RADS category 4B, represents nodules with a high suspicion for malignancy. These nodules are more likely to be cancerous and typically warrant prompt evaluation and management, such as biopsy or surgical resection.

The calcification and cancer areas have big differences in density distribution but could have similar forms and locations in the lungs. The calcification has much more density in Hounsfield Units (HU) which is shown in Fig 3.

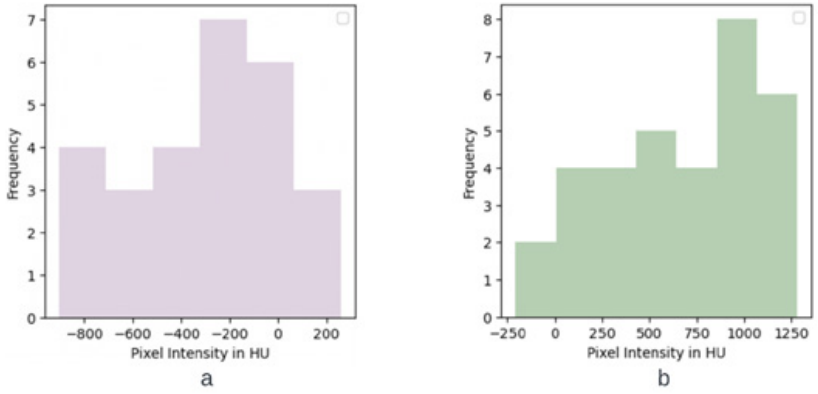

Figure 3 – Histogram of the density distribution of HU of (a) LR2 (b) calcification from the dataset for one CT image

Feature extraction

First, we calculated the Region of Interest (RoI) by multiplication the binary mask with the CT image, as in (1). This RoI will be used for all of the next calculations as an affected area.

$$RoI_{i,j} = CTimg_{i,j} \times mask_{i,j} \quad (1)$$

Where $CTimg$ is an original CT image,
$mask$ is a binary mask of cancer or calcification area,
i, j are indexes

The density is the key feature for classification between cancer and calcification. So first we calculated three mean values of the RoI: Average mean (2), Mean of positive (3).

$$Mean_{avg} = \frac{\sum_{i,j} CTimg_{i,j} \times mask_{i,j}}{\sum_{i,j} mask_{i,j}} \quad (2)$$

$$Mean_{pos} = \frac{\sum_{i,j} CTimg_{i,j} \times mask_{i,j} \times (CTimg_{i,j} > 0)}{\sum_{i,j} mask_{i,j} \times (CTimg_{i,j} > 0)} \quad (3)$$

Where $CTimg$ is an original CT image,
$mask$ is a binary mask of cancer or calcification area,
i, j are indexes

Additionally, we tried to simulate the approach the clinicians used. We calculate a mean value of the area near the centroid, as in (4) – (7)

$$R = round \frac{\sqrt{\frac{\sum mask}{\pi}}}{3} \quad (4)$$

Where R is the Radius of the area which will be used for mean calculation,
$mask$ is a binary mask of cancer or calcification area

$$(x_c, y_c) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i , \frac{1}{n} \sum_{i=1}^{n} y_i \right) \quad (5)$$

Where (x_c, y_c) are the coordinates of the centroid,
n is the number of vertices,
(x_i, y_i) are the coordinates of the i-th vertex

$$Circle_{i,j} = \begin{cases} 1, if\ (i - x_c)^2 + (j - y_c)^2 \leq R \\ 0 \end{cases} \quad (6)$$

Where $Circle$ is the binary mask with circle area which will be used for the next calculation
(x_c, y_c) are the coordinates of the centroid
(i, j) are the pixel coordinates

$$Mean_{round} = \frac{\sum_{i,j} CTimg_{ij} \times Circle_{i,j}}{\sum_{i,j} Circle_{i,j}} \quad (7)$$

Where $CTimg$ is an original CT image,
$Circle$ is the binary mask with circle

While density is the main difference between cancer and calcification, classes according to the Lung RAGS System have differences in size. We proposed to use the square of the affected area, as in (8), and the square of positive pixels from the affected area (9).

$$S_{all} = \sum mask \quad (8)$$

$$S_{pos} = \sum_{i,j} CTimg_{i,j} \times (CTimg_{i,j} > 0) \quad (9)$$

After the following calculations, we received an updated dataset with the radiological features described the lung cancer. The overview of the dataset has been provided in Table 2.

Table 2 – Updated dataset description

Name	Type	Description
Category id	ID	Unique ID of one patient
Mean	Real	The mean value of affected are in HU
Mean positive	Real non-negative	The mean value of positive pixels in affected are in HU
Mean round	Real	The mean value of area near the centroid of the affected are
S	Natural positive	The square of affected area
S positive	Natural non-negative	The square of positive pixels in affected area
Label	[0: calcinate, 1: LR2, 2: LR3, 3: LR4A, 3: LR4B]	The class with ordering according to Lung-RADS System or calcification
Cancer	[1: cancer, 0: non-cancer]	Cancer for all LRs, non-cancer for calcification

To ensure that all obtained features have an impact on the target variables, we calculated correlations between them and the target function. Since cancer and non-cancer are independent classes, we used Pearson correlation, as in (10), to calculate the correlation between variables and "cancer". The "label" values are ordinal. 0 denotes the absence of cancer, 1 denotes small areas, and so forth. Therefore, we applied Spearman correlation, as in (11). Fig 4. Shows the correlation between features and target variables.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (10)$$

Where r_{xy} is the correlation coefficient between variables x and y,
x_i and y_i are the values of variables x and y for the i-th observation,
$\bar{x}$ and $\bar{y}$ are the means of variables x and y,
n is the number of observations.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \quad (11)$$

Where d_i is the difference between the ranks of corresponding values in the two variables.
n is the number of observations.

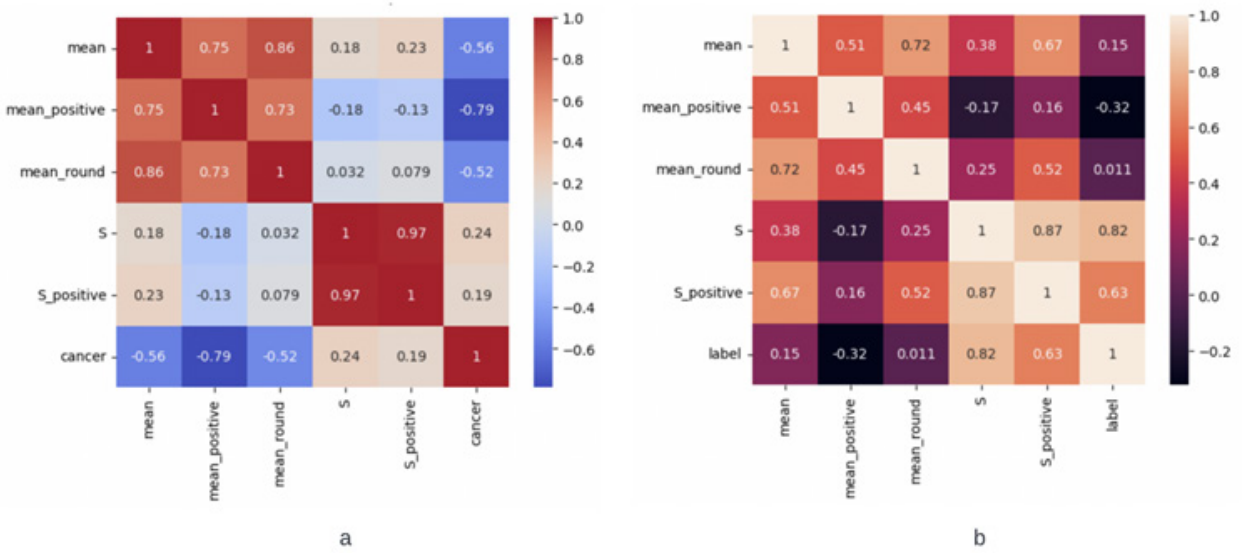

Figure 4 – Correlation matrixes (a) Pearson correlation between radiological features and cancer (b) Spearman correlation between radiological features and label

We randomly split the dataset to train and test sets. However, we worked with CT slices with calcinate or cancer. So the dataset contains continuous slices from one patient. So we split data by category ID to avoid the situation, when the CT images from one patient are in the train and test sets at the same time. We worked with imbalanced data with the prevailing cancer class. The distribution between classes in train and test sets is shown in Fig. 5.

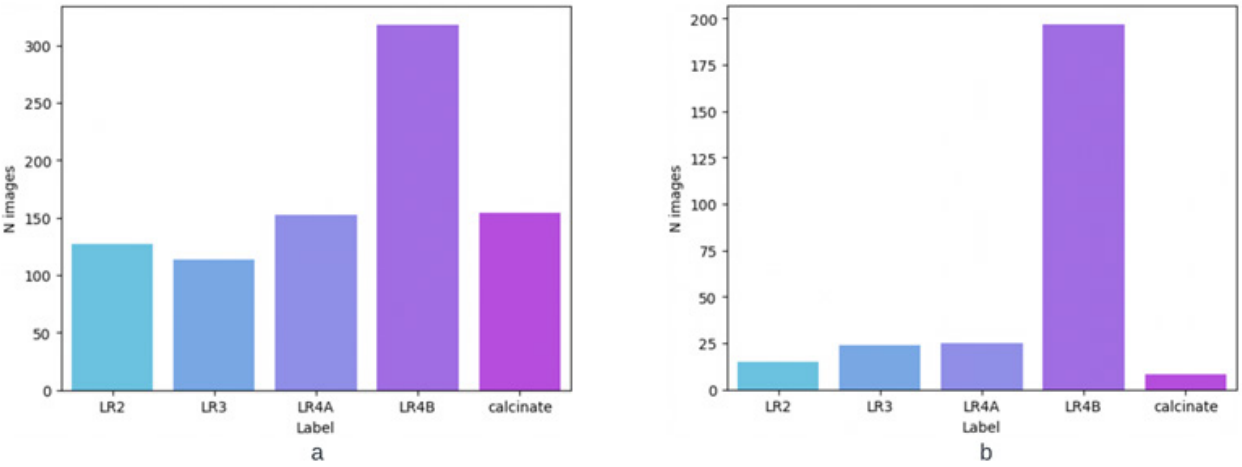

Figure 5 – The distribution according to the number
of images per label in the dataset (a) in the train set (b) in the test set

Machine learning models for lesion classification

We compared several commonly used machine learning algorithms to analyze radiological features of the lung, including logistic regression, decision tree, random forest, support vector machine (SVM), and Naïve Bayes classifier.

Logistic regression, although a linear algorithm, has demonstrated good performance in binary classification problems that require predicting the probability of belonging to a particular class. It also has the advantage of interpretable results, making it useful in clinical research.

A decision tree is a nonlinear classification algorithm that allows the construction of a tree structure where each node represents a decision based on a feature. This method is easy to interpret and can handle both numeric and categorical data.

Random forest is an ensemble method consisting of multiple decision trees. It creates a "forest" of trees trained on different subsets of data, which improves classification quality and reduces the risk of overfitting.

Support Vector Machine (SVM) is well suited for classification problems with linear and nonlinear relationships between features. It finds the optimal hyperplane separating classes, making it effective in solving complex classification problems.

The Naïve Bayes classifier, despite its simplicity, performs well in real-world applications, especially text classification and spam filtering. It is based on the assumption of feature independence, which makes it computationally efficient and easy to implement.

Metrics for model evaluation

We compared the models according to binary (cancer, non-cancer) and multi-class classification (Lung-RADS classes and calcification). We used standard metrics for the evaluation of classification, such as Accuracy, as in (12), Precision, as in , Recall, and F1 score. As we work with imbalanced data in the medical image processing field, the model with the minimum number of False Positive predictions is the most applicable for us. In case the number of False Positive decreases, the precision grows up.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (12)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (13)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (14)$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (15)$$

Where TP (True Positive) is the number of correctly predicted positive instances,
TN (True Negative) is the number of correctly predicted negative instances,
FP (False Positive) is the number of incorrectly predicted positive instances,
FN (False Negative) is the number of incorrectly predicted negative instances.

We used the same metrics for multi-class and binary classification, however, multi-class classification has been ordered, as calcification is accurate for non-cancer class, LR2 is a small and less dangerous lesion, and so on. So we additionally calculated the Quadratic Kappa for classification with ordical data, as in (16).

Where κ_w is the weighted quadratic kappa,

W_{ij} are the elements of the error weight matrix,

N_{ij} is the number of observed agreements between class i and class j,

N_{ij}^{exp} is the expected number of agreements between class i and class j that would occur by chance.

Results and discussion

The results of binary classification for cancer and non-cancer target values have been provided in Table 3. The confusion matrix for binary classification has been provided in Fig. 6.

Table 3 – Evaluation of binary classification among cancer and non-cancer

Model name	Accuracy	Precision	Recall	F1
Logistic Regression	0.9962	0.9962	1	0.998
Decision Tree	0.9888	0.9886	1	0.9942
Random Forest	0.9925	0.9923	1	0.9961
SVM	0.9925	0.9923	1	0.9961
Naïve Bayes	0.9776	0.9961	0.9808	0.9884

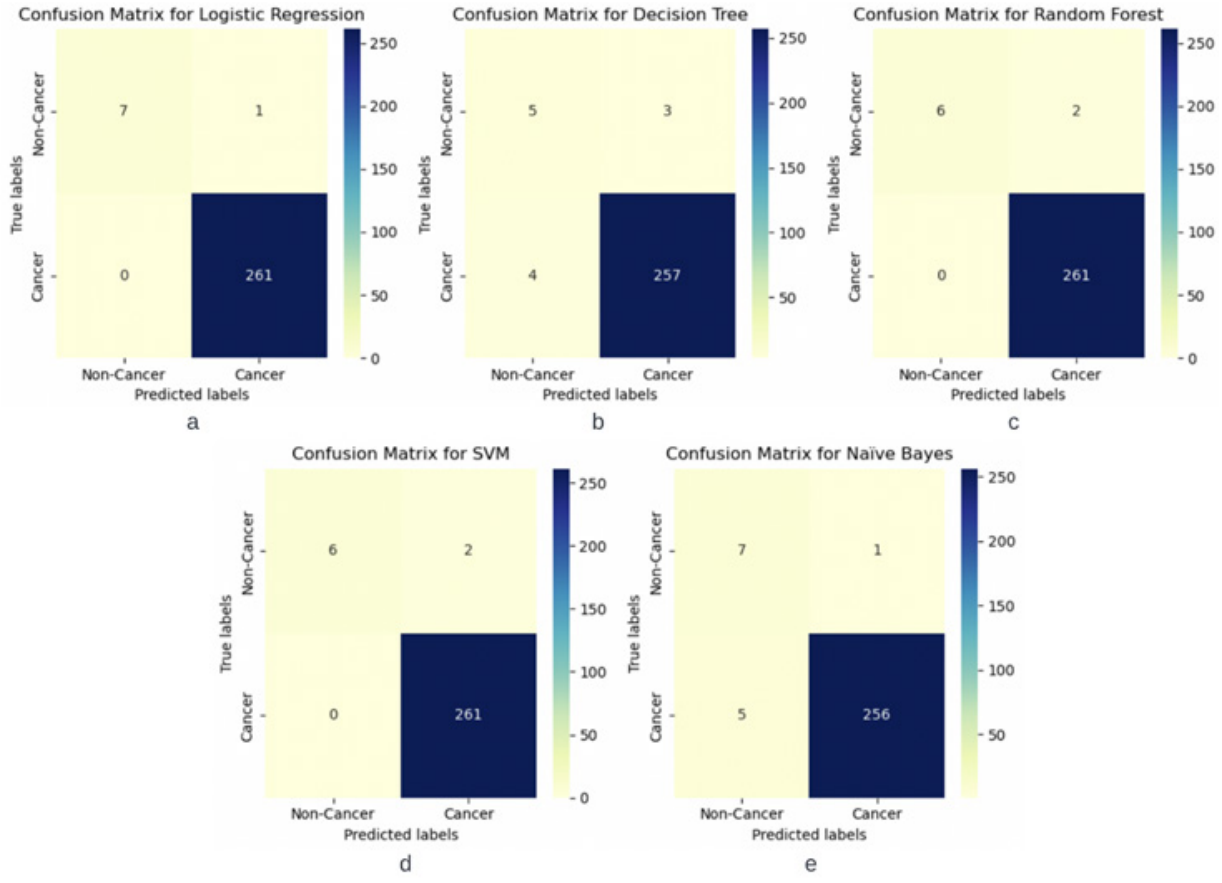

Figure 6 – Confusion matrixes for binary classification according to cancer and non-cancer cases for (a) Logistic Regression (b) Decision Tree (c) Random Forest (d) SVM (e) Naïve Bayes

The results of multi-class classification according to Lung-RADS and calcification with target values label (classes according to Lung-RADS and calcification) have been provided in Table 4. Fig. 7 shows the confusion matrix for all of the classes.

Table 4 – Evaluation of multi-class classification among calcification and classes in Lung-RADS

Model name	Accuracy	Precision	Recall	F1	Quadratic Kappa
Logistic Regression	0.6765	0.5391	0.6088	0.5558	0.7432
Decision Tree	0.684	0.5767	0.5922	0.5654	0.7554
Random Forest	0.684	0.5665	0.5899	0.5562	0.746
SVM	0.6579	0.5737	0.6593	0.5823	0.6653
Naïve Bayes	0.5501	0.5536	0.5313	0.4628	0.6254

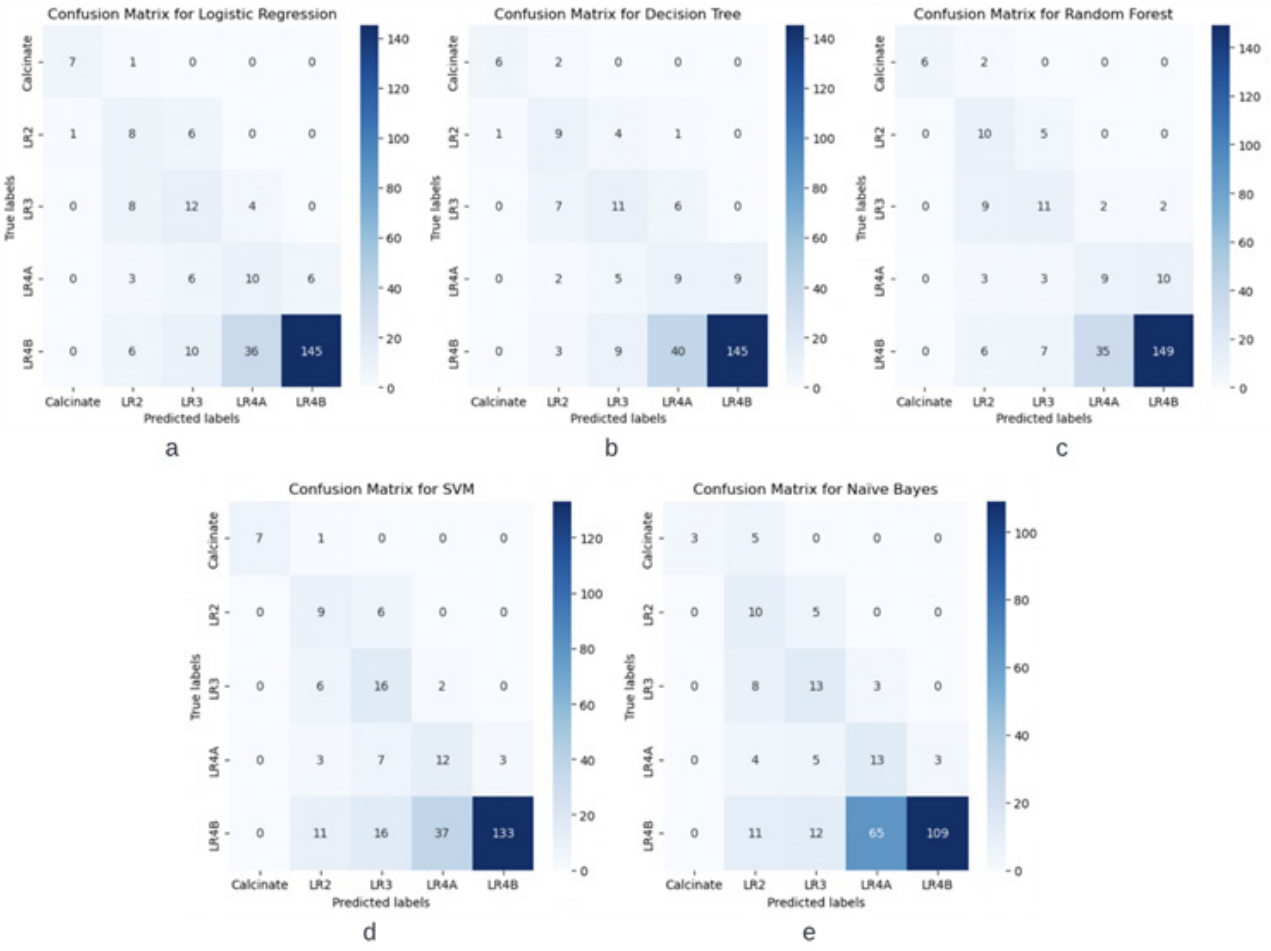

Figure 7 – Confusion matrixes for multi-class classification according to Lung-RADS classes and calcification for (a) Logistic Regression (b) Decision Tree (c) Random Forest (d) SVM (e) Naïve Bayes

We worked with data preliminarily labelled by doctor-clinicians according to the Lung-RADS System and calcification. The dataset also stored binary masks of damaged areas and CT scan values in Hounsfield units. We used annotated binary masks and their corresponding images to obtain regions of interest. Regions of interest represent areas affected by cancerous formations or calcifications. The obtained regions of interest were examined to extract radiological features characterizing the area, such as mean, mean positive, mean around centroid, area, and positive area.

We conducted a comparative analysis of five classical machine learning models working with vector data based on the obtained radiological features, mimicking the work of a clinician. We

compared Logistic Regression, Decision Tree, Random Forest, SVM, and Naïve Bayes in the context of binary and multi-class classification tasks based on Accuracy, Precision, Recall, and F1 for both classification tasks and supplemented with Quadratic Kappa for the multi-class classification task.

We separately solved the problem of binary classification according to cancer or non-cancer, where calcinate was considered non-cancerous. And the multi-class classification tasks according to the Lung RADS classes and calcifications. We consider Precision metric as the most effective for evaluating the quality of binary classification since its increase signifies a reduction in the number of false positive predictions, which is particularly important in medical tasks, especially when dealing with imbalanced data. Thus, logistic regression performed the best in the binary classification task. In the multi-class classification task for ordinal data, we find the Quadratic Kappa metric most suitable as it takes into account the class order. According to our experiments, the decision tree performed the best.

Several limitations are associated with this study. Firstly, we worked with a limited dataset that only included two possible types of anomalies: cancer and calcification. However, many other different lesions have a similar shape and location. Also, the presence of calcifications in the lungs may indicate a potentially existing condition. Secondly, increasing the size of the train set could significantly increase the quality of classification. Last but not least, we take into account only radiological features, but the doctor takes into account clinical features as well. However, we do not have this information because of ethical and privacy reasons.

Conclusion

Our study leveraged radiological features extracted from CT scans, alongside binary masks indicating areas of interest, to explore the performance of classical machine learning models in classifying lung abnormalities. The results indicated that logistic regression excelled in binary classification, demonstrating its potential for identifying cancerous and non-cancerous regions. Furthermore, the decision tree model showcased superior performance in multi-class classification tasks, particularly with ordinal data.

REFERENCES

1 Ferlay J., Ervik M., Lam F., et al. Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer, 2022. Available from: https://gco.iarc.fr/today.

2 Raza R., Zulfiqar F., Khan M. O., Arif M., Alvi A., Iftikhar M. A., & Alam T. Lung-EffNet: Lung cancer classification using EfficientNet from CT-scan images. Engineering Applications of Artificial Intelligence, 2023, vol. 126, p. 106902.

3 Tan M., & Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning, 2019, May, pp. 6105–6114. PMLR.

4 Cho J., Kim J., Lee K. J., Nam C.M., Yoon S.H., Song H. ... & Lee K.W. Incidence lung cancer after a negative ct screening in the national lung screening trial: Deep learning-based detection of missed lung cancers. Journal of Clinical Medicine, 2020, vol. 9, no. 12, p. 3908.

5 Huang G., Liu Z., Van Der Maaten L., & Weinberger K.Q. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

6 Chui K.T., Gupta B.B., Jhaveri R.H., Chi H.R., Arya V., Almomani A., & Nauman A. Multiround transfer learning and modified generative adversarial network for lung cancer detection. International Journal of Intelligent Systems, 2023, pp. 1–14.

7 Tiwari A., Hannan S.A., Pinnamaneni R., Al-Ansari A.R.M., El-Ebiary Y.A.B., Prema S. ... & Vidalón J.L.J. Optimized Ensemble of Hybrid RNN-GAN Models for Accurate and Automated Lung Tumour Detection from CT Images. International Journal of Advanced Computer Science and Applications, 2023, vol. 14, no.7.

8 Götz T.I., Göb S., Sawant S., Erick X.F., Wittenberg T., Schmidkonz C. ... & Ramming A. Number of necessary training examples for Neural Networks with different number of trainable parameters. Journal of Pathology Informatics, 2022, no.13, p. 100114.

9 Goel K., Gu A., Li Y., & Ré C. Model patching: Closing the subgroup performance gap with data augmentation, 2020, arXiv preprint arXiv:2008.06775.

10 Banerjee I., Bhimireddy A.R., Burns J.L., Celi L.A., Chen L.C., Correa R. ... & Gichoya J.W. Reading race: AI recognises patient's racial identity in medical images, 2021, arXiv preprint arXiv:2107.10356.

11 Gichoya J.W., Banerjee I., Bhimireddy A.R., Burns J.L., Celi L.A., Chen L.C. ... & Zhang H. (2022). AI recognition of patient race in medical imaging: a modelling study. The Lancet Digital Health, vol.4, no. 6, e406-e414.

12 Jayaraj D., & Sathiamoorthy S. Random forest based classification model for lung cancer prediction on computer tomography images. In 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), 2019, November, pp. 100–104. IEEE.

13 Armato III S.G., McLennan G., Bidaut L., et al. Data From LIDC-IDRI [Data set]. The Cancer Imaging Archive, 2015. https://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX.

14 Beucher S., & Meyer F. The morphological approach to segmentation: the watershed transformation. In Mathematical morphology in image processing, 2018, pp. 433–481).

15 Breiman L. Random forests. Machine learning, 2001, no. 45, pp. 5–32.

16 Paing May Phu, and Somsak Choomchuay. Improved random forest (RF) classifier for imbalanced classification of lung nodules. 2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST). IEEE, 2018.

17 Kareem H.F., AL-Husieny M.S., Mohsen F.Y., Khalil E.A., & Hassan Z.S. Evaluation of SVM performance in the detection of lung cancer in marked CT scan dataset. Indonesian Journal of Electrical Engineering and Computer Science, 2021, vol. 21, no. 3, p. 1731.

18 Otsu N. A threshold selection method from gray-level histograms. Automatica, 1975, vol. 11, no. 285–296, pp. 23–27.

19 Gabor D. Theory of communication. Part 1: The analysis of information. Journal of the Institution of Electrical Engineers-part III: radio and communication engineering, 1946, vol. 93, no. 26, pp. 429–441.

20 Haralick R.M., Shanmugam K., & Dinstein I.H. Textural features for image classification. IEEE Transactions on systems, man, and cybernetics, 1973, no. 6, pp. 610–621.

21 Cripsy J. Viji, and Divya T. Lung Cancer Disease Prediction and Classification based on Feature Selection method using Bayesian Network, Logistic Regression, J48, Random Forest, and Naïve Bayes Algorithms. 2023 3rd International Conference on Smart Data Intelligence (ICSMDI). IEEE, 2023.

22 Pearson K. VII. Note on regression and inheritance in the case of two parents. proceedings of the royal society of London, 1895, vol. 58, no. 347–352, pp. 240–242.

23 Spearman C. The proof and measurement of association between two things, 1961.

24 Nam D., Panina A., Pak A. Lung cancer segmentation dataset with Lung-RADS class, Mendeley Data, V1, doi: 10.17632/5rr22hgzwr.1

[1]Нам Д.,
т.ғ.м., докторант, ORCID ID: 0000-0002-9356-3114,
e-mail: d.nam@kbtu.kz

[1]Қазақстан-Британ техникалық университеті, Алматы қ., Қазақстан,

ӨКПЕДЕГІ КАЛЬЦИФИКАЦИЯЛАР МЕН ОБЫРДЫ LUNG-RADS ЖҮЙЕСІНДЕ РАДИОЛОГИЯЛЫҚ ЕРЕКШЕЛІКТЕР НЕГІЗІНДЕ ЖІКТЕУ

Аңдатпа

Өкпе обыры Қазақстанда және әлемде денсаулық сақтау саласындағы елеулі мәселелердің бірі. Бұл ауру өлімге әкелетін қатерлі ісіктер қатарында. Өкпе обырын ерте диагностикалау қиын, себебі оның бастапқы кезеңдерінде клиникалық белгілері байқалмайды. Сонымен қатар, өкпе обырының басқа өкпе ауруларымен ұқсас клиникалық көріністері оның дәл диагностикасын қиындатады. Дәстүрлі диагностикалық әдістер, мысалы, өкпені тесіп, биопсия жүргізу, инвазивті және науқас үшін ауыр процедуралар. Осыған байланысты компьютерлік томография (КТ) негізінде жалған оң жағдайларды азайту диагностика сапасын жақсартуда маңызды рөл атқарады. Бұл зерттеуде аннотацияланған компьютерлік томографиялардан алынған

радиологиялық ерекшеліктерге негізделген бес машинамен оқыту моделінің (логистикалық регрессия, шешім ағашы, кездейсоқ орман, векторларды қолдау әдісі және Байестің аңғал алгоритмі) салыстырмалы талдауы жүргізілді. Классикалық модельдерді таңдау олардың шешім қабылдау процесін нейрондық желілермен салыстырғанда жеңіл бақылауға болатындығымен түсіндіріледі. Модельдер бинарлы және көпкласты жіктеу тұрғысынан бағаланды. Бинарлы жіктеу барысында нақты түйіннің кальцификациялармен немесе обырмен байланысты екенін анықтау және биопсияның қажет екенін, тұрақты бақылаудың жеткілікті екенін шешу үшін Precision метрикасы қолданылды. Ал көпкласты жіктеу үшін Lung-RADS кластарын реттік сипатын ескеретін Quadratic Kappa сенімділік өлшемі пайдаланылды. Зерттеу жергілікті қазақстандық деректер мен жалпыға қолжетімді LIDC-IDRI деректер жиынтығының комбинациясына негізделген. Әртүрлі дерекқөздерді біріктіру диагностикалық мүмкіндіктерді кеңейтуге деген ұмтылысты көрсетеді.

Тірек сөздер: өкпе обырын жіктеу, радиологиялық ерекшеліктерді алу, реттік деректер, медициналық бейнелерді өңдеу, компьютерлік көру, машинамен оқыту.

[1]Нам Д.,
магистр техн. наук, PhD студент, ORCID ID: 0000-0002-9356-3114,
e-mail: d.nam@kbtu.kz

[1]Казахстанско-Британский технический университет, г. Алматы, Казахстан,

КЛАССИФИКАЦИЯ КАЛЬЦИФИКАЦИЙ И РАКА ЛЕГКОГО В СИСТЕМЕ LUNG-RADS НА ОСНОВЕ РАДИОЛОГИЧЕСКИХ ПРИЗНАКОВ

Аннотация

Рак легких представляет собой значительную проблему для здравоохранения как в Казахстане, так и в мире, являясь одной из самых смертельных форм рака. Диагностика рака легких сложна, так как симптомы часто остаются незаметными на ранних стадиях. Более того, рак легких имеет общие клинические признаки с различными другими легочными заболеваниями, что усложняет его точное выявление. Точная диагностика обычно требует прокола легкого для последующей биопсии, что является высокоинвазивной и болезненной процедурой для пациента. Поэтому крайне важно отличать ложноположительные случаи на этапе диагностики с использованием компьютерной томографии. Мы провели сравнительный анализ пяти моделей машинного обучения (логистическая регрессия, решающее дерево, случайный лес, метод опорных векторов и наивный байесовский алгоритм) на основе радиологических признаков, извлеченных из аннотированных компьютерных томографий. Мы выбрали классические методы машинного обучения, потому что их процесс принятия решений легче контролировать по сравнению с нейронными сетями. Мы оценили модели с точки зрения бинарной и многоклассовой классификации, чтобы определить, связано ли данное образование с кальцификацией или раком, а также его классификацию согласно Lung-RADS, что позволяет решить, требуется ли дальнейшая биопсия или достаточно только рутинного наблюдения. Мы использовали метрику Precision для оценки количества ложноположительных предсказаний в задаче бинарной классификации. Precision стал ключевой метрикой в нашей оценке, предоставляя информацию о количестве ложноположительных предсказаний именно в задаче бинарной классификации. Для аспекта многоклассовой классификации мы обратились к Quadratic Kappa, надежной мере, учитывающей порядковый характер классов Lung-RADS. Наш анализ основывался на комбинации местных казахстанских данных и общедоступного набора данных LIDC-IDRI, подчеркивая нашу приверженность использованию разнообразных источников данных для улучшения диагностических возможностей.

Ключевые слова: классификация рака легких, извлечение радиологических признаков, порядковые данные, обработка медицинских изображений, компьютерное зрение, машинное обучение.

Article submission date: 11.06.2024