ORIGINAL ARTICLE / ÖZGÜN MAKALE

Deep learning in distinguishing pulmonary nodules as benign and malignant

Pulmoner nodüllerin benign ve malign olarak ayrımında derin öğrenme

Muhammed Bilal Akıncı[1] ⓘ, **Mesut Özgökçe**[1] ⓘ, **Murat Canayaz**[2] ⓘ, **Fatma Durmaz**[1] ⓘ,
Sercan Özkaçmaz[1] ⓘ, **İlyas Dündar**[1] ⓘ, **Ensar Türko**[1] ⓘ, **Cemil Göya**[1] ⓘ

Institution where the research was done:
Van Yüzüncü Yıl University Faculty of Medicine, Van, Türkiye

Author Affiliations:
[1]Department of Radiology, Van Yüzüncü Yıl University Faculty of Medicine, Van, Türkiye
[2]Van Yüzüncüyıl University, Computer Engineering, Van, Türkiye

ABSTRACT

Background: Due to the high mortality of lung cancer, the aim was to find convolutional neural network models that can distinguish benign and malignant cases with high accuracy, which can help in early diagnosis with diagnostic imaging.

Methods: Patients who underwent tomography in our clinic and who were found to have lung nodules were retrospectively screened between January 2015 and December 2020. The patients were divided into two groups: benign (n=68; 38 males, 30 females; mean age: 59±12.2 years; range, 27 to 81 years) and malignant (n=29; 19 males, 10 females; mean age: 65±10.4 years; range, 43 to 88 years). In addition, a control group (n=67; 38 males, 29 females; mean age: 56.9±14.1 years; range, 26 to 81 years) consisting of healthy patients with no pathology in their sections was formed. Deep neural networks were trained with 80% of the three-class dataset we created and tested with 20% of the data. After the training of deep neural networks, feature extraction was done for these networks. The features extracted from the dataset were classified by machine learning algorithms. Performance results were obtained using confusion matrix analysis.

Results: After training deep neural networks, the highest accuracy rate of 80% was achieved with the AlexNET model among the models used. In the second stage results, obtained after feature extraction and using the classifier, the highest accuracy rate was achieved with the support vector machine classifier in the VGG19 model with 93.5%. In addition, increases in accuracy were noted in all models with the use of the support vector machine classifier.

Conclusion: Differentiation of benign and malignant lung nodules using deep learning models and feature extraction will provide important advantages for early diagnosis in radiology practice. The results obtained in our study support this view.

Keywords: Deep learning, lung cancer, solitary pulmonary nodule.

ÖZ

Amaç: Akciğer kanserinde mortalitenin yüksek olması nedeniyle benign ve malign olguları yüksek doğrulukla ayırt edebilen, tanısal görüntüleme ile erken tanıya yardımcı olabilecek evrişimsel sinir ağı modellerinin bulunması amaçlandı.

Çalışma planı: Kliniğimizde tomografisi çekilen ve akciğerinde nodül saptanan hastalar Ocak 2015 ve Aralık 2020 tarihleri arasında geriye dönük olarak tarandı. Hastalar iki gruba ayrıldı: benign (n=68; 38 erkek, 30 kadın; ort. yaş: 59±12.2 yıl; dağılım, 27-81 yıl) ve malign (n=29; 19 erkek, 10 kadın; ort. yaş: 65±10.4 yıl; dağılım, 43-88 yıl). Ayrıca kesitlerinde herhangi bir patoloji bulunmayan sağlıklı hastalardan oluşan bir kontrol grubu (n=67; 38 erkek, 29 kadın; ort. yaş: 56.9±14.1 yıl; dağılım, 26-81 yıl) oluşturuldu. Derin sinir ağları, oluşturduğumuz üç sınıflı veri setinin %80'i ile eğitildi ve verilerin %20'si ile test edildi. Derin sinir ağlarının eğitiminin ardından bu ağlardan özellik çıkarımı yapıldı. Veri setinden çıkarılan özellikler makine öğrenmesi algoritmaları ile sınıflandırıldı. Performans sonuçları karışıklık matrisi analizi kullanılarak elde edildi.

Bulgular: Derin sinir ağlarının eğitimi sonrasında kullanılan modeller arasında en yüksek doğruluk oranına %80 ile AlexNET modelinde ulaşıldı. Özellik çıkarımı ve sınıflandırıcı kullanımı sonrasında elde edilen ikinci aşama sonuçlarda ise en yüksek doğruluk oranına %93.5 ile VGG19 modelinde destek vektör makinesi sınıflandırıcısı ile ulaşıldı. Ayrıca destek vektör makinesi sınıflandırıcısının kullanılmasıyla tüm modellerde doğruluk oranlarında artışlar tespit edildi.

Sonuç: Benign ve malign akciğer nodüllerinin derin öğrenme modelleri ve özellik çıkarımı kullanılarak ayrıştırılması, radyoloji pratiğinde erken tanı açısından önemli avantajlar sağlayacaktır. Çalışmamızda elde edilen sonuçlar da bu görüşü destekler niteliktedir.

Anahtar sözcükler: Derin öğrenme, akciğer kanseri, soliter pulmoner nodül.

Corresponding author: Muhammed Bilal Akıncı.
E-mail: m.bilalakinci@gmail.com

The most common cancers in all age groups and both sexes in the world and in our country are breast, prostate, and lung cancers. Among these, the most common cause of death is lung cancer for all age groups and for both sexes.[1] The five-year survival rate for lung cancer is between 10 and 20%, which is quite low.[2-4] Five-year survival is high in patients with lung cancer detected and resected at an early stage. It is known that even a cure can be achieved in Stage I cancers.[2,5] Computed tomography (CT) is the most commonly used and most effective imaging method for detecting lung cancer.[4]

Lung lesions are divided into two main groups: malignant and benign lesions. About 70 to 80% of benign lesions are infectious granulomas, and 10% are hamartomas. Malignant lesions consist of primary lung cancer or metastases.[6]

Radiologists detect and characterize diseases through the qualitative features of medical images.[7] The qualitative features of nodular lesions detected in the lung include the size of the nodule, its location, shape and border features, whether it contains calcification and fat, its contrast status, and growth rate.[8] Computer-aided diagnostic systems using convolutional neural networks (CNNs) are excellent at automatically recognizing complex patterns in imaging data and are highly effective in providing quantitative rather than qualitative assessments of radiographic features. In studies by Gonçalves et al.[4] and Kim et al.,[9] there was an increase in the diagnostic accuracy of radiologists using computer-aided diagnosis systems. In this respect, the use of computer-aided diagnosis systems in the characterization of lung lesions will provide significant support for early and rapid diagnosis.[7]

Feature extraction must first be performed to detect and characterize nodules using deep learning. Feature extraction is the process of highlighting important points in the image. There are two commonly used methods for feature extraction. The first of these is the manual feature extraction process with image processing methods. In this process, it is necessary to work on each image separately, and therefore, this takes a long time. The second most widely used feature extraction method is the use of CNN in deep learning. Here, the images are circulated between the CNN layers, and certain features on each image are automatically obtained. It is among the most preferred methods in studies since it automates the work. In this method, it is necessary to investigate deep learning models that provide the best extraction of features. In our study, we aimed to differentiate lung nodules as benign and malignant with high accuracy by using CNN models with deep feature extraction, and we researched the best model for the dataset we created.

PATIENTS AND METHODS

Thoracic CT scans taken at the Van Yüzüncü Yıl University Faculty of Medicine, Department of Radiology between January 2015 and December 2020 were retrospectively scanned. Patients with pulmonary nodules on scanned CT images were evaluated. Among these patients, patients with histopathological diagnosis or at least two years of follow-up were included in the study. Patients with nodules smaller than 5 mm in size, with nodules that could not be clearly distinguished due to artifacts in the images, and who did not have radiological follow-up or histopathology were excluded from the study. Afterward, patients were grouped into two separate groups, according to their pathology and clinical and radiological results: benign and malignant. The criteria for determining the malignant group were that the lesion was diagnosed as histopathologically malignant or evaluated as malignant clinicoradiologically. The criteria for determining the benign group were that the lesion was diagnosed as histopathologically benign or there was no increase in size and no change in nature during clinicoradiological follow-up of at least two years.

Computed tomography images were obtained with multislice tomography devices with 128 detectors (Somatom Definition AS+128; Siemens AG, München, Germany) and 16 detectors (Somatom Emotion 16-slice; Siemens AG, München, Germany). Sections were obtained from the distal neck to the upper abdomen, with the patients in the supine position and holding their breath. Axial images with a cross-sectional thickness of 3 mm, which were transferred to the system after the previous imaging procedures, were evaluated on the high-resolution grayscale medical monitor used for routine CT examinations by two separate radiologists with four and 10 years of experience, respectively, in terms of compatibility with the study. Afterward, cases were classified as malignant and benign according to the histopathological or clinicoradiological data of the patient.

In the malignant group, 199 images were created using the follow-up tomography images of 29 patients (19 males, 10 females; mean age: 65±10.4 years; range, 43 to 88 years) and all pathological axial

sections within these images. In the benign group, 202 images were created using follow-up tomography images of 68 patients (38 males, 30 females; mean age: 59±12.2 years; range, 27 to 81 years) and all axial sections in which the nodule was present (Figure 1). Since these images created for training deep learning algorithms contained malignant or benign lesions, training was also required with normal sections. For this reason, we created a control group without lung pathology. Sixty-seven patients (38 males, 29 females; mean age: 56.9±14.1 years; range, 26 to 81 years) were included in this control group. To homogenize the distribution while creating normal sections, patients from different age groups were selected. In addition, different sections were obtained from the upper, middle, and lower parts of the lung in each patient. In this way, 343 images were created from 67 patients included in the control group (Figure 2).

Artificial intelligence training

The images obtained after archive scanning were collected in three different classes (normal, benign, and malignant), and then training was started in the Matrix laboratory (MATLAB; MathWorks Inc., Natick, MA, USA) environment with AlexNet, GoogleNet, VGG19, and ResNet models. The images from the three groups we created were divided into 80% training and 20% test data. The networks were trained with 80% training data, and attempts were made to verify the trained network using the 20% test data. Afterward, the accuracy values obtained for each class were determined, and each network trained with the dataset was recorded as a model (Figure 3).

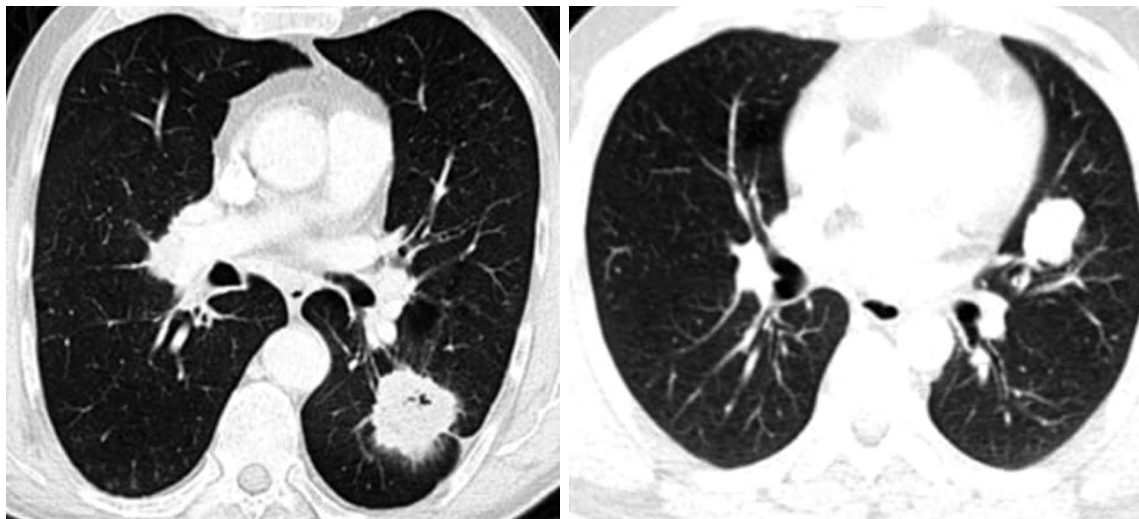In the next step, feature extraction was started, and 1,000 features were extracted from each network.

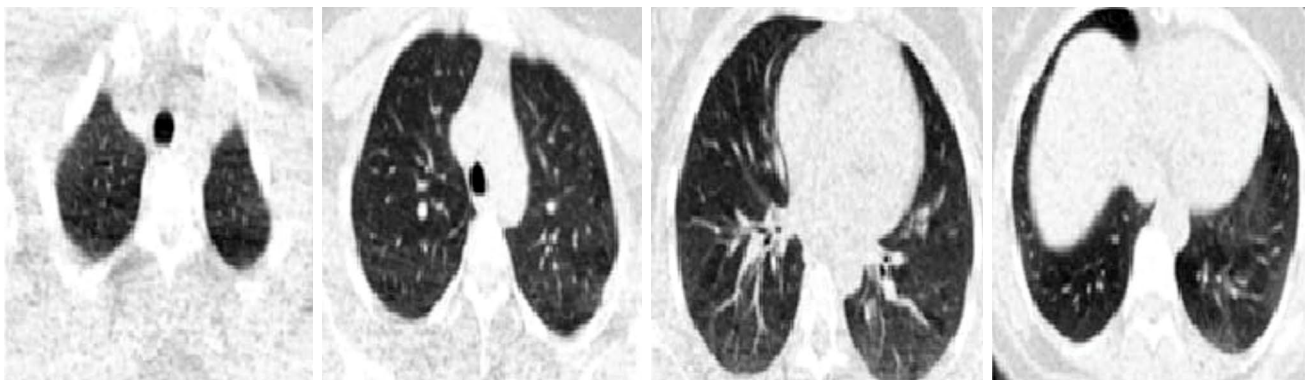**Figure 1.** Section sample of a malignant lesion on the left and a benign lesion on the right.

Figure 2. Normal section samples obtained from the upper-middle and lower sections of the lung.
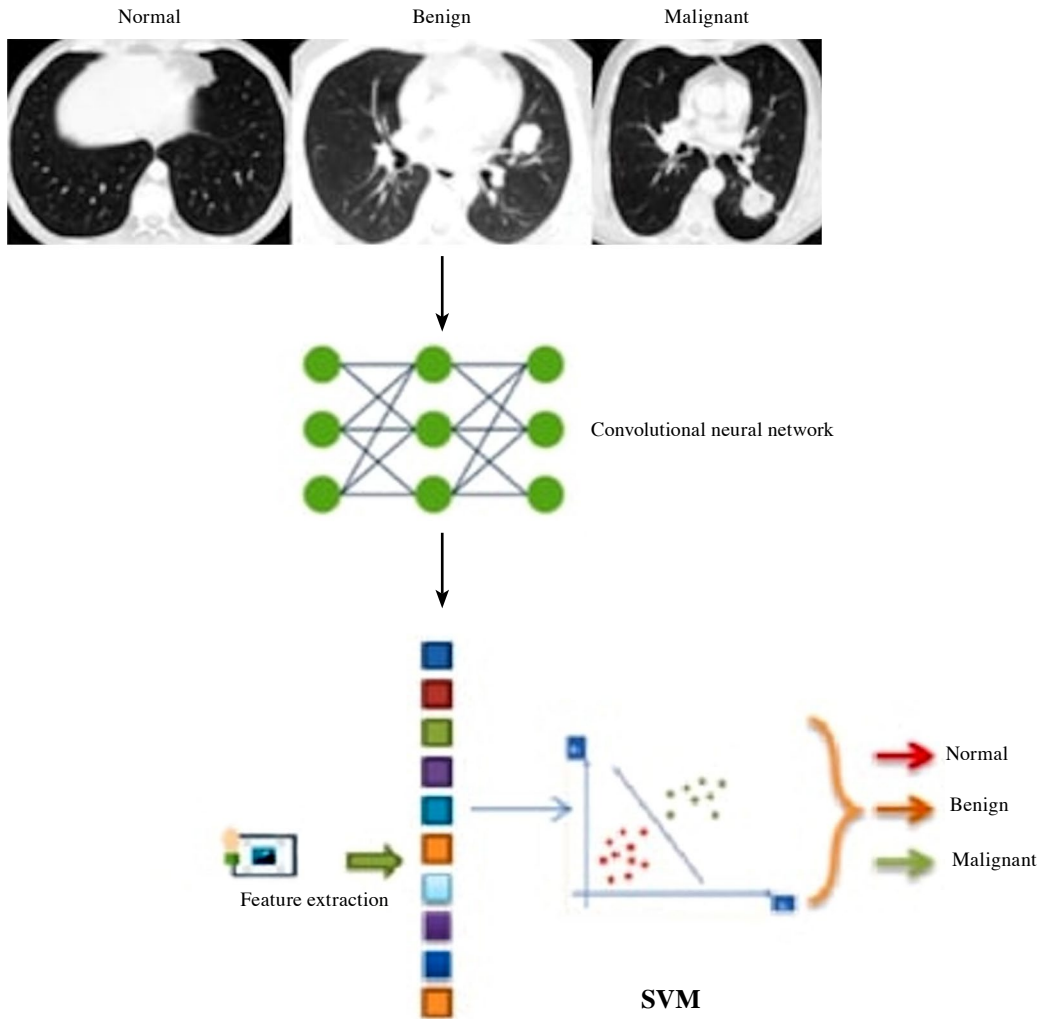
Figure 3. Working flowchart.

After this stage, each image had 1,000 features after passing through the network. These features were classified using SVM (support vector machine) and KNN (k-nearest neighbor) classifiers from machine learning algorithms.

Statistical analysis

Python programming language and the sklearn library written in this language were used in the statistical analysis studies carried out in our study. All statistical metrics were obtained using ready-made functions in this library.

RESULTS

The parenchyma window was used for all sections (Table 1). Before feature extraction and using SVM and KNN classifiers, the accuracy rates we obtained as a result of artificial neural network training were 78.33% for VGG19, 80% for AlexNet, 75% for ResNet, and 73.33% for GoogleNet.

In the results we obtained after feature extraction and use of classifier, it was found to be 93.5% for VGG19, 86.7% for AlexNet, 88.5% for ResNet, and 78% for GoogleNet after using the SVM classifier. After using the KNN classifier, it was found to be 92.2% for VGG19, 78.9% for AlexNet, 74.6% for ResNet, and 73% for GoogleNet (Table 2).

As a result of the first training, the AlexNet model was the best model with 80% accuracy. After feature extraction through models, statistical classifiers SVM and KNN were used. According to the results obtained from these classifiers, the VGG19 model achieved the best result with an accuracy rate of 93.5%. Retraining was required when new patient images were added to the dataset. These rates are expected to increase with new images.

Table 1. Dataset information

	Normal dataset			Benign dataset			Malign dataset		
	n	Mean±SD	Range	n	Mean±SD	Range	n	Mean±SD	Range
Number of patients	67			68			29		
Sex									
Male	38			38			19		
Female	29			30			10		
The average age		56.9±14.1	26-81		59±12.2	27-81		65±10.4	43-88
Number of images	343			202			199		

SD: Standard deviation.

Table 2. Comparison table before and after using classifier

CNN Algoritms	Before feature extraciton and using classifier accuracy rate	After using the SVM classifier accuracy rate	After using the KNN classifier accuracy rate
VGG19	78.33	93.5	92.2
AlexNet	80	86.7	78.9
ResNet	75	88.5	74.6
GoogleNet	73.33	78	73

CNN: Convolutional neural networks; SVM: Support vector machine; KNN: K nearest neighbor.

The metric values used for statistical analysis are given in Figure 4. Statistical analysis results for SVM and KNN are given in Tables 3 and 4.

$$S_e = \frac{TP}{TP + FN}$$

$$S_p = \frac{TN}{TN + FP}$$

$$Pre = \frac{TP}{TP + FP}$$

$$F\text{-}score = \frac{2TP}{2TP + FP + FN}$$

$$Accuracy = \frac{TP + T.V}{TP + TN + PP + FN}$$

	Actual value	
	Positive	Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Figure 4. Equations for metric values used for statistical analysis.

DISCUSSION

Precise and accurate detection and examination of pulmonary nodules is one of the best approaches to reduce deaths from lung cancer. With the development of artificial intelligence methods, pulmonary nodules have been classified according to whether they are malignant or benign. In this sense, the first machine learning methods were used, but the time-consuming and limited discrimination power of the applied methods was an important limitation of the studies.[10,11] Later on, CNN architectures that can automatically extract high-level features from images were developed.

In our retrospective study, we aimed to differentiate the nodules observed on lung tomography images obtained in our clinic as malignant and benign by using the automatic image analysis feature of CNNs. Similar to our study, Sun et al.[12] used LeNet to classify lung nodules, Hussein et al.[13] estimated the malignancy of nodules based on AlexNet, and Nibali et al.[14] used ResNet for benign/malignant lung nodule classification.[15] In our study, we used AlexNet, GoogleNet, ResNet, and VGG19 models as CNNs. These models were chosen because they achieved

Table 3. Statistical analysis results after using SVM classifier

	True positive	True negative	False positive	False negative	F-score	Sensitivity	Specificity	Precision	Accuracy
VGG19									
Benign	182	523	19	20	0,903226	0,90099	0,964945	0,905473	0,947581
Malignant	197	535	10	2	0,970443	0,98995	0,981651	0,951691	0,983871
Normal	317	382	19	26	0,933726	0,924198	0,952618	0,943452	0,939516
AlexNet									
Benign	157	506	36	45	0,794937	0,777228	0,933579	0,813472	0,891129
Malignant	187	533	12	12	0,939698	0,939698	0,977982	0,939698	0,967742
Normal	301	350	51	42	0,866187	0,877551	0,872818	0,855114	0,875000
ResNet									
Benign	167	503	39	35	0,818627	0,826733	0,928044	0,81068	0,900538
Malignant	185	538	7	14	0,946292	0,929648	0,987156	0,963542	0,971774
Normal	304	359	42	39	0,882438	0,886297	0,895262	0,878613	0,891129
GoogleNet									
Benign	141	480	62	61	0,696296	0,69802	0,885609	0,694581	0,834677
Malign	169	515	30	30	0,849246	0,849246	0,944954	0,849246	0,919355
Normal	270	329	72	73	0,788321	0,787172	0,820449	0,789474	0,805108

SVM: Support vector machine.

Table 4. Statistical analysis results after using KNN classifier

	True positive	True negative	False positive	False negative	F-score	Sensitivity	Specificity	Precision	Accuracy
VGG19									
Benign	181	514	28	21	0,880779	0,89604	0,948339	0,866029	0,93414
Malignant	198	535	10	1	0,972973	0,994975	0,981651	0,951923	0,985215
Normal	307	381	20	36	0,916418	0,895044	0,950125	0,938838	0,924731
AlexNet									
Benign	152	463	79	50	0,702079	0,752475	0,854244	0,658009	0,826613
Malignant	179	526	19	20	0,901763	0,899497	0,965138	0,90404	0,947581
Normal	256	342	59	87	0,778116	0,746356	0,852868	0,812698	0,803763
ResNet									
Benign	144	443	99	58	0,647191	0,712871	0,817343	0,592593	0,788978
Malignant	175	522	23	24	0,881612	0,879397	0,957798	0,883838	0,936828
Normal	236	334	67	107	0,73065	0,688047	0,832918	0,778878	0,766129
GoogleNet									
Benign	150	436	106	152	0,537634	0,496689	0,804428	0,585938	0,694313
Malign	170	609	36	29	0,839506	0,854271	0,944186	0,825243	0,922986
Normal	223	342	159	120	0,615172	0,650146	0,682635	0,58377	0,669431

SVM: Support vector machine.

success in competitions based on image analysis and classification.[16,17]

After training of CNN models with our dataset, we performed feature extraction in the next step and classified the features obtained with KNN and SVM classifiers, which are machine learning algorithms. The highest accuracy rate of 93.5% was achieved with the combination of VGG19 and SVM classifier.

Shen et al.[18] used a multi-crop pooling technique they called multi-crop CNN in their study in 2016. In the study, three different dimensional datasets were used, and the highest accuracy rate achieved was 87% in the highest dimensional dataset. This study provides important information in terms of the effect of dataset size on the increase in accuracy rates.

Nibali et al.[14] reached an accuracy rate of 89.9% in their study with the ResNet model in 2017. Dai et al.[15] achieved an accuracy rate of 91.47% with the artificial neural network model they developed in 2018. In these two studies, the publicly available lung tomography image database known as the Lung Image Database Consortium image collection (LIDC-IDRI) was used, and the classification of malignant and benign in both studies was based on the opinions of radiologists. The large number of datasets used in the studies by Nibali et al.[14] and Dai et al.[15] may have been effective in increasing the accuracy rates. The classification process in these studies by radiologists as low probability or high probability differs from our study. As stated above, some objective criteria were used while classifying malignant and benign in our study. In addition, while creating the groups, we paid attention to the similarity of average ages and number distribution. While forming the control group, we selected samples from different lung sections. Thus, our dataset became more homogeneous before starting training. This may have an effect on the accuracy rates we obtained using a smaller number of samples compared to the studies mentioned above.

Da Nóbrega et al.[10] also used the LIDC-IDRI database in 2018, like the mentioned researchers, but they used many different architectures and used a classifier after feature extraction, as in our study. In this comprehensive study, the highest accuracy rate was close to 90% for the SVM-radial basis function (RBF) classifier combination with the ResNET50 CNN network. In this study, the ResNET and VGG19 architectures and KNN and SVM classifiers were used, similar to our study. In the study by Da Nóbrega et al.,[10] the accuracy rates for these combinations were

86.85% for ResNet-KNN, 86.98% for ResNet-SVM, 86% for VGG19-KNN, and 82.03% for VGG19-SVM. In our study, accuracy was found to be 74.6% for ResNet-KNN, 88.5% for ResNet SVM, 92.2% for VGG19-KNN, and 93.5% for VGG19-SVM. The ResNet-KNN combination was not as successful in our study as it was for Da Nóbrega et al.[10] However, when other similar combinations were compared, the accuracy rates we obtained were found to be higher. This study shows that similar to our study, the accuracy rates obtained as a result of training CNN architectures will increase with the use of classifiers after deep feature extraction.

The main limitation of our study is the small number of patients. The retrospective nature of the study is another limitation. Additionally, thorax CT sections could not always be obtained with the same protocol, and images were taken using different CT devices.

In conclusion, the results show that deep learning techniques can distinguish benign nodules from malignant by performing image analysis, and they are promising in this respect. Our suggestions for future studies are that CNN models be trained with higher and homogenized sample numbers and suitable classifier combinations to increase the accuracy rates.

REFERENCES

1. Thandra KC, Barsouk A, Saginala K, Aluru JS, Barsouk A. Epidemiology of lung cancer. Contemp Oncol (Pozn) 2021;25:45-52. doi: 10.5114/wo.2021.103829.

2. Collins LG, Haines C, Perkel R, Enck RE. Lung cancer: Diagnosis and management. Am Fam Physician 2007;75:56-63.

3. Tanoue LT, Tanner NT, Gould MK, Silvestri GA. Lung cancer screening. Am J Respir Crit Care Med 2015;191:19-33. doi: 10.1164/rccm.201410-1777CI.

4. Gonçalves L, Novo J, Cunha A, Campilho A. Learning lung nodule malignancy likelihood from radiologist annotations or diagnosis data. J Med Biol Eng 2018;38:424-42. doi: 10.1007/s40846-017-0317-2.

5. Kartaloğlu Z. Soliter pulmoner nodüle yaklaşım. Turkish J Thorac Cardiovasc Surg 2008;16:274-83.

6. Ost D, Fein AM, Feinsilver SH. Clinical practice. The solitary pulmonary nodule. N Engl J Med 2003;348:2535-42. doi: 10.1056/NEJMcp012290.

7. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer 2018;18:500-10. doi: 10.1038/s41568-018-0016-5.

8. Soubani AO. The evaluation and management of the solitary pulmonary nodule. Postgrad Med J 2008;84:459-66. doi: 10.1136/pgmj.2007.063545.

9. Kim RY, Oke JL, Pickup LC, Munden RF, Dotson TL, Bellinger CR, et al. Artificial intelligence tool for assessment of indeterminate pulmonary nodules detected with CT. Radiology 2022;304:683-91. doi: 10.1148/radiol.212182.

10. da Nóbrega RVM, Rebouças Filho PP, Rodrigues MB, da Silva SPP, Dourado Júnior CMJM, de Albuquerque VHC. Lung nodule malignancy classification in chest computed tomography images using transfer learning and convolutional neural networks. Neural Comput Applic 2020;32:11065-82. doi: 10.1007/s00521-018-3895-1.

11. Monkam P, Qi S, Ma H, Gao W, Yao Y, Qian W. Detection and classification of pulmonary nodules using convolutional neural networks: A survey. IEEE Access 2019;7:78075-91.

12. Sun W, Zheng B, Qian W. Computer aided lung cancer diagnosis with deep learning algorithms. Presented at the SPIE medical imaging,Computer-Aided Diagnosis; 2016 Feb 27-3, San diego. 2016;9785:241-8. doi: 10.1117/12.2216307

13. Hussein S, Gillies R, Cao K, Song Q, Bagci U. TumorNet: Lung nodule characterization using multi-view Convolutional Neural Network with Gaussian Process 2017:1007-10. doi: 10.1109/ISBI.2017.7950686.

14. Nibali A, He Z, Wollersheim D. Pulmonary nodule classification with deep residual networks. Int J Comput Assist Radiol Surg 2017;12:1799-808. doi: 10.1007/s11548-017-1605-6.

15. Dai Y, Yan S, Zheng B, Song C. Incorporating automatically learned pulmonary nodule attributes into a convolutional neural network to improve accuracy of benign-malignant nodule classification. Phys Med Biol 2018;63:245004. doi: 10.1088/1361-6560/aaf09f.

16. Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. NeurIPS 2012:25. doi: 10.1145/3065386.

17. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015:1-9. doi: 10.1109/CVPR.2015.7298594.

18. Shen W, Zhou M, Yang F, Yu D, Dong D, Yang C, et al. Multi-crop Convolutional Neural Networks for lung nodule malignancy suspiciousness classification. Pattern Recognit 2017;61:663-73. doi: 10.1016/j.patcog.2016.05.029.