European Journal of Radiology Artificial Intelligence 4 (2025) 100047

Contents lists available at ScienceDirect

European Journal of Radiology Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/european-journal-of-radiology-artificial-intelligence

Using ensemble radiomic models to identify uncertainty in lung nodule classifications

Elisabeth Pfaehler [a,*,1], Lena Krieger [a], Ira Assent [a,b], Sven Nebelung [c], Alex Zwanenburg [d,e], Daniel Truhn [c]

[a] Forschungszentrum Jülich GmbH, Institute of Advanced Simulations, IAS-8: Data Analytics and Machine Learning, Jülich 52428, Germany
[b] Aarhus University, Department of Computer Science, Aarhus 8200, Denmark
[c] University Hospital Aachen, Department of Diagnostic and Interventional Radiology, Aachen 52074, Germany
[d] National Center for Tumor Diseases (NCT), NCT/UCC Dresden, a partnership between DKFZ, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, and Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Germany
[e] OncoRay – National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Helmholtz-Zentrum Dresden-Rossendorf, Dresden, Germany

ARTICLE INFO

Keywords:
Imaging biomarkers
Statistical uncertainty
Lung nodules
Classification
CT images

ABSTRACT

Aims: The performance of machine learning models for tasks such as tumor malignancy classification depends heavily on representativeness of the training data. Under-represented cases can lead to less reliable and error-prone predictions. To increase trust, it is crucial to determine when a prediction should be considered reliable.
Methods: Radiomic features were extracted from 918 lung nodules in the LIDC dataset. Images were discretized with a fixed bin size of 64 before texture analysis. Features highly correlated with nodule volume or each other were excluded, retaining 80 features. Features were normalized using min-max scaling. The dataset was split into 80 % training and 20 % testing data, where the model was applied and the uncertainty analysis was performed. An ensemble model was trained using 1000 bootstraps of 90 % of training data, Random Forest, Support Vector Machine, Gradient Boosting, AdaBoost, and ExtraTree classifiers were trained, each of them on 200 bootstraps. Aleatoric and epistemic uncertainties were calculated, and percentile-based thresholds were used to identify uncertain predictions. Evaluation metrics assessed overall performance and performance on certain versus uncertain cases.
Results: The ensemble model achieved 86 % accuracy. Depending on the uncertainty threshold, 11 %-32 % of test cases were flagged as uncertain. The proportion of certain incorrect predictions decreased from 8.1 % at the highest threshold to 2.1 % at the lowest. Thresholds near the 80th or 85th percentile balanced reducing manual review workload and identifying errors effectively.
Conclusion: The ensemble model successfully identifies most incorrect predictions as uncertain. Future work will explore the current missing validation on external datasets.

Introduction

Lung cancer is one of the most prevalent cancers worldwide [1]. Lung cancer screening with chest computed tomography (CT) can enable earlier detection and thus improve survival rates [2]. In clinical practice, radiologists identify suspicious lung nodules on the CT scan. If nodules are suspected to be malignant, biopsy procedures are usually performed. Since biopsies are time-consuming and burdensome for patients, a (semi-)automatic and non-invasive system for classifying nodules solely from CT images would provide substantial clinical value. In recent years, machine learning (ML) approaches for the automatic classification of tumor malignancies in CT images have been widely investigated. Two widely used approaches are (1) the use of radiomic features and (2) Convolutional Neural Networks (CNN) [3,4]. An advantage of radiomic features is their interpretability as they describe tumor characteristics such as volume or mean intensity. Textural

* Correspondence to: Institute of Advanced Simulations, IAS-8: Data Analytics and Machine Learning, Forschungszentrum Jülich GmbH, Wilhelm Johnen Straße, Jülich 52428, Germany.
E-mail address: e.pfaehler@fz-juelich.de (E. Pfaehler).
[1] ORCID: 0000–0002-6160–3011

https://doi.org/10.1016/j.ejrai.2025.100047
Received 9 May 2025; Received in revised form 1 October 2025; Accepted 20 October 2025
Available online 22 October 2025
3050-5771/© 2025 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

features, for instance, can describe rapid textural changes in voxel neighborhoods [5]. As model interpretability is a critical factor for clinical use, we concentrate in this work on radiomic features. Previous studies used radiomic features for the identification of malignant lung nodules [6,7] in general and small (diameter < 15 mm), malignant lung nodules specifically [8]., reporting classification accuracies between 73 % and 86 %.

Not all predictions made by classification models have the same certainty. For example, uncertain predictions may arise from lung nodules that differ substantially from the training data distribution, or from nodules that cannot clearly be defined as benign or malignant. Identifying such uncertain predictions can guide further direct clinical investigation, whereas certain predictions can be used to rule out diagnostic procedures. Uncertainty estimation has therefore received considerable attention within the context of deep learning [9]. However, uncertainty estimation of radiomic models for lung nodule classification remains unexplored.

Uncertainty in model decisions can be divided in aleatoric and epistemic uncertainty. Aleatoric uncertainty refers to uncertainty caused by randomness or noise in the data and cannot be reduced even when collecting more data. Epistemic uncertainty arises due to limited data or imperfect models and can potentially be reduced by collecting more data or changes in the model [10–12]. Ensemble models are one of several approaches for identifying uncertain predictions [10]. In this strategy, multiple classifiers are trained on the same task, such as detecting malignant lung nodules. To induce variability during training and increase model diversity, different classifiers are combined and/or bootstrapping is performed [13]. In an ensemble model, the predictions of all ensemble members are aggregated to calculate the final decision. One advantage of ensemble models lies in their intuitive interpretation: Ensemble members converge to the same decision when the input data is similar to the training data distribution.

In this work, we employ an ensemble of radiomic models to quantify uncertainty in lung nodule malignancy classification. The ensemble model is applied to a test set, and aleatoric and epistemic uncertainty are used to identify uncertain decisions. The distribution of malignant and benign nodules flagged as uncertain is analyzed and the tumor characteristics associated with these uncertain predictions are investigated.

Materials and methods

Dataset and data processing

The Lung Image Database Consortium (LIDC) dataset was used [14], consisting of thoracic CT scans of 1018 patients that had one or more lung nodules (total: 2625). To obtain segmentation and binary classification labels, the information of the LIDC dataset was used. In this dataset, up to four radiologists delineated the lesions independently and voted whether a nodule was highly unlikely, moderately unlikely, moderately likely, or highly likely to be malignant. The radiologists could also indicate if they were unsure of their decision. The classification problem was converted into a binary classification problem by assuming that all nodules labelled moderately and highly unlikely malignant were benign. All moderately and highly likely malignant nodules were considered malignant. Cases where the radiologist's decisions did not lead to a consensus were excluded. A consensus among radiologists was defined as more than half of the radiologists reviewing a lesion classified the lesion as malignant or benign. Moreover, all images with an original slice thickness greater than 2.5 mm were excluded. After discussion with two radiologists (DT and SN), we excluded the six biggest lung nodules (lesions with a size greater than 8000 mm^3). As our aim was to focus on cases normally appearing in lung cancer screening and the volume of these nodules was so large that they were likely detected earlier. In total, 918 lung nodules of 490 patients are included (an overview of the exclusion process is displayed in Suppl. Fig. 1). The final segmentation mask was obtained by majority voting among all available segmentations of the corresponding lesion and was used for feature calculation.

Before feature extraction, images were resampled to a common voxel size of $0.70 \times 0.70 \times 1.25$ mm using bicubic interpolation and the SimpleITK library (version 2.3.1).

Label uncertainties

As the labels of this dataset were obtained by visual analysis, incorrect decisions might occur. To address label uncertainty, each lesion was assigned a confidence-weight during the classification process. Before model training, the previously defined instance weights for class imbalance were multiplied by the confidence weights, that were defined as follows: a value of 1.0 was assigned if the reviewing radiologist classified the lesion as highly likely malignant/benign; a value of 0.8 was assigned if the reviewer classified the lesion as likely malignant/benign; and a value of 0.6 was assigned if the reviewer expressed uncertainty regarding the classification. For each lesion, the final confidence weight was obtained by averaging across the reviewing radiologists. To assess the influence of confidence weights, an additional analysis with only class imbalance weights is performed.

Radiomic feature calculation

489 radiomic features were calculated using RaCaT (version 1.28 [15]), which complies with the Image Biomarker Standardization Initiative [5]. Before textural feature calculation, images were discretized using a fixed bin number of 64. All features with a Spearman correlation coefficient above 0.9 with nodule volume were excluded. Only features not highly correlated with any other feature were kept. When two features demonstrated high correlation, the feature with the lower volume dependence was kept. From these 89 features, those found in previous work to be repeatable and reproducible were considered further [16–18]. Feature names of the remaining 81 features are listed in the supplemental Table 1. Before model training, features were normalized using min-max normalization.

Creating an ensemble model

Before training, the dataset was split into a training and test dataset (80 % ($n = 734$) and 20 % ($n = 184$) of the dataset) by stratified splitting at the patient level. The fraction of malignant nodules was 22 % in the training set and 24 % in the test set.

Generation of ensemble members

An ensemble model consists of multiple ensemble members, which are models themselves. Data and model diversity are used to generate diverse ensemble members.

In each of 1000 bootstraps, 90 % of the training data is randomly sampled with replacement. These 1000 bootstraps are divided into five sets of 200 bootstraps. Each set of 200 bootstraps is assigned to one of the five classifiers: Random Forest, Support Vector Machine (SVM), Gradient Boosting, AdaBoost, or ExtraTree. These classifiers were included as they demonstrated good results in preliminary tests on the training dataset and allowed for the use of instance weights during training. To account for data imbalance, instance weights equal to the ratio of benign and malignant nodules in the current bootstrap are set during training. Classifier hyperparameters (see Table 1) were determined once via grid search on the whole training dataset. The mean predicted value of all ensemble members was computed to derive the predicted class label. If the mean was greater than or equal to 0.5, the predicted was set to malignant. Otherwise, the prediction was set to benign.

Table 1
Parameters chosen via grid search for different classifiers. Parameters that are not listed were kept at their default.

Classifier	Parameter
Random Forest	max. depth: 4; min samples split: 5; nr. estimators: 24;
SVM	kernel: poly, Regularizazion param. C: 0.15, probability: True
Gradient Boosting	loss: exponential, learning rate: 0.01 7 nr. estimators= 100
Adam Boost	nr. estimators: 100, algorithm: "SAMME"
Extra Tree	max. depth= 5, min samples split= 5

Feature selection

A selection of features related to nodule malignancy is required to achieve adequate predictive model accuracy. Similar to classifier hyperparameters, the optimal number of features was determined globally within the training dataset, using 5-fold cross-validation. Feature set sizes between 1 and 10 were assessed. As the highest performance metric was achieved with four features, the four most informative features were selected in each bootstrap training set based on random forest feature importance in scikit-learn (version 1.3.0). For comparison, only the morphological feature volume was used in the process as volume has a well-known predictive value for lung nodule

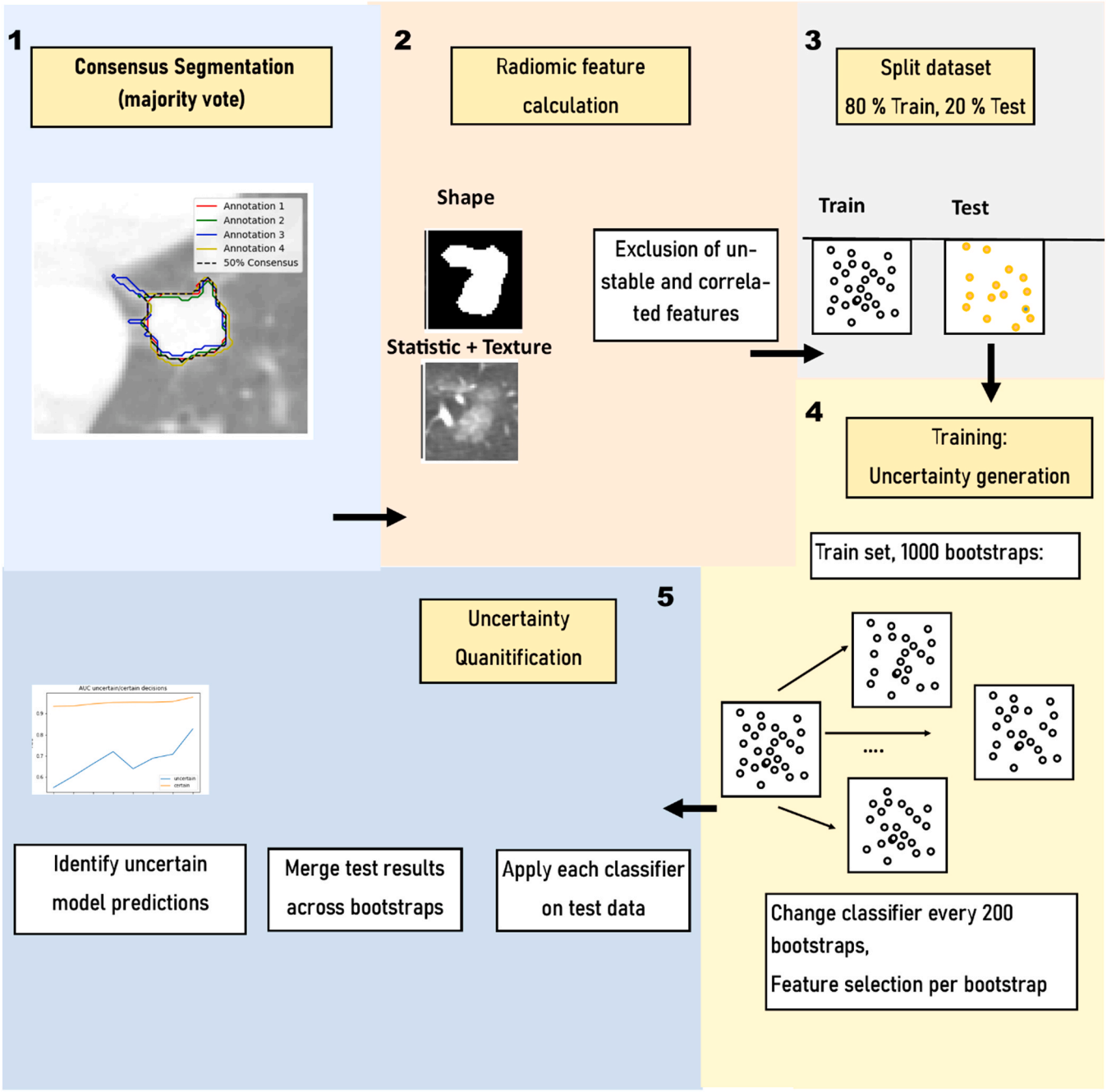

Fig. 1. Overview of the approach: (1) The consensus (majority vote) of four observers was defined as the final mask to (2) calculate radiomic features. Unstable and highly correlated features are eliminated. (3) Dataset is split in 80 % training and 20 % testing. (4) The training dataset is bootstrapped repeatedly, creating 1000 bootstraps. Every 200 bootstraps, a different classifier is applied. (5) Each bootstrap model is used to infer cases in the test dataset, predictions are ensembled, and cases with high uncertainty are identified.

malignancy.

Uncertainty quantification

Each ensemble member is applied to the test-set and the probability of each decision is stored. To identify uncertain cases, uncertainty values are computed: epistemic uncertainty is calculated as the variance of predicted probabilities across model samples, while aleatoric uncertainty is quantified as the conditional entropy of the predicted probabilities [12,19]. Fig. 1 displays an overview of our approach.

If classifier decisions vary highly across ensemble members, the model decisions can be regarded as uncertain. However, the optimal threshold to identify uncertain decisions must be determined. As uncertain model decisions require some additional work by physicians, it is important that: (1) most selected uncertain cases are indeed incorrect decisions, (2) all incorrect decisions are selected, and (3) the number of selected cases is as small as possible. In our approach, various percentiles of aleatoric and epistemic uncertainty are computed, and cases with one of both uncertainty values exceeding a given threshold are identified as uncertain. We tested thresholds at the 75th, 80th, 85th, 90th, and 95th percentiles.

All data analysis was performed in Python version 3.8.13. The radiomic features and the code used in this paper are accessible on GitHub (https://github.com/ElliMEP/uncertaintyEnsembles).

Evaluation metrics

We assess the number of false positive (FP), true positive (TP), false negative (FN), and true negative (TN) decisions. Besides accuracy, we compute Positive and Negative Predictive Value (PPV, NPV), True Positive Rate (TPR), and True Negative Rate (TNR). In addition, the Area under the receiver operating characteristic curve (AUC) is calculated. Mean and standard deviation values across bootstraps are calculated. To evaluate the results of the uncertainty analysis, the evaluation metrics for all selected certain/uncertain decisions are calculated separately across thresholds, and uncertainty values are compared between correct and incorrect ensemble model decisions (cases where the model identified a nodule as benign or malignant while the ground truth yielded the opposite value), and the number of malignant and benign lesions labelled as uncertain is assessed.

Analysis of individual models

To evaluate the performance of ensemble members, the evaluation metrics are computed for each model, and overall mean, minimum, and maximum values are determined. Additionally, the importance of the dataset composition on feature selection results is evaluated. Reliability diagrams for some folds and the ensemble model are displayed in the supplemental material Section 3.

Interpretability

To understand classifier decisions, SHapley Additive exPlanations (SHAP) [20] are calculated for each bootstrap performed with the Python library shap (version 0.44.0). Positive SHAP values indicate a contribution to the classification as 1 (malignant). The absolute value of the SHAP value illustrates the features' contribution to a certain class. To get an impression of the reasoning behind each ensemble member, SHAP values are calculated member-wise and compared. SHAP Beeswarm plots are displayed for four randomly selected bootstrap samples. To understand the impact of each feature on the ensemble model performance, mean SHAP values are calculated feature-wise across bootstraps.

Results

Analysis of ensemble and individual models

The ensemble model consisted of 1000 member models and was used to predict lung nodule malignancy for cases in the test dataset using a majority consensus. The ensemble model had an accuracy of 86 %, an AUC of 0.93, a PPV of 0.88, an NPV of 0.80, a TPR of 0.94, and a TNR of 0.69. The individual ensemble members had an accuracy of 85 $\pm$ 2 % (mean$\pm$standard deviation), an AUC of 0.90 $\pm$ 0.02, PPV of 0.89 $\pm$ 0.04, NPV of 0.75 $\pm$ 0.09, a TPR of 0.90 $\pm$ 0.06, and a TNR of 0.70 $\pm$ 0.13. The accuracy ranges between 75 % and 89 % while the AUC of the ensemble members lies between 0.82 and 0.95. PPV and NPV range between 0.78 and 0.98, and 0.51–1, respectively. TPR and TNR lie between 0.78 and 0.99, and 0.22–0.94.

Different features were selected across bootstraps (Fig. 2). The only feature consistently selected was volume. Joint maximum (GLCM2DAVG) was selected in 22 %, energy (intensity histogram), sum entropy, inverse different moment (both GLCM2DAVG), and volume density AEE (morphological feature) were selected in 15–18 % of bootstraps.

Interpretability: shap dependence plots

Fig. 3 displays the SHAP values for the test set of four bootstrap-samples. In each bootstrap, different features are selected, resulting in different SHAP values per feature and bootstrap, which is one reason for different classifier decisions.

Fig. 4 displays the mean absolute SHAP values across ensemble models feature-wise. Nodule volume resulted in the highest mean value, underlining that it had the highest impact on the decision. Some textural features, such as contrast or difference average (GLCM2DAVG) and the morphological feature asphericity resulted in relatively high SHAP values, but had still less impact than volume.

Uncertainty analysis

In this section, we compare the number of selected uncertain cases and the evaluation metrics for selected uncertain/certain cases across thresholds. Moreover, we compare uncertainty values between correct and incorrect decisions and count the number of malignant and benign lesions labelled as uncertain.

With decreasing threshold, the number of cases classified as uncertain increases (see Table 2, Fig. 5). The optimal threshold should balance identifying a small number of uncertain cases while classifying the highest number of incorrect decisions as uncertain. At the highest threshold (95th percentile), only few (20/184) uncertain cases are selected, and a substantial number of cases (15 out of 25 incorrect decisions) with incorrect decisions are flagged as certain. With decreasing threshold, the number of certain, but incorrect cases decreases, but at the cost of a rising number of cases with accurate classification considered uncertain. With decreasing threshold, the number of cases classified as uncertain increased from 20 at the highest threshold to 60 at the lowest threshold. Conversely, the number of certain yet incorrect decisions decreased from 15 (8.1 % of all lesions) to 4 (2.1 % of all lesions). In accordance with these results, the accuracy calculated for the uncertain cases only increased from 55 % for the highest threshold to 62 % for the lowest threshold. Higher thresholds result in an increasing number of correctly classified cases being considered uncertain. Thresholds at the 80th or 85th percentile may provide a good compromise between reducing the number of cases requiring manual evaluation and effectively identifying incorrect decisions. These thresholds resulted in 45 and 52 cases flagged as uncertain, of which 19 and 20 cases were incorrectly classified, respectively. All evaluation metrics for the different thresholds can be found in supplemental Table 2. The corresponding values when training without case-weighting are displayed in

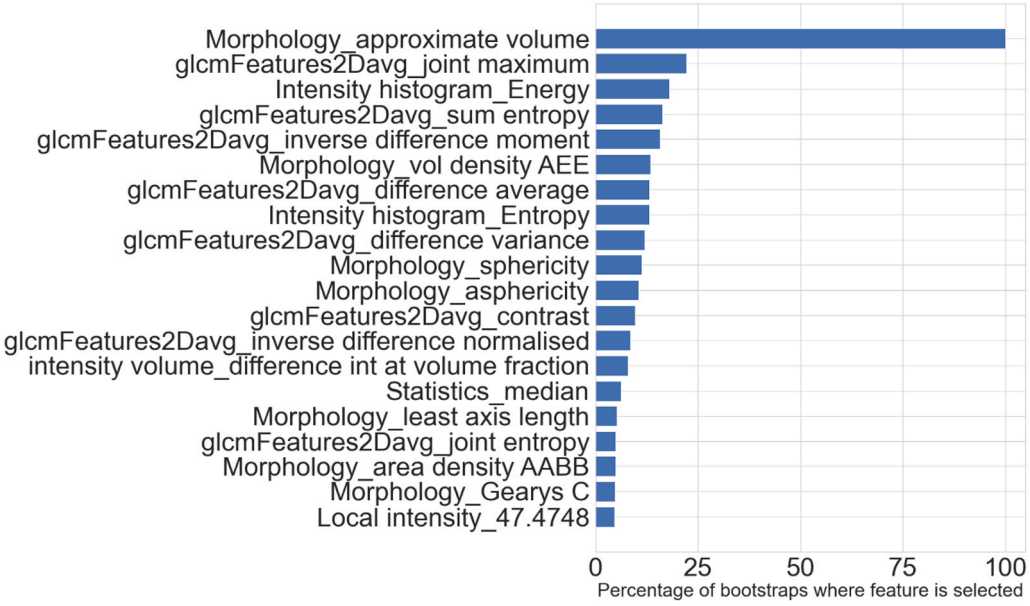

Fig. 2. The twenty most frequently selected features across bootstraps, including their frequency. As illustrated, nodule volume was selected in 100 % of the bootstrap samples, while other features varied highly across bootstraps.

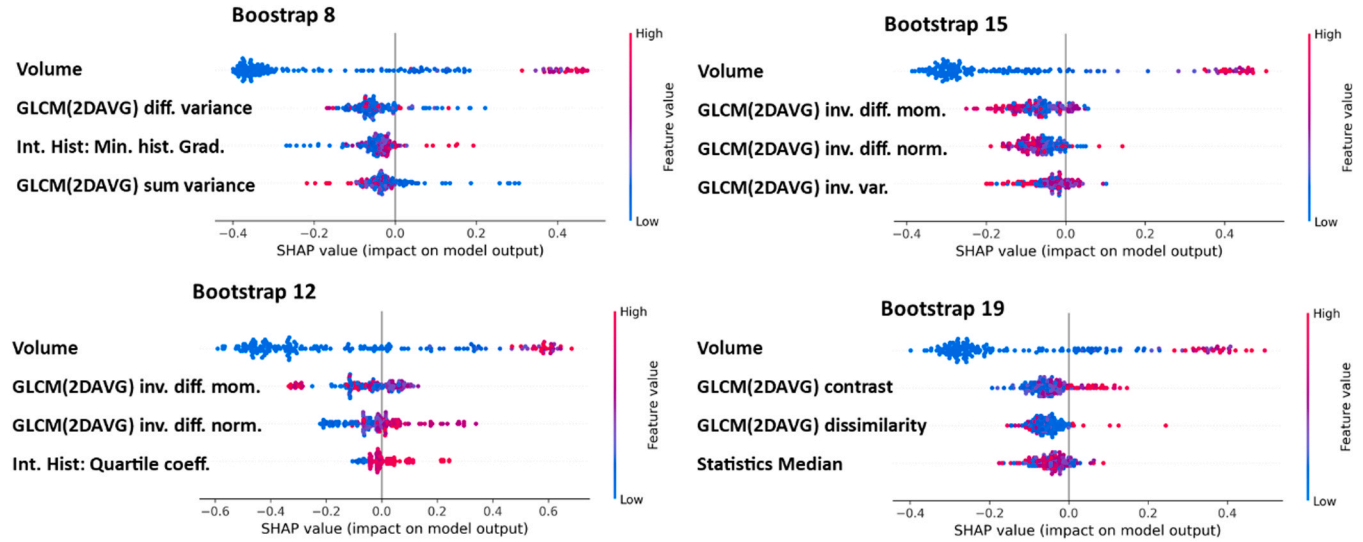

Fig. 3. SHAP values of four bootstrap samples. As displayed, in each sample different SHAP values are assigned which explains partly the difference in model decisions.

the supplemental Table 3.

Fig. 6 displays the distribution of epistemic and aleatoric uncertainty across correct and incorrect decisions. As displayed, many incorrect decisions yield a high aleatoric uncertainty indicating that the source of this uncertainty lies in the data itself. The mean epistemic uncertainty equals 0.028 for correct and 0.032 for incorrect cases while aleatoric uncertainty results in a mean value of for 0.24 correct and 0.29 for incorrect cases.

When comparing with the indicated uncertainty of the reviewers, in 4 (95th percentile), 6 (90th, 85th, 80th percentile), and 9 (75th percentile) of the cases all reviewers of the LIDC dataset labelled the lesions as highly likely malignant or benign. In 1 (95th and 90th percentile), 2 (85th and 80th percentile), and 4 (75th percentile), at least half of the reviewers indicated that they were uncertain about their decision. In all other cases, the reviewers labelled the lesions as likely malignant or benign.

All evaluation metrics as well as the uncertainty analysis when using only the morphological feature volume are listed in the supplemental Tables 4 and 5.

Feature analysis of uncertain decisions

In Fig. 7, nodule volume and nodule sphericity are displayed for certain and uncertain cases and 95th and 85th percentile threshold. The figures for the other thresholds are displayed in the supplemental Fig. 4. The majority of the most uncertain cases (uncertainty values > 95th percentile), yielded nodule volume between 200 and 350 m^3. With decreasing threshold, some cases flagged as uncertain yielded higher volumes. For the lower thresholds (percentile 85th and lower), also some smaller (<50 m^3) and larger (>1000 m^3) lesions were selected as uncertain. However, the majority of uncertain cases yielded a volume between 200 mm^3 and 400 mm^3. In supplemental Fig. 5 the feature

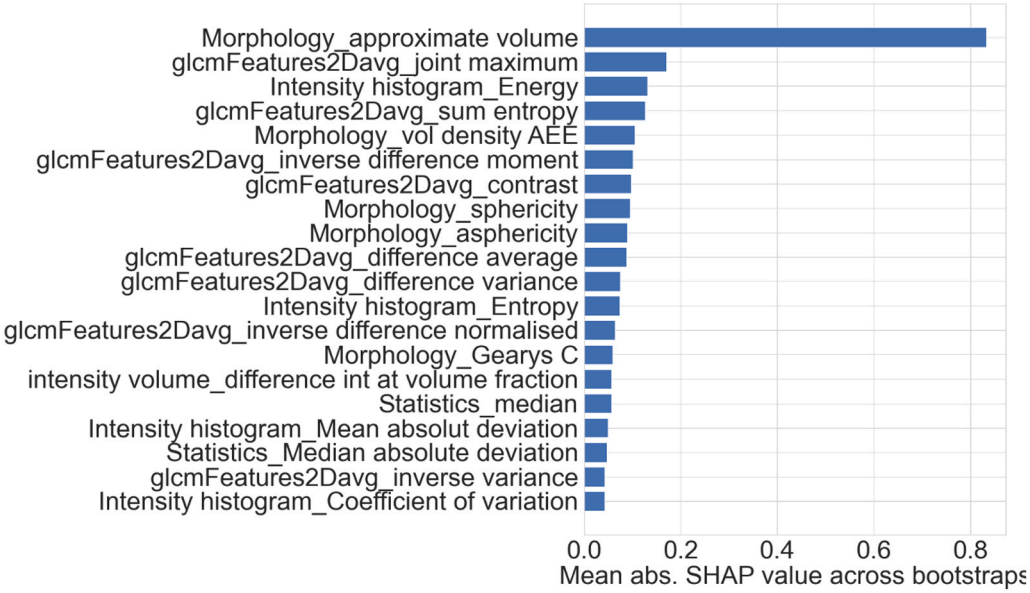

Fig. 4. Absolute SHAP values are added per feature and divided by the number of bootstraps (1000). Therefore, this plot displays the total impact one feature had in the decision process. Only the 20 features with the highest impact are displayed.

Table 2
Accuracy and AUC for the selected uncertain/certain cases (first two columns) for the test data. Last three columns: Number of selected uncertain cases and remaining false decisions. All other evaluation metrics are listed in the supplemental material, Table 2.

Threshold (percentile)	Accuracy (uncertain/ certain cases)	AUC (uncertain/ certain cases)	Nr. uncertain cases (malignant/benign)	Nr. uncertain/certain and incorrect decisions	Nr. certain malignant/benign but incorrect decisions
95th	55 %/90 %	0.63/0.94	20 (11/19)	9/16	9/7
90th	52.3 %/94 %	0.63/0.95	34 (18/16)	12/9	4/5
85th	58.%/ 96 %	0.65/0.96	45 (19/26)	19/6	3/3
80th	62 %/96 %	0.71/0.96	52 (21/31)	20/5	3/2
75th	65 %/97 %	0.68/0.96	60 (24/36)	21/4	2/2

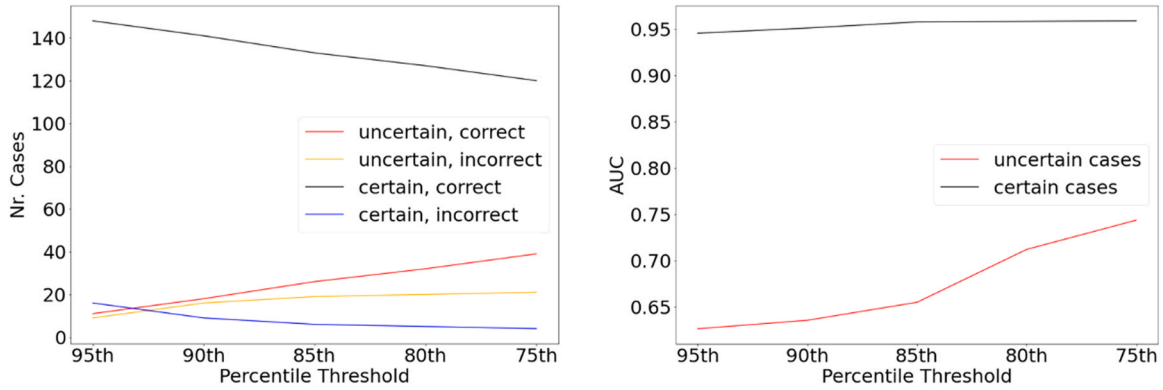

Fig. 5. (A) Number of selected uncertain cases depending on the chosen threshold; (B) AUC for certain and uncertain cases; As illustrated, the number of selected uncertain cases increases with the threshold; the AUC of the uncertain cases increases. In conclusion, the number of correct decisions selected as being uncertain increases.

distribution of malignant and benign cases in the training dataset is displayed. In the training dataset, benign and malignant cases yield volumes between 200 and 400 m^3 what might be the reason for the uncertainty of the classifier.

Discussion

Our main finding is that an ensemble approach using diverse classifiers and bootstrapping provided a mechanism to flag uncertain and often incorrect predictions. Identifying uncertain model decisions is of high interest: Unreliable predictions occur in cases where the model should not be trusted.

For further processing, several strategies are possible: One practical approach is to subject uncertain cases to a more detailed diagnostic evaluation. For instance, nodules flagged as uncertain could prompt a closer radiological review—potentially involving additional imaging modalities (such as, contrast-enhanced CT or PET/CT) or other clinical data, such as patient history or genomic biomarkers. Moreover,

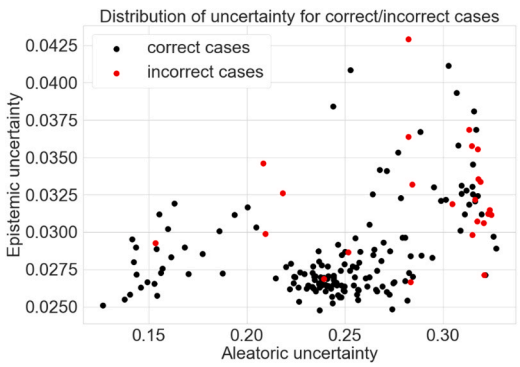

Fig. 6. Distribution of uncertainty values for correct and incorrect predictions. Incorrect cases yield in many cases a high aleatoric uncertainty.

particularly ambiguous cases might benefit from interdisciplinary tumor board discussions to determine whether biopsy or a short-term follow-up scan is warranted. By focusing expert attention on only the subset of nodules deemed uncertain, clinicians can streamline the workflow for straightforward cases while maintaining a high level of diagnostic confidence. This selective approach not only preserves valuable clinical resources but also enhances the reliability of automated predictions in routine practice.

Although our approach flagged many incorrect predictions as uncertain, approximately 13.6 % of incorrect predictions were not identified based on uncertainty. The reason for this effect is likely the absence of similar cases in the training dataset. Alternatively, the ground truth in our dataset is determined by radiologists and might yield some incorrect labels. With the use of a more balanced training dataset might help to also identify these incorrect predictions. We will investigate this in future work.

Many incorrect decisions were associated with high aleatoric uncertainty, indicating that this uncertainty originated from variability in the training data (such as ambiguous cases or overlapping class distributions). We observed that in the volume range where most uncertain cases appeared, both benign and malignant cases were included in the training data. Because aleatoric uncertainty reflects irreducible data noise, it cannot be mitigated simply by improving the model or gathering more data.

Nodule volume was the only radiomic feature selected in all bootstraps. This is in line with previous studies reporting the high predictive power of nodule volume [21,22]. In contrast, no other radiomic feature was selected in more than 25 % of the bootstraps, suggesting that these features possess limited or inconsistent predictive power across different sample subsets. This variability may be attributable to factors such as redundancy among features, or varying relevance depending on individual patient characteristics, further emphasizing the dominant role of nodule volume in the predictive modeling framework.

A considerable number of cases were selected as uncertain, including correct predictions. However, reviewing only these cases carefully would already reduce the workload of physicians by allowing them to prioritize their attention. In future work, we will investigate how to further decrease this number. Possible approaches are the inclusion of other features such as features extracted from CNNs or the application of uncertainty aware training strategies. The LIDC dataset is widely used for developing models for lung nodule classification. Previous studies have performed radiomic analyses on the same dataset and for the same task [23]. In [6], the authors compared the performance of including morphological, textural, or all radiomic features and trained a logistic regression model. The model trained only with morphological features resulted in the best performance, achieving an accuracy of 81 %. Shaffie et al. [24] combined morphological features, shape features modeled using spherical harmonics, and ‚appearance' features describing nodule texture, achieving an accuracy of 93.12 %. Garau et al. [25] trained an SVM classifier that achieved an accuracy of 89 % when applied to a dataset from a different center. In [26], an ensemble radiomic model resulted in an accuracy of 81.68 %. No uncertainty quantification was performed in this case. In terms of evaluation metrics, our results are comparable to previous works. However, in contrast to these works, we also evaluated the uncertainty of the ensemble model to increase the trust in the model.

Recent advances in CT image classification emphasize integrating uncertainty quantification to improve diagnostic accuracy and model reliability. Techniques to date concentrate on uncertainty methods for CNNs such as Monte Carlo Dropout, Deep Ensembles, and Bayesian methods are commonly used to estimate aleatoric and epistemic uncertainties [27,28]. In our work, we focussed on radiomic features as these features are easier to interpret than feature maps extracted from CNNs. A comparison with uncertainty approaches incorporated in CNNs will be conducted in future work.

This study has several limitations. First, ground truth labels were based on expert visual inspection, which is subjective and may contain inaccuracies due to inter- and intra-observer variability, potentially impacting model performance. Second, lung nodules were manually segmented, introducing annotation variability and bias; future work should employ automated segmentation to improve consistency and scalability. Third, the lack of an external validation dataset from a different institution restricts the assessment of generalizability across diverse populations and imaging protocols. Finally, while the ensemble of five classifiers with bootstrapping improves performance, it increases model complexity, complicating interpretability and explainability. Addressing these issues in future research will be essential to enhance

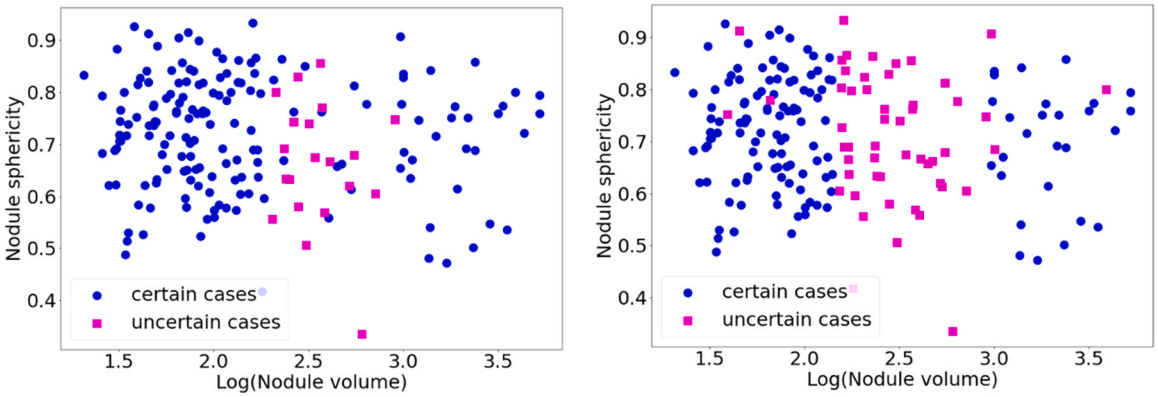

Fig. 7. Feature distribution of the logarithm (log10) of nodule volume and nodule sphericity for certain/uncertain cases for a threshold of 95th percentile (left) and 80th percentile (right). With decreasing threshold, some cases with smaller nodule volumes (<100 mm^3) are included. However, the majority of cases flagged as uncertain yield a nodule volume between 200 and 400 mm^3.

the model's clinical applicability. However, one strength is that the LIDC dataset has been acquired at various hospitals using different scanners and matches, thus creating a realistic scenario.

In summary, our results demonstrate that ensemble models can be used to assess radiomic model uncertainty and can identify a large number of incorrect decisions for the classification of malignant lung nodules.

Conclusion

Our ensemble-based radiomic approach provides an intuitive and effective way to quantify decision uncertainty for lung nodule malignancy classification. This uncertainty flagging can increase clinician confidence in automated pipelines and expedite the integration of radiomics in real-world practice. Future work will focus on optimizing ensemble configurations and assessing the method's robustness across diverse external datasets.

Funding sources

No additional funding was received for this work.

Author Contribution

EP performed formal analysis, data curation, investigation, and wrote the manuscript. LK performed data curation and part of the formal analysis, and reviewed the manuscript. IA and SN reviewed the manuscript, AZ contributed to investigation, data curation and reviewed the manuscript, DT wrote the manuscript and lead the study. All authors reviewed the manuscript.

Ethical statement

No ethical approval was necessary to perform the study.

CRediT authorship contribution statement

Ira Assent: Writing – review & editing, Project administration. **Lena Krieger:** Writing – review & editing, Software, Data curation. **Elisabeth Pfaehler:** Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Daniel Truhn:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Alex Zwanenburg:** Writing – review & editing, Supervision, Methodology, Investigation. **Sven Nebelung:** Writing – review & editing, Project administration.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ejrai.2025.100047.

Data availability

There is a link to the code in the manuscript.

References

[1] H. Sung, J. Ferlay, R.L. Siegel, et al., Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 71 (2021) 209–249, https://doi.org/10.3322/caac.21660.
[2] D.R. Baldwin, Prediction of risk of lung cancer in populations and in pulmonary nodules: significant progress to drive changes in paradigms, Lung Cancer 89 (2015) 1–3, https://doi.org/10.1016/j.lungcan.2015.05.004.
[3] P. Lambin, E. Rios-Velazquez, R. Leijenaar, et al., Radiomics: extracting more information from medical images using advanced feature analysis, Eur. J. Cancer 48 (2012) 441–446, https://doi.org/10.1016/j.ejca.2011.11.036.
[4] J.L. Causey, J. Zhang, S. Ma, et al., Highly accurate model for prediction of lung nodule malignancy with CT scans, Sci. Rep. 8 (2018) 9286, https://doi.org/10.1038/s41598-018-27569-w.
[5] A. Zwanenburg, M. Vallières, M.A. Abdalah, et al., The image biomarker standardization initiative: standardized quantitative radiomics for High-Throughput Image-based phenotyping, Radiology 295 (2020) 328–338, https://doi.org/10.1148/radiol.2020191145.
[6] J. Liu, A. Corti, V.D.A. Corino, L. Mainardi, Lung nodule classification using radiomics model trained on degraded SDCT images, Comput. Methods Prog. Biomed. 257 (2024) 108474, https://doi.org/10.1016/j.cmpb.2024.108474.
[7] B. Hunter, M. Chen, P. Ratnakumar, et al., A radiomics-based decision support tool improves lung cancer diagnosis in combination with the herder score in large lung nodules, eBioMedicine 86 (2022) 104344, https://doi.org/10.1016/j.ebiom.2022.104344.
[8] B. Hunter, C. Argyros, M. Inglese, et al., Radiomics-based decision support tool assists radiologists in small lung nodule classification and improves lung cancer early diagnosis, Br. J. Cancer 129 (2023) 1949–1955, https://doi.org/10.1038/s41416-023-02480-y.
[9] Shen Y., Zhang Z., Sabuncu M.R., Sun L. (2020) Real-Time Uncertainty Estimation in Computer Vision via Uncertainty-Aware Distribution Distillation.
[10] J. Gawlikowski, C.R.N. Tassi, M. Ali, et al., A survey of uncertainty in deep neural networks, Artif. Intell. Rev. 56 (2023) 1513–1589, https://doi.org/10.1007/s10462-023-10562-9.
[11] Chan M.A., Molina M.J., Metzler C.A. (2024) Estimating Epistemic and Aleatoric Uncertainty with a Single Model.
[12] Hofman P., Sale Y., Hüllermeier E. (2024) Quantifying Aleatoric and Epistemic Uncertainty with Proper Scoring Rules.
[13] Ilg E., Çiçek Ö., Galesso S., et al (2018) Uncertainty Estimates and Multi-Hypotheses Networks for Optical Flow.
[14] S.G. Armato, G. McLennan, L. Bidaut, et al., The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans, Med Phys. 38 (2011) 915–931, https://doi.org/10.1118/1.3528204.
[15] E. Pfaehler, A. Zwanenburg, J.R. De Jong, R. Boellaard, RaCaT: an open source and easy to use radiomics calculator tool, PLOS ONE 14 (2019) e0212223, https://doi.org/10.1371/journal.pone.0212223.
[16] E. Pfaehler, I. Zhovannik, L. Wei, et al., A systematic review and quality of reporting checklist for repeatability and reproducibility of radiomic features, Phys. Imaging Radiat. Oncol. 20 (2021) 69–75, https://doi.org/10.1016/j.phro.2021.10.007.
[17] C. Haarburger, G. Müller-Franzes, L. Weninger, et al., Radiomics feature reproducibility under inter-rater variability in segmentations of CT images, Sci. Rep. 10 (2020) 12688, https://doi.org/10.1038/s41598-020-69534-6.
[18] E. Pfaehler, J. Van Sluis, B.B.J. Merema, et al., Experimental multicenter and multivendor evaluation of the performance of PET radiomic features using 3-Dimensionally printed phantom inserts, J. Nucl. Med 61 (2020) 469–476, https://doi.org/10.2967/jnumed.119.229724.
[19] T. Nair, D. Precup, D.L. Arnold, T. Arbel, Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation, Med Image Anal. 59 (2020) 101557, https://doi.org/10.1016/j.media.2019.101557.
[20] S. Lundberg, S.-I. Lee, A Unifi.. Approach Interpret. Model Predict. (2017).
[21] M.L. Welch, C. McIntosh, B. Haibe-Kains, et al., Vulnerabilities of radiomic signature development: the need for safeguards, Radio. Oncol. 130 (2019) 2–9, https://doi.org/10.1016/j.radonc.2018.10.027.
[22] E. Pfaehler, L. Mesotten, I. Zhovannik, et al., Plausibility and redundancy analysis to select FDG-PET textural features in non-small cell lung cancer, Med Phys. 48 (2021) 1226–1238, https://doi.org/10.1002/mp.14684.
[23] K. Mehta, A. Jain, J. Mangalagiri, et al., Lung nodule classification using biomarkers, volumetric radiomics, and 3D CNNs, J. Digit Imaging 34 (2021) 647–666, https://doi.org/10.1007/s10278-020-00417-y.
[24] A. Shaffie, A. Soliman, H.A. Khalifeh, et al., Radiomic-Based framework for early diagnosis of lung cancer. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, Venice, Italy, 2019, pp. 1293–1297.
[25] N. Garau, C. Paganelli, P. Summers, et al., External validation of radiomics-based predictive models in low-dose CT screening for early lung cancer diagnosis, Med Phys. 47 (2020) 4125–4136, https://doi.org/10.1002/mp.14308.
[26] G. Tang, L. Du, S. Ling, et al., Multi-type classification of lung nodules based on CT radiomics and ensemble learning for diversity weighting, Quant. Imaging Med Surg. 14 (2024) 8942–8965, https://doi.org/10.21037/qims-24-1315.
[27] R. Zahari, J. Cox, B. Obara, Uncertainty-aware image classification on 3D CT lung, Comput. Biol. Med 172 (2024) 108324, https://doi.org/10.1016/j.compbiomed.2024.108324.
[28] V.E. Papageorgiou, G. Petmezas, P. Dogoulis, et al., Uncertainty CNNs: a path to enhanced medical image classification performance, Math. Biosci. Eng. 22 (2025) 528–553, https://doi.org/10.3934/mbe.2025020.