RESEARCH ARTICLE

Prediction of extranodal extension in head and neck squamous cell carcinoma by CT images using an evolutionary learning model

Tzu-Ting Huang[1,2†], Yi-Chen Lin[3†], Chia-Heng Yen[2], Jui Lan[4], Chiun-Chieh Yu[5], Wei-Che Lin[5], Yueh-Shng Chen[5], Cheng-Kang Wang[5], Eng-Yen Huang[1,6*] and Shinn-Ying Ho[3,7,8,9*]

Abstract

Background Extranodal extension (ENE) in head and neck squamous cell carcinoma (HNSCC) correlates to poor prognoses and influences treatment strategies. Deep learning may yield promising performance of predicting ENE in HNSCC but lack of transparency and interpretability. This work proposes an evolutionary learning method, called EL-ENE, to establish a more interpretable ENE prediction model for aiding clinical diagnosis.

Methods There were 364 HNSCC patients who underwent neck lymph node (LN) dissection with pre-operative contrast-enhanced computerized tomography images. All the 778 LNs were divided into training and test sets with the ratio 8:2. EL-ENE uses an inheritable bi-objective combinatorial genetic algorithm for optimal feature selection and parameter setting of support vector machine. The diagnostic performances of the ENE prediction model and radiologists were compared using independent test datasets.

Results The EL-ENE model achieved the test accuracy of 80.00%, sensitivity of 81.13%, and specificity of 79.44% for ENE detection. The three radiologists achieved the mean diagnostic accuracy of 70.4%, sensitivity of 75.6%, and specificity of 67.9%. The features of gray-level texture and 3D morphology of LNs played essential roles in predicting ENE.

Conclusions The EL-ENE method provided an accurate, comprehensible, and robust model to predict ENE in HNSCC with interpretable radiomic features for expanding clinical knowledge. The proposed transparent prediction models are more trustworthy and may increase their acceptance in daily clinical practice.

Keywords Head and neck squamous cell carcinoma, Extranodal extension, Radiomics, Evolutionary learning, Artificial intelligence

---

†Tzu-Ting Huang and Yi-Chen Lin contributed equally.

*Correspondence:
Eng-Yen Huang
hey1200@cgmh.org.tw
Shinn-Ying Ho
syho@nycu.edu.tw
Full list of author information is available at the end of the article

Huang *et al. Cancer Imaging*   (2023) 23:84

Page 2 of 11

Introduction

Extranodal extension (ENE) is a pathological diagnosis defined by the College of American Pathologists lip and oral cavity cancer protocol as "extension of metastatic tumor, present within the confines of the lymph node (LN), through the LN capsule into the surrounding connective tissue, with or without associated stromal reaction" [1]. ENE is a poor prognostic factor associated with increased locoregional failure, distant metastases, and reduced overall survival in patients with head and neck squamous cell carcinoma (HNSCC) [2–4].

The presence of ENE is critical in clinical decision-making. For patients with ENE-positive HNSCC, concurrent chemoradiotherapy may yield similar treatment outcomes to patients receiving surgery followed by adjuvant chemoradiation, while providing fewer treatment-related acute and late toxicities, and lower healthcare costs [5–8]. Therefore, developing an accurate, robust, and trustworthy prediction model to distinguish the ENE status before the definitive treatment is important to guide the best therapy for HNSCC patients.

Contrast-enhanced computed tomography (CT) scan is the most widely used method to predict ENE status for HNSCC patients in clinical practice. However, the literature revealed that this method has limited diagnostic performance, with reported sensitivity ranging from 43.7 to 69% and the area under the receiver operating characteristic curve (AUC) ranging from 0.6 to 0.69 [9–13]. Furthermore, high inter-observer variability is also reported [9, 11–13].

To improve the diagnostic performance of ENE by CT scanning, two studies applied deep learning methods to establish prediction models for ENE detection [14, 15]. Both studies showed excellent results with AUC of 0.91 and 0.82 for ENE prediction. Although deep learning models yield attractive results, these models often work as black-boxes with limited transparency and interpretability [16]. It is difficult for clinicians to correlate the results of these deep learning models with known radiomic features of ENE.

Identification of effective radiomic features plays a vital role in advancing prediction performance and providing interpretability associated with clinical knowledge. Lee et al. proposed an evolutionary learning (EL) method for establishing clinical-radiomic models to predict the early recurrence of hepatocellular carcinoma after resection, better than other well-known machine learning (ML) derived models [17]. This EL method aims to optimize the feature selection and model parameters in establishing ML models.

In this work, we use the novel EL approach to identifying a set of interpretable radiomic features. The proposed method EL-ENE uses the inheritable bi-objective combinatorial genetic algorithm (IBCGA) [18] with an intelligent evolutionary algorithm (IEA) [19] for optimal feature selection and parameter setting of support vector machine (SVM) to establish an interpretable model for predicting ENE by CT scanning.

Materials and methods

Patient selection, image acquisition, and characteristics

The medical records of consecutive patients with histologically proven HNSCC from 1 to 2009 to 31 October 2017 were reviewed retrospectively. Three hundred and sixty-four HNSCC patients who underwent neck LN dissection with preoperative contrast-enhanced diagnostic head and neck CT scans were enrolled. Exclusion criteria included previous neck surgery, preoperative chemotherapy/chemoradiotherapy, LN short axis<1 cm on CT images, and the time between staging CT to LN dissection over 6 weeks. The Institutional Review Board of our institution approved this study (201801181B0/201801181 B0C501/201801181B0C601).

The head and neck CT scans were performed on a 64-channel scanner (Aquilion 64, Toshiba Medical Systems, Tokyo, Japan), 80-channel scanner (Aquilion Prime, Canon Medical Systems, Otawara, Japan) or 256-channel scanner (Siemens Healthcare AG, Erlangen, Germany) with the following parameters: tube current 100–550 mAs; voltage 120 kVp; gantry rotation time 0.5 s; pitch 0.969 mm/rotation; detector collimation 80×0.5 mm; field of view 22 cm; and 3 mm axial reconstruction thickness. The CT images extend from the upper orbital rim through the upper thorax. Enhanced images were obtained 60 s after intravenous injection of 1.0 mL/kg CT contrast (Omnipaque 350, GE Healthcare, Princeton, New Jersey) at a rate of 2.0 mL/second. The CT scans were reviewed on a commercial Picture Archiving and Communication System (PACS) workstation. (Centricity RA 1000; GE Healthcare, Chicago, IL, USA).

All pathology specimens were collected and reviewed by one head and neck pathologist (J. Lan) to avoid interobserver variation. ENE was defined as tumor infiltrating from the capsule of a metastatic LN [1]. For each LN, a one-to-one matching between the pre-operative CT images and the pathology report was obtained according to the LN's laterality, anatomical level, and nodal size. If there were more than one LN with a similar size at the same region on the CT image where a definite correlation could not be derived, these LNs were not included in the study. The regions of interest (ROIs) were delineated manually at the edge of LNs on each slice in the axial plane and were recorded in the RT structure set (RTSS) label file. The segmentation process was done by one radiation oncologist (T.T. Huang) to ensure contouring consistency.

The CT images used included 364 patients with 778 3D LN images. The dataset contained 375 normal LNs, 139

Huang *et al. Cancer Imaging*   (2023) 23:84

Page 3 of 11

Table 1 Demographics (364 patients; 391 primary sites; 778 LNs)

Characteristic	Value
Age	
Mean	54.4 ± 10.8
Gender	
Male	333 (91.5%)
Female	31 (8.5%)
Primary cancer site	
Oral cavity	289 (73.9%)
Oropharynx	47 (12.0%)
Hypopharynx	36 (9.2%)
Larynx	17 (4.4%)
Salivary glands	2 (0.5%)
Pathological T stage	
T1	80 (20.5%)
T2	106 (27.1%)
T3	35 (9.0%)
T4	162 (41.4%)
Biopsy only	8 (2.0%)
Pathological N stage	
N0	134 (36.8%)
N1	62 (17.0%)
N2	164 (45.1%)
N3	4 (1.1%)
p16 status	
Positive	8 (2.2%)
Negative	337 (92.6%)
Unknown	19 (5.2%)
LN status	
Negative	375 (48.2%)
Metastatic with ENE(-)	139 (17.9%)
Metastatic with ENE(+)	264 (33.9%)

metastasis LNs, and 264 ENE LNs. The CT image format was Digital Imaging and Communications in Medicine (DICOM), and the size was 512*512 pixels. Among them, 22 patients had synchronous head and neck cancers with 391 primary sites. The most common primary disease site was the oral cavity. Only 2.2% of patients in the cohort had positive p16 status. The detailed patient characteristics are listed in Table 1.

The 778 3D LN images were divided into a training set and a test set by the approximate ratio 8:2. The training set had 618 LNs of 314 patients, including 296 negative LNs, 111 metastasis LNs, and 211 ENE LNs. The test set had 160 LNs of 50 patients, including 79 negative LNs, 28 metastasis LNs, and 53 ENE LNs.

The proposed method EL-ENE

The proposed method EL-ENE used an evolutionary learning approach to identifying a small set of radiomics features while maximizing the prediction accuracy. Figure 1 shows the flowchart of EL-ENE, including image pre-processing, feature extraction, feature selection, and ensemble classifier of SVM [20].
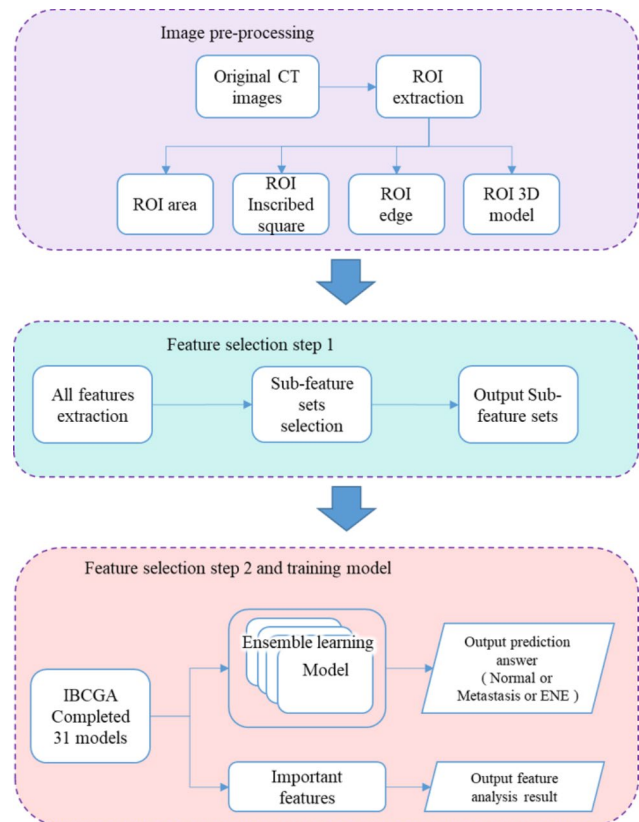

Fig. 1 The flowchart of EL-ENE including image pre-processing, feature extraction, feature selection, and ensemble classifier of support vector machine

Image pre-processing

The image pre-processing for extracting ROIs includes three main tasks, including (1) extraction of the volume of Interest (VOIs), (2) superimposition of CT image and RTSS annotation, and (3) extraction of ROIs from the DICOM images. First, the new Window Center was adjusted using Window Center, Rescale Slope, and Rescale Intercept in the DICOM file header. Then, the VOI was calculated using the new Window Center and Window Width, and was normalized into the range of [0, 255]. The coordinate information of the ROI was recorded in the RTSS annotation file. The normalized CT image was superimposed with the ROI coordinate, and the desired ROI contour position map in the DICOM file was obtained.

The boundary information of LNs is highly associated with the ENE. To ensure that the contours detected were complete, we extracted the accurate ROI using morphological operations, including dilation, fill, and erosion. Finally, the designed mask using the morphological operations was operated on the calibrated CT images, and the tomographic images of the LN sections were extracted. The Imdilate, imfill, and imerode functions in the Matlab tool were used to extract the ROI boundary. In addition,

we also extracted the ROI inscribed square and the ROI contour information for subsequent image analysis, e.g., feature extraction from the gray level change inside and outside the ROIboundary.

Feature extraction

We extracted gray level, geometric, morphological, and texture features from CT images of LNs as candidate features. There were 460 candidate features, which were categorized into six types of features with 26 feature subsets (Table 2). The six types were Gray-Level Co-occurrence Matrix (GLCM), Gray-Level Size Zone Matrix (GLSZM), Gray-Level, LN morphology, LN boundary, and Invariant moment.

The GLCM, GLSZM, and Invariant moment features were extracted from the largest inscribed square of the largest ROI section in the LN. The GLCM features reflect the texture distribution by counting the gray level changes between the two pixels at the space in various angles and distances. GLCM features contain four types of gray quantitative features [21] and 14 types of Haralick features [22], including Cluster Shade, Cluster Proximity, Contrast, Correlation, Different Entropy, Different Variance, Dissimilarity, Energy, Entropy, Homogeneity

Table 2 The 26 subsets belonging to six feature types

	Feature type	Subset of features	Number of features
1	Gray-Level Co-occurrence Matrix	Cluster Shade	20
		Cluster Proximity	20
		Contrast	20
		Correlation	20
		Different Entropy	20
		Different Variance	20
		Dissimilarity	20
		Energy	20
		Entropy	20
		Homogeneity Normalized	20
		Homogeneity	20
		InfoCorrelation1	20
		InfoCorrelation2	20
		Max Probability	20
		Sum Average	20
		Sum Entropy	20
		Sum of variance	20
		Variance	20
2	Gray-Level Size Zone Matrix	GLSZM	11
3	Gray-Level	Gray level	11
4	LN morphology	3D Morphology	29
		2D Morphology	24
5	LN boundary	Edge 3	6
		Edge 5	6
		Edge 10	6
6	Invariant moment	Invariant moment	7

Normalized, Homogeneity, InfoCorrelation1, InfoCorrelation2, Max Probability, Sum Average, Sum Entropy, Sum of variance, and Variance. Each feature type contains 20 features calculated from 20 GLCMs with different angles and distances. The parameter range of GLCM was set to 16 grey levels, the directions were 0°, 45°, 90, and 135°, and the distances were integers from 1 to 5.

GLSZM calculated the change of gray levels in ROI by quantizing the gray area in the image [23]. Unlike GLCM, GLSZM calculated a matrix for the domains connected in all directions for the same gray levels, regardless of rotation and distance. The parameter of GLSZM for the gray level range was set to 16 levels, and 11 features can be obtained, including small area emphasis, large area emphasis, low intensity emphasis, high intensity emphasis, low intensity small area emphasis, low intensity large area emphasis, high intensity small area emphasis, high intensity large area emphasis, intensity variance, size zone variance, and zone%. Invariant moments are often used as optical character recognition and shape recognition features in images. Their moment invariance was not changed by the rotation, translation, and scaling of images [24, 25]. Through the second-order and third-order central moments, seven invariant moments were obtained as features.

Gray-Level and 2D LN morphology features were extracted from the largest ROI section in the LN. For the gray level features of ROI, they show the statistical analysis of the numerical changes of the gray levels in the ROI, including ten features such as Mean, Median, Variance, Standard deviation, Maximal of gray levels, Minimal of gray levels, Skewness, Kurtosis, Energy, and Entropy. The 2D morphological features are the surface configuration of image objects, which are essential in distinguishing LNs. 24 features were collected, including Area, Perimeter, Major Axis Length, Minor Axis Length, Orientation, Convexity, Convex Area, Convex Perimeter, Maximum radius, Bounding Box Area, Defects Ratio, Perimeter Area Ratio, Aspect Ratio, Bending Energy, Eccentricity, Equivalent Diameter, Solidity, Extent, Compactness, Rectangularity, Elongation, Roundness, Ellipiticty, and Sphericity.

3D LN morphology features were extracted from the 3D LN model. First, a series of LN CT images were stacked. Then, the height of the stacked 3D LN model was corrected by using the actual width of the pixel (e.g., 0.4680 mm) and slice thickness (e.g., 3 mm) recorded in DICOM head file. Finally, Delaunay triangulation was used to smooth the surface of interpolated LNs (using the interp3 function of Matlab). Twenty-nine features were collected, including Volume, Surface, Equivalent diameter, Extent, three Principal Axis Length, three Orientation, Eccentricity, Solidity, Convex volume, Convex surface, Convexity, Compactness, Rectangularity,

Huang *et al. Cancer Imaging*   (2023) 23:84

Page 5 of 11

Elongation, Roundness, Area volume ratio, three Aspect radius, Maximum radius, Bounding box volume, Ellipiticty, Defect ratio, Gaussian Curvature sum, and Mean Curvature sum.

The LN boundary was extracted from the ROI section in the LN. For boundary features, the gray level changes inside and outside the ROI area are related to whether the LNs expand outside the LNs. The Imdilate function of Matlab was used to extract the ROI boundary area, which is dilated with disc-shaped structural elements, considering the radius of the disc shape with 3, 5, and 10 pixels. For each ROI boundary area on CT images, 6 LN boundary features were extracted, including Mean inside the ROI, Mean outside the ROI, Standard variance inside the ROI, Standard variance outside the ROI, differences of Mean, and Standard variance between inside and outside the ROI.

Feature selection

Due to the large number of candidate features, EL-ENE used a coarse-to-fine feature selection. The coarse step is to independently evaluate each of the 26 feature subsets using the classification accuracy of SVM in terms of 10-fold cross-validation (10-CV). For each feature subset, three SVM models were established to evaluate feature subsets. The three models predicted a LN as (1) normal or metastatic, (2) ENE or non-ENE, and (3) normal, metastatic, or ENE. For each model, we selected the top five feature subsets ranked by prediction accuracy. Experimental results revealed seven feature subsets with 89 features, including Sum of variance, GLSZM, Gray-Level, 3D Morphology, Edge 3, Edge 5, and Edge 10.

The fine step used an IBCGA [18, 19] cooperated with SVM to select a minimal number of features while maximizing prediction accuracy. IBCGA selects m form n (=89 in this study) features and determines the parameter setting of SVM for training the prediction models. Since IBCGA is a non-deterministic algorithm, the obtained SVM models with identified features were not always the same. EL-ENE establishes an ensemble SVM classifier consisting of 31 SVM models with different sets of features that predicts LNs as normal, metastasis, or ENE.

The customized IBCGA

EL-ENE uses an evolutionary learning approach to optimizing the system parameters in designing an interpretable classifier. The customized IBCGA algorithm was used to select a small number m from a large number n of radiomics features and determine two parameter values of the SVM model, cost C and γ of the kernel function.

The simultaneous optimization of feature selection and SVM parameters play a vital role in modeling. The m features can be ranked according to the prediction contribution using the main effect difference. Some applications of IEA and IBCGA in designing prediction models for biomedicine research can refer the studies [26–29].

In EL-ENE, the fitness function of IBCGA is to maximize the prediction accuracy of 10-CV on the training dataset. The best value of m was automatically determined belonging to the range [r_{end}, r_{start}]. The parameter settings of IBCGA were as follows: N_{pop}=50, P_s = 1.0, P_c = 0.8, P_m = 0.05, G_{max} = 100, r_{start} =70, and r_{end}=5. The main steps of IBCGA are as follows.

Step 1. Initialization: Generate a population of N_{pop} individuals randomly where each contains $r = r_{start}$ selected features, (n-r_{start}) unselected features, C and γ. $G = 0$.

Step 2. Evaluation: Evaluate all individuals using the fitness function.

Step 3. Selection: Select $P_s \times N_{pop}$ individuals by a tournament selection method to form a mating pool.

Step 4. Crossover: Perform the orthogonal array crossover of IEA [19] on randomly selected $P_c \times N_{pop}$ individuals.

Step 5. Mutation: Randomly select $P_m \times N_{pop}$ individuals excluding the best one to mutate using a bit-swap operation.

Step 6. Termination test: Increase the number G by one. If $G = G_{max}$, output the best individual in the population as X_r, $G = 0$, and go to Step 7. Otherwise, go to Step 2.

Step 7. Inheritance: If $r > r_{end}$, randomly mutate a binary gene from 1 to 0 for each individual, decrease the value of r by 1, and go to Step 2.

Step 8. Output: Let X_m with m selected features be the best individual among X_r where $r = r_{end}$, r_{end}+1, …, r_{start}.

Radiologists' review protocol

Three neuroradiologists with more than 4 years of experience in head and neck imaging were recruited for assessing the status of ENE. LNs in the test data sets were annotated with serial numbers for review. Five radiomic features were applied for judging ENE presence, including irregular nodal enhancement, poorly defined nodal margins, infiltration of the adjacent fat plane, central necrosis, and matted nodes.

According to the 5 imaging features, the observers concluded the probability of ENE based on a 5-point rating score: 1, definitely not ENE; 2, likely not ENE; 3, equivocal ENE; 4, likely ENE and 5, definitely ENE. Scores 1 and 2 were deemed negative ENE while scores 3–5 were considered positive ENE [9, 11].

Model evaluation and statistical analysis

The diagnostic performance of the prediction model was evaluated on the independent test data set using AUC, sensitivity, specificity, accuracy, positive and negative

Huang *et al. Cancer Imaging* (2023) 23:84

Page 6 of 11

predictive values. The statistics were performed by R version 4.02 (The R Foundation for Statistical Computing, Vienna, Austria) and SPSS version 22.0 software (SPSS, Chicago, IL).

Results
Subset feature evaluation
Three types of prediction ability were tested for selecting the promising subset features as candidate ones for the feature selection of IBGGA. Table A1 listed the top five subset features with high 10-CV accuracy, which can distinguish the metastatic LNs. The top five subset features were Gray-level, Edge 10, Sum of variance, 3D morphology, and Edge 5. Table A2 listed the top five subset features which can distinguish the ENE LNs. The top five subset features were 3D morphology, Gray-level, Edge 5, GLSZM, and Edge 10. Table A3 listed the top five subset features which can distinguish three classes of LNs. The top five subset features were 3D morphology, Gray-level, Edge 3, Sum of variance, and Edge 10.

The 3D morphology, Gray-level, and Edge 10 were selected in three types of evaluations. The Sum of variance and Edge 5 were selected in two of them, and Edge 3 and GLSZM were selected once. The results show that morphology, gray level, and edge features were important in distinguishing the LN types. The seven subsets with 89 features were selected as the input feature for the EL-ENE method.

Feature selection results
A set of features were selected from a total of 89 features through IBCGA. Then, the ensemble classifier with 31 stable models with different feature combinations was

Table 3 The selection times of features in the ensemble classifier consisting of 31 SVMs

Subset feature	Feature name	Times of selection
GLSZM	Low intensity small area emphasis	31
3D morphology	Orientation2	30
GLSZM	Zone%	29
GLSZM	High intensity large area emphasis	29
Grey level	Median	28
Grey level	Max Pixel Value	28
3D morphology	Orientation3	27
Grey level	Variance	26
GLSZM	small area emphasis	24
3D morphology	Solidity	21
3D morphology	Max radius	19
3D morphology	Area	18
Grey level	Energy	17
Sum of variance	D1A45	16
3D morphology	Compactness	16

established, and the final model predicted the answer of the LN types by voting on 31 models. In the 31 models, Gray-Level features were the most frequently selected subset features, followed by Edge features, Sum of variance, and 3D morphology features. Among them, 3D morphology features were mainly suitable for distinguishing ENE LNs.

Each of the 31 models had a satisfactory prediction ability. From the analysis of the subset features of the 31 models, the features that were selected more than 16 times represent that they had a significant influence on the voting process and were the most influential. The top-rank features in the best combination feature set were shown in Table 3. The GLSZM subset feature contained Low intensity small area emphasis, Zone%, High intensity large area emphasis, and small area emphasis. Among them, small area emphasis had the smallest p-value, 3.454e-09.

Four Grey-level subset features were selected, including Median, Max Pixel Value, Variance, and Energy, where Variance had the smallest p-value, 1.821e-18. The normal LNs had the most significant value of Variance, and ENE LNs had the smallest value of Variance. In addition, the same results were found in the analysis of D1A45 (distance 1, direction 45 degrees) in Sum of variance of GLCM.

The difference between the small area emphasis, Variance, and Sum of variance D1A45 was that the small area emphasis focused on the grey level change related to the size of the area changed, the Variance focused on the grey level change of the entire image, and the Sum of variance D1A45 focused on the grey level change in specific distances and angles.

Six 3D morphology subset features were selected, including Orientation2, Orientation3, Solidity, Max radius, Area, and compactness. Solidity had the smallest p-value, 7.182e-42. Solidity represented the irregularity of the surface. The boxplot of the small area emphasis, Variance, Sum of variance D1A45, and Solidity in three types of LNs were shown in Fig. 2.

Figure 3 showed the inscribed squares in ROI and 3D models of the three types of LNs, including normal LNs (no. 152), metastatic LNs (no. 246), and ENE LNs (no. 61). Although it was difficult to distinguish the difference in texture with the human eye [30], the analysis revealed valuable information that normal LNs have the largest values of the small area emphasis, Variance and Sum of variance D1A45. For the 3D features, normal LNs had the smallest value of Solidity, and ENE LNs had the most significant value of Solidity.

Prediction performance of EL-ENE model and radiologists
The EL-ENE method established 31 independent prediction models. All the results were counted, the voting
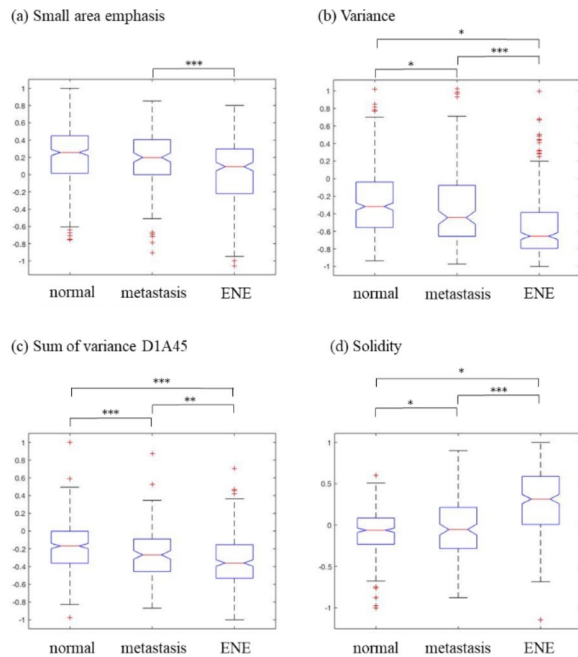
Fig. 2 The boxplots of (**a**) Small area emphasis, (**b**) Variance, (**c**) Sum of variance D1A45, and (**d**) Solidity in the normal, metastatic, and extranodal extension lymph nodes
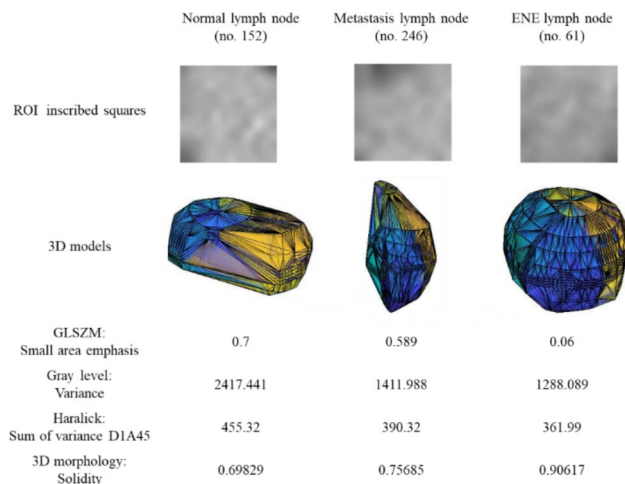

Fig. 3 The regions of interest inscribed squares and 3D models of normal lymph nodes (no. 152), metastatic lymph nodes (no. 246), and ENE lymph nodes (no. 61)

method was adopted, and the final answer was decided by a majority. The EL-ENE ensemble model was trained by 618 LNs and independently tested by 160 LNs.

The EL-ENE ensemble model achieved test accuracy of ENE prediction 80.00%, sensitivity 81.13%, specificity 79.44%, PPV 66.15%, NPV 90.32%, and AUC: 82.51%. For metastasis prediction, the prediction model achieved accuracy 77.50%, sensitivity 70.37%, specificity 84.81%, PPV 82.61%, NPV 73.63%, and AUC 83.41%. (Table 4; Fig. 4)

The radiologists' prediction performance for ENE prediction achieved test accuracy 70.44%, sensitivity 75.64%, specificity 67.91%, PPV 50.04%, NPV 85.43%, and AUC 71.78%. For metastasis prediction, the prediction performance was as follows: accuracy 73.79%, sensitivity 76.54%, specificity 70.94%, PPV 74.53%, NPV 74.67%, and AUC 71.87%. The EL-ENE model performs significantly better prediction performance than two of the radiologists (p-value=0.0006 and 0.002), and no statistically significant difference with the third radiologist (p-value=0.654). (Table 4)

Discussion

We have proposed an evolutionary learning method for establishing a transparent and interpretable ensemble classifier to predict metastatic and ENE LNs in HNSCC patients. This model shows superior classification ability to the radiologists while providing exquisite interpretable information to physicians. Many selected radiomics features can find reasonable clinical or pathological relevance. For example, small area emphasis, the most popular feature selected in the classification model, may represent the invasion of tumor cells and necrotic changes in a metastatic or ENE LN [31]. A larger small area emphasis value means a finer texture in the small area. In our data, normal LNs possess the largest value of small area emphasis. As the pathological changes of cancer cell invasion and necrosis development progress, this value decreased in the metastatic LNs and became the smallest value in the ENE LNs. Our study also finds that the 3D morphology features, which are rarely mentioned in published literature, are powerful for detecting ENE LNs. These implicit and subtle features may provide further clinical insights for ENE image evaluation in the future. The results revealed that an interpretable model can not only provide excellent prediction ability but also correlate the association between the radiomics features and novel clinical knowledge.

Interpretability is critical for clinical prediction models. Understanding the correlation between the input data, the prediction results, and the principles of decision-making behind the EL algorithms may gain the trust and confidence of the prediction models to the clinical practitioners [32] because clinical decision-making is based on logical reasoning, rigorous inference, and solid evidence [33–35]. Due to the lack of interpretability, clinicians may be more conservative in applying black-box algorithms to support clinical decision-making, especially in the high-stake clinical scenarios [36]. It is also challenging to detect or even be aware of potential model errors or biases in an opaque prediction model [37]. Furthermore, an interpretable prediction model might discover comprehensible novel information for future clinical practice [38]. That is to say, clinicians may learn new knowledge
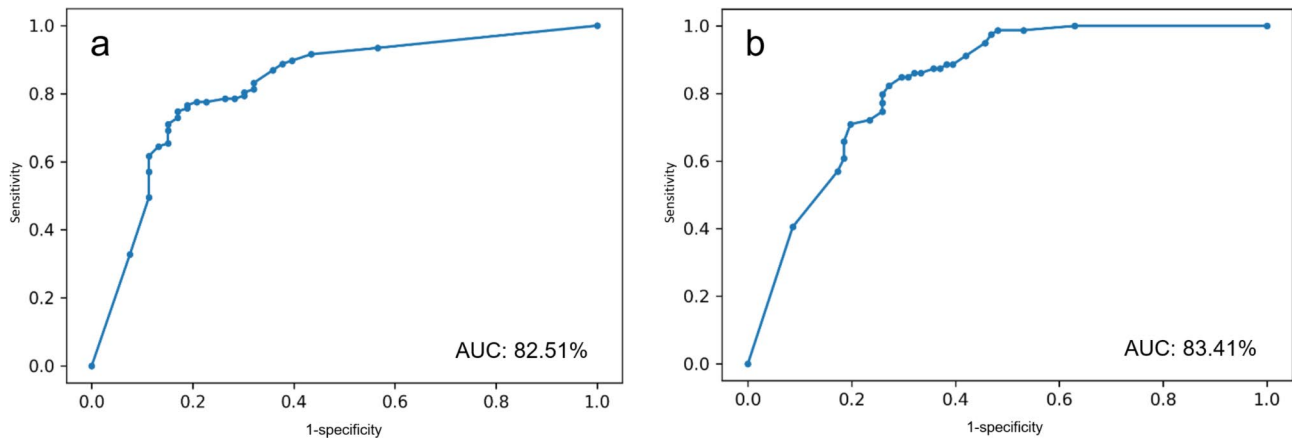
Huang *et al. Cancer Imaging* (2023) 23:84

Page 8 of 11

Fig. 4 The receiver operating characteristic curve of extranodal extension prediction and nodal metastasis prediction models on an independent test set

Table 4 The prediction performance of the EL-ENE model and radiologists

	EL-ENE model				Radiologists	
	ENE Prediction		**Metastatic Prediction**		**ENE Prediction**	**Metastatic Prediction**
	Training Set	**Test Set**	**Training Set**	**Test Set**	**Test Set**	**Test Set**
Accuracy	82.52%	80.00%	78.96%	77.50%	70.44%	73.79%
Sensitivity	79.15%	81.13%	66.77%	70.37%	75.64%	76.54%
Specificity	84.27%	79.44%	92.23%	84.81%	67.91%	70.94%
PPV	72.81%	66.15%	89.92%	82.61%	50.04%	74.53%
NPV	88.46%	90.32%	71.58%	73.63%	85.43%	74.67%

from the interpretable models by analyzing their "thinking process". Consequently, interpretable EL-based medical applications are more trustworthy, robust, creative, and more feasible for clinical practice.

Deep learning has been heavily applied to medical image research for constructing appealing high-accuracy diagnostic and prediction models in individual studies in recent years [39–45]. In ENE detection, Kann and his colleagues developed the first deep learning 3D convolutional neural network model with impressive diagnostic performance and comprehensive external validation [15, 46, 47]. Both this deep learning model and our EL-ENE model outperformed most radiologists in ENE detection. Their AUCs are numerically higher than our model, although we cannot directly compare results from different data sets. Deep learning algorithms such as convolutional neural networks can automatically and adaptively learn complex imaging features and establish sophisticated models [48]. Therefore, with sufficient data, these models might catch essential features beyond the predefined radiomic features and potentially achieve better prediction outcomes.

However, the widespread adoption of deep learning models into daily clinical practice is yet to be established [49]. One major reason for this disproportionate phenomenon is that the data-driven nature of deep learning models is often referred to as black-box algorithms [40]. Therefore, most early deep learning applications have

the inherent shortcomings of intransparency and uninterpretability. These model defects may erode the physicians' confidence in deep learning models and further restrict the wide acceptance of these models into clinical practice. Recently, there has been increasing research on interpretable deep learning models to mitigate the opacity and uninterpretability of deep learning models [50, 51]. Various methods have been developed for building more interpretable deep learning models with promising results [51]. With the rapid progress of interpretable ML, a more comprehensive deep learning algorithm might create more trustworthy prediction models and increase the adoption of its applications into clinical practice in the future.

The diagnostic and prediction power of ML models is not always unlimited. In our case, the physical limitations of the diagnostic CT images may restrict its accuracy for recognizing metastatic or ENE LNs. For example, the z-axis resolution in standard diagnostic helical CT images with 2–3 mm slice thickness may not be sufficient to identify subtle micro ENE [52]. Moreover, uncertainties from CT homogeneity, Hounsfield number accuracy, image linearity, noise interference, and artifact may further hamper the diagnostic ability of CT images [53]. Therefore, if a ML model provides exaggerated results beyond our expectations, we should carefully examine that model for potential errors or biases. Undoubtedly, an

interpretable ML model is also more applicable for this purpose.

This study has several limitations. First, all images were collected in a single institution. The generalizability of this model should be further validated. Second, some LN data were discarded during data collection because a definite correlation between CT images and pathology reports could not be established. Finally, the CT slice thickness is 2–3 mm which may limit the special resolution. Some subtle image features might be blurred due to this relatively thick CT slice thickness.

Future research is warranted to overcome the above limitations. First, external validation is essential for evaluating model generalizability and the robustness and consistency of selected radiomics features. This process could further strengthen the reliability of this explainable EL-ENE model and increase confidence in applying this model in clinical practice. Second, modern medical imaging, such as high-resolution CT or magnetic resonance imaging, might further improve the model's performance. These advanced medical images possess more clinical information and better resolution for discriminating subtle image features such as micro ENE. With a similar model build-up process, we can build an enhanced EL-ENE model with these modern medical images with potentially better performance.

Conclusions

In addition to the pursuit of accurate ENE prediction models, a transparent ML algorithm may provide more comprehensible and robust models for medical applications. Furthermore, these models may explore novel features to expand our clinical knowledge. We believe that more clinicians will be pleased to adopt these trustworthy applications into their daily practice in the future.

List of abbreviations

10-CV	10-fold cross-validation
AUC	Area under the receiver operator characteristic curve
CT	Computed tomography
DICOM	Digital imaging and communications in medicine
EL	Evolutionary learning
ENE	Extranodal extension
HNSCC	Head and neck squamous cell carcinoma
IBCGA	Inheritable bi-objective combinatorial genetic algorithm
IEA	Intelligent evolutionary algorithm
LN	Lymph node
ML	Machine Learning
ROI	Regions of interest
RTSS	RT structure set
SVM	Support vector machine
VOI	Volume of interest

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40644-023-00601-7.

Supplementary Material 1

Data Availability
The datasets generated during and/or analyzed in the current study are available from the corresponding authors upon reasonable request.

Declarations

Ethics approval and consent to participate
The Institutional Review Board of Kaohsiung Chang Gung Memorial Hospital approved this study (201801181B0/201801181B0C501/201801181B0C601).

Consent for publication
All authors gave consent for publication.

Competing interests
All authors have no conflicts of interest to disclosure.

Author details
[1]Department of Radiation Oncology and Proton & Radiation Therapy Center, Kaohsiung Chang Gung Memorial Hospital, Chang Gung University College of Medicine, No. 129, Dapi Road, Niaosong District, Kaohsiung, Taiwan
[2]Institute of Computer Science and Engineering, National Yang Ming Chiao Tung University, No. 1001 University Road, Hsinchu, Taiwan
[3]Institute of Bioinformatics and Systems Biology, National Yang Ming Chiao Tung University, No. 75 Po- Ai Street, Hsinchu, Taiwan
[4]Department of Anatomic Pathology, Kaohsiung Chang Gung Memorial Hospital, Chang Gung University College of Medicine, No. 123, Dapi Road, Niaosong District, Kaohsiung, Taiwan
[5]Department of Diagnostic Radiology, Kaohsiung Chang Gung Memorial Hospital, Chang Gung University College of Medicine, No. 123, Dapi Road, Niaosong District, Kaohsiung, Taiwan
[6]School of Medicine, College of Medicine, National Sun Yat-sen University, No. 70, Lienhai Rd, 80424 Kaohsiung, Taiwan
[7]Department of Biological Science and Technology, National Yang Ming Chiao Tung University, No. 1001 University Road, Hsinchu, Taiwan
[8]Center for Intelligent Drug Systems and Smart Bio-Devices (IDS 2 B), National Yang Ming Chiao Tung University, No. 75 Po-Ai Street, Hsinchu, Taiwan
[9]College of Health Sciences, Kaohsiung Medical University, No. 100, Shih-Chuan 1st Road, Sanmin District, Kaohsiung, Taiwan

Huang *et al. Cancer Imaging*   (2023) 23:84

Page 10 of 11

References

1. Raja R, Seethala MIW, Martin MD, Bullock J, Diane MD, Carlson L, Robert MD, Ferris L, Louis MDPhD, Harrison B, Jonathan MD, McHugh B, Jason MD, Pettus MD, Mary S, Richardson MD, Jatin Shah MD, Lester DR, Thompson MD, Bruce M. Wenig, MD. Protocols for the examination of specimens from patients with carcinomas of the lip and oral cavity. 2017; 4.0.0.1:[Available from: https://documents.cap.org/protocols/cp-headandneck-lip-oralcavity-17protocol-4001.pdf.

2. Ferlito A, Rinaldo A, Devaney KO, MacLennan K, et al. Prognostic significance of microscopic and macroscopic extracapsular spread from metastatic tumor in the cervical lymph nodes. Oral Oncol. 2002;38(8):747–51. https://doi.org/10.1016/s1368-8375(02)00052-0.

3. Grandi C, Alloisio M, Moglia D, Podrecca S, et al. Prognostic significance of lymphatic spread in head and neck carcinomas: therapeutic implications. Head Neck Surg. 1985;8(2):67–73. https://doi.org/10.1002/hed.2890080202.

4. Schuller DE, McGuirt WF, McCabe BF, Young D. The prognostic significance of metastatic cervical lymph nodes. Laryngoscope. 1980;90(4):557–70. https://doi.org/10.1288/00005537-198004000-00001.

5. Johnson DE, Burtness B, Leemans CR, Lui VWY, et al. Head and neck squamous cell carcinoma. Nat Rev Dis Primers. 2020;6(1):92. https://doi.org/10.1038/s41572-020-00224-3.

6. Chen WY, Chen TC, Lai SF, Liang TH, et al. Outcome of bimodality definitive chemoradiation does not differ from that of trimodality upfront neck dissection followed by adjuvant treatment for > 6 cm lymph node (N3) head and neck cancer. PLoS ONE. 2019;14(12). https://doi.org/10.1371/journal.pone.0225962. e0225962.

7. Sher DJ, Fidler MJ, Tishler RB, Stenson K, et al. Cost-effectiveness analysis of Chemoradiation Therapy Versus Transoral robotic surgery for Human Papillomavirus-Associated, clinical N2 Oropharyngeal Cancer. Int J Radiat Oncol Biol Phys. 2016;94(3):512–22. https://doi.org/10.1016/j.ijrobp.2015.11.006.

8. Ling DC, Chapman BV, Kim J, Choby GW, et al. Oncologic outcomes and patient-reported quality of life in patients with oropharyngeal squamous cell carcinoma treated with definitive transoral robotic surgery versus definitive chemoradiation. Oral Oncol. 2016;61:41–6. https://doi.org/10.1016/j.oraloncology.2016.08.004.

9. Maxwell JH, Rath TJ, Byrd JK, Albergotti WG, et al. Accuracy of computed tomography to predict extracapsular spread in p16-positive squamous cell carcinoma. Laryngoscope. 2015;125(7):1613–8. https://doi.org/10.1002/lary.25140.

10. Prabhu RS, Magliocca KR, Hanasoge S, Aiken AH, et al. Accuracy of computed tomography for predicting pathologic nodal extracapsular extension in patients with head-and-neck cancer undergoing initial surgical resection. Int J Radiat Oncol Biol Phys. 2014;88(1):122–9. https://doi.org/10.1016/j.ijrobp.2013.10.002.

11. Chai RL, Rath TJ, Johnson JT, Ferris RL, et al. Accuracy of computed tomography in the prediction of extracapsular spread of lymph node metastases in squamous cell carcinoma of the head and neck. JAMA Otolaryngol Head Neck Surg. 2013;139(11):1187–94. https://doi.org/10.1001/jamaoto.2013.4491.

12. Tran NA, Palotai M, Hanna GJ, Schoenfeld JD, et al. Diagnostic performance of computed tomography features in detecting oropharyngeal squamous cell carcinoma extranodal extension. Eur Radiol. 2023;33(5):3693–703. https://doi.org/10.1007/s00330-023-09407-4.

13. Multidisciplinary O, Sahin KA, Wahid N, Taku, et al. Multi-Specialty Expert Physician Identification of Extranodal Extension in Computed Tomography Scans of Oropharyngeal Cancer Patients: prospective Blinded Human Inter-Observer performance evaluation. medRxiv. 2023. https://doi.org/10.1101/2023.02.25.23286432.

14. Ariji Y, Sugita Y, Nagao T, Nakayama A, et al. CT evaluation of extranodal extension of cervical lymph node metastases in patients with oral squamous cell carcinoma using deep learning classification. Oral Radiol. 2020;36(2):148–55. https://doi.org/10.1007/s11282-019-00391-4.

15. Kann BH, Aneja S, Loganadane GV, Kelly JR, et al. Pretreatment identification of Head and Neck Cancer nodal metastasis and Extranodal Extension using deep learning neural networks. Sci Rep. 2018;8(1):14036. https://doi.org/10.1038/s41598-018-32441-y.

16. Price WN. Big data and black-box medical algorithms. Sci Transl Med. 2018;10(471). https://doi.org/10.1126/scitranslmed.aao5333.

17. Lee IC, Huang JY, Chen TC, Yen CH, et al. Evolutionary learning-derived Clinical-Radiomic Models for Predicting Early recurrence of Hepatocellular Carcinoma after Resection. Liver Cancer. 2021;10(6):572–82. https://doi.org/10.1159/000518728.

18. Ho SY, Chen JH, Huang MH. Inheritable genetic algorithm for biobjective 0/1 combinatorial optimization problems and its applications. IEEE Trans Syst Man Cybern B Cybern. 2004;34(1):609–20. https://doi.org/10.1109/tsmcb.2003.817090.

19. Shinn-Ying Ho. L.-S.S.a.J.-H.C., *Intelligent evolutionary algorithms for large parameter optimization problems*. IEEE Trans Evol Comput. 2004;8(6):19. https://doi.org/10.1109/TEVC.2004.835176.

20. Chih-Chung Change C-JL. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol. 2011;2(3):27. https://doi.org/10.1145/1961189.1961199.

21. Clausi DA. An analysis of co-occurrence texture statistics as a function of grey level quantization. Can J Remote Sens. 2002;28(1):17. https://doi.org/10.5589/m02-004.

22. Haralick RM. Sam Shanmugam and its'hak Dinstein, *Textural features for image classification*. IEEE TRans Syst Man Cybern. 1973;3:11.

23. Thibault G, Angulo J, Meyer F. Advanced Statistical Matrices for texture characterization: application to cell classification. IEEE Trans Biomed Eng. 2014;61(3):630–7. https://doi.org/10.1109/TBME.2013.2284600.

24. Chen C-C. Improved moment invariants for shape discrimination. Pattern Recogn. 1993;26(5):683–6. https://doi.org/10.1016/0031-3203(93)90121-C.

25. Ming-Kuei H. Visual pattern recognition by moment invariants. IRE Trans Inform Theory. 1962;8(2):179–87. https://doi.org/10.1109/TIT.1962.1057692.

26. Tsai MJ, Wang JR, Ho SJ, Shu LS, et al. GREMA: modelling of emulated gene regulatory networks with confidence levels based on evolutionary intelligence to cope with the underdetermined problem. Bioinformatics. 2020;36(12):3833–40. https://doi.org/10.1093/bioinformatics/btaa267.

27. Yerukala Sathipati S, Sahu D, Huang HC, Lin Y, et al. Identification and characterization of the lncRNA signature associated with overall survival in patients with neuroblastoma. Sci Rep. 2019;9(1):5125. https://doi.org/10.1038/s41598-019-41553-y.

28. Yerukala Sathipati S, Ho SY. Identifying a miRNA signature for predicting the stage of breast cancer. Sci Rep. 2018;8(1):16138. https://doi.org/10.1038/s41598-018-34604-3.

29. Yerukala Sathipati S, Ho SY. Identifying the miRNA signature associated with survival time in patients with lung adenocarcinoma using miRNA expression profiles. Sci Rep. 2017;7(1):7507. https://doi.org/10.1038/s41598-017-07739-y.

30. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, et al. Radiomics in medical imaging-"how-to" guide and critical reflection. Insights Imaging. 2020;11(1):91. https://doi.org/10.1186/s13244-020-00887-2.

31. Kuno H, Garg N, Qureshi MM, Chapman MN, et al. CT texture analysis of cervical lymph nodes on contrast-enhanced [(18)F] FDG-PET/CT images to differentiate nodal metastases from reactive lymphadenopathy in HIV-Positive patients with Head and Neck squamous cell carcinoma. AJNR Am J Neuroradiol. 2019;40(3):543–50. https://doi.org/10.3174/ajnr.A5974.

32. Larasati RaDL. Anna, Building a Trustworthy Explainable AI in Healthcare, in INTERACT 2019/ 17th IFIP: International Conference of Human Computer Interaction. Workshop: Human(s) in the loop -Bringing AI & HCI together. 2019, Cardiff and Ubiquity Press: Paphos, Cuprus.

33. Scott I, Cook D, Coiera E. Evidence-based medicine and machine learning: a partnership with a common purpose. BMJ Evid Based Med. 2021;26(6):290–4. https://doi.org/10.1136/bmjebm-2020-111379.

34. Trimble M, Hamilton P. The thinking doctor: clinical decision making in contemporary medicine. Clin Med (Lond). 2016;16(4):343–6. https://doi.org/10.7861/clinmedicine.16-4-343.

35. Arocha JF, Wang D, Patel VL. Identifying reasoning strategies in medical decision making: a methodological guide. J Biomed Inform. 2005;38(2):154–71. https://doi.org/10.1016/j.jbi.2005.02.001.

36. Scheetz J, Rothschild P, McGuinness M, Hadoux X, et al. A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. Sci Rep. 2021;11(1):5193. https://doi.org/10.1038/s41598-021-84698-5.

37. Quinn TP, Jacobs S, Senadeera M, Le V et al. The three Ghosts of Medical AI: can the Black-Box Present deliver? 2020. arXiv:2012.06000.

38. Wells L, Bednarz T. Explainable AI and reinforcement Learning-A systematic review of current Approaches and Trends. Front Artif Intell. 2021;4:550030. https://doi.org/10.3389/frai.2021.550030.

39. Aggarwal R, Sounderajah V, Martin G, Ting DSW, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med. 2021;4(1):65. https://doi.org/10.1038/s41746-021-00438-z.

40. Ravi D, Wong C, Deligianni F, Berthelot M, et al. Deep learning for Health Informatics. IEEE J Biomed Health Inform. 2017;21(1):4–21. https://doi.org/10.1109/JBHI.2016.2636665.

41. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. Radiographics. 2017;37(2):505–15. https://doi.org/10.1148/rg.2017160130.
42. Ibrahim A, Vaidyanathan A, Primakov S, Belmans F, et al. Deep learning based identification of bone scintigraphies containing metastatic bone disease foci. Cancer Imaging. 2023;23(1):12. https://doi.org/10.1186/s40644-023-00524-3.
43. Li S, Wan X, Deng YQ, Hua HL, et al. Predicting prognosis of nasopharyngeal carcinoma based on deep learning: peritumoral region should be valued. Cancer Imaging. 2023;23(1):14. https://doi.org/10.1186/s40644-023-00530-5.
44. Nakagawa J, Fujima N, Hirata K, Tang M, et al. Utility of the deep learning technique for the diagnosis of orbital invasion on CT in patients with a nasal or sinonasal tumor. Cancer Imaging. 2022;22(1):52. https://doi.org/10.1186/s40644-022-00492-0.
45. Lu CF, Liao CY, Chao HS, Chiu HY, et al. A radiomics-based deep learning approach to predict progression free-survival after tyrosine kinase inhibitor therapy in non-small cell lung cancer. Cancer Imaging. 2023;23(1):9. https://doi.org/10.1186/s40644-023-00522-5.
46. Kann BH, Likitlersuang J, Bontempi D, Ye Z, et al. Screening for extranodal extension in HPV-associated oropharyngeal carcinoma: evaluation of a CT-based deep learning algorithm in patient data from a multicentre, randomised de-escalation trial. Lancet Digit Health. 2023;5(6):e360–9. https://doi.org/10.1016/S2589-7500(23)00046-8.
47. Kann BH, Hicks DF, Payabvash S, Mahajan A, et al. Multi-institutional validation of deep learning for pretreatment identification of Extranodal Extension in Head and Neck squamous cell carcinoma. J Clin Oncol. 2020;38(12):1304–11. https://doi.org/10.1200/JCO.19.02031.
48. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging. 2018;9(4):611–29. https://doi.org/10.1007/s13244-018-0639-9.
49. Varghese J. Artificial Intelligence in Medicine: chances and Challenges for wide clinical adoption. Visc Med. 2020;36(6):443–9. https://doi.org/10.1159/000511930.
50. Xu Y, Hu M, Liu H, Yang H, et al. A hierarchical deep learning approach with transparency and interpretability based on small samples for glaucoma diagnosis. NPJ Digit Med. 2021;4(1):48. https://doi.org/10.1038/s41746-021-00417-4.
51. Gulum MA, Trombley CM, Kantardzic M. A review of Explainable Deep Learning Cancer Detection Models in Medical Imaging. Appl Sci. 2021;11(10):4573.
52. *Spatial Resolution in CT* Journal of the International Commission on Radiation Units and Measurements, 2012. 12(1): p. 107–120.DOI: https://doi.org/10.1093/jicru/ndt001.
53. Chad Dillon MWBI, Jessica Clements MS, Diana Cody MS, Dustin Gress PhD, Kalpana Kanal MS, James Kofler PhD, Michael PhD, McNitt-Gray F, James Norweck PhD, Doug Pfeiffer MS, Thomas MS, Ruckdeschel G, Keith MS. J. Strauss, MS, FACR; James Tomlinson, MS;, 2017 Computed Tomography QUALITY CONTROL MANUAL. 2017, American College of Radiology.

Publisher's Note