2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 978-1-6654-6263-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/Confluence56041.2023.10048891

Ensemble learning based Deep Learning Architecture for malignancy detection of mediastinal lymph nodes in CT images

Posina Praneeth
Electronics and Communication
Koneru Lakshmaiah Education Foundation
Hyderabad, India
190340065@klh.edu.in

Abothula Dhana Lakshmi
Electronics and Communication
Koneru Lakshmaiah Education Foundation
Hyderabad, India
190349110@klh.edu.in

Hitesh Tekchandani
Electronics and Communication
Koneru Lakshmaiah Education Foundation
Hyderabad, India
hitesh@klh.edu.in

Shrish Verma
Electronics and Communication
National Institute of Technology
Raipur, India
shrishverma@nitrr.ac.in

Narendra D. Londhe
Electrical Engineering
National Institute of Technology
Raipur, India
nlondhe.ele@nitrr.ac.in

***Abstract*— The detection of malignancy in lymph nodes (LNs) in its early stages is important for appropriate treatment planning. The tumour growths are usually identified using invasive procedures like biopsy which are painful and complex in nature. Hence, in this study the authors have taken computed tomography images which are non-invasive in nature for detection of malignancy in LNs. For this purpose, the authors have proposed an ensemble learning based deep learning (DL) architecture. The proposed ensemble architecture comprises of various significant custom built DL models. The proposed architecture achieves accuracy = 98.56%, sensitivity = 99.09%, specificity = 98.04%, and AUC = 98.56%.**

Keywords—Ensemble learning, Stacking, Deep learning, Lymph nodes, Cancer, CT.

I. INTRODUCTION

The mediastinum is an anatomical area that contains various vital tissues and organs, including the oesophagus, blood vessels, nerves, thymus, trachea, and lymphatic tissue. Therefore, mediastinal anatomy is intricate, making invasive procedures such as fine needle aspiration cytology (FNAC) and biopsy challenging in the diagnosis of anomalous mediastinal lymph nodes (MLNs) [1]. Additionally, sometimes invasive diagnosis procedures require hospital stay, anaesthesia, causes inflammation and allergy. Medical imaging based diagnosis approaches such as MRI, X-ray, CT, and USG are non-invasive techniques for cancer detection. Computed tomography (CT) is an accurate, painless, inexpensive, and more reliable imaging technique for cancer detection. Hence, this work focuses on malignancy detection of MLNs in CT images. However, manual evaluation of medical images necessitates highly specialized anatomical knowledge and is also stressful and time consuming [2]. Hence, in this study the authors proposed an automated method for cancer detection of MLNs in CT images. DL is the revolutionary technique that does not require manual feature selection and their calculation [3]. Thus, in this work the authors proposed DL based methodology for the intended task of malignancy detection.

II. RELATED WORKS

To overcome the difficulties of manual inspection of malignancy in LNs, various researchers have proposed several image processing and feature extraction based malignancy detection approaches, such as size-based [4]–[8], CT texture feature-based [9]–[12], and radiomics feature-based [13]–[15] techniques.

Nodal size cannot be used as a significant approach for cancer detection for LNs because they can enlarge for a variety of reasons, including inflammation, infection, or cancer. Normal sized nodes appear as a single huge nodal cluster, which can lead to an incorrect diagnosis. Furthermore, inter/intra observer variability might affect size assessment because of the requirement of specialized anatomical knowledge, imprecise boundaries, varied guidelines, and standardization of different measuring software. The existing Nodal size-based approaches [4]–[8] have attained poor sensitivity (41-67%) and specificity (65-94%).

Classification approaches based on texture analysis rely on a variety of techniques to quantify an image's pixel interrelationships, gray-level patterns, and spectral features. Due to the small size of the LNs, there were insufficient voxels for effective heterogeneity analysis, resulting in compromised texture analysis's discriminative capabilities. As a result, texture analysis approaches [9]–[12] have shown less promising results, such as accuracy (56-71%), sensitivity (52-81%), and specificity (60-97%).

Radiomics approaches produce quantitatively large numbers of features from medical data using sophisticated feature extraction algorithms. Despite some practical benefits, radiomics techniques have many limitations including lack of standardisation for validating results, reproducibility, reliance on reconstruction algorithms, feature selection bias, and feature instability. Radiomics techniques [13]–[15] have achieved sensitivity, specificity, and accuracy ranging from (68.01-94.8%), (73.35-92%), and (91.1%) respectively.

Deep learning is the most significant advancement in the technology domain. With recent breakthroughs in medical imaging using DL techniques, we may achieve fast and accurate results. Some researchers have worked on DL approaches for malignancy detection in LNs [1], [2], [16]–

Authorized licensed use limited to: University at Buffalo Libraries. Downloaded on January 12,2025 at 00:24:00 UTC from IEEE Xplore. Restrictions apply.

[18]. Due to the usage of pre-trained weights from generalised images that are distinctive from medical images, there is a problem of insufficient learning and a lack of flexibility with the architecture and number of parameters. These limitations of related literature necessitate, more reliable DL approaches for improved performance. Ensemble learning significantly improves machine learning results by blending multiple models, when compared to a single model, this technique produces greater predictive performance [19]. Hence, in this paper, the authors proposed ensemble learning based DL architecture for the intended task of malignancy detection in MLN CT images.

III. Methodology

In any DL model, noise, bias, and variance are the three main source for inaccuracy. One model cannot compete with these. Using multiple models for predicting the target rather than building just one model will overcome these issues. This method of combining multiple models to create a single, high performance predictive model is called ensemble learning. These individual models in ensemble learning are called the weak learners. These weak learners are combined to create a strong learner, which generalises to more accurately predict all the target classes. The final ensembled model is generally named as meta learner [19]. The proposed ensemble learning based methodology is shown in Fig.1. In the proposed methodology we have designed ensemble learning based meta model using three Convolutional neural network (CNN) based custom models. (The details of used dataset are mentioned in section IV). Furthermore, we have also experimented with different ensemble methodologies like majority voting, weighted voting and stacking. The details of these different ensemble methodologies are explained in below sub sections.

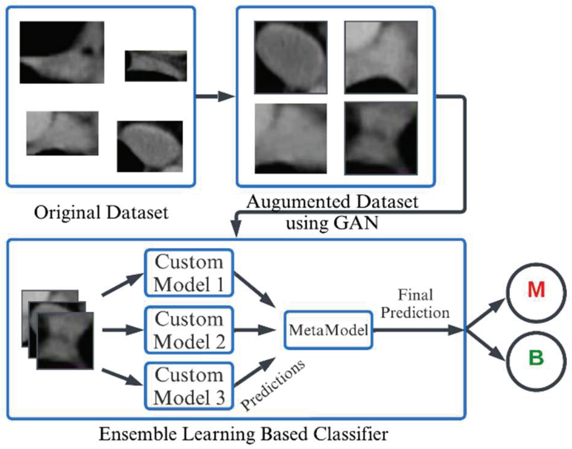


Fig. 1. Workflow of the proposed ensemble learning based approach. “Malignant, M”, “Benign, B”.

A. Majority Voting

Majority voting [20] is one of the averaging techniques popularly used in ensemble learning. In majority voting, the class label that is predicted most number of times by each individual classifier is considered the final predicted class label for the given task. The final prediction can be determined using the formula mentioned in (1):

$$P_f(x)= mode\{P_1(x), P_2(x), P_3(x)\} \quad (1)$$

Where, ‘P_f’ is final prediction, and ‘P_1’, ‘P_2’ & ‘P_3’ are predictions of custom models 1,2 & 3 respectively.

B. Weighted Voting

The main drawback of majority voting is that each model contributes equally to the final ensemble prediction, despite the fact that some models are known to perform much better or worse than other models. A weighted voting based ensemble learning is an improvement to the majority voting in which the weighting of each member's contribution to the final prediction is determined by the individual model's performance [20]. Weighted voting based ensemble learning can be formulated using the (2):

$$P_f(x)=arg\ max[W_1\times P_1(x)+ W_2\times P_2(x)+ W_3\times P_3(x)] \quad (2)$$

Where, ‘P_f’ is final prediction, ‘W_1’, ‘W_2’ & ‘W_3’ are the weights assigned to custom models 1, 2 & 3 respectively, and ‘P_1’, ‘P_2’ & ‘P_3’ are predictions of custom models 1, 2 & 3 respectively. respectively.

C. Stacking

Stacking, also known as stacked generalisation [21], is an ensemble learning method that combines predictions from various models that have been fitted on the training data using a meta model. The workflow of the proposed stacking methodology is shown in below Fig.2. (Explanation of the figure is provided in section IV). The model used to integrate the predictions is known as a level-1 model, while the ensemble members are known as level-0 models. Any DL predictive model can be utilized to combine and learn from the predictions, however logistic regression is commonly used for binary classification. This ensures that the complexity remains at the lower level ensemble models as the higher level models

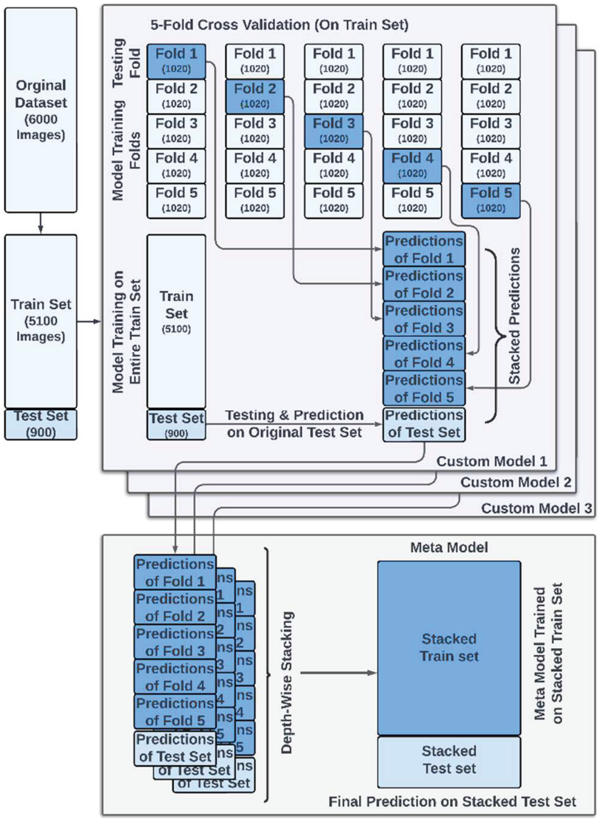


Fig. 2. Workflow of Stacking methodology.

Authorized licensed use limited to: University at Buffalo Libraries. Downloaded on January 12,2025 at 00:24:00 UTC from IEEE Xplore. Restrictions apply.

learn how to harness the range of predictions made. Stacking is a popular meta-learning approach. This method utilizes a meta-learner to determine which classification models are reliable and which are not. The different DL models chosen as ensemble members contribute to the diversity. As a result, it is preferable to utilize a set of models that are trained or built differently, so that the meta model considers assumptions of all kinds and, therefore, reduces the prediction errors by learning and combining the results in the best possible way.

D. Proposed Individual Ensemble Classifiers

For the given task of malignancy detection in MLNs using ensemble learning we have proposed three CNN based DL models explained in the following subsections. These models are considered as weak learners. They are chosen in such a way that they have least correlation in classification and prediction error.

a) Custom Model 1: The proposed Model 1 is composed of three sets of convolutional, and maxpooling layers, followed by flatten and fully connected layers. For the proposed model state-of-art Leaky ReLu activation function is utilized. The loss function used is categorical cross-entropy. Further, we have used the state-of-art optimizer Rectified Adam *[22]*. The model architecture of custom model 1 is presented in Fig.3.

b) Custom Model 2: The proposed Model 2 is pruned version of model 1 in which one convolutional layer and one maxpooling layer are present followed by flatten and fully connected layers, with same activation function, loss function and optimizer as utilized for custom model 1. The model architecture of custom model 2 is presented in Fig.3.

The limitation of deep neural networks is that, as the size of the network grows, there are more parameters to learn, thus increasing the likelihood of overfitting. If we design a very large, extremely deep network, each layer may simply memorise the output resulting in a neural network that fails to generalise to new input. Apart from the risk of overfitting, the training time will also increase as the network size increases.

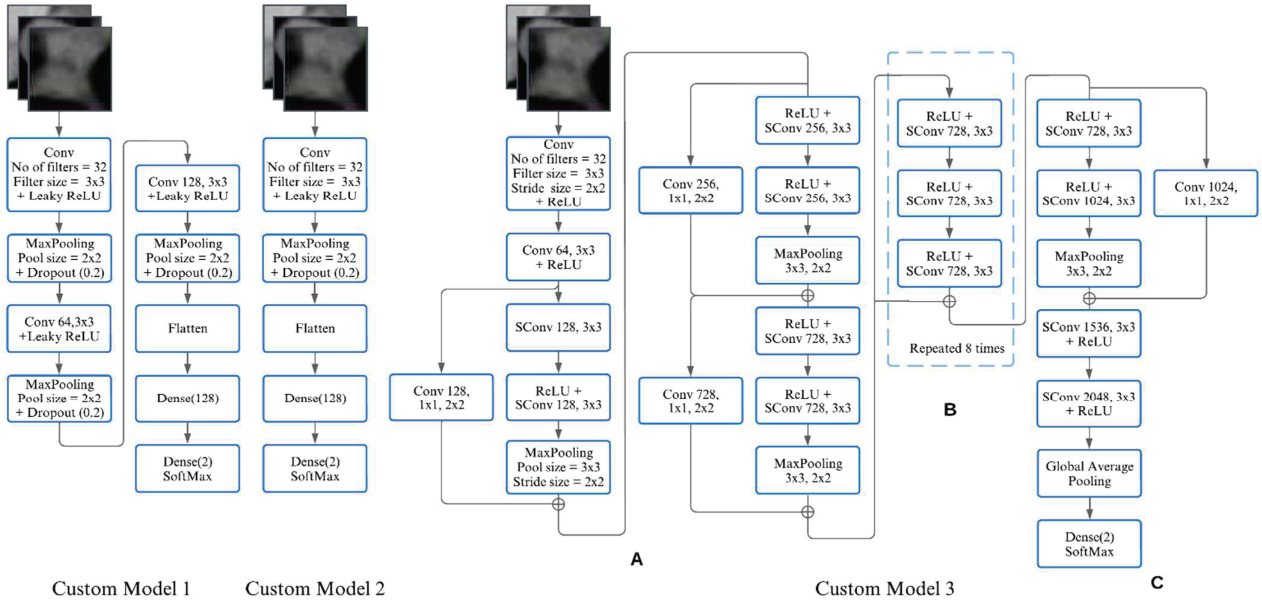
CNN's primary layers excel at capturing fundamental and universal traits such as curves, edges, corners, etc. The deeper layers have more complex representations, they learn to identify entire objects from various perspectives and angles. Since, the majority of CT pictures are taken from a fixed perspective, thus highly complex architectures may not be required. Hence, to tackle these limitations we experimented with these less complex architectures. Furthermore, there is no information compression because there are fewer levels and less pooling involved.

c) Custom Model 3: The proposed model 3 is based on XceptionNet *[23]* concept. This model utilizes the depth-wise separable convolutions (DSConvs). DSConvs are computationally more efficient alternatives to classical convolutions. The DSConv is significantly and effectively utilized throughout the custom model 3. In addition, residual or skip connections, are added across the flow of the network. The original DSConvs execute channel-wise spatial convolution first and then 1x1 convolution, while the modified DSConvs (which is proposed in [23]) execute 1x1 convolution first and then channel-wise spatial convolution. The proposed custom model 3 comprise of three blocks (A, B, & C which can be observed in Fig.3). The first module contains only conventional convolution layers and no DSConv layers. All of the modules in the entry flow (i.e., block A), except for the first, contain residual skip connections. A pointwise convolution layer is added to the output from the main path by the parallel skip connections. The middle flow (i.e., block B) is formed by repeating the convolution module multiple times. This module does not contain skip connections. The exit flow (i.e., block C) contains only two convolution modules, one of which has no skip connection.

IV. Dataset details AND Experimental setup

A. Dataset details

For this proposed study the utilized MLNs dataset is downloaded from [11]. There are a total of 271 MLNs CT images in this dataset in which 138 benign and 133 malignant images. Since the DL algorithms are data driven techniques and the dataset size is limited, the GAN-based augmentation



Custom Model 1 Custom Model 2 Custom Model 3

Fig. 3. Model architectures of Custom models 1, 2 & 3.

Authorized licensed use limited to: University at Buffalo Libraries. Downloaded on January 12,2025 at 00:24:06 UTC from IEEE Xplore. Restrictions apply.

technique from our previous work [2] was utilized to significantly augment the dataset. (Since the details of GAN-based augmentation are appropriately mentioned in our previous work [2], they are not discussed in this paper). Post augmentation, we have a total of 6000 images available, of which 3000 images are benign and 3000 images are malignant. Because the size of the images in the source dataset ranges from 16x24 to 122x112, they are resized to 50x50 before being fed to the GAN network.

B. Experimental Setup

The Google Colab platform and the Keras [24] DL libraries were utilized to perform all the experiments. In this proposed study 5-fold cross validation is utilized for train test split of dataset. Majority voting is applied on all the five fold sets and tested using the final test set to obtain five individual evaluation metrics. These metrics are then averaged to calculate final evaluation metrics for majority voting. For weighted voting ensemble, weightage is assigned to each model's contribution to the final prediction based on the individual model performance. The final evaluation metrics are calculated by averaging the individual metrics obtained for each folding set.

Unlike majority voting and weighted voting, five fold is applied to stacking in a slightly different way according to the need of stacking methodology. Post training on five fold sets, predictions for each testing fold are stacked vertically for each model. Thereafter, the prediction stacks of all three models are depth-wise stacked. Then, the depth-wise stacked predictions are translated to a 2d matrix. These stacked predictions from all three models act as new column features which are utilized to train the meta model. Predictions made on the final test set by all three models which are trained on the whole train set are also stacked horizontally. The final meta model prediction is made on this stacked test set.

V. Results and Discussion

Accuracy (ACC), sensitivity (SEN), specificity (SPE), and area under the curve (AUC) are the evaluation metrics calculated. Table 1 summarises results of all the proposed models evaluated using 5-fold cross-validation. The proposed custom models 1, 2, and 3 attain an average accuracy of 96.06%, 97.16%, and 97.4% respectively. It is also observable from the table 1 that less complex models namely model 1 and 2 performed equally well with sophisticated model 3 however model 3 slightly does better than other two models.

TABLE I. Prediction performance of custom model1,2,&3 on each fold.

Model	**Fold**	***ACC***	***SEN***	***SPE***	***AUC***
		(in %)			
Custom Model 1	Fold 1	96	96.13	95.86	96
	Fold 2	96.55	96.81	96.30	96.56
	Fold 3	95.67	95.45	95.86	95.66
	Fold 4	96.11	97.95	94.34	96.15
	Fold 5	96	96.13	95.86	96
Custom Model 2	Fold 1	97.22	96.59	97.82	97.20
	Fold 2	97	96.81	97.17	96.99
	Fold 3	97.11	98.63	95.65	97.14
	Fold 4	97.33	98.40	96.30	97.35
	Fold 5	97.22	96.81	97.60	97.21
Custom Model 3	Fold 1	97.55	97.72	97.39	97.55
	Fold 2	96.22	96.59	95.86	96.23
	Fold 3	97.33	96.36	98.26	97.31
	Fold 4	98	97.27	98.69	97.98
	Fold 5	97.88	97.72	98.04	97.88

The results of all the implemented ensemble methods, namely majority voting, weighted voting, and stacking using 5-fold cross validation are presented in Table 2. These performance metrics are calculated from predictions made on the final test set using models 1, 2, and 3. The proposed approach of stacking achieves ACC = 98.56%, SEN = 99.09%, SPE = 98.04%, and AUC = 98.56%. The table infers that majority voting provided equal weightage to all the models does not extract the full potential of high performing models. Whereas, weighted voting with more weightage to significantly high performing models was effective up to a certain extent. weightage assigned to custom models 1,2, and 3 are 0.1, 0.3, 0.4 respectively. These values are decided based on the individual model performance. However, multi-level training of stacking helped in producing better performance by the proposed approach.

TABLE II. Evaluation parameters of Majority voting, Weighted voting, And Stacking ensembles.

Ensemble Method	***ACC***	***SEN***	***SPE***	***AUC***
	(in %)			
Majority Voting	97.95	98	97.91	97.95
Weighted Voting	98.06	98	98.13	98.06
Stacking	98.56	99.09	98.04	98.56

Table 3 compares the performance evaluation metrics from related works. The proposed approach achieves a significant improvement of 3.56% in ACC, 5.09% in SEN, 1.04% in SPE, and 3.56% in AUC, as shown in Table.

TABLE III. Summary of performance evaluation metrics obtained from related works.

Authors	***ACC***	***SEN***	***SPE***	***AUC***
	(in %)			
Michael et al. [10]	-	53	97	83
Hamid [9]	71	81	80	87
Hongkai et al. [17]	86	84	88	91
Pham et al. [11]	70	75	90	89
Pham et al. [14]	-	68.01	73.35	75
Pham et al. (raw images) [18]	87.07	90.42	83.96	-
Hitesh et al. [16]	63.14	71.03	55.69	63
Hitesh et al. [1]	90	91	90	90
Hitesh et al. [2]	95	94	97	95
Proposed Approach	98.56	99.09	98.04	98.56

VI. Conclusion and Future Scope

In this paper the authors have proposed the ensemble learning based DL Architecture for the diagnosis of malignant and benign MLNs in CT images. This study shows that the proposed ensemble learning architecture performs well compared to the individual custom models. Furthermore, the proposed methodology also performs significantly well

Authorized licensed use limited to: University at Buffalo Libraries. Downloaded on January 12,2025 at 00:24:00 UTC from IEEE Xplore. Restrictions apply.

compared to the existing related works. In the proposed ensemble learning approach, we have taken three custom models which can be extended to more models in future works. Additionally, the proposed methodology can also be implemented for organ at risk detection in the nearby anatomy of lymph nodes. Moreover, the proposed work can be implemented for other radiological modalities like X-ray, PET scan, MRI, and USG.

REFERENCES

[1] H. Tekchandani, S. Verma, and N. D. Londhe, “Mediastinal lymph node malignancy detection in computed tomography images using fully convolutional network,” *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, pp. 187–199, Jan. 2020, doi: 10.1016/j.bbe.2019.05.002.

[2] H. Tekchandani, S. Verma, and N. Londhe, “Performance improvement of mediastinal lymph node severity detection using GAN and Inception network,” *Computer Methods and Programs in Biomedicine*, vol. 194, p. 105478, Oct. 2020, doi: 10.1016/j.cmpb.2020.105478.

[3] H. Tekchandani, S. Verma, N. D. Londhe, R. R. Jain, and A. Tiwari, “Computer aided diagnosis system for cervical lymph nodes in CT images using deep learning,” *Biomedical Signal Processing and Control*, vol. 71, p. 103158, Jan. 2022, doi: 10.1016/j.bspc.2021.103158.

[4] J. M. Seely, J. R. Mayo, R. R. Miller, and N. L. Müller, “T1 lung cancer: prevalence of mediastinal nodal metastases and diagnostic accuracy of CT.,” *Radiology*, vol. 186, no. 1, pp. 129–132, Jan. 1993, doi: 10.1148/radiology.186.1.8416552.

[5] H. Libshitz and R. McKenna, “Mediastinal lymph node size in lung cancer,” *American Journal of Roentgenology*, vol. 143, no. 4, pp. 715–718, Oct. 1984, doi: 10.2214/ajr.143.4.715.

[6] H. C. Steinert *et al.*, “Non-small cell lung cancer: nodal staging with FDG PET versus CT with correlative lymph node mapping and sampling.,” *Radiology*, vol. 202, no. 2, pp. 441–446, Feb. 1997, doi: 10.1148/radiology.202.2.9015071.

[7] D. W. von Haag, D. M. Follette, P. F. Roberts, D. Shelton, L. D. Segel, and T. M. Taylor, “Advantages of Positron Emission Tomography over Computed Tomography in Mediastinal Staging of Non-Small Cell Lung Cancer,” *Journal of Surgical Research*, vol. 103, no. 2, pp. 160–164, Apr. 2002, doi: 10.1006/jsre.2002.6354.

[8] L. K. Toney and H. J. Vesselle, “Neural Networks for Nodal Staging of Non–Small Cell Lung Cancer with FDG PET and CT: Importance of Combining Uptake Values and Sizes of Nodes and Primary Tumor,” *Radiology*, vol. 270, no. 1, pp. 91–98, Jan. 2014, doi: 10.1148/radiol.13122427.

[9] H. Bayanati *et al.*, “Quantitative CT texture and shape analysis: Can it differentiate benign and malignant mediastinal lymph nodes in patients with primary lung cancer?,” *European Radiology*, vol. 25, no. 2, pp. 480–487, 2014, doi: 10.1007/s00330-014-3420-6.

[10] M. B. Andersen, S. W. Harders, B. Ganeshan, J. Thygesen, H. H. T. Madsen, and F. Rasmussen, “CT texture analysis can help differentiate between malignant and benign lymph nodes in the mediastinum in patients suspected for lung cancer,” *Acta Radiologica*, vol. 57, no. 6, pp. 669–676, 2016, doi: 10.1177/0284185115598808.

[11] T. D. Pham, Y. Watanabe, M. Higuchi, and H. Suzuki, “Texture Analysis and Synthesis of Malignant and Benign Mediastinal Lymph Nodes in Patients with Lung Cancer on Computed Tomography,” *Sci Rep*, vol. 7, no. 1, Art. no. 1, Feb. 2017, doi: 10.1038/srep43209.

[12] X. Gao *et al.*, “The method and efficacy of support vector machine classifiers based on texture features and multi-resolution histogram from 18F-FDG PET-CT images for the evaluation of mediastinal lymph nodes in patients with lung cancer,” *European Journal of Radiology*, vol. 84, no. 2, pp. 312–317, Feb. 2015, doi: 10.1016/j.ejrad.2014.11.006.

[13] X. Yang *et al.*, “A new approach to predict lymph node metastasis in solid lung adenocarcinoma: a radiomics nomogram,” *J Thorac Dis*, vol. 10, no. Suppl 7, pp. S807–S819, Apr. 2018, doi: 10.21037/jtd.2018.03.126.

[14] T. D. Pham, “Complementary features for radiomic analysis of malignant and benign mediastinal lymph nodes,” in *Proceedings - International Conference on Image Processing, ICIP*, 2018, vol. 2017-Septe, no. 4, pp. 3849–3853. doi: 10.1109/ICIP.2017.8297003.

[15] Y. Zhong, M. Yuan, T. Zhang, Y.-D. Zhang, H. Li, and T.-F. Yu, “Radiomics Approach to Prediction of Occult Mediastinal Lymph Node Metastasis of Lung Adenocarcinoma,” *American Journal of Roentgenology*, vol. 211, no. 1, pp. 109–113, Jul. 2018, doi: 10.2214/AJR.17.19074.

[16] H. Tekchandani, S. Verma, and N. D. Londhe, “Severity Assessment of Lymph Nodes in CT Images using Deep Learning Paradigm,” in *International Conference on Computing Methodologies and Communication*, 2018, pp. 686–691.

[17] H. Wang *et al.*, “Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18 F-FDG PET/CT images,” *EJNMMI research*, vol. 7, no. 1, p. 11, 2017.

[18] T. D. Pham, “From Raw Pixels to Recurrence Image for Deep Learning of Benign and Malignant Mediastinal Lymph Nodes on Computed Tomography,” *IEEE Access*, 2021, doi: 10.1109/ACCESS.2021.3094577.

[19] C. Zhang and Y. Ma, Eds., *Ensemble Machine Learning*. Boston, MA: Springer US, 2012. doi: 10.1007/978-1-4419-9326-7.

[20] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. CRC Press, 2012.

[21] D. H. Wolpert, “Stacked generalization,” *Neural Networks*, vol. 5, no. 2, pp. 241–259, Jan. 1992, doi: 10.1016/S0893-6080(05)80023-1.

[22] L. Liu *et al.*, “On the Variance of the Adaptive Learning Rate and Beyond.” arXiv, Oct. 25, 2021. doi: 10.48550/arXiv.1908.03265.

[23] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions.” arXiv, Apr. 04, 2017. doi: 10.48550/arXiv.1610.02357.

[24] F. Chollet and others, “Keras: The Python Deep Learning library,” *Astrophysics Source Code Library*, p. ascl:1806.022, Jun. 2018.

Authorized licensed use limited to: University at Buffalo Libraries. Downloaded on January 12,2025 at 00:24:06 UTC from IEEE Xplore. Restrictions apply.