공학석사 학위논문

흉부방사선 영상에서 이상탐지에 강인한 정상군 선별을 위한 딥러닝 기반 다중학습 앙상블 모델 개발

Development of deep learning based ensemble model with multi-task learning for screening normal group with robust unseen classes in chest radiographs

울산대학교대학원

의 과 학 과

김 준 식

흉부방사선 영상에서 이상탐지에 강인한 정상군 선별을 위한 딥러닝 기반 다중학습 앙상블 모델 개발

지도교수 김 남 국

이 논문을 공학석사 학위 논문으로 제출함

2024년 2월

울산대학교대학원
의 과 학 과
김 준 식

김준식의 공학석사학위 논문을 인준함

심사위원 이 상 민 (인)

심사위원 김 남 국 (인)

심사위원 이 준 구 (인)

울 산 대 학 교 대 학 원

2024 년 2 월

Abstract

One of the primary responsibilities of radiologists in routine clinical settings is to diagnose chest radiographs (CXR) to determine the presence or absence of disease in patients. In scenarios where a large volume of chest radiographs are taken daily, the majority of patients presenting are diagnosed as normal. In such cases, the time and effort required for specialists to interpret every image are substantial, and the use of artificial intelligence models to classify definitively normal groups can significantly reduce the workload of doctors. However, the risk of missing significant diseases in patients is a critical issue. Therefore, an algorithm that accurately classifies normal and diseased states is necessary, especially one that avoids misdiagnosing diseased groups as normal. To address this, we propose the Strictly Chest X-ray Normal Network (SCXNN), a model that emulates the interpretation of radiologists and is more robust to diseases. The SCXNN model is a multi-task learning model that simultaneously performs reconstruction and classification for 13 different diseases. The first decoder in this model aims to learn the location and characteristics of lesions, enabling the encoder to learn the patterns of the disease. The second classifier, a Support Vector Machine (SVM), uses the logits from each disease model to classify groups into definite normal and abnormal. For data, we used 24,714 chest X-ray images collected from Asan Medical Center (AMC) in Seoul from 2011 to 2018, of which 5,400 were images with masks (Hard Label) and 19,314 were images without masks (Weak Label). The Hard Label data was split into training, validation, and test sets in an 8:1:1 ratio, and all Weak Label images were used for training. These images included or were classified as 'Normal' if they did not contain any of the 13 diseases such as Cardiomegaly, Advanced Tuberculosis, etc. Additionally, 355 chest X-ray images collected from January to March 2021 at AMC were used as external data for temporal validation datasets, and all mask and class labeling was performed by radiologists. Recognizing the problem of not sufficiently learning the characteristics of normal lung tissue when targeting only the diseased areas in the disease reconstruction process, we applied dilation, a type of

morphological operation, in the mask processing step. This approach allowed for learning not only the diseased areas in the image but also the patterns of surrounding normal tissues affected by the disease, contributing to improved accuracy in disease reconstruction. The SCXNN model showed consistent performance in both internal and temporal validation datasets as well as external public datasets. This demonstrates the robustness of SCXNN, providing stable results regardless of the dataset type, which is crucial for ensuring patient health and safety in real clinical settings. Furthermore, the use of SCXNN as a diagnostic support tool for radiologists can contribute to reducing the workload of medical staff. This will enhance the speed and accuracy of the diagnostic process, improving the quality of medical services and allowing medical staff to focus more time on patient care, thereby enhancing the overall efficiency of medical services. The characteristics of SCXNN are expected to contribute to the further expansion of the use of AI tools in the clinical decision-making process.

차 례

표 및 그림 목차

1. Introduction

Deep learning is a field of machine learning based on artificial neural networks and has achieved significant success in the computer vision domain. The convolutional neural network (CNN)[1], which won the large-scale ImageNet visual recognition competition in 2012 by significantly reducing error rates, is well-known as a prominent algorithm in deep learning. CNNs have demonstrated excellent performance not only in image classification but also in various tasks such as object recognition, semantic segmentation, image restoration, depth prediction, and visual question-answering.

In the medical field, CNNs can be utilized for various research purposes using tasks similar to those in computer vision[2]. Particularly, the interpretation of chest radiographs (CXR) is suitable for conducting such research[3]. Every day, a substantial number of CXRs are captured, and radiologists in the field of radiology interpret them. They determine whether the patient who underwent the CXR belongs to the normal group or the abnormal group with a disease[4,5]. Ensuring that diseases are not misclassified as normal is a critical issue, as it directly relates to patient well-being and health.

CXR is a commonly used diagnostic tool in the medical field. However, interpreting radiographic images requires deep domain expertise and can be time-consuming, leading to the risk of errors. In clinical practice, radiologists use CXRs to assess the presence of diseases and their progression. Deep learning technology has been introduced to support the interpretation of radiographic images, greatly assisting in tasks such as disease classification, detection, and region delineation. These deep learning approaches help alleviate the burden on radiologists and enhance the accuracy of CXR interpretation[6].

In the process of classifying normal and abnormal groups using CXR and identifying diseases, radiologists consider various factors. They first identify if there are any abnormal areas in the CXR. If there are areas different from normal lungs, they examine the characteristics and patterns of these areas (such as size, shape, location, texture, etc.) to determine the nature of the disease and its extent. Expertise and experience of radiologists are crucial for such interpretations. Radiologists must maintain high concentration levels, and the number of CXRs to be interpreted daily can be substantial. However, the majority of CXRs captured daily belong to the normal group. If artificial intelligence can first classify definite normal cases, it

can reduce the workload of radiologists.

To address the challenge of identifying definite normal cases in chest radiographs (CXRs), we have developed an artificial intelligence algorithm. The disease reconstruction decoder of this algorithm helps the encoder learn various features and patterns of diseases by reconstructing only the specific disease when present in the CXR. The disease classification decoder then determines the presence or absence of disease in the input chest radiograph. We chose image reconstruction as it allows artificial intelligence to learn the features and patterns of abnormal areas that radiologists observe when interpreting diseases. By jointly learning these features and patterns, our artificial intelligence algorithm can demonstrate robust performance in disease classification.

Using this disease classification algorithm, we perform our ultimate goal of classifying definite normal cases. By classifying normal cases using the disease classification algorithm, we can create a more robust algorithm than simply binary classification as normal or abnormal. This unique approach can effectively mitigate instances where individuals with diseases are mistakenly predicted as normal.

The key contributions of our research are as follows:

- Learning various features and patterns used by radiologists in disease classification.
- Training on a high-resolution CXR dataset and validating on an internal dataset, a temporal validation dataset, and an external dataset.

Employing a deep learning algorithm that classifies diseases first to classify definite normal cases.

2. Background

2.1. CNN

In deep learning utilizing image processing and artificial intelligence (AI), CNN (Convolutional Neural Network) plays a pivotal role. CNN is a specialized design for processing pixel data and was developed for image analysis. It is used in tasks such as image recognition, video recognition, image classification, medical image analysis, and also performs recommendation systems and natural language processing.

CNN can be viewed as a specific subset of neural networks. It consists of multiple layers, with each layer performing various operations (e.g., convolution, pooling, loss computation). The initial layer, known as the input layer, includes neurons directly connected to the pixels of the input image[7]. Subsequent layers are convolutional layers that apply convolution to the input data using filters, also known as kernels, whose sizes can vary based on the designer's choice. Each neuron responds only to a small region, known as a receptive field, rather than the entire layer. The output in the convolutional layer is referred to as an activation map, representing the results of applying specific filters to the input. Convolutional layers are typically followed by activation layers that apply non-linear effects. The next layer may be a pooling layer, which can reduce dimensions and often employs strategies like max pooling or average pooling. Finally, high-dimensional abstraction is achieved through fully connected layers. The weights and kernels of this neural network are continuously adjusted during the training process using the backpropagation method.

2.2. *DenseNet*

DenseNet[8] is a model composed of Dense Blocks. A Dense Block is structured in a way where every layer is connected to all previous layers. In other words, each layer receives input from all the previous layers and its output is passed on to all the subsequent layers. This interconnectedness of all layers allows for the direct propagation of lower-level features, enabling their reuse and mitigating the gradient vanishing problem. Additionally, DenseNet

differs from ResNet[9] in the way connections are merged. While ResNet adds the outputs of two paths (sum), DenseNet preserves information by concatenating the outputs of previous layers along the channel dimension. The model is constructed by stacking Dense Blocks, and between these blocks, convolution and pooling operations are applied to reduce the image size

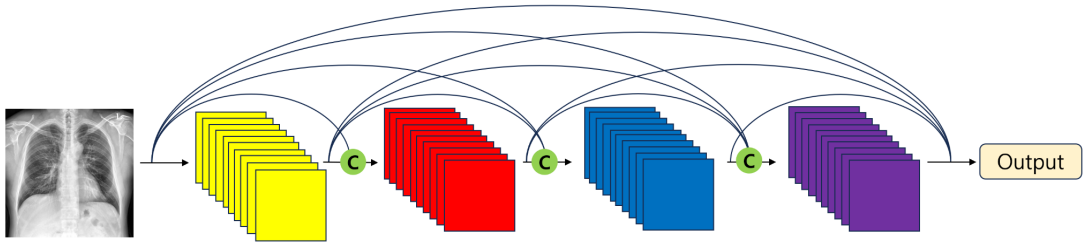

Figure 1. Densenet

2.3. U-Net

U-Net[10], which was presented at the prestigious MICCAI conference in the field of medical image processing, was developed for the purpose of medical image segmentation. It also won the ISBI cell tracking competition in the same year. The structure of U-Net is similar to conventional convolutional neural networks, where feature maps pass through convolutional layers and pooling layers, reducing spatial resolution and increasing depth. This process is repeated four times, resulting in a 32x32x512 feature map at the bottom. Now, transpose convolutions are used to restore the original image size. U-Net refers to the downsampling and upsampling processes as the contracting path and expanding path.

What sets U-Net apart is the use of skip connections. The top skip connection cuts out a portion from the center of the 568x568x64 feature map and passes it to the expanding path. On the right, the data received through skip connections is combined with the data coming from below to create a tensor. This tensor passes through convolutional layers to output the final segmentation map. This segmentation map includes object class probability and background class probability.

In the original U-Net, the convolutional layers use 3x3 filters with a stride of 1 and no padding. The pooling layers use 2x2 filters to reduce the size of the feature maps. Since U-Net was developed for medical image segmentation, the input images are typically grayscale, and the final output consists of two-channel maps representing objects and backgrounds. In actual implementations, various modifications can be applied to accept multi-channel input images or segment multiple classes of objects.

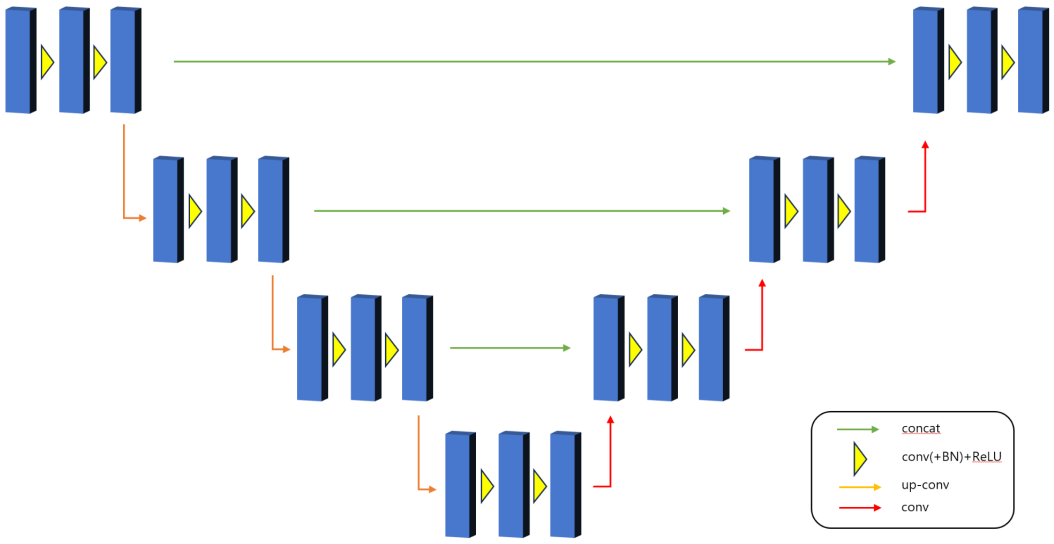

Figure 2. U-Net

2.4. Loss Function

Cross-Entropy Loss[11] is a widely used fundamental loss function in classification problems. This function measures how closely the model's predicted probabilities align with the actual labels. It has the characteristic of assigning high loss values to incorrect predictions, so when the model makes significant errors, it incurs a large penalty. Because of this characteristic, the model focuses more on making accurate predictions. Cross-Entropy loss is particularly useful for quantifying the difference between probability distributions. As a result, it trains the model to generate predictions that are closer to the actual distribution. However, it's

important to exercise caution when dealing with severe class imbalances.

The Mean Squared Error (MSE) Loss[11] calculates the average of the squared differences between predicted values and actual values. This is suitable not only for regression tasks but also for tasks like image reconstruction. In reconstruction tasks, the goal is to minimize the pixel differences between the original image and the reconstructed image. To achieve this, MSE Loss calculates the squared differences at each pixel location and then computes the average of these differences. One advantage of MSE Loss is that it makes it easy to interpret how close the model's predictions are to the actual values. Additionally, because it uses the square of errors, it has the property of penalizing larger errors more, encouraging the model to reduce significant errors.

2.5. SVM

A Support Vector Machine (SVM)[11] is a supervised learning model used for classification and regression analysis. The core principle of SVM is to find the 'decision boundary' that best separates the data, aiming to maximize the margin between two classes. The margin represents the distance between the decision boundary and the closest training data points, which are referred to as 'support vectors.' Mathematically, SVM attempts to solve the following optimization problem with the following constraints:

$$min_{w,b}\frac{1}{2}\|w\|^2, y_i(w \cdot x_i + b) \geq 1, \forall i$$

Here, w represents the normal vector to the decision boundary, b is the bias, x_i represents data points, and $y_i$is the label of the corresponding data point (-1 or +1).The linear decision function is as follows:

$$f(x) = w \cdot x + b$$

If $f(x) > 0$, the data point x belongs to class +1; otherwise, it belongs to class -1.

Linear SVM separates data linearly, but many datasets cannot be linearly separated. To address this, SVM uses the "kernel trick" to map the data to a higher-dimensional space where linear separation is possible. The kernel function $K(x, x')$ measures the similarity between two data points. Common kernels include polynomial kernels, RBF kernels, sigmoid kernels, among others. SVM performs well in various fields, particularly on small datasets or high-dimensional datasets. It is not limited to classification and can also be used for regression or anomaly detection tasks. Kernel selection and parameter tuning are crucial, and training time can be long for large datasets

2.6. Image Dilation

Image morphology[12] is a technique in the field of image processing that analyzes and processes the morphological structure of images using structural elements. Fundamentally, it is based on the theory of mathematical morphology and is useful for simplifying, removing, or enhancing objects within an image. Morphological operations are primarily applied to binary or grayscale images and are used to analyze or modify characteristics such as the size, shape, connectivity, and holes of objects.

These operations all involve the use of structural elements, also known as kernels or masks, which are applied to the image to define the relationships between each pixel and its neighboring pixels. The size and shape of the structural element have a significant impact on the processing results.

One of the types of morphological operations, dilation, expands the boundaries of an image. This is achieved by adding pixels at all positions that overlap with the structural element, the reby increasing the size of objects and filling in small gaps or spaces.

2.7. ***Contrast Limited Adaptive Histogram Equalization(CLAHE)***

Contrast Limited Adaptive Histogram Equalization (CLAHE)[13] is a widely used technique for

enhancing the contrast of images, particularly in fields such as medical imaging and satellite imagery. CLAHE is a variation of traditional Histogram Equalization (HE) that is effective in improving contrast in localized regions of an image. The fundamental principle of CLAHE involves dividing the image into small "tiles" or "blocks" and independently applying histogram equalization to each block. This enhances contrast in localized areas of the image. Additionally, CLAHE includes a "contrast limiting" feature that restricts areas with excessively high contrast during the histogram equalization process. This prevents certain parts of the image from becoming overly bright or dark, suppresses noise amplification, and helps maintain a natural contrast in the image.

CLAHE achieves optimal contrast in various parts of the image by independently processing each block. This is particularly useful for images with uneven lighting or diverse textures. In medical imaging, CLAHE is used to make fine structures more clearly visible in images such as X-rays and MRI scans. CLAHE was developed with the goal of improving image contrast while preserving natural image quality, and its effectiveness has been demonstrated in various fields.

3. materials and Methods

3.1. Data acquisition

From 2011 to 2018, we utilized 34,851 chest X-ray images from Asan Medical Center (AMC). Among the entire dataset, 6,628 images were labeled with hard labels, while the remaining 18,086 images were labeled with weak labels. The dataset with hard labels was split into training, validation, and test sets in an 8:1:1 ratio, and all images with weak labels were used for training. The images we used contain one or more of the following diseases: Cardiomegaly, Advanced Tuberculosis, Interstitial Opacity, Pneumothorax, Pleural Effusion, Active Tuberculosis, Nodule, Consolidation, Atelectasis, Mediastinal Widening, Support Device, Pleural Calcification, and Pneumoperitoneum. Images without diseases were classified as

"Normal." For external data, we used chest X-ray data from Asan Medical Center (AMC) taken from January 2021 to March 2021, totaling 355 images. This external dataset serves as an internal dataset, captured at different times for temporal validation. The masking and class labeling of each dataset were performed by radiologists. You can inspect the data distribution in the images for training and validation.

	Class	Hard Label	Weak Label	Total		Class	Hard Label	Weak Label	Total
Critical	Pneumothorax	331	1669	2000	Urgent	Advanced Tuberculosis	324	0	324
	Pneumoperitoneum	100	3959	4059	Non-urgent	Atelectasis	98	4175	4273
	Mediastinal Widening	98	223	321		Support Device	585	0	585
Urgent	Pleural Effusion	1364	636	2000		Pleural Calcification	100	4146	4246
	Nodule	914	1546	2460		Interstitial Opacity	280	1732	2012
	Consolidation	1540	0	1540	Cardiomegaly	Cardiomegaly	312	0	312
	Active Tuberculosis	582	0	582		Normal			10137

Table 1 : This table provides information about the dataset and categorizes 13 diseases into Critical, Urgent, Non-urgent, and Cardiomegaly based on their risk levels. The classification of risk levels for each disease is a key aspect of the study and summarizes the dataset's characteristics.

3.2. Pre-Processing

We apply the following preprocessing steps to the DICOM data we used: (1) Adjust image intensity, setting the 0.5 percentile intensity to the minimum value and the 99.5 percentile intensity to the maximum value. (2) Perform min-max normalization to scale values between 0 and 255. (3) Pad empty image areas with zeros, considering aspect ratio. (4) Apply the CLAHE (Contrast Limited Adaptive Histogram Equalization) algorithm to ensure uniform contrast distribution. (5) Resize all images to 1024 x 1024 pixels and save them in PNG format. Augmentation[14] for training includes image shift, rotation, gamma correction, sharpening, blurring, and random Gaussian noise.

3.3. Methods

The proposed Strictly Chest X-ray Normal Network (SCXNN) is a multi-task learning model[15] that is trained for the reconstruction[16,17] and classification of 13 different diseases. It employs a second classifier, the Support Vector Machine (SVM), to classify a certain subset of normal cases among the 13 diseases. The first decoder aims to learn (1) the location of the lesions and (2) specific patterns of the lesions, with the goal of enabling the encoder to learn the patterns of the diseases. The second decoder is responsible for determining the presence or absence of diseases, and the SVM is used to identify definite normal cases among the patient group. You can see the model's diagram in the figure below

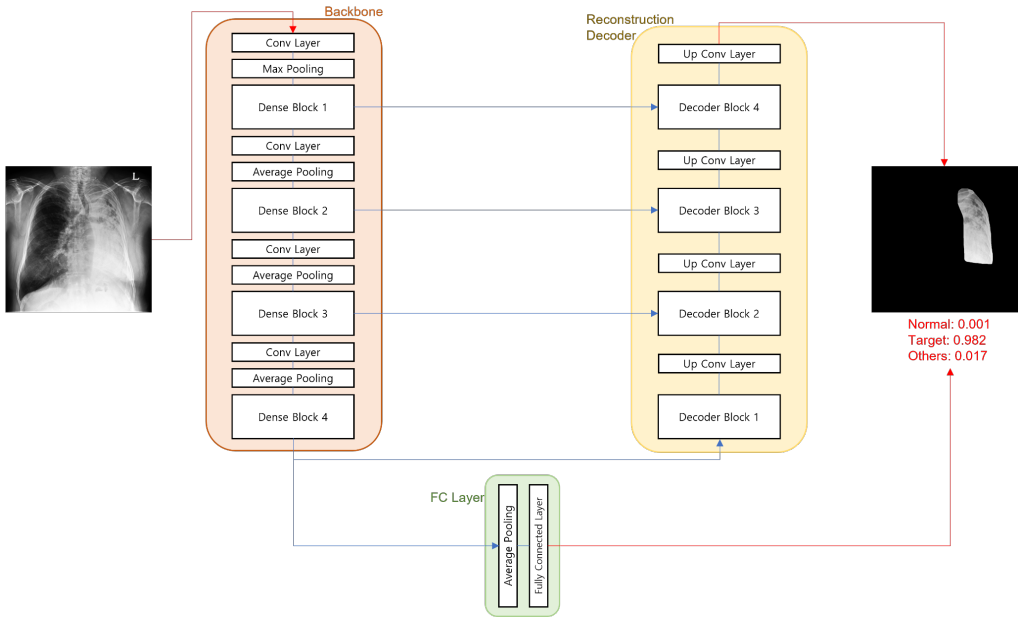

Figure 3. represents a model used in training, which predicts three classes: normal, target, and others, while also performing reconstruction. This figure highlights the model's capability to classify instances into these three classes and its concurrent reconstruction process.

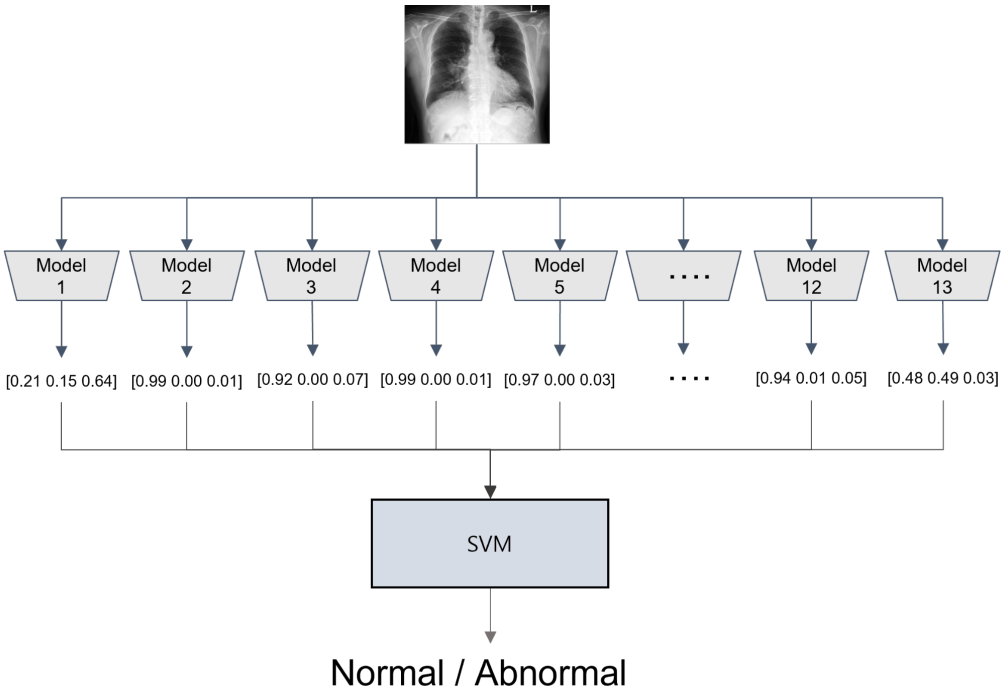

Figure 4. illustrates an ensemble of 13 models, which are combined using SVM to create a 2-class model for distinguishing between normal and abnormal cases. This ensemble approach

showcases how multiple models can be integrated to enhance binary classification performance.

3.3.1 Backbone network

Our proposed model uses DenseNet-121 as the backbone network. We chose DenseNet-121 for its ability to learn patterns of various diseases in chest X-ray images. DenseNet, like VGGNet[18], ResNet, EfficientNet[19], etc., is a widely used computer vision model with the capacity suitable for learning various disease patterns.

DenseNet focuses on the concept of shorter connections that allow direct connections between the input layer and output layer. This efficient training format helps prevent the vanishing gradient problem[20] that occurs during the backpropagation process. It also enhances feature propagation by passing features from shallow layers to deep layers and allows feature reuse, reducing the number of parameters.

DenseNet-121 consists of 1x1, 3x3, and 7x7 convolution layers, as well as 2x2, 3x3, and 7x7 pooling layers. It includes a 1000-dimensional fully connected layer (FC layer)[21] at the end. In our research, we created a model for each of the 13 diseases. These 13 models are then ensembled. Each model is designed as a 3-class problem, distinguishing between normal, the target disease, and other diseases. Therefore, in our proposed model, all layers except for the FC layer remain the same as the original DenseNet-121. The FC layer, on the other hand, is replaced with a 3-dimensional FC layer to align with the objectives of our research. The initial weights of the model are pretrained on the ImageNet dataset[22]. We first train the model with hard labels and then proceed with the final training using weak labels[23].

3.3.2. Reconstruction decoder

The reconstruction decoder assists the encoder, which is the backbone network, in learning the location of the target lesions that each model aims to find in chest X-rays, as well as the features such as texture and brightness, darkness around the disease. The lesions in our dataset

can occur at specific locations or randomly in the X-ray images. To learn both the lesion's location and the context, such as whether the disease brightens or darkens over time, segmentation alone may not be sufficient. Therefore, we chose to perform a reconstruction task[24-27] instead of segmentation[28-31]. By using reconstruction, the encoder can not only learn the lesion's location but also gain insights into the surrounding context and textures of the lesion. This approach allows us to consider various factors when making judgments about diseases, including their contextual information.

3.3.3. Second classifier

We used a second classifier to improve classification performance in order to identify a clearly normal group[32-34]. The second classifier utilizes Support Vector Machine (SVM), a machine learning model commonly used in pattern recognition and data analysis. SVM is primarily employed for classification and regression analysis, especially for distinguishing one category of data from another. It works by maximizing the margin, which is the distance between the support vectors (input data) and the decision boundary, to classify the data. When the margin is maximized, the robustness of classification performance is also maximized.

3.3.4. Training and testing

The training of our proposed model follows the following steps: (1) Fine-tuning the weights pretrained on ImageNet to train the backbone network. (2) Training the reconstruction and classification simultaneously using hard-label data. During this step, the backbone learns features such as the location of abnormalities and texture patterns. (3) Training with weak labels, where the decoder responsible for reconstruction is frozen, and only the FC layer for classification is trained. (4) Training SVM using the logits from each of the 13 models. During SVM training, the labels are transformed into binary classification: normal and abnormal, where abnormal includes both target and other disease classes.

Chest X-ray images with diseases may have one or more diseases, making it a multi-class classification problem. Therefore, we use three classes for classification in each of the 13 models: target disease, other disease (others), and normal. The goal is to improve the model's

ability to classify a clear normal group. Finally, using the three logits obtained from each of the 13 models, SVM classifies normal and abnormal. The ultimate objective of our model is to identify a clear normal group without any of the specified 13 diseases.

During training, two loss functions are used: cross-entropy loss and Mean Squared Error (MSE) loss. The final loss function combines these two losses by multiplying the MSE loss, which is used for reconstruction, by a hyperparameter λ (λ = 0.8 in our experiments) and adding it to the cross-entropy loss. The expressions for each loss function are as follows:

$$cross\ entropy\ loss = l_1(x, y) = \{l_1, l_2, \ldots, l_n\}^T,$$

$$l_n = -w_n log \frac{exp(x_n, y_n)}{\sum_{c=1}^{C} exp(x_{n,c})} \cdot 1\{y_n \neq ignore\ index\}$$

$$MSE\ loss = l_2(x, y) = mean(l_1, l_2, \ldots, l_N)^T, l_n = (x_n - y_n)^2, N = batch\ size$$

$$Total\ loss = l_1 + \lambda \times l_2$$

The optimizer used is Adaptive Moment Estimation (Adam)[35], with an initial learning rate of 0.001. If there is no weight update for 50 epochs, the learning rate is halved, and training continues. The momentum is set to 0.9, and weight decay is set to 0.0005. A batch size of 4 is used, and the maximum number of epochs is set to 300. If the learning rate becomes lower than 0.0001 and there is no weight update for 50 epochs, the training process is halted.

During the model's testing phase, each image is down-sampled to 1024×1024, similar to the training phase. The 13 models produce logits, which are then passed through an SVM to obtain the final classification result of normal and abnormal.

4. Results

4.1. Result on datasets

In this study, the performance of the AI model was evaluated using four different datasets: internal, temporal validation, external 1[36], and external 2[37]. The results in Table 2 demonstrate

the robustness of the model, showing consistent performance metrics across all datasets. In the internal dataset, the model exhibited high sensitivity and specificity, correctly identifying 1325 out of 1326 abnormal cases and accurately classifying 109 out of 162 normal cases (sensitivity 0.99, specificity 0.67). This indicates that the model excels at detecting abnormal cases and does not misclassify normal cases as abnormal. The model also maintained high performance in the temporal validation dataset, correctly identifying 299 out of 300 abnormal cases and 121 out of 225 normal cases (sensitivity 0.99, specificity 0.54). This highlights the model's robustness and reliability across different time settings and variations in medical devices. In external dataset 1, the AI correctly classified 127 out of 128 abnormal cases and 66 out of 106 normal cases (sensitivity 0.99, specificity 0.63), demonstrating the model's adaptability and accuracy. Finally, in external dataset 2, the model exhibited high accuracy, accurately identifying 628 out of 645 abnormal cases and 47 out of 71 normal cases (sensitivity 0.97, specificity 0.66). Despite a slight increase in false positives compared to other datasets, the overall performance of the model remained robust and reliable

			Reference standard	
			normal	abnormal
AI	Internal dataset	normal	109	4
		abnormal	53	1325
	Temporal Validation dataset	normal	121	1
		abnormal	104	299
	External dataset 1	normal	66	1
		abnormal	40	127
	External dataset 2	normal	47	17
		abnormal	24	628

Table 2. This table presents the model's normal/abnormal results for internal, temporal valid, external1, and external2 data partitions. The values within the table represent the number of observations in each data partition, serving to compare the model's performance across different data splits.

4.2. Results on risk based screening

In this study, disease groups were classified based on the severity of the diseases into 'critical,' 'urgent,' 'non-urgent,' and 'cardiomegaly'[38], and the diseases were evaluated according to their respective severity levels. The most significant finding was that there were no mispredictions of normal cases in the 'critical' disease group. This indicates that the model exhibited very high accuracy in identifying 'critical' diseases. Furthermore, the model performed quite well for the 'urgent' and 'non-urgent' disease groups as well, demonstrating its ability to effectively distinguish disease states with varying levels of severity. Particularly, the model showed high prediction accuracy for the 'urgent' disease group and maintained stable performance for the 'non-urgent' disease group. While the prediction accuracy for the 'cardiomegaly' disease group was relatively lower, it may reflect the more complex patterns associated with this disease group. These results can serve as important foundational data for future model improvements. Overall, the model used in this study demonstrated high performance in effectively distinguishing disease states with varying levels of severity, and the outstanding predictive ability in the 'critical' disease group holds significant clinical relevance.

			Reference standard			
			critical	urgent	Non-urgent	cardiomegaly
AI	Internal dataset	Normal	0	3	1	0
		Target	82	564	186	60
		Others	25	379	26	3
	Temporal Validation dataset	Normal	0	1	0	0
		Target	22	28	47	6
		Others	12	59	44	11

Table 3. This table displays the results of the model's predictions for risk levels on internal and

temporal valid data. The values within the table indicate the number of observations in each data partition, demonstrating how the model predicts risk levels.

4.3. Ablation study

4.3.1 Backbone selection

In the ablation study of this study, various deep learning architectures were compared and analyzed to select the optimal model. This comparative analysis included architectures such as 'VGG 16', 'Resnet 50', 'GoogLeNet'[39], 'Densenet 121', 'Densenet 169', 'Efficientnet B2', among others. The performance of each model was evaluated using metrics such as Accuracy, AUC[40], Specificity and Sensitivity. Specificity is a metric that represents the proportion of true negatives that the model accurately classified as negatives. It measures how accurately the model classifies the negative class, specifically how well it identifies normal cases as normal. Sensitivity, on the other hand, is a metric that represents the proportion of true positives that the model accurately classified as positives. It measures how accurately the model classifies the positive class, specifically how well it identifies abnormal cases as abnormal. The analysis results showed that the 'Densenet 121' model performed the best in accurately predicting normal cases as normal and disease cases as disease cases. This model particularly achieved a high score of 0.971 in the Sensitivity metric, and it also demonstrated excellent performance in Accuracy and AUC with scores of 0.865 and 0.904, respectively. These results indicate that 'Densenet 121' is highly effective in accurately identifying diseases. Therefore, in this study, the 'Densenet 121' model was selected as the final model. The high prediction accuracy and reliability of this model are expected to play a crucial role in disease diagnosis and classification. Additionally, these results can serve as a criterion for model selection in future similar studies.

	Accuracy	AUC	Specificity	Sensitivity

VGG 16	0.7655	0.726	0.709	0.822
Resnet 50	0.828	0.780	0.723	0.933
GoogLeNet	0.786	0.757	0.692	0.880
Densenet 121	**0.865**	**0.904**	**0.758**	**0.971**
Densenet 169	0.841	0.881	0.730	0.952
Efficientnet B2	0.821	0.878	0.729	0.913

Table 4. This table showcases the experimental results for selecting the Backbone network. Specificity represents the ratio of correctly predicting normal cases as normal, while Sensitivity represents the ratio of correctly predicting abnormal cases as abnormal. This table evaluates the accuracy and reliability of the model.

4.3.2. Hyperparameter λ selection

In this study, experiments were conducted by adding the classification loss, which is the cross-entropy loss, and the reconstruction loss, which is the MSE (Mean Squared Error) loss, to the final loss of the model. In these experiments, the two losses were combined by multiplying the MSE loss by a weight parameter λ. The value of λ was set to 0.8, and this choice was made for the following reasons. Experimental results showed that the setting of λ=0.8 produced the best results. This configuration allowed the model to perform accurate classification using the cross-entropy loss while also improving the reconstruction of input data through the MSE loss. Specifically, the model using λ=0.8 demonstrated excellent performance in various performance metrics, including Accuracy, AUC, Specificity, and Sensitivity. This indicates that the λ value is a crucial hyperparameter that enables better model performance, and it explains why λ is included in the final loss of the model. Therefore, the model using λ=0.8 effectively demonstrated the ability to simultaneously handle classification and reconstruction tasks by combining cross-entropy loss and MSE loss.

	Accuracy	AUC	Specificity	Sensitivity

$\lambda = 1.0$	0.786	0.745	0.629	0.944
$\boldsymbol{\lambda = 0.8}$	**0.865**	**0.904**	**0.758**	**0.971**
$\lambda = 0.6$	0.762	0.777	0.652	0.872
$\lambda = 0.5$	0.759	0.789	0.624	0.895
$\lambda = 0.4$	0.716	0.752	0.581	0.852
$\lambda = 0.2$	0.668	0.802	0.533	0.803

Table 5. This table presents experimental results for selecting the hyperparameter lambda when combining classification loss and reconstruction loss. Specificity and Sensitivity indicate how well the model predicts normal and abnormal cases. This table provides insights into hyperparameter tuning for improving model performance.

4.3.3. Data pre-processing selection

In this study, three different methods were compared and analyzed during the image preprocessing stage. These three methods are Reconstruction, CLAHE (Contrast Limited Adaptive Histogram Equalization), and With-Dilation. Each of these methods was used in the image preprocessing process, and their results were evaluated using performance metrics such as Accuracy, AUC, Specificity, and Sensitivity. The experimental results showed that the last row, which used all three methods Reconstruction, CLAHE and With-Dilation, exhibited the best performance. This row recorded higher Accuracy, AUC, Specificity and Sensitivity scores compared to the other methods, indicating that simultaneously using these three methods during image preprocessing is crucial for improving model performance. Therefore, it was concluded that using Reconstruction, CLAHE and With-Dilation together in the image preprocessing stage is the optimal strategy, and these preprocessing methods can help the model achieve better performance

		Accuracy	*AUC*	*Specificity*	*Sensitivity*
hist-equalization + Without	*only classification*	*0.608*	*0.748*	*0.608*	*0.766*
	Segmentation	*0.621*	*0.784*	*0.621*	*0.785*

Dilation	*Reconstruction*	*0.651*	*0.801*	*0.651*	*0.806*
Reconstruction + Without Dilation	*CLAHE*	*0.720*	*0.874*	*0.720*	*0.857*
Reconstruction + CLAHE	*With Dilation*	*0.865*	*0.904*	*0.758*	*0.971*

Table 6. This table displays experimental results for selecting data preprocessing methods. Specificity and Sensitivity represent how well the model predicts normal and abnormal cases. This table illustrates the impact of data preprocessing on model performance.

4.3.4. Performance of 2 class and 3 class

The proposed artificial intelligence model primarily deals with a multi-class classification problem involving three classes: 'normal', 'target'(specific disease), and 'others'. This approach aims to classify medical images into these three classes, rather than a classic binary classification problem of 'normal' and a specific disease. This approach allows the model to predict medical images that do not belong to the target disease class as 'others', capturing valuable information about different diseases that are not the primary target.

Experimental results have shown that this multi-class approach outperforms binary classification (refer to Table 6). Extending the model to multi-class classification can be more effective in various disease discrimination and identification tasks, indicating its potential benefits over binary classification in medical image analysis.

4.3.5. Novel data results

In this study, we considered three diseases, 'pneumonia', 'edema' and 'fracture', which were not present in the internal dataset used for training when evaluating the external 1 and external 2 datasets. These diseases represented conditions that the model encountered for the first time, and experiments were conducted to investigate how the model would handle these new diseases. The experimental results showed that the majority of the 13 models tended to predict these new diseases as 'Other disease'. However, the models rarely predicted them as

'Normal'. This highlights that changing the problem from a 2-class model ('Normal' and 'Target') to a 3-class model ('Normal', 'Target' and 'Other disease') effectively addressed the issue that arose when the model initially encountered these new diseases.

Therefore, the experimental results emphasize the importance of considering various diseases during the initial model design and demonstrate that using a 3-class classification model can enhance the model's performance when dealing with previously unseen conditions. This approach is expected to be useful in real-world medical image classification scenarios.

Internal dataset		Reference standard			
		2 Class		3 Class	
		normal	abnormal	normal	abnormal
AI	normal	78	381	109	4
	abnormal	84	978	53	1325

Table 7. This table provides results for selecting 13 models with 3 classes instead of 2 classes, demonstrating the reasons behind this choice. It shows results for normal, abnormal, and intermediate risk levels, with data counts to assess the model's classification capabilities.

Internal dataset		Reference standard							
		2 Class				3 Class			
		critical	urgent	Non-urgent	cardiomegaly	critical	urgent	Non-urgent	cardiomegaly
AI	Normal	43	218	98	22	0	3	1	0
	Target	64	728	115	41	82	564	186	60
	Others	-	-	-	-	25	379	26	3

Table 8. This table presents results for selecting 13 models with 3 classes instead of 2 classes, focusing on risk levels. It displays results for various risk levels, with data counts to evaluate the model's risk level classification capabilities.

	Critical									Urgent		
	Pneumothorax Model			Pneumoperitoneum Model			Mediastinal Widening Model			Pleural Effusion Model		
	Normal	Target	Others	Normal	Target	Others	Normal	Target	Others	Normal	Target	Others
pneumonia	0.103	0.103	0.793	0.000	0.100	0.900	0.000	0.100	0.900	0.100	0.233	0.667
edema	0.000	0.143	0.857	0.000	0.143	0.857	0.000	0.429	0.571	0.143	0.000	0.857
fracture	0.000	0.279	0.721	0.140	0.140	0.721	0.140	0.140	0.721	0.140	0.140	0.721
	Urgent											
	Nodule Model			Consolidation Model			Active Tuberculosis Model			Advanced Tuberculosis Model		
	Normal	Target	Others	Normal	Target	Others	Normal	Target	Others	Normal	Target	Others
pneumonia	0.233	0.100	0.667	0.000	0.100	0.355	0.233	0.000	0.767	0.000	0.333	0.667
edema	0.143	0.429	0.429	0.143	0.000	0.122	0.143	0.429	0.429	0.000	0.143	0.857
fracture	0.000	0.279	0.721	0.000	0.000	1.000	0.250	0.250	0.500	0.000	0.140	0.860
	Non-urgent											
	Atelectasis Model			Support Device Model			Pleural Calcification Model			Interstitial Opacity Model		
	Normal	Target	Others	Normal	Target	Others	Normal	Target	Others	Normal	Target	Others
pneumonia	0.000	0.300	0.700	0.100	0.333	0.567	0.100	0.233	0.667	0.000	0.100	0.900
edema	0.143	0.429	0.429	0.429	0.143	0.429	0.000	0.429	0.571	0.000	0.000	1.000
fracture	0.279	0.140	0.581	0.286	0.286	0.429	0.000	0.279	0.721	0.000	0.140	0.860
	Cardiomegaly											
	Cardiomegaly Model											

	Normal	Target	Others
pneumonia	0.000	0.100	0.900
edema	0.000	0.143	0.857
fracture	0.000	0.140	0.860

Table 9. This table displays experimental results for evaluating the performance of 13 models on disease groups not covered during training. The prediction distribution is used to assess the model's predictive abilities on disease groups that were not included in the training data.

4.3.6. Model Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping)[41] is a visualization technique used to generate class activation maps for deep learning models. It helps in visually understanding which regions of an image the model considers important for making predictions. Grad-CAM is particularly useful for visualizing which image regions influence the model's decision during image classification.

In the results visualized using Grad-CAM, all the models were found to focus on and emphasize the image regions related to the target disease. This indicates that all the models effectively understood and utilized important features related to the target disease for accurate prediction. Therefore, the fact that all the models pay attention to and predict the target disease results suggests that our model is effectively using crucial information for detecting and classifying the target disease. This underscores the potential of our model to provide reliable diagnoses.

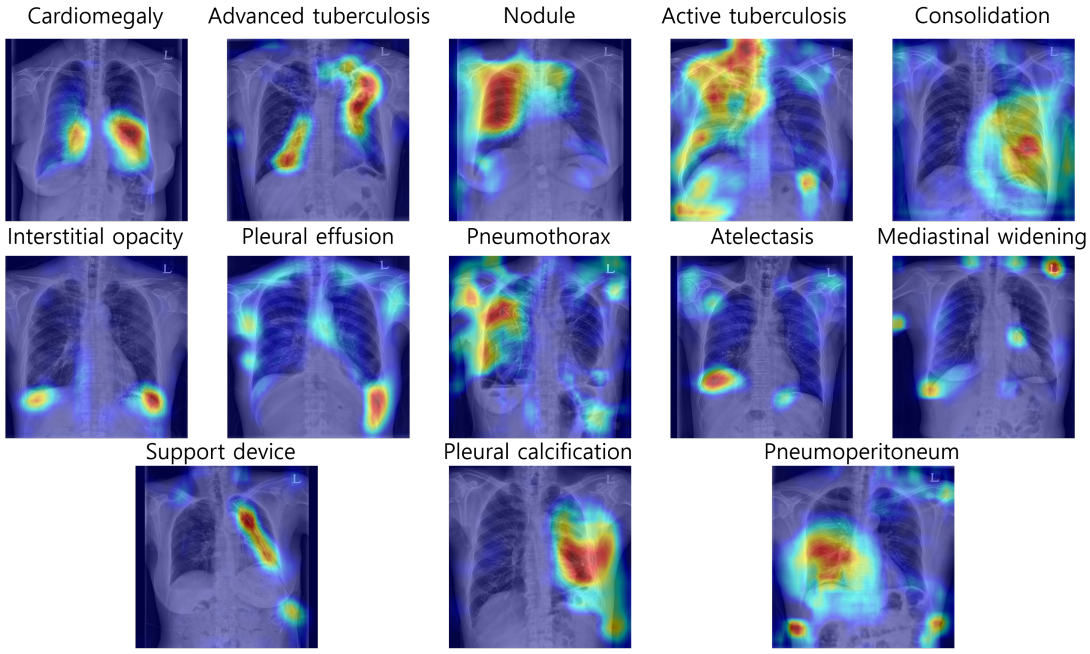

Figure 5. showcases the visualization of Grad-CAM (Gradient-weighted Class Activation Mapping) for the 13 models. Grad-CAM highlights the regions in images that are most relevant to the model's predictions. This figure provides insights into where and how the models focus their attention when making predictions.

5. Discussion

In this study, we developed a specialized deep learning algorithm and AI model based on deep learning to enhance the efficiency of radiology evaluations by identifying normal cases in chest radiographs (CXR). Our algorithm, utilizing advanced techniques such as DenseNet121 and U-Net, demonstrated high efficiency in distinguishing normal cases from pathological ones effectively. This differentiation is crucial, given the high volume of daily chest radiographs, a significant portion of which are normal.

The use of DenseNet121 leveraged its depth and feature reuse capability, enabling subtle feature extraction necessary for accurate screening. This was complemented by the U-Net architecture, which reinforced the localization of important relevant features for precise selection.

Our approach included rigorous data augmentation techniques like adding Gaussian noise, CLAHE, and mask expansion to adapt to various imaging conditions, enhancing the model's robustness and adaptability.

However, this study has limitations as it relies on data from a single institution, potentially impacting the generalizability of research findings. Future research should focus on validating the algorithm in diverse datasets and clinical settings. Additionally, it's essential to note that our model is designed not to replace but complement human experts, reducing the workload of radiologists while maintaining high standards of patient care.

These results and conclusions emphasize the model's robustness and reliability, consistent performance across diverse datasets, and high predictive accuracy in disease classification based on risk. This bodes well for the advancement and utilization of AI technology in the field of healthcare, contributing to improved patient diagnosis and disease management.

6. Conclusion

The developed deep learning algorithm represents a significant advancement, particularly in the efficient screening of normal chest radiographs in the field of radiology. By accurately identifying normal cases, this algorithm has the potential to significantly reduce the workload of radiologists and healthcare professionals. Such a transformation can lead to faster diagnosis, reduced medical costs, and improved patient outcomes.

The results across internal, temporal validation, external 1, and external 2 datasets have demonstrated the robustness and reliability of the model, with high predictive accuracy in disease classification based on risk.

Future research should focus on further refining the algorithm, ensuring its applicability and reliability across diverse populations and imaging technologies. The direction of evolution for the model will involve improving its performance and expanding its classification capabilities for various disease categories. The integration of these AI tools into clinical practice will contribute to enhancing the efficiency and accuracy of medical diagnosis, ultimately leading to the delivery of better healthcare services.

Reference

[1] LeCun, Yann, and Yoshua Bengio. "Convolutional networks for images, speech, and time series." *The handbook of brain theory and neural networks* 3361.10 (1995): 1995.

[2] Alom, Md Zahangir, et al. "A state-of-the-art survey on deep learning theory and architectures." *electronics* 8.3 (2019): 292.

[3] Thirukrishna, J. T., et al. "Survey on diagnosing CORONA VIRUS from radiography chest X-ray images using convolutional neural networks." *Wireless Personal Communications* 124.3 (2022): 2261-2270.

[4] Ragab, Mahmoud, et al. "Machine learning with quantum seagull optimization model for COVID-19 chest X-ray image classification." *Journal of Healthcare Engineering* 2022 (2022).

[5] Bhuvaneswari, P., and A. Brintha Therese. "Feature extraction and classification of COPD chest X-ray images." *International Journal of Computer Aided Engineering and Technology* 12.3 (2020): 301-317.

[6] Heiliger, Lars, et al. "Beyond medical imaging-A review of multimodal deep learning in radiology." *Authorea Preprints* (2023).

[7] Sun, Le, et al. "Spectral–spatial feature tokenization transformer for hyperspectral image classification." *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022): 1-14.

[8] Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[9] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[10] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer International Publishing, 2015.

[11] Bishop, Christopher M., and Nasser M. Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. No. 4. New York: springer, 2006.

[12] Haralick, Robert M., Stanley R. Sternberg, and Xinhua Zhuang. "Image analysis using mathematical morphology." *IEEE transactions on pattern analysis and machine*

intelligence 4 (1987): 532-550.

[13] Yadav, Garima, Saurabh Maheshwari, and Anjali Agarwal. "Contrast limited adaptive histogram equalization based enhancement for real time video system." *2014 international conference on advances in computing, communications and informatics (ICACCI)*. IEEE, 2014.

[14] Shorten, Connor, and Taghi M. Khoshgoftaar. "A survey on image data augmentation for deep learning." *Journal of big data* 6.1 (2019): 1-48.

[15] Graham, Simon, et al. "One model is all you need: multi-task learning enables simultaneous histology image segmentation and classification." *Medical Image Analysis* 83 (2023): 102685.

[16] Bank, Dor, Noam Koenigstein, and Raja Giryes. "Autoencoders." *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook* (2023): 353-374.

[17] Chen, Zhaomin, et al. "Autoencoder-based network anomaly detection." *2018 Wireless telecommunications symposium (WTS)*. IEEE, 2018.

[18] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[19] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." *International conference on machine learning*. PMLR, 2019.

[20] Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training recurrent neural networks." *International conference on machine learning*. Pmlr, 2013.

[21] Popescu, Marius-Constantin, et al. "Multilayer perceptron and neural networks." *WSEAS Transactions on Circuits and Systems* 8.7 (2009): 579-588.

[22] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009.

[23] Howard, Jeremy, and Sebastian Ruder. "Universal language model fine-tuning for text classification." *arXiv preprint arXiv:1801.06146* (2018).

[24] Thanh, Nguyen Chi, and Tran Quoc Long. "Self-supervised Visual Feature Learning for Polyp Segmentation in Colonoscopy Images Using Image Reconstruction as Pretext Task." *2021 8th NAFOSTED Conference on Information and Computer Science (NICS)*.

IEEE, 2021.

[25] Wu, Yali. "Classification of Ancient Buddhist Architecture in Multi-Cultural Context Based on Local Feature Learning." *Mobile Information Systems* 2022 (2022).

[26] Sun, Jiayuan, Jiewen Zhu, and Luping Ji. "Shared Coupling-bridge for Weakly Supervised Local Feature Learning." *arXiv preprint arXiv:2212.07047* (2022).

[27] Liang, Lili, Di Miao, and Han Meng. "Multi-scale Super-resolution Image Reconstruction based on Visual Feature and Scale Feature Transformation." *2023 19th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*. IEEE, 2023.

[28] Huang, Xiaodong, et al. "Lw-TISNet: light-weight convolutional neural network incorporating attention mechanism and multiple supervision strategy for tongue image segmentation." *Sensing and Imaging* 23.1 (2022): 6.

[29] Mohanakurup, Vinodkumar, et al. "Breast cancer detection on histopathological images using a composite dilated Backbone Network." *Computational Intelligence and Neuroscience* 2022 (2022).

[30] Haque, Ayaan, et al. "Generalized multi-task learning from substantially unlabeled multi-source medical image data." *arXiv preprint arXiv:2110.13185* (2021).

[31] Huang, Chao, et al. "3D U 2-Net: A 3D universal U-Net for multi-domain medical image segmentation." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer International Publishing, 2019.

[32] Wang, Shi-jin, et al. "Empirical analysis of support vector machine ensemble classifiers." *Expert Systems with applications* 36.3 (2009): 6466-6476.

[33] Ghoshal, Asish, et al. "An ensemble svm model for the accurate prediction of non-canonical microrna targets." *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*. 2015.

[34] Jassim, Omar Abdullatif, Mohammed Jawad Abed, and Zenah Hadi Saied Saied. "Deep Learning Techniques in the Cancer-Related Medical Domain: A Transfer Deep Learning Ensemble Model for Lung Cancer Prediction." *Baghdad Science Journal* (2023).

[35] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).

[36] Johnson, Alistair EW, et al. "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs." *arXiv preprint arXiv:1901.07042* (2019).

[37] Irvin, Jeremy, et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. No. 01. 2019.

[38] Nam, Ju Gang, et al. "Development and validation of a deep learning algorithm detecting 10 common abnormalities on chest radiographs." *European Respiratory Journal* 57.5 (2021).

[39] Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[40] Bradley, Andrew P. "The use of the area under the ROC curve in the evaluation of machine learning algorithms." *Pattern recognition* 30.7 (1997): 1145-1159.

[41] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

국문 요약

일상적인임상환경에서 방사선전문의의 주요책임중 하나는 환자의 질병 유무를 판단하기 위해 흉부 방사선 사진(CXR)을 진단하는 것이다. 이렇게 하루동안 많은 양의 흉부 방사선 영상이 촬영되는 상황에서, 내원하는 환자들의 대다수가 정상환자로 판독된다. 이러한 상황에서 모든 영상을 전문의가 판독하는 데에는 상당한 시간과 노력이 필요하며, 특히 확실한 정상군을 분류하는데 인공지능 모델의 도움은 의사의 업무 부담을 줄여줄 수 있는 매우 유용한 방법이다. 하지만, 이때 환자의 중요한 질병을 놓치는 위험은 매우 중요한 문제이다. 따라서 정상과 질병을 정확하게 분류하는 알고리즘이 필요한데, 특히 영상을 판독할 때 질병군을 정상군으로 오진 하지 않게 하는 것이 필요하다. 우리는 정상과 비정상을 분류하기 위한 모델을 만들기위해 기존 알고리즘같은 이진분류 인고지능 알고리즘을 사용하는것이 아닌 방사선의사의 판독을 모방하며, 질병에 더욱 강건한 모델을 만들기위해 Strictly Chest X-ray Normal Network(SCXNN)을 제안한다. SCXNN 모델은 13 가지 다른 질병에 대해 재구성과 분류를 동시에 수행하는 멀티태스크 학습 모델이다. 이 모델은 첫번째 디코더는 병변이 발생하는 위치와 병변의 특징을 학습하여 인코더가 병변의 패턴을 학습하는것을 목표로 한다. 두 번째 분류기인 서포트 벡터 머신(SVM)은 각 질병에 대한 모델의 로짓을 이용하여 확실한 정상,비정상 그룹을 분류합니다. 데이터의 경우 2011 년부터 2018 년까지 서울아산병원에서 수집한 총 24,714 건의 흉부 X-ray 데이터 중 5,400 건은 마스크가 적용된 이미지(하드 레이블), 19,314 건은 마스크가 없는 이미지(위크 레이블)였습니다. 하드 레이블 데이터는 훈련, 검증, 테스트 세트로 8:1:1 비율로 나뉘어졌고, 위크 레이블 이미지는 모두 학습에 사용되었습니다. 이 이미지들은 Cardiomegaly, Advanced Tuberculosis 등 13 가지 질병을 포함하거나 질병이 없는 'Normal'로 분류되었습니다. 또한, 2021 년 1 월부터 3 월까지 수집된 355 건의 서울아산병원 흉부 X-ray 데이터를 외부 데이터로 사용하여 시간적 검증 데이터셋으로 활용했으며, 모든 마스크 및 클래스 레이블링은 영상의학과 의사가 수행하였습니다. 질병의 재건 과정에서 질병 부위만을 대상으로 할 경우 정상 폐 조직의 특성을 충분히 학습하지 못하는 문제를 인식하였습니다. 이를 해결하기 위해, 마스크 처

리 과정에 모폴로지 연산의 일종인 팽창(dilation) 연산을 적용하였습니다. 이러한 접근은 영상에서 질병이 있는 부위뿐만 아니라 질병이 영향을 미치는 주변 정상 조직 패턴까지 함께 학습할 수 있도록 하여, 질병 재건의 정확도를 향상시키는 데 기여하였습니다. SCXNN 모델은 내부 및 시간적 검증 데이터셋과 외부 공개 데이터셋 모두에서 일관된 성능을 보여주었습니다. 이는 데이터셋의 종류에 관계없이 SCXNN 이 안정적인 결과를 제공하는 강건한 모델임을 입증하는 것입니다. 이러한 강건성은 실제 임상 환경에서 환자의 건강과 안전을 보장하는 데 중요한 역할을 할 것으로 기대됩니다. 더불어, 영상의학과 의사의 진단 지원 도구로서 SCXNN 의 활용은 의료진의 워크로드를 줄이는 데 기여할 수 있습니다. 이는 진단 과정의 신속성과 정확성을 높여 의료 서비스의 질을 향상시키며, 의료진이 보다 많은 시간을 환자 치료에 집중할 수 있게 함으로써 전반적인 의료 서비스의 효율성을 증진시킬 것입니다. SCXNN 의 이러한 특성은 임상 의사 결정 과정에서 인공지능 도구의 활용을 더욱 확대하는 데 기여할 것으로 사료됩니다.