Attention mechanisms for physiological signal deep learning: which attention should we take?

Seong-A Park¹, Hyung-Chul Lee^{1,2}, Chul-Woo Jung^{1,2}, and Hyun-Lim Yang^{1,*}

¹ Department of Anesthesiology and Pain Medicine,

Seoul National University Hospital, Seoul, Republic of Korea

² Department of Anesthesiology and Pain Medicine, Seoul National University

College of Medicine, Seoul, Republic of Korea

hlyang@snu.ac.kr

Abstract. Attention mechanisms are widely used to dramatically improve deep learning model performance in various fields. However, their general ability to improve the performance of physiological signal deep learning model is immature. In this study, we experimentally analyze four attention mechanisms (e.g., squeeze-and-excitation, non-local, convolutional block attention module, and multi-head self-attention) and three convolutional neural network (CNN) architectures (e.g., VGG, ResNet, and Inception) for two representative physiological signal prediction tasks: the classification for predicting hypotension and the regression for predicting cardiac output (CO). We evaluated multiple combinations for performance and convergence of physiological signal deep learning model. Accordingly, the CNN models with the *spatial* attention mechanism showed the best performance in the classification problem, whereas the channel attention mechanism achieved the lowest error in the regression problem. Moreover, the performance and convergence of the CNN models with attention mechanisms were better than stand-alone self-attention models in both problems. Hence, we verified that convolutional operation and attention mechanisms are complementary and provide faster convergence time, despite the stand-alone self-attention models requiring fewer parameters.

Keywords: Physiological signal · Attention · Deep learning.

1 Introduction

Deep learning has dramatically improved the predictability of various phenomena based on input data of past events. For natural language processing, recurrent neural networks (RNNs) are particularly effective in analyzing time-series sequences [2,6,11]. For image processing, convolutional neural networks (CNNs) that mimic human visual cognitive functions have grown in popularity [10,23,24]. However, both methods have shortcomings, such as the RNN's vanishing gradient and information loss problems [21], which limits performance, and the CNN's

^{*} Corresponding author.

locality of pixel dependency [15], which make it goes deeper. To overcome these roadblocks, attention mechanisms have been used to enable neural models to pay closer attention to the most important parts of the data while ignoring irrelevant parts [7]. It gives higher weight to parts that are more relevant to produce output, and lower weights to parts that are not. Bahadnau et al. [1] introduced this idea to machine translation, resulting in superior performance over canonical RNNs. Similar concept of attention mechanism was also introduced, e.g., Luong et al. [18], and the other types of attention mechanisms were also proffered which tailored to computer vision applications [12, 26, 27].

In recent days, self-attention-based mechanisms had been replaced the canonical deep learning architectures and are positioned as a mainstream of AI research. Vaswani et al. [25] proposed a deep learning model that skipped the RNN and applied a self-attention mechanism by itself (so-called Transformer), achieving superior performance in machine translation and document generation. Dosoviskiy et al. [3] proposed a vision transformer, which a variant of the Transformer for image classification tasks, outperforming canonical CNNs with substantially fewer computations. Subsequently, self-attention-based deep learning was used to predict protein structures [14], compiler graph optimizers [30], and audio generation methods [13].

Consequently, the application of deep learning to physiological signal analysis has been considered [4]. For example, Hannun et al. [8] built a CNN that detects arrhythmia from electrocardiogram (ECG), showing human expert-level performance. As in other domains, attention mechanisms have been used to improve performance in physiological signal analysis. Mousavi et al. [22] proposed an attention-based CNN+RNN network to predict sleep stages from single-channel electroencephalogram. Yang et al. [28] built a CNN with attention blocks to predict stroke volume from arterial blood-pressure waveform. Unfortunately, all of these methods were tuned for specific signal types or tasks, and the best attention mechanisms for general field use for physiological signal analysis was not determined.

In this study, we experimentally determine which CNN architectures and attention mechanisms are the best for analyzing physiological signals. We focus on attention mechanisms used in computer vision, as the various features of physiological signal processes are similar, and the challenges of accurately predicting and classifying the presence of signal and object anomalies are closely related. Hence, we considered the three types of CNN models which popular for image processing and four types of attention mechanisms which suggested for computer vision tasks. Notably, a physiological signal generally has a smaller dimension than does an image, and the attention mechanism designed for computer vision may reduce efficiency by adding unnecessary calculations. Additionally, in a computer vision problem, discriminating feature detection is the main task, whereas in physiological signal analysis, not only is detecting discriminating features important, but detecting signal trends is also crucial. Therefore, for effective and efficient use of attention mechanisms, it is necessary to analyze how each attention mechanisms affects physiological signal analysis. To the best of our knowledge, this study is the first attempt to identify the most effective attention mechanism for physiological signal analysis using deep learning. We believe that our work will enable generalizable physiological signal deep learning, including the development of prototypes.

2 Methods

In this study, we analyze the efficacy of three CNN architectures (e.g., VGG-16 [23], ResNet-18 [10], and Inception-V1 [24]) with four types of attention mechanisms (e.g., squeeze-and-excitation (SE) [12], non-local (NL) [26], convolutional block attention module (CBAM) [27], and multi-head self-attention (MSA) [25]) for physiological signal deep learning. Each model uses unique feature extraction modules. VGG module includes two or three consecutive convolution layers and a pooling layer. ResNet module contains two consecutive convolution layers and a residual path. Inception module includes three convolution layers and a pooling layer in parallel. The CNN models used in this study are tailored to modality and dimension differences between image and physiological signal data. Detailed reduction criteria are described in Appendix.

The SE module is a *channel* attention mechanism. It encodes features with a squeeze part and decodes it with an excitation part to increase the quality of feature representation by considering the interdependency of channel information. The NL module is a *spatial* attention mechanism that calculates global feature information with covariance-like self-attention, which can overcome the locality of pixel dependency of CNN model, in which they fail to extract relational features between the first and last points of the input segment. CBAM is a *channel* + *spatial* attention mechanism. It performs channel-wise attention which is similar to SE module and performs spatial attention mechanism in that it sequentially reduces the feature size using multiple pooling and convolutional layers. The MSA module [25] is a stand-alone spatial self-attention method comprising multiple scaled dot-product attention layers in parallel, which use input data itself for queries, keys, and values. It analyzes how the given input data are self-related and helps extract enriched feature representations. The first three attention modules are harmonized to CNN models, but MSA does not use intermediate convolutional layers. A total of 13 types deep learning models (i.e., three pure CNN-based models, nine attention involved CNN-based models, and an MSA-based model) are compared.

Each model is trained to solve two representative physiological signal problems: classification for predicting intraoperative hypotension and regression for predicting intraoperative cardiac output (CO). Unexpected hypotension is a critical event that requires prompt intervention. Many risk factors have been revealed, but they do not help reduce its incidence or duration. Therefore, early prediction and prevention are crucial. Several studies have attempted to predict hypotension using deep learning [9,17]. We followed their methods of predicting hypotension events within 5 min of occurrence.

ECG, plethysmography (PPG), and demographic data were used as input variables for classification task. The output variable was binary, the positive label was defined as hypotension (mean arterial blood pressure $\leq 65 \text{ mmHg}$) lasting >1 min, and the negative label for otherwise. A pair of 20-s input segments of ECG and PPG waveforms and demographic data were extracted to predict events within 5-min. For preprocessing, we removed segments with ECG outside a range of -2 to 4.5mV or a PPG range of zero (unitless) or less.

CO, the volume of blood being pumped by the heart per minute, is used to monitor and optimize systemic oxygen and drug delivery in critically ill or high-risk surgical patients. Especially for surgical patients, it is directly related to postoperative complications; hence, immediate treatment to keep CO levels between 4 and 8L/min during surgery may improve patient outcomes [5]. However, accurate CO monitoring requires invasive catheters, which may lead to severe complications. Some previous deep learning works attempted to predict CO using the data of invasive medical devices [19, 28, 29]. However, we sought a non-invasive method. Our model allows us to monitor CO for general patients by eliminating the invasiveness.

The input variables of the regression task were the same as those of the hypotension prediction model. The output variable was stroke volume index (SVI) instead of CO so that we could return a prompt result and correct the interpatient biases. Note that $SVI = CO/(\text{heart rate (HR)} \times \text{body surface area})$. To remove outliers, only values with CO / HR between 20 and 200mL/beat were used. The 20-s segments of input were extracted to predict immediate SVI values. Preprocessing for input segments was the same as the classification task.

3 Experiments

Training and testing datasets were obtained from VitalDB [16], an open-source physiological signal database containing perioperative physiological signs of more than 6,000 surgical patients. We extracted the required tracks for each task and conducted minimal preprocessing to determine CNN models and attention mechanisms having the best model effects using real-world physiological signal data.

To measure the effectiveness of the three attention mechanisms in each CNN model, performance variations were recorded by changing the attention fraction of the attention mechanism. Attention fraction is defined as the number of attention mechanisms divided by the number of CNN modules times 100. The 0, 50, and 100% attention fractions were considered in our experiments. Each attention mechanism was applied as the end-stage of each module. Note that for a 50% attention fraction, one attention mechanism was embedded in every two CNN modules. Notably, the MSA-based model did not include a convolutional module and do not have a standardized architecture; hence, we explored various MSA-based model types using a grid search. The search spaces for self-attention had input and output dimensionalities of 16, 32, and 64, parallel attention layers (number of heads) of two, four, six, and eight, inner-layer dimensionalities of

32, 64, and 128, and identical layers (number of layers) of one, two, and three. Through the hyperparameter search, we fixed other options as to be the best performance except number of heads and number of layers and recorded performance by changing the unit of number of heads and layers. Note that unit for number of heads increase by 2 and for number of layers increase by 1. The best setting of our MSA-based model was input and output dimensionality of 32, inner-layer dimensionality of 128. A single convolutional layer was added to the input layer of each MSA-based model to match the variable dimensionality of the self-attention models.

The input data of two tasks were two-channel (ECG and PPG) 100-Hz waveforms of 20-s. Patient demographic information was concatenated after the first fully connected layer. Detailed model architectures are illustrated in Fig 1. It presents the final baseline model used. The green box (attention module) was replaced with the module required for each experiment.



(a) VGG-based model (b) ResNet-based model (c) Inception-based model (d) MSA-based model



For classification task, all models were trained with binary cross-entropy loss. The Adam optimizer was used for all models, apart from the inception-based one, which used RMSProp. The area under the receiver operating characteristics curve (AU-ROC) was used to evaluate the classification model. For the regression task, all models were trained with root mean squared error loss and the Adam optimizer. The mean absolute percentage error (MAPE) was calculated to measure model performance. Both classification and regression models

were generated with a learning rate of 0.001 set to decrease by 0.1 times every 20 epochs. A batch size of 128 was used. To derive more reliable results, all models were repeated five times for training, and their performances were compared based on mean and standard deviation. We also measured the elapsed times of model convergence at given performances. The elapsed times to reach 0.7 AU-ROC for classification and 27.0% of MAPE for regression task were considered. All experiments, apart from those of the MSA-based models, were performed using Tensorflow 2.4.1 with Python 3.9 on a 32-core AMD EPYC 7542 processor and a single NVIDIA RTX 5000 GPU. For self-attention models, we used two NVIDIA RTX 5000 GPUs with NVLink connections to supplement GPU memory.

4 Results

Totals of 3,211 and 801 cases were extracted for hypotension and CO prediction, respectively. A randomly sampled 20% of cases were used for testing. For the hypotension prediction problem, 289,775 and 74,779 samples containing 4.74 and 4.03% positive events were collected for training and testing, respectively. The CO prediction problem collected 271,288 and 64,659 samples, providing a mean SVI and a standard deviation of 42.11 ± 13.25 and 41.71 ± 12.37 , respectively, for training and testing. Patient demographic information was not different (*P*-value > 0.05) between training and testing, except that the weight and height of patients in the hypotension testing were slightly larger (Table 1).

Hypotension prediction (Classification)					
Characteristic	Training dataset	Testing dataset	<i>P</i> -value		
Age, years ^{\dagger}	61.0 (49.0-69.8)	60.0 (52.0-70.0)	0.258		
Sex, $\#$ of male (%)	1409(54.8%)	368~(57.3%)	0.278		
Height, cm^{\dagger}	162.6 (156.3-168.7)	$163.4\ (157.2-170.0)$	0.040		
Weight, kg^{\dagger}	60.0 (53.4-68.6)	61.3(53.0-68.3)	0.030		
Cardiac output prediction (Regression)					
Characteristic	Training dataset	Testing dataset	P-value		
Age, years ^{\dagger}	61.0 (52.0-70.0)	62.0 (50.0-69.0)	0.660		
Sex, $\#$ of male (%)	394~(61.3%)	90~(57.0%)	0.367		
Height, cm^{\dagger}	163.8 (157.8-169.8)	162.3 (155.4-169.2)	0.178		
Weight, kg^{\dagger}	61.5(54.2-69.5)	61.1 (53.7-68.0)	0.478		

Table 1. Patient demographics of training and testing datasets

[†] Data are represented as median (interquartile range).

The model performance variances of each CNN model and attention mechanism are illustrated in Fig 2 and 3. Regarding the classification task for predicting hypotension of Fig 2, ResNet-based model showed overall higher performance with a 50% attention fraction. ResNet-based model with NL module showed the



Fig. 2. Performance and convergence time in hypotension prediction problem. (a) is comparison of AU-ROC in the classification task. (b) is comparison of elapsed time to converge AU-ROC = 0.7



Fig. 3. Performance and convergence time in CO prediction problem. (a) is comparison of MAPE in the regression task. (b) is comparison of elapsed time to converge MAPE = 27.0%

best AU-ROC of 0.854. When examining the elapsed time needed to converge 0.7 of the AU-ROC, ResNet-based model was the fastest. Additionally, the SE module added negligible additional computing overhead, but the overall CNN performance increased. There was an obvious tendency of increased performance when using *spatial* attention (i.e., NL or CBAM module).

During CO regression prediction, as shown in Fig 3, the VGG-based model showed an overall low error. The VGG-based model with a 50% attention fraction of the CBAM module showed the best MAPE of 17.3%. However, ResNet-based model had the best convergence time to achieve 27.0% of MAPE. The computational overhead of the SE module in the three CNN models was also negligible in the regression problem, whereas it played a major role in reducing errors. Moreover, the convergence time was shortened in the ResNet-based model with SE module. There was also a clear tendency of decreasing error when using *channel* attention mechanisms (i.e., SE or CBAM module).

These experimental results can be better understood when contrasted with the problem defined. To predict hypotension within 5 min of occurrence, the most important feature is hemodynamic flow changes across 20-s of input data. Therefore, *spatial* attention plays an important role in model performance. In the prompt-CO regression problem, the waveform shape from a single beat was most important as CO is closely related to heart dynamics and the elasticity or compensation of blood vessels. Notably, each patient has a different beat pattern. Therefore, it is crucial to properly analyze the shape of the beat waveform. *Channel* attention extracts various features from the input and improves performance by helping diversify feature representations.

In both problems, the model performance was generally better when using 50% of the attention fraction rather than 100% or the fully self-attention-based model. Similar results were reported for computer vision problems [20]. We confirmed that convolution and self-attention were complementary in physiological signal deep learning, as with computer vision. Furthermore, good performance cannot be achieved by using only one building block.

5 Conclusion

In this study, we determined the best CNN and attention mechanism pairing for building deep learning models for physiological signal analysis. An attention mechanism should be selected by determining which characteristics from the raw physiological signal should be addressed to solve the problem. Convolution and attention mechanisms are complementary; therefore, there may be an ideal attention fraction for optimal performance. The ResNet-based model showed moderate performance and fast convergence in both experimental tasks. Therefore, ResNet-based model with an attention mechanism is the best candidate for prototype model. Recent studies suggest using a combined MSA with CNN for higher performance. We plan to compare physiological signal analysis performance using multiple models in a future paper. Acknowledgement. This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number : HI21C1074); and the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, Republic of Korea, and the Ministry of Food and Drug Safety) (Project Number: 202011B23)

References

- Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015 (2015)
- Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014. pp. 1724–1734 (2014)
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021 (2021)
- Faust, O., Hagiwara, Y., Hong, T.J., Lih, O.S., Acharya, U.R.: Deep learning for healthcare applications based on physiological signals: A review. Computer Methods and Programs in Biomedicine 161, 1–13 (2018)
- Giglio, M., Marucci, M., Testini, M., Brienza, N.: Goal-directed haemodynamic therapy and gastrointestinal complications in major surgery: a meta-analysis of randomized controlled trials. British Journal of Anaesthesia 103(5), 637–646 (2009)
- Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm networks. In: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005. vol. 4, pp. 2047–2052 (2005)
- Guo, M., Xu, T., Liu, J., Liu, Z., Jiang, P., Mu, T., Zhang, S., Martin, R.R., Cheng, M., Hu, S.: Attention mechanisms in computer vision: A survey. CoRR abs/2111.07624 (2021)
- Hannun, A.Y., Rajpurkar, P., Haghpanahi, M., Tison, G.H., Bourn, C., Turakhia, M.P., Ng, A.Y.: Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature medicine 25(1), 65 (2019)
- Hatib, F., Jian, Z., Buddi, S., Lee, C., Settels, J., Sibert, K., Rinehart, J., Cannesson, M.: Machine-learning Algorithm to Predict Hypotension Based on High-fidelity Arterial Pressure Waveform Analysis. Anesthesiology 129(4), 663–674 (10 2018)
- He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

- 10 Park et al.
- Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation 9(8), 1735–1780 (11 1997)
- Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
- Huang, C.Z.A., Vaswani, A., Uszkoreit, J., Simon, I., Hawthorne, C., Shazeer, N.M., Dai, A.M., Hoffman, M.D., Dinculescu, M., Eck, D.: Music transformer: Generating music with long-term structure. In: International Conference on Learning Representations, ICLR 2019 (2019)
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al.: Highly accurate protein structure prediction with alphafold. Nature 596(7873), 583–589 (2021)
- Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25 (2012)
- Lee, H.C., Jung, C.W.: Vital recorder—a free research tool for automatic recording of high-resolution time-synchronised physiological data from multiple anaesthesia devices. Scientific reports 8(1), 1–8 (2018)
- Lee, S., Lee, H.C., Chu, Y.S., Song, S.W., Ahn, G.J., Lee, H., Yang, S., Koh, S.B.: Deep learning models for the prediction of intraoperative hypotension. British Journal of Anasethesia 126, 808–817 (2021)
- Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015. pp. 1412–1421 (2015)
- Moon, Y.J., Moon, H.S., Kim, D.S., Kim, J.M., Lee, J.K., Shim, W.H., Kim, S.H., Hwang, G.S., Choi, J.S.: Deep learning-based stroke volume estimation outperforms conventional arterial contour method in patients with hemodynamic instability. Journal of clinical medicine 8(9), 1419 (2019)
- 20. Park, N., Kim, S.: How do vision transformers work? In: International Conference on Learning Representations (2022)
- Pascanu, R., Mikolov, T., Bengio, Y.: On the difficulty of training recurrent neural networks. In: International conference on machine learning. pp. 1310–1318 (2013)
- Sajad Mousavi, F.A., Acharya, U.R.: Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach. arXiv preprint arXiv:1903.02108 (2019)
- Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems 30. pp. 5998–6008 (2017)
- Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
- Woo, S., Park, J., Lee, J., Kweon, I.S.: CBAM: convolutional block attention module. In: Computer Vision - ECCV 2018. vol. 11211, pp. 3–19 (2018)
- 28. Yang, H.L., Jung, C.W., Yang, S.M., Kim, M.S., Shim, S., Lee, K.H., Lee, H.C.: Development and validation of an arterial pressure-based cardiac output algorithm using a convolutional neural network: Retrospective study based on prospective registry data. JMIR Med Inform 9(8), e24762 (Aug 2021)

- Yang, H.L., Lee, H.C., Jung, C.W., Kim, M.S.: A deep learning method for intraoperative age-agnostic and disease-specific cardiac output monitoring from arterial blood pressure. In: 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE). pp. 662–666 (2020)
- 30. Zhou, Y., Roy, S., Abdolrashidi, A., Wong, D., Ma, P., Xu, Q., Liu, H., Phothilimtha, P., Wang, S., Goldie, A., Mirhoseini, A., Laudon, J.: Transferable graph optimizers for ml compilers. In: Advances in Neural Information Processing Systems. vol. 33, pp. 13844–13855 (2020)

A Appendix

The original CNN models (e.g., VGG-16, ResNet-18, and Inception-V1) used 224×224 image inputs and analyzed them with two-dimensional (2D) convolutions. However, in our research, the dimensionality of input data should be one-dimensional (1D). Therefore, all 2D convolutional operations were replaced with 1D convolutions. Additionally, input sizes were much smaller at 224×224 = 50,176 vs. 2,000. We thus reduced the model depth to prevent overfitting caused by superfluous immoderate trainable parameters. Let our input data size of 2,000 to be 2D. 2,000 $\approx 45 \times 45$. Thus, the ratio between image data used in the original CNN studies and our physiological data was 224 / $45 \approx 5$. Therefore, we used the model reduction ratio of five for each CNN model. The main characteristics of CNN models was the modules they contained. The VGG module included two or three consecutive convolution layers and a pooling layer. The ResNet module contained two consecutive convolution layers and a residual path. The inception module included three convolution layers and a pooling layer in parallel. To maintain each model's identity, we set cutoff criteria while preserving the modules. Fig A1 illustrates the shallow part of each CNN model.



Fig. A1. Example of the levels of each model

Note that the level indicates a section divided while maintaining the module's property. To find the optimal subset of the CNN model for our study, we compared the number of training parameters by dividing the model by levels. The fully connected part (the classification or regression part) of the original model was added to the subset model. Additional concatenating layers for patient demographic data were added in the last fully connected part. Table A1 presents the number of trainable parameters divided by the level of each model. ResNet and Inception models showed fewer trainable parameters as they were divided at shallow levels, whereas VGG showed more parameter increases owing to the growth of feature sizes entering the fully connected layer without global-average

Level	VGG-based model	ResNet-based model	Inception-based model	
	Trainable param.	Trainable param.	Trainable param.	
1	192,128,065	26,048	117,744	
2	189,891,329	51,008	$297{,}584$	
3	$130,\!614,\!785$	134,080	$538,\!592$	
4	90,639,361	$233,\!152$	$806,\!504$	
5	$40,\!567,\!296$	563, 136	1,089,280	
6	-	$957,\!888$	$1,\!404,\!864$	
7	-	$2,\!273,\!216$	$1,\!884,\!096$	
8	-	$3,\!849,\!152$	$2,\!538,\!432$	
Default	40 567 206	2 940 152	2 417 964	
param.	40,507,290	3,649,152	3,417,204	
Default/5	8 133 450	760.830	683 453	
param.	0,100,409	103,000	000,400	

 Table A1. Trainable parameters for each level of each model

pooling. We selected a VGG five levels (full model), a ResNet six levels, and an inception with four levels as our baseline CNN architecture.