CAPSTONE PROJECT DOCUMENT

BLOCK-WISE ATTENTION-DRIVEN SOFT SEGMENTATION FOR IMBALANCED MULTI-LABEL CHEST X-RAY CLASSIFICATION

DSP391m - Group 4

Authors:
Nguyen Huu Duy - SE183995
Hoang Khuong Duy - SE184883
Tran Khanh Nguyen - SE183486
Nguyen Khac Vuong - SE183769
Supervisor: Dr. Huynh Cong Viet Ngu
Department of Artificial Intelligence FPT University
Ho Chi Minh, Vietnam

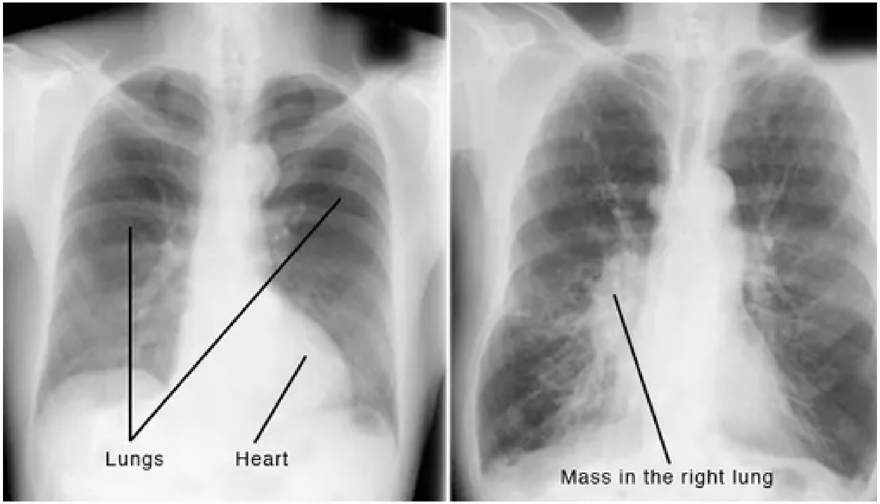
Outlines

1. Introduction
2. Related Work
3. Main Contribution
4. Experiment
5. Demo
6. Conclusion
7. References

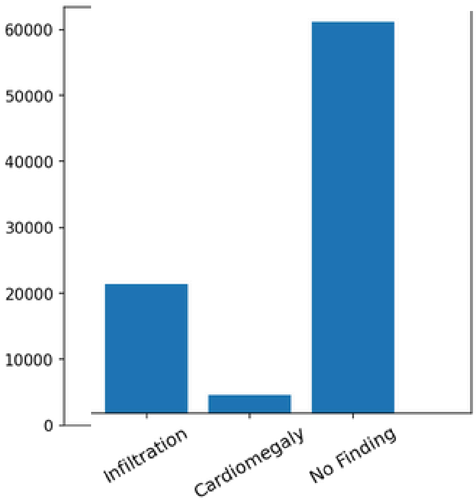
1. Introduction

Background Context

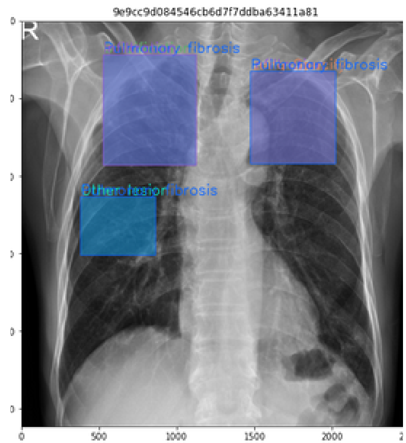
- Chest X-ray can help doctors see clearly the inside of the lungs and heart
- AI models to support more effective diagnosis.



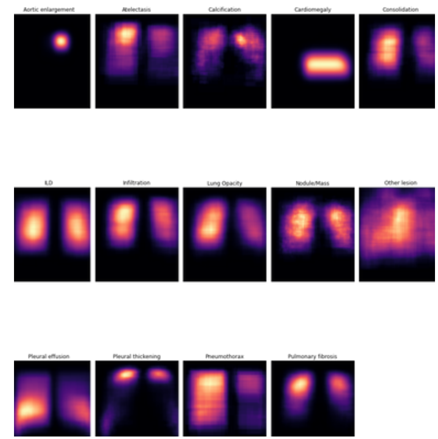
Common Problems



Class Imbalance



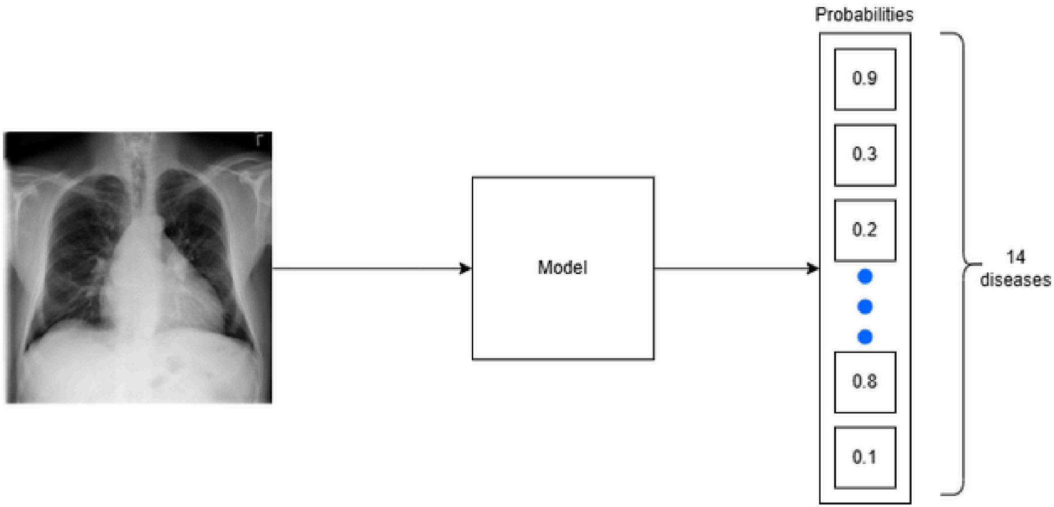
Annotation Limitations



Feature Discrimination

Research Questions

- How to improve the performance of multi-label thoracic disease classification by **extracting specialized feature** with low resource consumption? → **Lightweight attention modules** integrated into CNN backbones
- Are current traditional evaluation methods suitable for the classification problem for **imbalance dataset?** → **Focal Loss**



Our contributions

- **Solutions**
  - We proposed an end-to-end model comprising following methods:
    - Apply Attention Modules inspired by **Large Kernel Attention (LKA) as Soft Segmentation**
    - **Block-Wise Attention** within CNNs (DenseNet121, VGG16)
    - Two-Stage Training Strategy (**BCE Loss + Focal Loss**)
- **Achievements**
  - Achieves **0.8818** average AUC, surpassing state-of-the-art methods
  - **Computational efficiency**

2. Related Work

Learning Techniques

- **Common Workflow**
  - **Data Pre-processing**
    - Clean and normalize raw data
    - Prepare data format for model input
  - **Learning Representation Space**
    - Encode data into meaningful embeddings
    - Capture essential patterns or features
  - **Learning the Classifier**
    - Map embeddings to class labels
    - Minimize prediction errors

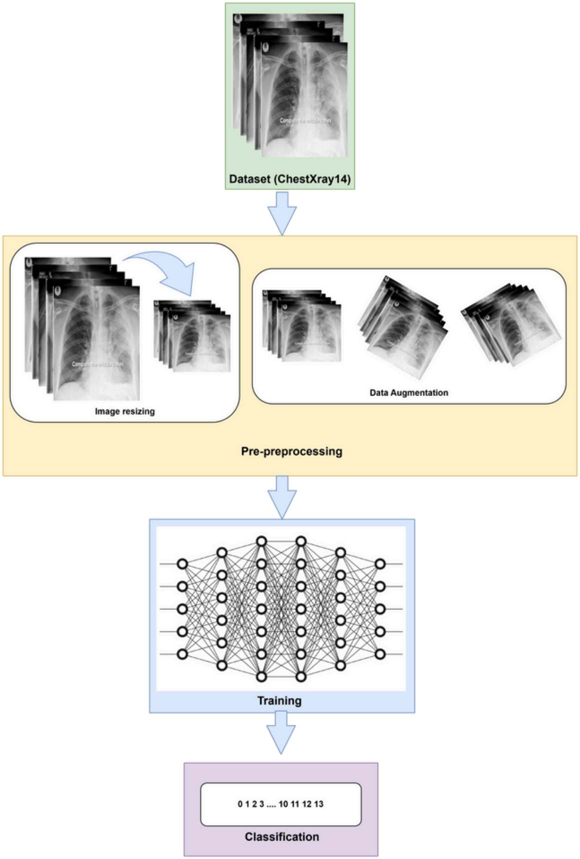
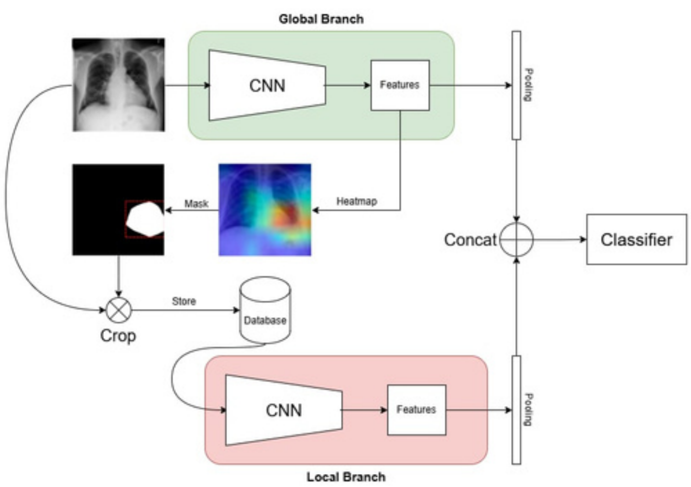


Fig: The workflow of Learning Technique for ChestXray14 multi-label classification.

AG-CNN Architecture (1/2)

- **Their Contributions:**
  - **Global branch:** Analyzes the entire X-ray image
  - **Local branch:** Focuses on pathological regions through attention mechanism
  - **Fusion branch:** Combines information



[*] L. Guan et al. “AG-CNN: Attention Guided Convolutional Neural Network for Thoracic Disease Classifica tion”. In: IEEE Transactions on Medical Imaging 37.6 (2018), pp. 1324–1333.

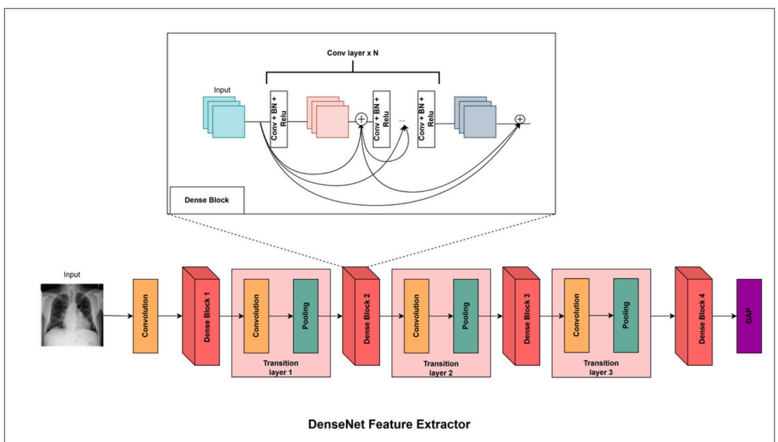
AG-CNN Architecture (2/2)

- **Advantages:**
  - Learns to ignore irrelevant background and emphasizes diagnostic regions.
  - Global, local, and fusion branches capture both context and fine-grained lesion details, boosting classification accuracy.
- **Limitations:**
  - Three branches increase computational, making real-world deployment challenging

[*] L. Guan et al. “AG-CNN: Attention Guided Convolutional Neural Network for Thoracic Disease Classifica tion”. In: IEEE Transactions on Medical Imaging 37.6 (2018), pp. 1324–1333.

DenseNet121 Extractor (1/2)

- **Core Components:**
  - **Dense block:** Convolution layer, Batch Normalization and ReLU function.
  - **Transition layer:** Use to connect dense blocks, has 2 main purpose:
    - Reduce number of feature map
    - Downsampling the dimension of feature map

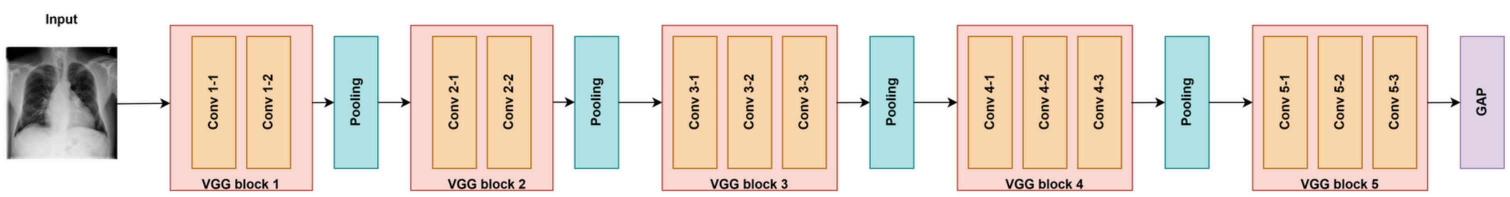


DenseNet121 Extractor (2/2)

- **Advantages:**
  - Reduce vanishing gradient
  - Feature reuse
  - Fewer parameters
- **Limitations:**
  - High Memory Consumption
  - Computational Complexity
  - Implementation Complexity

VGG16 Extractor (1/2)

- **Core Components:**
  - Contains 16 weight layers: 13 Convolutional layers and 3 Fully Connected layers
  - Uses small (3×3) convolution filters with stride 1 and padding 1
  - Includes 5 Max Pooling layers (2×2, stride 2) for downsampling
  - Uses ReLU activation after each convolution



VGG-16 Feature Extractor

VGG16 Extractor (2/2)

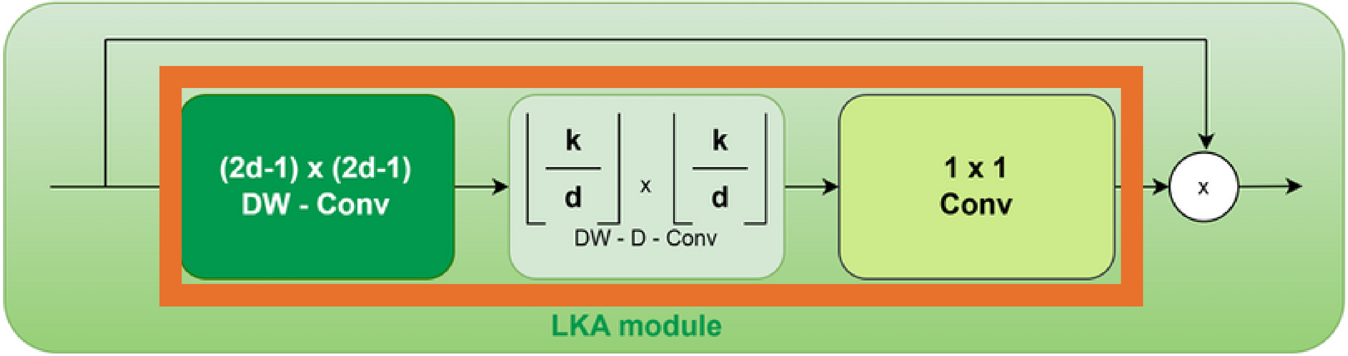
- **Advantages:**
  - Basic architecture
  - Strong in feature extractor
  - Suitable for pretraining
- **Limitations:**
  - Lots of parameter
  - High risk of overfitting because of not having skip connection

Large Kernel Attention (LKA)

- Based on the attention mechanism of Visual Attention Network*
- Mathematical Formulation of LKA's Attention Map

$$a = A(u) = W_p * (K_2 * (K_1 * u))$$

where $K_1 \in \mathbb{R}^{C\times1\times k_1\times k_1}$: Depthwise convolution kernel with $k_1 = 5$, $K_2 \in \mathbb{R}^{C\times1\times k_2\times k_2}$: Dilated depthwise convolution with $k_2 = 7$, dilation $d = 3$, $W_p \in \mathbb{R}^{C\times C\times1\times1}$: Pointwise convolution for channel mixing.



[*] Guo, M.H., Lu, C.Z., Liu, Z.N., Cheng, M.M., Hu, S.M.: Visual attention network. Computational visual media 9(4), 733–752 (2023)

Receptive Field in LKA

- The receptive field of LKA is:

$$R = k_2 + (k_2 - 1)(d - 1) - \text{padding} = 7 + (7 - 1)(3 - 1) - 6 = 7 + 12 - 6 = 13$$

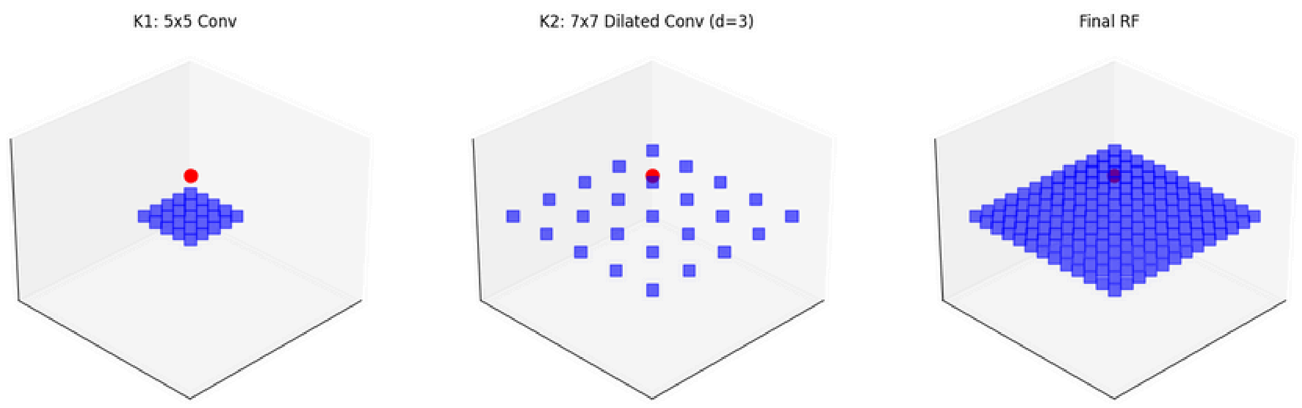
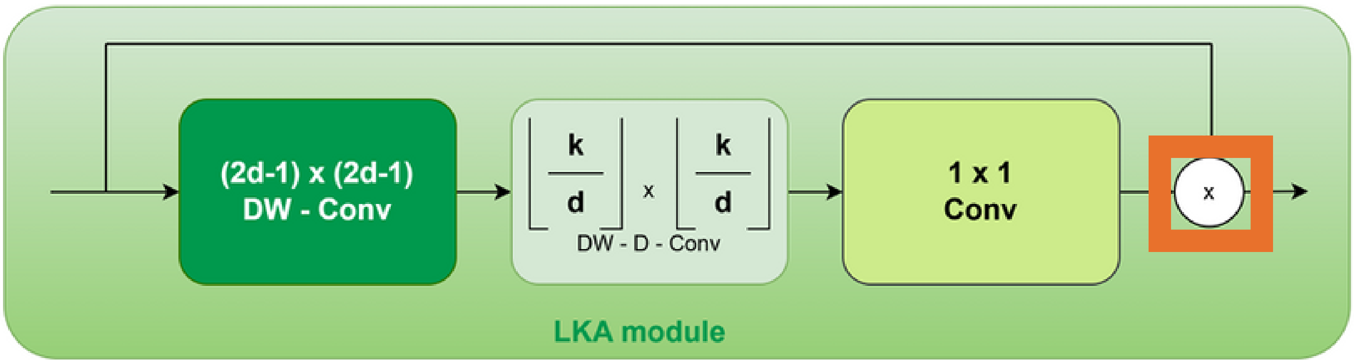


Fig: Illustration of the receptive field of the LKA modules.

LKA as Soft Segmentation

- Using element-wise product, the attention map acts as a soft mask to highlight important regions of the input.



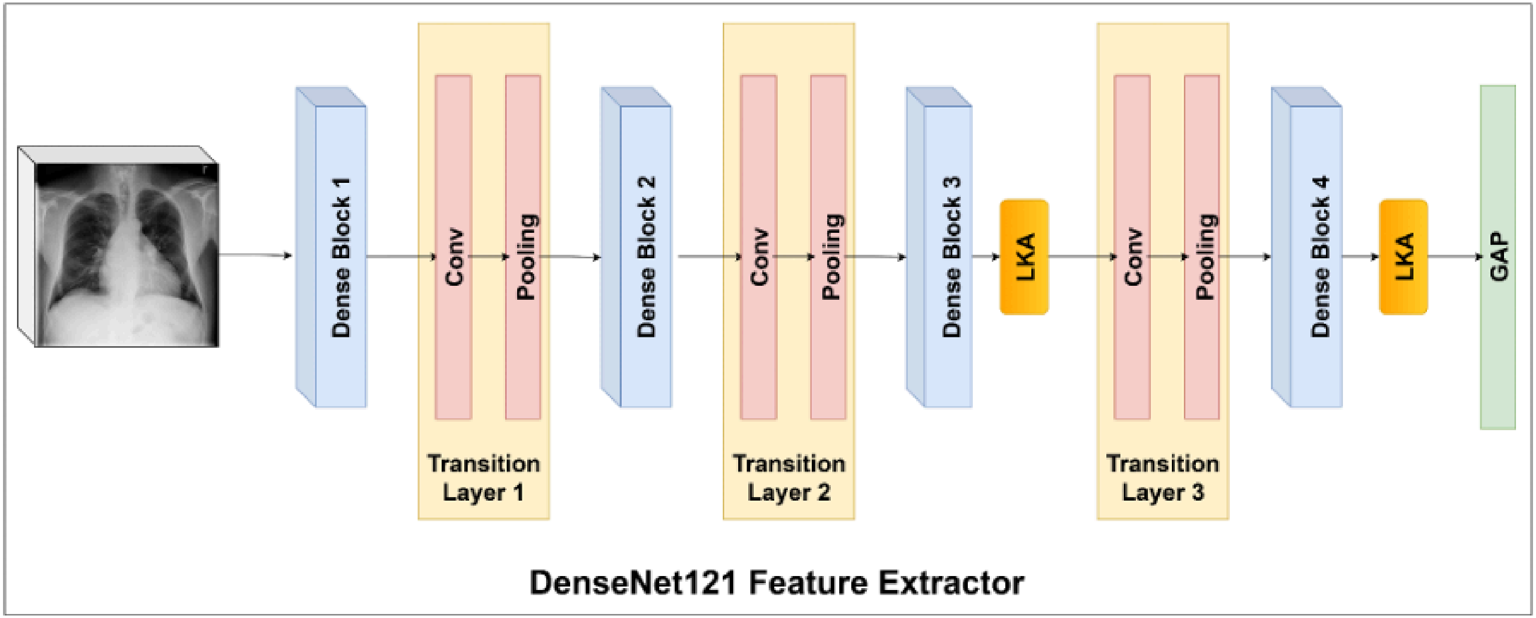
$$y_{i,j,c} = u_{i,j,c} \cdot \boxed{a_{i,j,c}}$$

Soft Mask

3. Main Contribution

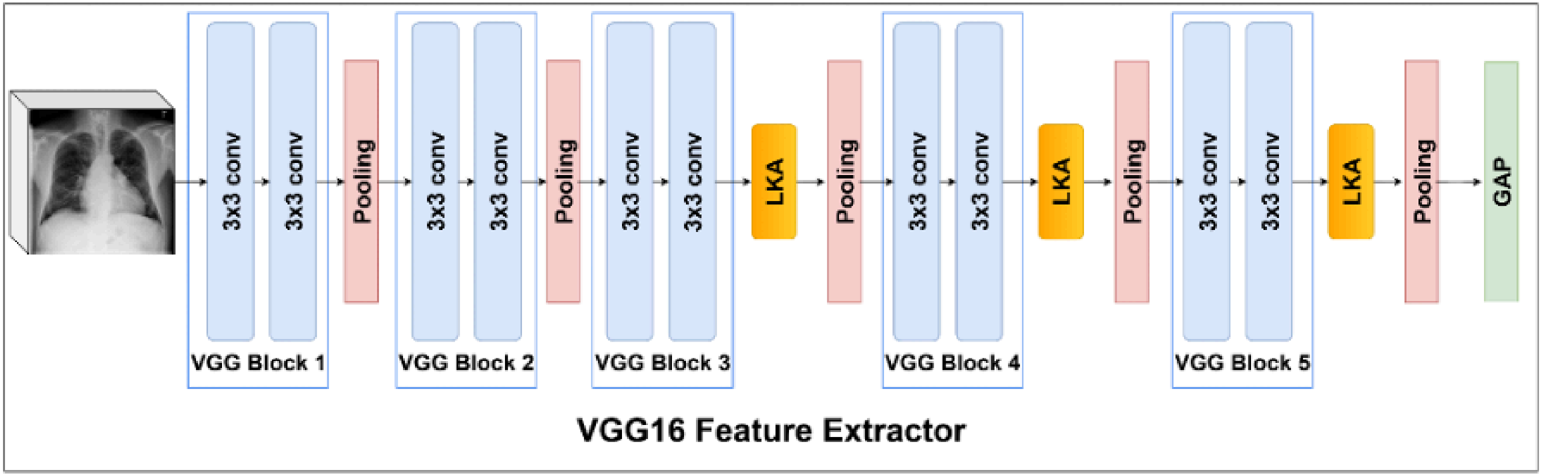
Block-Wise DenseNet121

- Taking advantage of DenseNet121's strengths, we integrate LKA into the following blocks.



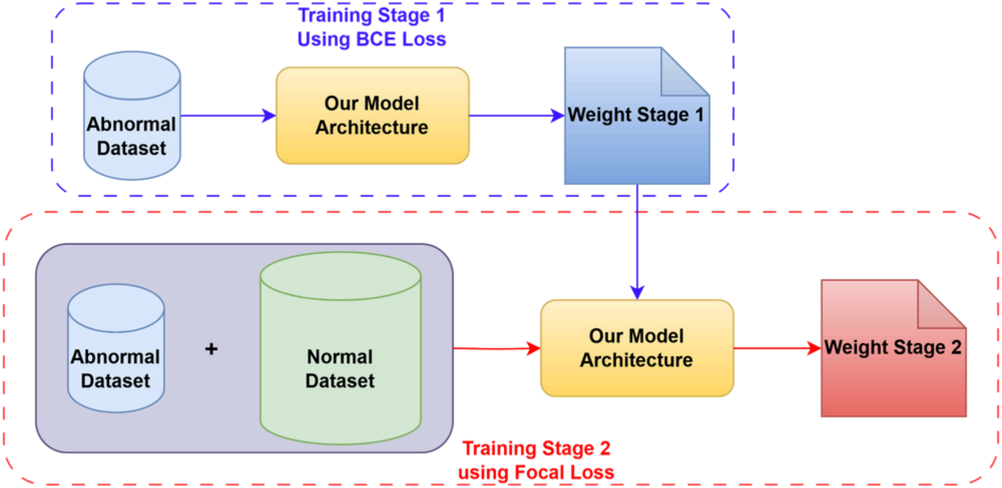
Block-Wise VGG16

- Taking advantage of VGG16's strengths, we integrate LKA into the following blocks.



Training Strategy

- **Based** on the **training strategy of SynthEmsemble***
- **Change:** Employs **Focal Loss** in the 2nd stage instead of standard loss
- **Insight:** To better handle the severe class imbalance



[*] Ashraf, S.N., Mamun, M.A., Abdullah, H.M., Alam, M.G.R.: Synthensemble: a fusion of cnn, vision transformer, and hybrid models for multi-label chest x-ray classification. In: 2023 26th International Conference on Computer and Information Technology (ICCIT). pp. 1–6. IEEE (2023)

Loss Function

- **Stage 1:** Abnormal Data Pretraining using **binary cross-entropy loss**

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{N_c}\left[y_{i,c}\log(\sigma(z_{i,c})) + (1 - y_{i,c})\log(1 - \sigma(z_{i,c}))\right]$$

- **Stage 2:** Full Data Training with **Focal Loss***

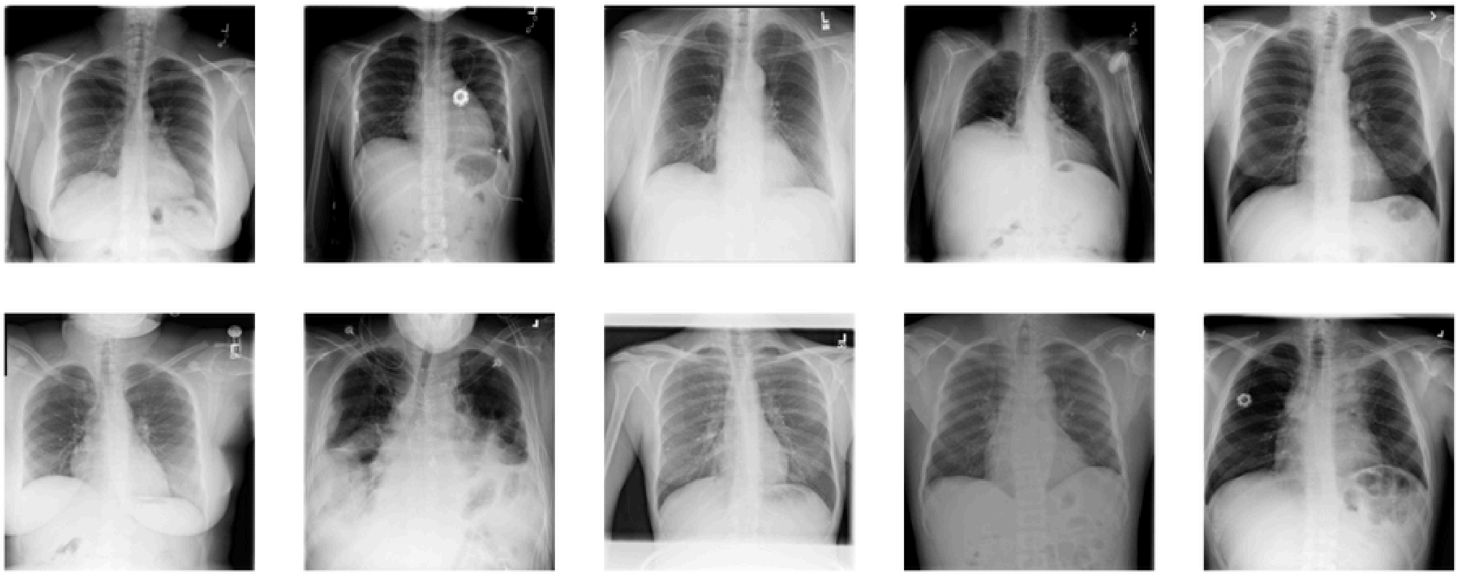
$$L_{FL} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{N_c}\alpha_c(1 - p_{i,c})^{\gamma}y_{i,c}\log(p_{i,c})$$

[*] Lin, T.Y., Goyal, P., Girshick, R., He, K., Doll´ar, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)

4. Experiment

Dataset: ChestXray14*

- **112,120** frontal-view X ray images from **30,805** unique patients.
- Each image labeled for the presence of **14** thoracic pathologies.



[*] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)

Imbalance Dataset

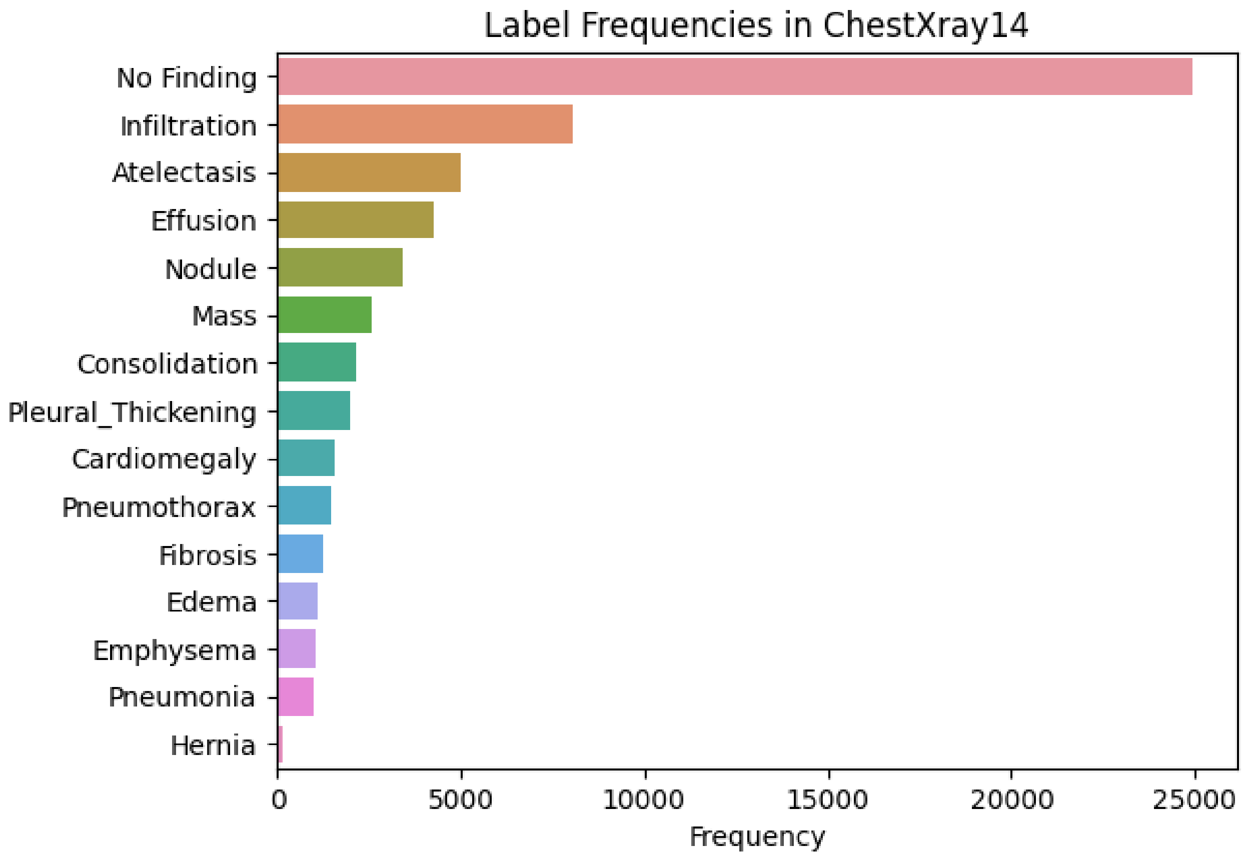


Fig: Diseases Frequencies

Setting Training

- **Metric:** AUC
- **Preprocessing:**
  - **Image resizing:** All images resized to 224×224 pixels.
  - **Normalization:** Normalized using ImageNet statistics.
  - **Data Augmentation:** Applied during training to improve generalization:
    - Random rotation (±10°)
    - Horizontal flipping
    - Random cropping and zooming
    - Patient-wise splitting: Ensured no patient appears in both training and test sets to avoid data leakage, divide into 70-10-20.
- **Accelerator:** P100 on Kaggle

Splitting Dataset

Stage	Training Set	Validation Set	Test Set
Stage 1	36,506 images	5,215 images	-
Stage 2	78,544 images	11,220 images	22,356 images

Performance Comparison SOTA

FPT Education
FPT UNIVERSITY

TABLE I: Comparison of AUC scores with state-of-the-art methods on ChestXray14 dataset.

Pathology	Wang et al. [1]	CheXNet [2]	SynthEnsemble [3]	AG-CNN (ResNet-50) [4]	AG-CNN (DenseNet-121) [4]	Ours (BCELoss 2 Stage)	Ours (BCELoss + FocalLoss)
Atelectasis	0.716	0.8094	0.83390	0.844	0.853	0.853	0.858
Cardiomegaly	0.807	0.9248	0.91954	0.937	0.939	0.941	0.945
Effusion	0.784	0.8638	0.88977	0.904	0.903	0.909	0.909
Infiltration	0.609	0.7345	0.74102	0.753	0.754	0.749	0.755
Mass	0.706	0.8676	0.87315	0.893	0.902	0.899	0.907
Nodule	0.671	0.7802	0.80611	0.827	0.828	0.836	0.838
Pneumonia	0.633	0.7680	0.77648	0.776	0.774	0.807	0.815
Pneumothorax	0.806	0.8887	0.90164	0.919	0.921	0.93	0.932
Consolidation	0.708	0.7901	0.81575	0.842	0.842	0.832	0.833
Edema	0.835	0.8878	0.91034	0.919	0.924	0.935	0.935
Emphysema	0.815	0.9371	0.92946	0.941	0.932	0.955	0.956
Fibrosis	0.769	0.8047	0.83347	0.857	0.864	0.841	0.856
Pleural Thickening	0.708	0.8062	0.81270	0.836	0.837	0.846	0.853
Hernia	0.767	0.9164	0.91723	0.903	0.921	0.942	0.954
Mean AUC	0.738	0.841	0.85433	0.868	0.871	0.8768	0.8818

Magenta indicates the highest value, and Cyan indicates the second-highest value for each pathology.

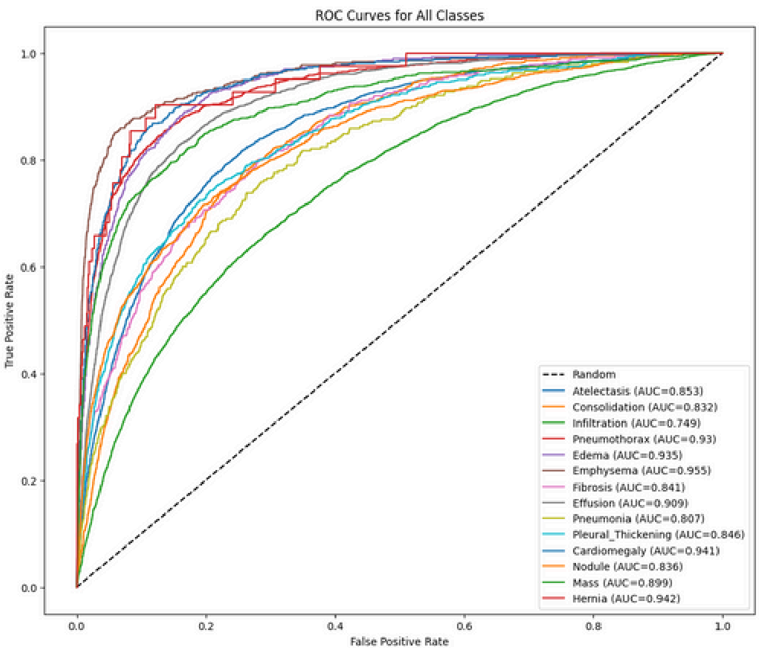
[1] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)

[2] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017)

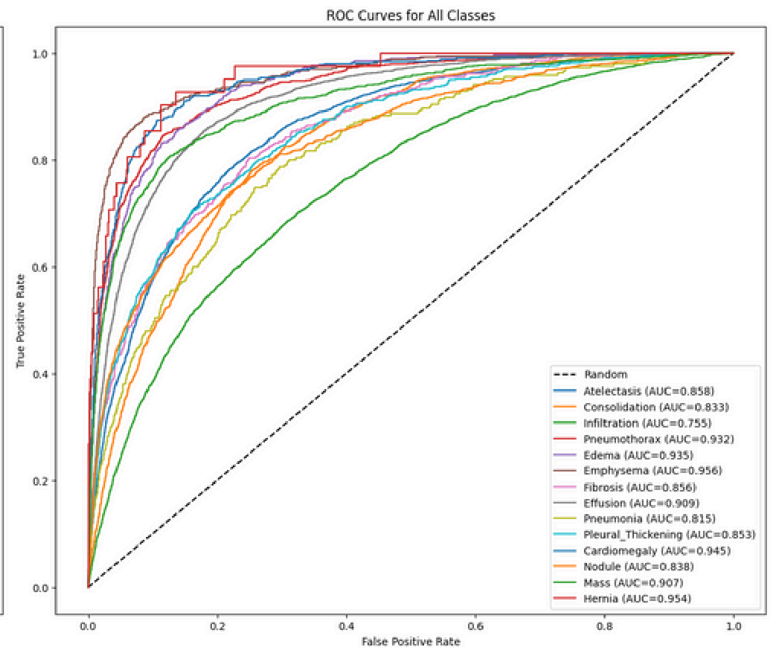
[3] Ashraf, S.N., Mamun, M.A., Abdullah, H.M., Alam, M.G.R.: Synthensemble: a fusion of cnn, vision transformer, and hybrid models for multi-label chest x-ray classification. In: 2023 26th International Conference on Computer and Information Technology (ICCIT). pp. 1–6. IEEE (2023)

[4] Guan, Q., Huang, Y., Zhong, Z., Zheng, Z., Zheng, L., Yang, Y.: Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. arXiv preprint arXiv:1801.09927 (2018)

ROC AUC



(a) ROC AUC of Ours (BCELoss 2 Stage).



(b) ROC AUC of Ours (BCELoss + FocalLoss).

Fig: ROC AUC of Ours.

Parameters and Training time Compared to SynthEnsemble (1/2)

Ours						
Model variant	**Trainable params**	**Total params**	**Training stage 1**	**Training stage 2**	**Total**	**GFLOPS**
Densenet121+LKA full	9609870	9609870	6563.2s	8588.8s	**15152s**	**3. 64**
Densenet121+LKA 2 block	9223054	9223054	6375.6s	8239s	**146.146**	**3. 14**
Densenet121+LKA after	8095630	8096530	5611.4s	6957.3s	**12568.7s**	**2. 92**
VGG16+LKA full	15453966	15453966	14225.2s	10206.1s	**24431.3s**	**16. 70**
VGG16+LKA 3 block	15418702	15418702	8628.6s	5682.1s	**14310.7s**	**15. 93**
VGG16+LKA after	15031886	15031886	4851.2s	5492.6s	**10343.8s**	**15. 39**

Parameters and Training time Compared to SynthEnsemble (2/2)

Synth Ensemble				
Model	**Trainable Params**	**Total Params**	**FLOPS (GFLOPS)**	**Training time**
CoAtNet	1128320	73904264	14. 55	Stage 1: 21045.6s Stage 2: 6811.2s
DenseNet121	1144512	8014720	2. 87	Stage 1: 40958.1s
MaxViTV2	1122944	116128376	23. 88	Stage 1: 13791.6s
SwinV2	842240	49725140	9. 08	Stage 1: 31071.3s
VOLO D2	319872	57923696	14. 24	Stage 1: 32398.5s
convnextv2	1647488	198007040	34. 4	Stage 1: 35538.9 Stage 2: 11597.6s
Total	**6205376**	**503703236**	**99. 02**	**~193212.8s**

Ablation Studies (1/3)

- **Impact of Loss Function in Single vs Two-Stage Training**

Model Variant	One-Stage BCE	One-Stage Focal	Two-Stage
DenseNet121	0. 8227	0. 8044	0. 8402
DenseNet121 + AWBs	0. 8340	0. 8369	**0. 8498**
VGG16	0. 8275	0. 8094	0. 8401
VGG16 + AWBs	0. 8238	0. 8247	**0. 8818**

- **Statement:** Two-Stage strategy enables the model to better learn and focus on relevant features → improved performance.

Ablation Studies (2/3)

- **Impact of Attention Module Integration Position**

Model	Position	Mean AUC
DenseNet121 + LKA	After	0. 8475
DenseNet121 + LKA	Full Block	0. 8416
DenseNet121 + LKA	Last 2 Block	**0. 8498**
VGG16 + LKA	After	0. 8470
VGG16 + LKA	Full Block	0. 8666
VGG16 + LKA	Last 3 Block	**0. 8818**

- **Statement:** Optimal placement of attention modules <=> higher-level features are consolidated

→ yields the best performance

Ablation Studies (3/3)

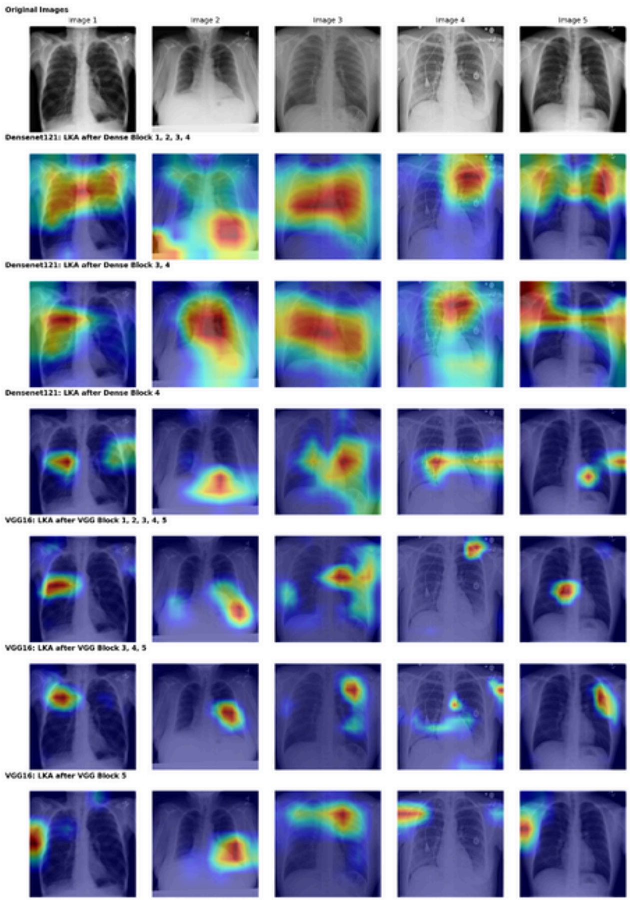


Fig: GRADCAM for Illustration of Attention Module based on Different Methods.

GRADCAM

- **"Soft segmentation"**

→ The model adaptively focuses on **important areas**

→ Improving the **refine segmentation** process.

5. Demo

System Workflow

- After training we use the best weight for demo, we will highlight important area via GRAD-CAM and give the predicted results of model and ground truth labels.

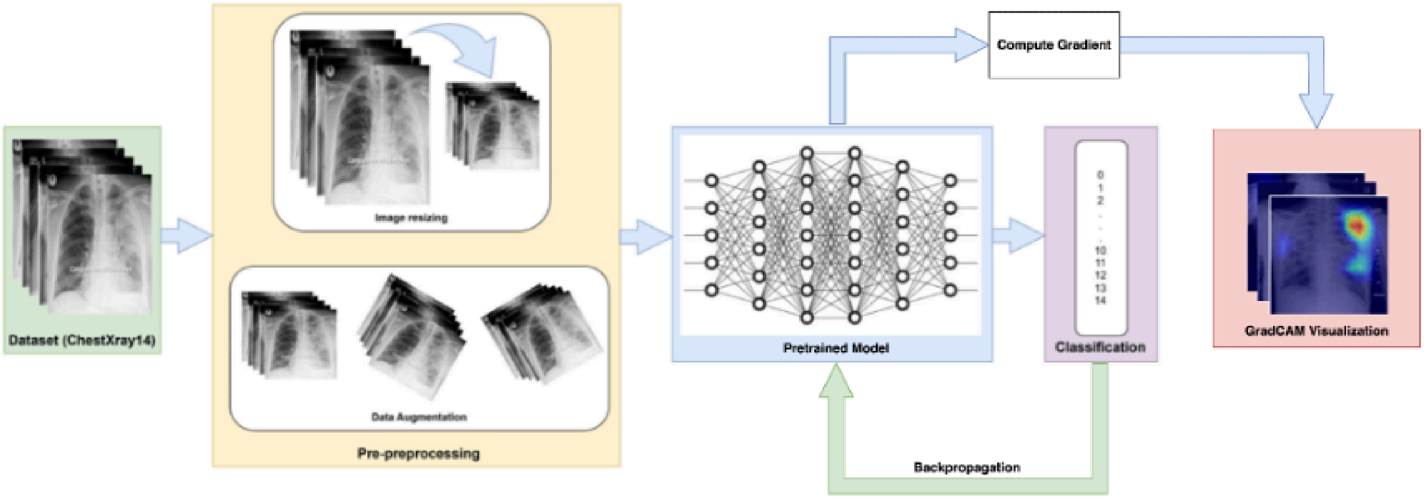


Fig: Our system workflow

6. Conclusion

Our contribution

- **Solutions**
  - We proposed an end-to-end model comprising following methods:
    - Apply Attention Modules as Soft Segmentation
    - Block-Wise Attention within CNNs
    - Two-Stage Training Strategy
- **Achievements**
  - Achieves 0.8818 average AUC, surpassing state-of-the-art methods
  - Computational efficiency.

7. References

References (1/4)

FPT Education
FPT UNIVERSITY

1. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)
2. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data 6(1), 317 (2019)
3. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
4. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., S´anchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis 42, 60–88 (2017)
5. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017)

References (2/4)

6. Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., Lyman, K.: Learning to diagnose from scratch by exploiting dependencies among labels. arXiv preprint arXiv:1710.10501 (2017)

7. Ashraf, S.N., Mamun, M.A., Abdullah, H.M., Alam, M.G.R.: Synthensemble: a fusion of cnn, vision transformer, and hybrid models for multi-label chest x-ray classification. In: 2023 26th International Conference on Computer and Information Technology (ICCIT). pp. 1–6. IEEE (2023)

8. Taslimi, S., Taslimi, S., Fathi, N., Salehi, M., Rohban, M.H.: Swinchex: Multilabel classification on chest x-ray images with transformers. arXiv preprint arXiv:2206.04246 (2022)

9. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)

10. Manzari, O.N., Ahmadabadi, H., Kashiani, H., Shokouhi, S.B., Ayatollahi, A.: Medvit: a robust vision transformer for generalized medical image classification. Computers in biology and medicine 157, 106791 (2023)

References (3/4)

11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

12. Yan, C., Yao, J., Li, R., Xu, Z., Huang, J.: Weakly supervised deep learning for thoracic disease classification and localization on chest x-rays. In: Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics. pp. 103–110 (2018)

13. Yan, Y., Kawahara, J., Hamarneh, G.: Melanoma recognition via visual attention. In: Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26. pp. 793–804. Springer (2019)

14. Guo, M.H., Lu, C.Z., Liu, Z.N., Cheng, M.M., Hu, S.M.: Visual attention network. Computational visual media 9(4), 733–752 (2023)

15. Lin, T.Y., Goyal, P., Girshick, R., He, K., Doll´ar, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)

16. Guan, Q., Huang, Y., Zhong, Z., Zheng, Z., Zheng, L., Yang, Y.: Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. arXiv preprint arXiv:1801.09927 (2018)

References (4/4)

17. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)

18. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708)

19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

20. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Gradcam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)

THANK YOU FOR YOUR ATTENTION !

Q&A