Federated Learning and Deep Learning Architecture for Pneumonia Classification using Chest X-Ray Images

Mirza Md. Nazmus Sakib
Department of Computer Science and Engineering
BRAC University, Bangladesh
Email: mirza.mohammad.nazmus.sakib@g.bracu.ac.bd

Irfana Afifa
Department of Computer Science and Engineering
BRAC University, Bangladesh
Email:irfana.afifa@g.bracu.ac.bd

Md Sabbir Hossain
Department of Computer Science and Engineering
BRAC University, Bangladesh
Email:md.sabbir.hossain1@g.bracu.ac.bd

Md Mostakin Alam
Department of Computer Science and Engineering
BRAC University, Bangladesh
Email:md.mustakin.alam@g.bracu.ac.bd

Annajiat Alim Rasel
Department of Computer Science and Engineering
BRAC University, Bangladesh
Email:annajiat@gmail.com

***Abstract*—Pneumonia Detection has been a real problem for the last few centuries. Detecting Pneumonia has been a job for the skilled, such as doctors and medical practitioners. Visiting doctors in this time in many countries is very tough with Covid-19 on the rise and stricter lockdown regulations. Deep Learning has helped build many systems and algorithms over the years to detect pneumonia using X-ray images. Such Deep Learning models are first trained on many X-ray images that would be collected from multiple hospitals and diagnostic centers and then can be deployed centrally for people to use them. However, building such models is impeded by the problem of garnering mass data from hospitals due to data confidentiality between patients and hospitals. For that, we propose a system where detecting Pneumonia would be done using a Deep Learning model with a Federated Learning approach and achieve an accuracy of around 90**

***Index Terms*—Pneumonia, Deep Learning, Federated Learning, X-ray images, dataset.**

I. Introduction

Pneumonia is a serious hazard in underdeveloped countries, where billions of people live below the poverty line and where the environment is unsafe for health. The World Health Organization (WHO) estimates that air pollution contributes to more than 4 million deaths from illnesses each year, including pneumonia. Pneumonia affects more than 150 million people annually, most of whom are children under the age of five. In rural and underdeveloped regions, the problem could be made worse by a lack of infrastructure and medical resources. In Africa, there is a 2.3 million people medical workforce shortfall. Residents of these communities typically do not have access to the timely and specialized medical care needed to properly treat pneumonia. Even in the case of an emergency, treatment might be highly expensive. Later-stage pneumonia diagnosis increases the cost and difficulty of treatment. If pneumonia is not identified and treated very away, it can be dangerous. This is quite important, especially with relation to newborns. Early diagnosis of pneumonia is crucial for lowering disease severity, avoiding unneeded hospitalizations, and saving money on medication. This in-depth study aims to develop a model that can outperform the individual federated Learning and deep learning models employed in the study for predicting the results of a chest X-ray disease assessment.Treatment can be quite costly, even in the event of an emergency. Diagnosing pneumonia at a later stage makes therapy more expensive and challenging. Pneumonia can be dangerous if not diagnosed and treated in time. This is of paramount importance, particularly with regards to infants The final output and a thorough understanding of the classification aspects are the main goals of this endeavor. [1] In the early detection of pneumonia is essential to reducing the severity of the disease, preventing unnecessary hospitalizations, and reducing drug costs. The of this in-depth research is to create a model for predicting the outcome of a chest X-ray disease examination that can outperform the individual deep learning models used in the study. [2].

II. Related Works

Research on lung related diseases has also been done on various other domains of datasets apart from the x-ray image dataset. Several researchers have proposed different algorithms for the diagnosis of lung diseases based on sound data. One of the parameters used for the detection of pulmonary sound is entropy. There are differences in the sound of a normal respi-

ratory system and a system with the pathologies of pneumonia. A. Rizal et al [3] discussed several measures of entropy for the classification of pneumonia based on pulmonary sounds. The paper revealed that the usage of a single entropy was not enough to achieve high accuracy. Therefore, seven entropies were applied which achieved 94.95using multilayer perceptron. In this paper [4], the researcher completed the research in two stages. In the first stage, heatmaps of different CNN models were generated and combined in the ensembled model. Then, They used the XAI technique to identify the region of interest for the classification. By which explainability and interpretability problems can be reduced. They ensembled the best performing models and then tested them on a small dataset of pediatric X-rays. In the second stage, a new ensembled model is generated and trained with a smaller dataset. They believed that their newly created model had higher accuracy than the other pneumonia detection dataset. In another study [5], the authors created a dataset consisting of 35,389 chest x-ray images and trained a prediction model which is capable of detecting pneumonia. The Bayesian network is used to create an XAI model from different CNN models. The findingsshow that multi-source data have improved efficiency and provide an intuitive description of diagnostic results. The researchers proposed and built XAI approaches for COVID-19 classification models in the research, as well as comparing them. The findings suggest that by providing more detailed information from the learned XAI models' outputs, quantitative and qualitative visualizations might help physicians comprehend and make better judgments. The results outperformed those of state-of-the-art approaches, and the method outperformed the commonly used ensemble techniques. This study created [6] an automated CAD system that employs deep transfer learning to categorize chest X-ray pictures into two categories: "Pneumonia" and "Normal". The ensemble architecture uses the decision scores from three CNN models, GoogleNet, ResNet-18, and DenseNet-121, to construct a weighted average ensemble. The classifier weights were calculated by combining precision, recall, f1-score, and AUC using the hyperbolic tangent function. On the Kermany dataset, the framework achieved 98.81a unique approach, resulting in a weighted average ensemble strategy. The scores of four typical assessment metrics, precision, recall, f1-score, and area under the curve, are fused to generate the weight vector, which was frequently set experimentally in studies in the literature, an approach that is prone to inaccuracy. Using a five-fold cross-validation scheme [7], the suggested approach was tested on two publicly available pneumonia X-ray datasets provided by Kermany et al. and the Radiological Society of North America (RSNA).

III. Research Methodology

The top level overview of the PNEXAI model in Figure 1 gives an overview of each step we have done to train and evaluate our models.

Initially, we gathered and preprocessed our data, which comprised data scaling and augmentation. Then, we trained

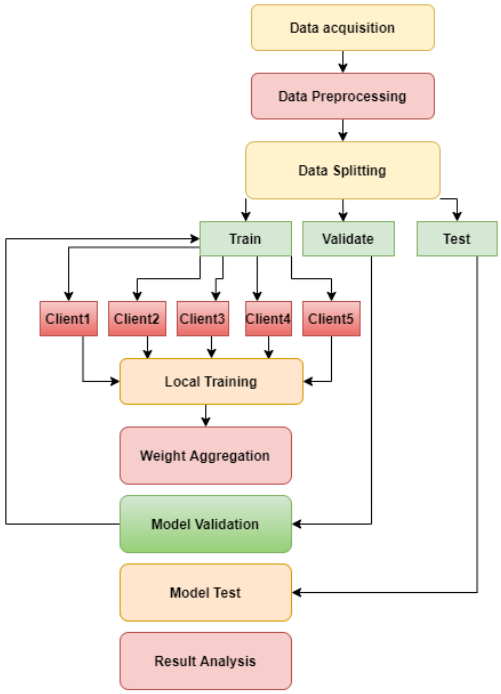
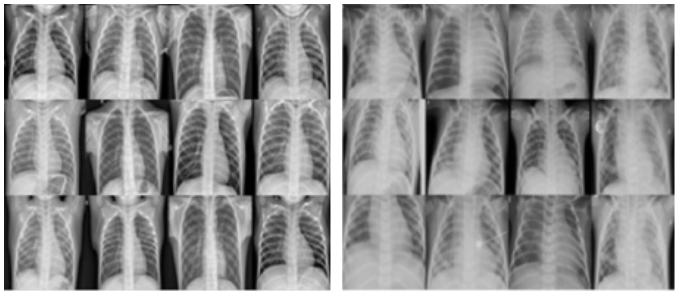


Fig. 1: Proposed PNEXAI Model Top level overview.

our data set using deep learning models (VGG16, VGG19, ResNet50, ResNet101, and Inception V3). Subsequently, the dataset was examined and validated. We selected the top three architectural designs to join the ensemble in the subsequent step [8]. The ensemble model was then trained and evaluated to determine its precision. Explainable AI (XAI) was used to analyze the ensemble model as a last step.

A. Details of the Dataset

Kermany et al. presented a publicly accessible chest Xray data set, which we utilised in our research. The data was acquired at the Women and Children's Medical Center in Guangzhou [9]. It contains 5,842 X-ray images that fall into two distinct categories: normal and pneumonia.



(a) Normal Images (b) Pneumonia affected images

Fig. 2: Sample Data of the Dataset.

B. Data Sample

One of the main symptoms of pneumonia in chest x-ray images is, the alveoli get filled with secretion and appear as a white portion on the chest radiograph.Figure 2(a) shows the normal and 2(b) shows pneumonia affected X-ray images obtained from the dataset.

C. Data Classification

We classified the Train and test images into 8:2 in our study. We have taken 4,263 images as a training set where 3,198 Pneumonia affected images and 1,065 Normal Images. On the other hand, We have taken 1,828 images as a trainingset where 1,279 Pneumonia affected images and 549 Normal Images. Table I represents the distribution of our dataset.

TABLE I: Distribution of our Dataset

Class	Train Set	Test Set	Total
Pneumonia	3,876	390	4,266
Normal	1,342	234	1,576
Total	5,218	624	5,842

D. Data Pre-processing

1) Image Resize: The aim of our model is to detect pneumonia from chest x-ray images with better accuracy. To do so, We have used Chest x-ray images of pneumonia patients of both pneumonia positive and negative images. As VGG, ResNet and Inception models receive 224*224 size images, we have resized our input images into 224*224 shape. For this resizing process we have used some python frameworks such as TensorFlow [12], Scikit Image [13] and Caffe [14]. To convert the image data into pixel values ImageDataGenerator class of Keras has been used.

2) Normalization and Scaling Images: Normalization is the process of reducing data redundancy and removing less important image information. We have used the PCA technique for this normalization. PCA or Principal Component Analysis is a method by which a large data variable is converted into a small data variable with most of the information [15]. It generates Eigen flat fields and merged them to normalize the Chest X-ray image projection. The systematic errors of projection intensity normalization are reduced by using dynamic flat fields [16] [17]. We have done this task by using the Keras ImageDataGenerator class.

3) Data Augmentation: We applied some data augmentation to the images. Generally, it is advised not to makebig modifications to medical image datasets as the images should represent the actual data as closely as possible. As a result, the amount of augmentation was kept as limited as possible. As chest x-ray images are nearly symmetrical from the horizontal view, we applied an x-axis flip on the x-ray images. Furthermore, we varied the brightness of the images just slightly. All the data augmentation was only done on the training dataset so that better training can be achieved. The test set was kept as it was.

E. Model specification

1) VGG16: The VGG16 architecture contains about 16 convolution layers, as the name suggests. The default VGG16 architecture takes an image of shape 224x224x3 as input and provides a volume of 7x7x512 feature slices at the end of the convolutional layers. All the convolution blocks follow a common pattern: multiple stacked convolution layers followed by a max pool layer by the end of it. The originally proposed VGG16 architecture had 2 Fully Connected (FC) layers by the end of the convolution layers. The first FC hidden layer had 4096 FC neurons and the final output FC layer had 1000 FC neurons, each corresponding to one of all 1000 classes it had to classify [18].

2) VGG19: VGG 19 is the upgraded version of the VGG-16. It consists of 16 convolution layers and 3 fully connected layers [19]. This is a deeper CNN with more layers. To reduce the number of parameters in such deep networks, it uses small 3×3 filters in all convolutional layers and is best utilized with its 7.3dataset and can classify images into more than 1000 objects [20]. The features of this deep CNN architecture are, its input size is 224*224*3, the size of convolution kernel is 3*3 with stride size of 1 pixel. For the spatial resolution preservation, spatial padding is used.2*2 pixel windows are used to perform max pooling with stride 2.

3) ResNet50: ResNet [6] (2015) proposed the residual block with bypass layer, which allows the gradient to flow more easily, even with deeper layers. ResNet-50 has 25.5 million parameters across 49 convolution layers and one fullyconnected layer. Each residual block element-wise adds the current feature map with the feature map from the previousresidual block. There is also a bottleneck layer with 1*1convolution that shields a large number of channels for the more expensive 3*3 layer. The pre-trained ResNet 50 Py Torch model achieved a top-1 accuracy of 76of 92.9of the ResNet family balancing computational complexity and prediction accuracy.

4) ResNet101: ResNet101 is a variant of the ResNet model which consists of 101 deep layers. A pretrained model is loaded and trained on Imagenet Dataset. By classifying images from 1000 object categories it has learned high feature representation [21]. This network take input size of 224*224*3.

5) Inception V3: GoogleNet (Inception-V1) [22] (2014) is very parameter-efficient. It has 7 million parameters across 57 convolutional layers and only one fully connected layer. GoogleNet has nine inception modules. Each inception module consists of four branches with 1×1, 3×3, 5×5 convolutions and down-sampling. Two auxiliary loss layers inject loss from the intermediate layers and prevent gradient vanishing. At inference time, the auxiliary layers can be removed.

6) Ensemble Modeling: Ensemble modeling is a process of multilayer diverse base models which are used to predict an outcome either by using many different modeling algorithms or using different training data sets. The aim of this modeling is to reduce the generalization error of a prediction and have the possibility of higher accuracy results than a single classifier. In Ensembled model, multiple models are combined

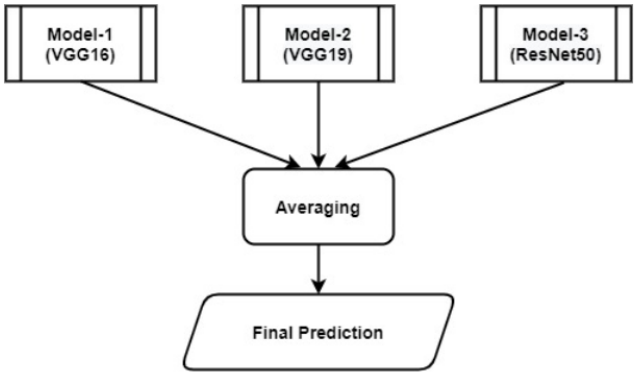


Fig. 3: Output layers of the ensemble model hierarchy for PNEXAI.

and act as a single model. In each model the inputs and outputs remain the same for averaging layer modification [23]

In the proposed model, the output probability of VGG16, VGG19 and ResNet50 will be taken as input for averaging layers and it will generate an average value for two labels , one is Pneumonia positive and another is pneumonia negative. Based on the result of the averaging layer pneumonia patients can be identified from the chest x-ray images. If the probability of the first index is greater than the second index, the patient is pneumonia negative, otherwise it is pneumonia positive. Figure 3 represents the Output layers of the ensemble model hierarchy for PNEXAI. F. Model explainability through Explainable AI (XAI) Explainable Artificial Intelligence is a technique where a more explainable model is generated with a high level of learning performance [24]. By this technique human can easily understand ,trust and manage the emerging generation of intelligent partners. XAI allows supervised care by adapting deep learning techniques to obtain explainable attributes. It also contains procedures for acquiring more hierarchical, generalizable, and explanatory representations, and other pattern inference tools for asserting an understandable paradigm with any black -box testing model. Generally two types of techniques are used for explainable AI systems, these are ante-hoc and post-hoc techniques [25]. Ante-hoc techniques are applied in the AI models from the start of the implementations. Two most used ante-hoc techniques are Reverse Time Attention Model (RETAIN) and Bayesian Deep Learning (BDL). Post-hoc techniques involve explainability during testing stages. The training stages are carried out normally. Post-hoc model analysis is a very common approach towards explaining AI in production.Local interpretable Model-Agnostic Explanations (LIME) and Black Box Explanation through Transparent Approximation (BETA) are two types of Post-hoc techniques. In our research, we used the LIME technique. As a result, model integrity and complexity are traded off in LIME.. LIME is most important for AI systems. To trust the AI system, Models must be explainable to users. AI interpretability reveals what's going on inside these systems and aids in the detection of potential problems including information leakage, model bias, robustness, and causality. LIME provides a generic framework for uncovering black boxes and explaining why AIgenerated predictions or recommendations are made.

IV. Implementation and Result Analysis

A. The Performance Matrices

The confusion matrix is an array that contains correct and incorrect predictions of the algorithm and the actual situation [6]. Elements of confusion Matrix are-

• True Positive (TP): Number of individuals who really have pneumonia as indicated by the model.

• False Negative (FN): Number of individuals who have pneumonia but classified as healthy.

• False Positive (FP): Number of individuals who are actually healthy, however, classified as pneumonia, as per the model.

• True Negative (TN): Number of individuals who are actually healthy and classified as healthy, indicated by the model.

B. Transfer Learning Models

We have achieved the accuracy rate of 97.17by VGG16, 97.69 by VGG19, 97.35by ResNet50, 95.63by ResNet101, and 94.86by Inception V3, respectively

TABLE II: Comparison between the used individual architectures

Architecture	Accuracy	Precision	Recall	f1-score
Inception v3	94.86%	92.30%	95.58%	93.73%
VGG16	**97.17%**	**95.56%**	**97.57%**	**96.49%**
VGG19	**97.69%**	**96.94%**	**97.22%**	**97.07%**
ResNet101	95.63%	93.42%	96.11%	94.63%
ResNet50	**97.35%**	**96.19%**	**97.18%**	**96.67%**

C. PNEXAI (Ensemble of VGG16, VGG19 and ResNet50)

In our "PNEXAI" model, we combined our three best performed architecture which are VGG16, VGG19 and ResNet50 for ensemble modeling. We can see from the illustration of figure 5 that the PNEXAI performs better than the individual models in terms of accuracy, precision, recall and f1-score Our proposed PNEXAI model achieved an accuracy of 98.4698.48shows the confusion matrix of our proposed PNEXAI model in which, 840 of the 853 images affected by pneumonia were categorized as pneumonia, 13 of which were listed as normal, which is false. In contrast,which are VGG16, VGG19 and ResNet50 for ensemble modeling. Our proposed PNEXAI model achieved an accuracy of 98.4698.48from the 315 normal chest pictures, and 5 were classified incorrectly. We can see from the illustration of figure 5 that the PNEXAI performs better than the individual models in terms of accuracy, precision, recall and f1-score.

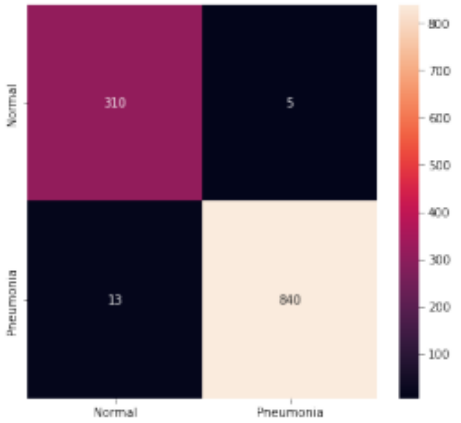


Fig. 4: PNEXAI's Confusion Matrix of

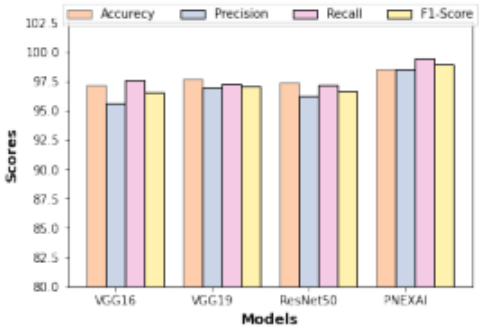


Fig. 5: PNEXAI's Confusion Matrix of

D. Using LIME the Model explainability

After implementing Explainable AI on "PNEXAI", we have noticed that there are 3 types of masksthe Green mask represents the Non infected responding regions and the Red mask represents the infected responding regions represented by 3 different colors Yellow, Red and Green. Here, Yellow borders represent the interpretable regions, the Green mask represents the Non infected responding regions and the Red mask represents the infected responding regions.

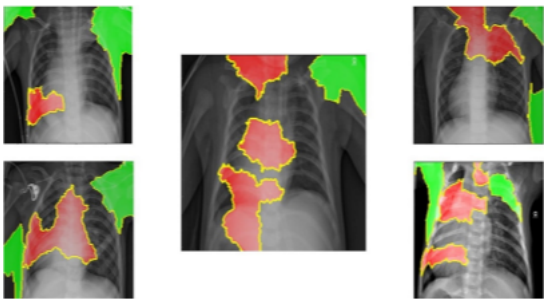


Fig. 6: Output of XAI for Pneumonia cases

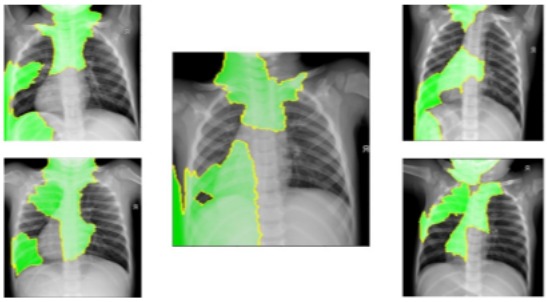


Fig. 7: Output of XAI for Normal cases

V. Future Work and Conclusion

This is the vital to have faster medical monitoring in order to diagnose Pneumonia faster. The major purpose of this study is to create a computer-aided diagnostic system that can aid in the identification of pneumonia and hence prevent unfavorable consequences (such as mortality). In this research, we can see that deep learning models can identify pneumonia very effectively. Chest X-ray images are used in the developed model PNEXAI. First, we trained our dataset using VGG16, VGG19, ResNet50, Resnet101, and Inception V3 which obtained 97.17models (VGG16, VGG19 and ResNet50) and achieved 98.46overall accuracy. For the identification of the affected regions and a better understanding of the classification, Explainable AI (XAI) is applied to the PNEXAI model. We can gather more chest x-ray images in the future to enhance the dataset, which could improve pneumonia detection accuracy and correctly identify the pneumonia-affected regions of the patient. By doing so, an effective method for detecting pneumonia will be discovered.

References

[1] Robert E Black, Simon Cousens, Hope L Johnson, Joy E Lawn, Igor Rudan, Diego G Bassani, Prabhat Jha, Harry Campbell, Christa Fischer Walker, Richard Cibulskis, et al. Global, regional, and national causes of child mortality in 2008: a systematic analysis. The lancet, 375(9730):1969–1987, 2010.

[2] Saraladevi Naicker, Jacob Plange-Rhule, Roger C Tutt, and John B Eastwood. Shortage of healthcare workers in developing countries–Africa. Ethnicity disease, 19(1):60, 2009.

[3] Arata Andrade Saraiva, Nuno M Fonseca Ferreira, Luciano Lopes de Sousa, Nator Junior C Costa, Jose Vigno M Sousa, DBS Santos, ´ Antonio Valente, and Salviano Soares. Classification of images of childhood pneumonia using convolutional neural networks. In BIOIMAGING, pages 112–119, 2019.

[4] Julie Knoll Rajaratnam, Jake R Marcus, Abraham D Flaxman, Haidong Wang, Alison Levin-Rector, Laura Dwyer, Megan Costa, Alan D Lopez, and Christopher JL Murray. Neonatal, postneonatal, childhood, and under-5 mortality for 187 countries, 1970–2010: a systematic analysis of progress towards millennium development goal 4. The Lancet, 375(9730):1988–2008, 2010

[5] Okeke Stephen, Mangal Sain, Uchenna Joseph Maduh, and Do-Un Jeong. An efficient deep learning approach to pneumonia classification in healthcare. Journal of healthcare engineering, 2019, 2019

[6] Achmad Rizal, Risanuri Hidayat, and Hanung Adi Nugroho. Entropy measurement as features extraction in automatic lung sound classification. In 2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC), pages 93–97. IEEE, 2017.

[7] Helena Liz, Manuel Sanchez-Monta ´ n˜es, Alfredo Tagarro, Sara ´ Dom´ınguez-Rodr´ıguez, Ron Dagan, and David Camacho. Ensembles of convolutional neural network models for pediatric pneumonia diagnosis. Future Generation Computer Systems, 122:220–233, 2021.

[8] Hao Ren, Aslan B Wong, Wanmin Lian, Weibin Cheng, Ying Zhang, Jianwei He, Qingfeng Liu, Jiasheng Yang, Chen Jason Zhang, Kaishun Wu, et al. Interpretable pneumonia detection by combining deep learning and explainable models with multisource data. IEEE Access, 9:95872–95883, 2021.

[9] Qinghao Ye, Jun Xia, and Guang Yang. Explainable ai for covid-19 ct classifiers: An initial comparison study. arXiv preprint arXiv:2104.14506, 2021.

[10] Luka Raci ˇ c, Tomo Popovi ´ c, Stevan ´ Sandi, et al. Pneumonia detection ˇ using deep learning based on convolutional neural network. In 2021 25th International Conference on Information Technology (IT), pages 1–4. IEEE, 2021

[11] Rohit Kundu, Ritacheta Das, Zong Woo Geem, Gi-Tae Han, and Ram Sarkar. Pneumonia detection in chest x-ray images using an ensemble of deep learning models. Plos one, 16(9):e0256630, 2021.

[12] Mart´ın Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In 12th USENIX symposium on operating systems design and implementation (OSDI 16), pages 265–283, 2016.

[13] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent ¨ Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. the Journal of machine Learning research, 12:2825–2830, 2011

[14] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM international conference on Multimedia, pages 675–678, 2014.

[15] Zakaria Jaadi. A step-by-step explanation of principal component analysis (pca)

[16] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1-3):37–52, 1987

[17] Herve Abdi and Lynne J Williams. Principal component analysis. ´ Wiley interdisciplinary reviews: computational statistics, 2(4):433–459, 2010.

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

[19] Yufeng Zheng, Clifford Yang, and Alex Merkulov. Breast cancer screening using convolutional neural network and follow-up digital mammography. In Computational Imaging III, volume 10669, page 1066905. International Society for Optics and Photonics, 2018

[20] Jian Xiao, Jia Wang, Shaozhong Cao, and Bilong Li. Application of a novel and improved vgg-19 network in the detection of workers wearing masks. In Journal of Physics: Conference Series, volume 1518, page 012041. IOP Publishing, 2020.

[21] Siyan Tao, Yao Guo, Chuang Zhu, Huang Chen, Yue Zhang, Jie Yang, and Jun Liu. Highly efficient follicular segmentation in thyroid cytopathological whole slide image. In International Workshop on Health Intelligence, pages 149–157. Springer, 2019

[22] Explainable ai - an introduction.

[23] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai—explainable artificial intelligence. Science Robotics, 4(37):eaay7120, 2019.

[24] Ali Yazdizadeh, Zachary Patterson, and Bilal Farooq. Ensemble convolutional neural networks for mode inference in smartphone travel survey. IEEE Transactions on Intelligent Transportation Systems, 21(6):2232–2239, 2019

[25] Ajay Shrestha and Ausif Mahmood. Review of deep learning algorithms and architectures. IEEE Access, 7:53040–53065, 2019.