EVALUATION OF THE CLASSIFICATION OF PNEUMONIA USING MACHINE AND DEEP LEARNING MODELS

Sabrina Leung, Erica Soo

1. INTRODUCTION

Pneumonia is a prevalent infectious disease with considerable morbidity and mortality amongst children [1]. According to WHO, pneumonia makes up for 15% of all deaths of infants under the age of 5 [2]. Pneumonia can be caused by viruses or bacteria and is characterised by inflammation in the alveoli of the lungs which can potentially result in severe consequences [1, 3]. Diagnosis is most commonly done through a physical examination of a chest radiograph by an expert clinician, with early diagnoses leading to a more successful treatment process [1, 3]. However, doing this manually is quite difficult even for trained professionals and is also dependent on the quality of their diagnosing tools [3]. Thus, there is an ongoing need for a smart automated system that is capable of accurately detecting pneumonia from chest X-rays, assisting and simplifying the decision-making processes of clinicians.

Machine Learning (ML) and Deep Learning (DL) models are being used in various fields and have been successfully applied to healthcare issues, allowing for an improvement in clinical predictions [3]. These models present the opportunity to classify pneumonia in a patient through x-ray image analysis. This report aims to implement and evaluate the following two machine learning methods and one deep learning model: Support Vector Machine (SVM), RandomForest and convolutional neural networks (CNNs).

2. RELATED WORKS

An extensive literature review depicts that through recent years there has been significant progress made in developing Machine Learning and Deep Learning models capable of diagnosing pneumonia from x-ray images.

The implementation of ML algorithms for this type of application is highlighted in a study conducted by Singh et al [4]. Through this, the following two machine learning classifiers were explored and compared: Support Vector Machines (SVM), and K-Nearest Neighbour (KNN). An SVM is supervised learning algorithm that is able to plot data in high dimensional space equal to the number of features provided [4]. From this, hyperplanes are utilized, with the distance from the plane indicative of the likelihood a certain feature belonging to a class [5]. This model works well when the classes are well separated, and when there is a higher number of dimensions in space [6]. However, there is a chance of underperformance when the number of features for each data point exceeds the training data samples [6]. KNN is another supervised learning algorithm that utilizes 'feature similarity' in order to predict what the test data should be classified as [4]. It is quite easy and fast to use; however, its inaccuracy and computational time increases as the volume of data increases, thus making it impractical for this application [4]. Despite the disadvantages of these ML algorithms, on a data set of 3000 x-ray images where 7 GLCM texture features were extracted to train the classifiers, an accuracy of 94.6% and 92.6% was achieved for SVM and KNN respectively. These high levels of accuracy for pneumonia detection is not always the case as demonstrated by a study conducted by Sousa et al [7]. Using 156 chest X-ray images with 17 texture features extracted, an accuracy of approximately 50% for both SVM and KNN was achieved [7]. When only optimal features obtained through Recursive Feature Elimination (RFE) were used for training, the accuracy increased to 77% and 70% respectively [7]. This highlights the need for a larger dataset as well as the advantages of picking only the best features to train the classifier. Another classifier that is making rounds in the field of pneumonia diagnosis is Random Forests (RF), another supervised learning method. This method ensembles a community of decision trees, where each tree is grown randomly and then makes a decision by taking into account the class assignment of each tree and then obtaining an average [8]. The ability of this algorithm to use multiple classifiers before making a single decision allows for it to make reliable decisions without overfitting, however a large number of trees can make computational time too slow and ineffective for real-time predictions [8]. Akugundogdu demonstrated that this technique could achieve accuracy of 97.11%, specificity of 99.09% and sensitivity of 91.79% when performed on 5856 X-ray images (1583 normal and 4273 pneumonia), with features extracted by a discrete wavelet transform [8]. Hence, it is evident that these prevalent ML algorithms have a high potential in classifying pneumonia from X-ray images.

Similarly, numerous papers have explored the success of DL models, specifically neural networks, for this type of application. Convolutional neural networks (CNNs) are a specialized subset of neural networks which are built specifically for feature extraction by identifying visual patterns in images with minimal preprocessing [9] It is particularly effective for analyzing large amounts of data from their unprocessed state [10]. CNNs involve a network of layers, where each layer takes the output of each preceding layer as input. Each layer applies a linear or non-linear operation on the image, with the extracted features becoming more refined over each layer [11]. A common method of implementation involves transfer learning, where a pretrained model is used as the starting point for this task. This enables generalized features to be identified without needing to undergo redundant training, which increases computational efficiency. Fine-tuning can then be implemented for the specific data being trained [12]. Wang et al. compared the outcomes of notable CNN architectures on a prominent chest x-ray dataset (ChestX-Ray14), consisting of 1,315 labelled pneumonia images and 6,041 normal images [13]. The four architectures ResNet-50, AlexNet, GoogLeNet, and VGG-16 achieved area under the receiver operating characteristic (AUROC) scores were 0.63, 0.55, 0.60, and 0.51 respectively, where 1 indicates perfect classification. These scores indicate that deeper architectures such as ResNet-50 have greater success compared to the shallower VGG-16 model. Guendal et al. [14] studied the success of DenseNet-121 (with 121 layers), another deep architecture, which achieved an AUROC score of 0.76 on the same ChestX-Ray14 dataset. While it may be deduced that deeper models are more successful across all datasets, it has been suggested that the success of the model is influenced by the dataset, and specifically the size and split of the data. This has been confirmed by Rajamaran et al. who implemented a VGG-16 model on a different dataset and achieved a classification accuracy of 93.6%, and sensitivity and specificity of 96.2% [15]. Similar results have also been achieved by Eid et al. [16] in which a ResNet-50 model achieved an accuracy of 96.3%. Thus, while there are numerous CNN architectures available for use, their success dependent on a range of factors outside of the specific layers being implemented. The primary benefit of deep learning is the fact that computational performance increases as the scale of data increases, with a model consistently learning as more information is provided. However, this process often takes longer to train, though it is reduced through transfer learning. Particularly for large-scale datasets, this is far more beneficial than traditional machine learning models which are restricted by the parameters set during training. Therefore, CNNs pose great potential in effectively classifying pneumonia.

3. MACHINE LEARNING – SABRINA

As established from the literature review, it is clear that machine learning algorithms are prominent in the field of biomedical image classification. Thus, for the task at hand of classifying pneumonia from x-ray images, two of the most prevalent methods will be implemented: Support Vector Machines and Random Forest.

3.1. Support Vector Machines

As mentioned previously, a support vector machine (SVM) is a supervised machine learning algorithm where each data is plotted as a point in n-dimensional space (n = number of features) [4]. In order to perform a classification, a hyper-plane/s which are essentially the optimal boundaries providing the largest margin between the classes is found and implemented [4].

3.2. Random Forest

Random Forest is an algorithm based on randomly grown decision trees. This ensemble method decides by averaging the class possibilities obtained from each of the trees [8]. Thus, when new data is given to the algorithm, it is evaluated by all the trees which, each of which provides an independent class prediction [8]. The class with the most votes will be the final classification.

3.3. Methodology

In this section, the implementation of these two algorithms will be contextualized and described. It will be divided into the following sections: Datasets, Preprocessing, Features Extraction, Training, Classification and Evaluation .

a) Datasets:

The dataset used is collected by Kermany et al. and contains 5856 chest X-rays images acquired from children [17]. These X-rays had children who were either normal, had bacterial pneumonia or viral pneumonia. These were then broken down into 3 versions – full, partial and small. For this implementation, the smallest dataset (1951 images) was chosen to optimize computational time. It should be noted the dataset was already split into training and test images for convenience, with a 70/30 split respectively.

b) Pre-processing:

The classifier first extracts the labels from the images, providing a binary classification, with normal X-rays classed as 0 and X-rays with pneumonia classed as 1. This was done as SVM and RF are one of the most common binary classification algorithms [18]. To test the capability of the two algorithms in detecting the type of pneumonia in the X-rays, another dataset was created where the pneumonia class was divided into two – bacterial and viral (multi-class

dataset). Additionally, once the classes for every image in the dataset was obtained, the images were then resized (256, 256). This was done as the raw images all have different image sizes from one another, and ML algorithms require that all the input data to be the same size [3]. Furthermore, resizing allows for the reduction of unnecessary information in the background, allowing for better feature extraction and lower computational costs [1, 3]. Alongside resizing, the datasets (binary and multi-class) were also shuffled to avoid any bias/patterns being formed prior to training the ML models.

c) Features Extraction:

Next, the following texture features were extracted from all the images: contrast, dissimilarity, homogeneity, energy, correlation and ASM at 4 different angles (0, 45, 90, 135º). As depicted in the literature review, both Sigh et al. and Sousa et al. utilized texture features in their classifiers, indicating its potential in allowing for the differentiation between classes. Additionally, the following statistical features were also extracted: entropy, mean, minimum, maximum, variance, skewness and kurtosis. This was done as it was demonstrated by Masud et al. to have an 82.13% accuracy when used on ensemble learning algorithms such as Random Forest [19]. By combining these feature sets, it can be inferred that a higher accuracy may be achieved on the classifiers as there are more features for the algorithms to learn from. However, it should be noted that sometimes more irrelevant input features can result in a worse performance by the algorithm. Hence, a Recursive Feature Elimination (RFE), a feature selection algorithm was implemented. This technique involves the use of another machine learning algorithm, in this case Logistic Regression, to select the most relevant features from the main feature set. The number of features selected was manually set, with 7 features deemed as optimal. This technique was shown by Sousa et al. and Luo et al. to improve the performance of their classifiers as it allowed for less space and computational time, and was overall more effective [20].

d) Training, Classification and Evaluation:

The training data (70%) with the extracted labels and optimal features for both datasets (binary and multi-class) were then used to train both SVM and Random Forest classifiers. Once trained, the testing data (30%) were then used on the models to gain a prediction of whether the input image had pneumonia or not. Mediocre accuracy values were obtained for both ML algorithms on both datasets (binary and multi-class), with cross validation displaying similar results. To optimize the classifiers, the hyperparameters were tuned. All possible combinations of various parameter values were evaluated through a 3-fold cross validation, with the best combination retained. The models were then retrained using the optimized parameters show in Tables 1 and 2, and tested on the testing data. Confusion matrices were then plotted, and evaluation metrics was used to quantify the performance of the predictive models. These will be discussed further in Section 5.

Table 1: Optimized hyperparameters for SVM for Binary and Multi-Class Dataset

Hyperparameters	Binary Dataset	Multi-Class Dataset
C	0.1	100
Degree	2	1
Gamma	1	0.1
Kernel	Poly	Rbf
Tolerance	0.01	0.001

Table 2: Optimized hyperparameters for Random Forest Binary and Multi-Class Dataset

Hyperparameters	Binary Dataset	Multi-Class Dataset
Bootstrap	True	True
Max. depth	60	60
Max. features	sqrt	sqrt
Min. samples leaf	4	1
Min. samples split	8	10
N estimators	200	200

4. DEEP LEARNING – ERICA

From the literature, it is evident that CNNs have great potential in radiology with improved efficiency and ability to obtain higher training and validation accuracy [12]. This CNN implementation applies transfer learning with both ResNet50 and VGG-16 architectures, two of the most successful networks identified in the literature. ResNet50 architecture is a model which is based around 'shortcut connections' which avoids the vanishing gradient that occurs when a network becomes too deep, meaning the model can be trained on up to 3000 layers [21]. In comparison, VGG-16 is a far shallower model which only uses 3x3 convolutions, demonstrating the potential simplicity of a classification model. This model reduces the number of trainable variables to encourage faster learning and more robustness to overfitting [22]. Transfer learning was chosen as it allows the model to learn faster by starting with a higher accuracy. As full training is such a computationally expensive task and because the dataset was relatively small (which is often the case in medical imaging), implementing a pre-trained model requires less computational effort, data, and training time. The motivation of fine-tuning was based on the observation that while shallower layers are often more generic, such as edges, deeper layers can be trained for a specific purpose [11].

4.1. Methodology

The CNNs were implemented as a binary classifier using the Tensorflow Keras libraries on the small dataset collected by

Kermany et al., consisting of 2,762 pneumonia chest x-rays and 1,140 normal chest x-rays [17].

a) Pre-processing:

Extensive pre-processing was not required as CNNs have the benefit of being able to receive raw image data as input [12]. The classifier first loads the train and test directories for both normal and pneumonia classes. Both the normal and pneumonia classes in the train directory were split in an 80/20 ratio to create separate training and validation subsets. Once these subsets were defined, images were resized to a standardised size of (224, 224), the recommended dimensions for the both the ResNet and VGG models [12]. This also ensures all images were given the same weighting, and reduce background noise [3]. Augmentation of each image also involved normalising each pixel by dividing each value by 255.

b) Model

The CNN models were then implemented. The fine-tuning and parameters were based on a seminal study by He et al. [21] which explores the optimisation of a residual network (ResNet) in comparison with other architectures including VGG. After the pretrained ResNet-50 model was implemented, the first convolutional layer was removed, and fine-tuning was implemented for modification. This involved implementing a global average pooling layer for the dimensional reduction of data to prevent overfitting. The output was then flattened for two full-connected layers, one with a ReLu activation and the other with a softmax activation. These decisions were based on the work of Xu et al. which achieved an accuracy of 86.7% [23]. The VGG model was constructed separately, using the same fine-tuning architecture. This method replicated a study by Asnaoui et al. in which a single fine-tuning method was proposed for numerous CNN architectures including ResNet and VGG, resulting in high accuracies for both ResNet50 (96.61%) and VGG16 (85.94%).

c) Training

Before training, the layers from the original base model of both models the ResNet and VGG models were frozen, as retraining would be redundant [24]. Hence, only the fine tuning layers were trained, conserving time and computational efficiency. The models were then trained on the training dataset. This training used 10 epochs, a batch size of 32, and learning rate of 0.001. These values were chosen as they have proven successful values in a study by Hashmi et al., in which pneumonia classification using a DenseNet model achieved an accuracy of 98.43% [12]. They were deemed an appropriate trade-off between achieving an accurately trained model while maintaining some computational efficiency.

d) Classification and Evaluation

The ResNet and VGG models were then evaluated on the validation and test datasets using the Model.evaluate() method. This produced loss and accuracy values for the datasets. To visualise the improvement of the model during training, the accuracy and loss were plotted against each epoch for both the train and validation data. A confusion matrix was then plotted, and values for precision, recall, and f1-score were calculated to evaluate the performance of the classification models.

5. EVALUATION

Observing the confusion matrices shown in Figure 1, it is evident that both SVM and Random Forest were quite good at identifying x-ray images with pneumonia (class = 1) in the binary dataset, with approximately 380 and 370 cases respectively out of 390, identified correctly. However, it is also noticeable that when it came to normal images (class = 0), the algorithms predicted only half correctly, while the other half was classed as pneumonia. In terms of the multi-class dataset, it is clear that both machine learning classifiers behaved quite similarly, identifying normal (class = 0) and bacterial pneumonia (class =1) relatively well. On the other hand, viral pneumonia (class =2) was misidentified half the time as bacterial pneumonia, thus highlighting the limitations of these algorithms in multi-class classification. The outcomes of the CNN classifiers showed similar success true positive cases, identifying pneumonia, with ResNet correctly identifying 334 cases and VGG correctly identifying 274 cases. This said, both models also produced a high number of false positives, incorrectly classifying normal images as pneumonia 208 and 171 times respectively, of a total of 234 normal cases. From these values, it is evident that both CNN models have significant bias towards pneumonia. While this model is not perfect, the bias towards pneumonia is preferrable as false negative outcomes have a far greater cost [25]. That is, if a patient was to have pneumonia but was classified as normal, this may lead to ineffective treatments which can have detrimental health consequences.

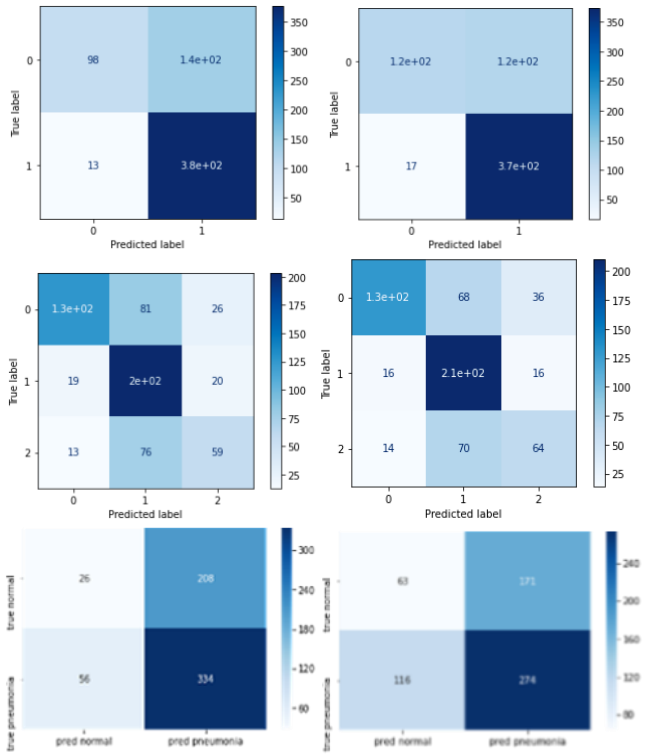

Figure 1: Confusion Matrices (going across): a) SVM classifier on the binary dataset, b) Random Forest classifier on the binary dataset, c) SVM classifier on the multi-class dataset, d) Random Forest multi-class dataset, e) CNN (ResNet) Classifier, f) CNN (VGG16) Classifier

To evaluate the performance of the classifiers more quantitively, the following metrics was used:

Accuracy – This represents the portion of all true predicted classes that are actually correct i.e. how well did the classifier predict the pneumonia and normal cases.

$$\frac{TP + TN}{Total\ Test\ Sample}$$

Precision – This represents the proportion of positives detected that are truly positive. In this case, it demonstrates how many pneumonia cases were actually predicted correctly.

$$\frac{TP}{TP + FP}$$

Recall – This represents the proportion of actual positives detected, highlighting if all the pneumonia cases were predicted.

$$\frac{TP}{TP + FN}$$

F1-Score – This is a balanced average between precision and recall, illustrating how robust the models are.

$$\frac{2TP}{2TP + FP + FN}$$

These results for each classifier are summarized in Table 3.

Table 3: Evaluation metrics for the Classifiers

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SVM (Binary)	76	81	69	70
Random Forest (Binary)	78	81	72	74
SVM (Multi-Class)	62	64	59	60
Random Forest (Multi-Class)	65	66	62	62
CNN (ResNet)	75	62	86	72
CNN (VGG16)	54	61	70	64

Out of the machine learning algorithms, it is clear from Table 3 that Random Forest has the highest scores across the board for both the binary and multi-class classification, achieving an accuracy of 78% and 65% respectively. Although this is the case, it is worth noting that SVM follows closely behind with an accuracy of 76% and 62%. This is lower than expected when compared to literature where Singh et al. demonstrated an accuracy of 94.6% for SVM and similarly Akgundogdu depicted an accuracy of 97.11% for Random Forest [4, 8]. Of the CNN models, the ResNet performed far better than VGG across all metrics. ResNet achieved an accuracy of 75%, slightly lower than the 98.13% accuracy achieved by Eid et al. [16]. In comparison, VGG achieved an accuracy of 54%, far lower than expected in comparison with Rajamaran et al., who achieved an accuracy of 96.3% for classification of pneumonia using VGG [15]. When comparing between ML algorithm and CNNs, it is evident that there is no significant difference across the metrics used for evaluation. This is surprising as deep learning models have shown to outperform ML classifiers in most applications, as well diagnosing pneumonia from chest x-rays [26]. This was demonstrated in a study conducted by Mamlook et al. which illustrated that the CNN implemented had an accuracy of 98.46%, whilst Random Forest had an accuracy of 97.61% [3].

The reason behind these low accuracy, precision and recall scores could be due to the small dataset used. Whilst it reduced computational time and data storage, most of the classifiers utilized in literature were trained on extremely large datasets. This is because a small dataset may trigger over-fitting, thus resulting in a low performance [27]. In the future, these implementations of ML algorithms and CNNs classifiers can be done on the full dataset containing 5856 images. Another potential reason for the poor performance of the models compared to literature could be due to the fact that a bias was formed towards the majority class. In the dataset that was utilized, there was 336 normal chest x-rays, 649 with bacterial pneumonia and 342 with viral pneumonia. Hence, a bias would have been formed towards pneumonia images, particularly bacterial pneumonia images, thus affecting the training of the classifiers. To avoid this in the future as well as increase the robustness of the models, Masud et al. suggests the use of data augmentation [19]. This involves applying image transformations such as cropping, scaling, rotation, and reflection to the images in order to increase the data set sizes. Thus, in the future, this can be done for the normal chest x-rays and those with viral pneumonia in order to balance the training dataset as well as increase the dataset in general to avoid over-fitting.

It is evident that machine learning and deep learning both have the potential for great success in the classification of pneumonia. While both methods have been effective for this purpose, they differ in many respects. While machine learning algorithms can be trained on a small amount of data, deep learning methods require large amounts of labelled data to achieve greater accuracy [28]. Hence, it is possible that the size of chosen dataset was insufficient for training the CNN models in this study. Additionally, machine learning models require features and parameters to be more explicitly identified in a fragmented approach, whereas deep learning generates novel features autonomously using end-to-end principles [28]. These different requirements of each learning model explain the discrepancies between the results and those found in the literature.

6. CONCLUSION

In this work, we have implemented multiple models for the classification of pneumonia from chest x-ray images using machine learning methods. The methods implemented were support vector machines and random forest algorithms (both binary and multi-class classifications), and Convolutional Neural Networks (using ResNet and VGG architectures). Through a comparison of each model based on the evaluation metrics of accuracy, precision, recall, and f1-score, it was evident that no model achieved significantly greater performance than the others. These results were surprising, as there is a general consensus that deep learning models tend to outperform more traditional models in the context of image classification, while in this study the RF binary classifier achieved the highest accuracy of 78%.

Future work in this area may therefore consider use of a larger and augmented dataset to predict better results in the deep learning models. Other investigations may also explore further optimisation of the models, including the hyperparameters used in the SVMs and RFs, and the fine-tuning of the CNN models.

7. REFERENCES

[1] J. V. S. d. Chagas, D. de A. Rodrigues, R. F. Ivo, M. M. Hassan, V. H. C. de Albuquerque, and P. P. R. Filho, "A new approach for the detection of pneumonia in children using CXR images based on an real-time IoT system," *Journal of Real-Time Image Processing*, 2021/03/16, 2021.

[2] Who.int. "Pneumonia," https://www.who.int/news-room/fact-sheets/detail/pneumonia.

[3] R. E. A. Mamlook, S. Chen, and H. F. Bzizi, "Investigation of the performance of Machine Learning Classifiers for Pneumonia Detection in Chest X-ray Images." pp. 098-104.

[4] N. Singh, R. Sharma, and A. Kukker, "Wavelet Transform Based Pneumonia Classification of Chest X- Ray Images." pp. 540-545.

[5] J. Yao, A. Dwyer, R. M. Summers, and D. J. Mollura, "Computer-aided diagnosis of pulmonary infections using texture analysis and support vector machine classification," *Acad Radiol*, vol. 18, no. 3, pp. 306-14, Mar, 2011.

[6] S. Karamizadeh, S. M. Abdullah, M. Halimi, J. Shayan, and M. j. Rajabi, "Advantage and drawback of support vector machine functionality." pp. 63-65.

[7] R. T. Sousa, O. Marques, F. A. A. M. N. Soares, I. I. G. Sene, L. L. G. de Oliveira, and E. S. Spoto, "Comparative Performance Analysis of Machine Learning Classifiers in Detection of Childhood Pneumonia Using Chest Radiographs," *Procedia Computer Science*, vol. 18, pp. 2579-2582, 2013/01/01/, 2013.

[8] A. Akgundogdu, "Detection of pneumonia in chest X-ray images by using 2D discrete wavelet feature extraction with random forest," *International Journal of Imaging Systems and Technology*, vol. 31, no. 1, pp. 82-93, 2021.

[9] N. M. Elshennawy, and D. M. Ibrahim, "Deep-Pneumonia Framework Using Deep Learning Models Based on Chest X-Ray Images," *Diagnostics (Basel, Switzerland)*, vol. 10, no. 9, pp. 649, 2020.

[10] Z. Yue, L. Ma, and R. Zhang, “Comparison and Validation of Deep Learning Models for the Diagnosis of Pneumonia,” *Computational Intelligence and Neuroscience*, vol. 2020, pp. 1-8, 09/18, 2020.

[11] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, “Convolutional neural networks: an overview and application in radiology,” *Insights into Imaging*, vol. 9, no. 4, pp. 611-629, 2018/08/01, 2018.

[12] M. F. Hashmi, S. Katiyar, A. G. Keskar, N. D. Bokde, and Z. W. Geem, “Efficient Pneumonia Detection in Chest Xray Images Using Deep Transfer Learning,” *Diagnostics (Basel, Switzerland)*, vol. 10, no. 6, pp. 417, 2020.

[13] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases." pp. 3462-3471.

[14] S. Guendel, S. Grbic, B. Georgescu, S. K. Zhou, L. Ritschl, A. Meier, and D. Comaniciu, “Learning to recognize Abnormalities in Chest X-Rays with Location-Aware Dense Networks,” 03/12, 2018.

[15] S. Rajaraman, S. Candemir, I. Kim, G. Thoma, and S. Antani, “Visualization and Interpretation of Convolutional Neural Network Predictions in Detecting Pneumonia in Pediatric Chest Radiographs,” *Applied Sciences*, vol. 8, pp. 1715, 09/20, 2018.

[16] M. Eid, and Y. Elawady, “Efficient Pneumonia Detection for Chest Radiography Using ResNet-Based SVM,” *European Journal of Electrical Engineering and Computer Science*, vol. 5, pp. 1-8, 01/14, 2021.

[17] D. Z. Kermany, Kang; Goldbaum, Michael, "Large Dataset of Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images," M. Data, ed., 2018.

[18] R. Kumari, and S. Srivastava, “Machine Learning: A Review on Binary Classification,” *International Journal of Computer Applications*, vol. 160, pp. 11-15, 02/15, 2017.

[19] M. Masud, A. Bairagi, A. Nahid, N. Sikder, S. Rubaiee, A. Ahmed, and D. Anand, “A Pneumonia Diagnosis Scheme Based on Hybrid Features Extracted from Chest Radiographs Using an Ensemble Learning Algorithm,” *Journal of Healthcare Engineering*, vol. 2021, pp. 1-11, 02/25, 2021.

[20] Y. Luo, Z. Tang, X. Hu, S. Lu, B. Miao, S. Hong, H. Bai, C. Sun, J. Qiu, H. Liang, and N. Na, “Machine learning for the prediction of severe pneumonia during posttransplant hospitalization in recipients of a deceased-donor kidney transplant,” *Ann Transl Med*, vol. 8, no. 4, pp. 82, Feb, 2020.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition." pp. 770-778.

[22] K. Simonyan, and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv 1409.1556*, 09/04, 2014.

[23] X. Xu, X. Jiang, C. Ma, P. Du, X. Li, S. Lv, L. Yu, Q. Ni, Y. Chen, J. Su, G. Lang, Y. Li, H. Zhao, J. Liu, K. Xu, L. Ruan, J. Sheng, Y. Qiu, W. Wu, T. Liang, and L. Li, “A Deep Learning System to Screen Novel Coronavirus Disease 2019 Pneumonia,” *Engineering*, vol. 6, no. 10, pp. 1122-1129, 2020/10/01/, 2020.

[24] S. Misra, S. Jeon, S. Lee, R. Managuli, I.-S. Jang, and C. Kim, “Multi-Channel Transfer Learning of Chest X-ray Images for Screening of COVID-19,” *Electronics*, vol. 9, pp. 1388, 08/27, 2020.

[25] D. van Ravenzwaaij, and J. P. A. Ioannidis, “True and false positive rates for different criteria of evaluating statistical evidence from clinical trials,” *BMC Medical Research Methodology*, vol. 19, no. 1, pp. 218, 2019/11/27, 2019.

[26] S. S. Yadav, and S. M. Jadhav, “Deep convolutional neural network based medical image classification for disease diagnosis,” *Journal of Big Data*, vol. 6, no. 1, pp. 113, 2019/12/17, 2019.

[27] A. Althnian, D. AlSaeed, H. Al-Baity, A. Samha, A. B. Dris, N. Alzakari, A. Abou Elwafa, and H. Kurdi, “Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain,” *Applied Sciences*, vol. 11, no. 2, pp. 796, 2021.

[28] H. Mohammad-Rahimi, M. Nadimi, A. Ghalyanchi-Langeroudi, M. Taheri, and S. Ghafouri-Fard, “Application of Machine Learning in Diagnosis of COVID-19 Through X-Ray and CT Images: A Scoping Review,” *Frontiers in cardiovascular medicine*, vol. 8, pp. 638011-638011, 2021.