AUTOMATED PNEUMONIA DIAGNOSIS THROUGH CHEST X-ray IMAGE ANALYSIS

Tshepo Nkambule
School of Computer science and Applied mathematics
University of the Witwatersrand
Johanesburg, South Africa
1611821@students.wits.ac.za

***Abstract*—In this paper we present an automated pneumonia diagnosis system by using the following supervised learning methods kNN, logistic regression and the multi layer perceptron. The presented dataset is of high dimensionality and principal component analysis is employed in order to reduce dimensionality while keeping features that are highly discriminative, The multilayer perceptron achieves a peak performance in accuracy of 86.31%.**

I. Introduction

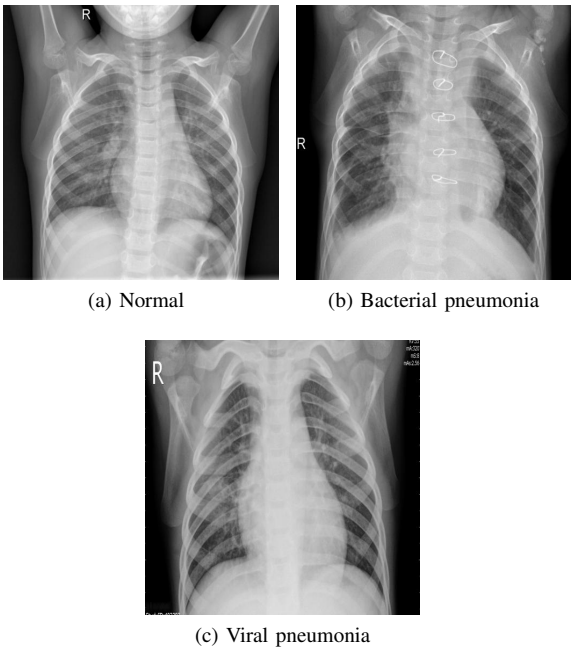
Medical imaging refers to a set of techniques that aim to produce visual representations of internal aspects of the human body in a noninvasive manner. Medical imaging seeks to reveal the internal organs masked by the skin and bones so that they can be analysed for any abnormalities and diagnose and treat disease.

Pneumonia is inflammation of the tissues in one or both of the lungs typically caused by a bacterial infection [1]. Although pneumonia can be a manageable disease through a series of drugs such as antibiotics and antivirals it is crucial that an early diagnosis is made [2]. The globally accepted medical imaging technique used for diagnosing pneumonia is X-ray imaging, in which chest X-ray images are analysed for any abnormalities [3]. Diagnosing pneumonia from chest X-ray images is a tedious and challenging task even for expert radiologist. This is because the appearance of pneumonia in X-ray images is often unclear, meaning it can be confused with other respiratory diseases and benign abnormalities. The above challenges lead to inconsistencies amongst different radiologist in the diagnosis of pneumonia [4].

The aforementioned challenges faced by radiologists in diagnosing pneumonia present a pressing need for an automated system to aid radiologists in analysing chest X-ray images for diagnosing pneumonia. Thus the purpose of this research project is to develop an automated pneumonia diagnosis system that indicates whether a given chest X-ray image represents a patient that has pneumonia or not. The following supervised learning methods will be used to develop such a system k − Nearest Neighbours (kNN), logistic regression (LR) and multilayer perceptron (MP).

Identify applicable funding agency here. If none, delete this.

The research task can be formally cast as *Given a chest X-ray input image, classify the image as a normal chest X-ray image or a pneumonia chest X-ray image.*. Figure 1 below shows an example of a normal chest X-ray image along with examples of pneumonia chest X-rays.



(a) Normal (b) Bacterial pneumonia

(c) Viral pneumonia

Fig. 1: Normal chest X-ray and Pneumonia chest X-rays

This paper is structured as follows: Section II will present an overview of the background and related work, Section III will outline the research methodology, Section IV will present experiments and results and Section V will discuss the results.

II. BACKGROUND AND RELATED WORK

A. Introduction

The previous section presented the problem of diagnosing pneumonia by analysing a chest X-ray image of a particular patient. This section will present related work by viewing

what other authors have done to accomplish similar tasks. The outline of this section is as follows: Section II-B will present a brief description on the supervised learning machine learning paradigm, Section II-C will present what image features and feature extraction techniques have been used with medical images, Section II-D will discuss dimensionality reduction and feature selection and Section II-E will highlight some of the machine learning models used for medical imaging diagnosis.

B. Supervised learning

The field of machine learning consist of three main paradigms supervised learning, unsupervised learning and reinforcement learning. To give a definition for supervised learning we consider a set $\boldsymbol{\Xi}$ in which every member of $\boldsymbol{\Xi}$ represents a tuple of the form $(\boldsymbol{x}, y)$. The members of $\boldsymbol{\Xi}$ obey some mapping $f(\boldsymbol{x}) = y$. Supervised learning is a method of learning/estimating the function $f(\boldsymbol{x})$ as some hypothesis function $h(\boldsymbol{x})$ by using m samples from $\boldsymbol{\Xi}$. If we can find a hypothesis function $h(\boldsymbol{x})$ that agrees with $f(\boldsymbol{x})$ for the members of $\boldsymbol{\Xi}$ then $h(\boldsymbol{x})$ will be a good guess for $f(\boldsymbol{x})$. An example of supervised learning is curve fitting [5].

C. Image features

As shown in figure 1, X-ray images are always greyscale images which means they have no colour component. The most intuitive features that can be extracted from an image are the greyscale pixel intensity values. In the case of a greyscale image the intensity values represent how dark or bright a pixel at a given coordinate (x, y) in the image is. Several image processing techniques can be employed for feature extraction in images. [6] developed a lung cancer detection system based on X-ray images in which features are extracted as follows: image enhancement (histogram equalization), lung region extraction, lung segmentation and edge detection. [7] compares wavelet decomposition and curvelet decomposition as feature extraction mechanisms on X-ray images for lung cancer diagnosis.

While it is clear that the greyscale pixel intensity values of grey scale image are used as features, the $2d$ array representation of digital images is not convenient for feature representation. A convenient way to represent features would be to stack all the pixels next to each other by concatenating all the rows of the 2d array to form a long $1d$ array. This means that for an input greyscale image of size 28×28 the resulting feature vector would have dimemsions 1×784.

D. Dimensionality reduction and feature extraction

As mentioned above that greyscale digital images are often transformed from their $2d$ array representation to a $1d$ array representation, with a 28×28 greyscal image being a 1×784 feature vector. The immediate caveat of the $1d$ array feature representation is that the image feature space is of high dimensionality even for low resolution images. If the dimensionality of the feature space is too high any subsequent machine learning methods that use the feature will be computationally inefficient [8].

Dimensionality reduction aims to reduce the dimensionality of the feature space by discarding features that account for little variance in a dataset and keeping those that account for significant amount of variance in the dataset. In [8] principal component analysis (PCA) is proposed for feature extraction as it has been widely used in other disiplines such as image processing, pattern recognition, data compression and computer vision. The quality of a principal component p_j is the amount of variance it accounts for given by its associated eigen value λ_j

$$\pi_j = \frac{\lambda_j}{\sum_{j=1}^{n} \lambda_j} \tag{1}$$

In PCA the number of features kept is decided by the desired total variance explained which can be expressed as a percentage of the cumulative sum of (1), a cutoff total variance of 70% is common [9]. In [10] PCA was used to reduce the dimensionality of the mnist digits dataset from 784 to 281 in order to efficiently perform digit recognition, the reatained features accounted for total variance of over 99%.

E. Supervised learning models

Supervised learning algorithms that have been previously used for medical imaging diagnosis. In [7] and [11] support vector machines (SVM) were used for lung cancer diagnosis and brain abnormally detection from MR image respectively. [6] employed a neural network for lung cancer detection using chest X-ray images.

III. Research Methodology

This section will outline the aims of this research project and introduce the research hypothesis and specify the methodology that will be followed towards testing the research hypothesis and achieving desired aims.

A. Research hypothesis

As discussed in the previous sections there exist a pressing need for an automated system to aid radiologists in analysing chest X-ray images for diagnosing pneumonia. The purpose of this research is to classify chest X-ray images as either normal indicated by a label of 0 or as pneumonia indicated by a label of 1 using the following supervised learning models.

1) k-Nearest Neighbours (kNN) - A lazy algorithm that does not learn a discriminative function from training set but memorizes the trainining set and queries it when a new data point is to be classified. A new data point is assigned the modal label from its k nearest neighbours using some distance metric.

2) Logistic regression (LR) - A statistical model that uses the logistic function in order to model a binary dependent variable. A new data point is classified by

evaluating the logistic function with relevant parameters at that point, the output of the function is the probability that the data point belongs to a particular class.

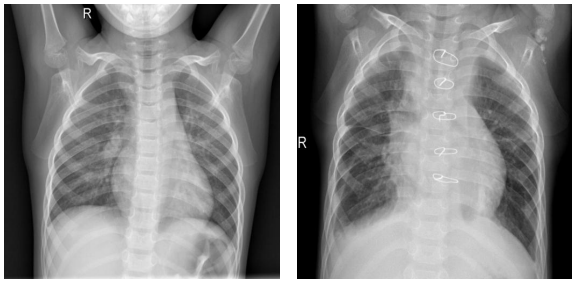
3) Multilayer perceptron (MLP) - An artificial feedforwad neural network, which is an extension of the perceptron and has at least three layers (minimum of one hidden layer).

The research hypothesis can then be stated as follows

Research Hypothesis: Normal chest X-ray images and pneumonia chest X-ray images can be differentiated using the aforementioned supervised learning techniques. The multilayer perceptron model will outperform all other models.

B. Methodology

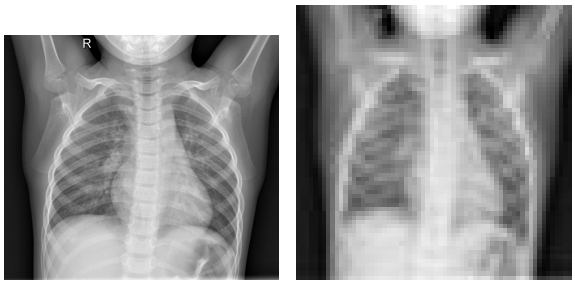
*1) **Data collection**:* In order to perform the desired task a suitable dataset is required. The dataset used in this project was adapted from kaggle.com. There are 5 856 X-Ray images (JPEG) and 2 categories (Pneumonia/Normal). There are 1 583 images that belong to the normal class and 4 273 images that belong to the pneumonia class. The images are of varying sizes with some having high resolution such 2090×1858 while others are 712×439. More details about the dataset and its integrity are covered by [12]. An example of images from the dataset is shown in figure 2 below.



(a) Normal (b) Pneumonia

Fig. 2: Normal chest X-ray and Pneumonia chest X-ray

*2) **Data preparation**:* Before the data can be used all the input images must be resized to be of the same dimensions. Due to computational cost and memory limitation all images in the dataset were resized to be 64×64 using a gaussian filter with anti-liasing (see scikit-image resize).



(a) Original 2090×1858 (b) Resized image (64×64)

Fig. 3: Image resizing

Once the images have been resized the dataset is split in an approximatley $70 : 30$ ratio of training data to testing data. After the split the training set consists of 4 172 images while the testing set consists of 1 684 images.

*3) **Dimensionality reduction and feature extraction**:* The resized images have dimensions 64×64, which means a single image corresponds to a feature vector of dimensions $1 \times 4\ 096$. We observe that the dimensionality of the feature space is too high and will make the models computationally inefficient. In order to perform dimensionality reduction we use principal component analysis with the requirement that the total retained variance must be at least 90%. A visualization of the total variance retained with a different number of principal components is shown in figure 4 below.

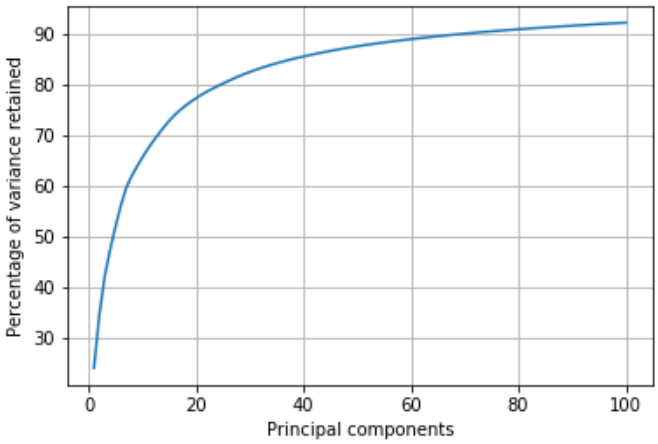


Fig. 4: Variance retained with first 100 principal components

The first 69 principal components were able to retain 90% variance which means that the remaining 427 principal components could be dropped without losing significant amounts of variation. The feature vector now has the following dimensions 1×69. Once the dataset had been projected into a lower dimensional space it was the normalized using

$$\boldsymbol{X}_j = \frac{\boldsymbol{X}_j - \mu_j}{\sigma_j} \quad j = 1, 2, 3, ..., 69 \tag{2}$$

At this stage the dataset is ready for application with various supervised learning methods.

*4) **Implementation of models**:* In the models presented here 2 of them, kNN and logistic regression were implemented using the *numpy* library while the multilayer perceptron is an off the shelve classifier adapted from the *scikit-learn* library.

- kNN the chosen distance metric for this algorithm was the euclidean distance, which for a query data point $\boldsymbol{q}$ the distance to any other point in the training set $\boldsymbol{X}$ can be computed as

$$\boldsymbol{D(q, X)} = \sqrt{(\boldsymbol{X} - \boldsymbol{q})^T(\boldsymbol{X} - \boldsymbol{q})} \tag{3}$$

  In order to determine the optimal value for k a parameter sweep can be done in a suitable search space.

- Logistic regression will use the sigmoid function along with a parameter set $\boldsymbol{\Theta} = [\theta_0, \theta_1, ..., \theta_{69}]$ in order to determine the probability that a query data point $\boldsymbol{x}$ belongs to a particular class and use this value as a threshold for classification.

$$h_{\boldsymbol{\Theta}}(\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{\Theta}^T \boldsymbol{x}}} \tag{4}$$

$$\hat{y} = \begin{cases} 1 & \text{if } h_{\boldsymbol{\Theta}}(\boldsymbol{x}) > 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

  The cost function $J(\boldsymbol{\Theta})$for this model was chosen to be the cross entropy cost function which takes the form

$$Q_{\boldsymbol{\Theta}}(\boldsymbol{x}) = -ylog(h_{\boldsymbol{\Theta}}(\boldsymbol{x})) - (1 - y)log(1 - h_{\boldsymbol{\Theta}}(\boldsymbol{x})) \tag{6}$$

$$J(\boldsymbol{\Theta}) = \frac{1}{N} \sum_{n=1}^{N} Q_{\boldsymbol{\Theta}}(\boldsymbol{x}^{(n)}) \tag{7}$$

  The learning algorithms that were used to learn paramaters $\boldsymbol{\Theta} = argmin_{\boldsymbol{\Theta}} J(\boldsymbol{\Theta})$ are gradient descent and particle swarm optimization.

  - Gradient descent learning : A first order iterative optimization technique for finding the local minimum of a multivariate differentiable function. It uses the fact that the direction if fastest decrease is in the opposite direction of the gradient of the function. Therefore for our objective function the iterative minimization is

$$\boldsymbol{\Theta}_{t+1} = \boldsymbol{\Theta}_t - \alpha \nabla J(\boldsymbol{\Theta}_t) \tag{8}$$

$$\nabla J(\boldsymbol{\Theta}) = \frac{1}{N} \sum_{n=1}^{N} \nabla Q_{\boldsymbol{\Theta}}(\boldsymbol{x}^{(n)}) \tag{9}$$

    The α is called the learning rate and corresponds to how rapidly we want to decrease our function by taking a proportion of steps that are in the direction is fastest decrease, it has the range $\alpha \in (0, 1]$. In practical the number of datapoints N is usually quite large and it may be computationally inefficient to consider the entire set at each iteration, this leads to a variation of the above algorithm in which the gradient in 9 is approximated by a sample of m with $m < N$points and takes the form.

$$\nabla J(\boldsymbol{\Theta}) = \frac{1}{m} \sum_{n=1}^{m} \nabla Q_{\boldsymbol{\Theta}}(\boldsymbol{x}^{(n)}) \tag{10}$$

    When the gradient in 10 is used the algorithm is called mini batch gradient descent. The batch size m is usually chosen such that it is a power of 2 in order to speec up computation. A special case of mini batch gradient descent is when $m = 1$ which is known as stochastic gradient descent and the gradient is approximated using a single data point. The stopping criteria used for gradient descent learning is the number of iterations and a tolerance. The algorithm stops when the maximum number of iterations is reached or the convergence criteria is satisfied.

  - Genetic algorithm learning : An optimization technique that is motivated by the process of natural selection. In this technique an initial set of possible solutions is randomly generated. Each of the possible solutions can be thought of as an individual in a population. The fitness of an individual is measured by the value they give for the cost function in 7. Now from the initial population we perform tournament selection and identify the most fit individuals (gives minimum value for cost function in current generation). We then crossover (linearly combine) these fit individuals to generate a set m new solutions that will replace the m least fit individuals in the population. This process is continued until a specified number of generations is reached. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are the 2 most fit members of the population then a new member $\boldsymbol{z}$ can be generated by using

$$\boldsymbol{z} = \beta \boldsymbol{x} + (1 - \beta)\boldsymbol{y} \quad \beta \in (0, 1) \tag{11}$$

- Multilayer perceptron model is an extension of the perceptron, this model uses the logistic function as an activation function for neurons and will output a probability value that will be used as a threshold for classification. The underlying learning algorithm is stochastic gradient descent. The architecture of the hidden layer in this network is such that it always consist of 50% of the number of neurons in the input layer.

*5) **Model hyper-parameters**:*

- kNN : To determine the optimal value of k for this model we perform a parameter sweep on the following search

space $\{3, 5, 7, 9, 11, 13\}$ and evaluate the error on the training set.

- LR : For gradient descent learning the hyper-parameter is the learning rate. The default number of iterations for gradient descent is 500. There is a trade-off between the number of allowed iterations and the learning rate. If the learning rate is too low and the number of allowed iterations is also low,then the number of iterations may not be sufficient to reach a local minimum. While if the learning rate is set too high and the number of iterations is also high the learning process becomes computationally inefficient as we may overshoot the minimum and not meet the convergence criteria, which means the algorithm will run the maximum number of iterations. The learning rate will be empirically determined by trying values in the range $[0.001, 0.3]$. In the case of genetic algorithm the important parameters are the initial population size N, tournament selection and the crossing over policy.

- MP : the learning rate will be emperically determined while the architecture consist of three layers with one hidden layer that has 34 neurons.

*6) **Testing models**:* The models were tested by evaluating their accuracy on the training dataset and testing datase, the time taken to build or obtain model parameters from the training dataset was also considered. The computational efficiency of the models was also evaluated by gauging how they perfom with high dimensional feature set, when PCA is removed.

IV. Experiments

The experiments were conducted in order to determine model parameters where applicable. All experiments were run on the training dataset.

A. kNN

The specified search space for a k value is $\{3, 5, 7, 13\}$. We perform a parameter sweep in the search space and evaluate the training error given by each unique value of k. The training errors are shown in table I below.

k	3	5	7	9	11	13
Training error	0.023	0.028	0.029	0.027	0.028	0.03

TABLE I: Training error for various k values

From table I it is clear that the optimal value of k in the given search space is 3.

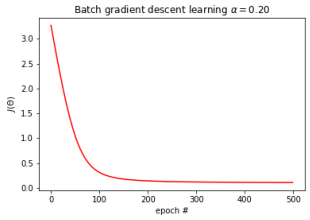
B. LR

- Gradient descent learning: As stated the hyper-parameter α was chosen using a parameter sweep in the search space $[0.001, ..., 0.5]$ with a step size of 0.1. The different values were used with batch gradient descent and fitted to the model in order to determine the training error for a given value of α. Table II below shows the associated error values of the different learning rates.

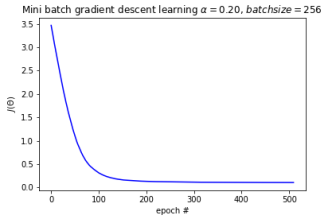
α	0.001	0.101	0.201	0.301	0.401
Training error	0.04	0.03	0.03	0.03	0.03

TABLE II: α values with error

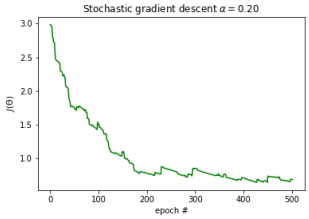
From Table II the ideal value of α could be any of the following valuess $0.201, 0.301, 0.401$. Although other values may be suitable as the final parameters depend on where the algorithm starts, in some instances the random initial parameters are close to the local minimum and low α values in some instances the random initial parameters are far from the local minimum in which high α values would be required. We choose $\alpha = 0.201$. We can explore the three variations of gradient descent with the given learning rate by plotting the cost function. See figure 5 below.



(a) Batch



(b) Mini-batch ($m = 256$)



(c) Stochastic

Fig. 5: Gradient descent variations

We observe that full batch gradient descent and mini batch gradient descent are quite similar and have little variance in the cost function while stochastic gradient descent has a significant amount of variation in cost function.

- Genetic algorithm learning: The maximum number of allowed generations was chosen to be 500 (same as the number of iterations for gradient descent learning). In this learning algorithm an initial set of N parameters are randomly generated and then the 2 most fit candidates are chosen to produce a set of m new solutions according to 11. The paramaters N and m were determined by trial and error and after several attempts were chosen to be $N = 50$ and $m = 10$.

N	23	100	40	50
m	6	12	8	10
time(s)	6.42	22.8	10.78	15.93
Training error	2.9	0.0292	0.06	0.10

TABLE III: Genetic algorithm parameters

C. MP

The multilayer perceptron has the logistic function as an activation function and stochastic gradient descent as a learning algorithm in which the learning rate is set to 0.02.

V. Results

A. kNN

The 3-NN model achieved a trainig accuracy of 97.72% and a testing accuracy of 84.58%. The confusion matrix for this models is shown below.

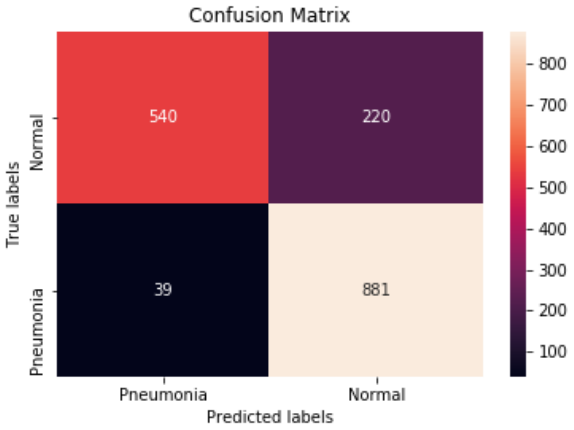


Fig. 6: 13-NN confusion matrix

B. LR

The best logistic regression model is the one that uses batch gradient descent learning and it achieved a training accuracy of 97.03% and a testing accuracy of 85.48. The confusion matrix is shown below.

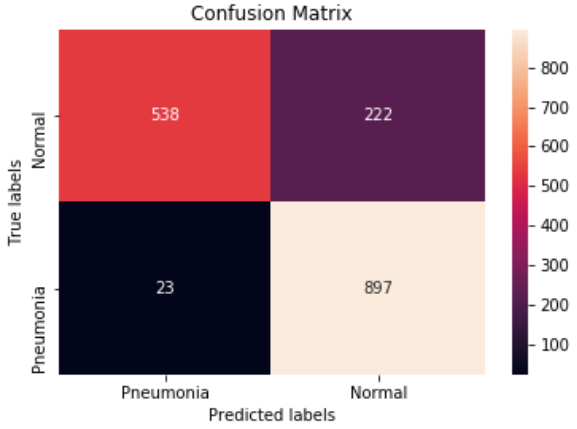


Fig. 7: Batch gradient descent logistic regression

In some rare instances the genetic algorithm logistic regression model does outperform the batch gradient descent model and it has in some experiments achieved a peak testing accuracy of 86%.

C. MP

The multilayer perceptron model achieved a training accuracy of 99.83 and a testing accuracy of 86.31%. The confusion matrix is shown below

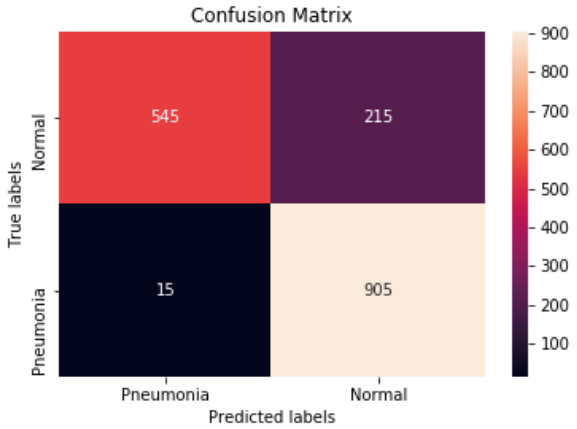


Fig. 8: MP confusion matrix

D. Model Comparison

Model	Training accuracy %	Testing accuracy %	time (seconds)
knn	97.72	84.58	1.57
LR	96.93	85.42	0.81
MP	99.83	86.31	0.05

TABLE IV: Model performance with dimensionality reduction

Model	Training accuracy %	Testing accuracy %	time (seconds)
knn	96.57	89.29	154.46
LR	93.84	83.15	41.19
MP	100	87.8	2

TABLE V: Model performance without dimensionality reduction

VI. Discussion

From the results in the previous section it is observed that the multilayer perceptron model outperformed all other models and achieved a peak accuracy of 86.31% on the testing set. This means that the stated research hypothesis can be accepted as results are consistent with it. The high performance of the logistic regression and multilayer perceptron suggest that the given dataset is linearly separable as these classifiers are known to work well in binary classification in which the data is linearly separable. The main factor that might have capped the accuracy of models is data loss. In this project data was lost in two significant parts, in the data preparation phase in which the chest X-ray images were resized to be 64×64, reducing the resolution of the images significantly degraded image quality in which important details for classification might have been lost, data was also lost in the dimensionality reduction step (This can be accounted for by the higher accuracy achieved by kNN and MP in table V when there is no dimensionality reduction).

VII. Conclusion

In this work were presented the problem of classifying chest X-ray images in order to aid radiologist in establishing a diagnosis for pneumonia. A challenge faced in tackling this problem was the high dimensionality of the feature space which was succesfully remedied by principal component analysis, the developed calcification algorithms are suitable for practical use in medical imaging for diagnosis. In futute the following may be attempted

- Identification of underlying cause of pneumonia (bacteria or virus) from chest X-ray.
- Automated detection of several respiratory diseases from chest X-ray images.

References

[1] Chandra, Tej Bahadur, and Kesari Verma. "Pneumonia Detection on Chest X-Ray Using Machine Learning Paradigm." In Proceedings of 3rd International Conference on Computer Vision and Image Processing, pp. 21-33. Springer, Singapore, 2020.

[2] Aydogdu, Müge, Ezgi Ozyilmaz, Handan Aksoy, G. Gursel, and Numan Ekim. "Mortality prediction in community-acquired pneumonia requiring mechanical ventilation; values of pneumonia and intensive care unit severity scores." Tuberk Toraks 58, no. 1 (2010): 25-34.

[3] World Health Organization. Standardization of interpretation of chest radiographs for the diagnosis of pneumonia in children. No. WHO/VB/01.35. World Health Organization, 2001.

[4] Neuman, Mark I., Edward Y. Lee, Sarah Bixby, Stephanie Diperna, Jeffrey Hellinger, Richard Markowitz, Sabah Servaes, Michael C. Monuteaux, and Samir S. Shah. "Variability in the interpretation of chest radiographs for the diagnosis of pneumonia in children." Journal of hospital medicine 7, no. 4 (2012): 294-298.

[5] Nilsson, Nils J. "Introduction to machine learning: An early draft of a proposed textbook." (1996).

[6] Kumar, Vinod, and Anil Saini. "Detection system for lung cancer based on neural network: X-Ray validation performance." International Journal of Enhanced Research in Management Computer Applications 2, no. 9 (2013): 40-47

[7] Gindi, A., Tawfik A. Attiatalla, and Moustafa M. Sami. "A comparative study for comparing two feature extraction methods and two classifiers in classification of early stage lung cancer diagnosis of chest x-ray images." Journal of American Science 10, no. 6 (2014): 13-22.

[8] Song, Fengxi, Zhongwei Guo, and Dayong Mei. "Feature selection using principal component analysis." In 2010 international conference on system science, engineering design and manufacturing informatization, vol. 1, pp. 27-30. IEEE, 2010.

[9] Jolliffe, Ian T., and Jorge Cadima. "Principal component analysis: a review and recent developments." Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374, no. 2065 (2016): 20150202.

[10] Singh, Vineet, and Sunil Pranit Lal. "Digit recognition using single layer neural network with principal component analysis." In Asia-Pacific World Congress on Computer Science and Engineering, pp. 1-7. IEEE, 2014.

[11] Zhang, Yu-Dong, and Lenan Wu. "An MR brain images classifier via principal component analysis and kernel support vector machine." Progress In Electromagnetics Research 130 (2012): 369-388.

[12] Kermany, Daniel S., Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown et al. "Identifying medical diagnoses and treatable diseases by image-based deep learning." Cell 172, no. 5 (2018): 1122-1131.