Predictions of Pneumonia Disease using Image Analytics in Orange Tool

Hafiz Muneeb Ahmad
Department of Computer Science
Riphah International University
Lahore, Pakistan
ahmad.muneeb8470@gmail.com

Muhammad Sohail
Department of Computer Science
Riphah International University
Lahore, Pakistan
maliksohail284@yahoo.com

Muhammad Muneeb Ahmad
Department of Computer Science
Riphah International University
Lahore, Pakistan
muneeb.biomeds14@gmail.com

Shahzaib Iqbal
Department of Computer Science
Riphah International University
Lahore, Pakistan
shahzaibiqbal43@gmail.com

Arshad Sarfaraz
Department of Computer Science
Riphah International University
Lahore, Pakistan
arshadsarfaraz45@yahoo.com

Kainat Noor
Department of Computer Science
Riphah International University
Lahore, Pakistan
Nkainat801@gmail.com

***Abstract*—Pneumonia is an infection that causes the air sacs to swell in one or both lungs. Air bags can be filled with fluid or pus (pap material), which can cause mucus or pus, fever, chills and difficulty breathing. Many organisms, including bacteria, viruses and fungi, can cause pneumonia. In this paper we used the data mining tool orange and work on image analytics project. We used the two algorithms, that one is deep learning model Neural Network and then we applied the classification model that named is Logistic Regression. For this method we take the Chest X-rays images data set,that is taken by kaggle. In Chest X-ray two types of images, one is normal images and the second images are effected with pneumonia disease. After this we trained and test the model, and predict the how many patients that have Pneumonia according to the given data set. That two algorithms predict the disease Pneumonia in the form of Yes or No. After that experiments we see the results, that is Logistic Regression give us the better accuracy on the prediction of disease.**

Keywords---Pneumonia, Data mining tool, Neural Network, Logistic Regression, Chest X-rays, Predictions, Orange.

I. INTRODUCTION

"Pneumonia are unique of the principal causes of infant mortality. The identify the indicators, the basic steps of identifying the maintenance system, select the software and equipment it is used."[1]. Pneumonia results in swelling of the lungs which, if not pickled in time, may be life aggressive.So the expert knowledge and skill are mandatory to deliver the report or image the chest X-ray carefully. For babies and young children, people older than 65 years, and people with health situations or cooperated immune systems, it is most dangerous. Many germs have the potential to cause pneumonia. Symptoms of the pneumonia are the chest pain,when you are breath or cough,Fever sweating and shaking chills, Nausea, vomiting or diarrhea, shortness of breath etc.

Images are often use to diagnose disease like ultra sound, X-rays etc. Pictures are the 1st procedure or process to detect the internal disorders of body.These picture provide a better internal view and its help doctors for diagnose the internal disease.

So we see the Chest X-ray Fig 1, that is described the normal lungs and effected lungs with Pneumonia.

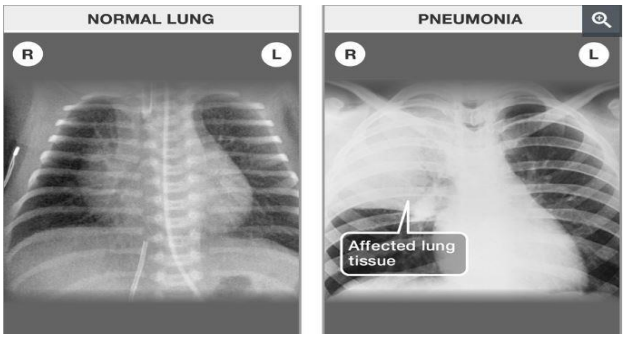


Fig 1: Normal Lung & effected lungs with pneumonia

In figure 1 the left one is normal Chest X-ray and 2nd one is Pneumonia Chesy X-ray. The lungs with pneumonia that is effected part will appear in white colour in the chest x-ray. The white shadow in the effected X-ray picture that is caused by filled in lungs air sacs.

So in the field of computer science there are many softwares and tools which are made for predicting these types of disease using some models. In this paper we are also working on some models to predict the disease on images dataset provided by patient Chest X-rays. We are using orange tool to predict the pneumonai with the help of algorithms. We apply the two different models that are predicted the pneumonia in form of yes or not. After that when we apply the neural network that is deep learning algorithm, the accuracy is 95% and the predict the pneumonia disease. When we apply the logistic regression algorithm that is classification algorithm, the accuracy is 97% and give us the maximum pneumonia predictions.

II. LITERATURE RIVIEW

“In this paper [2] the author proved that the proposed methodologies discussed the sample group, disease heterogeneous nature and scientific batch effects challenges correctly and advanced a very fast then reliable classifier using RNA sequencing for UIP classification”

“The proposed system consists of five convolution layers with 2 kernels and detections of LeakyReLU, followed by a pooling layer with size equal to the final feature maps and three dense layers. The last dense layer has seven outputs, corresponding to the classes considered: safe, ground glass opacity (GGO), micronodules, consolidation, reticulation, cracks and a GGO / reticulation combination” [3].

“In this paper [4] the author develop AN formula that may observe respiratory illness from chest X-rays at level extraordinary active radiologists. Our formula could be a 121-layer virtual neural network trained on Chi X-Net, Chest X-ray 14, which is currently the most important publicly accessible chest X-ray dataset that contains one hundred, X-ray images of fourteen diseases. Four tutorial radiologists read, explain a checklist, and we compare ChiXnet's performance to that of a radiologist. We find that CheXNet F1 exceeds the average performance of a medical specialist on metrics. Our tendency is to increase Che Cheets Net to observe 14 diseases in 14 chest x-rays and to come up with the best state results related to 14 diseases.”.

“This work investigates opportunities for applying machine learning solutions for automatic detection and localization of respiratory illness on chest x-ray pictures. we have a tendency to propose AN ensemble of 2 convolutional neural networks, particularly RetinaNet and Mask R-CNN for respiratory illness detection and localization. we have a tendency to valid our answer on a recently free dataset of twenty six, 684 pictures from Kaggle respiratory illness Detection Challenge and were score among the highest third-dimensional of submitted solutions”[5].

“In this paper author described the COVID-19 and they used the covid data set for the pneumonia prediction. This paper investigates deep learning strategies for mechanically analyzing question chest X-ray pictures with the hope to bring exactitude tools to health professionals towards screening the COVID-19 and identification confirmed patients. during this context, coaching datasets, deep learning architectures and analysis methods are experimented from in public open sets of chest X-ray pictures. Chest X-ray take a look at pictures of COVID-19 infected patients area unit with success diagnosed through detection models preserved for his or her performances. The power of planned health indicators is highlighted through simulated situations of patients presenting infections and health issues by combining real and artificial health information”[6].

“This study proposes a convolutional neural network model trained from scratch to classify and discover the presence of respiratory illness from a set of chest X-ray image samples. It is troublesome to get an oversized quantity of respiratory illness dataset for this classification task; so, we tend to deployed many knowledge augmentation algorithms to enhance the validation and classification accuracy of the CNN model and achieved outstanding validation accuracy”[7].

III. RESEARCH QUESTIONS

Q 1: What are the predictions of normal and effected patients of the pneumonia?

Q 2: In which algorithms give us the better accuracy and predictions?

IV. METHODOLOGY

The study of methodology that in this section we will discuss about dataset, tools and algorithms that I have used in this paper.

A. TOOLS DISCRIPTION

In ths section we chose the data mining tool orange for the predictions and check the accuracy of the effected paitients.The tools ara accessible for permitted resources on the internet. After that we sufferunf the internet we chose the two algorithms of classification and deep learning and meaure the accuracy and predctions.

ORANGE: It also a data mining tool and easily available over the internet. It has an feature of open source and it provides the best visulization than weka tool. It also provides the feature such as drag and drop. This tool is very useful for the executions of machine learning algorithms. It provides amazing GUI and add-ons facility. It has python scripting as well. This tool is very helpful for researchers and many researches had already done in it.

B. DATA SET EXPLANATION

Data set have taken from kaggle and data set name are pneumonia Chest X-Ray images data set. In which data set are described how much data set are used.

Chest X-Ray images are contain two different datasets. One is the train dataset images in which included normal images of chest X-Ray and pneumonia images of X-Ray. In train data we take the 50 images of normal chest X-Ray and the 80 images of pneumonai chest X-Ray.

Second is the test dataset images in which also included normal images of chest X-ray which is not effected by pneumonai and also included effected images of chest X-Ray. In test data we take the 20 images of normal chest X-Ray and the 50 images of pneumonai chest X-Ray.

Following the Table 1 we describe the above information in tabular form and also see the picturse of normal chest x-ray and effected by pneumonai chest x-rays.

Table 1: X-Ray Images Detail

Chest X-Ray Datasets	Normal Images	Pneumonai Images	Total Images
Train Images	50	80	130
Test Images	20	50	70

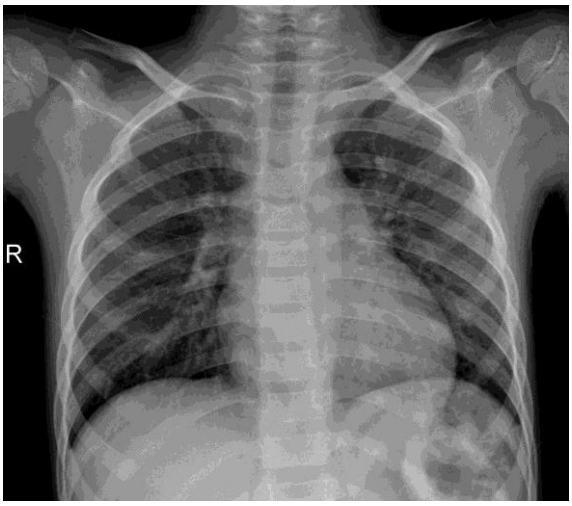


Fig 2: Normal Chest X-Ray

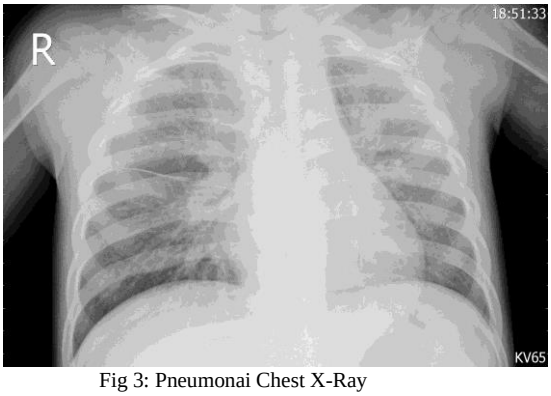


Fig 3: Pneumonai Chest X-Ray

Firstly we import our dataset using imort images. After that we use Image embedding to inhance the

variables of dataset. After embedding data we use concatenate node and concatenate data of normal and effected datatables.

C. ALGORITHMS DESCRIPTION

After import the images and experiments we chose the algorithms for train the model. Ist we apply the deep learning model that Neural Network and then we apply the second classification algorithm that are Logistic Regression.

- Neural Network

Neural networks are a set of algorithms that are freely modeled according to the principle of the human brain, which are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling, or raw input methods. Working of Neural network is a way similar to the working of human brain neuron network. Neural network "neurons" are mathematical applications that accumulate and catalogue information according to the design of a particular designer. The Neural Network consist of layers that are connected with nodes.

- Logistic Regression

Logistic regression is a classification algorithm used to establish observations on a set of classes. Some examples of classification problems are email or non spam, fake or not fake online transactions, malignant or benign. Logistic regression changes its output using the sigmid logistic function to return potential values.

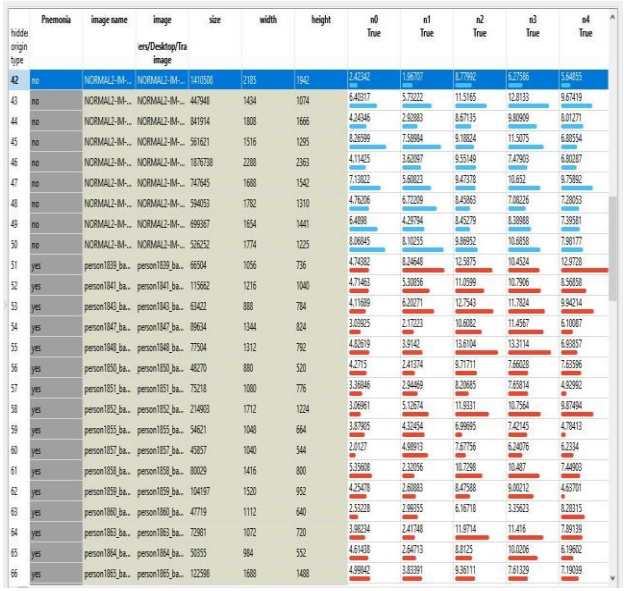
V. EXPERIMENTAL WORK IN ORANGE

In this section we will discuss about the experiment of pneumonia dataset and perform the orange tool.

First we have drag the import images icon and import the dataset images that already taken from kaggle. After that we drag the image viewer and image embedding icons. Image embedding used for the more features of data that is more important for our predictions and accuracy.

We should select the target variables that is categorical data for the pneumonai prediction in form of yes or no. ‘Yes’ are those value that is effected by pneumonai and ‘no’ are those values that are not effected by pneumonia.

We use concatenate icon for merge the train and test dataset images. After that we see the data table that is data together for the better results. We create the train file in excel sheet and write the 130 values in form of yes or no, because we train the model for the test data then we conclude the better results.



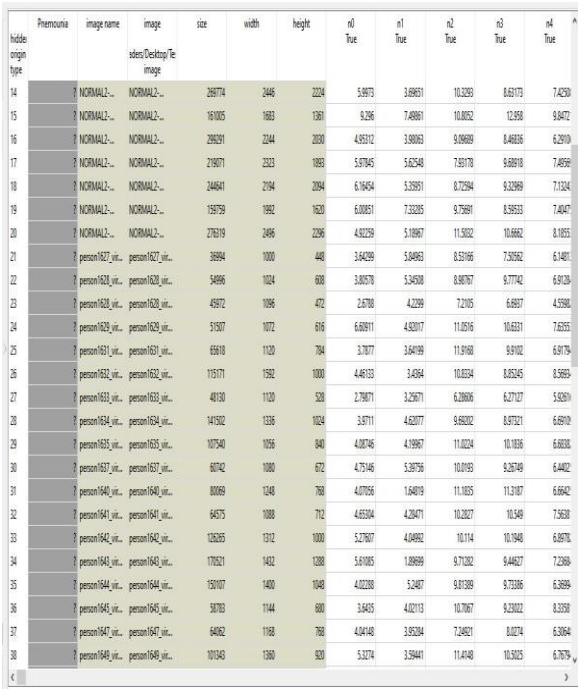
	Pneumonia	image name	image	size	width	height	n0 True	n1 True	n2 True	n3 True	n4 True
hidden origin type			ers/Desktop/Tra image								
42	no	NORMAL2-IM-...	NORMAL2-IM-...	1410508	2105	1942	2.42342	1.96707	8.77992	6.27586	5.94055
43	no	NORMAL2-IM-...	NORMAL2-IM-...	447948	1434	1074	6.40317	5.73222	11.5165	12.8133	9.67419
44	no	NORMAL2-IM-...	NORMAL2-IM-...	841914	1808	1666	4.24346	2.92883	8.67135	9.80909	8.01277
45	no	NORMAL2-IM-...	NORMAL2-IM-...	561621	1516	1295	8.26599	7.59884	9.18824	11.5075	6.88554
46	no	NORMAL2-IM-...	NORMAL2-IM-...	1876738	2288	2363	4.11425	3.62097	9.55149	7.47903	6.80287
47	no	NORMAL2-IM-...	NORMAL2-IM-...	747645	1688	1542	7.13822	5.69023	9.47378	10.652	9.75892
48	no	NORMAL2-IM-...	NORMAL2-IM-...	594053	1782	1310	4.76206	6.72209	8.45663	7.00226	7.28053
49	no	NORMAL2-IM-...	NORMAL2-IM-...	699367	1654	1441	6.4898	4.29794	8.45279	8.38988	7.39581
50	no	NORMAL2-IM-...	NORMAL2-IM-...	526252	1774	1225	8.06845	6.10255	9.86952	10.6858	7.98177
51	yes	person1839_ba...	person1839_ba...	66504	1056	736	4.74382	8.24648	12.5875	10.4524	12.9728
52	yes	person1841_ba...	person1841_ba...	115662	1216	1040	4.71493	5.30856	11.0599	10.7906	8.56858
53	yes	person1843_ba...	person1843_ba...	63422	888	784	4.11699	6.20271	12.7543	11.7824	9.94214
54	yes	person1847_ba...	person1847_ba...	89634	1344	824	3.02925	2.17223	10.6082	11.4567	6.10087
55	yes	person1848_ba...	person1848_ba...	77504	1312	792	4.02619	3.9142	13.6104	13.3114	6.93857
56	yes	person1850_ba...	person1850_ba...	48270	880	520	4.2715	2.41374	9.71711	7.66028	7.63596
57	yes	person1851_ba...	person1851_ba...	75218	1080	776	3.39845	2.94469	8.20685	7.65814	4.92992
58	yes	person1852_ba...	person1852_ba...	214903	1712	1224	3.06961	5.12674	11.9331	10.7564	9.87494
59	yes	person1855_ba...	person1855_ba...	54621	1048	664	3.87925	4.32454	6.99895	7.42145	4.78413
60	yes	person1857_ba...	person1857_ba...	45857	1040	544	2.0127	4.98913	7.67756	6.24076	6.2334
61	yes	person1858_ba...	person1858_ba...	80029	1416	800	5.35808	2.32056	10.7298	10.487	7.44903
62	yes	person1859_ba...	person1859_ba...	104197	1520	952	4.25478	2.69883	8.47588	9.00212	4.63701
63	yes	person1860_ba...	person1860_ba...	47719	1112	640	2.53228	2.99355	6.16718	3.59623	8.28315
64	yes	person1862_ba...	person1862_ba...	72981	1072	720	3.98234	2.41748	11.9714	11.416	7.89139
65	yes	person1864_ba...	person1864_ba...	50255	984	552	4.61438	2.64713	8.8125	10.0206	6.19692
66	yes	person1865_ba...	person1865_ba...	122598	1688	1488	4.98842	3.83391	9.36111	7.61329	7.19039

Fig 4: Merge data for train the model

In this fig we see that 50 values of ‘no’ that shows the normal paitients and 80 values of ‘yes’ that shows the effected persons, these vlaues that we create the excel sheet and merge in the table form in orange tool for train the model.

After that we use the column icon and apply algorithms Neural Network and Logistic regression to train these models.

Similar process is applied on test dataset and connects it with select column of train dataset to predict the data. Similar as above method we create the test file in excel sheet and write the 70 values in form of Question mark ?, because we test the model that are given values true or not and predict the results with the best algoritms.



	hidden: origin type	Pneumonia	image name	image sders/Desktop/Te image	size	width	height	n0 True	n1 True	n2 True	n3 True	n4 True
14		?	NORMAL2-...	NORMAL2-...	269774	2446	2224	5.9973	3.69651	10.3293	8.63173	7.4250
15		?	NORMAL2-...	NORMAL2-...	161005	1683	1361	9.296	7.48861	10.8052	12.958	9.0472
16		?	NORMAL2-...	NORMAL2-...	299291	2244	2030	4.95312	3.98063	9.09689	8.46836	6.2910
17		?	NORMAL2-...	NORMAL2-...	219071	2323	1893	5.97845	5.62548	7.83178	9.68918	7.4599
18		?	NORMAL2-...	NORMAL2-...	244641	2194	2094	6.16454	5.39351	8.72594	9.32969	7.1324
19		?	NORMAL2-...	NORMAL2-...	158759	1992	1620	6.00851	7.33285	9.75691	8.59533	7.4047
20		?	NORMAL2-...	NORMAL2-...	276319	2496	2296	4.92259	5.18967	11.5032	10.6662	8.1855
21		?	person1627_vir...	person1627_vir...	36994	1000	448	3.64299	5.84963	8.53166	7.30562	6.1401
22		?	person1628_vir...	person1628_vir...	54896	1024	608	3.80578	5.34508	8.98767	9.77742	6.9126
23		?	person1628_vir...	person1628_vir...	45972	1096	472	2.6788	4.2299	7.2105	6.6897	4.5598
24		?	person1629_vir...	person1629_vir...	51507	1072	616	6.00911	4.92017	11.0516	10.6331	7.6255
25		?	person1631_vir...	person1631_vir...	65618	1120	784	3.7077	3.64199	11.9168	9.9102	6.9179
26		?	person1632_vir...	person1632_vir...	115171	1592	1000	4.46133	3.4364	10.8334	8.85245	9.5695
27		?	person1633_vir...	person1633_vir...	48130	1120	528	2.79871	3.25671	6.28606	6.27127	5.9261
28		?	person1634_vir...	person1634_vir...	141502	1336	1024	3.8711	4.62077	9.69202	8.87321	6.6910
29		?	person1635_vir...	person1635_vir...	107540	1056	840	4.08746	4.19967	11.0224	10.1036	6.6838
30		?	person1637_vir...	person1637_vir...	60742	1080	672	4.75146	5.39756	10.0193	9.26749	6.4402
31		?	person1640_vir...	person1640_vir...	80069	1248	768	4.07956	1.64819	11.1835	11.3187	6.6642
32		?	person1641_vir...	person1641_vir...	64575	1088	712	4.65304	4.28471	10.2827	10.549	7.5638
33		?	person1642_vir...	person1642_vir...	126265	1312	1000	5.27607	4.04982	10.114	10.1048	6.8978
34		?	person1643_vir...	person1643_vir...	170521	1432	1288	5.61085	1.89699	9.71282	9.44627	7.2368
35		?	person1644_vir...	person1644_vir...	150107	1400	1048	4.02288	5.2487	9.81389	9.73386	6.3899
36		?	person1645_vir...	person1645_vir...	58783	1144	680	3.6435	4.02113	10.7067	9.23022	8.3258
37		?	person1647_vir...	person1647_vir...	64062	1168	768	4.04148	3.95284	7.24921	8.0274	6.3064
38		?	person1648_vir...	person1648_vir...	101343	1360	920	5.3274	3.59441	11.4148	10.5025	6.7679

Fig 5: Merge data for test the model

For prediction In this fig we see that 20 values of 'no' that shows the normal paitients and 50 values of 'yes' that shows the effected persons, and total values are 70 that represent in question mark ?. these vlaues that we create the excel sheet and merge in the table form in orange tool for prediction the images that are effected or not.

After that we use test and score nodes to check the accuracy of algorithms and apply some visualization nodes for better results. Also attached the prediction node for predict the normal and effected patients.

At the end in fig 6 we show that the work flow of orange with different nodes is attached for better accuracy and prediction results.

All over the project work done in data mining tool orange for the predictions of pneumonia in chest X-Ray detest. Orange nodes with connects each other and see that test and score with predictions and visualization widget's.

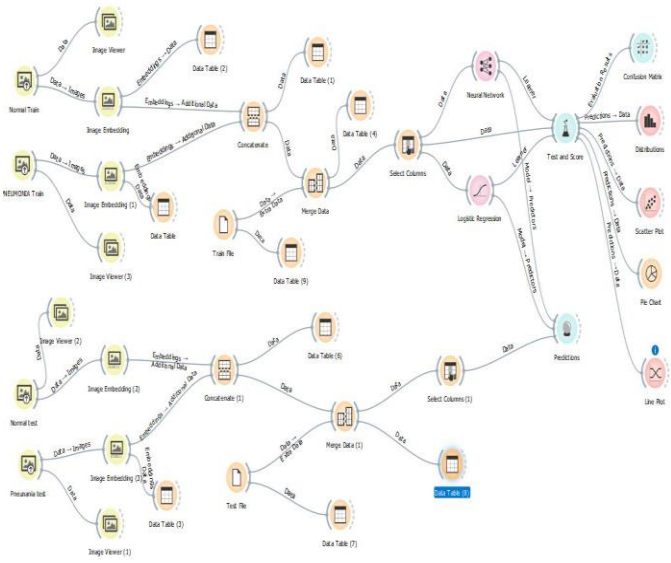


Fig 6: Work flow diagram in orange tool

Work flow of the orange tool show the image analytics project with different nodes of attached the two algorithms with best vizulization nodes.

VI: EXPERIMENTAL RESULTS

Now in this section we will discuss the results that we get the after experiment in orange tool.

When we apply the prediction node on the test data set model attached with two nodes of algorithms then we see the different results of prediction in form of ssyes or not. Show the results in fig 7 below:

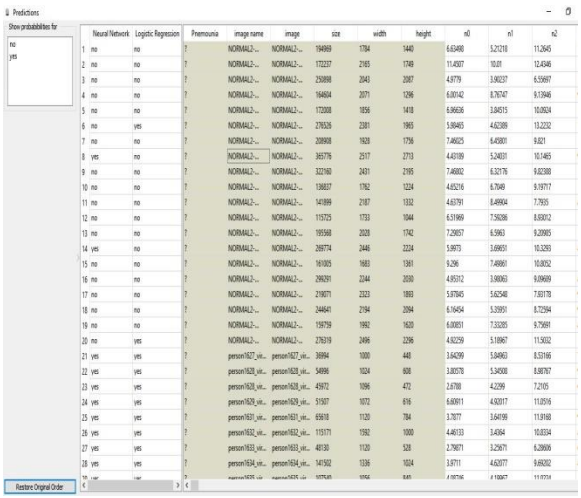


Fig 7: Prediction Results

In this section we report our results of predictions followed by the given algorithms. In logistic regression give us the better results for predict the pneumonai in form of yes or not. Because logistic regression accuracy is better then Neural Network. In the other hand the Neural Network also perom the good predictions, but not more than logistic regression model. Logistic regression are better understand the predict pneumonai disease than the neural network.

If we zoom the fig 7, we see that Ist 20 values that are described the prediction of pneumonai that the form of “no or yes”. ‘No’ stand for normal chest X-Ray and ‘yes’ satnd for effected by pneumonai chest X-ray. All these values predict by the train models.

After that we show the accuracy table by the test and score node that attached with the logistic regression and Neural Network algorithms.

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Neural Network	0.987	0.954	0.954	0.956	0.954
Logistic Regression	0.999	0.977	0.977	0.977	0.977

Fig 8: Accuracy of Algorithms

In the above “fig 8” we compare the results of two different algorithms that we apply on dataset on the basis of different X-Ray images, basically we are comparing the accuracy of these two algorithms and find out that Logistic Regression give us the better accuracy during prediction.

In NN model the classification accuracy is 95%, which not more efficient from the Logistic Regression percentage.

In Logistic Regression algorithem the classification accuracy is 97%, which is more efficient from the Neural Network percentage.

After that for better understanding we attach the different nodes of visulization with test and score icon and perform the different results in the form of graph that describes the all normal and pneumonai results. The more efficient results in orange that we use the satter plot for the better visulization of algorithms accuracy.

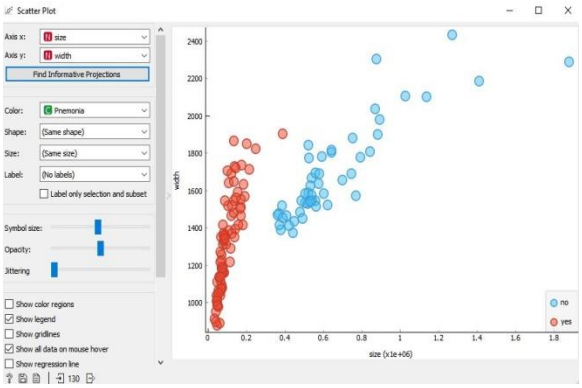


Fig 9: Scatter plot on Pneumonai results

Scatter plots are surely one of the best method visualizations in Orange.

Fig 10: Confusion matrix on LR results

In confusion matrix fig shows the positvie and negative reults. In green colour values shows the positive and efficient accuracy results and the red colour give us the negative or wrong results.

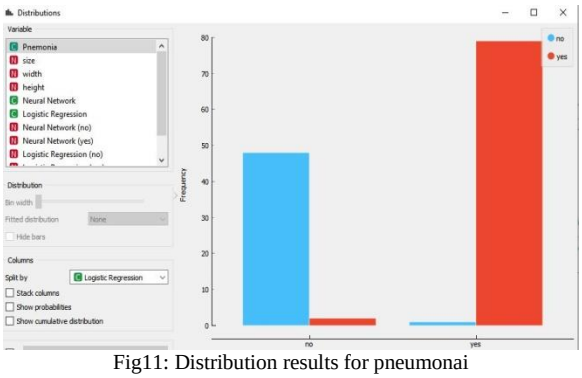


Fig11: Distribution results for pneumonai

Red line shows the ‘Yes’ values for Pneumonia symptoms and blue line shows the ‘No’ values in Logistic Regression model.

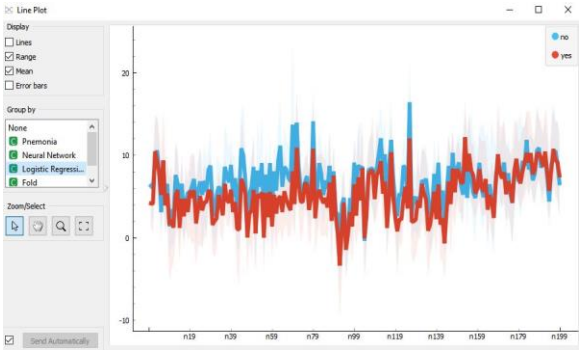


Fig 12: Line plot in LR results

All the visulization figures shows the same results but different graph styles for better understanding.

VII. CONCLUSION

We conclude that in this paper we use image analytics project on different chest X-Ray Images for prediction of pneumonai. In this paper we use two different algorithms that are neural network and logistic regression. Ist of all we apply neural network algorithm on train data set to train our model and 2nd algorithm that we use is logistic regression. After prediction we conclude that logistic regression give us 97% prediction accuracy and neural network give us 95% accuracy. Both algorithms are better for predictions but in this project we can say that the logistic regression give us better results for prediction on pneumonia disease.

VIII. FUTURE WORK

In furture research, we can use any kind of data set for prediction and apply more algorithms to generate different result. We use the different tool like Knime, Python, Rapid Miner, Weka in classification and clustring algorithms and compare the different kinds of results for more efficient work and predictions.

REFERENCES

[1] A. K. Jaiswal, P. Tiwari, S. Kumar, D. Gupta, A. Khanna, and J. J. Rodrigues, "Identtifying pneumonia in chest X-rays: A deep learning approach," *Measurement*, vol. 145, pp. 511-518, 2019.

[2] Y. Choi *et al.*, "Identification of usual interstitial pneumonia pattern using RNA-Seq and machine learning: challenges and solutions," *BMC genomics*, vol. 19, no. 2, p. 101, 2018.

[3] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1207-1216, 2016.

[4] P. Rajpurkar *et al.*, "Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225 [cs, stat]*, 2017.

[5] I. Sirazitdinov, M. Kholiavchenko, T. Mustafaev, Y. Yixuan, R. Kuleev, and B. Ibragimov, "Deep neural network ensemble for pneumonia localization from a large-scale chest x-ray database," *Computers & Electrical Engineering*, vol. 78, pp. 388-399, 2019.

[6] K. Hammoudi *et al.*, "Deep Learning on Chest X-ray Images to Detect and Evaluate Pneumonia Cases at the Era of COVID-19," *arXiv preprint arXiv:2004.03399*, 2020.

[7] L. Wang and A. Wong, "COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images," *arXiv preprint arXiv:2003.09871*, 2020.