**PNEUMONIA DETECTION**

A Course Project report submitted

in partial fulfillment of requirement for the award of degree

**BACHELOR OF TECHNOLOGY**

in

**ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING**

by

**D.BALA VARSHITHA (2103A52012)**

**P.TEJASWI (2103A52029)**

**K.SRINIDHI (2103A52051)**

Under the guidance of

**Mr. Ramesh Dadi**

Assistant Professor, Department of CSE.

**

**Department of Computer Science and Artificial Intelligence**



**Department of Computer Science and Artificial Intelligence**

**CERTIFICATE**

This is to certify that project entitled **“PNEUMONIA DETECTION”** is the bonafied work carried out by **Bala Varshitha,Tejaswi and Srinidhi** as a Course Project for the partial fulfillment to award the degree **BACHELOR OF TECHNOLOGY** in **ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING** during the academic year 2023-2024 under our guidance and Supervision.

**Mr. D.Ramesh Dr. M.Sheshikala**

Asst. Professor, Assoc. Prof. & HOD (CSE), S R University, S R University,

Ananthasagar, Warangal. Ananthasagar, Warangal.

**ACKNOWLEDGEMENT**

We express our thanks to Course co-coordinator **Mr.** **D.Ramesh, Asst. Prof.** for guiding us from the beginning through the end of the Course Project. We express our gratitude toHead of the department CS&AI, **Dr. M.Sheshikala, Associate Professor** for encouragement, support and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean, School of Computer Science and Artificial Intelligence, **Dr C. V. Guru Rao**, for his continuous support and guidance to complete this project in the institute.

Finally, we express our thanks to all the teaching and non-teaching staff of the department for their suggestions and timely support.

**ABSTRACT**

This paper surveys and examines how computer-aided techniques can be deployed in detecting pneumonia. It also suggests a hybrid model that can effectively detect pneumonia while using the real-time medical image data in a privacy preserving manner. This paper will explore how various preprocessing techniques such as X-rays can detect and classify multiple diseases.

The survey also examines how different machine learning technologies like convolution neural network (CNN), k-nearest neighbor (KNN), Support vector machine(SVM),Random forest and Decision Tree can be used in detecting pneumonia disease.

In this article, we have performed a comprehensive review of the literature to find how we can combine hospitals and medical institutions to train the machine learning models from their datasets so that the ML algorithms can detect disease more efficiently and correctly. We have proposed the future work of using transfer learning combined with federated knowledge that could help the medical institutions and hospitals form a combined approach of performing medical image detection using real-time datasets. We have also explored the scope, future work and limitations of the proposed solution.

**TABLE OF CONTENTS**

**Chapter No. Title Page No.**

1. Introduction
   1. Overview 1
   2. Problem Statement 2
   3. Existing system 2
   4. Proposed system 2-3
   5. Objectives 3
   6. Architecture 4
2. Literature survey

2.1.1 Survey 5

1. Data pre-processing

3.1. Dataset description 6-8

3.2. Data cleaning 9

3.3. Data Augmentation 9

3.4. Data Visualization 9

1. Methodology
   1. Logistic Regression 10-11

4.2 KNN 12-13

4.3 SVM 14-15

* 1. Decision Tree 16-17

4.5 Random Forest 18

1. Results and discussion 19
2. Conclusion and future scope 20
3. References 21

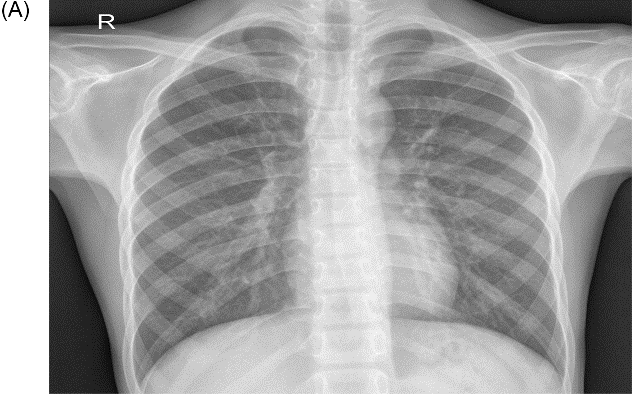
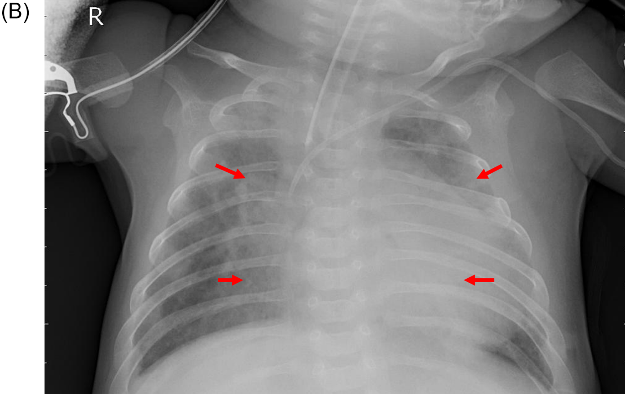
**INTRODUCTION**

* 1. **OVERVIEW**

Pneumonia is an acute pulmonary infection that can be caused by bacteria, viruses, or fungi and infects the lungs, causing inflammation of the air sacs and pleural effusion, a condition in which the lung is filled with fluid. It accounts for more than 15% of deaths in children under the age of five years . Pneumonia is most common in underdeveloped and developing countries, where overpopulation, pollution, and unhygienic environmental conditions exacerbate the situation, and medical resources are scanty. Therefore, early diagnosis and management can play a pivotal role in preventing the disease from becoming fatal.

Radiological examination of the lungs using computed tomography (CT), magnetic resonance imaging (MRI), or radiography (X-rays) is frequently used for diagnosis. X-ray imaging constitutes a non-invasive and relatively inexpensive examination of the lungs.  The white spots in the pneumonic X-ray (indicated with red arrows), called infiltrates, distinguish a pneumonic from a healthy condition. However, chest X-ray examinations for pneumonia detection are prone to subjective variability.

Thus, an automated system for the detection of pneumonia is required. In this study, we developed a computer-aided diagnosis (CAD) system that uses an ensemble of deep transfer learning models for the accurate classification of chest X-ray images.

** **

**NORMAL PNEUMONIC**

**[1]**

* 1. **PROBLEM STATEMENT**

According to the report released during World Pneumonia Day, it is estimated that more than 11 million infant children below the age of 5 years are likely to die from pneumonia by the year 2030. According to the report released by “our World in data” , children below five have the highest death rate caused by pneumonia . In 2017, 808,920 children died due to pneumonia, and this figure is 16 folds more than the deaths caused by cancer a year.

The Chest X-ray is commonly used in detecting pulmonary diseases like pneumonia. The problem of lack of experts can be addressed through the use of different computer-aided diagnosis techniques. Technological advancements in artificial intelligence (AI) have proven to be helpful in the diagnosis of disease.

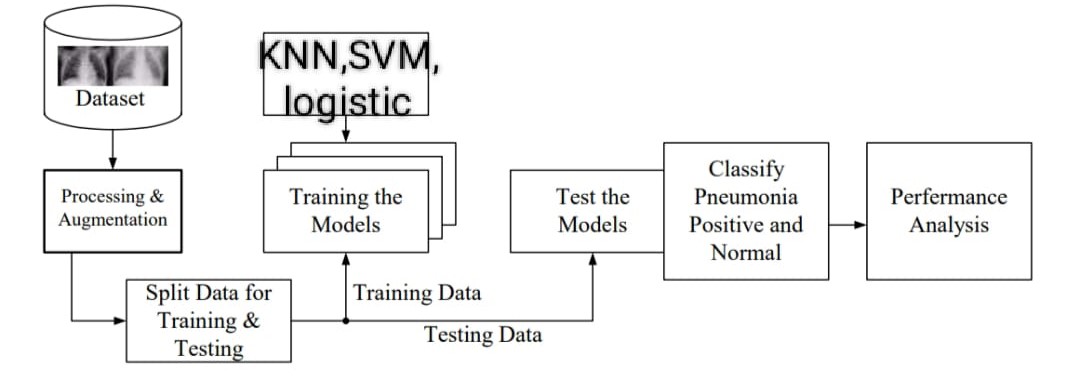
* 1. **EXISTING SYSTEM**

According to the search results, there are different types of existing systems for pneumonia detection based on machine learning and deep learning techniques. Some of them are:

* **Faster R-CNN**: A two-stage object detector that uses a region proposal network (RPN) to generate candidate regions and then classifies them using a convolutional neural network (CNN)
* **ResNet**: A CNN architecture that uses residual connections to overcome the problem of vanishing gradients and improve accuracy. It can have different depths such as 50 or 101 layers
* **CheXNet**: A CNN model that is trained on a large dataset of chest X-rays and can detect 14 thoracic diseases, including pneumonia 3.
* **DECNET**: A CNN model that uses a deformable convolution layer to adapt to the shape and size of the lung regions and improve the detection of pneumonia.
  1. **PROPOSED SYSTEM**

We have anaylsed the various work done on medical image detection in the previous section. The experiments were performed based on available datasets. It has been observed that the machine learning models effectively detects medical images when the model is fed with a larger quantity of data. The use of ML algorithms has been proven effective in detection while compared to the traditional procedures mentioned in the literature review. ML models need a higher volume of data for effective training capable of achieving higher accuracy in detection.

[2]

****

**Proposed system**

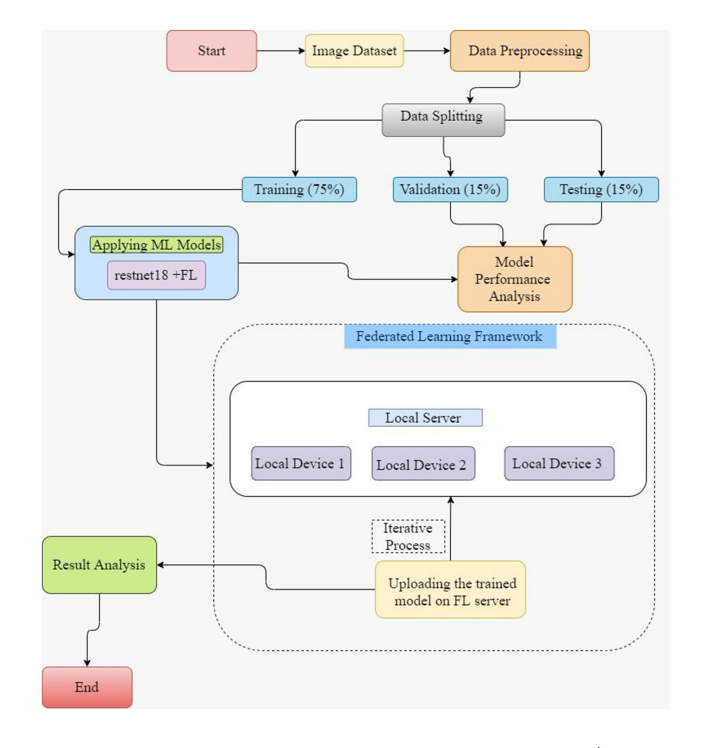
**1.5 OBJECTIVES**

The main objective of this research is to develop a pneumonia detection system. The objective of this project is to build a model that can accurately detect the presence or absence of pneumonia. This study focuses on building a deep learning model based on deep learning that could directly scan the X-ray and detect whether the person is suffering from pneumonia or not. It helps to avoid human interaction and also reduce the cause of disease and gives accurate result whether there is pneumonia or not.

A feature can appear anywhere in a digital image, pixel values are stored in a two-dimensional (2D) grid, i.e., an array of numbers, and a small grid of parameters called kernel, an optimizable feature extractor, is applied at each image position, CNNs are highly efficient for image processing. Extracted features can evolve hierarchically and progressively more complicated as one layer feeds its output into the next layer. Training is the process of adjusting parameters like kernels to reduce the discrepancy between outputs and ground truth labels using optimization algorithms like backpropagation and gradient descent, among others.

**[3]**

**1.6 ARCHITECTURE**



[4]

**2.LITERATURE SURVEY**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.NO** | **AUTHOR** | **YEAR** | **APPROACH** | **RESULT** |
| 1. | TatianaGabruseva, Dmytro Poplavskiy,Alexander Kalinin | 2020 | Computervision andpattern recognition | 77% accuracy |
| 2. | Dimpy Varshni, Kartik Thakral, Lucky Agarwal | 2019 | CNN based feature extraction | 66.7% accuacy |
| 3. | PranavRajpurkar, Jeremy Irvin, Kaylie Zhu | 2017 | Radiologist-level pneumonia detection | 91% accuracy |
| 4. | Vikash Chouhan, Sanjay Kumar Singh, Aditya Khamparia | 2020 | Novel transfer learning based approach. | 99% accuracy |
| 5. | AshitoshTilve, Shrameet Nayak | 2020 | Pneumonia detection using deep learning | 65% accuracy |
| 6. | Marcelo Fiszman, Wendy W Chapman | 2000 | Automatic detection of acute bacterial pneumonia | 84.5% accuracy |
| 7. | MohammadFarukh HashmiSatyarth Katiyar, | 2020 | Deep transfer learning | 54.2% accuracy |
| 8. | Abdullah Faqih Al Mubarok, Jeffrey AM Dominique | 2019 | deep convolutional architecture | 67% accuracy |
| 9. | TawsifurRahman, MuhammadEH Chowdhury, | 2020 | transfer learning with deep CNN | 78.9% accuracy |
| 10. | Ian Pan, Alexandre Cadrin-Chênevert, | 2019 | pneumonia detection in chest xray |  |

[5]

**3. DATA PRE-PROCESSING**

**3.1 DATASET DESCRIPTION**

**About the dataset:**

* The dataset is organized into 3 folders (train, test, val) and contains subfolders for each image category (Pneumonia/Normal). There are 5,863 X-Ray images (JPEG) and 2 categories (Pneumonia/Normal).
* For the analysis of chest x-ray images, all chest radiographs were initially screened for quality control by removing all low quality or unreadable scans.
* There are 4872 training images for pneumonia and 1341 training images for normal condition.
* There are 390 testing images for pneumonia and 234 testing images for normal condition.

The normal chest X-ray (left panel) depicts clear lungs without any areas of abnormal opacification in the image. Bacterial pneumonia (middle) typically exhibits a focal lobar consolidation, in this case in the right upper lobe (white arrows), whereas viral pneumonia (right) manifests with a more diffuse ‘‘interstitial’’ pattern in both lungs.

Chest X-ray images (anterior-posterior) were selected from retrospective cohorts of pediatric patients of one to five years old from Guangzhou Women and Children’s Medical Center, Guangzhou. All chest X-ray imaging was performed as part of patients’ routine clinical care.

For the analysis of chest x-ray images, all chest radiographs were initially screened for quality control by removing all low quality or unreadable scans. The diagnoses for the images were then graded by two expert physicians before being cleared for training the AI system. In order to account for any grading errors, the evaluation set was also checked by a third expert.

***Target:***

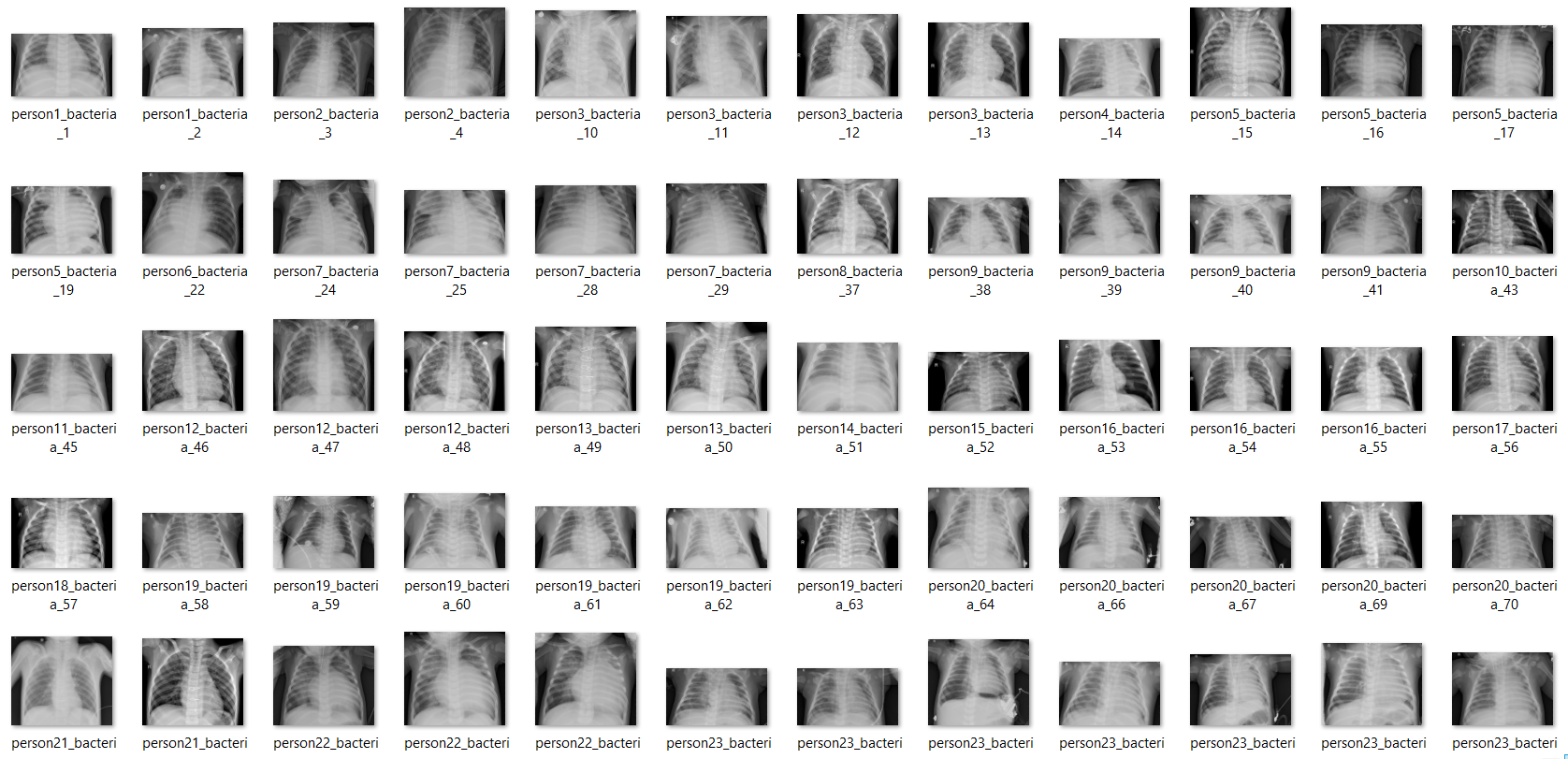
**1.Target Variable:**Pneumonia/Normal

**2.Pneumonia/Normal:**If pneumonia is present then value is 1 else it is 0.

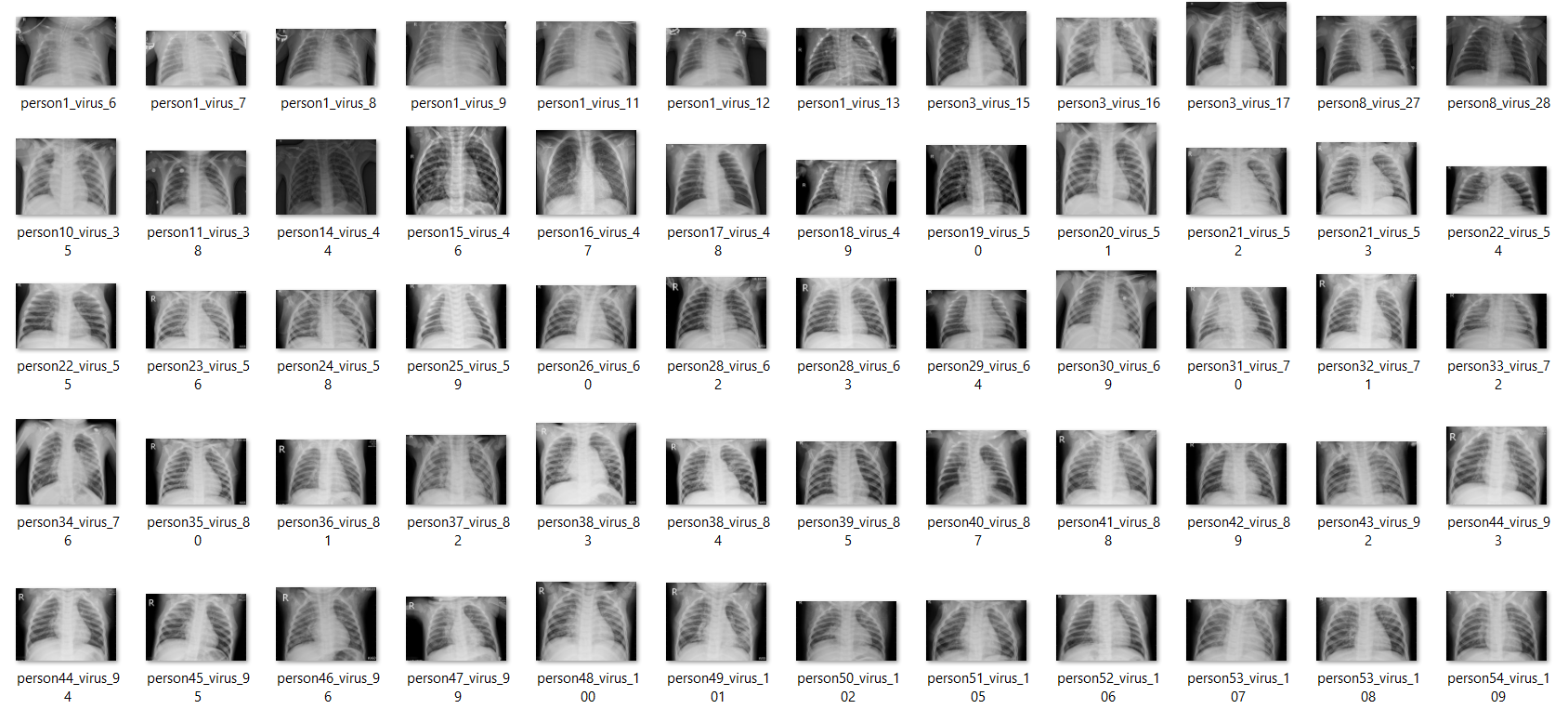
**[6]**

**Dataset:**

**Training dataset**

****

**Testing dataset:**

****

**[7]**

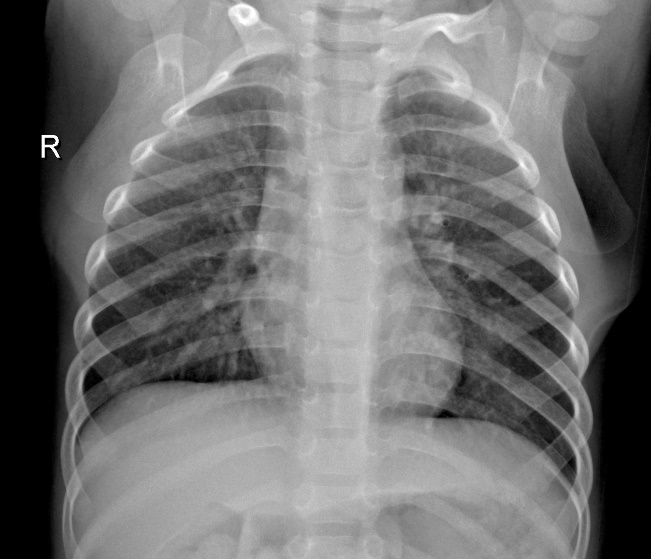
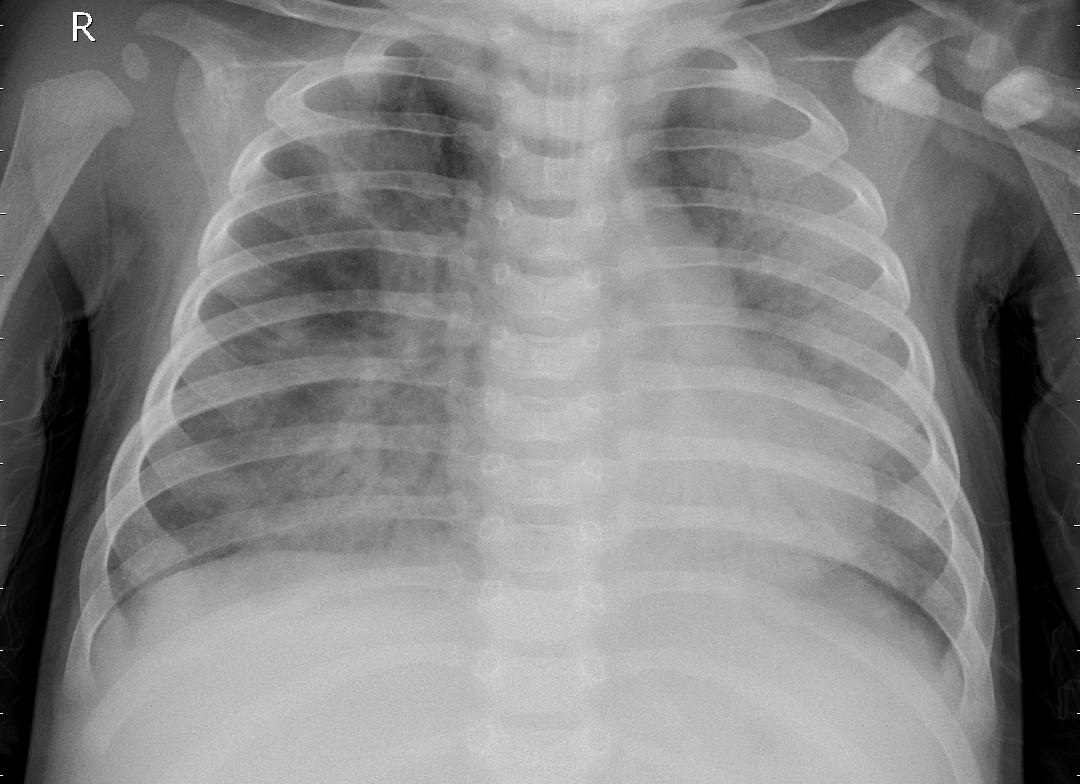
**Features:**

**X:**

[array([[[ 23, 23, 23], [ 33, 33, 33], [ 44, 44, 44], …, [117, 117, 117], [121, 121, 121], [131, 131, 131]], [[ 36, 36, 36], [ 62, 62, 62], [ 79, 79, 79], …, [121, 121, 121], [138, 138, 138], [126, 126, 126]], [[ 66, 66, 66], [ 87, 87, 87], [ 99, 99, 99], …, [129, 129, 129], [121, 121, 121], [121, 121, 121]], …, [[ 0, 0, 0], [ 0, 0, 0], [ 0, 0, 0], …, [ 30, 30, 30], [ 0, 0, 0], [ 0, 0, 0]], [[ 0, 0, 0], [ 0, 0, 0], [ 0, 0, 0], …, [ 27, 27, 27], [ 0, 0, 0], [ 0, 0, 0]], [[ 0, 0, 0], [ 0, 0, 0], [ 0, 0, 0], …, [ 23, 23, 23], [ 0, 0, 0], [ 0, 0, 0]]], dtype=uint8), array([[[ 56, 56, 56], [ 59, 59, 59], [ 68, 68, 68], ..., [124, 124, 124], [118, 118, 118], [117, 117, 117]], [[ 56, 56, 56], [ 62, 62, 62], [ 70, 70, 70], …, [124, 124, 124], [120, 120, 120], [117, 117, 117]], [[ 54, 54, 54], [ 62, 62, 62], [ 69, 69, 69], …, [125, 125, 125], [123, 123, 123], [119, 119, 119]], ...,

**Y:**

[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0,0,0,0,0,1,1,0,0,0,0……………….]

Normal Pneumonic

[8]

**3.2 DATA CLEANING**

Data quality has become an important issue. This issue becomes more and more important, where the need for effective decision making is high. In this context, the need for data cleaning to improve data quality is becoming crucial. Duplicate records elimination is a challenging data cleansing task. Here, we present a duplicate records elimination approach to improve the quality of data. We propose a deep learning-based approach for duplicate records detection using a sentence embeddings model. Also, we propose an algorithm for duplicated records correction. Then, we apply the proposed duplicate records elimination approach to analyse the effect of data cleaning on the quality of decisions.

* In sum: There are 5,863 X-Ray images (JPEG) and 2 categories (Pneumonia/Normal).
* There are 4872 training images for pneumonia and 1341 training images for normal condition.
* There are 390 testing images for pneumonia and 234 testing images for normal condition.

**3.3 DATA AUGMENTATION**

Technique of artificially increasing the training set by creating modified copies of a dataset using existing data.

**3.4 DATA VISUALISATION**

There are 5,863 X-Ray images (JPEG) and 2 categories (Pneumonia/Normal).

There are 4872 training images for pneumonia and 1341 training images for normal condition. There are 390 testing images for pneumonia and 234 testing images for normal condition.

If pneumonia is present then value is 1 else it is 0.

**[9]**

**4. METHODOLOGY**

**4.1 LOGISTIC REGRESSION**

[Linear regression models](https://www.ibm.com/in-en/topics/linear-regression) are used to identify the relationship between a continuous dependent variable and one or more independent variables. When there is only one independent variable and one dependent variable, it is known as simple linear regression, but as the number of independent variables increases, it is referred to as multiple linear regression. For each type of linear regression, it seeks to plot a line of best fit through a set of data points, which is typically calculated using the least squares method.

Similar to linear regression, logistic regression is also used to estimate the relationship between a dependent variable and one or more independent variables, but it is used to make a prediction about a categorical variable versus a continuous one. A categorical variable can be true or false, yes or no, 1 or 0, et cetera. The unit of measure also differs from linear regression as it produces a probability, but the logit function transforms the S-curve into straight line.

There are three types of logistic regression models, which are defined based on categorical response.

* **Binary logistic regression**: In this approach, the response or dependent variable is dichotomous in nature—i.e. it has only two possible outcomes (e.g. 0 or 1).Within logistic regression, this is the most commonly used approach, and more generally, it is one of the most common classifiers for binary classification.
* **Multinomial logistic regression:** In this type of logistic regression model, the dependent variable has three or more possible outcomes; however, these values have no specified order. A multinomial logistic regression model can help the studio to determine the strength of influence a person's age, gender, and dating status may have on the type of film that they prefer.
* **Ordinal logistic regression:** This type of logistic regression model is leveraged when the response variable has three or more possible outcome, but in this case, these values do have a defined order. Examples of ordinal responses include grading scales from A to F or rating scales from 1 to 5.

Terminologies involved in Logistic Regression:

* **Independent variables:** The input characteristics or predictor factors applied to the dependent variable’s predictions.
* **Dependent variable:**The target variable in a logistic regression model, which we are trying to predict.
* **Logistic function:** The formula used to represent how the independent and dependent variables relate to one another. The logistic function transforms the input variables into a probability value between 0 and 1, which represents the likelihood of the dependent variable being 1 or 0. **[10]**
* **Coefficient:** The logistic regression model’s estimated parameters, show how the independent and dependent variables relate to one another.
* **Intercept:** A constant term in the logistic regression model, which represents the log odds when all independent variables are equal to zero.
* **Maximum likelihood estimation:**The method used to estimate the coefficients of the logistic regression model, which maximizes the likelihood of observing the data given the model.

**RESULTS:**

Accuracy: 0.906441717791411

**[11]**

**4.2 K-NEAREST NEIGHBOR (KNN)**

* K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
* K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
* K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
* K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
* K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.
* It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
* KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

The K-NN working can be explained on the basis of the below algorithm:

* **Step-1**: Select the number K of the neighbors
* **Step-2**: Calculate the Euclidean distance of K number of neighbors
* **Step-3**: Take the K nearest neighbors as per the calculated Euclidean distance.
* **Step-4**: Among these k neighbors, count the number of the data points in each category.
* **Step-5**: Assign the new data points to that category for which the number of the neighbor is maximum.
* Step-6: Our model is ready.

**ADVANTAGES OF KNN**

* It is simple to implement.
* It is robust to the noisy training data
* It can be more effective if the training data is large.

[12]

**DISADVANTAGES OF KNN**

* Always needs to determine the value of K which may be complex some time.
* The computation cost is high because of calculating the distance between the data points for all the training samples.

Below are some points to remember while selecting the value of K in the K-NN algorithm:

* There is no particular way to determine the best value for “K”, so we need to try some values to find the best out of them. The most preferred value for K is 5.
* A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model.
* Large values for K are good, but it may find some difficulties.

**RESULTS:**

predicted value for training value - 0.9220347648261759

predicted value for testing value - 0.9164110429447853

[13]

**4.3 SUPPORT VECTOR MACHINE (SVM)**

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

SVM can be of two types:

* **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
* **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

**Hyperplane:**

There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM.

The dimensions of the hyperplane depend on the features present in the dataset, which means if there are 2 features (as shown in image), then hyperplane will be a straight line. And if there are 3 features, then hyperplane will be a 2-dimension plane.

We always create a hyperplane that has a maximum margin, which means the maximum distance between the data points.

**Support Vectors:**

The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector.

[14]

## **ADVANTAGES OF SVM**

* Effective in high-dimensional cases.
* Its memory is efficient as it uses a subset of training points in the decision function called support vectors.
* Different kernel functions can be specified for the decision functions and its possible to specify custom kernels.

**RESULTS:**

Training Accuracy - 0.9930981595092024

Testing Accuracy - 0.9754601226993865

[15]

**4.4 DECISION TREE**

* Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
* In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
* The decisions or the test are performed on the basis of features of the given dataset.
* It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
* It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
* In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.
* A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision tree:

* Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
* The logic behind the decision tree can be easily understood because it shows a tree-like structure.

**DECISION TREE TERMINOLOGIES**

* **Root Node:** Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.
* **Leaf Node:** Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.
* **Splitting:** Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.
* **Branch/Sub Tree:** A tree formed by splitting the tree.
* **Pruning:** Pruning is the process of removing the unwanted branches from the tree.

**[16]**

* **Parent/Child node:**The root node of the tree is called the parent node, and other nodes are called the child nodes.

**Steps for decision tree**

* **Step-1:** Begin the tree with the root node, says S, which contains the complete dataset.
* **Step-2:** Find the best attribute in the dataset using **Attribute Selection Measure (ASM).**
* **Step-3:** Divide the S into subsets that contains possible values for the best attributes.
* **Step-4:** Generate the decision tree node, which contains the best attribute.
* **Step-5:** Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

**RESULTS:**

Training Accuracy - 1.0

Testing Accuracy - 0.8964723926380368

[17]

**4.5 RANDOM FOREST**

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, “Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.” Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

* There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
* The predictions from each tree must have very low correlations.

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

**Step-1:** Select random K data points from the training set.

**Step-2:** Build the decision trees associated with the selected data points (Subsets).

**Step-3:** Choose the number N for decision trees that you want to build.

**Step-4:** Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

**RESULTS:**

Training Accuracy - 1.0

Testing Accuracy - 0.9562883435582822

[18]

**5.RESULTS AND DISCUSSIONS**

The accuracy value for decision tree and random forest are more accurate than other machine learning models .So we can say that pneumonia detection values can be obtained accurately by using decision tree and random forest methods. And the others methods obtain around 99% accuracy score.

The overall accuracy values are :

**LOGISTIC REGRESSION:**

Total Accuracy - 0.906441717791411

**KNN:**

Training Accuracy - 0.9220347648261759

Testing Accuracy - 0.9164110429447853

**SVM:**

Training Accuracy - 0.9930981595092024

Testing Accuracy - 0.9754601226993865

**DECISION TREE:**

Training Accuracy - 1.0

Testing Accuracy - 0.8964723926380368

**RANDOM FOREST:**

Training Accuracy - 1.0

Testing Accuracy - 0.9562883435582822

[19]

**6.CONCLUSION AND FUTURE SCOPE**

In conclusion, various machine learning algorithms can be used to predict pneumonia accurately. Some of the commonly used algorithms include decision trees, support vector machines, artificial neural networks, and k-nearest neighbors. These algorithms use different techniques and methods to classify the data and make predictions.

The performance of these algorithms can be evaluated based on various metrics such as accuracy, sensitivity, specificity, and F1-score. The choice of the best algorithm for pneumonia prediction depends on the nature of the data, the size of the dataset, and the available computational resources.

The future scope for prediction of pneumonia using machine learning techniques is promising. With the increasing availability of digital healthcare data and advancements in machine learning algorithms, new predictive models can be developed to improve the accuracy of pneumonia diagnosis.

Overall, with the help of machine learning algorithms, it is possible to effectively diagnose pneumonia and potentially reduce the morbidity and mortality rates associated with this disease.

**[20]**

**7.REFERENCES**

[1]WHO Pneumonia World Health Organization. [DB/OL] https://www.who.int/news-room/factsheets/detail/pneumonia. 2022-06-12

[2] UNICEF for every child. [DB/OL]

https://data.unicef.org/topic/childhealth/pneumonia. 2022-06-12

[3] Dejun Zhang, FuquanRen, Yushuang Li, et al.Pneumonia Detection from Chest X-ray Images Based on Convolutional Neural Network [J]. Electronics. 2021, 10: 1512.1-17.

https://doi.org/10.3390/electronics10131512

[4] DimpyVarshni, KartikThakral,Lucky Agarwal, et al. Pneumonia Detection Using CNN based FeatureExtraction [C], In: 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT). 2019-02-22

[5] RohitKundu, Ritacheta Das,Zong Woo Geem, et al. Pneumonia detection in chest X-ray images using an ensemble of deep learning models [J].PLOS ONE.Online.2021-09-07. doi: 10.1371/journal.pone.0256630

[6] V. SirishKaushik,AnandNayyar,GauravKataria, et al. Pneumonia Detection Using Convolutional Neural Networks (CNNs) [C]. In: Proceedings of First International Conference on Computing, Communications, and Cyber-Security (IC4S 2019),2019: 471–483

[7] M. Elkhodr, B. Alsinglawi, M. Alshehri. A privacy riskassessment for the internet of things in healthcare [OL]. Applications of Intelligent Technologies in Healthcare, Springer,Berlin, Germany, 2018-11-10: 47-54.

[8] YamashitaR, NishioM, DoRKG, et al. Convolutional neural networks: an overview and application in radiology [J]. Insights Imaging, 2018, 9: 611–629.

https://doi.org/10.1007/s13244-018-0639-9

[21]