Task 2 Business understanding.

Identifying your business goals

Background

Chest X-rays are one of the most common methods to diagnose a variety of lung diseases such as COVID-19, tuberculosis, or pneumonia. Traditional diagnostic methods are usually time-consuming and highly dependent on the experience and availability of medical personnel. The application of machine learning can automatically detect diseases much faster than the traditional diagnosis methods to reduce errors and improve patient care.

Business Goals

Creating a model that helps to identify diseases from X-ray images.

Accelerate the diagnosis of lung diseases by providing real-time results.

To support areas without access to experienced radiologists.

Business Success Criteria

Model accuracy exceeds 85% in all categories.

The tool is easy to use.

Project success measure: the model is working and able to accurately classify chest X-ray images into the correct categories with a minimum accuracy of 85% on unseen test data.

Assessing Your Situation

Inventory of Resources

Data: A dataset of chest X-ray images with categories of COVID-19, tuberculosis, Pneumonia and Normal.

Technical resources: TensorFlow framework, Jupyter Notebook.

Team skills: Machine learning, image processing, data analysis.

Requirements, Assumptions, and Constraints

Requirements: Images must be of standard size (224x224 pixels).

Assumptions: The dataset is large and diverse enough for the model to be able to adapt to real data.

Limitations: Computational resources are limited (the work is done on a personal computer), which may affect model complexity and training time.

Risks and contingency plans

Risks:

Poor model performance (The model won’t achieve 85% accuracy).

Model training issues.

Computational resources.

Backup plans:

Optimize model architecture or try different models.

Use pre-trained models.

Use cloud-based solutions to train the model.

Terminology

COVID-19: An infectious disease caused by the SARS-CoV-2 virus.

Tuberculosis: An infectious disease caused by bacteria

Pneumonia: Inflammation and fluid in your lungs caused by a bacterial or viral infection.

Costs and Benefits

Costs: Computational resources, time required to process data and train the model.

Benefits: Speeding up and improving availability of medical diagnostics.

Defining Your Data-Mining Goals

Data-Mining Goals

Train a ML model that can identify the categories from the chest X-ray images.

Achieve at least 85% accuracy in all categories.

Test the model on new data and validate its performance.

Data-Mining Success Criteria

Accuracy: Accuracy of model predictions exceeds 85%.

Generalizability: The model can generalize to new, previously unseen data.

Performance: The model runs at a reasonable speed and is optimized for low computing power.

Task 3 Data understanding

1. Gathering Data

Outline Data Requirements
The dataset aims to support the development of machine learning models for classifying respiratory conditions into four categories:

Pneumonia
Tuberculosis (TB)
Corona (COVID-19)
Normal (healthy individuals)
Key data requirements include:

Labeling: Each X-ray image must be categorized into one of the four classes.
Image Format: Images should be in .jpg or .png formats to ensure compatibility with machine learning frameworks.
Image Resolution: Images may have variable resolutions but should be resized to uniform dimensions (e.g., 224x224 pixels) for preprocessing and model training.
Diversity: The dataset should include representative samples of different severities and variations within each class to ensure robust model performance.
Data Volume: Adequate samples across all four categories to prevent data imbalance issues.

Verify Data Availability
The dataset is already organized into directories based on the four classes. Initial review confirms that the data is available and structured appropriately. However, further steps are needed to verify:

The total number of images per category.
The quality of the images (e.g., clarity, lack of corruption).
Metadata or annotations accompanying the images, if available.

Define Selection Criteria
Inclusion Criteria:
Images must belong to one of the four classes.
Files should be readable and not corrupted.
Exclusion Criteria:
Duplicate or mislabeled images.
Images with excessive noise or poor quality that might interfere with analysis.

2. Describing Data

The dataset consists of X-ray images stored in four directories based on the medical condition:

Pneumonia:

Features: Lung inflammation, fluid accumulation.
Visual cues: Hazy or cloudy regions in the lungs.
Clinical relevance: Represents bacterial or viral lung infections, providing distinct visual patterns compared to other conditions.
Tuberculosis (TB):

Features: Granulomas, nodules, scarring.
Visual cues: Localized abnormalities often concentrated in specific lung areas.
Clinical relevance: Important for detecting chronic bacterial infections in the lungs.
Corona (COVID-19):

Features: Ground-glass opacities, patchy consolidations.
Visual cues: Peripheral lung involvement and diffused patterns.
Clinical relevance: Helps differentiate COVID-19 from other respiratory infections, particularly viral pneumonia.
Normal:

Features: Clear lung structures, no visible abnormalities.
Visual cues: Uniform and symmetrical lung patterns.
Clinical relevance: Serves as the baseline class for healthy individuals.
The dataset is expected to include images of varying quality and resolution, which will require preprocessing.

3. Exploring Data

Initial Observations:

Class Distribution: It is crucial to check the number of samples in each class. If one class (e.g., COVID-19) has significantly fewer images, it may introduce bias into the model.
Visual Differences: The dataset includes distinct visual patterns (e.g., haziness for pneumonia, nodules for TB), which can aid classification.
Potential Exploration Steps:

Calculate the number of images in each class directory.
Visualize sample images from each class to confirm that they align with their descriptions.
Analyze image properties:
Resolution: Check for consistency or variations.
Brightness and contrast: Determine if normalization is necessary.

Challenges Identified:

Class imbalance: If one or more classes have fewer samples, it may require data augmentation.
Overlap in features: Pneumonia and COVID-19 share some visual similarities, making it challenging for the model to distinguish between them.

4. Verifying Data Quality

Completeness:

Check that all four classes are adequately represented.
Ensure no missing or incomplete files (e.g., broken image links).
Accuracy:

Verify that the images are correctly labeled. Mislabeling could significantly impact model performance.
Consistency:

Images should be in the specified formats (.jpg or .png).
Standardize image dimensions (e.g., resizing to 224x224 pixels).
Relevance:

All images must be relevant to the specified classes. Irrelevant or noisy data should be removed.
Steps for Data Cleaning:

Remove corrupted or low-quality images.
Standardize the image format and dimensions.
Address class imbalance by generating augmented images for underrepresented categories (e.g., flipping, rotating, or applying transformations).

Project Plan: Detection of COVID-19, NORMAL, PNEUMONIA, and TUBERCULOSIS Using Chest X-Rays

Detailed Plan

Task 1: Dataset Preparation

Load the Kaggle dataset, verify its structure, and split it into training, validation, and test sets. The dataset is already preprocessed and categorized into four classes: Pneumonia, Tuberculosis, Corona, and Normal.

- **Time Allocation**: Mairon (3 hours), (2 hours), Raigo (2 hours).
- **Total**: 7 hours.

Task 2: Exploratory Data Analysis (EDA)

Perform statistical and visual analyses of the dataset. Identify class distributions, inspect sample images, and check for any biases or imbalances.

- **Time Allocation**: Mairon (7 hours), Marttis (7 hours), Raigo (7 hours).
- **Total**: 21 hours.

Task 3: Model Development and Training

Train a classification model using transfer learning with pre-trained CNN architectures (e.g., ResNet). Fine-tune hyperparameters and monitor performance metrics during training.

- **Time Allocation**: Mairon (11 hours), Marttis (11 hours), Raigo (11 hours).
- **Total**: 33 hours.

Task 4: Model Evaluation and Optimization

Evaluate the model on the test set using metrics like accuracy, F1-score, and confusion matrix. Address class-specific performance issues through optimization techniques.

- **Time Allocation**: Mairon (7 hours), Marttis (7 hours), Raigo (7 hours).
- **Total**: 21 hours.

Task 5: Report Writing and Poster

Document the methodology and results, and create poster for an engaging presentation.

- **Time Allocation**: Mairon (5 hours), Marttis (5 hours), Raigo (5 hours).
- **Total**: 15 hours.

Tools and Methods

- **Tools**: Python (TensorFlow, Keras), Jupyter Notebook, Matplotlib for visualization, PowerPoint for presentations.
- **Methods**: Transfer learning, CNN classification, and data analysis.

Regular check-ins will ensure smooth project execution.