**NAME : Kishor P Patil**

**Roll No: 391042**

**Batch : A2**

**PRN : 22210749**

**Mini Project Report**

**Title:** *BioBERT for Clinical Insight: Multi-Class Disease Classification from Symptom Descriptions*

**1. Objective**

The main goal of this mini project was to build a disease classification system using **BioBERT**, a biomedical language model. The model takes free-text symptom descriptions as input and predicts the most likely disease from a set of 31 diseases across three domains: **Neurological**, **Respiratory**, and **ENT (Ear, Nose, Throat)**.

**2. Dataset and Preprocessing**

* The dataset was sourced from **Hugging Face**, containing **256,000+** medical records.
* Only the **Disease Name** and **Symptom Description** columns were used.
* From the complete dataset, a subset of **30,815 records** was created, covering 31 diseases across:
  + Neurological & Psychological (10,937 samples)
  + Respiratory (10,379 samples)
  + ENT (9,499 samples)

**Preprocessing steps:**

* Removed unwanted characters, converted text to lowercase.
* Tokenized and padded symptom text using **BioBERT tokenizer**.
* Encoded disease names into numerical labels for training.

**3. Model Used: BioBERT**

BioBERT is a pre-trained transformer model designed for biomedical text. Here's how it was used:

* **Input:** Token embeddings + positional embeddings + segment embeddings .
* **Architecture:**
  + 12-layer Transformer encoder.
  + Multi-head self-attention for capturing symptom patterns.
  + Feedforward neural network for non-linear transformation.
* **Output:** Classification layer that predicts one of 31 disease classes.

**4. Training and Evaluation**

* Loss Function: Cross-Entropy
* Optimizer: AdamW
* Metrics used: Accuracy, Precision, Recall, F1-Score

**Results:**

* Training Accuracy: **91.6%**
* Testing Accuracy: **89.7%**
* Precision: **88.2%**
* Recall: **88.6%**
* F1 Score: **88.4%**

Compared to traditional models like Decision Trees and SVMs, BioBERT showed much better performance in understanding and classifying symptom descriptions.

**5. Tools & Libraries Used**

* Python, Jupyter Notebook
* Hugging Face Transformers
* PyTorch
* Scikit-learn (for evaluation metrics)

**6. Conclusion**

This project showed how BioBERT can be fine-tuned to classify diseases from symptom descriptions with high accuracy. It successfully captures complex symptom patterns, making it useful for clinical decision support systems or initial symptom checkers.