1. Dataset Understanding

1.1 Structure of Dataset

<u>Key Insights</u>

The dataset has -
6 columns: 1 integer (index), 4 text-based features (description, sample_name, transcription, keywords), and 1 categorical feature (medical_specialty).

4999 Rows representing individual medical records, with each row detailing a patient case, its specialty, and associated transcription.

<u>Features</u>

A. description : A **brief summary** or reason for the medical visit. May contain concise context.
B. medical_speciality : The **target label**: the medical department or specialty (e.g., Allergy, Bariatrics). This is what we want to predict.
C. sample_name : A **name or title** for the transcription (possibly a human-readable name, not very useful).
D. transcription : The **full detailed transcription** of the medical note/report. This will the **main input** to the model.
E. keywords : A list of **keywords** related to the note and specialty. May help with feature engineering or label enhancement.

1.2 EDA on Dataset

<u>Unique classification label</u>

Their are 40 unique classification label comprising of -

```
[' Allergy / Immunology', ' Bariatrics',
       ' Cardiovascular / Pulmonary', ' Neurology', ' Dentistry',
       ' Urology', ' General Medicine', ' Surgery', ' Speech - Language',
       ' SOAP / Chart / Progress Notes', ' Sleep Medicine',
       ' Rheumatology', ' Radiology', ' Psychiatry / Psychology',
       ' Podiatry', ' Physical Medicine - Rehab',
       ' Pediatrics - Neonatal', ' Pain Management', ' Orthopedic',
       ' Ophthalmology', ' Office Notes', ' Obstetrics / Gynecology',
       ' Neurosurgery', ' Nephrology', ' Letters',
       ' Lab Medicine - Pathology', ' IME-QME-Work Comp etc.',
       ' Hospice - Palliative Care', ' Hematology - Oncology',
       ' Gastroenterology', ' ENT - Otolaryngology', ' Endocrinology',
       ' Emergency Room Reports', ' Discharge Summary',
       ' Diets and Nutritions', ' Dermatology',
       ' Cosmetic / Plastic Surgery', ' Consult - History and Phy.',
       ' Chiropractic', ' Autopsy']
```

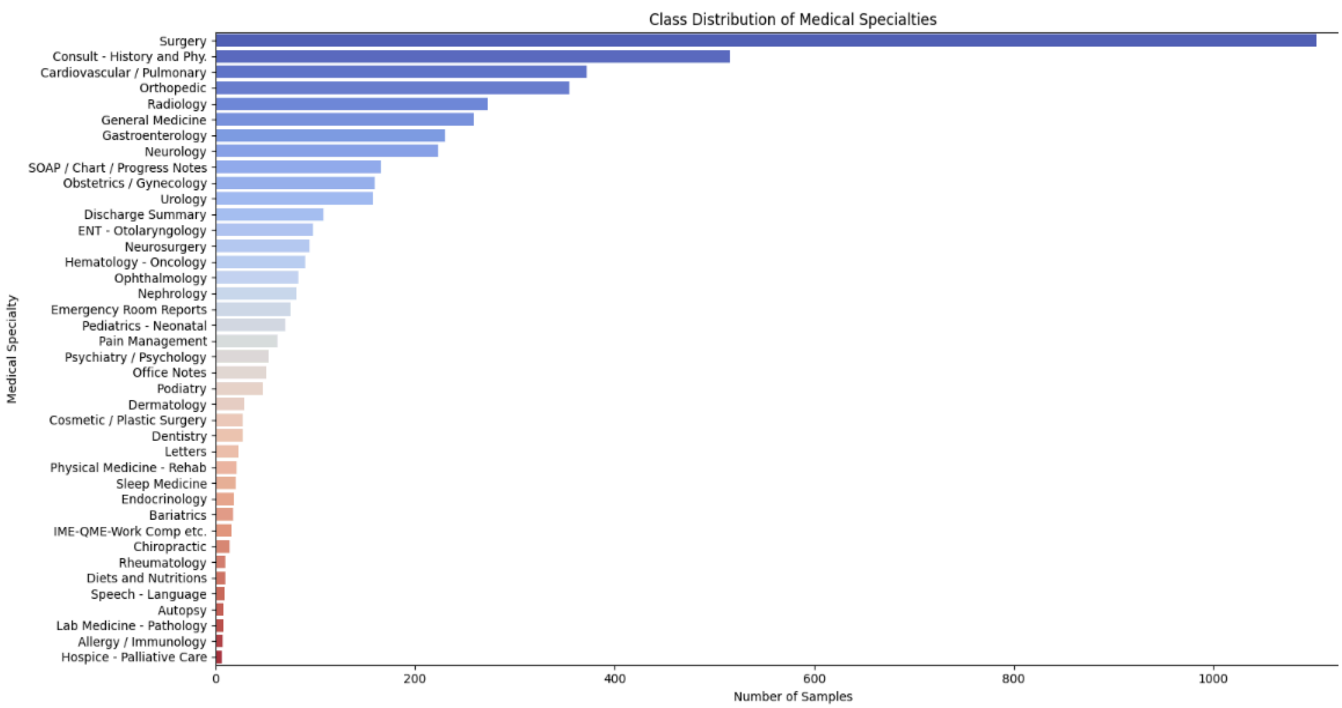
Identifying category of classification task single label or multi label classification task

```
#checking multilabel classification or single lable classification
df['medical_specialty'].apply(lambda x: ',' in x).sum()
```

```
0
```

It means it's a single lable classification task

Class Distribution of Classification label for bias free fine tuning



Key Insights

Imbalanced Dataset

The dataset is significantly **imbalanced**. A few specialties such as Surgery, Consult - History and Physical, and Cardiovascular / Pulmonary account for a large portion of the dataset, while many other classes such as Hospice - Palliative Care, Autopsy, and Allergy / Immunology have very few samples.

This imbalance may cause the model to be biased toward predicting the majority classes, hampering the evaluation metrics like F1-score or precision-recall in addition to accuracy.

My Approach to this problem

This will be addressed at model fine-tuning using class-weighted loss or possibly data augmentation after establishing a good baseline.

2. Data Preprocessing

2.1 Text Cleaning

Applied basic preprocessing on transcription, description and keyword

- Converted to lowercase.
- Removed special characters and extra whitespace.
- Removed nulls and empty text entries.

2.2 Handling Missing Values & Duplicates

- Dropped rows with missing transcriptions or labels.
- Removed duplicate transcriptions.
- Filled missing keywords with empty strings.

2.3 Label Cleaning & Consolidation

- Stripped and lowercase medical_speciality.
- Rare classes (appearing <2 times) were grouped into a new label: ‘other’

2.4 Feature Engineering

Constructed a new feature full_text by concatenating cleaned:

1. description
2. transcription
3. keywords

This helped enrich the context for training.

2.5 Train-Validation Split

3. Model Training and Fine-Tuning

I tried multiple approaches were tested to evaluate the effectiveness of various models both classical and transformer-based on the domain-specific medical transcription dataset.

3.1 Without Data Augmentation

Model	Description	Accuracy	Macro F1 Score
TF-IDF + SVM	Classical model using TF-IDF vectorized full_text and SVM classifier	0.8280	0.5647
BERT CLS + Logistic Regression	Extracted [CLS] token from frozen BERT, used with logistic regression	0.7219	0.3866
BERT CLS + SVM	Same as above, used SVM instead of logistic regression	0.7091	0.4079
Fine-tuned BERT	BERT fine-tuned with classifier head using Hugging Face Transformers	0.6985	0.3694

3.2 With Data Augmentation

Model	Description	Accuracy	Macro F1 Score
TF-IDF + SVM	Classical model retrained on oversampled data	0.8259	0.6144
BERT CLS + SVM	SVM classifier on frozen BERT CLS vectors with balanced data	0.6964	0.3647

3.3 Reasoning Behind Model Choices

Aspect	**Justification**
TF-IDF + SVM	Strong baseline for text classification, interpretable, fast to train.
BERT CLS (Frozen)	Leveraging rich semantic embeddings without fine-tuning; fast, avoids overfitting
Fine-tuned BERT	Adapt model weights to domain-specific language; allows end-to-end learning.
Oversampling	Address severe class imbalance by increasing minority class representation

3.4 Key Findings

1. TF-IDF + SVM achieved the highest accuracy and F1 without tuning large models.
2. Frozen BERT models performed worse than classical TF-IDF, suggesting that pretrained representations may not align well with this domain without adaptation.
3. End-to-end fine-tuning of BERT unexpectedly underperformed, likely due to:
    - Small dataset size
    - Overfitting
    - High class imbalance
4. Data Augmentation via oversampling improved F1 scores, particularly for TF-IDF-based models, indicating better minority class recognition.

4. Challenges Faced & Solutions

Challenge	**Solution**
Severe class imbalance	Performed oversampling using target fraction (~0.05)
BERT fine-tuning overfitting	Reduced epochs, smaller LR, used dropout, still struggled due to data size
Feature sparsity in minority classes	Merged rare classes into '**other**', then oversampled to stabilize
Slow training times	Used frozen embeddings + external classifiers (Logistic, SVM) for efficiency

5. Conclusion

1. Classical TF-IDF with SVM remained the most reliable and consistent performer in this scenario.

2. Pretrained models require larger labeled data or more aggressive domain adaptation (e.g., using BioBERT).

3. Fine-tuning alone is not sufficient when data is both small and imbalanced augmentation and regularization are crucial.

4. The architecture can be extended to incorporate external domain-specific language models (like ClinicalBERT) for better results in future iterations.