Disease diagnosis using classification and NLP

Rebecca Ceppas de Castro, Fulya Tastan, Philip Barron, Mohammad Rafiqul Islam, Nina Adhikari, Viraj Meruliya

Motivation

Patient comes into a hospital feeling unwell.

Healthcare workers ask questions about symptoms and history,

?

narrow down the possible diagnoses,

conduct tests, and form a diagnosis

and begin treatment

Stakeholders

- Medical professionals
- Epidemiology experts
- Health organizations
- Data scientists and ML researchers
- End users (patients)

Dataset

We use a **synthetic** dataset provided in *"DDXPlus: A New Dataset for Automatic Medical Diagnosis"* [1]

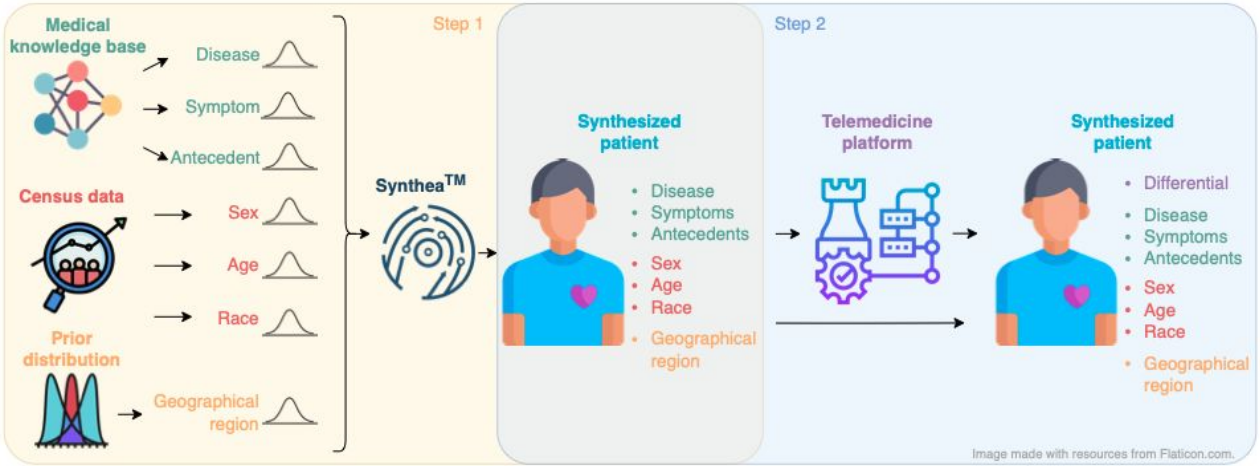


Figure 1: Overview of the data generation process of DDXPlus.[2]

- 1.3 million patients
- 49 types of pathologies
- 110 symptoms
- 113 antecedents

1. https://figshare.com/articles/dataset/DDXPlus_Dataset/20043374
2. https://arxiv.org/pdf/2205.09148

Dataset description

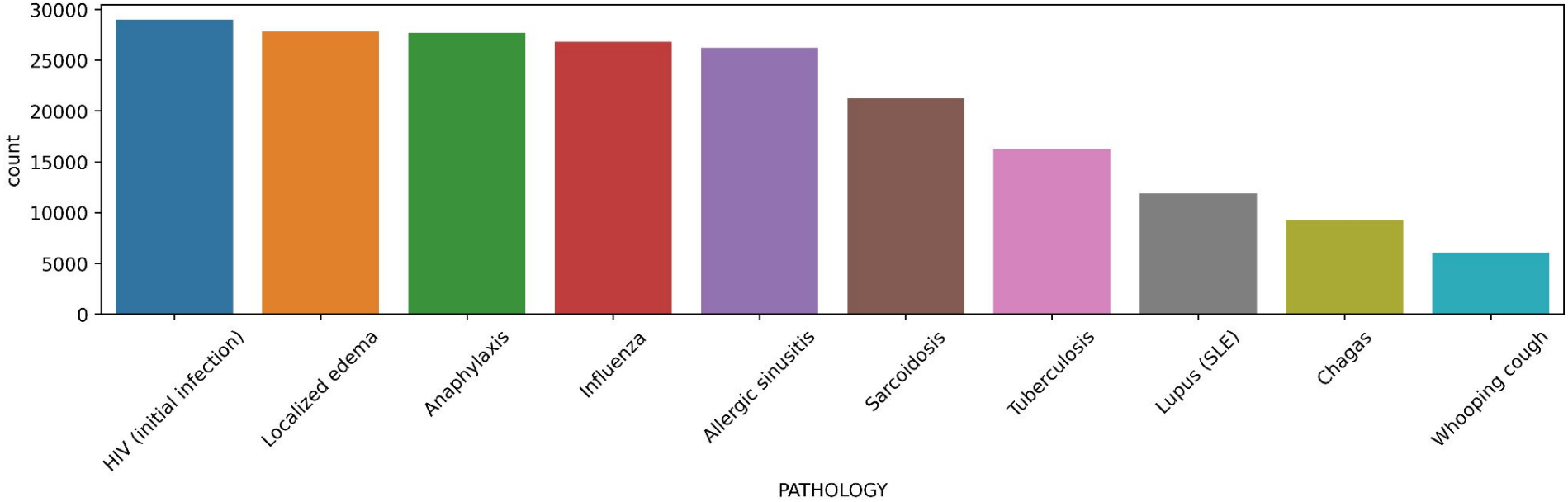
The columns of this dataset include a mixture of binary, multiple-choice, and categorical-type questions used to find symptoms and antecedents of the patients.

Our dataset provides a differential diagnosis, as well as the true pathology for each patient.

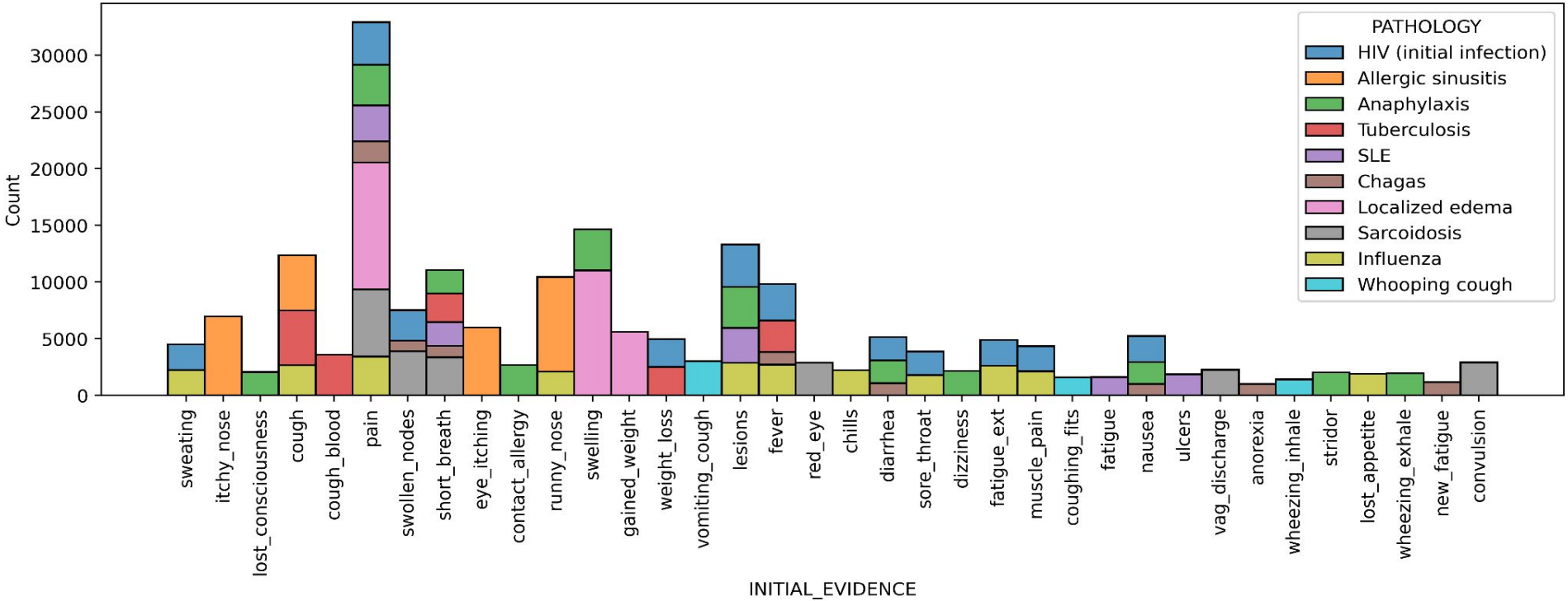
- We restricted our exploration to 10 immune and infectious diseases.

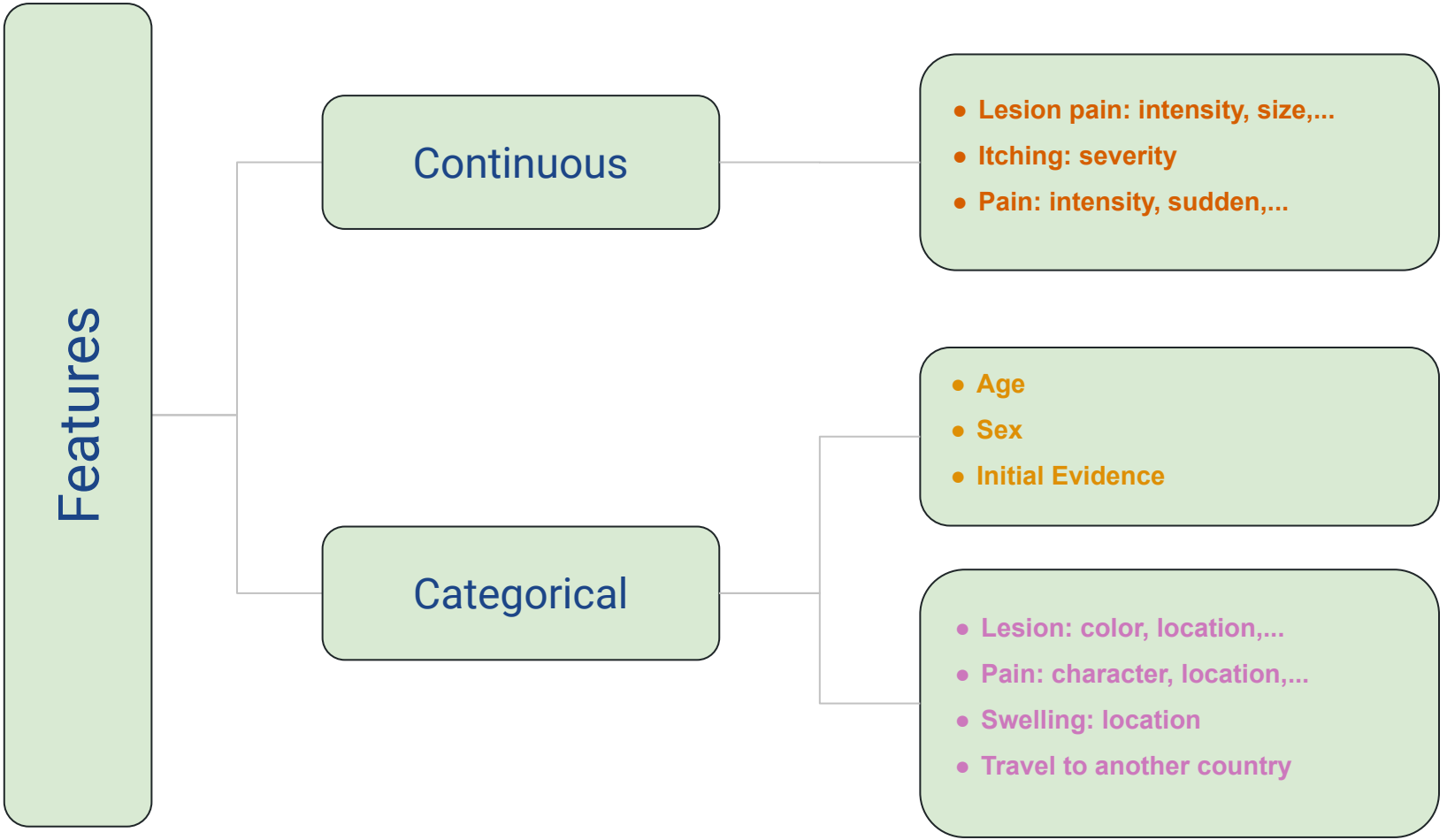
EDA / Visualization

- For these particular subset of diseases, the dataset contains ~200K rows.
- Distribution of these diseases:



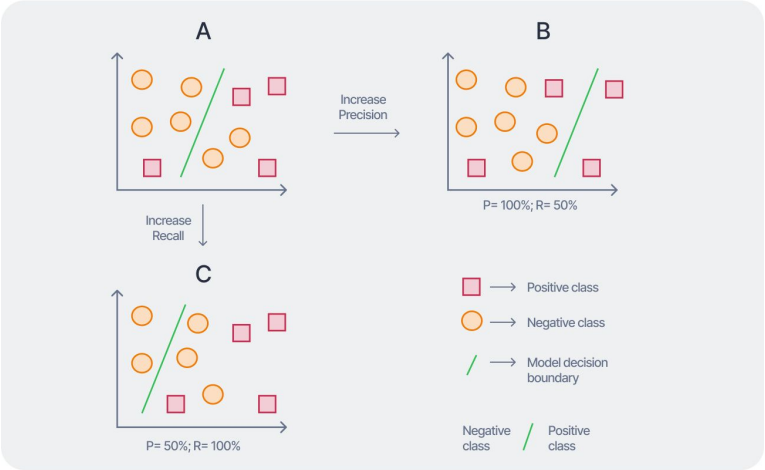
- Initial Evidence is one of the first inputs taken from patient
- Build a model based on this input + further questions to pin down the disease.



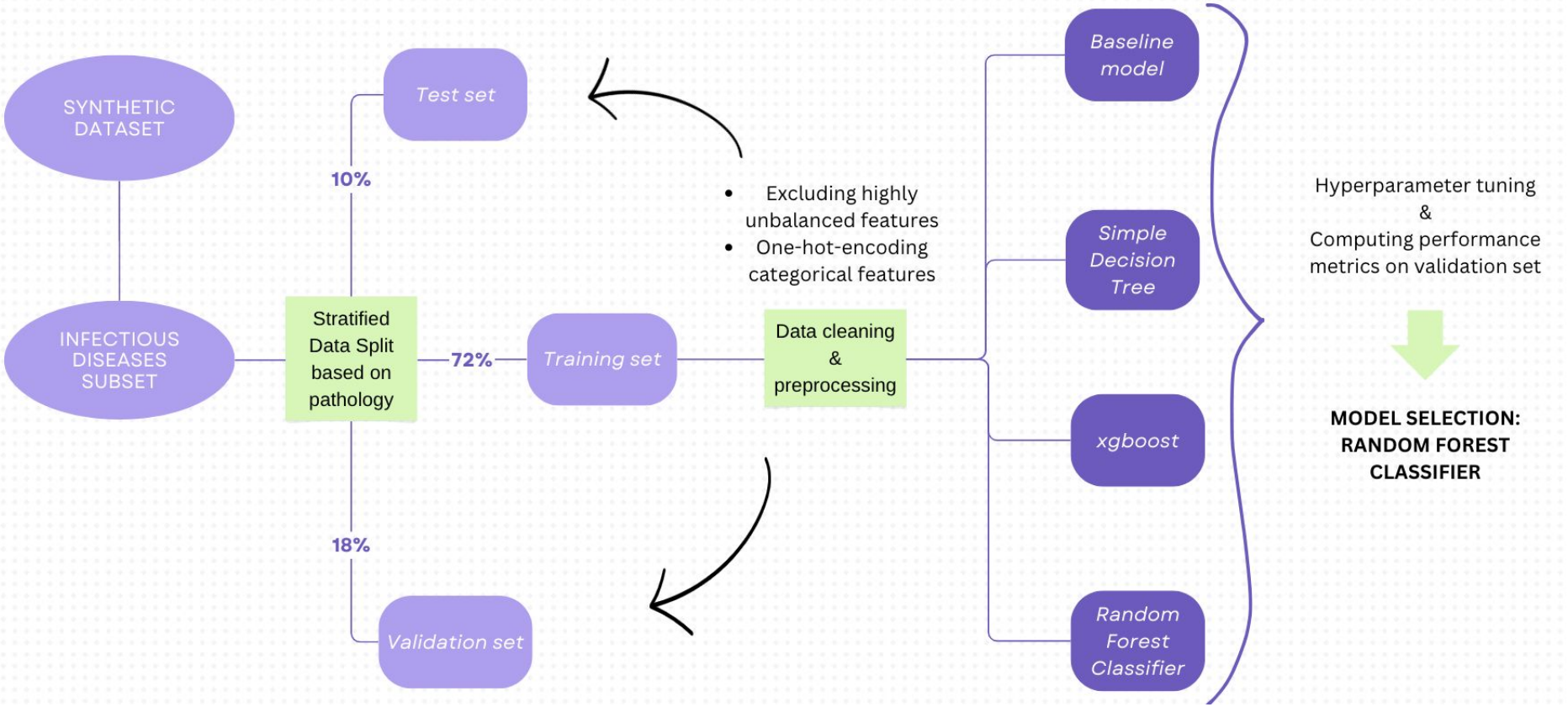

Features
Continuous
Categorical
Lesion pain: intensity, size,...
Itching: severity
Pain: intensity, sudden,...
Age
Sex
Initial Evidence
Lesion: color, location,...
Pain: character, location,...
Swelling: location
Travel to another country

Performance metrics

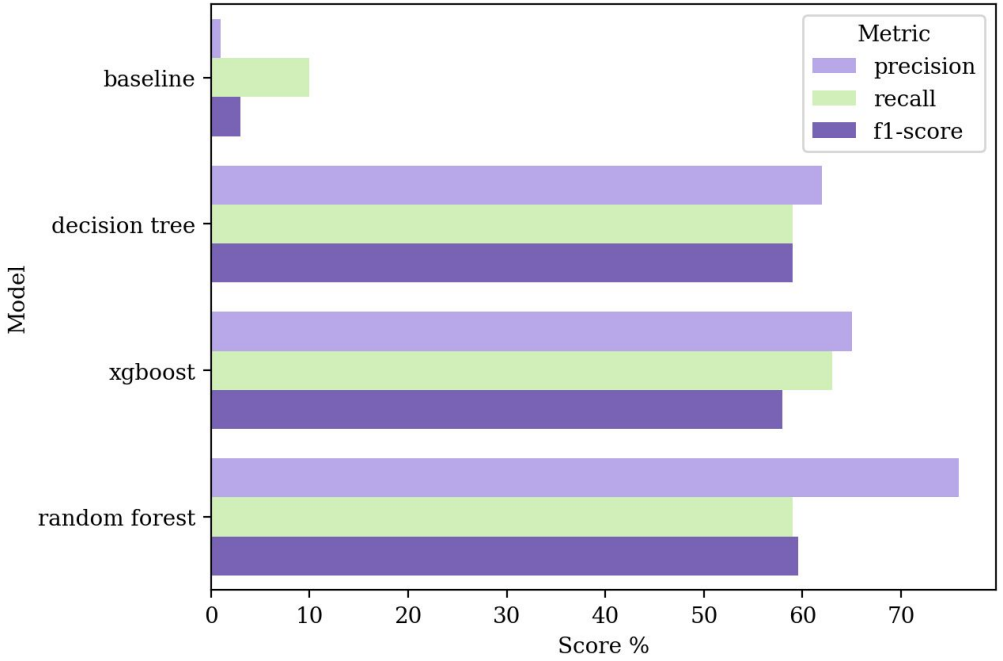
- Individual disease scores + average scores for the entire dataset
- Recall: given that a patient has a certain disease, how often does the model diagnose it correctly?
- Precision: what fraction of the positive diagnosis are correct?
- **f1-score:** what is the accuracy of the (unbalanced) classification model



www.v7labs.com


SYNTHETIC DATASET
INFECTIOUS DISEASES SUBSET
Stratified Data Split based on pathology
10%
Test set
72%
Training set
18%
Validation set
Excluding highly unbalanced features
One-hot-encoding categorical features
Data cleaning & preprocessing
Baseline model
Simple Decision Tree
xgboost
Random Forest Classifier
Hyperparameter tuning & Computing performance metrics on validation set
MODEL SELECTION: RANDOM FOREST CLASSIFIER

Model approach and selection



Xgboost & random forest models were comparable in f1-score (~ 58% and 60%).

Decided to go with random forest for interpretability of results.

Further feature selection, taking 50% top important features identified by initial random forest model.

K-fold cross validation on final model to test reliability.

Results

- Random Forest:

  F1 59.58%, Precision 75.83%, Recall 59.04%, and Accuracy 60.4%

Limitations

- Dataset limitations
- Inability to assign importance to symptoms other than “Initial evidence”

NLP Approach

- 2-3-sentence long paragraphs generated from the symptoms:

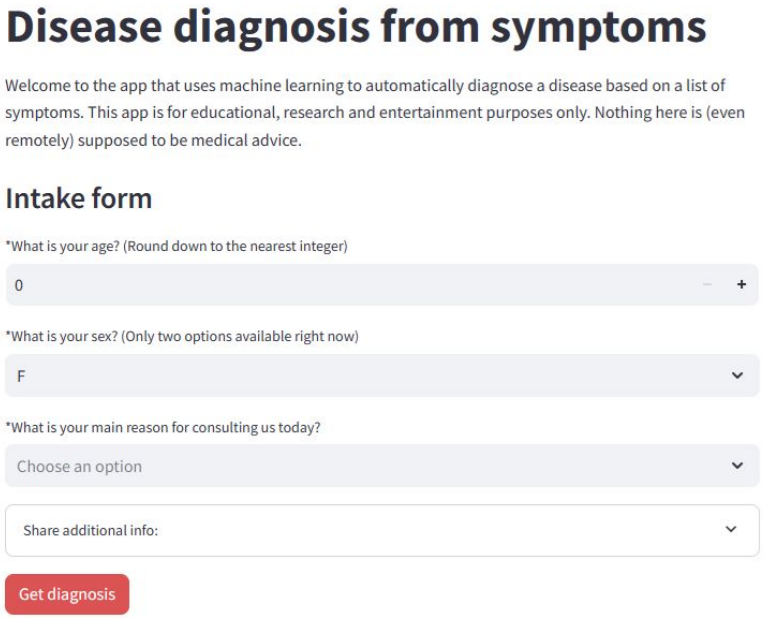
Sentence	Label
“I have symptoms of coughing up blood, increased sweating, exhausting pain in my forehead, skin lesions or rashes on my forehead, diffuse muscle pain, loss of appetite or early fullness, cough presence, nasal congestion or runny nose, and chills or shivers.”	Tuberculosis
“My 2-year-old daughter has a cough, increased sweating, nasal congestion, pain in the forehead that is exhausting, fever, extreme fatigue, sore throat, and loss of appetite with pink rashes on the forehead.”	Influenza

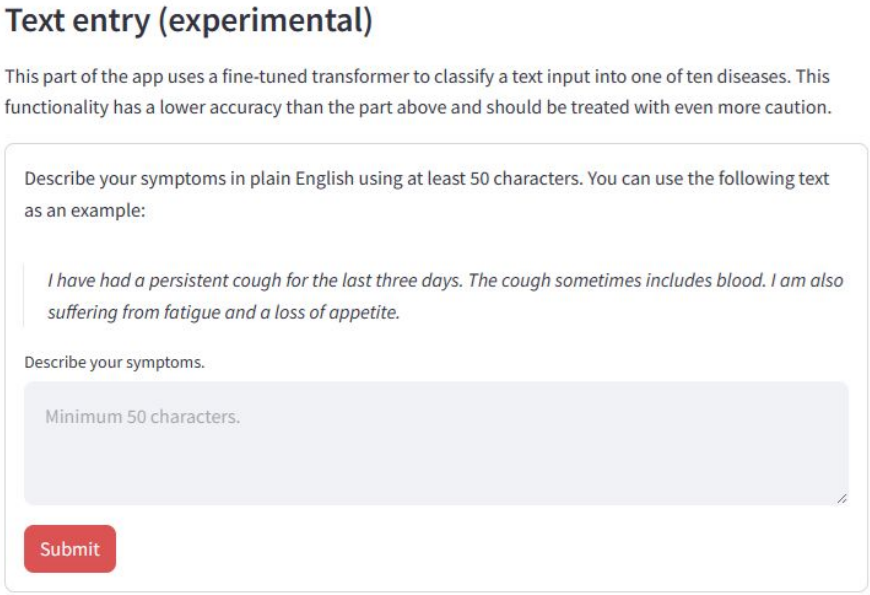
- Hugging Face’s DistilBERT transformer fine-tuned on these paragraphs

- Achieved a categorical accuracy of **58.68%**  on the test set

App

- Prototype of product available at disease-pred.streamlit.app





Future work

- Include more disease classes
- Other datasets, especially those based on real-world data
- Include image/voice input
- A chatbot that interacts with the user to ask follow-up questions

Thank you

to Roman Holowinsky, Alec Cott, Steven Gubkin and the Erdös Institute. We also appreciate the help from our mentor: Yuchen Luo, for her support in completing this project.