PATIENT DIAGNOSIS WITH MACHINE LEARNING

Caitlin Ortega Ruble

How diagnosis is done

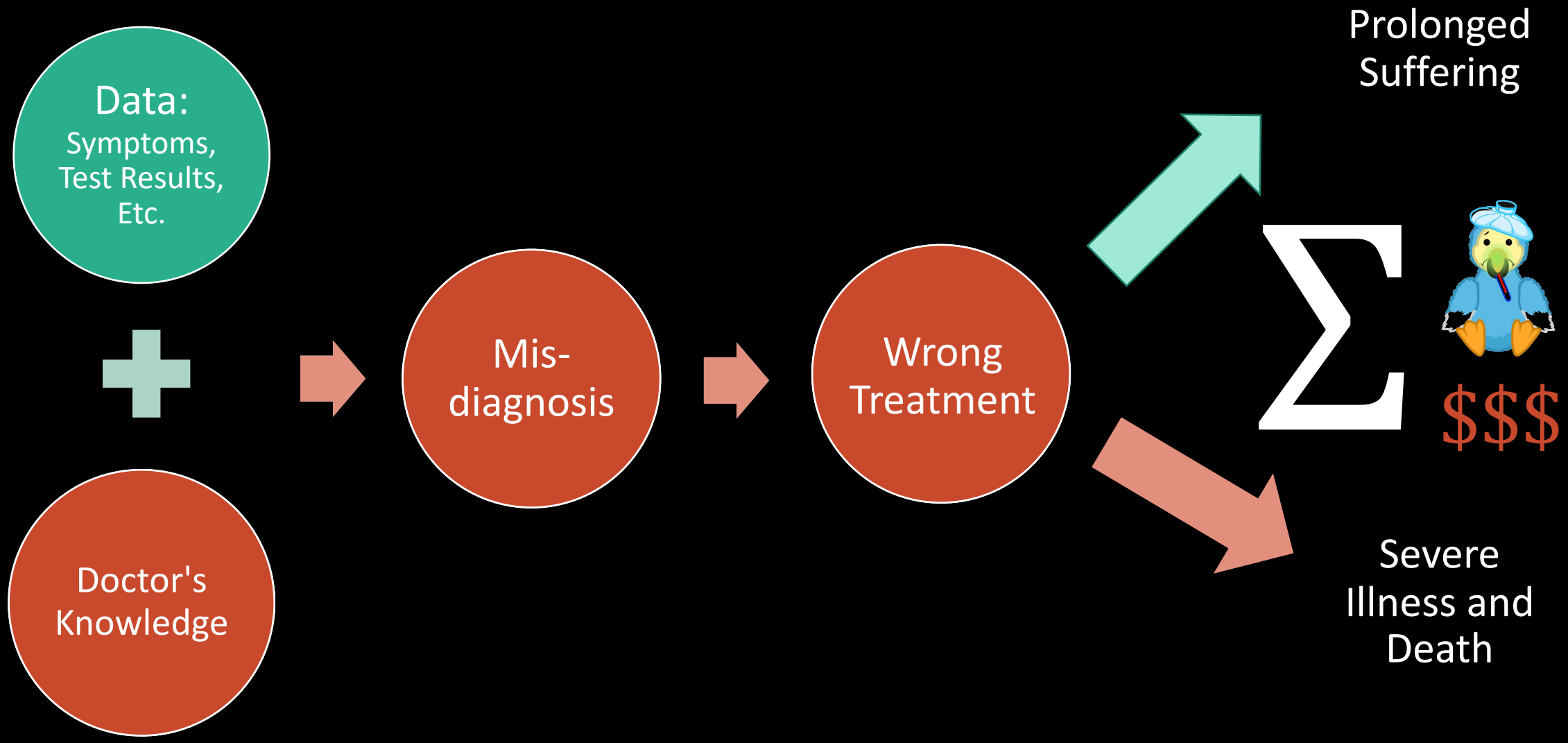
Data:
Symptoms,
Test Results,
Etc.

+

Doctor's
Knowledge

→ Diagnosis → Treatment

How can it go wrong?



What can we do?

Data:
Symptoms,
Test Results,
Etc.

+

Machine
Learning
Model

→ Right Diagnosis → Right Treatment → $$$

The data: "Disease Prediction Using Machine Learning" from Kaggle

Data:

- 4920 patient observations
- 131 symptoms
- 41 diseases

Some light tidying to get the data ready for analysis and modeling

Data Handling

Duplicate Features

- One symptom had two associated columns with nearly identical names
- One column was empty
- Empty column dropped, other column renamed

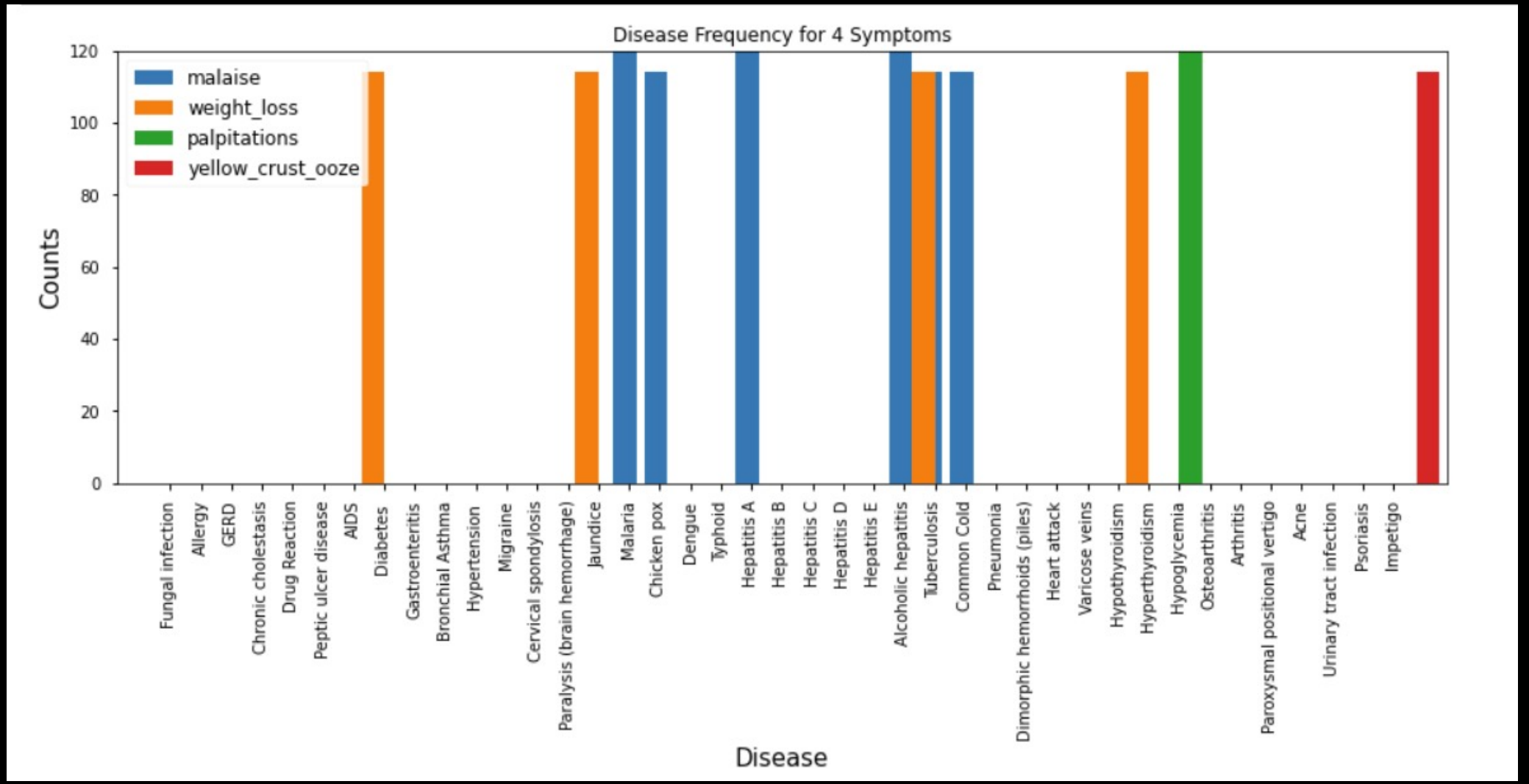
Formatting

- Several disease names had simple misspellings
  - Replaced with correct
- Some disease names had inconsistent formatting
  - Capitalized first letter

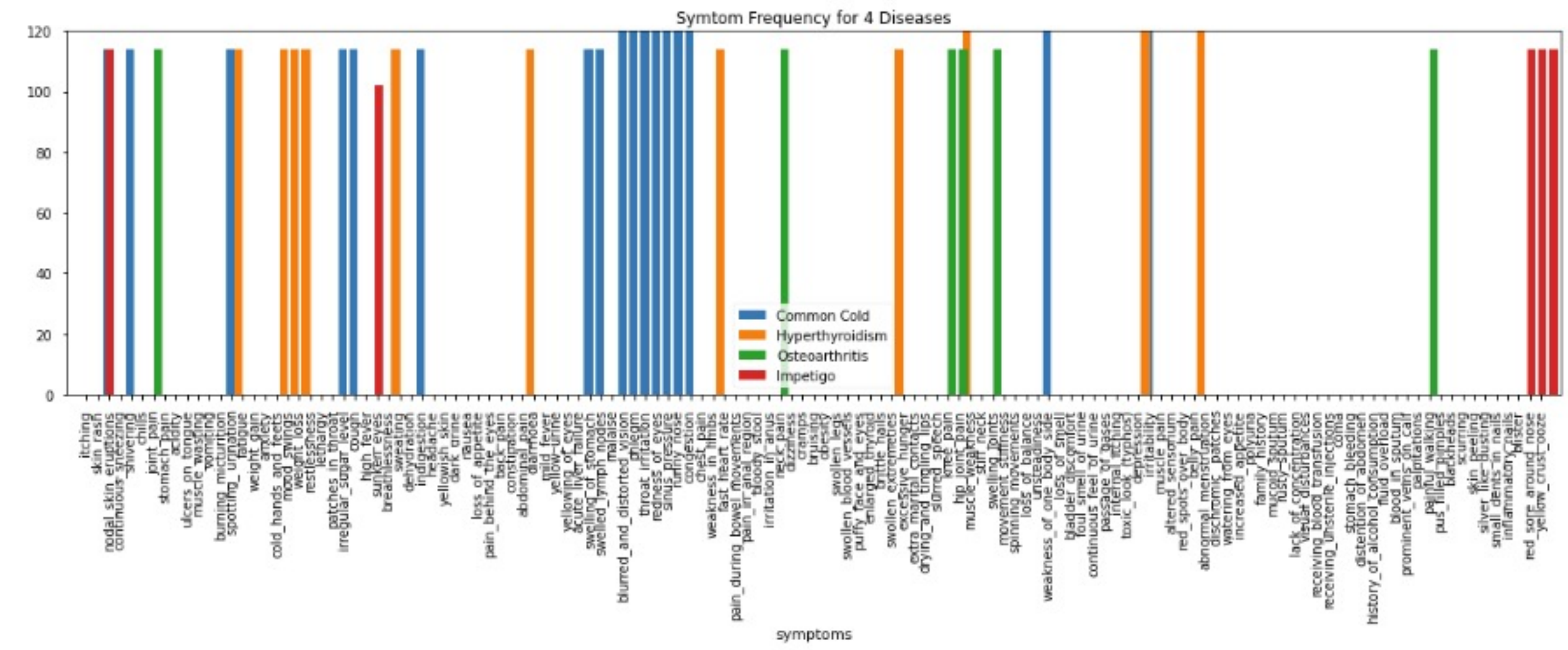
Label Encoding

- Disease names were label encoded prior to modeling

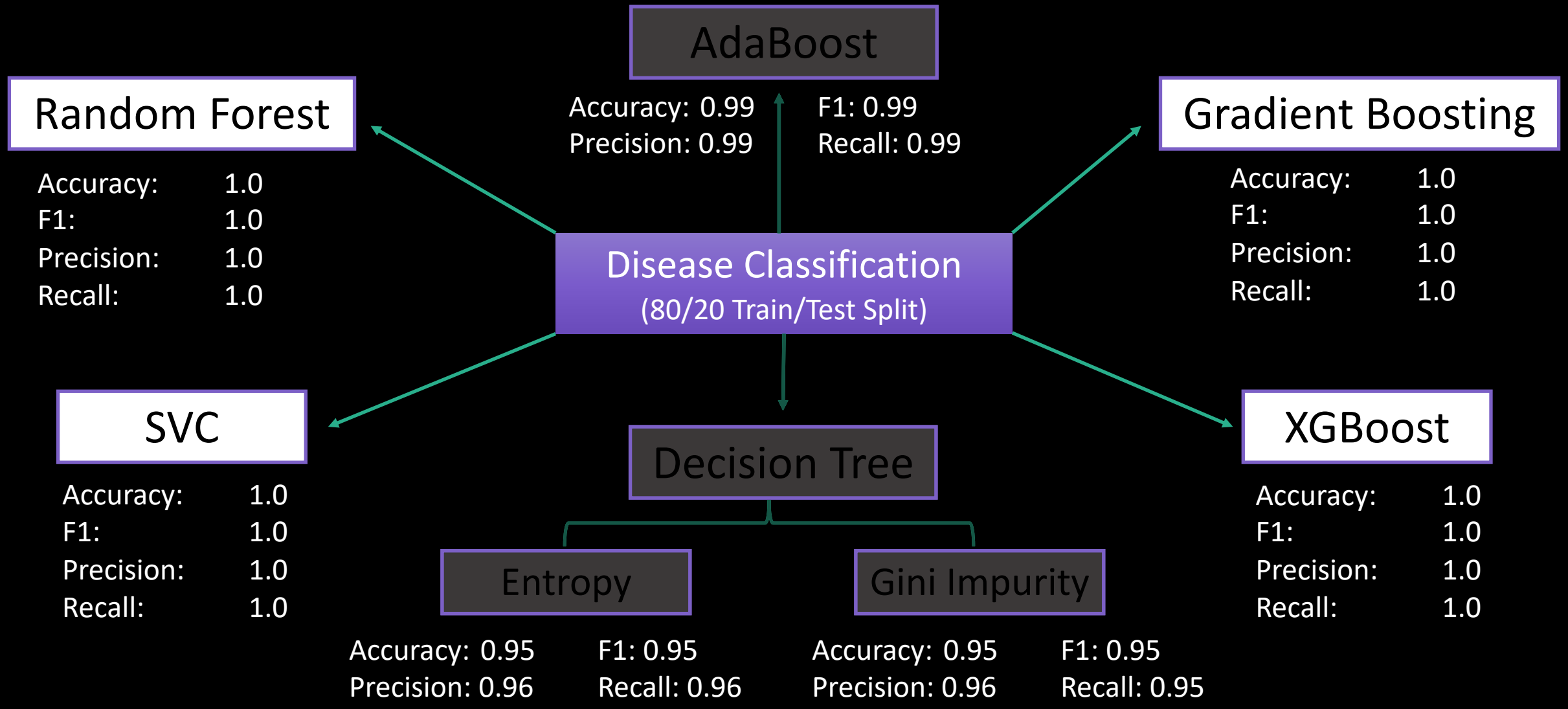
Symptoms can be indicated in one or more diseases



Each disease has a unique combination of symptoms



7 classification models were tested and 4 had perfect test accuracy scores



Training time and prediction time per patient to differentiate models