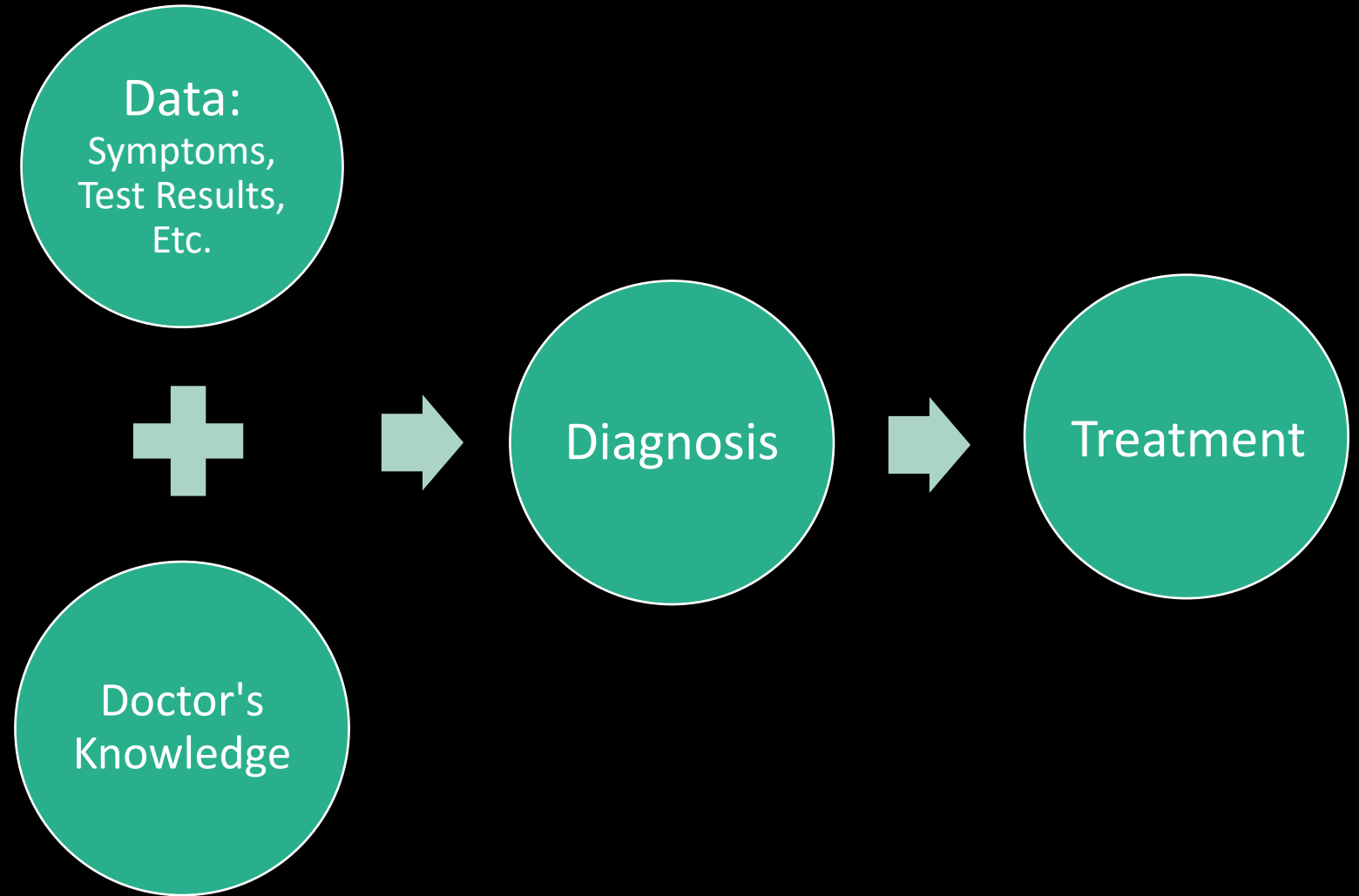




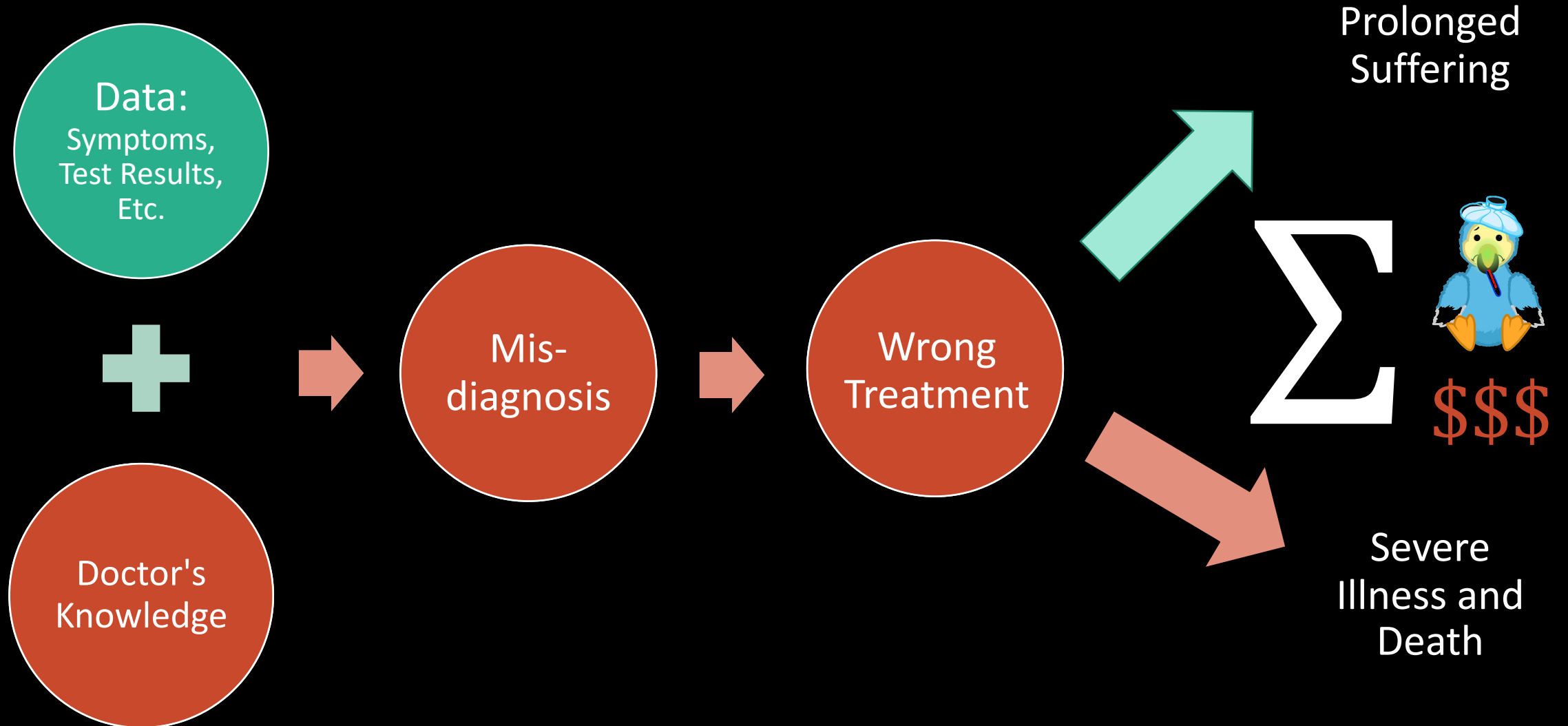
PATIENT DIAGNOSIS WITH MACHINE LEARNING

Caitlin Ortega Ruble

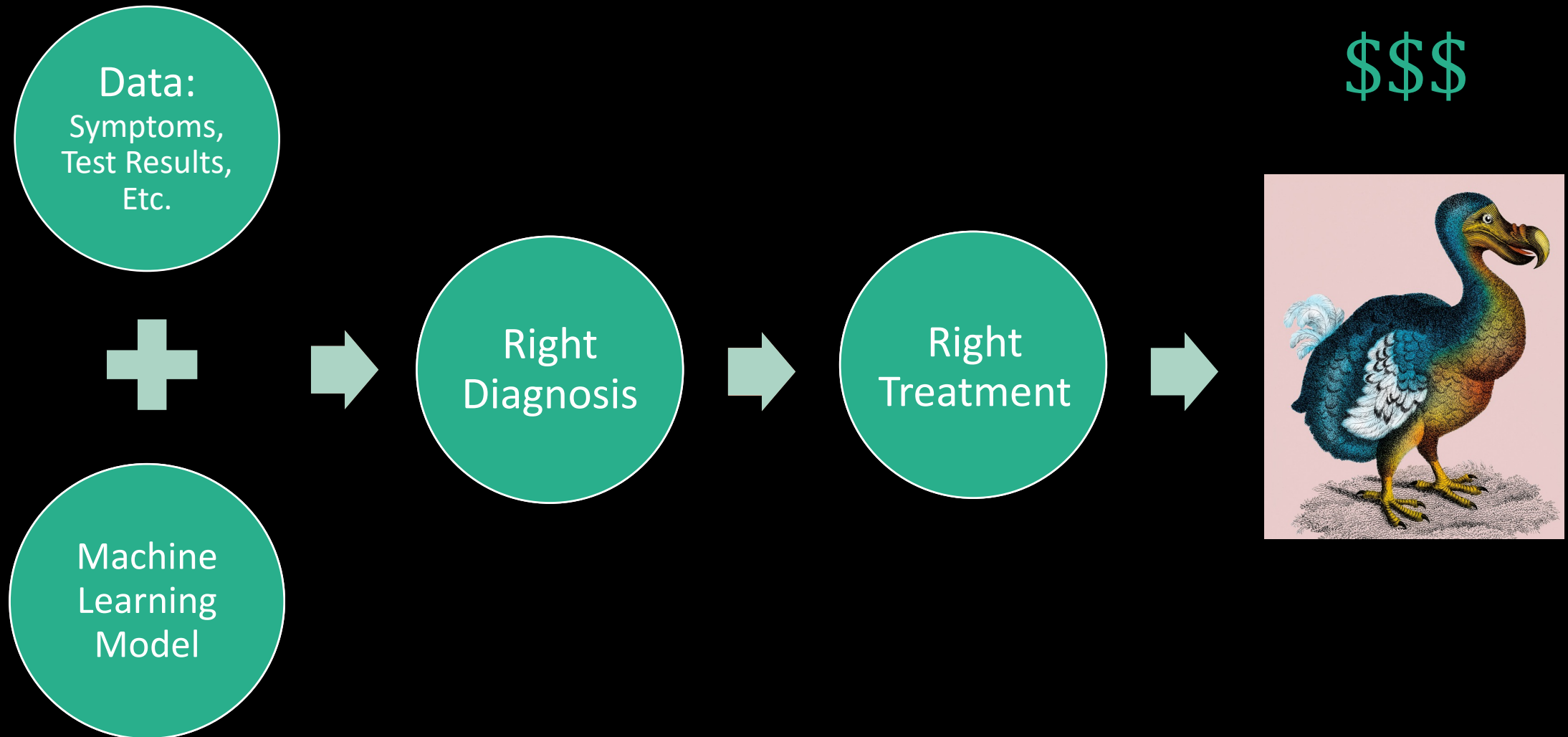
How diagnosis is done



How can it go wrong?



What can we do?



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

```
graph TD; A[Data Handling] --> B[Duplicate Features]; A --> C[Formatting]; A --> D[Label Encoding];
```

Duplicate Features

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

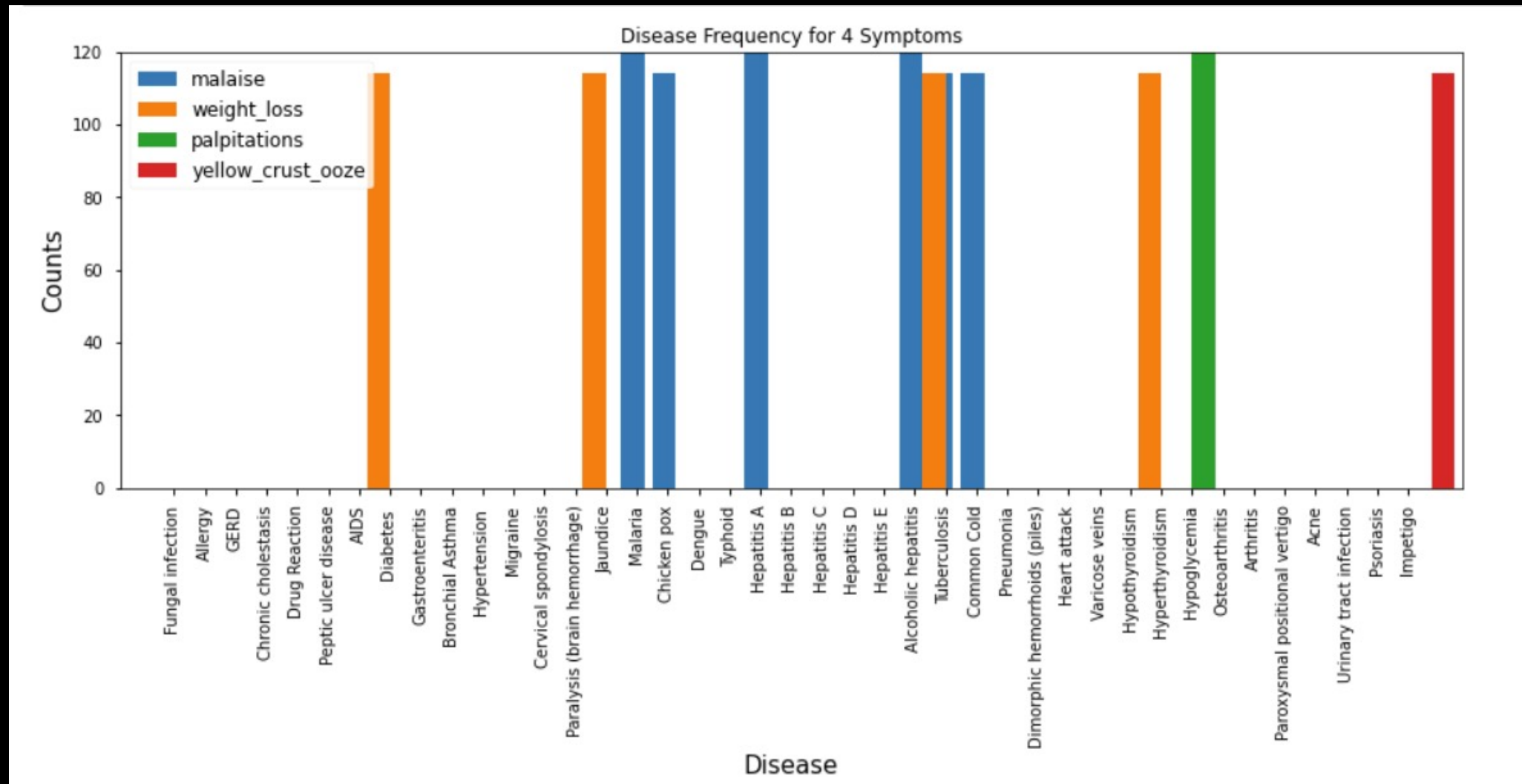
Formatting

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

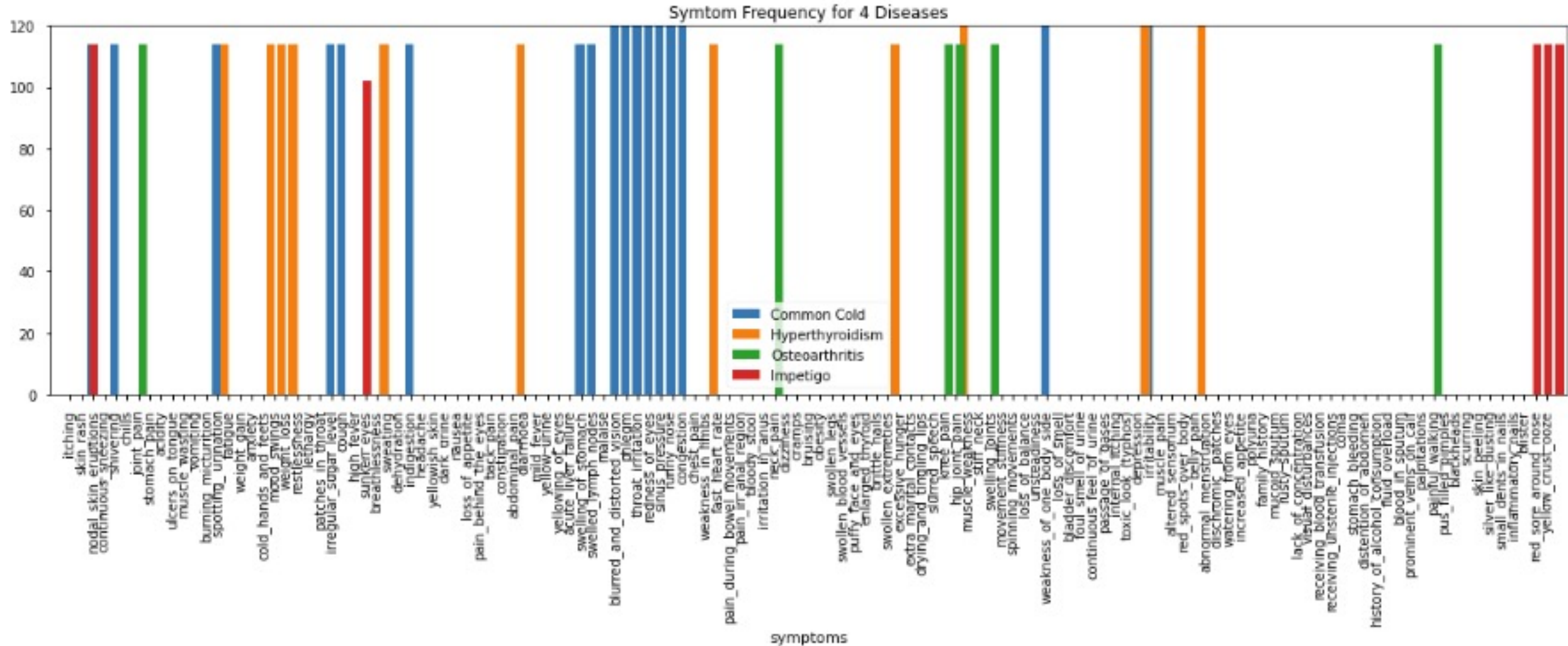
Label Encoding

- Disease names were label encoded prior to modeling

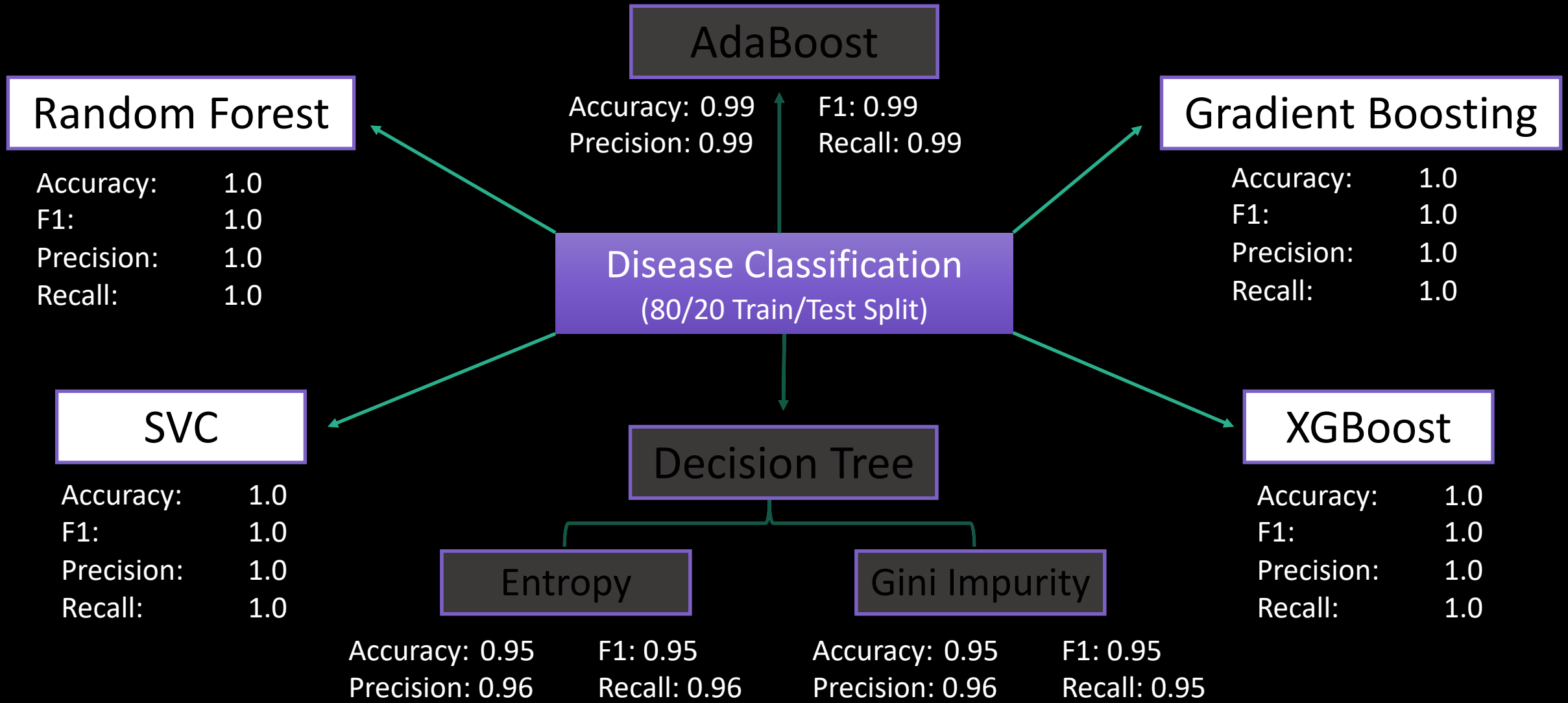
Symptoms can be indicated in one or more diseases



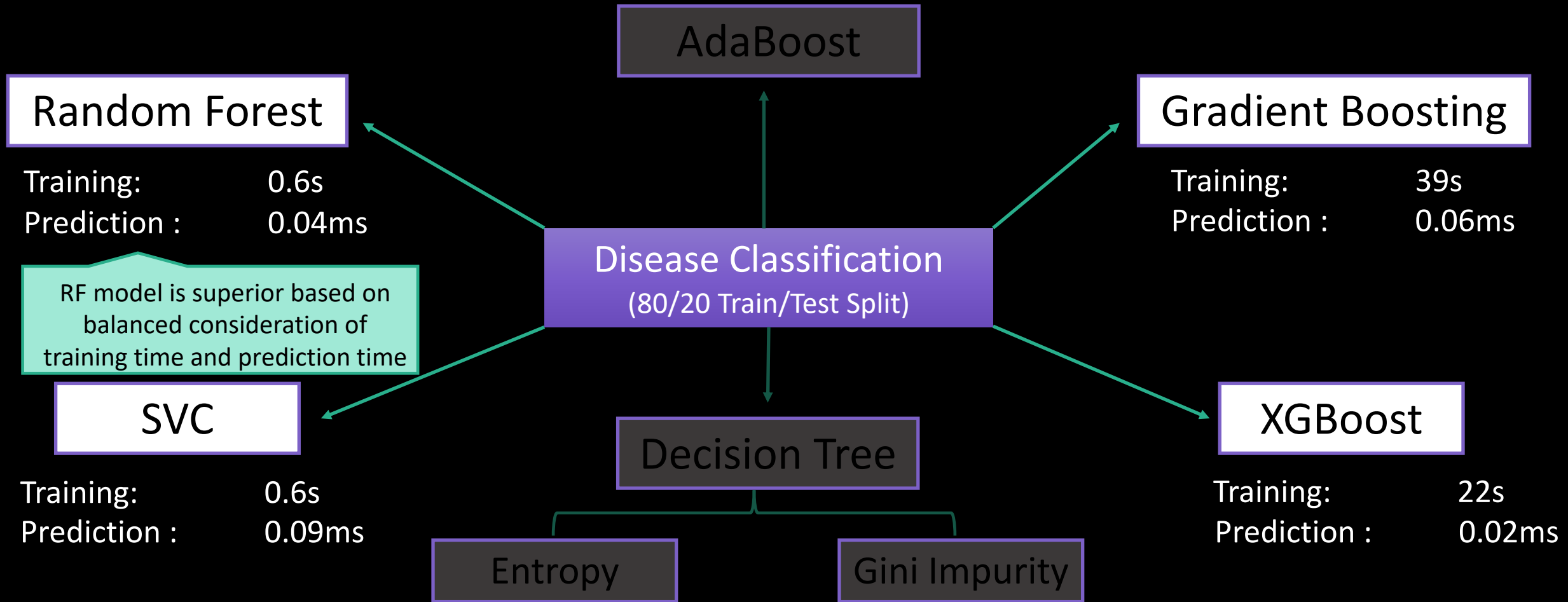
Each disease has a unique combination of symptoms



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



Training time and prediction time per patient to differentiate models



An analysis of feature importances reassures that the RF model is attributing similar weights to each symptom and that the model is valid

