EDA

In total, our dataset has 4962 samples, with 42 classes and 133 features. This gives us high complexity and only 120 samples per class in the training set (Fig. X). To manage problems with the complexity, we will compare between Random Forest and statistical tests for feature selection. Our chosen models, Random Forest and Gradient Boosting, are also more suited to higher-dimensional data. Both also have hyperparameters that can be tuned to reduce any overfitting that may occur. For our neural network, strategies such as dropout and hyperparameter tuning can also be used. Should we fail to reach an accuracy above 0.8, we can use data augmentation to produce more information. On preliminary testing, a decision tree achieves an accuracy of 0.977, indicating that this data will be sufficient to train our models with.

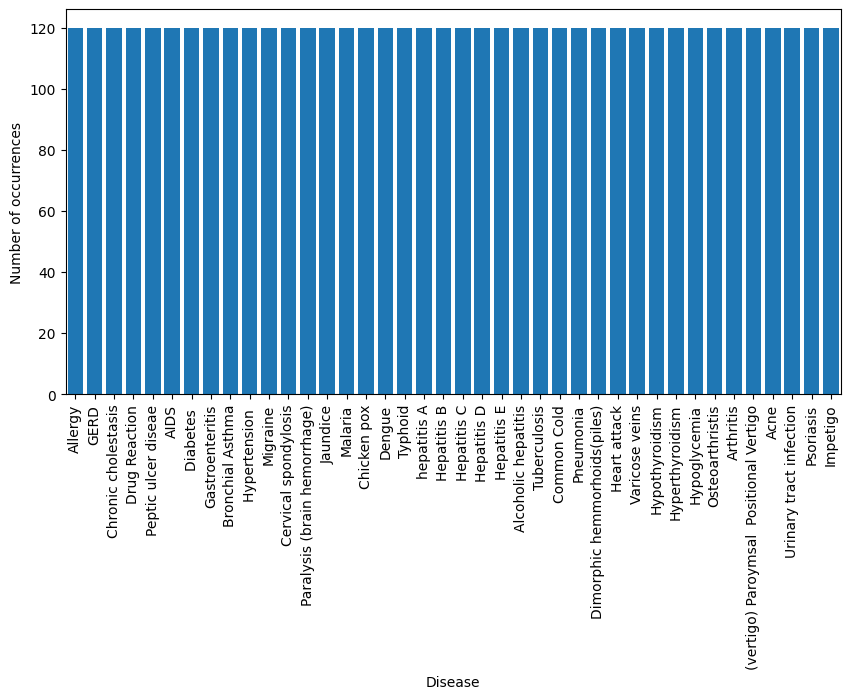


Figure X. Number of samples per Disease Class.

Timeline

We have chosen the following deadlines:

* **Saturday, March 8 2025** – Feature selection and data cleaning
  + Jamie will work on data cleaning, while Emily and Gabriel conduct feature selection.
* **Saturday, March 15 2025** –Model building and training
  + Emily will work on a neural network model and get initial metrics. Gabriel will be working on Random forest, Gradient Boosting and AdaBoost. Jamie will finish SVC, XGBoost and other models.
* **Saturday, March 22 2025** – Hyperparameter tuning and performance optimization
  + Each member will maximize performance for their assigned model and produce visuals. A draft project report will be completed.
* **Monday, April 2025** - Project Report
  + Project Report work to be divided