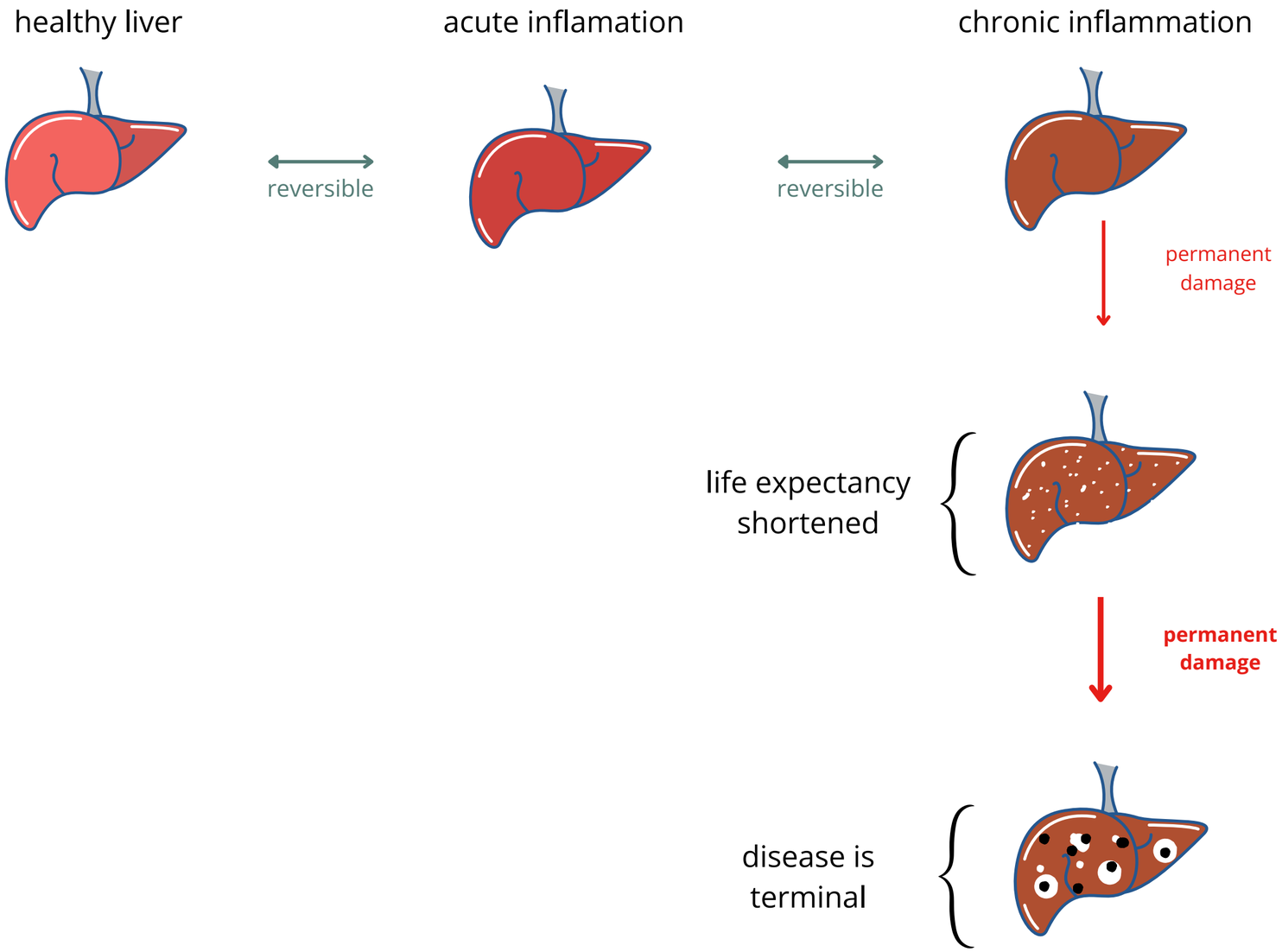
D INFK

Prognosis of hepatitis disease using ML techniques

LiverLogics

ETH zürich

The different stages of the disease



The dataset

Collection of **medical records** created in **1989** used for analyzing **factors** affecting the **survival** of patients diagnosed with **hepatitis**

- **19 features**
- **categorical** and **boolean** attributes
- **binary classification** (die or survive)

Research questions I

Motivation

Can we develop a model to forecast a patient's **survival outcome** based on specific **characteristics**?

Dataset hypothesis

Small dataset size

Modern deep learning methods will perform similarly to the simpler methods

Health hypothesis

A **small subset of the features** will have strong **enough predictive power** to give good accuracy in predicting the occurrence of a death event.

Research questions II

Motivation

Can we find a **correlation** between the **outcome** of the disease and the **measurements** that were taken during treatment ?

Dataset hypothesis

The majority of **missing values** account for **expensive** measurements / Lab tests

Methodology overview

Exploratory data analysis

Tackling of dataset imbalance

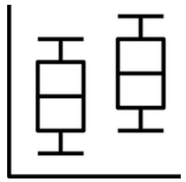
Models comparison (**baseline** vs **more complex** models)

Imputation of missing values

Feature importance: interpretability/explain ability

Ensemble techniques

Outlier detection



Cost-feature importance association

Error analysis

Challenges I

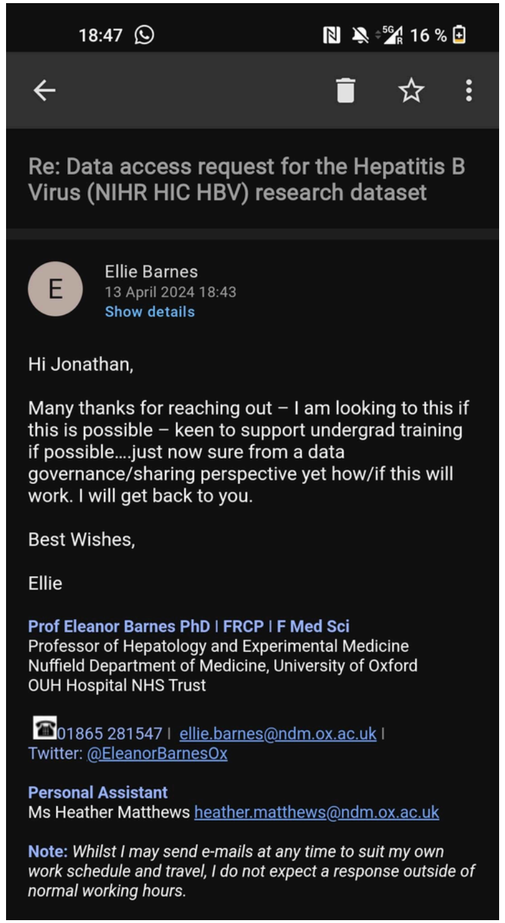
Limited dataset size

Only **155** instances

Datasets **compatibility issues**:

- not public
- differentt measurements
- differentt labels

restricted access



18:47 16 %

Re: Data access request for the Hepatitis B Virus (NIHR HIC HBV) research dataset

E Ellie Barnes
13 April 2024 18:43
Show details

Hi Jonathan,

Many thanks for reaching out – I am looking to this if this is possible – keen to support undergrad training if possible....just now sure from a data governance/sharing perspective yet how/if this will work. I will get back to you.

Best Wishes,

Ellie

Prof Eleanor Barnes PhD | FRCP | F Med Sci
Professor of Hepatology and Experimental Medicine
Nuffield Department of Medicine, University of Oxford
OUH Hospital NHS Trust

01865 281547 | ellie.barnes@ndm.ox.ac.uk |
Twitter: @EleanorBarnesOx

Personal Assistant
Ms Heather Matthews heather.matthews@ndm.ox.ac.uk

Note: *Whilst I may send e-mails at any time to suit my own work schedule and travel, I do not expect a response outside of normal working hours.*

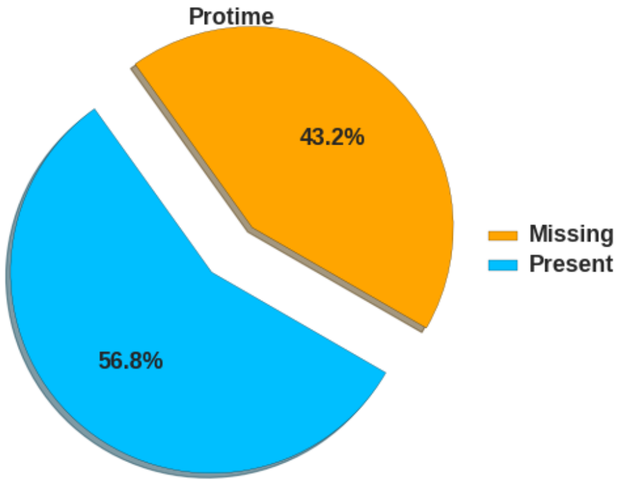
different measurements

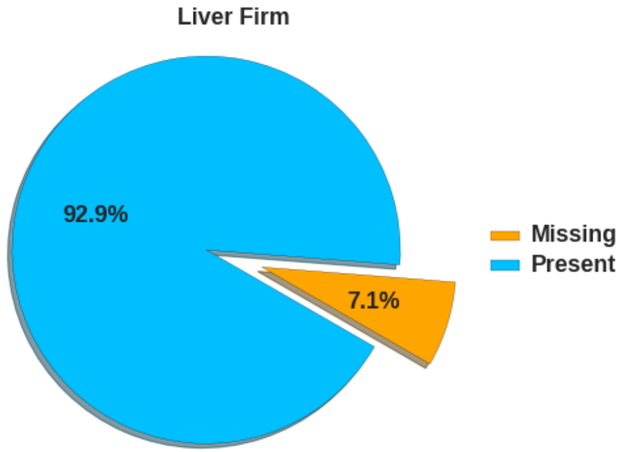
Variable Name	Role	Type	Demographic	Description	Units	Missing Values
Age	Feature	Integer	Age			no
Gender	Feature	Binary	Gender	[Male], [Female]		no
BMI	Feature	Integer		Body Mass Index		no
Fever	Feature	Binary		[Absent], [Present]		no
Nausea/Vomting	Feature	Binary		[Absent], [Present]		no
Headache	Feature	Binary		[Absent], [Present]		no
Diarrhea	Feature	Binary		[Absent], [Present]		no
Fatigue & generalized bone ache	Feature	Binary		[Absent], [Present]		no
Jaundice	Feature	Binary		[Absent], [Present]		no
Epigastric pain	Feature	Binary		[Absent], [Present]		no

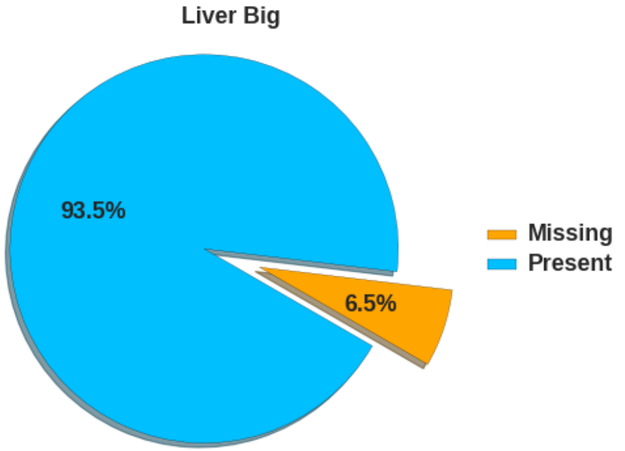
ETH zürich

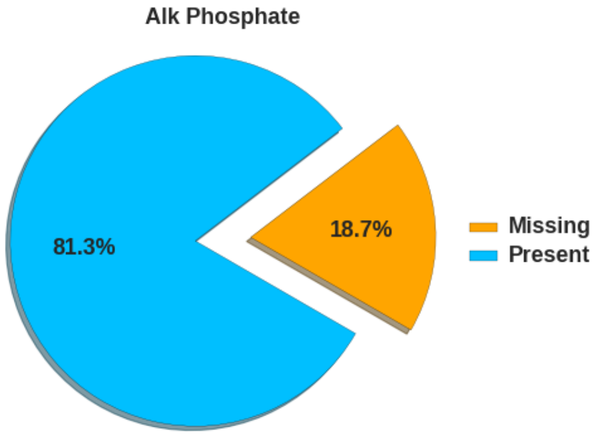
Challenges II

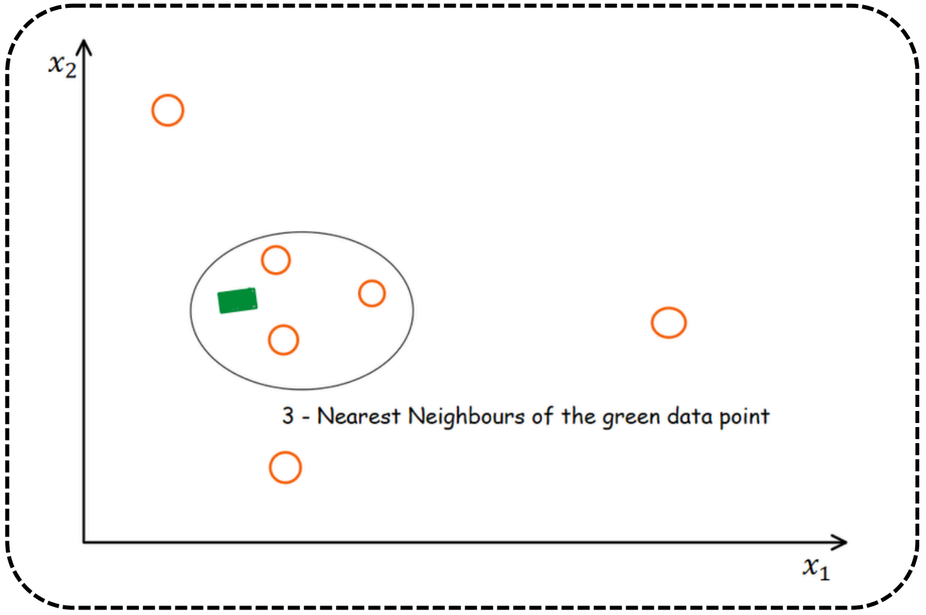
Missing values

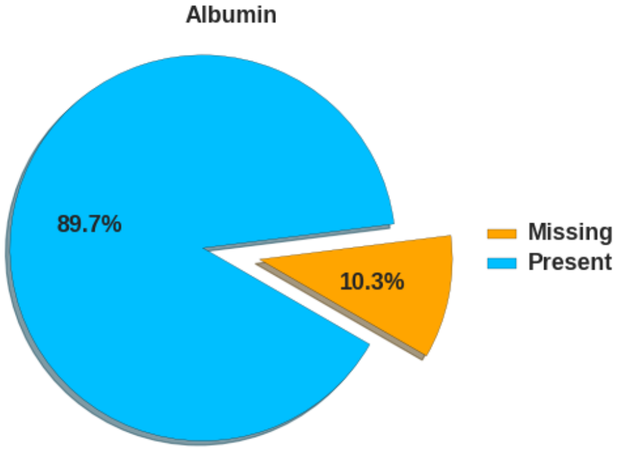






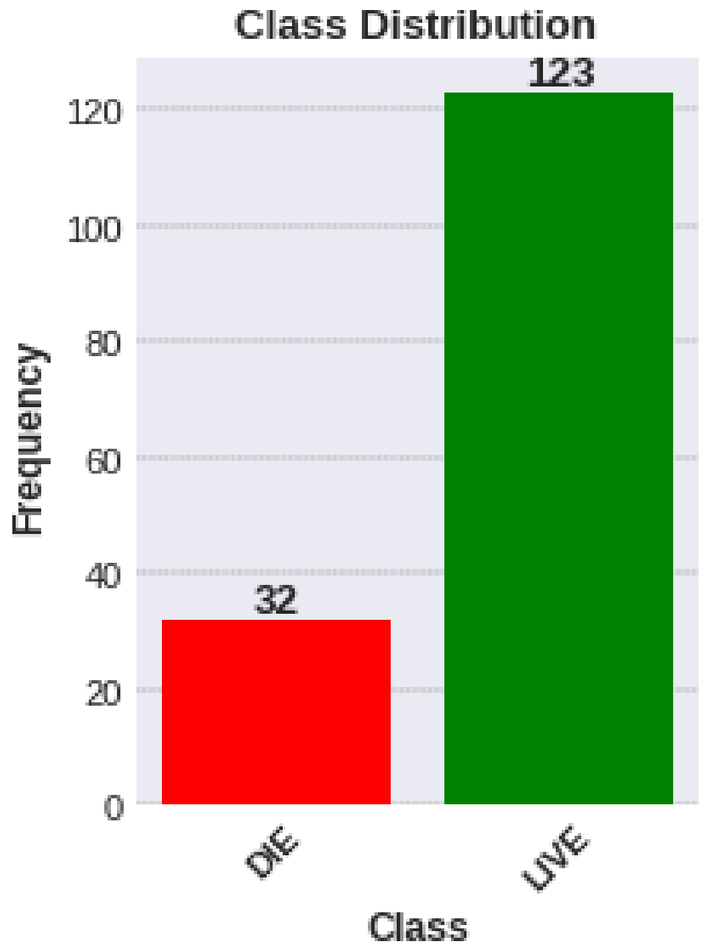






Challenges III

Class imbalance



✓ Simplistic approaches

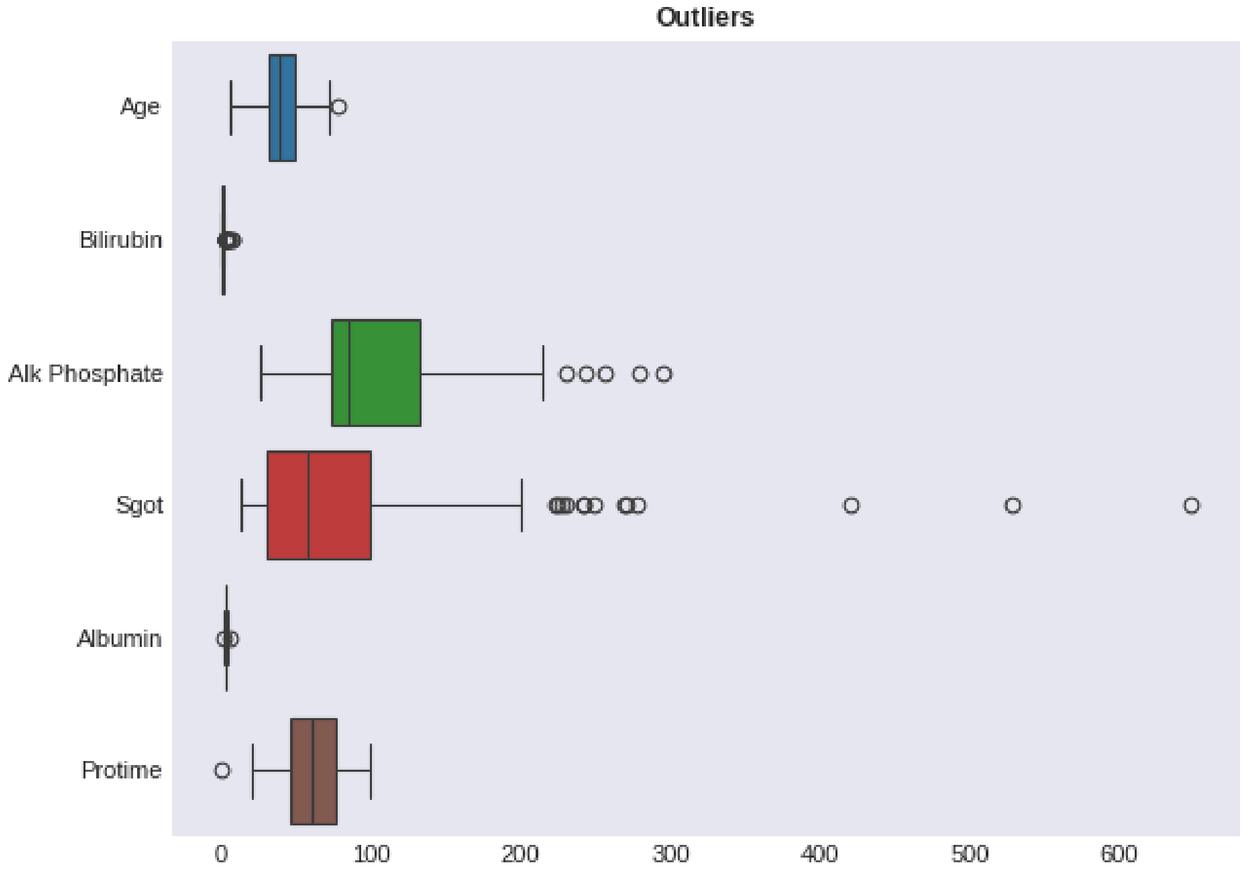
Oversampling minority class
Downsampling majority class

✓ More sophisticated methods

SMOTE

Challenges IV

Outliers



We did **NOT** exclude any outliers:

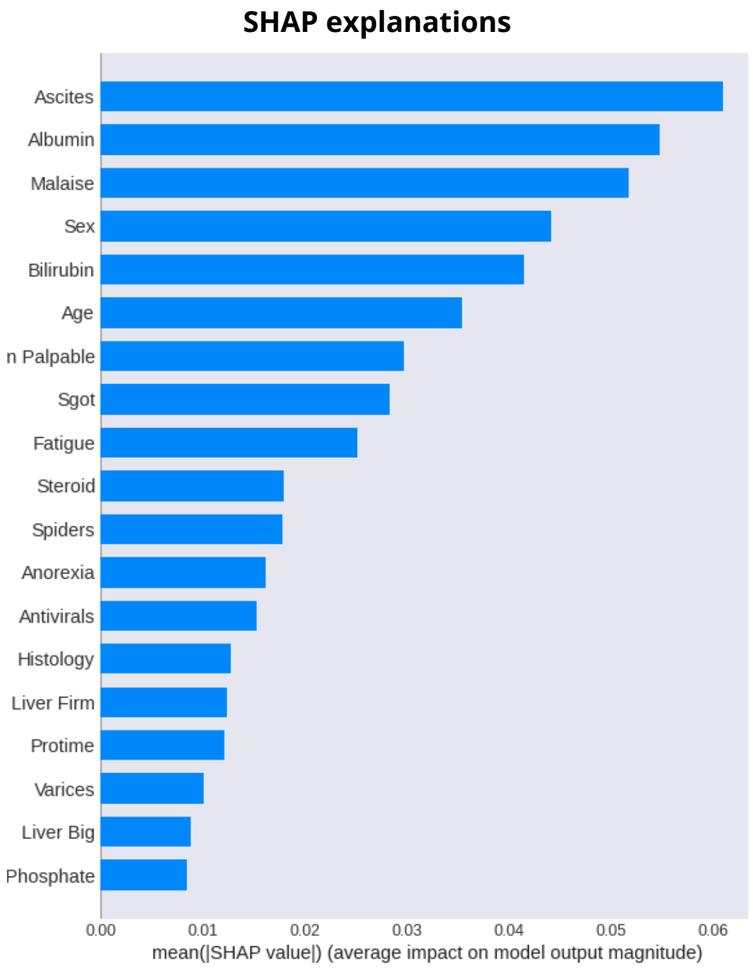
- ✓ They do not constitute measurement error
- ✓ Very limited dataset size

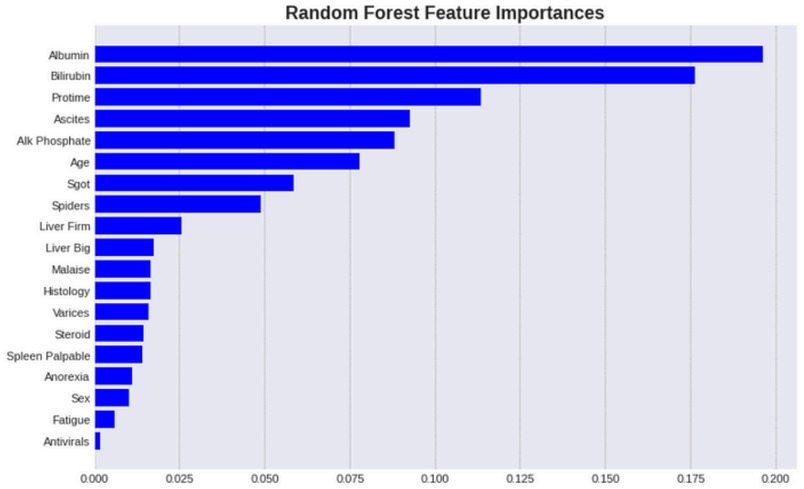
Challenges V

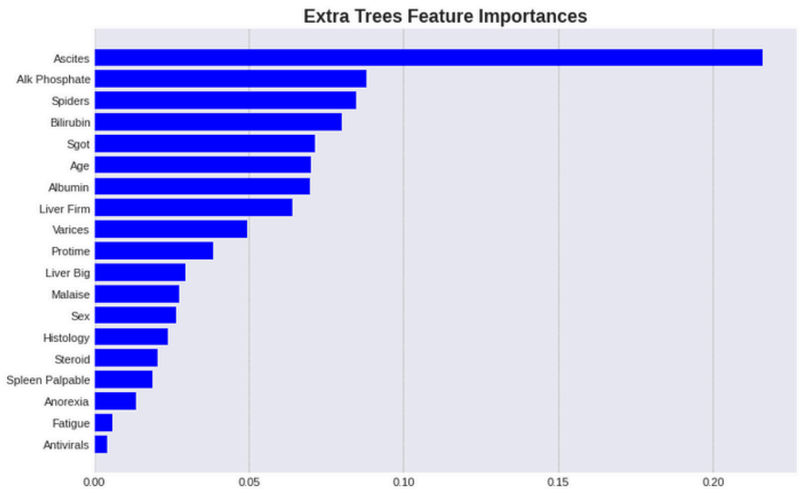
✓ Missing Features

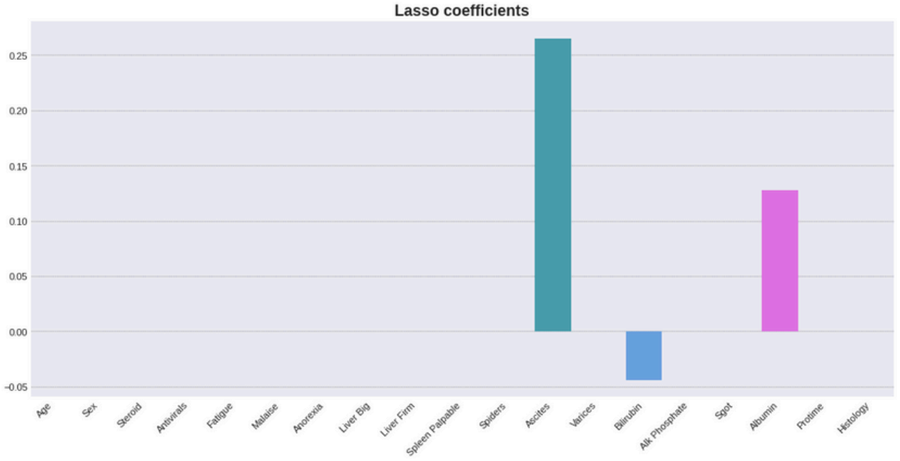
- Limited to symptoms of the patients and lab tests
- Lab Tests miss measurements of important liver enzymes
- Imaging studies and liver biopsies are not taken into account

Feature importance









Important Features

- ✓ Albumin
- ✓ Ascites
- ✓ Bilirubin
- ✗ Fatigue
- ✗ Antivirals
- ✗ Anorexia

Medical explanation of most imporant features

- **Liver enzymes: ALT, ALP, AST**
    - Alanine Aminotransferase (ALT)
    - Alk Phosphate (ALP)
    - Aspartate Aminotransferase (SGOT/AST)

- **Pro(thrombin) time:** blood test that measures how long it takes for blood to clot.

- **Bilirubin:** pigment produced by the breakdown of red blood cells.

- **Ascites:** accumulation of fluid in the abdominal cavity.

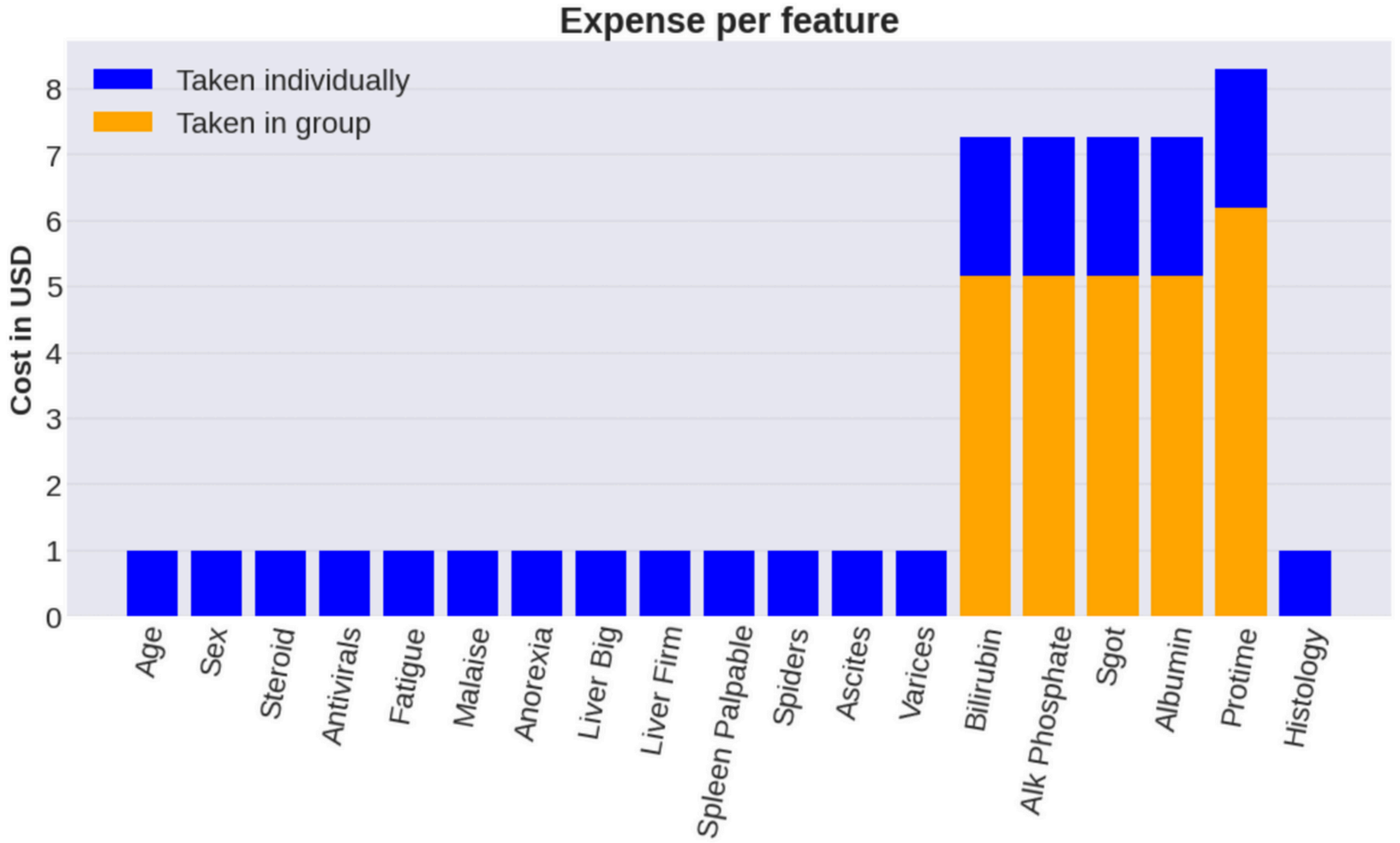
- **Albumin:** protein produced by the liver.

Medical explanation of less imporant features

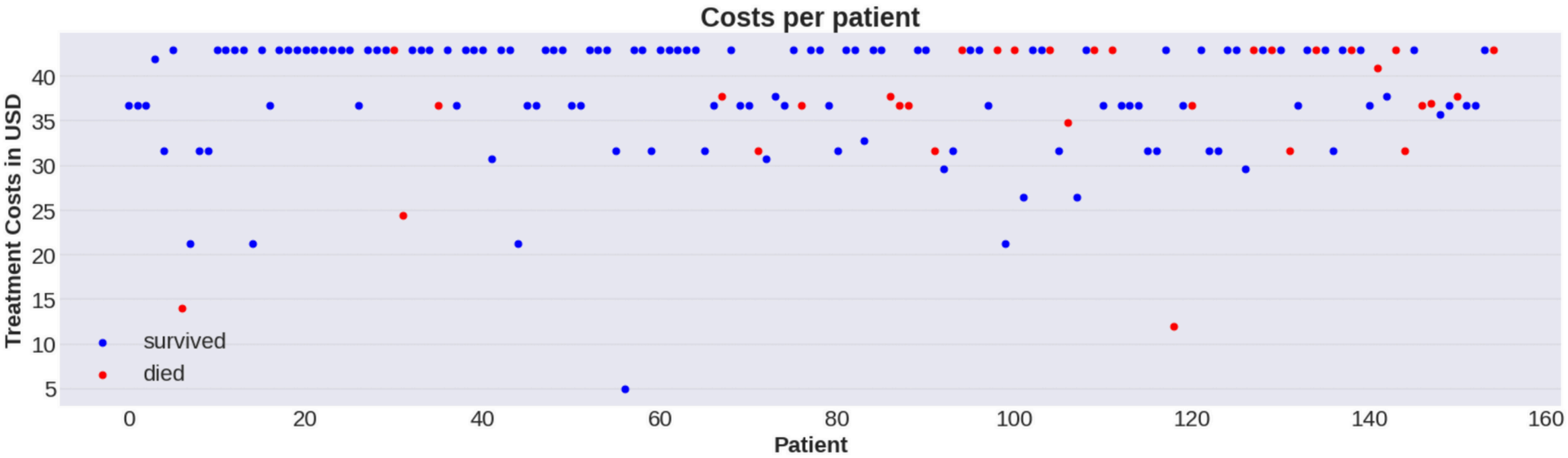
- **Antivirals:** medications used to treat viral infections.
- **Malaise:** term used to describe discomfort, and general unease.
- **Anorexia:** mental health disorder characterized by an intense fear of gaining weight.
- **Fatigue:** tiredness

Cost Analysis

We are given an additional costs dataset, listing the price for every measurement (feature) in USD



General Cost Information



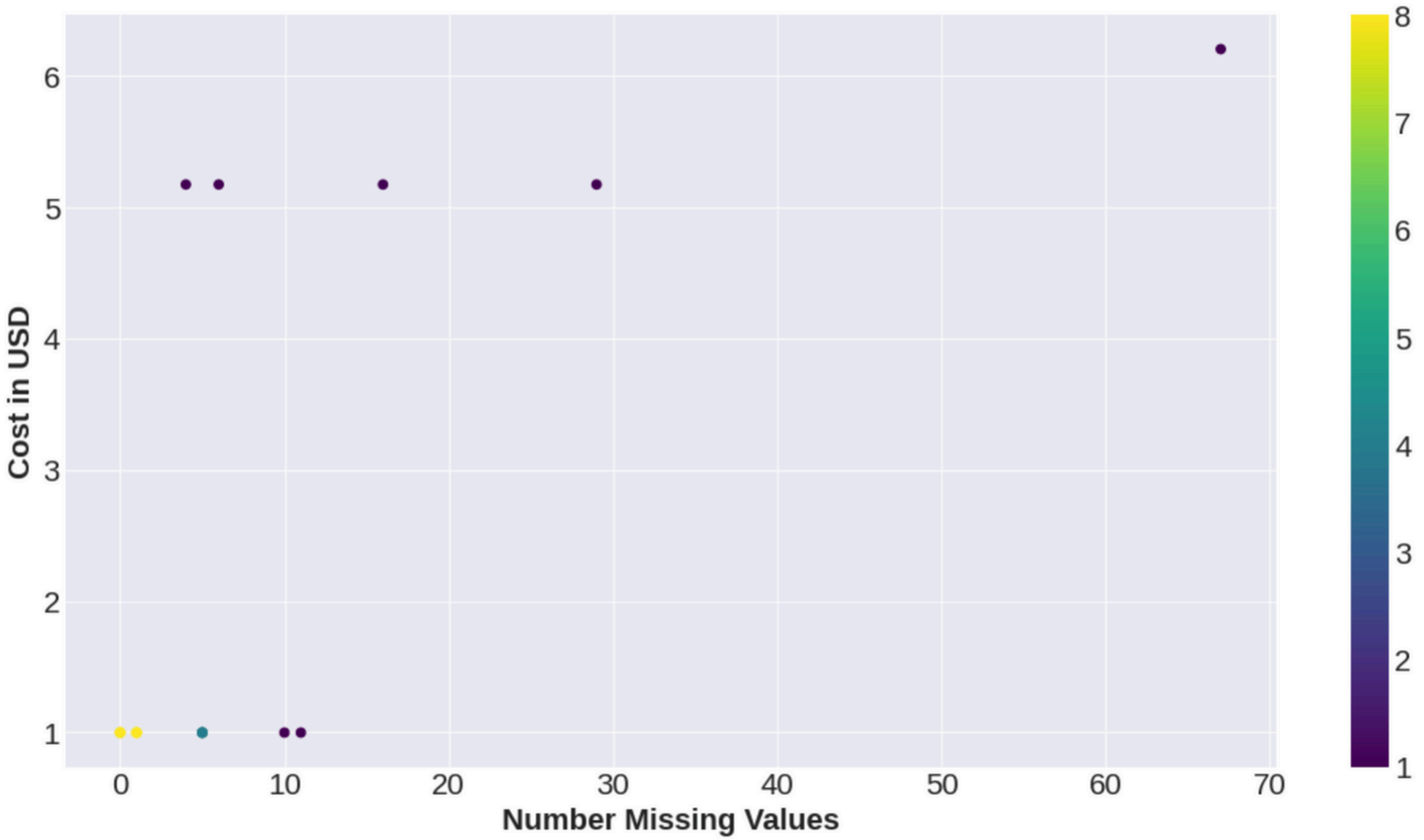
Mean expenses for patients:

- survived: 38.44 $
- dead: 36.95 $

Relative amount of patients where all possible measurements were taken:

- survived: 54.5 %
- dead: 40.6 %

Correlation between Missing Values and Cost



Pearson Correlation Coefficient
for evaluation

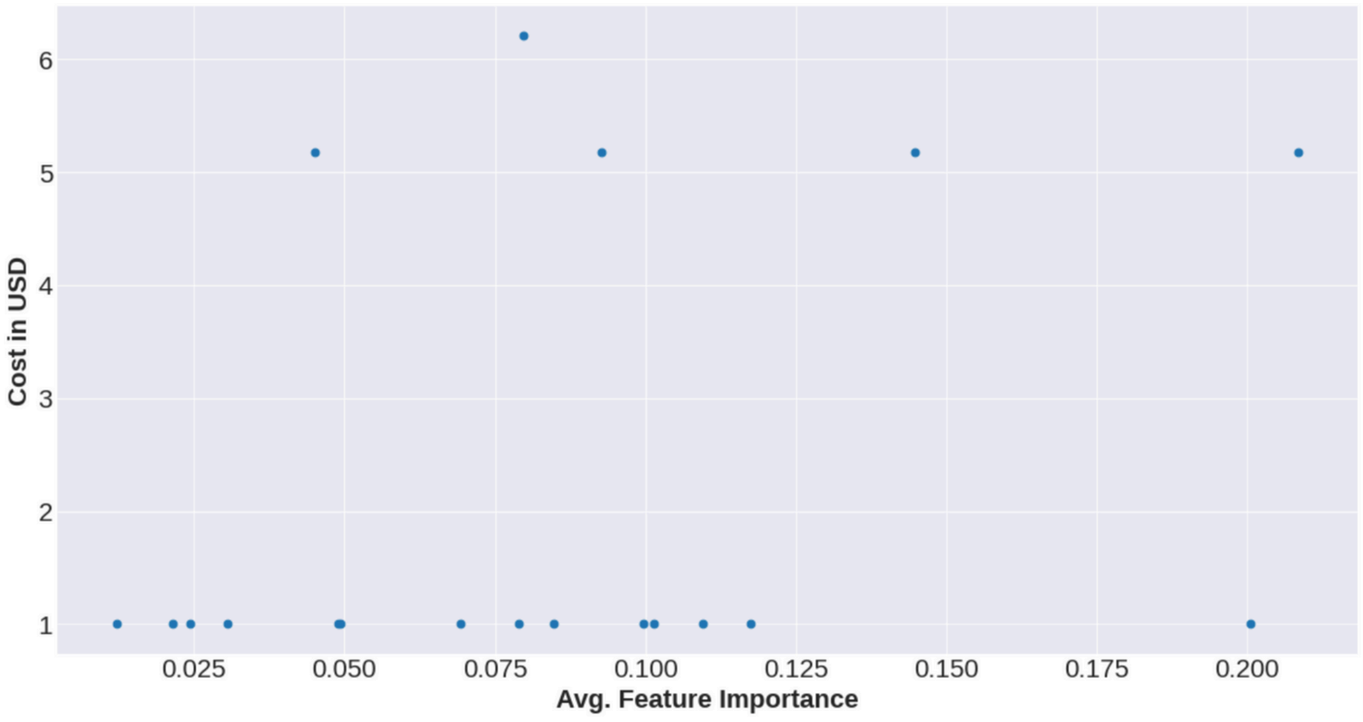
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where

- n is sample size
- x_i, y_i are the individual sample points indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$.

Pearson Correlation Coefficient **r = 0.7**

Correlation between averaged Feature Importance and Cost



Predictiveness values based on the **averaged feature importances** over all ML methods

Pearson Correlation Coefficient
for evaluation

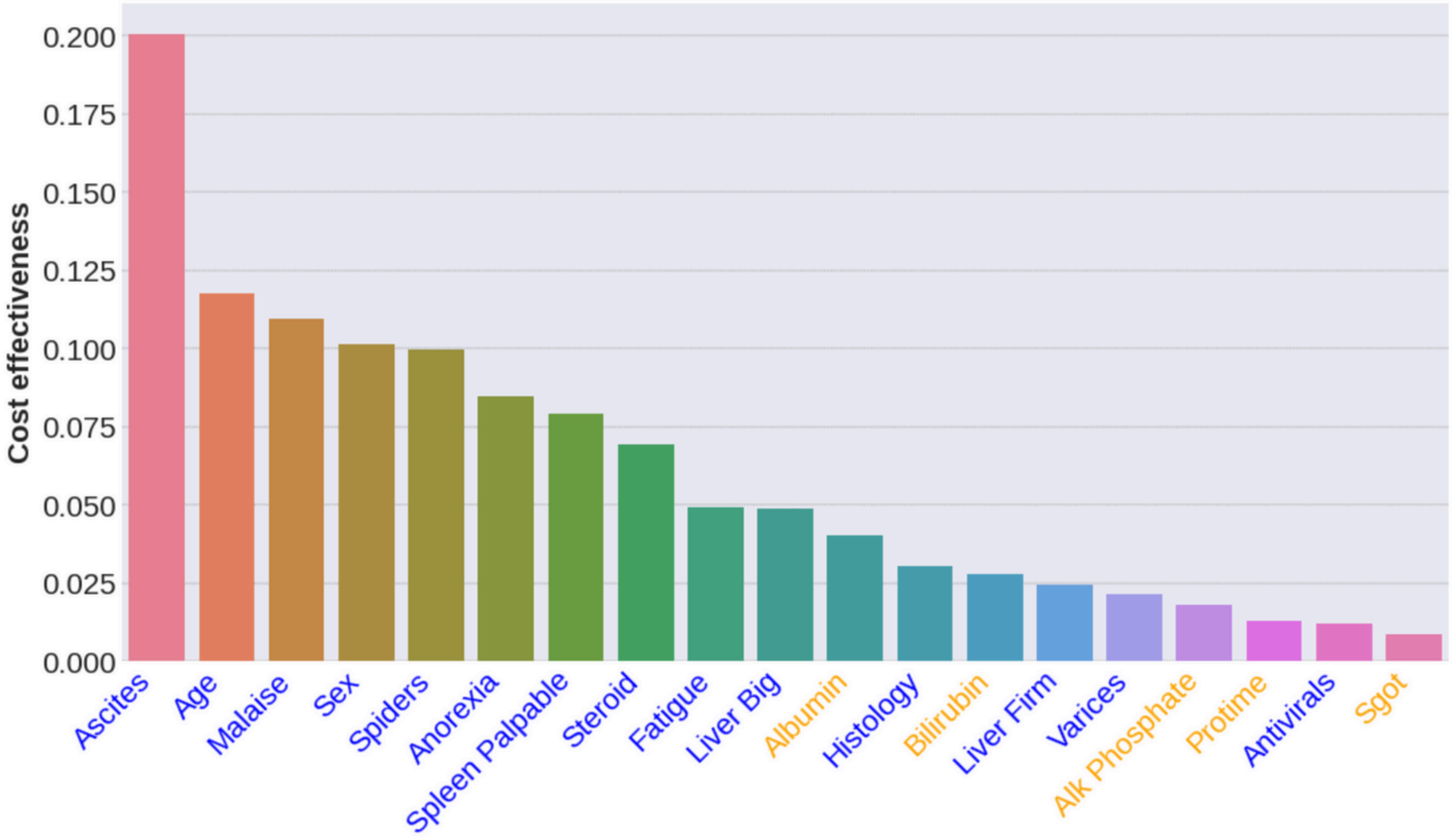
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where

- n is sample size
- x_i, y_i are the individual sample points indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$.

Pearson Correlation Coefficient **r = 0.3**

Cost Effectiveness



Predictiveness values based on the **averaged feature importances** over all ML methods

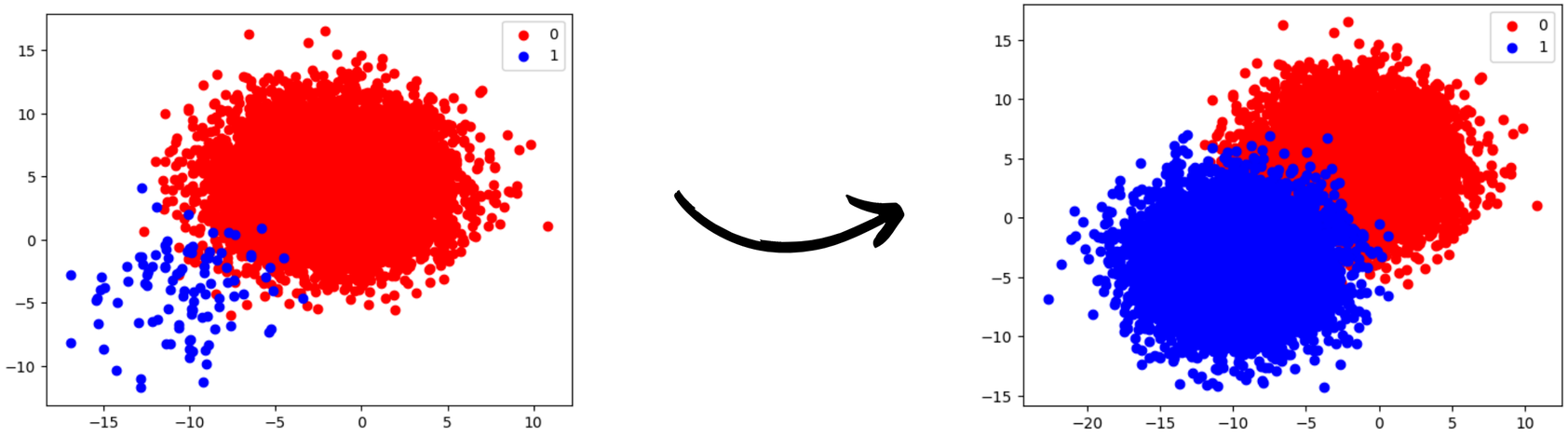
- **Expensive** blood measurements highlighted in yellow

Cost analysis conclusion

- ✓ Patients that survived had **more expensive measurements** taken from them
- ✓ The expensive blood measurements were disproportionately often not taken
  - ➡ For patients with critical symptoms, not all sensitive blood measurements are needed to confirm the **severity of the disease**

Dealing with unbalanced data

Oversampling minority class

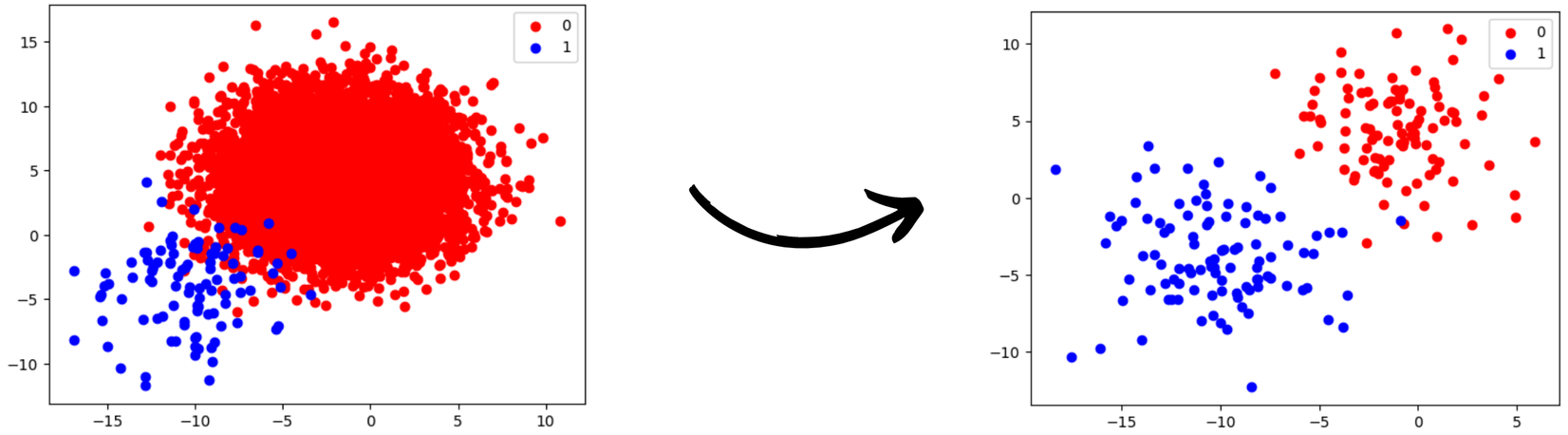


- **Balance** the dataset
- **Increase** the total number of datapoints

⚠ Create “fake” data → need to be **careful** and **deliberate**

Dealing with unbalanced data

Undersampling majority class



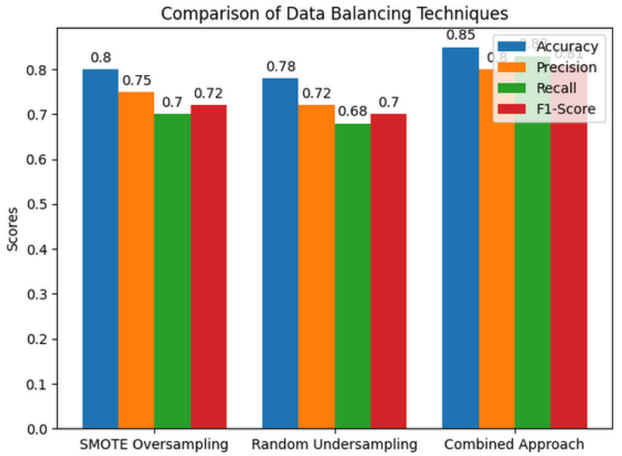
- **Balance** the dataset
- **No synthetic** data points are created

⚠ Fewer data points are used

Dealing with unbalanced data

Methodology: Balancing data with combined techniques

Empirical wisdom: A **combination** of both is usually the best approach, as the number of data points **does not decrease**, while the fraction of **synthetic** data points is **not too high**



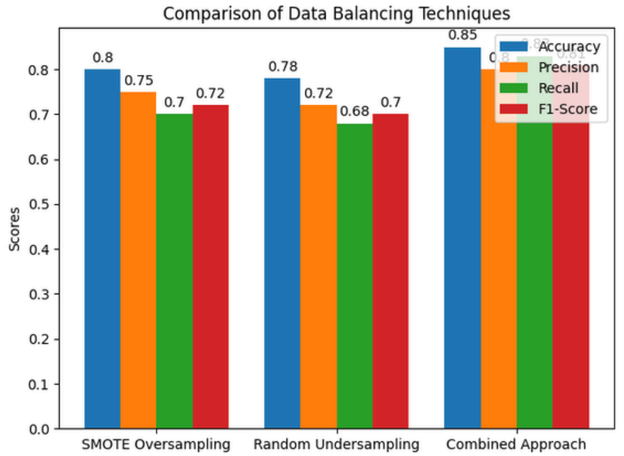
Dealing with unbalanced data

Methodology: Balancing data with combined techniques

Empirical wisdom: A **combination** of both is usually the best approach, as the number of data points **does not decrease**, while the fraction of **synthetic** data points is **not too high**

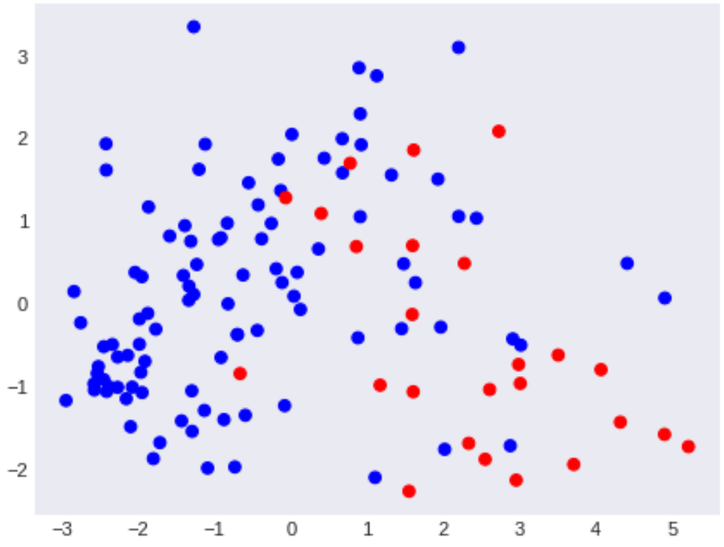
Our experimentations

- ✓ SMOTE oversampling
- ✓ Undersampling
- ✓ Both

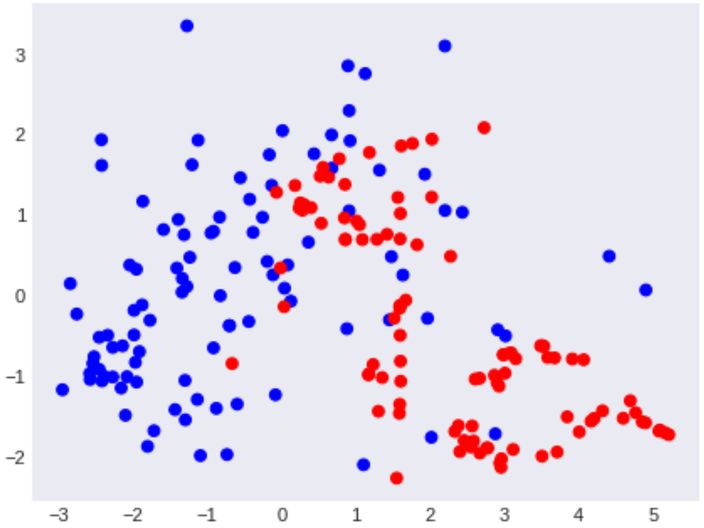


SMOTE

1. Selects a minority point **a** at random and finds its **k** nearest minority neighbors.
2. A new point is created by randomly choosing one of the neighbors **b**, connecting **a** and **b** to form a line segment.
3. The new point is then randomly chosen on the line segment.



Original data set (visualized using top-2 PCA)



SMOTE performed on top-2 PCA (k=5)

Evaluation metrics

🎯 Accuracy is a **poor** metric for unbalanced data

✓ F1-score
✓ balanced accuracy

Our data set is too small to make sound decisions using a validation set

✓ Use **cross-validation** to find the best potential model architectures and hyper-parameters

✓ Once discovered, train model from scratch and **test** on split dataset

Results

Mean and Std. Dev of 25 runs

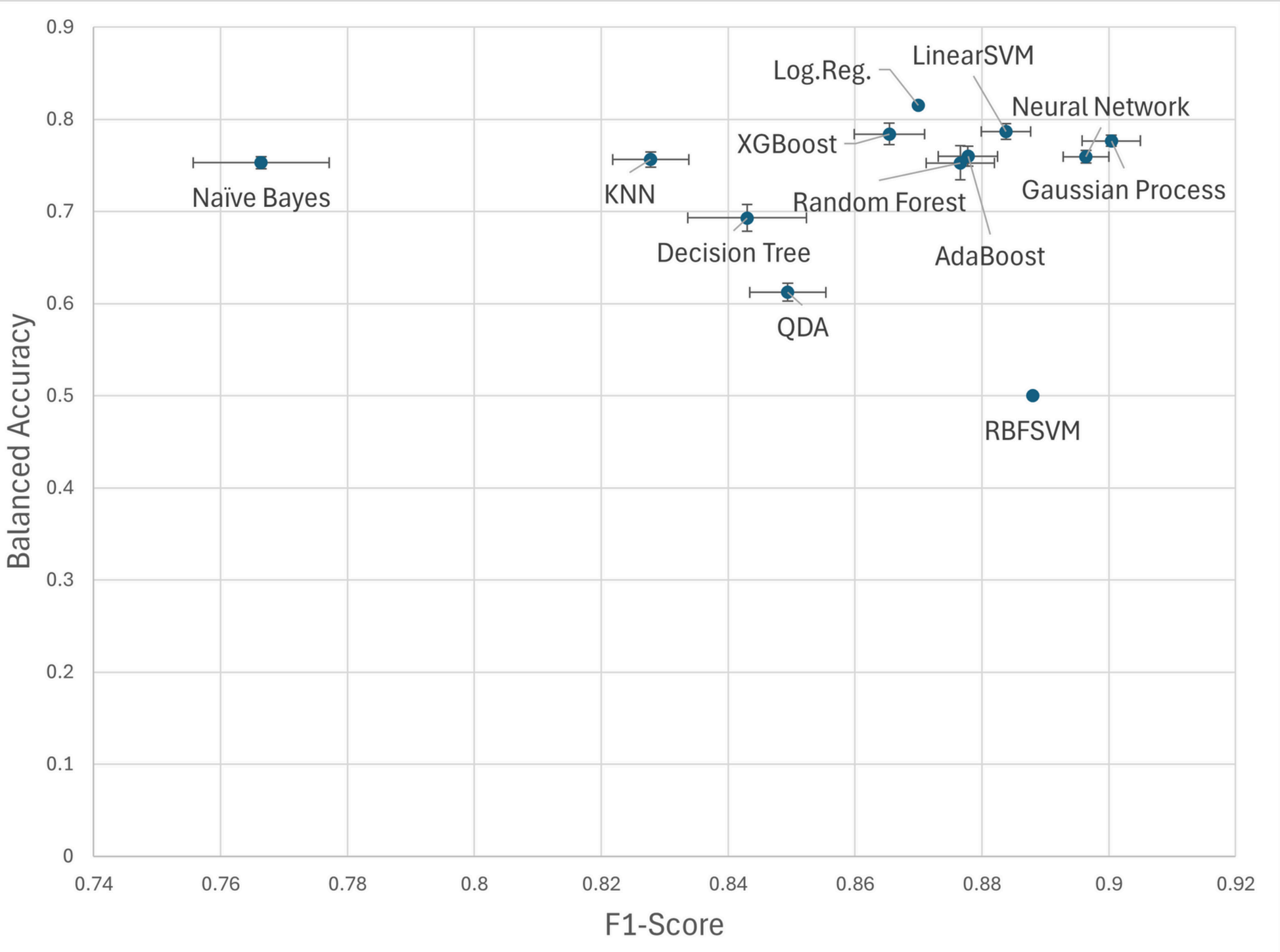
Model	F1 Score Whole Dataset (Mean ± Std. Dev)	F1 Score reduced feature set (Mean ± Std. Dev.)	Balanced Accuracy Whole Dataset (Mean ± Std. Dev.)
KNN	0.827 ± 0.006	0.830 ± 0.003	0.756 ± 0.008
Linear SVM	0.884 ± 0.004	0.833 ± 0.005	0.787 ± 0.008
RBF SVM	0.888 ± 0	0.873 ± 0.001	0.5 ± 0
Gaussian Process	**0.900 ± 0.005**	0.854 ± 0.004	0.777 ± 0.006
Decision Tree	0.843 ± 0.009	0.823 ± 0.007	0.693 ± 0.015
Random Forest	0.877 ± 0.005	0.856 ± 0.009	0.753 ± 0.019
NN	**0.896 ± 0.004**	0.870 ± 0.005	0.759 ± 0.011
AdaBoost	0.878 ± 0.005	0.838 ± 0.008	0.760 ± 0.011
Naive Bayes	**0.766 ± 0.011**	**0.883 ± 0.003**	0.753 ± 0.007
QDA	**0.849 ± 0.006**	**0.883 ± 0.008**	0.612 ± 0.010
XGBoost	0.865 ± 0.006	0.799 ± 0.001	0.784 ± 0.012
Log. Reg - L2	0.870 ± 0	0.843 ± 0.003	**0.815 ± 0**

Models

Conclusions:
F1-score not sufficient to find the best model

Best models:

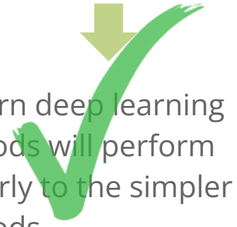
- Gauss. Process
- Linear SVM
- Log. Regression



Hypothesis confirmed

Dataset hypothesis

Small dataset size



Modern deep learning methods will perform similarly to the simpler methods

Health hypothesis

A **small subset of the features** will have strong **enough predictive power** to give good accuracy in predicting the occurrence of a death event.

Takeaway messages

✓ **There seems to be a difference between measurements important for diagnosis and measurements important for predicting outcome:**

- Expensive blood measurements are not as predictive / important as we assumed - especially when normalized over the costs
- Other (cheap) measurements have significantly higher predictive power

Thank you for your attention!

The team*

Maria Gkoulta

Andreas Psaroudakis

Alexander Morgenroth

Ziyad Diab

Jonathan Seele

* All team members contributed equally

ETH zürich

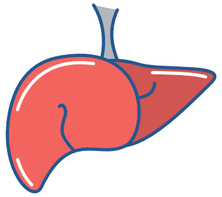
References

- https://archive.ics.uci.edu/dataset/46/hepatitis
- https://en.wikipedia.org/wiki/Hepatitis
- https://hep.org/hep-abc/
- https://scikit-learn.org/stable/
- https://shap.readthedocs.io/en/latest/
- https://arxiv.org/abs/1106.1813
- https://pubmed.ncbi.nlm.nih.gov/35708657/
- https://www.who.int/health-topics/hepatitis#tab=tab_1
- https://medlineplus.gov/hepatitis.html

Appendix

The disease

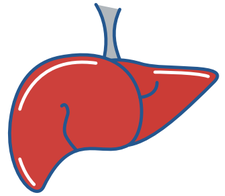
healthy liver



- Cells function optimally, detoxifying the blood and producing vital proteins

reversible

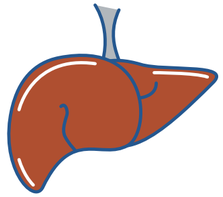
acute inflamation



- The essential processes are disrupted

reversible

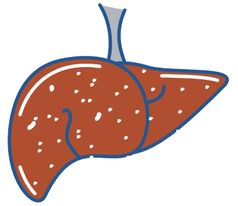
chronic inflammation



- The liver tissue is damaged over time

permanent damage

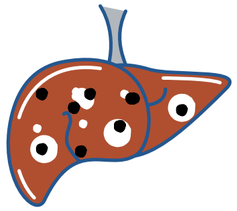
life expectancy shortened



- Chronic inflammation leads to cirrhosis
- The liver's ability to regenerate is hindered

permanent damage

disease is terminal



- Chronic hepatitis can evolve into liver failure or liver cancer

Hepatitis stages of progression

ETH zürich

The Diagnosis and Severity of Hepatitis

Symptoms

- Fatigue
- Nausea
- Vomiting
- Anorexia

Lab Tests

- Liver enzyme tests
- Bilirubin levels
- Viral loads
- Serology tests
- Prothrombin time

Imaging Studies

- Ultrasound / CT / MRI
- Evaluate structure and condition of the liver
- Detect abnomalies such as cirrhosis or cancer

Liver Biopsy

- Sample of liver tissue
- Assess the extent of liver damage and inflammation

Correlation matrix