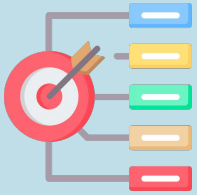
Prognosis for liver metastasis in unresectable metastatic colorectal cancer

Dylan Nico Ambrosi
Francesco Botrugno
Anas Shamoon

Code:
`https://github.com/anasshamoon12002/tumor-analysis-liver-metastasis.git`

OBJECTIVE



build a system that can **predict survival** after surgery of liver metastasis in **metastatic colorectal cancer**

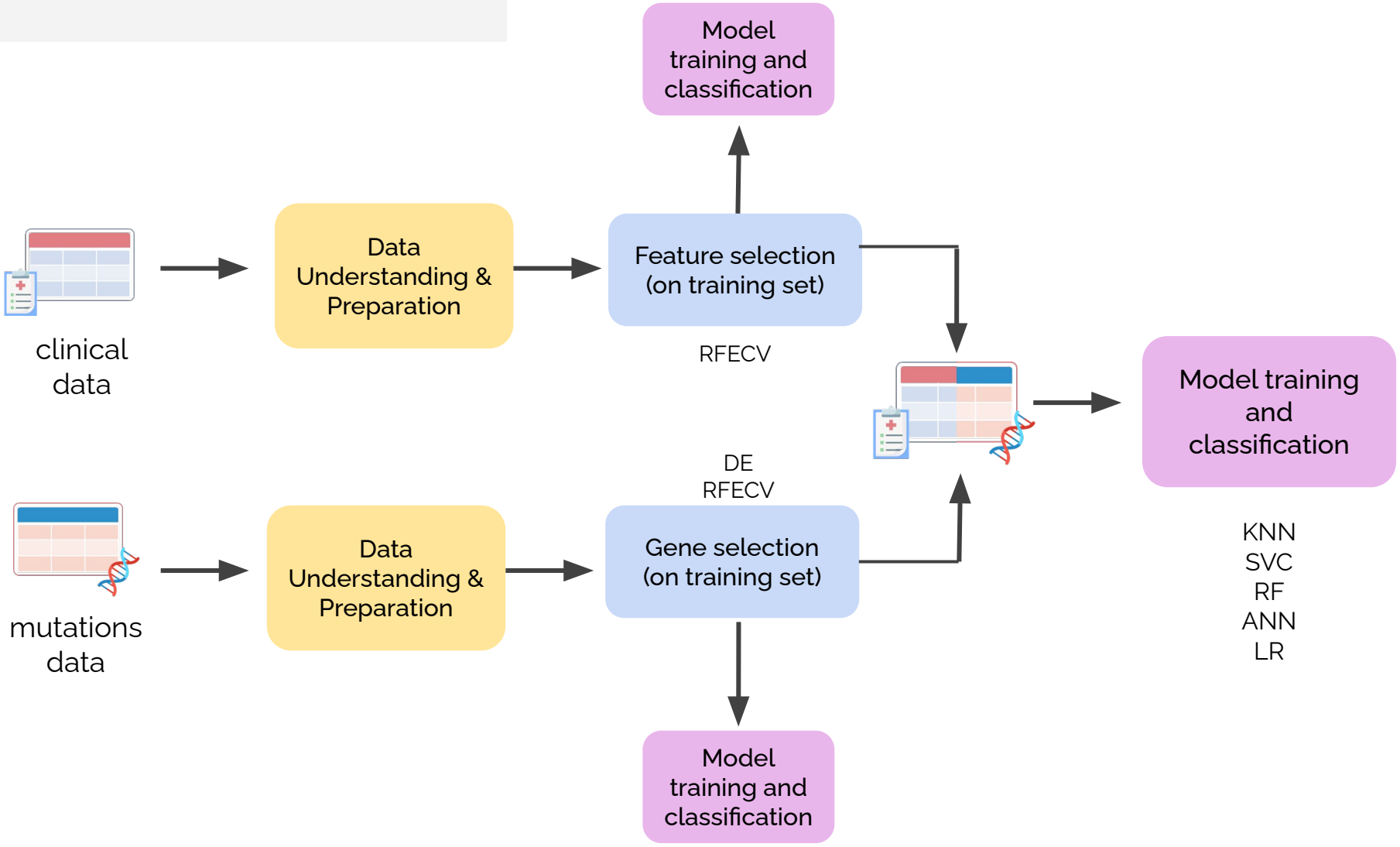
LIMITATIONS

small sample size

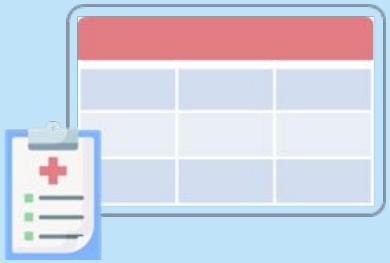
data simplification

State of the art

Most used method to predict survival in health care are ML models like **SVM**, **Random Forest**, **Logistic Regression** and **Neural networks**



Data Understanding



Clinical Data

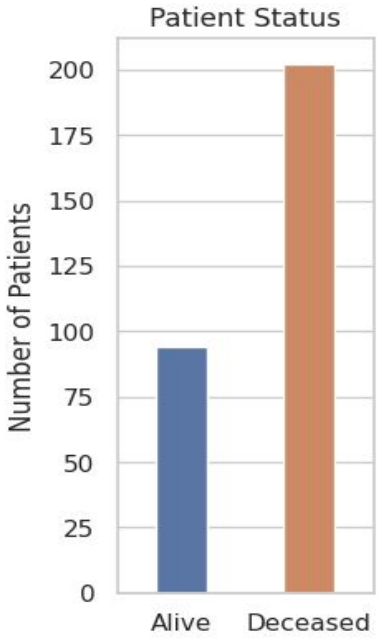
- 296 patients
- 68 attributes:
  - age
  - chemotherapy
  - tumor stage
  - metastasis location
  - symptoms
  - ...

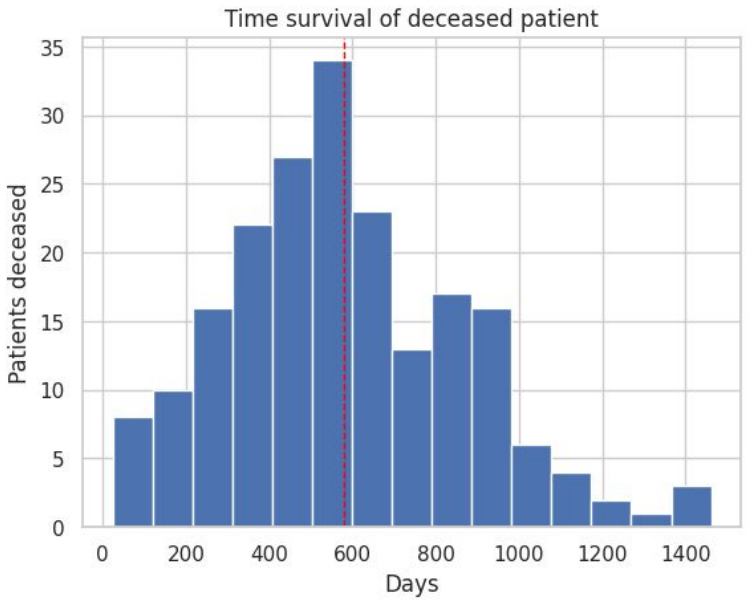
Mutation Data

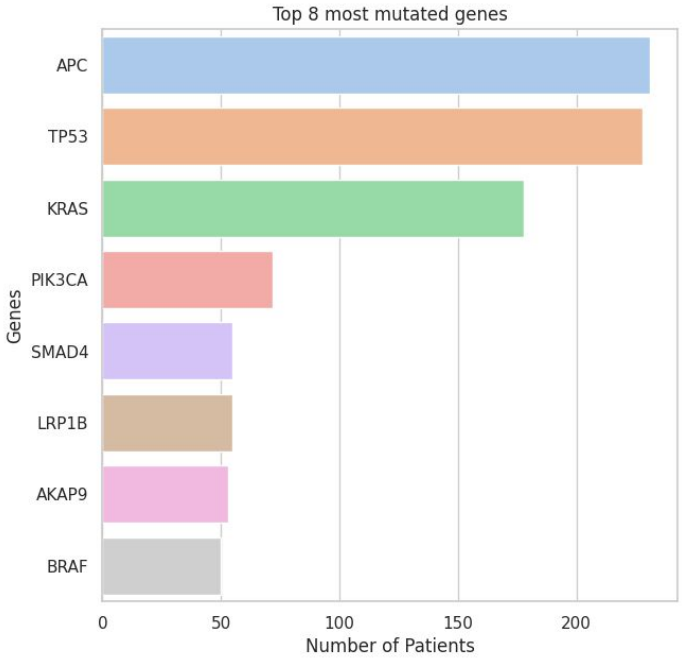
- ≈ 1.000 rows per patient
- 595 different genes
- two different technologies for most of the genes
- categorical values

Data Understanding

- Distributions
- Correlations
- Missing values







Data Preparation

Missing values
total elimination of rows with NaN or unknown values instead of data imputation (lost 30 patients)

Data mutation extraction
decided to keep only “NGS Q3” technology and its numeric values

Technology	Biomarker	Conclusion	TestResult	NGS_PercentMutated
NGS Q3	BRCA1	No Result	Mutated, Variant of Unknown Significance	47
CNA	BRCA1	Amplification Not Detected	Amplification Not Detected	

Gene selection with DE

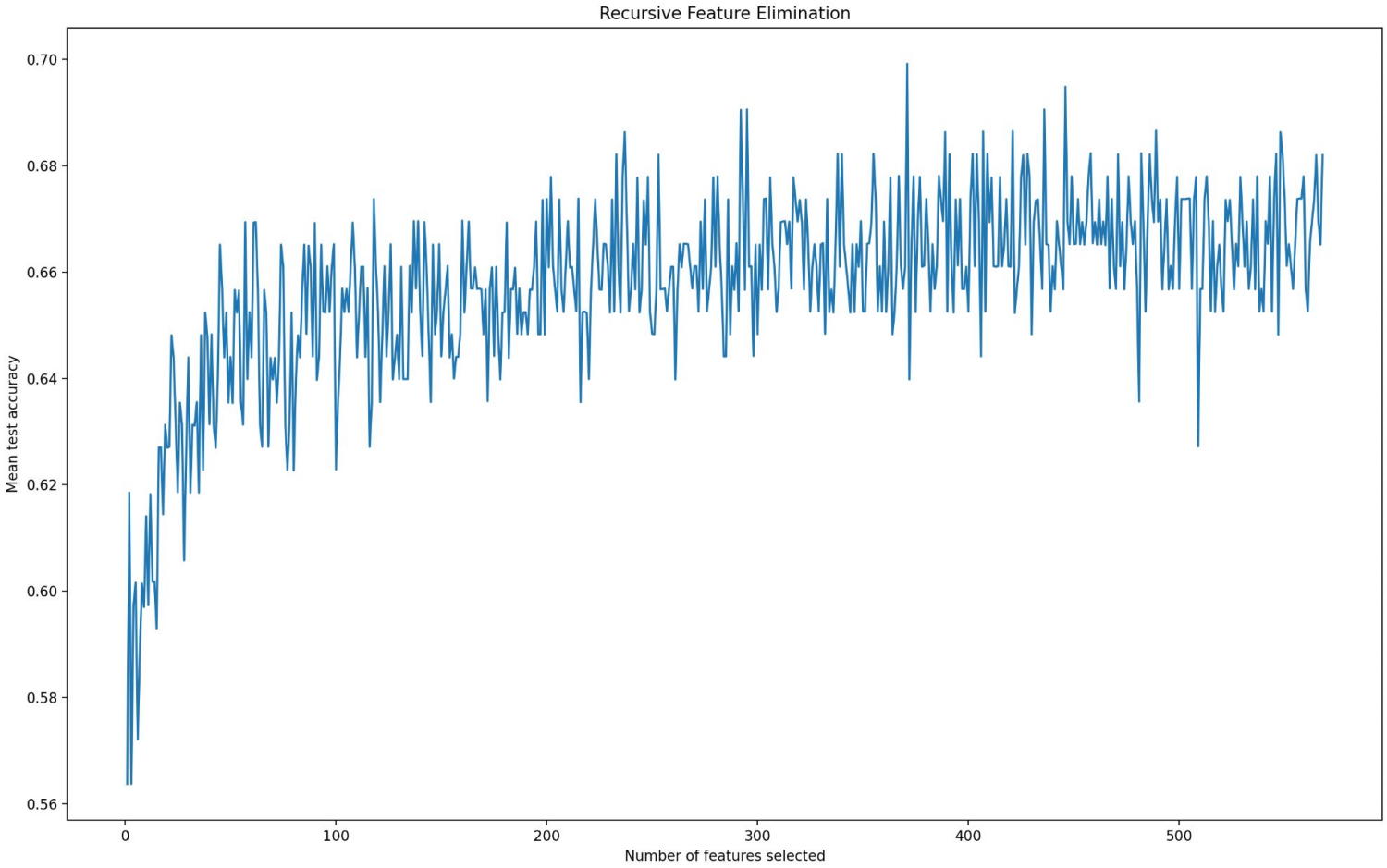
- Removed genes not available for every patient
- Removed genes with no mutation for anyone
- Differential Expression on training set based on P.Value and logFC
- Collected top 20 genes

```
"KDM5A"   "PIK3R1"  "TET1"
"ATP1A1"  "ERCC3"   "PIK3CG"
"FUS"     "CARD11"  "LRIG3"
"NFE2L2"  "CASP8"   "FAS"
"ERC1"    "KMT2C"
"IDH2"    "NF1"     "RET"
"SMARCA4" "PMS1"    "PIK3CA"
```

Experiments on Mutations Dataset

- Data with biomarkers and their percentages
- Features including ‘dos’ from clinical data
- RFECV proposing ‘dos’ and 4 other genes as the best features
- ‘dos’ as a very good predictor but has to be dropped

RFECV on Mutations Dataset

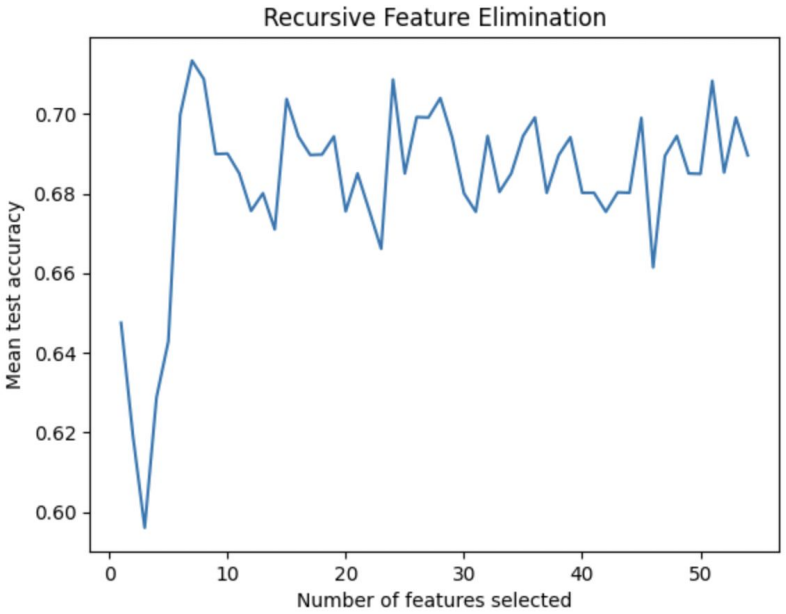


Classification on Mutations Dataset

Model	Acc (TS)	F1 score (TS)	AUC (TS)
RF	71.67%	0.82	0.55
KNN	71.67%	0.82	0.55
SVC	71.67%	0.82	0.55
ANN	71.67%	0.82	0.54

RFECV on clinical data

- Estimator: RandomForest
- Selected features: **{age, outcome of surgery, RTK/RAS, WNT, HIPPO, CELL_CYCLE, TP53}**



- High percentage in CELL_CYCLE, high division/grow, higher risk
- TP53 encodes tumor suppressor protein

RFECV on mutations data

- Estimator: RandomForest
- Run with different seeds and combine results
- Selected features: **{FUS, ATM, ERCC3, ERC1}**

    - ATM: helps prevent cancer ; regulates variations of protein like **p53**; eligibility criteria in **115 clinical trials!**
    - ERCC3: DNA nucleotide repair (even small mutation -> large damage); **27 clinical trials**.

source: My Cancer Genome

Classification on clinical data

Model	Acc (TR/TS)	F1 score (TR/TS)	AUC (TR/TS)
RF	100/70%	1.0/0.8	1.0/0.61
KNN	78/69%	0.85/0.78	0.71/0.59
SVC	73/75%	0.82/0.83	0.64/0.66

Table 3: Accuracy, `F1` score and `AUC` curve of the models on clinical data

Classification on merged data

Model	Acc (TR/TS) (%)	F1 score (TR/TS)	AUC (TR/TS)
LR	75/77%	0.82/0.85	0.75/0.72
RF	1.0/68%	1.0/0.38	1.0/0.58
KNN	74/70%	0.46/0.30	0.64/0.56
SVC	80/68%	0.65/0.34	0.74/0.56
ANN	76/78%	0.83/0.85	0.71/0.68

Table 5: Accuracy, `F1` score and `AUC` curve of the models on merged data

Conclusions

- ➔ Best model obtained on merged data
- ➔ Best results with ANN

Model	Acc (TR/TS) (%)	F1 score (TR/TS)	AUC (TR/TS)
ANN	76/78%	0.83/0.85	0.71/0.68

Future work

- ❑ Use both NGS and CNA mutations results
- ❑ Do data imputation with a domain expert
- ❑ Collect more data (especially for alive patients)

THE END

Thank you for your attention