Machine Learning Approaches to Predict 6-Month Mortality Among Patients With Cancer

Tan, W. Q., Ang, Z. E., Ling, Z. J., Loh, J. W. Z., Chee, C. E., Chan, N., Chong, W. Q., Walsh, R. J., Ow, S., Wong, R., Choong, C., Ooi, M., Wong, A., Tan, T. H., Yee, C. M., Zhu, K., Tan, J. H., Liu, H. J., Yuen, Y. C., Boo, A. R., Chua, C. T., Wang, Y. J., ... Barr Kumarakulasinghe, N..

National University Cancer Institute, National University Health System, Singapore

Regional Health System Office, National University Health System, Singapore

INTRODUCTION

The importance of predicting short-term mortality

Recognising when patients are at the end-of-life is vital to trigger the cascade of treatment decisions that ensure patient centred and value driven care. Crucially, it prevents unnecessary cancer directed treatment with its well-known toxicities and allows the initiation of serious illness conversations between healthcare providers, patients, and their caregivers.

The challenges faced today

However, recognising patients at high risk of short-term mortality is challenging due to varying accuracy and consistency amongst healthcare providers.

This project's objective

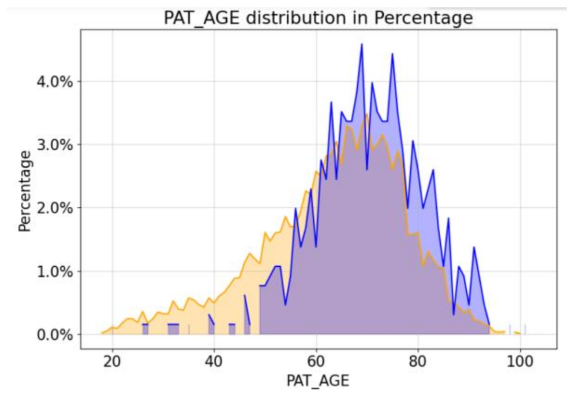
This study, done in Singapore, uses patient demographics, lab data and co-morbidities that is readily available in our Next Generation Electronic Medical Record (NGEMR),EPIC to predict mortality within 180 days after the last encounter.

LITERATURE REVIEW

- Various machine learning (ML) algorithms and models have been tested to predict short-term mortality among cancer patients, with tree-based modelling generally outperforming other ML algorithms.
- Many ML models are built for a specific subset of cancer patients (specific type of cancer, specific stages of cancer, post-operative or disseminated cancer patients)
- Predictions of mortality made are between a range of 30 days to 5 years in these studies.
- Datasets utilized range from publicly available datasets (e.g., SEER, MIMIC-III) where datasets are large, to self-collected data that are specific for the study, where datasets tend to be smaller.

PATIENT DEMOGRAPHICS

Mean Age



Alive (Orange): 63.5 [55.0-73.0]

Deceased (Blue): 70.1[63.0-77.8]

Gender :

Alive:	Male **35.9%**	Female **64.1%**
Deceased:	Male **55.5%**	Female **44.5%**

Latest Specialty Visited

Specialty Visited	Alive	Deceased
Chemotherapy Unit	90.7%	81%
Medical Oncology	7.9%	9.6%
Radiation Oncology	1.4%	9.3%

METHODOLOGY

The population is defined as patients with outpatient oncology encounter(s) between 01 October 2023 and 01 April 2024. Out of 10,985 patients, 654 (5.95%) patients died within 180 days after the latest encounter. Data was extracted from our institution's Electronic Health Record (EHR) system

Extract all completed/arrived outpatient oncology encounters between 01 October 2023 to 01 April 2024. Exclude patients age <18 (n=26,968)

Keep last encounter for each unique patient (n=10,985)

Extract patient age, gender, all diagnosis, labs taken in past 180 days, death date (if available)

- **Aggregate lab data** within 3 months and 6 month before encounter as count, min, max, std, first and last.
- **Aggregate diagnosis** data according to Elixhauser comorbidity categories.
- **Convert death date to death indicator** if death is within 180 days after encounter (5.95% death rate)

Split data into 70% train (death rate = 5.8%) and 30% test (death rate = 6.3%)

Impute missing diagnosis and lab counts and std with 0, and missing min, max, first, last using the median of train dataset for both train and test dataset.

Perform **feature selection** by removing highly correlated variables (correlation threshold 0.95), and variables with no mutual information.

Use data without imputation | Perform standard scaling for both train and test dataset, using mean and std from train dataset.

Train Random Forest | Train XGBoost | Train SVM

Optimise hyperparameters using randomised search using 5-fold CV

Predict test dataset

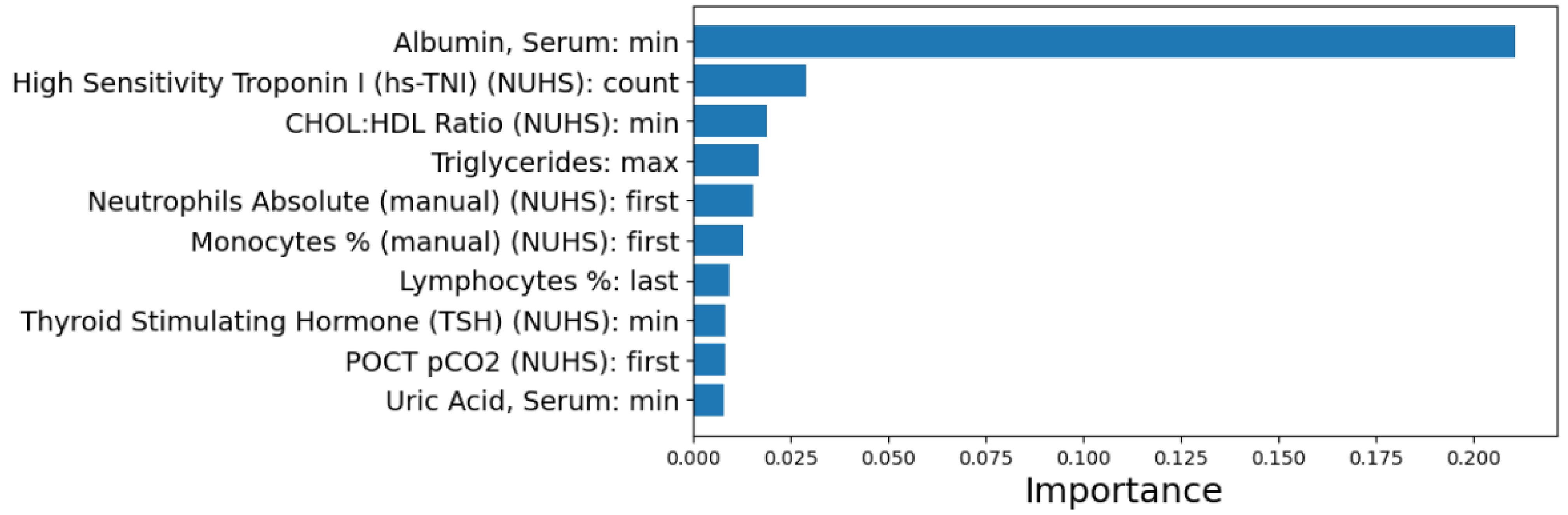
RESULTS

The models were evaluated based on AUROC scores. XGBoost and Random Forest outperformed the rest of the models.

Model	3 months of historical lab data, AUROC	6 months of historical lab data, AUROC
Random Forest	0.92	0.92
SVM	0.78	0.77
XGBoost (all features)	0.92	0.92
XGBoost (top 100 features)	0.91	0.90
XGBoost (top 50 features)	0.91	0.89

The feature importance plot shows that lowest albumin level recorded in the past 3 months contributes very significantly to the ML model's prediction.

Feature Importance Graph



DISCUSSION

- ML models can predict 6 month mortality from the EPIC EHR with good performance.
- 3 months laboratory data is sufficient to provide good mortality predictions.
- The model was able to achieve similar performance despite fewer features which can facilitate deployment in other institutions
- Lowest albumin level recorded in past 3-6 months plays a major role in mortality predictions
- Further model refinement, interpretability and seamless deployment are needed promote trust and acceptance amongst stakeholders.

Many thanks to Ravi B. Parikh, Medical Director, Winship Data and Technology Applications Shared Resource for sharing their work done for our learning and application.