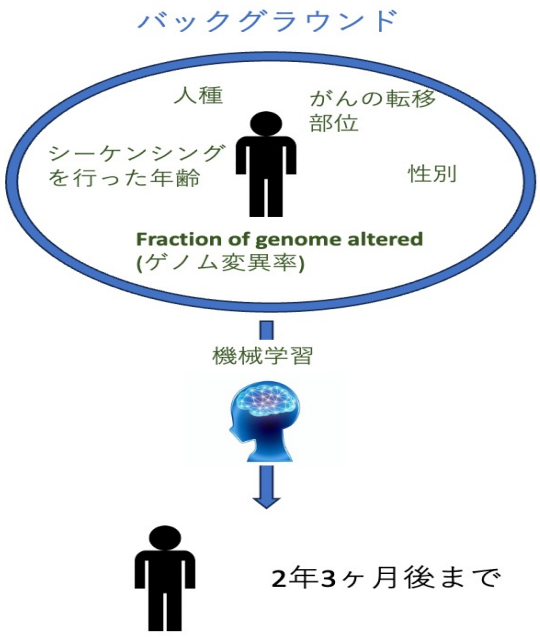
東京理科大学大学院　生命科学研究科　生命科学専攻　修士1年　下村功真
指導教員：波江野 洋教授

機械学習を用いたがん患者の寿命予測

背景・目的

がんは世界中で主要な死因の一つであり、患者の生存予測は治療を進めるにあたって極めて重要な指針の一つになる。従来の統計的手法による予測では患者データから個別に寿命を予測することは難しい。
一方、機械学習を用いると患者データから個別に精度の高い寿命の予測を行うことができ、効果的な治療の選択に繋がると考えられる。

本研究の目的は、XGBoostを含む５種の機械学習・統計的手法でがん患者の寿命を予測し、最も予測性能の高いモデルを特定することである。これにより、臨床現場で治療法の選択、スケジュールの手助けなどの実用化に向けた基礎を築き、がん治療の個別化および最適化を目指す。



本研究では、予測性能の高いモデルを見出すことで個別化医療を促進することを目的としている。

材料・方法

実装ツール
・プログラミング言語：python
・ライブラリ：numpy, pandas, optuna, matplotlib, tensorflow, XGBoost, scikit-learn

研究フロー

I.データセットの収集 → II.データの処理 → Ⅲ. 機械学習 → Ⅳ.予測性能の比較

結果・考察

1.データセット

Nguyenらの論文(Cell 2022)では25,000人の患者の臨床シーケンスデータを用いて、がんの転移パターンとゲノム的特徴を紐付ける研究が行われた (Table. 1)。本研究では目的変数をOverall Survival(OS)と設定し、このOSを予測した。OSはシーケンシングを行ってから患者が亡くなるまでの期間(月単位)と設定した。

2.データの処理

モデルが各特徴量を均等に扱えるようにデータの標準化を行った。続いて、特徴量同士の相関(Table. 2)が大きいデータ(相関係数の絶対値の閾値 :0.95)を削減した後、主成分分析(PCA)を行った。これにより、多重共線性の問題を解決し、データの重要な情報を保持したまま次元の削源を実現した。

Table.1 Nguyenらのデータ（がん種と症例数）

がん種	症例数
Lung Adenocarcinoma(肺がん)	1361
Colon Adenocarcinoma(大腸がん)	747
Pancreatic Adenocarcinoma(膵臓がん)	917
Prostate Adenocarcinoma(前立腺がん)	515
Breast Invasive Ductal Carcinoma(乳がん)	724
High-Grade Serous Ovarian Cancer(卵巣がん)	313

Table.2 肺がん患者における特徴量同士の相関(一部抜粋)

	Mutation_Count	Sample_coverage	TMB	Tumor_Purity	IMPACT410	IMPACT468	Indeterminate	Stable
Mutation_Count	1.00000	0.05037	0.98999	0.01312	0.04540	0.03607	0.03615	−0.03439
Sample_coverage	0.05037	1.00000	0.04181	0.11051	0.17070	−0.05717	0.04154	−0.06033
TMB	0.98999	0.04181	1.00000	0.01821	0.05958	−0.04698	0.03929	0.00280
Tumor_Purity	0.01312	0.11051	0.01821	1.00000	0.02188	−0.05291	0.10049	−0.04889
IMPACT410	0.04540	0.17070	0.05958	0.02188	1.00000	−0.59859	0.05891	0.20983
IMPACT468	0.03607	−0.05717	−0.04698	−0.05291	−0.59859	1.00000	−0.04104	−0.43496
Indeterminate	0.03615	0.04154	0.03929	0.10049	0.05891	−0.04104	1.00000	−0.52654
Stable	−0.03439	−0.06033	0.00280	−0.04889	0.20983	−0.43496	−0.52654	1.00000

3. 機械学習

はじめに、各患者の観測値の分布を確認した(Fig. 1)。次に、PCAを行い、累積寄与率が90 %を超える主成分までを特徴量として選択した。これらの特徴量をUMAPに適用し、得られたUMAPデータを用いてK-meansによるクラスタリングを実施した。また、UMAPデータを用いてIsolation Forestによる異常検知を行い、これらのK-means、UMAP、Isolation Forestの結果を新たな特徴量として追加し、生存期間の予測に用いた。学習にはXGBoost、RandomForest、Lasso、Ridge、ElasticNetの５種のモデルを使用した。今回特に注目しているXGBoostは、RandomForestと同様に決定木を組み合わせた機械学習モデルであり、勾配ブースティングと呼ばれる逐次学習によって損失関数を最小化する方法である。

4. 予測性能の比較

それぞれのモデルで訓練データを用いた学習を実施した後、テストデータを用いた二乗平均平方根誤差(RMSE)を用いて精度の比較を行った(Table. 3)。

考察

本研究の結果、どのモデルもRMSEの値が高く、予測精度は期待に及びませんでした。この主な原因として、データ数の不足や治療データの欠如、そして転移部位情報が予測に大きく寄与しなかったことが挙げられます。特に、症例数が少なかったことでモデルが十分に学習できず、予測精度が低下したと考えられます。また、治療データが含まれていないため、モデルが生存期間の予測において不確実性を増している可能性があります。さらに、Fig.2で示されたように、転移部位情報の寄与が低かったことも、モデルが生存期間の変動を十分に捉えられなかった要因として考えられます。

Fig.1　肺がん患者のヒストグラム
(縦軸：人数、横軸：OS(month))

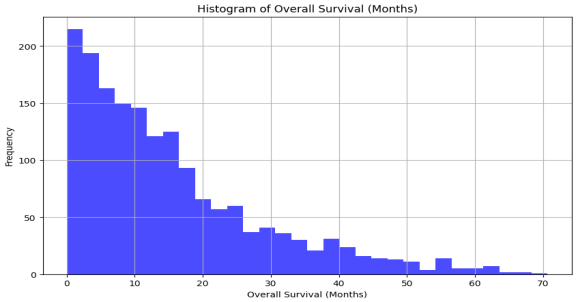


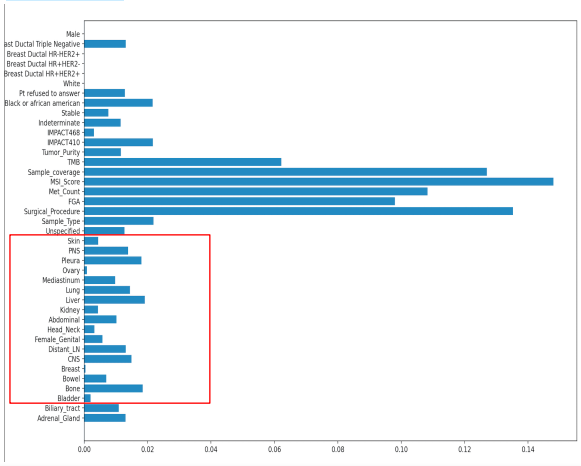
Table.3　RMSE(month)

モデル	Elastic Net	Lasso	Ridge	RandomForest Regressor	XGBRegressor
肺がん	10.809	10.809	10.477	11.352	12.657
大腸がん	11.324	11.324	10.339	11.653	12.992
乳がん	14.781	14.781	14.786	15.079	18.797
卵巣がん	9.066	9.066	9.069	9.631	11.117
膵臓がん	9.272	9.272	9.083	9.039	10.704
前立腺がん	14.635	14.630	14.630	15.763	16.805

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2}$$

（ y_i:実際の値、 $\hat{y}_i$:予測値、 n:データの総数 ）

Fig.2　Feature importance(breast)



総括・展望

今後は、データ量の増強や治療データの統合を通じて、モデルの予測精度を向上させます。また、次元削減技術を活用し、新しい特徴量を導入することで、モデルの性能をさらに高めます。

私はこの研究結果やデータ解析技術を活かし、臨床現場で適用可能な治療法を提示することで、患者さんの助けになりたいと考えています。