

知識発見と特徴選択のための自動化予測モデリング②

～最新のAutoMLツールとの比較～

Ioannis Tsamardinos, Paulos Charonyktakis, Georgios Papoutsoglou, Giorgos Borboudakis, Kleanthi Lakiotaki, Jean Claude Zenklusen, Hartmut Juhl, Ekaterini Chatzaki & Vincenzo Lagani



npj Precision Oncology volume 6, Article number: 38 (2022)

doi: <https://doi.org/10.1038/s41698-022-00274-8>

Table 1. AutoML tools functionalities. The presence of a functionality in each tool is marked by a check symbol or by a short name/acronym. Asterisks and short names are explained in the text.

	JADBio	auto-sklearn	TPOT	TPOT-MDR	GAMA	AP
CASH*	✓	✓	✓	✓	✓	✓
BioSignature Identification*	✓			✓*		
(Multiple) Signatures Identification	✓					
Selected Feature Interpretations	ICE plot					
Selected Feature Added Value	✓					
Explains individual predictions						✓
Predictive Performance Estimation	GCV, BBC-CV	Holdout, CV*	CV	CV	CV*	CV
Confidence Intervals	✓					✓
Classification (nominal) outcome	✓	✓	✓	✓	✓	✓
Regression (continuous) outcome	✓	✓	✓	✓	✓	
Time-to-event (survival) outcome	✓					
Optimization of Clinical Thresholds	✓					
Accepts missing values	✓	✓	✓	✓	✓	✓
Handles clustered samples	✓	✓				

*See [research paper](#) for details on asterisks

比較に使用したデータと方法

トランスクリプトミクス (271 マイクロアレイ、23 RNA-seq)、エピゲノミクス (23)、およびメタボロミクス (43) データを含む 360 の高次元データセット。データセットは 125 の疾患または表現型に関連しており、細胞増殖疾患 (つまり、さまざまな種類の癌) が 154 のデータセットで最も代表的なグループです。JADBio は、パフォーマンス (JADBio-P)、解釈可能性 (JADBio-I)、特徴選択 (JADBio-FS)、積極的・特徴選択 (JADBio-AFS) の 4 つの異なる基準に対して最適化されています。auto-sklearn¹⁷、TPOT¹⁸、GAMA²⁰、AutoPrognosis²¹、および Random Forest (RF²², scikit-learn implementation³¹) と比較します。各データセットは 2 つの同じサイズのサンプルセットに分割され、それぞれがトレーニングセットとテストセットとして使用され、合計 720 の分析が行われました。詳細は **Methods** に記載。

Results

JADBio は知識の発見を可能にし、高い予測モデルを提供します。

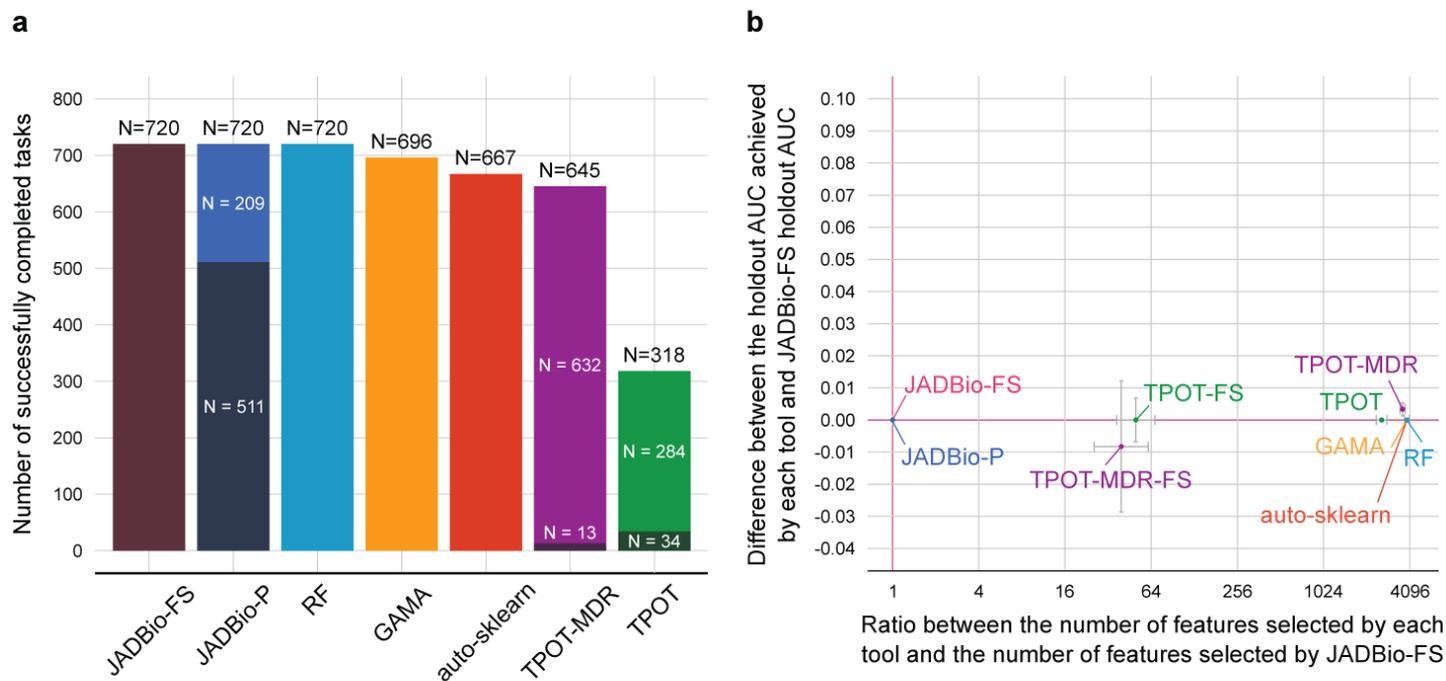


Fig. 4 予測性能に関する定量的比較の結果①

a. 各ツールの 720 個のタスクのうち、正常に完了したタスクの数。AutoPrognosisは一度もランを完了しないため、表示されません。JADBio-P と TPOT-MDR は特徴選択を試みます。特徴選択がwinning modelにつながったランの部分は、それぞれの列が濃い色で示されています。JADBio-FSは常に特徴選択を強制します。b. JADBio-FS と他のすべてのツールとの相対的な次元削減とパフォーマンスの向上に関する比較。JADBio-FS に関する次元削減 (x 軸) は、各ツールによって選択された特徴の数と JADBio-FS によって選択された特徴の数 (log2 スケール) の比率として計算されます。1より大きい値は、ツールがJADBioよりも多くの特徴を選択することを示します。パフォーマンスの向上は y 軸で報告され、各ツールが達成したホールドアウト AUCとJADBio-FSが達成したホールドアウトAUCの差として計算されます。0より大きい値は、ツールのパフォーマンスが JADBio-FS よりも優れていることを示します。各ツールについて、各軸の中央値上にポイントをプロットしています。標準誤差を示すエラーバーが表示されます。中央値は、完了したランでのみ計算されます。クランベリー色の線はベースライン値を示します。つまり、相対的な次元削減は 1 に等しく、パフォーマンスの向上は 0 に等しくなります。

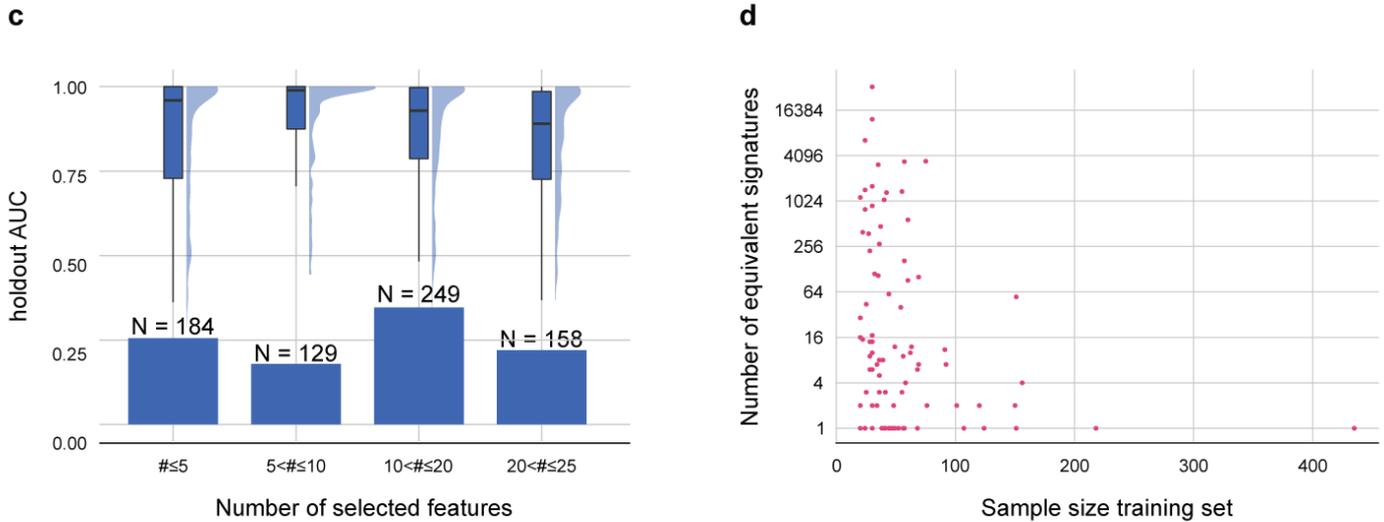


Fig. 4 予測性能に関する定量的比較の結果②

c. 異なる数の特徴選択に対する JADBio-FS のホールドアウト AUC 分布。ホールドアウト AUC 分布は、ボックスプロットとバイオリンプロットの両方で報告されます。d. トレーニングセットのサンプルサイズ (x 軸) と、JADBio-FS によって識別された同等のシグネチャーの数 (y 軸、log₂ 変換) を対比させたもの。複数のシグネチャーを識別できるアルゴリズムが選択されたランのみ表示されます。1 より大きい y 座標を持つポイントによって表される 98 ランのうち、合計 68 ランで複数のシグネチャーが識別されました。

結果、JADBio が、実験に含まれる他のすべてのツールに対して同等の予測性能を達成するために必要なバイオマーカーの測定数を約 4000 個少なくするモデルを導き出すことができました。

JADBio はパフォーマンスの過大評価を回避します

トレーニング セットから取得したパフォーマンス推定値が、ホールドアウトセットで計算されたパフォーマンス推定値を正確に反映しているかどうかを調査します。これは、モデルの臨床的有用性を正しく評価するために重要です。

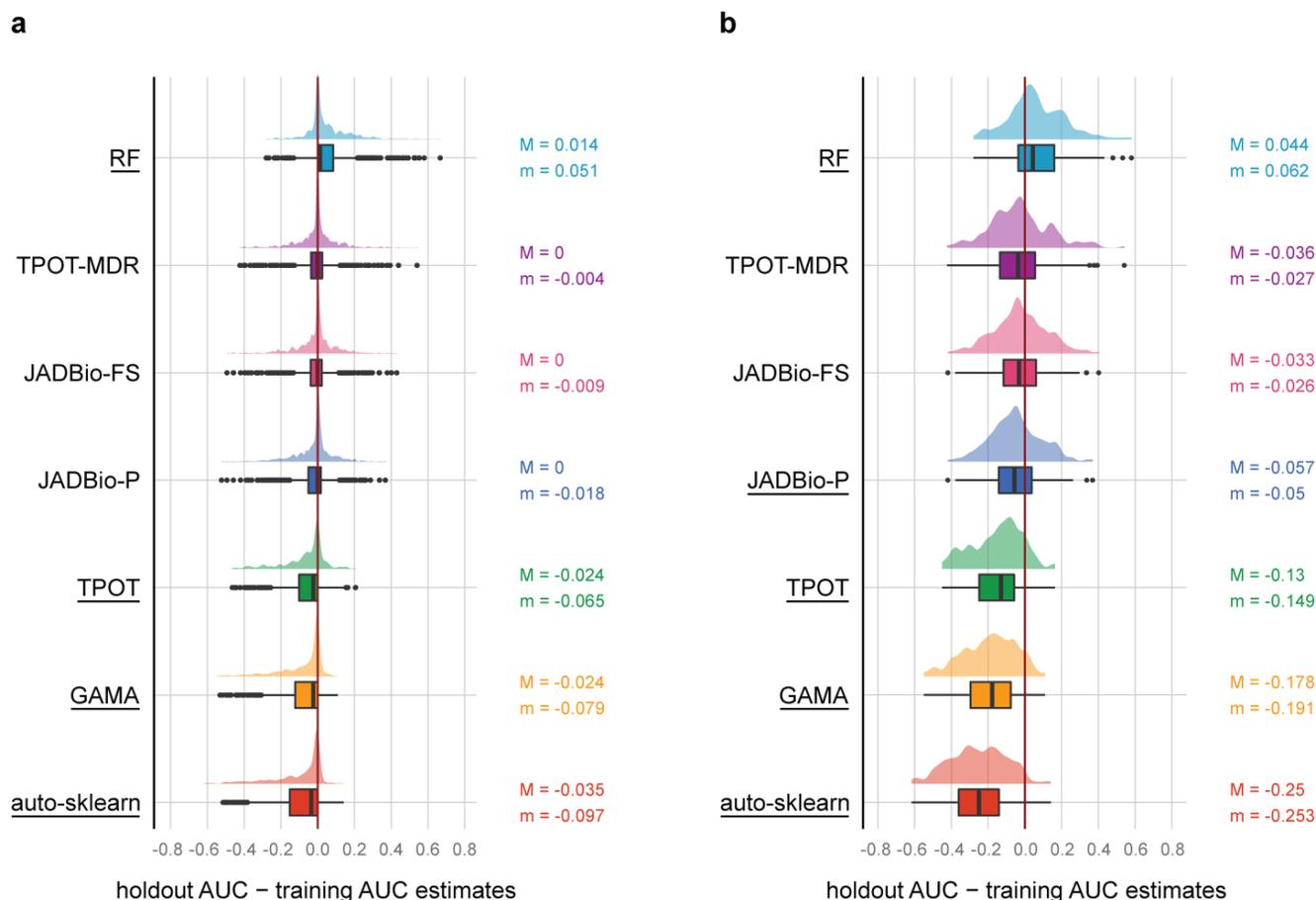


Fig. 5 バイアスに関する定量的比較の結果①

バイアスは、ホールドアウト AUC 推定値からトレーニングセットで推定されたAUCを引いたものとして定義されます。ゼロラインの右側の分布は保守的な推定値に対応し、ラインの左側の結果は楽観的なものです。a. 720 の分類タスク (360 のデータセットごとに 2 つのタスク) にわたるAUCバイアス分布。各ツールの分布は、ボックスプロットとバイオリンプロットの両方でレポートされます。バイアスの中央値がゼロから統計的に有意に異なるツールは、名前が下線付きで表示されます。中央(M)と平均(m)のバイアス値は、各ボックスプロットの軸の右側に表示されます。b. aと同じですが、ホールドアウトセットでRandom Forest (RF)が 0.8 以下のAUCを達成した196回のランについてのみです。これらのより困難なタスクでは、TPOT、GAMA、およびauto-sklearnは、14 ~ 25 AUC ポイントのパフォーマンスを体系的に過大評価します。

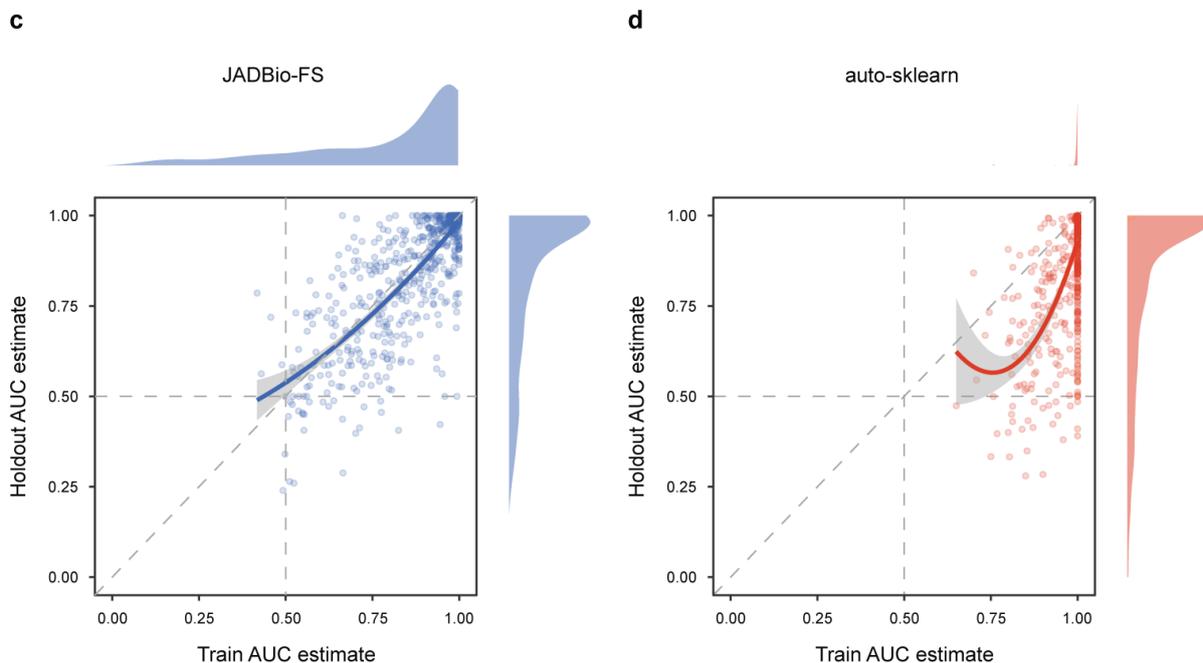


Fig. 5 バイアスに関する定量的比較の結果②

c. と d. JADBio-FS ツールと auto-sklearn ツールのそれぞれについて、トレーニングセットで推定された AUC (x 軸) とホールドアウトセットで推定された AUC (y 軸) を報告する散布図。各ポイントは 1 回のランに対応します。Locally Weighted Scatterplot Smoothing (LOESS) 回帰線は、対応する散布図の主な傾向を示します。auto-sklearn ポイントと LOESS ラインはほとんどが対角線より下にあり、トレーニングセットのパフォーマンス推定値が体系的に過度に楽観的であることを示しています。JADBio-FSでは、対角線に近いLOESS回帰線で示されるように、バイアスが大幅に減少していることがわかります。

Fig. 5のパネル(c)と(d)は、AutoMLプラットフォーム間の推定プロパティの違いをよりよく視覚化しています。各ドットは 1 回のランに対応し、x軸はトレーニングセットのAUC推定値、y 軸はホールドアウトセットの推定値を報告しています。推定トレーニングAUCは、達成されたホールドアウトAUCに対してプロットされます。対角線の上/下の点は、パフォーマンスを過小評価/過大評価するランに対応します。LOESS回帰線は太字で示されています。auto-sklearn はほとんどの場合、パフォーマンスを過大評価することに注意してください。パネルの右側にあるドットのストライプは、AUC = 1のauto-sklearn推定に対応していますが、ホールドアウトAUCは0.5未満から満点1のまでの範囲です対照的に、JADBio-FSのランは対角線の周りに対称的に分布しており、トレーニングセットからのJADBio-FSの推定値が平均して正確 (ゼロバイアス) であることを示しています。もちろん、推定にはまだばらつきがあるため、JADBioは信頼区間も報告しています。他のJADBio設定ではJADBio-FSよりもパフォーマンス推定値が保守的であるが、これは調査した構成数が少ないためと考えられます (Supplementary Table 9/元論文参照)。

これらの結果をまとめると、JADBioがトレーニングセットのみから信頼できる汎化パフォーマンス推定値を推定できるが、auto-sklearn、GAMA、および TPOTはこの点で失敗していることがわかります。

JADBioは、独立したホールドアウトセットを必要とせず、アウトオブサンプルパフォーマンスの正確な推定値を提供します。このことは、**Fig. 5**と**Supplementary Table 6**(元論文参照)の結果からも裏付けられます。ユーザーがパフォーマンスを推定ためにサンプルを失う必要はありません。最終的なモデルは、すべてのサンプルでトレーニングされます。JADBioは、パフォーマンスとその不確実性(信頼区間)の推定を自動的に処理します。つまり、最終的なモデルを統計的に検証するために、別途テストセットを用意する必要はないということです。これは、希少な疾患やコストのために少数の患者を含むことが多いオミクスデータセットで特に重要です。ただし、推定は同じ動作環境内でのみ有効であるという免責事項があります。バッチ効果またはその他の分布の変更が可能である場合は、別の外部検証セットを使用する必要があります。JADBioのこの特性は、解析の完全自動化に必要なものです。代わりに、GAMA、TPOT、および auto-sklearn による推定は、体系的に過大評価することが示されています(**Fig. 5**)。これは、そのようなツールは、トレーニングセットのパフォーマンス推定のために考案された実験プロトコルを採用していないためと考えられます。



【お問い合わせ先】

フィルジェン株式会社 バイオインフォマティクス部

TEL : 052-624-4388 (9:00~17:00)

FAX : 052-624-4389

E-mail : biosupport@filgen.jp