WRITE-UP SUMMARY

Introduction

Auto Machine Learning (AutoML) has massive potential to transform healthcare by automating complex data analysis, increasing accurate diagnostics, and accelerating medical discoveries. The DRT/CloudLeap team is excited to evaluate AutoML's efficacy as part of the precisionFDA AutoML App-a-thon challenge. Our results show that open-source AutoML tools can improve efficiency and performance with noted caveats.

Data and Preprocessing

We performed Exploratory Data Analysis and made the following observations.

Category	Observations
VOLUME	• Only 377 patient observations were provided.
INCOMPLETENESS	• 135 (35.8%) observations missing WHO_GRADING, 99 (26.3%) missing RACE, and 43 (11.4%) with NULL value in SEX. • 51 'UNKNOWN' values for CANCER_TYPE
BIAS	• WHITE accounts for 93.88% of RACE.
BALANCE	• Balanced datasets observed for WHO_GRADING vs SURVIVAL_STATUS as no WHO_GRADING was more prone to patient mortality.
IMBALANCE	• SURVIVAL_STATUS '1' accounts for 86.47% of all observations.
NULL	• No NULL value present among patient gene expressions.

AutoML tools provide different approaches for data preprocessing. For example, PyCaret automates the following actions upon ingestion, while H2O AutoML and Auto-SKLearn require manual preprocessing.

- Dimensionality reduction and data transposing
- Numeric value and categorical feature analysis
- Imputing missing values
- Normalization
- One-hot encoding
- Detecting outliers
- Fixing class imbalance
- Splitting (0.8/0.2 ratio) data into training & testing sets
- Creating n-folds for cross-validation.
- Handling class imbalance (8 over-sampling & 11 under-sampling)

AutoML Selection

We selected three open-source AutoML tools (see table below). We chose Google Colab for development, which provides powerful computing (GPUs/TPUs) and collaboration utilizing open-source libraries such as Pandas and NumPy.

AutoML Tool	Description	Integration with ML Model
H2O AutoML	Supports R & Python	- User-friendly, lightweight, well-documented - Suited for rapid prototyping
PyCaret	Python-based, low-code, providing end-to-end automation.	- Wide range of algorithms & features, simple to use, versatile - Integrates well with Python libraries
Auto-SKLearn	Automatic model selection, easy hyperparameter tuning, scalability & meta-learning.	- Balanced option with good performance - Experienced library compatibility issues

We conducted rigorous hyperparameter tuning across frameworks. In PyCaret, we initialized classifiers, conducted preprocessing, and configured key parameters like parallel processing, cross-validation, and addressed class imbalances. We leveraged parallel computing using Apache Spark, compared and identified the best-performing model, and then refined it with model blending and stacking. In H2O, we trained a range of ML models while optimizing hyperparameters and employing cross-validation for generalization and avoidance of overfitting. We selected the top-performing model based on scoring metrics. Within Auto-SKLearn, we developed a variety of ML models, incorporating resampling techniques and hyperparameter tuning through Bayesian optimization and ensemble methods. We addressed class imbalances and used the holdout technique for model evaluation and optimization.

Collectively, these efforts yielded substantial performance improvements documented in the following Jupyter Notebooks.

- PyCaret: Link
- H2O: Link
- Auto-SKLearn: Link

PyCaret prioritizes simpler models like Logistic Regression and Naive Bayes. Conversely, H2O specializes in more complex ensemble models such as GBM and XGBoost, offering superior performance and predictive capabilities. Auto-SKLearn tries to simplify the process, potentially reducing manual intervention. The table below lists the top five algorithms each tool uses and the best performing model.

AutoML Tool	Top 5 Algorithms	Best Performance
PyCaret	1. Logistic Regression 2. Random Forest Classifier 3. Light Gradient Boosting Machine 4. Extra Trees Classifier 5. SVM - Linear Kernel	Logistic Regression 89.47% Accuracy True Positives: 97/99 True Negatives: 5/15
H2O	1. Gradient Boosting Machine 2. Generalized Linear Model 3. XGBoost 4. Stacked Ensemble 5. Distributed Random Forest Model	Gradient Boosting Machine 85.05% Accuracy True Positives: 68/70 True Negatives: **6**/17 (highest survival prediction)

AutoML Tool	Top 5 Algorithms	Best Performance
Auto-SKLearn	1. Random Forest Classifier 2. Gradient Boosting Machine 3. Multi-Layer Perceptron 4. Extra Trees 5. k-Nearest Neighbors	Random Forest Classifier 86.84% Accuracy True Positives: 65/67 True Negatives: 2/9

Different frameworks yield different outcomes for similar models. This discrepancy arises from different hyperparameter optimization methods. Combining predictions from multiple models could enhance overall performance, particularly for minority class predictions. For instance, consolidating predictions of class '0' from PyCaret and H2O models could improve accuracy and provide a comprehensive view. Currently, PyCaret predicts 5 class '0's and H2O predicts 8, with 2 overlapping on the test dataset. Combining these results could accurately identify 11 class '0's. The Auto-SKLearn predictions are all 1's for the provided test data.

Lessons Learned

The data quality issues (missing & unknown values) in the training dataset could introduce unwanted training bias. One could try to re-classify these observations to the defined CANCER_TYPE, but the results would still raise ethics/bias concerns.

The low volume of the training dataset could adversely impact the performance and effectiveness of AutoML modeling. A larger training dataset provides more representative samples and improves coverage/diversity. It can help prevent overfitting, where the model learns to memorize the training data rather than generalize to new, unseen data. In our experiments, gene expression was the dominant feature and impacted the model the most.

A key obstacle to integrating AutoML in healthcare is its "black-box" nature; however, it is feasible to improve transparency and lower technical barriers for AutoML adoption by using open-source libraries and improving documentation geared to nonspecialists.

It is still important to keep human oversight of AI/ML a top priority despite its great potential in healthcare. The FDA has established an AI Governance and Advisory Board and developed AI Playbook to oversee AI usage for mission support. We believe focusing on transparency and collaboration are critical to effectively harnessing the power of AI to improve public health, while ensuring safety, efficacy, and security.

Ethics/Bias/Transparency Concerns

The UNKNOWN CANCER_TYPE in the training dataset raised potential cognitive bias. If the training dataset is not sufficiently investigated for quality, diversity, or fit to be generalized to real-world data (e.g., microarray vs RNA-seq), it can produce an inaccurate model. The AutoML model and tuning methods should be documented and easily accessible for external validation from trained biomedical professionals, statisticians, and data scientists in all stages of model generation and maintenance. An independent review board should validate the model and provide feedback. We suggest publishing the model in a peer-reviewed scientific journal. Any bias discovered should be eliminated if possible; noted and made available if not.