BioData Mining

SHORT REPORT Open Access

Benchmarking AutoML frameworks for disease prediction using medical claims

Roland Albert A. Romero[1], Mariefel Nicole Y. Deypalan[1], Suchit Mehrotra[1], John Titus Jungao[1], Natalie E. Sheils[1], Elisabetta Manduchi[2] and Jason H. Moore[2*]

*Correspondence:
jason.moore@csmc.edu
[1]OptumLabs, Minnetonka 55343, MN, USA
[2]Department of Computational Biomedicine, Cedars-Sinai Medical Center, 700 N. San Vicente Blvd., Pacific Design Center Suite G540, West Hollywood 90069, CA, USA

Abstract

Objectives: Ascertain and compare the performances of Automated Machine Learning (AutoML) tools on large, highly imbalanced healthcare datasets.

Materials and Methods: We generated a large dataset using historical de-identified administrative claims including demographic information and flags for disease codes in four different time windows prior to 2019. We then trained three AutoML tools on this dataset to predict six different disease outcomes in 2019 and evaluated model performances on several metrics.

Results: The AutoML tools showed improvement from the baseline random forest model but did not differ significantly from each other. All models recorded low area under the precision-recall curve and failed to predict true positives while keeping the true negative rate high. Model performance was not directly related to prevalence. We provide a specific use-case to illustrate how to select a threshold that gives the best balance between true and false positive rates, as this is an important consideration in medical applications.

Discussion: Healthcare datasets present several challenges for AutoML tools, including large sample size, high imbalance, and limitations in the available features. Improvements in scalability, combinations of imbalance-learning resampling and ensemble approaches, and curated feature selection are possible next steps to achieve better performance.

Conclusion: Among the three explored, no AutoML tool consistently outperforms the rest in terms of predictive performance. The performances of the models in this study suggest that there may be room for improvement in handling medical claims data. Finally, selection of the optimal prediction threshold should be guided by the specific practical application.

Keywords: Automated machine learning, AutoML, Machine learning, Healthcare, Medical claims, Class imbalance

Background and significance

Leveraging big data growth in biomedicine and healthcare, machine learning (ML) has helped improve health outcomes, cut healthcare costs, and advance clinical research [1–4]. Studies applying ML to healthcare data range from models for disease prediction or for improving quality of care, to applications such as detection of claim fraud [2, 5–8]. Clinical big data used in various studies range from electronic health records, medical records, and claims data. Many studies are limited to a single healthcare or hospital system [9–12].

Despite the demonstrated benefits of machine learning, different models need to be trained in the context of the problem to achieve good performance [13]. For each model, domain experts such as clinicians need to collaborate with data scientists to design ML pipelines [14]. Automated machine learning (AutoML) is an emerging field [15] that aims to simplify this labor-intensive process [16] which can accelerate the integration of ML in healthcare scenarios [1]. State-of-the-art AutoML platforms allow domain experts to design decently performing ML pipelines without deep knowledge of ML or statistics while at the same time easing the burden of tedious manual tasks such as model selection and hyperparameter optimization for data scientists [14].

With ML being adopted across industries, standardized benchmarks and datasets are needed to compare competing systems [17]. These benchmark suites need to have datasets that highlight strengths and weaknesses of established ML methods [18]. Despite the emergence of numerous AutoML tools, there is still a need for standardized benchmarks in the field. Multiple studies to benchmark various AutoML tools[14, 19–21] have been done. Notably, Gijsbers et al. [22] presented an open-source AutoML benchmark framework to provide objective feedback on the performance of different AutoML tools. Gijsbers et al. compared four AutoML tools across 39 public data sets, twenty-two of which are binary classifications, with a mixture of balanced and imbalanced data. Of these, only two have very low prevalence for one class, at around 1.8% each. Most of these studies on benchmarks tested public datasets that have sample sizes in the order 103 and feature sizes between 10-100. In contrast, our study uses a population of over 12 million and over 3,500 features.

Although different AutoML tools will perform differently depending on the problem, there is a need to have benchmarks on datasets that have similar characteristics to healthcare data. Highly imbalanced and large datasets are common in healthcare and thus, these benchmarks will prove useful for accelerating the model-building process by identifying a good baseline model.

A review of published papers for AutoML showed that despite the potential applications and demonstrated need [23], little work has been done in applying AutoML to the field of healthcare [7]. Waring et al. determined the primary reasons for the lack of AutoML solutions for healthcare to be: (1) the lack of high-quality, representative, and diverse datasets, and (2) the inefficiency of current AutoML approaches for large datasets common in the biomedical environment. In particular, disease prediction problems often involve highly imbalanced datasets [24] which do not lend themselves well to predictive modelling. Disease prevalences are much lower than those of the public datasets used by Gjisbers et. al; the datasets we consider in this paper have positive prevalence ranging from 0.053% - 0.63%. The extremely low prevalence does not give the models enough samples from one class to train on.

Objectives

To advance the use of AutoML tools in healthcare, there is a need to first assess their performance in representative datasets. Doing so brings to light the challenges and limitations of using these tools on healthcare data and serves as the basis for future improvements to better address problems in healthcare. In this study, we generated a dataset using claims data with 12.4M rows and 3.5k features. Using this, we compared the performance of different AutoML tools for predicting outcomes for different diseases of interest on datasets with high class imbalance.

Materials and methods

Population

The population used in this analysis consisted of 12,425,832 people who were continuously enrolled in Medicare or Commercial plans from January 1, 2018 to December 31, 2019. Continuous enrollment in this period was required since the identification of the disease cohorts and the creation of features are heavily reliant on historic claims data. While it would have been ideal to ensure the completeness of each person's claims history, imposing a longer continuous enrollment criterion would have made fewer people eligible. Although features were created based on claims data from 2016 to 2018, completeness can only be guaranteed for data in 2018.

We aimed to predict if a person will have the first occurrence of a specific disease at any point from January 1, 2019 to December 31, 2019. Patients who had prior diagnoses of the target disease before the prediction time were excluded. For example, those that had a diabetes diagnosis prior to 2019 would be excluded from the cohort for which we are predicting diabetes.

Target diseases

We aimed to predict the occurrence of six diseases – lung cancer, prostate cancer, rheumatoid arthritis (RA), type 2 diabetes (T2D), inflammatory bowel disease (IBD), and chronic kidney disease (CKD) – in the prediction year. Claims-based definitions were created for each target disease. Table 1 gives definitions for each disease, along with the

Table 1 Definitions for flagging disease outcomes and the respective prevalences in the final cohort table. Abbreviations used: Chronic Kidney Disease (CKD), Type 2 Diabetes (T2D), Inflammatory Bowel Disease (IBD), Rheumatoid Arthritis (RA), International Classification of Diseases, Tenth Revision (ICD-10)

Disease	ICD-10 Code	Definition	Prevalence	Number of cases
Lung Cancer	C34	Two lung cancer claims at least 30 days apart, no history of any cancer	0.053%	6,539
Rheumatoid Arthritis (RA)	M05, M06 (Except M064)	At least one RA claim*	0.10%	12,174
Prostate Cancer	C61	Two prostate cancer claims at least 30 days apart, no history of any cancer	0.12%	14,925
Type 2 Diabetes (T2D)	E11	Two T2D claims at least 30 days apart	0.59%	73,540
Inflammatory Bowel Disease (IBD)	K51, K52	Two IBD claims at least one day apart	0.32%	39, 502
Chronic Kidney Disease (CKD)	N18	Two CKD claims at least 30 days apart	0.63%	78,786

corresponding prevalence and cohort size, presented in order of increasing prevalence. Disease flags are based on the International Classification of Diseases, Tenth Revision (ICD-10). Since the presence of a given ICD-10 code in a claim may simply be due to an event such as a screening test being ordered rather than truly indicative of a diagnosis, we required the presence of that disease code in at least two claims within a specified time period for most of the diseases under consideration. The second occurrence of the ICD-10 code is considered the confirmatory diagnosis for most diseases.

Data creation

Features were derived from the administrative claims history of members from 2016 to 2018. Each claim corresponds to a patient visit and contains information that describes the healthcare services rendered such as diagnosis codes, procedure codes, medical supplies and equipment, and costs incurred. In this study, only the diagnosis codes were used as features. One claim can be associated with up to 12 diagnoses which corresponds to 12 unique ICD-10 codes, sequenced based on the severity of the illness. Only the first three diagnoses in each claim were considered to ensure that only the most clinically relevant diagnoses to the health service being availed were used. Other ICD-10 codes are coded primarily for billing purposes, and typically have very little to no relevance to the procedure or service.

Each diagnosis corresponds to an ICD-10 code, which can be up to 7 digits long. For each of the first three diagnoses, only the first three characters of the ICD-10 codes were used. The first three characters correspond to a broader classification of the diagnosis. For example, E10.2 corresponds to Type 1 Diabetes with kidney complications while E10.65 is for Type 1 diabetes with hyperglycemia. Taking only the first three characters, these two ICD-10 codes would fall under "Type 1 Diabetes". Using only the first three characters of the ICD-10 codes allows us to create adequately sized groups of patients that have the same disease.

For each claim in the patient's entire history from 2016 to 2018, the first three characters of the first three ICD-10 codes were taken. From these first three characters, indicator flags were created based on the presence or absence of these codes in four time periods of varying lengths. Thus, each code corresponds to four flags in our data set. Table 2 shows the time windows considered.

Binning diagnoses flags in different time windows was done to introduce a temporal component to the predictors. Older diagnoses were generally less relevant to the prediction of a disease. The presence of a particular diagnosis code in an earlier window does not guarantee that it will be present in the succeeding periods. Disease flags are only determined by the presence of a patient's claims with the relevant ICD codes related to a condition within time window, independently.

Table 2 Time periods for creating feature flags

Time window	Start date	End date
1	Oct 1 2018	Dec 31 2018
2	July 1 2018	Sep 30 2018
3	Jan 1 2018	Jun 30 2018
4	Jan 1 2016	Dec 31 2017

Demographic information such as gender, state-level socioeconomic index, and age in 2018 were also used as features in the analysis. In total, 3,511 features were created.

Benchmark framework

The flowchart in Fig. 1 shows the framework used to benchmark the different AutoML systems adapted from [22] and modified to include a bootstrapping procedure to obtain 95% confidence intervals for each of the metrics considered. The features used for each model depended on the target outcome; flags corresponding to the ICD-10 code of the disease being predicted were excluded. For example, for lung cancer, all four features across different windows for the ICD code C34 were dropped. For each target disease, we generated a training set of 300,000 samples taken from the population of 12 million, maintaining the disease prevalence. The three AutoML tools (AutoSklearn [25], H2O [26] and TPOT [27]) and a random forest model were trained on the same training set for each disease. Random forests were used as a baseline in this study primarily because it was also used as the baseline model in the framework of Gijsbers et al. [22]. In addition, random forests are good baseline models because they generate reasonable predictions without much parameter tuning, and can handle large numbers of inputs and features. Another difference between our framework and the reference framework is that for each AutoML model, we optimized for different metrics – average precision (area under precision-recall curve (AUCPR) approximation), balanced accuracy, and area under the receiver operating characteristic curve (ROC AUC). H2O was optimized for AUCPR and AUC, the latter corresponding to ROC AUC. We did not optimize H2O for balanced accuracy because this metric was not included in its base built-in scorers. This resulted in multiple models per target disease per tool instead of having a single model optimized for ROC AUC. The random forest model was considered the baseline for comparison. The default settings were used for each tool, except for the maximum run-time which we set at 48 hours for each model. All models were trained on identical 16-CPU 8-core Intel Xeon (2.3 GHz) machines with 256GB RAM. The trained models were then used to predict outcomes for the remaining holdout dataset consisting of 11.7 million samples. For each model and target disease, bootstrapping was performed on the predictions to obtain 95% confidence intervals for each model metric. Samples were taken with replacement (both stratified and not stratified) from the holdout validation set to obtain 500 sets of 150,000 observations

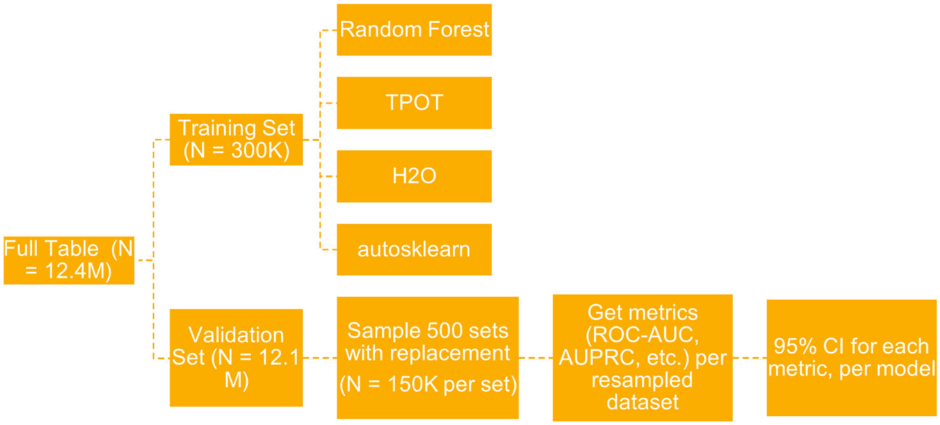
Random Forest
TPOT
H2O
autosklearn
Training Set (N = 300K)
Full Table (N = 12.4M)
Validation Set (N = 12.1 M)
Sample 500 sets with replacement (N = 150K per set)
Get metrics (ROC-AUC, AUPRC, etc.) per resampled dataset
95% CI for each metric, per model

Fig. 1 Flowchart of framework for benchmarking AutoML tools adapted from Gjisbers et al.

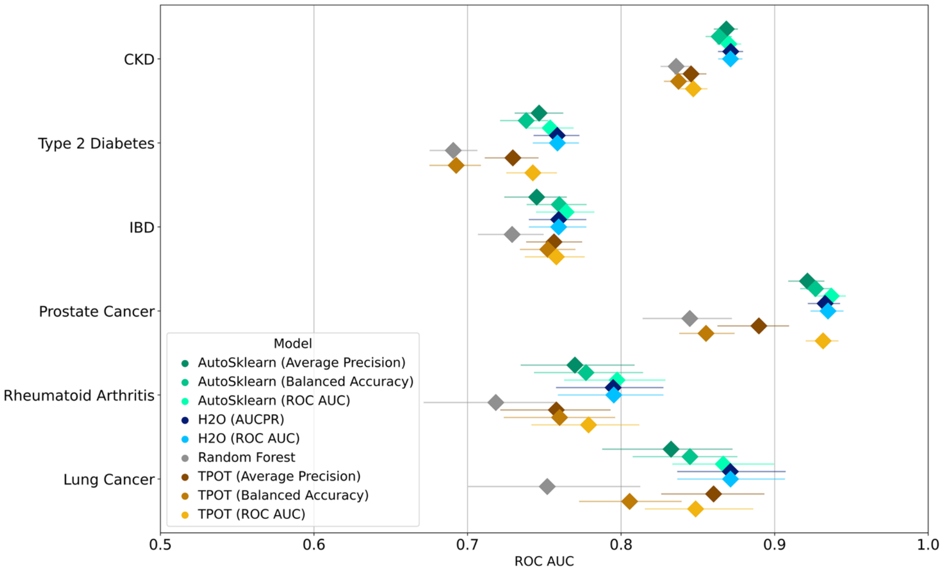

Fig. 2 ROC AUC performance of different AutoML models trained for various disease outcomes from stratified bootstrap samples. Median values are indicated by diamond markers and 95% CIs are indicated by lines

each. Metrics were then computed for the predictions of each model on each resampled dataset, yielding 500 values per metric per model which were used to derive the 95% confidence intervals. We note that, due to the large dataset size and consequent time and resource requirements, we ran each AutoML tool once for each choice of optimization metric, so these are confidence intervals for the performance on the holdout data for each of these specific AutoML runs.

Results

The bootstrapped metrics for the performance of the different models on the holdout set are shown in Figs. 2 and 3 for ROC AUC and AUCPR (the latter as approximated

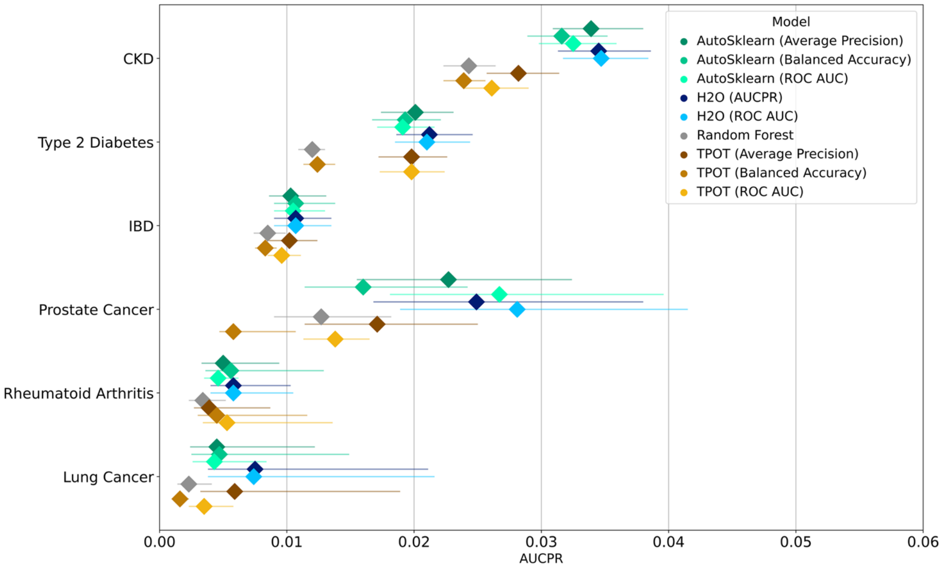

Fig. 3 AUCPR performance of different AutoML models trained for various disease outcomes from stratified bootstrap samples. Median values are indicated by diamond markers and 95% CIs are indicated by lines

by the average precision), respectively. The same results can be seen in tabular form in Supplementary Tables 1 and 2, Additional File 1.

In both figures, diamond markers indicate the median metric scores for each model, while circle markers denote the lower and upper limits of the 95% confidence intervals calculated through bootstrapping. These figures show metrics computed using stratified bootstrap samples. There is minimal difference between the results of getting the metrics from either stratified or non-stratified bootstrap samples. The results for non-stratified samples can be seen in Figs. 1 and 2 in Supplementary File 1. For ROC AUC, we observe varying performances across different diseases. In general, no single AutoML framework outperforms the rest consistently and with a wide margin. Also, we observe that disease prevalence is not directly correlated to model performance; models with highest ROC AUC scores were those for prostate cancer which is the second least prevalent disease (0.12% prevalence). We also observe narrow confidence intervals for the models trained for predicting CKD, which has the highest prevalence. Wider confidence intervals correspond to lower disease prevalence, with the widest intervals observed for lung cancer (0.053% prevalence). Note that this is not always the case; for prostate cancer, all AutoSklearn and H2O models, and the TPOT model optimized for ROC AUC trained have relatively narrow confidence intervals.

Since model scores and performance varied across diseases, we normalize the median ROC AUC scores based on the median random forest model performance as done by Gjisbers et al. The results are shown in Table 3. The best performing models across diseases are either H2O models or the AutoSklearn model optimized for ROC AUC. However, for each disease the difference between the best model and the other models are small. In terms of ROC AUC improvements relative to the random forest models, greater improvements are observed for the less prevalent diseases. The median improvements for all AutoML models per disease are 1.136, 1.100, 1.083, 1.041, 1.078, and 1.036 for lung cancer, prostate cancer, rheumatoid arthritis, IBD, Type 2 Diabetes, and CKD, respectively.

Table 3 Median performance ROC AUC scores for different AutoML models scaled according to median random forest performance. Models with the best performance for each disease are indicated in bold

Metric: ROC AUC	Lung Cancer	Prostate Cancer	Rheumatoid Arthritis	Type 2 Diabetes	IBD	CKD
Model						
AutoSklearn (Average Precision)	1.107	1.091	1.072	1.081	1.022	1.039
AutoSklearn (Balanced Accuracy)	1.124	1.097	1.082	1.069	1.042	1.034
AutoSklearn (ROC AUC)	1.152	**1.109**	**1.110**	1.091	**1.048**	1.041
H2O (AUC)	**1.159**	1.107	1.107	**1.098**	1.042	1.042
H2O (AUCPR)	**1.159**	1.104	1.106	**1.098**	1.042	**1.043**
Random Forest	1.000	1.000	1.000	1.000	1.000	1.000
TPOT (Average Precision)	1.144	1.053	1.055	1.056	1.037	1.012
TPOT (Balanced Accuracy)	1.071	1.013	1.058	1.003	1.032	1.002
TPOT (ROC AUC)	1.128	1.103	1.084	1.075	1.040	1.013
Random Forest	1.000	1.000	1.000	1.000	1.000	1.000

Due to the imbalance of the datasets, we also measure model performance on AUCPR. Low AUCPR scores are observed for all models as seen in Fig. 3. The models for prostate cancer which had narrow confidence intervals in terms of their ROC AUC scores have wider confidence intervals for their bootstrapped AUCPR scores. Generally, H2O models had the highest median AUCPR scores. Taking note of the range of AUCPR values, however, there is no single model that outperforms the rest significantly across different diseases.

Table 4 shows the performance increases of the models relative to the median baseline scores of the random forest model. Despite the low AUCPR scores, we generally observe improvements in AUCPR compared to the baseline models except for TPOT models optimized for balanced accuracy, especially those trained for predicting prostate and lung cancer. The median AUCPR improvements for all AutoML models per disease are 2.000, 1.567, 1.515, 1.224, 1.650, and 1.319 for lung cancer, prostate cancer, rheumatoid arthritis, IBD, Type 2 Diabetes, and CKD, respectively.

Beyond ROC AUC values, selecting the thresholds for each model is an essential step in evaluating a model for practical purposes. This is especially true when working with imbalanced data [28]. Despite the AutoML output models being ready to generate hard predictions, in practice, one must still consider the threshold that will give the best balance between true positive rate and false positive rate depending on the problem being solved. The actual ROC curves generated using the full validation set are shown in Fig. 4.

To illustrate, consider the case of predicting lung cancer, which has the lowest prevalence among the six diseases explored in this study. Lung cancer is often detected at the advanced stage when prognosis is poor and survival rates are low, thus making it one of the leading causes of cancer-related deaths in the United States. Several strategies that aim to detect the disease at an early stage where intervention is most effective are in place, chief of which are the rule-based screening guidelines provided by the National Comprehensive Cancer Network (NCCN) and the United States Preventive Services Task Force (USPSTF). However, even with these methods in place, only about 2% of annual

Table 4 Median AUCPR scores for different AutoML models scaled according to median random forest performance. Models with the best performance for each disease are indicated in bold

Metric: Average Precision	**Lung Cancer**	**Prostate Cancer**	**Rheumatoid Arthritis**	**Type 2 Diabetes**	**IBD**	**CKD**
Model						
AutoSklearn (Average Precision)	1.957	1.787	1.471	1.675	1.212	1.395
AutoSklearn (Balanced Accuracy)	2.043	1.260	1.647	1.608	1.259	1.300
AutoSklearn (ROC AUC)	1.870	2.102	1.353	1.592	1.235	1.337
H2O (AUC)	3.217	**2.213**	**1.706**	1.750	**1.259**	**1.428**
H2O (AUCPR)	**3.261**	1.961	**1.706**	**1.767**	**1.259**	1.420
TPOT (Average Precision)	2.565	1.346	1.147	1.650	1.200	1.160
TPOT (Balanced Accuracy)	0.696	0.457	1.324	1.033	0.976	0.984
TPOT (ROC AUC)	1.522	1.087	1.559	1.650	1.129	1.074
Random Forest	1.000	1.000	1.000	1.000	1.000	1.000

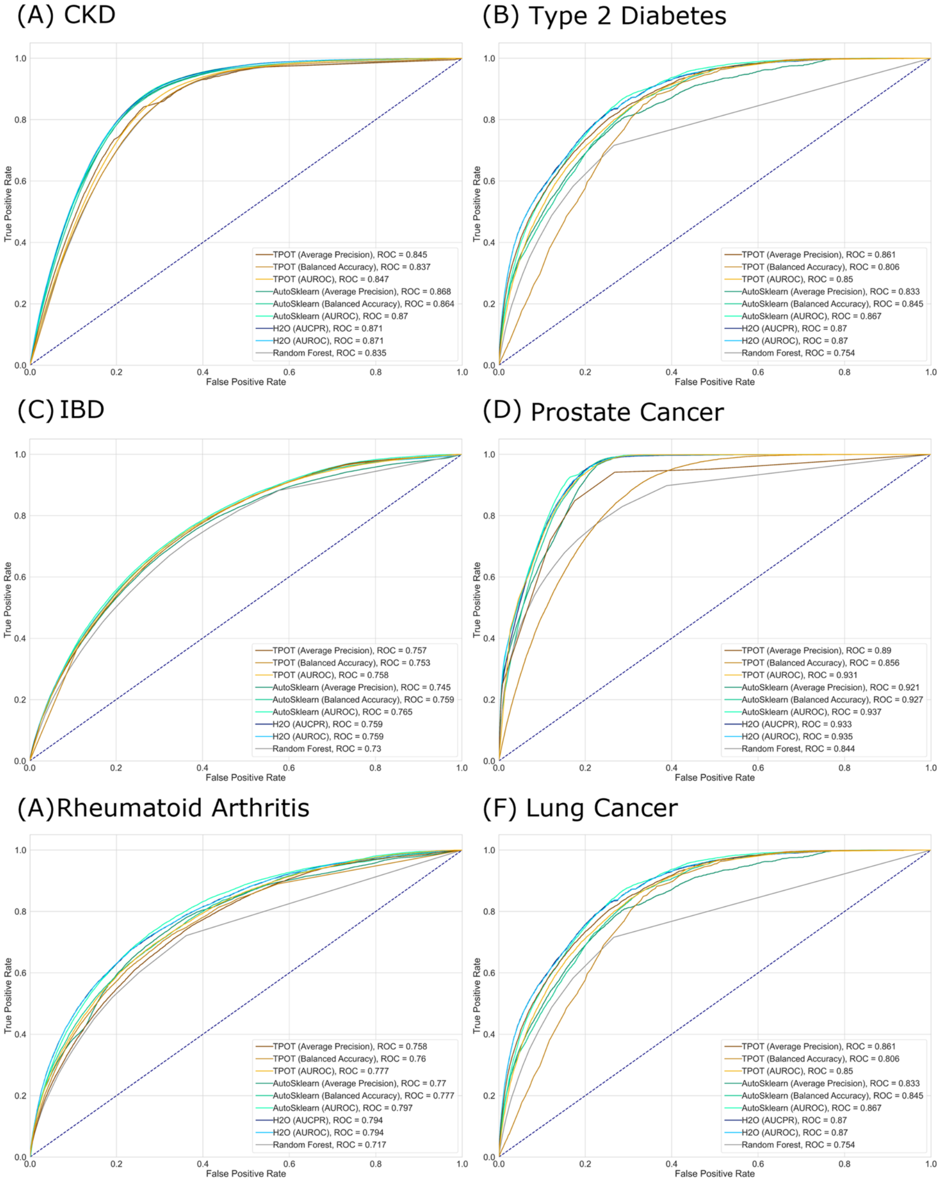

Fig. 4 Receiver operating characteristic (ROC) curves of models trained for predicting different diseases. ROC curves are generated using prediction scores on full validation set (N = 12,125,832)

lung cancer incidences are detected through screening. Patients who are considered eligible for screening based on the NCCN and USPSTF guidelines undergo a low-dose computed tomography (LDCT) annually. Though LDCT can detect lung cancer at a treatable stage, it also poses several health risks especially to those who are otherwise clear of the disease. These include unnecessary treatment, complications and a theoretical risk of developing cancer from exposure to low-dose radiation. Thus, in building a predictive model for lung cancer, these associated costs must be considered together with the objective of identifying as many positive cases as possible. In other words, for this kind of problem, there is a need to minimize the number of false positives while trying to achieve a high true positive rate (TPR). After training an AutoML model using any tool, caution should be exercised when still deploying the models. Models typically provide predictive probabilities and selecting the correct threshold for the application is necessary. Identifying the correct thresholds depending on the trade-offs between TPR and FPR can be done by looking at the respective ROC AUC curves as seen in Fig. 4

We show different confusion matrices for the best performing model for predicting lung cancer in terms of ROC AUC in Supplementary Table 3, Additional File 1. Thresholds are chosen based on deciles of actual predicted probability values for the full validation dataset. Identifying the optimal threshold will depend on the costs of true positives, false positives and false negatives. We consider hypothetical dollar costs for the same model noting that costs in terms of medical risks and quality of life are not included. We assume the per person cost of getting the disease is $300,000 annually if not detected early (equivalent to the cost of a false negative), while if detected early, the cost will be $84,000 annually (equivalent to the cost of a true positive). For this situation, we also consider two hypothetical tests, one priced at $100 per test and LDCT which costs about $500 on average. We compute savings based on the baseline situation where no tests are administered (each person with lung cancer is associated with the cost of a false negative). Figure 5 plots the savings for each hypothetical test cost per person for different decile probability thresholds. The optimal thresholds for the model depend on the situation where the model will be used. For the $100 test, we see the optimal cut-off is at the 70th percentile while for the $500 test, it is at the 90th percentile. For the $500 test, this cut-off is the only one that leads to positive savings. These cut-offs correspond to a FPR = 0.3, TPR = 0.9, and FPR = 0.1 and TPR = 0.52, respectively.

Discussion

Since AutoML software packages are attractive out-of-the-box tools to build predictive models in the context of healthcare data, we examined and compared the performance

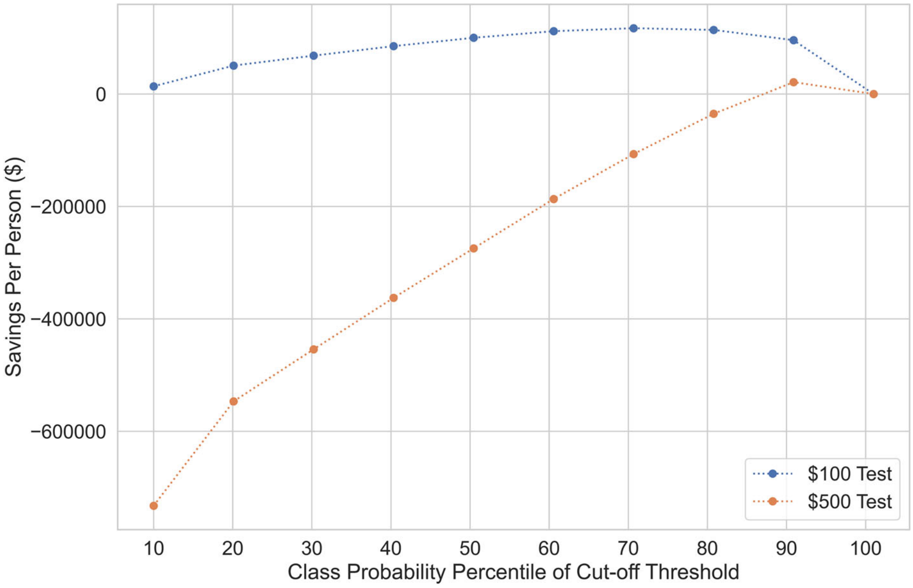

Fig. 5 Average savings per person for different cut-off thresholds for the H2O (AUROC) model for different test costs. True positive costs are set at $84,000, while false negative costs are set at $300,000. False positive costs are only from the test costs

of three of these tools (AutoSklearn, H2O, and TPOT) on a large medical claims dataset for six different disease outcomes. However, these datasets present several challenges. First, the sample size ($\sim$ 12.5M) is much larger than the typical size of datasets analyzed with AutoML. In this work we used a stratified sample of 300k for training, which is still quite large for AutoML given that these approaches are computationally intensive because they are iterating over many different algorithms. For example, the number of generations completed by TPOT within the 48-hour time limit varied greatly for each target disease and scoring metric. The number of generations completed ranged from 7 to 38, with an average of 18.88 across 18 models. Running time for the different AutoML models varied depending on the initial conditions and target conditions. However, for most methods, the running time hit 48 hours. Improvements in terms of scalability of these AutoML methods are certainly desirable in the context of medical claims data. Once training several AutoML models each on a different and relatively large subsample of the dataset becomes computationally feasible, combining the resulting models into an ensemble may provide further performance improvements.

A second challenge is the extremely low case prevalence characteristic of healthcare data; in our examples, this varied from 0.053% to 0.63%. This may be the main culprit for the low AUCPR scores we observed across the methods and diseases. Improvements in terms of handling highly imbalanced datasets are crucial for healthcare applications. One direction for future work is to explore combinations of over- and under-sampling techniques with ensemble approaches in the spirit of [28].

Another challenge which may partly account for the poor performances observed among the models stems from the limitations inherent to the features available in healthcare databases. Since claims are coded for billing purposes, some healthcare services are tied to a certain ICD-10 code which may not necessarily be indicative of the presence of a certain disease. For example, individuals who are eligible for cancer screening will have the screening procedure billed under a cancer ICD-10 code regardless of the result. Hence, individuals who do not have cancer will still have cancer codes in their claims history. This means that simply flagging the presence of these ICD-10 codes is not an accurate representation of the person's medical history. Using fewer selected features may help improve model performance. For example, retaining only features corresponding to ICD-10 codes clinically related to the disease being predicted can reduce the size of the feature set and allow the models to more easily establish relationships between the features and the target.

Conclusion

AutoML tools generally fast track the ML pipeline and the models they generate can serve as starting points for building predictors. However, the performance of these tools on the medical claims datasets used in this study suggest that there may be room for improvement in how AutoML tools handle data of this scale and with such high imbalance. To address the limitations of the data, further feature selection, resampling and imbalance-learning ensembles are possible next steps.

Despite the advantages of using AutoML tools for model selection and optimization, care must still be taken in identifying the optimal output thresholds depending on the research question.

Abbreviations

AutoML	Automated machine learning
ML	Machine learning
CKD	Chronic kidney disease
T2D	Type 2 diabetes
IBD	Inflammatory bowel disease
RA	Rheumatoid arthritis
ICD-10	International classification of diseases, tenth revision
ROC AUC	Area under the receiver operating characteristic curve
AUCPR	Area under the precision-recall curve

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13040-022-00300-2.

Additional file 1. ROC AUC performance of different AutoML models trained for various disease outcomes from non-stratified bootstrap samples. Median values are indicated by diamond markers and 95% CI limits are indicated by circles.

Acknowledgements

Not applicable.

Authors' contributions

RAR was responsible for methodology, software, formal analysis, writing - original draft, and creation of visualizations.
MND was responsible for the methodology, software, formal analysis, data curation, and writing - original draft.
SM contributed to the conceptualization of the project, methodology, data curation, writing - review and editing, and supervision.
JTJ contributed to the project design, methodology, writing - review and editing, and supervision.
NS was responsible for conceptualization, writing - review and editing, and supervision.
EM contributed suggestions on analysis plans and TPOT settings, result checks, and manuscript organization and revisions.
JM helped conceive of and design the project, plan the analyses, and interpret the results. He also assisted in writing and editing the manuscript.

Funding

JM and EM were supported by the National Institutes of Health under award number LM010098.

Availability of data and materials

The datasets generated and analysed in the current study are not publicly available since they contain private health information. However, if required, access to data for the editors can be made available upon request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

Roland Albert Romero, Mariefel Nicole Deypalan, Suchit Mehrotra, John Titus Jungao, and Natalie Sheils are employees of OptumLabs part of UnitedHealth Group. Natalie Sheils owns stock in the company. The other authors have no conflicts of interest to disclose.

Received: 7 December 2021 Accepted: 27 June 2022
Published online: 26 July 2022

References

1. Mustafa A, Rahimi Azghadi M. Automated machine learning for healthcare and clinical notes analysis. Computers. 2021;10(2):. https://doi.org/10.3390/computers10020024.
2. Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease prediction by machine learning over big data from healthcare communities: IEEE Access; 2017, pp. 1–1. https://doi.org/10.1109/ACCESS.2017.2694446.
3. Luo G, Stone BL, Johnson MD, Tarczy-Hornoch P, Wilcox AB, Mooney SD, Sheng X, Haug PJ, Nkoy FL. Automating construction of machine learning models with clinical big data: Proposal rationale and methods. JMIR Res Protoc. 2017;6(8):175. https://doi.org/10.2196/resprot.7757.
4. Osawa I, Goto T, Yamamoto Y, Tsugawa Y. Machine-learning-based prediction models for high-need high-cost patients using nationwide clinical and claims data. NPJ Dig Med. 2020;3(1):148. https://doi.org/10.1038/s41746-020-00354-8.

5. Srinivasan U, Arunasalam B. Leveraging big data analytics to reduce healthcare costs. IT Prof. 2013;15:21–28. https://doi.org/10.1109/MITP.2013.55.
6. Christensen T, Frandsen A, Glazier S, Humpherys J, Kartchner D. Machine learning methods for disease prediction with claims data. In: 2018 IEEE International Conference on Healthcare Informatics (ICHI). New York: IEEE Press; 2018. p. 467–4674.
7. Waring J, Lindvall C, Umeton R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artif Intell Med. 2020;104:101822. https://doi.org/10.1016/j.artmed.2020.101822.
8. Popescu M, Khalilia M. Improving disease prediction using ICD-9 ontological features. In: 2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011). IEEE; 2011. https://doi.org/10.1109/fuzzy.2011.6007410. https://doi.org/10.1109%2Ffuzzy.2011.6007410.
9. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. BMJ Open Respir Res. 2017;4(1):. https://doi.org/10.1136/bmjresp-2017-000234. https://bmjopenrespres.bmj.com/content/4/1/e000234.full.pdf.
10. Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, Hall MK. Prediction of in-hospital mortality in emergency department patients with sepsis: A local big data–driven, machine learning approach. Acad Emerg Med. 2016;23(3):269–78. https://doi.org/10.1111/acem.12876. https://onlinelibrary.wiley.com/doi/pdf/10.1111/acem.12876.
11. Shameer K, Johnson KW, Yahi A, Miotto R, Li L, Ricks D, Jebakaran J, Kovatch P, Sengupta PP, Gelijns S, et al. Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: a case-study using mount sinai heart failure cohort. In: Pacific Symposium on Biocomputing 2017. Hackensack: World Scientific; 2017. p. 276–87.
12. Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease prediction by machine learning over big data from healthcare communities. IEEE Access. 2017;5:8869–79. https://doi.org/10.1109/ACCESS.2017.2694446.
13. Wolpert DH, Macready WG. No free lunch theorems for optimization. IEEE Trans Evol Comput. 1997;1(1):67–82.
14. Zöller M-A, Huber MF. Benchmark and survey of automated machine learning frameworks. J Artif Intell Res. 2021;70: 409–72.
15. Hutter F, Kotthoff L, Vanschoren J. Automated Machine Learning: Methods, Systems, Challenges. New York: Springer; 2019.
16. Yao Q, Wang M, Chen Y, Dai W, Li Y-F, Tu W-W, Yang Q, Yu Y. Taking human out of learning applications: A survey on automated machine learning. arXiv preprint arXiv:1810.13306. 2018.
17. Mattson P, Reddi VJ, Cheng C, Coleman C, Diamos G, Kanter D, Micikevicius P, Patterson D, Schmuelling G, Tang H, et al. Mlperf: An industry standard benchmark suite for machine learning performance. IEEE Micro. 2020;40(2):8–16.
18. Olson RS, La Cava W, Orzechowski P, Urbanowicz RJ, Moore JH. Pmlb: a large benchmark suite for machine learning evaluation and comparison. BioData Min. 2017;10(1):1–13.
19. Milutinovic M, Schoenfeld B, Martinez-Garcia D, Ray S, Shah S, Yan D. On evaluation of automl systems. In: Proceedings of the ICML Workshop on Automatic Machine Learning, vol. 2020. Vienna; 2020.
20. Hanussek M, Blohm M, Kintz M. Can AutoML outperform humans? An evaluation on popular OpenML datasets using AutoML Benchmark. 2020. 2009.01564. Accessed 15 Dec 2020.
21. Balaji A, Allen A. Benchmarking Automatic Machine Learning Frameworks. 2018. 1808.06492. Accessed 15 Dec 2020.
22. Gijsbers P, LeDell E, Thomas J, Poirier S, Bischl B, Vanschoren J. An open source automl benchmark. arXiv preprint arXiv:1907.00909. 2019.
23. Luo G. A review of automatic selection methods for machine learning algorithms and hyper-parameter values. Netw Model Anal Health Inform Bioinforma. 2016;5(1):1–16.
24. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Making. 2011;11(1):1–13.
25. Feurer M, Klein A, Eggensperger K, Springenberg JT, Blum M, Hutter F. Auto-sklearn: efficient and robust automated machine learning. In: Automated Machine Learning. Vienna: Springer; 2019. p. 113–34.
26. LeDell E, Poirier S. H2o automl: Scalable automatic machine learning. In: Proceedings of the AutoML Workshop at ICML, vol. 2020. Vienna; 2020.
27. Olson RS, Moore JH. In: Hutter F, Kotthoff L, Vanschoren J, editors. TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning. Cham: Springer; 2019, pp. 151–60. https://doi.org/10.1007/978-3-030-05318-5_8. https://doi.org/10.1007/978-3-030-05318-5_8.
28. Schubach M, Re M, Robinson PN, Valentini G. Imbalance-aware machine learning for predicting rare and common disease-associated non-coding variants. Sci Rep. 2017;7(1):1–12.

Publisher's Note