Lab on a Chip

A Web-based Automated Machine Learning Platform to Analyze Liquid Biopsy Data

Journal:	*Lab on a Chip*
Manuscript ID	LC-ART-01-2020-000096.R1
Article Type:	Paper
Date Submitted by the Author:	12-Apr-2020
Complete List of Authors:	Shen, Hanfei; University of Pennsylvania, Bioengineering Liu, Tony; University of Pennsylvania, Computer Science Cui, Jesse; University of Pennsylvania, Computer Science Borole, Piyush; University of Pennsylvania, Bioengineering Benjamin, Ari; University of Pennsylvania, Computer Science Kording, Konrad; University of Pennsylvania, Bioengineering Issadore, David; University of Pennsylvania, Bioengineering

Lab on a Chip

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

A Web-based Automated Machine Learning Platform to Analyze Liquid Biopsy Data[†]

Hanfei Shen,[a] Tony Liu,[b] Jesse Cui,[b] Piyush Borole,[c] Ari Benjamin,[a] Konrad Kording,[ad] and David Issadore*[ae]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Liquid biopsy (LB) technologies continue to improve in sensitivity, specificity, and multiplexing and can measure an ever growing library of disease biomarkers. However, clinical interpretation of the increasingly large sets of data these technologies generate remains a challenge. Machine learning is a popular approach to discover and detect signatures of disease. However, limited machine learning expertise in the LB field has kept the discipline from fully leveraging these tools and risks improper analyses and irreproducible results. In this paper, we develop a web-based automated machine learning tool tailored specifically for LB, where machine learning models can be built without the user's input. We also incorporate a differential privacy algorithm, designed to limit the effects of overfitting that can arise from users iteratively developing a panel with feedback from our platform. We validate our approach by performing a meta-analysis on 11 published LB datasets, and found that we had similar or better performance compared to those reported in the literature. Moreover, we show that our platform's performance improved when incorporating information from prior LB datasets, suggesting that this approach can continue to improve with increased access to LB data. Finally, we show that by using our platform the results achieved in the literature can be matched using 40% of the number of subjects in the training set, potentially reducing study cost and time. This self-improving and overfitting-resistant automatic machine learning platform provides a new standard that can be used to validate machine learning works in the LB field.

Introduction

Diseases are often localized in parts of the body that are difficult to access, such as a tumor developing in the brain or an infection spreading in the spine, which makes measurements of molecular biomarkers for diagnostics and clinical monitoring challenging. In the last decade, there has been an enormous interest, and much success, in measuring the sparse molecular biomarkers - rare circulating cells, microvesicles, nucleic acids, proteins, and metabolites- that are shed from diseased cells into the circulation[1–5]. Such biomarkers can be found in accessible body fluids such as blood or urine. Increasingly, these liquid biopsies (LB) rely not on any single biomarker, but instead are measuring multiplexed panels of biomarkers that can more accurately predict and comprehensively capture a disease state than is possible using a single marker[6–9]. By identifying signatures of disease in multiplexed panels, rather than measuring only a single marker, these approaches can: 1. mitigate the effects of variability in biomarker expression across individuals, 2. diagnose diseases that are phenotypically heterogeneous, and 3. diminish sensitivity to variability of baselines levels of biomarkers in healthy individuals[10–14]. Biomarkers for LB have been measured using a variety of platforms, including next generation sequencing (NGS), mass spectrometry, microarrays, as well as emerging microfluidic approaches. The rapid development of each of these technologies, used separately and in combination with one another, has helped drive the trend to measure increasingly large numbers of biomarkers from clinical samples[15–18], which demands more sophisticated computational analysis to be interpreted.

Machine learning, a set of computational approaches that can reduce large numbers of measurements into lower-dimensional

[a] Department of Bioengineering, University of Pennsylvania, Philadelphia, PA 19104, USA. E-mail: daveissadore@gmail.com; Tel: +215 962 5206
[b] Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, USA
[c] Department of Chemical and Biomolecular Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA
[d] Department of Neuroscience, University of Pennsylvania, Philadelphia, PA 19104, USA
[e] Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA

[†] Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 10.1039/cXCP00000x/

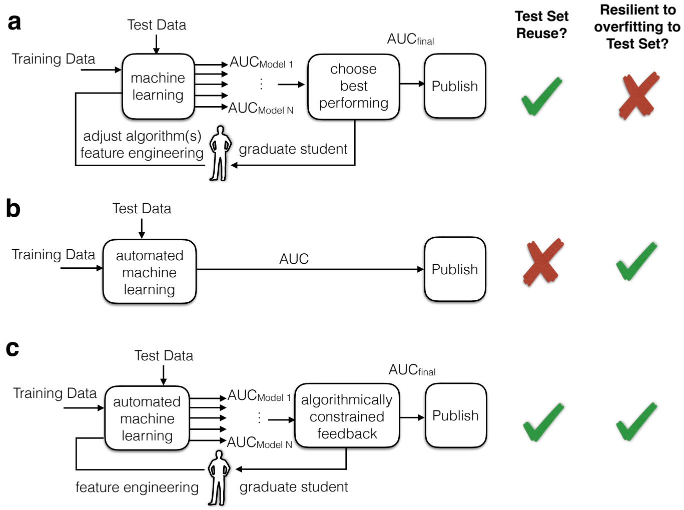


Fig. 1 Automated Machine Learning for LB. a. Adaptive data analysis is a common mistake in performing machine learning, where the same test set is reused to evaluate multiple models amongst which the best performing is selected, causing overfitting to the test set. b. An illustration of how automated machine learning can prevent overfitting via removing human factors in the machine learning process and prevent the reuse of the test set. c. A schematic of our platform (AutoML) that provides automated machine learning and controls overfitting to the test set caused by user's reusing a test set, via a computational algorithm which limits the information from the test set that is fed back to the user.

outputs, has proven particularly successful decoding patterns in biomarker panels to provide clinically actionable information[19,20]. A growing set of studies have used machine learning to identify biomarker signatures in a wide range of diseases, such as cancer[21–27], brain injury[28], neurological diseases[29], and infectious disease[30]. There are however major challenges that have kept machine learning from being fully leveraged in the field of LB. Primarily, there are not clear guidelines for a non-machine learning expert to navigate the many choices that must be made to successfully apply machine learning to a LB dataset. These include the choice of which features to measure, how to pre-process the data (e.g. whether or not to take the logarithm of a protein concentration), which machine learning algorithms to use (e.g. Random Forest, Support Vector Machine, etc. . .), and which hyperparameters (algorithm-regulating parameter that is defined before training) associated with these algorithms to select. Moreover, there are many pitfalls in the application of machine learning that can confound the results[31], and it is possible to generate misleading results[32,33]. One such common problem is data leakage. In data leakage, information outside of the training set inadvertently 'leaks' into the model. A common data leakage mistake that researchers make is to try many different machine learning models (data pre-processing, types of machine learning, choice of hyperparameters) and evaluate these choices on the test set, and then report the best result (**Fig.1a**). It has been proven that such an adaptive optimization of machine learning can lead to overfitting to the test set in both informative and non-informative datasets[34]. Liquid biopsies typically measure a number of molecular markers that is 1-2 orders greater than the number of subjects measured, primarily due to the the cost and time required to acquire clinical samples. This high feature-to-sample ratio makes the analysis of liquid biopsy data particularly prone to overfitting.

A promising solution to the challenges of applying machine learning to LB is automated machine learning. In such an approach, the preprocessing of the data, selection of the type of machine learning used, and the determination of the hyperparameters for those algorithms is performed automatically without input from the user (**Fig. 1b**). In recent years, these automated software packages have been shown to both reduce the barrier of entry for non-machine learning experts and to reduce the risk of improper use of machine learning in a variety of fields[35]. To ensure that a machine learning model generalizes to prospectively collected data and is not fit to attributes of the test set, it is ideal when using these approaches that the test set only be used once[36–39](**Fig.1b**). However, because of the expense of collecting samples in the field of LB, there is a strong motivation to reuse this data during the iterative development of a biomarker panel.

In this paper, we present a web-based, easy to use automated machine learning tool that features the capability to continuously improve given access to more LB data and that prevents overfitting to the test set (**Fig.1c**). It is specifically designed for LB data. Using this automated machine learning platform, we performed a meta-analysis on 11 published LB datasets and found that the automated approach had similar or better performance (AUC) compared to those reported in the literature. Moreover, we showed that automated machining learning's performance improved when the system had its model-fitting initialized based upon a library of previously analyzed LB datasets, suggesting that this approach can be further improved as those in the field make use of it. Additionally, we show that by using our platform the results achieved in the literature can be matched using only 40% of the number of subjects in the training set, potentially reducing the cost and time of an LB study. All of the 11 datasets that were included in this study are small to medium-sized binary classification (e.g. cancer or non-cancer) datasets. These datasets vary significantly in sizes, with a range of feature sizes from 7 to 28,541 and number of samples ranging from 34 to 303. Moreover, we implement our automated machine learning platform as a simple to use online platform. We have incorporated a feature into our AutoML to limit the overfitting that can arise from users iteratively developing a panel with feedback from our automated machine learning, using differential privacy - a notion of privacy preservation in data analysis which ensures that the probability of observing any outcome from an analysis is essentially unchanged by modifying any single dataset element[36]. This feature works by limiting the information provided to the user when evaluating the test set to control the effect of overfitting.

Methods

Dataset collection

A total of 11 datasets were selected from the LB literature[21–30]. Two of these studies were published work from our lab at The University of Pennsylvania[21,28]. These studies classified disease for various types of cancers, traumatic brain injury, depression, and inflammatory disease. The features measured included protein, metabolites, DNA and RNA biomarkers. The subjects measured included both human and mice, and all samples were blood samples, including serum and plasma. There were two articles from which we extracted multiple datasets[25,28]. However, in each of these papers the sets of data were independent cohorts of samples. Area Under the Curve (AUC) of the Receiver Operating Characteristic curve (ROC) was selected as the parameter to compare the performance of the automated machine learning with the published results. We use AUC as a metric to compare our results to the literature, rather than accuracy, because depending on the clinical context it is often more relevant to optimize for higher sensitivity or specificity rather than maximize accuracy. In the 11 studies involved in this meta-analysis, the AUC values were only reported in 10 out of 11 studies. As such, the remaining dataset is only used in the parts of our study that do not require direct comparison to the results in the literature.

Automated machine learning workflow

In this study, we incorporate an already published algorithm Autosklearn[35] into our AutoML platform. We hypothesized that Autosklearn is particularly well suited to LB, as it minimizes the effects of overfitting with a variety of methods, including internal cross validation within the training set, using only simple machine learning algorithms from scikit-learn[40] that perform well when fitting the small to medium-sized datasets typical of early LB studies[35]. Additionally, it incorporates automatic optimization of hyperparameters to regulate classifiers, and the automatic use of Ensemble Selection - a sophisticated algorithm which construct ensembles of several machine learning models[41]. By adopting simple algorithms and utilizing Ensemble Selection to construct ensemble from models, Autosklearn mitigates the problem that liquid biopsy datasets tend to have more features than subjects which makes LB datasets prone to overfitting.

Autosklearn treats automated machine learning as a combined algorithm selection and hyperparameter optimization problem[35]. A combination of algorithms and their hyperparameters are selected that minimize an averaged loss function evaluated using k-fold cross validation. Autosklearn uses a machine learning framework that consists of 4 data preprocessors, 14 feature preprocessors and 15 classifiers, with a total of 110 hyperparameters that require Bayesian optimization for each dataset[42]. A tree-based Bayesian optimization method[43–45] is applied to optimize these hyperparameters because of its capability for high dimension hyperparameter optimization with high time efficiency[46]. We also use built in procedures of Autosklearn including meta-learning[47], which compares the subject dataset to pre-trained datasets which suggest instantiations of the ML framework, and an automatic ensemble constructor, which construct an ensemble after a library of models has been built by the ML framework.[35]

The workflow of our automated machine learning is as follows. First, a web interface collects training data from the user. Next, a vector of 38 different metafeatures[35] are extracted from the subject dataset. These metafeatures are parameters that characterize the dataset[48,49], including data size, feature number and data skewness[35]. L1 distances are calculated between this metafeature vector with those extracted from 140 datasets about a variety of topics in the data repository, each of these 140 datasets (from OpenML[50]) has been pre-trained by Autosklearn, its metafeatures and a best performing ML instantiation (algorithm-hyperparemeter combination) are remembered by Autosklearn. The top 25 datasets in the repository, which are most similar to the new dataset, are selected based on the L1 distance, and their ML instantiations are suggested to the Bayesian optimization process. After a library of models had been completed within the computational budget given, Autosklearn automatically construct an ensemble.

We control the user's ability to overfit to the test set using a differential privacy (DP) algorithm[36]. In our implementation of DP, rather than directly reporting the accuracy of a model's evaluation of a test set we instead compared that accuracy to the accuracy attained on the training set, and report whether this difference was smaller than a pre-determined threshold value τ modified by adding it with a Laplacian noise $Lap(2*\sigma)$, where σ is a predetermined noise level parameter. If that difference was smaller than the sum of the modified threshold with a second Laplacian noise $Lap(4*\sigma)$ the accuracy of the training set was reported as the test accuracy. If that difference was larger than the threshold, the test set was reported plus a third Laplacian noise term $Lap(\sigma)$(SI **Fig.1b**) This procedure is intended to avoid the model overfitting to the test set, and only add features if they significantly improve the model. In this experiment we utilized a threshold value $\tau = 0.1$ and three different scales of Laplacian noise: σ= 0.005, 0.01 and 0.02.

Evaluation of AutoML using published Liquid Biopsy datasets

To evaluate AutoML and compare it to custom machine learning algorithms performed in the literature, each dataset was partitioned into a training set and a test set and then subjected to automated training and evaluation. If the partitioning of the training set and test set that was used in the source literature was reported, we used the same partitioning of the training and test sets. Otherwise, the dataset was separated into training and test sets using stratified-partitioning which maintains the ratio of samples with each label in both subsets, and the sizes of the training and test set were both kept consistent with what was reported in the literature. For each dataset, automated machine learning was constrained to 15 minutes. Once the model has been constructed, it was used to generate prediction results for the test set, and the resulting prediction outcome is quantified with AUC. Each instance of training the model is stochastic, and as such each dataset was trained and evaluated with Autosklearn five times with the same training-test sets partitioning to evaluate variance in the performance. Each machine learning model generated was deleted after each training and evaluation process to prevent data leakage across repetitions of training.

Evaluating the impact on performance of AutoML's access to prior LB datasets

We evaluated whether automated machine learning's performance can be improved by including LB-specific datasets in the data repository we use to initialize model generation. To address this question, we used a leave-one-out analysis where we iteratively loaded ten out of eleven datasets into the AutoML's data repository and then evaluated the capability to classify the remaining dataset. Autosklearn is capable of performing meta-learning, a function which stores datasets it has previously evaluated in a repository and are used to initiate the analysis of subsequent datasets (i.e. algorithm and hyperparameter selection). Specifically, for each of the eleven datasets, we compared the performance when the data repository had been loaded with the ten remaining data LB sets, when it was naive to LB datasets and only had access to the default repository of Auto-sklearn, and with ten control datasets that were chosen to be of a similar size and structure to our LB datasets but were unrelated to LB. To integrate a dataset into the data repository, our AutoML trained on it for 6 hours. To evaluate AutoML on the left-out datasets, 15 minutes were given to train the model and 5 repetitions of training and evaluation were performed. The same training-test partitions were used in each evaluation. To avoid data leakage between each analysis the meta-learning update was erased to restore Autosklearn to its naive version. The above procedures were repeated for each of the 11 datasets. For our control group of non-LB datasets, we used 10 random non-liquid-biopsy datasets gathered from UCI Machine learning Repository[51] that were matched according to sample and feature sizes with the 11 LB dataset used in this study. The purpose of the control experiment was to evaluate whether the enhancement in performance that came from adding LB datasets to the repository was specific to them being from the LB field.

Evaluating the impact on performance of reduction of training set size

To evaluate the effect of reducing sample size of the training data on the prediction of AutoML on LB datasets, smaller subsets of the original training set of each LB dataset were sub-sampled. In each of these experiments, AutoML was used with access to the LB dataset to initiate model generation. Specifically, each training set was subjected to stratified partitioning to generate subsets that contain between 90% and 20% of its original subject size using random selection. Each subset was subject to AutoML training for 15 minutes and evaluated with the test set. This procedure was repeated 3 times for each subject size, randomly selecting the subjects to be excluded, and the average AUC score was recorded and normalized by its corresponding literature-reported AUC score. One dataset[21] was excluded from this study because of its extremely small training dataset size (n=10).

Evaluating the AutoML safeguards against overfitting

To evaluate our AutoML system's resistance to overfitting, we simulated the process of overfitting. In this simulation, we generated a synthetic dataset that was designed such that its features had no correlation with the classifications, i.e. the features were completely uninformative for classifying the state of the subject. This synthetic dataset contained 600 samples and 1000 features, and was separated into a training set (n = 200), a test set (n = 200) and a fresh test set (n = 200). Because all the features were completely random and thus non-informative, any accuracy in classifying the test higher than 0.5 by a significant amount ($Atest > 0.55$) suggest the existence of overfitting in the training procedure.

We simulated adaptive data analysis, i.e. "overfitting by graduate student", that is, adaptively updating the model based on feedback that comes from evaluating the model on the test set with the aim of achieving improved prediction performance. The dataset was divided into 100 sets of features, each containing 10 features. For each attempt at classification, a collection of these sets were selected to perform the classification(**SI Fig.1a**). To this end, AutoML first attempted to perform a classification using a model based on one of the 100 sets of features. An additional set of data was then added to the existing panel, and it was included into the panel if the addition of that data led to a significantly higher ($p<0.05$) accuracy. To calculate these statistics, each evaluation was performed in triplicate. As we queried additional panels to add to our existing panel, we expected there to be overfitting to the test set data. This procedure was performed with AutoML without the DP algorithm and with the DP algorithms of various configurations (threshold: $\tau = 0.1$, scales of Laplacian noise: $\sigma = 0.005$, 0.01 and 0.02) to explore the effectiveness of the DP algorithm in inhibiting overfitting. To compare the susceptibility to overfitting of our AutoML to individual machine learning algorithms, the same procedure was also performed using Linear discriminant analysis (LDA), a commonly used algorithm in LB studies[52–54]. To evaluate the ability of our system to to control overfitting, we used the simulated dataset and compared the performance of our AutoML to the direct data reuse case and the case where a conventional non-automated machine learning algorithm was used.

Results and Discussion

Our AutoML platform either matched or exceeded the performance of algorithms presented in the literature. AutoML achieved greater AUC for seven out of ten datasets, and for the other four the difference was not significant ($p>0.05$). This comparison of AUC scores suggests that for the majority of datasets, naive AutoML, i.e. AutoML without access to prior LB data, is capable of generating similar or better prediction models than a customized model by a researcher (**Fig.2a**). Additionally, in most cases, the performance of our AutoML platform improved when our platform's data repository included LB-specific datasets, compared to the "naive" case that used only the data incorporated into the autosklearn package.(**Fig.2b**) There were two datasets[21,26] for which AutoML underperformed

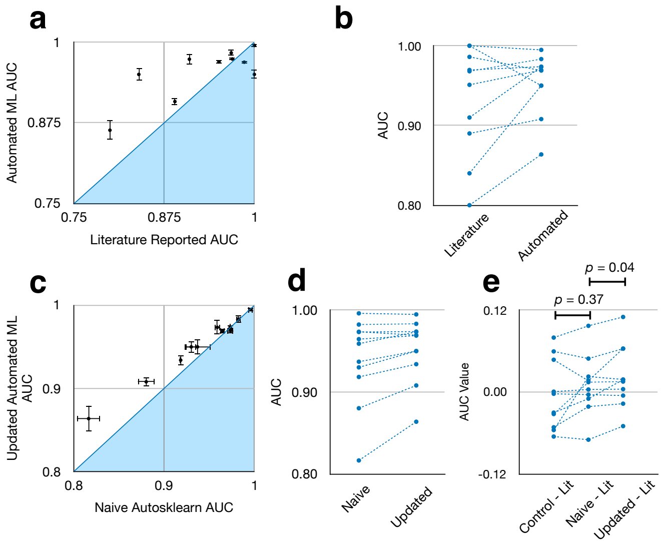


Fig. 2 Characterization of automated machine learning using published data. a.Comparison of the AUC of automated machine learning (AutoML) versus literature reported values. Error bars represent standard error from N = 5 independent AutoML training and evaluations for each dataset using the same training and test set partitioning. The majority of data fell in the upper left hand half of the plot, indicating that AutoML outperformed the literature reported values most of the time. b. Pairwise comparison between literature and AutoML. c. Comparison of the AUC of AutoML, updated with LB datasets loaded into its data repository versus naive AutoML. d. Pairwise comparison between updated and naive AutoML. e. Comparison of the AUC differences between updated AutoML and literature reported values (Updated - Lit), native values and literature values (Naive-Lit), and AutoML updated with a control set of non-LB data and literature reported values (Control - Lit).

the published method, with a difference in AUC of greater than 0.01. There are several potential contributing factors to these differences. In one of these two datasets[26], the data source did not clearly identify whether individual samples belonged to the training set or test set, and therefore our random stratification could have led to this difference. For the other dataset, a study that actually came from our lab, the sample number in both training set(10) and test set(24) were very low, resulting in large changes in AUC associated with one patient that switches from being correctly classified to being non-correctly classified. A paired t-test was performed to compare the performance of our AutoML, updated with a repository of prior LB data, with the results presented in the literature, and it was found that using our platform significantly improved the results.($p<0.05$)(**Fig.2c**). A paired t-test was performed to compare the performance of our AutoML, with and without access to prior LB datasets, and it was found that inclusion of the LB data into the repository significantly improved the results ($p<0.05$)(**Fig.2d**).

Similarities shared by some or all of the datasets used in this study, including label number, data purpose, content, and classifier choice could potentially contribute to the improvement of prediction after meta-learning update. First, all 11 datasets included were binary datasets. Second, 7 out of the 11 datasets[21–27] were used for the purpose of cancer classification. Third, the information of these datasets was mainly miRNA and protein information. Another similarity shared by liquid biopsy datasets is the homogeneity of number of samples and features, where features are typically in the thousands and the number of subjects is no more than a few hundred. The above similarities could explain why including LB specific datasets in the data repository improved the results of classifying prospective LB datasets.

Additionally, a control case was considered where non-LB data was added that matched the size and structure of the LB datasets to evaluate the impact of incorporating non-LB datasets on the performance of our AutoML's performance on LB datasets (**SI Fig.2a and b**). A paired t-test was performed to compare the enhancement of our AutoML's AUC scores over the literature-reported AUCs with access to prior LB datasets and with access to a control set of non-LB data, and it was found that inclusion of the LB data into the repository significantly improved the results compared to the control.($p < 0.05$) (**Fig.2e**) And, it was found that the control data did not make a significant change in such enhancement compare to using the default data in Autosklearn ($p > 0.05$). The 10 non-liquid-biopsy datasets that were used to update our AutoML was size-matched to the 11 liquid biopsy datasets employed in this study. Using these non-LB datasets, Auto-ML did not perform better than the naive case ($P > 0.05$), demonstrating that AutoML benefited specifically from other LB datasets, which share similarities of data characteristics and/or "preferred" algorithms and hyperparameters adopted by Autosklearn.

We demonstrated that our AutoML can achieve the same performance as demonstrated in the literature using a training set with fewer subjects than used in the literature (**Fig.3**). When only 40% of the training set was used to generate our model, the median value of the AUC to classify the test set across all of the datasets evaluated in this study, matched that reported in the literature ($AUC = AUC_{lit}$). Therefore, by using AutoML, researchers can potentially require fewer subjects to be recruited for the training set, which has the potential to reduce the cost and time for LB studies.

When we modeled overfitting caused by adaptive machine learning, we found that our AutoML platform was able to reduce overfitting compared to using standard reuse (SR) of the test set. When we implemented SR significant overfitting to the test set was observed, which increased with subsequent queries of the test set and the addition of features to the panel (**Fig.4a**). By performing 100 queries of evaluating the test set using AutoML without differential privacy, 50 non-informative features were selected that were able to achieve an accuracy predicting the test set of $A = 0.63$. Such significant overfitting with SR implies that Autosklearn is not completely invulnerable to overfitting to the test set caused by adaptively reusing test set data. When a fresh test set was evaluated, it resulted in an accuracy of 0.5, consistent with the fact that none of the features were informative (**Fig.4a**).

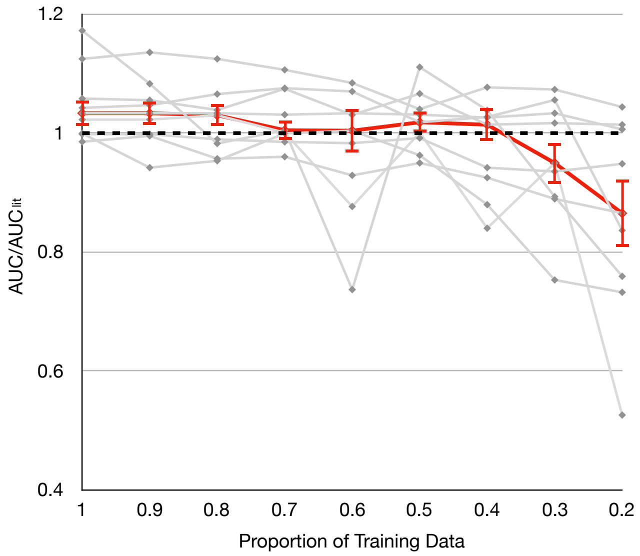


Fig. 3 An experiment was performed where our AutoML algorithm was trained with only a fraction of the data in training set. The purpose of this experiment was to evaluate whether our platform could achieve similar results to the literature with smaller amounts of training data. The performance of the automated platform in evaluating the test set (AUC) divided by the literature reported value (AUC_{lit}) was recorded for each individual LB dataset (grey) using varying fractions of the training set. The medians performance of all of the datasets, as well as the standard error (red) is also plotted. The dashed line signifies when the performance matches that demonstrated in the literature ($AUC = AUC_{lit}$)

We additionally compared our AutoML's capability to avoid overfitting to a commonly used algorithm LDA. When we used LDA, a total of 130 noninformative features were selected and an accuracy of 0.67 was achieved classifying the test set, after 100 queries, which was greater than that achieved by SR of AutoML (**Fig.4a**). In addition, it was demonstrated (**Fig.4c**) when reusing the test set data, LDA requires a lower (n = 11) number of queries to exceed the pre-defined overfitting threshold of $Atest$ = 0.55 comparing to that of SR with AutoML (n = 42). In summary, AutoML with SR was able to limit overfitting significantly better than a simple algorithm like LDA. And, this overfitting could be reduced further by incorporating differential privacy.

The DP algorithm that we incorporated to control the effects of overfitting successfully inhibited overfitting. We further demonstrated that the amount of overfitting could be tuned, such that a tradeoff between useful feedback and overfitting can be controlled, using various scales of Laplacian noises (σ= 0.005, 0.01 and 0.02). When the highest Laplacian noise term that we considered - σ=0.02 was applied, even at 100 queries a test set accuracy of only 0.53 was reached (**Fig.4b and c**). Additionally, AutoML with DP(σ= 0.02) did not reach the pre-defined overfitting threshold of $Atest$ = 0.55 at 100 queries, while LDA and AutoML with SR exceeded this threshold at query 11 and 42 respectively (**Fig.4d**). It was also demonstrated that the number of queries required to achieve significant overfitting ($Atest$ > 0.55) compared to the fresh test set (p<0.05) increased with the increase in Laplacian noise (**Fig.4e**), demonstrating that

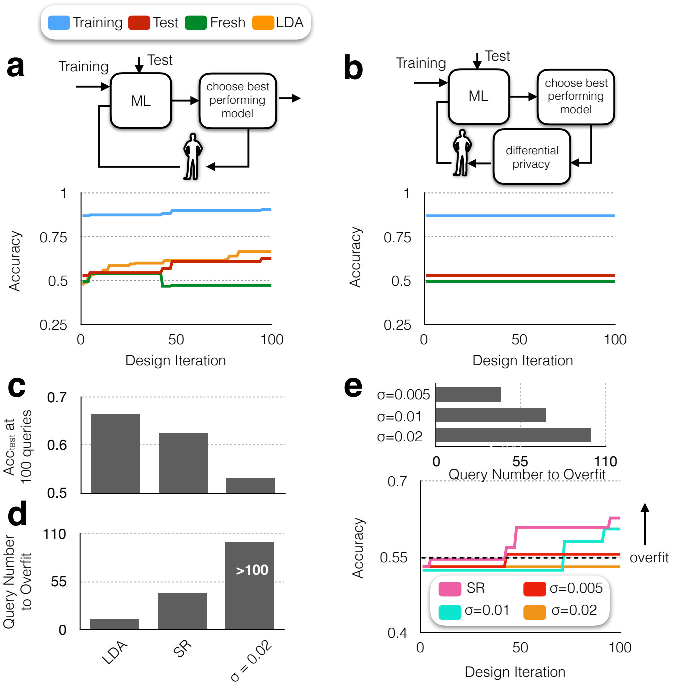
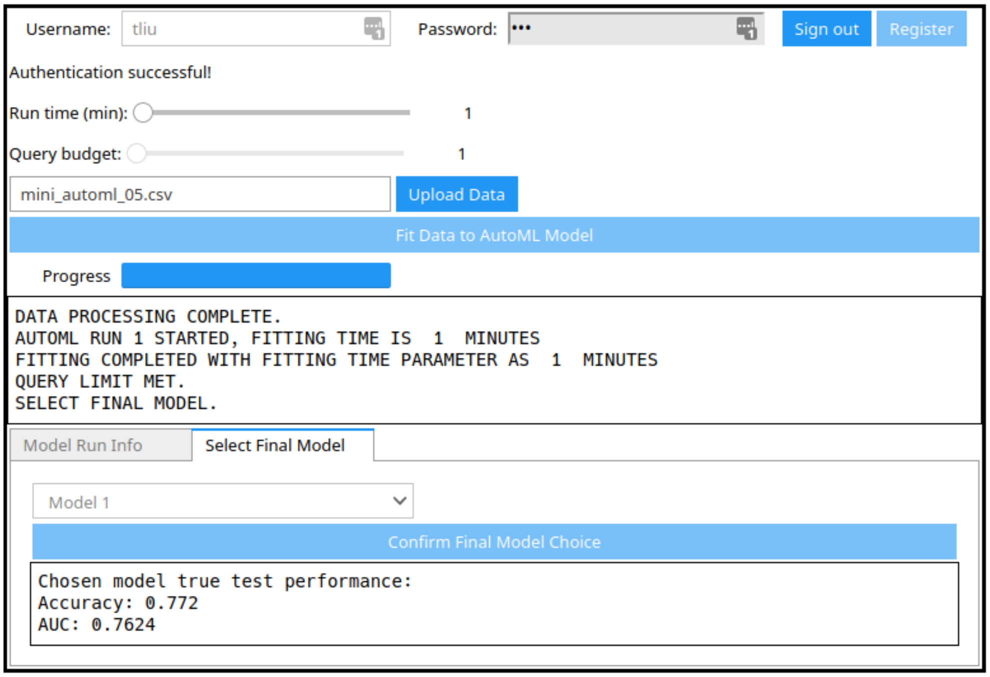


Fig. 4 Evaluating the capability of our AutoML to control overfitting both with and without Differential Privacy (DP) algorithm a.The accuracy of the machine learning model was reported for the classification of the training set, the test set, and a fresh test set that has never been seen before in each design iteration with Standard Reuse (SR) of the test set data. Additionally, the results of linear discriminant analysis (LDA), which is more susceptible to overfitting than our AutoML. b. The accuracy of the training set, the test set, and the fresh test set of AutoML implemented with DP. The configuration of DP in this example is with a threshold of 0.1 and a Laplacian noise with a scale of σ= 0.02. c. Comparisons of test set accuracy at 100 queries of LDA, SR, and DP with Laplacian noise scale σ= 0.02 and a threshold of 0.1. d. Comparisons of the number of design queries required to achieve overfitting ($Atest$ > 0.55) of LDA, SR and DP with Laplacian noise scale σ= 0.02 and a threshold of 0.1. e. The accuracies of test set for LDA, SR and three DP experiments with Laplacian noise with scales σ= 0.005, 0.01 and 0.02, and a threshold of 0.1. Additionally, the number of queries required to achieve overfitting (A > 0.55) of the these three DP experiments were compared.

the amount of information fed back to the user in each evaluation of the test set can be traded off with the number of allowed queries(**Fig.4e**). As expected, similar to SR, all DP experiments have fresh test set accuracies A = 0.50 (**Fig.4b, SI Fig.3a and b**).

To facilitate the use of automatic machine learning in the LB research community, we built a web-based interface (Our web interface can be found here: `https://asklb.page.link/run`) for our LB platform (**Fig.5**). Using Jupyter notebooks for implementation, we provide a graphical user interface (GUI) so that researchers with little experience in these technologies or in machine learning may still perform predictive analysis. Users register with our service and upload their dataset— a .csv file containing all the samples with test set after training set, through the GUI. We then store users' information and datasets with their approval for the meta-learning and differential privacy processes; the dataset will be incorporated into auto-sklearn's meta-learned



Username: tliu
Password: •••

Authentication successful!

Run time (min): 1

Query budget: 1

mini_automl_05.csv

```
DATA PROCESSING COMPLETE.
AUTOML RUN 1 STARTED, FITTING TIME IS  1  MINUTES
FITTING COMPLETED WITH FITTING TIME PARAMETER AS  1  MINUTES
QUERY LIMIT MET.
SELECT FINAL MODEL.
```

Model 1

```
Chosen model true test performance:
Accuracy: 0.772
AUC: 0.7624
```

Fig. 5 Web-based Graphical User Interface. A web-based graphic user interface (GUI) is created to make our automatic machine learning platform accessible to researchers who are non-experts in machine learning. The user is allowed to define several parameters, including the run time to train the dataset and the maximum number of queries on the test set to allow. At the end of each individual query, the GUI reports the accuracy modified by differential privacy (DP) algorithm and information about the machine learning model.

model initialization. The user is asked to input the numbers of samples in the test set, so that the splitting of the training and test sets can be performed automatically. Subsequently, the user determines the time budget for the service to train the data. To enable the differential privacy algorithm, the users may then select the number of queries they want to make to iteratively train models and perform feature engineering to their dataset. At the end of each run, they are given a test accuracy score, with added noise from the differential privacy algorithm, and can then re-upload adjusted dataset features accordingly based on this information. After their query budget is exhausted, the users will select a final model they wish to use, and the true test set performance, including the test set accuracy and AUC score, is revealed to them. Our platform will increase the accessibility of modern machine learning with built-in safeguards against overfitting to LB researchers, regardless of their expertise in these methods. To minimize the risk of unintended disclosure of uploaded dataset, the user is requested to remove all identifiable information from the dataset prior to uploading. Specifically, it is required that the headers of the dataset are completely removed and it is suggested to have labels of all the samples replaced with numeric values. Furthermore, through exposure to varied LB datasets, our platform will continuously and automatically improve its performance on prediction tasks within the field over time.

Conclusion

In this study we demonstrated that an automated, web based machine learning platform, which is resistant to overfitting to the test set, can generate and evaluate machine learning models to accurately classify diseases based on LB data. We demonstrated in a meta-analysis of literature reported data that our automated system can match and, in many cases, beat the performance of machine learning platforms built specifically for that study. Because of this capability, we recommend that our platform be used to either replace custom machine learning development or to be used as a companion gold standard, which custom developed machine learning algorithms can be compared to in their evaluation. Moreover, because of its automation and accessibility, it is our aim that this tool will lower the barrier of entry for LB experts that wish to use ML. Additionally, the community will continue to benefit from the use of a shared tool, as performance tends to increase the more LB-specific data that our platform is exposed to.

There are several noteworthy limitations to this study. First, the datasets included were all binary datasets. We could not identify enough published multiclass datasets to power a sufficient study. Second, we applied our AutoML to a relatively narrow selection of liquid biopsy topics including cancer, traumatic brain injury, depression, and inflammatory disease due to what was available in the literature. Third, due to the fact that our AutoML platform only utilizes supervised machine learning, it shares the common limitation of supervised learning, namely, the performance of the model requires proper categorization and correct labeling of the samples. Therefore, our AutoML is not capable of addressing human mistake which reduces the informativeness of the training set. Several potential follow up studies can be conducted to further explore the potential of our AutoML platform, including an evaluation of the change of its performance over times as more liquid biopsy datasets from users are integrated. Furthermore, different versions of the platform can be tuned and evaluated to specialize in more topics other than liquid biopsy, so that a broader range of research can benefit from an efficient, automated and overfitting-resistant machine learning platform.

Just like in statistics where multiple ways of analyzing data allows p-hacking, flexibility in machine learning allows practitioners to obtain misleading results. The field thus needs to develop standards that minimize this problem. We see three potential solutions. 1. use of a true "lockbox" test set that is only made available after the machine learning model is finalized, 2. preregistration of machine learning models, including hyperparameter settings, before any data is obtained, 3. the use of differential privacy as we have presented here. Though there is no conceptual problem with lockbox test sets, they make research slower and more expensive. Pre-registration does not allow meaningful experimentation with models and ideas. Here, we have offered a third possibility which allows both rapid research and efficient control of overfitting. We believe that this differential privacy solution thus provides flexibility to the experimenter while ensuring the results achieved are legitimate and transferable. The field, just like the other domains of applied machine learning, needs strong standards that can help stem the flood of machine learning results that do not generalize to the real world[32–34]. We believe that differential privacy, as implemented in this platform, may be a good way of simultaneously allowing rapid progress and minimizing the prevalence of misleadingly positive results.

Conflicts of interest

In accordance with our policy on Conflicts of interest, David Issadore is a founder and holds equity in Chip Diagnostics.

References

1 E. Crowley, F. Di Nicolantonio, F. Loupakis and A. Bardelli, *Nature reviews Clinical oncology*, 2013, **10**, 472.

2 J. V. Carter, N. J. Galbraith, D. Yang, J. F. Burton, S. P. Walker and S. Galandiuk, *British journal of cancer*, 2017, **116**, 762.

3 I. A. Cree, L. Uttley, H. B. Woods, H. Kikuchi, A. Reiman, S. Harnan, B. L. Whiteman, S. T. Philips, M. Messenger, A. Cox *et al.*, *BMC cancer*, 2017, **17**, 697.

4 M. J. Duffy and K. O'Byrne, *Advances in clinical chemistry*, Elsevier, 2018, vol. 86, pp. 1–21.

5 M. Z. Bidin, A. M. Shah, J. Stanslas and C. T. S. Lim, *Clinica Chimica Acta*, 2019.

6 E. Domenici, D. R. Willé, F. Tozzi, I. Prokopenko, S. Miller, A. McKeown, C. Brittain, D. Rujescu, I. Giegling, C. W. Turck *et al.*, *PLoS one*, 2010, **5**, e9166.

7 T. Hisamatsu, S. Okamoto, M. Hashimoto, T. Muramatsu, A. Andou, M. Uo, M. T. Kitazume, K. Matsuoka, T. Yajima, N. Inoue *et al.*, *PloS one*, 2012, **7**, e31131.

8 N. Fukutake, M. Ueno, N. Hiraoka, K. Shimada, K. Shiraishi, N. Saruki, T. Ito, M. Yamakado, N. Ono, A. Imaizumi *et al.*, *PLoS One*, 2015, **10**, e0132223.

9 A. Hohl, J. da Silva Gullo, C. C. P. Silva, M. M. Bertotti, F. Felisberto, J. C. Nunes, B. de Souza, F. Petronilho, F. M. S. Soares, R. D. S. Prediger *et al.*, *Journal of critical care*, 2012, **27**, 523–e11.

10 J. Zielenski and L.-C. Tsui, *Annual review of genetics*, 1995, **29**, 777–807.

11 J.-H. Kang, B. Mollenhauer, C. S. Coffey, J. B. Toledo, D. Weintraub, D. R. Galasko, D. J. Irwin, V. Van Deerlin, A. S. Chen-Plotkin, C. Caspell-Garcia *et al.*, *Acta neuropathologica*, 2016, **131**, 935–949.

12 T. Bird, S. Sumi, E. Nemens, D. Nochlin, G. Schellenberg, T. Lampe, A. Sadovnick, H. Chui, G. Miner and J. Tinklenberg, *Annals of neurology*, 1989, **25**, 12–25.

13 J. Ko, S. N. Baldassano, P.-L. Loh, K. Kording, B. Litt and D. Issadore, *Lab on a Chip*, 2018, **18**, 395–405.

14 J. A. Eastham, E. Riedel, P. T. Scardino, M. Shike, M. Fleisher, A. Schatzkin, E. Lanza, L. Latkany, C. B. Begg, P. P. T. S. Group *et al.*, *Jama*, 2003, **289**, 2695–2700.

15 J. V. Pagaduan, V. Sahore and A. T. Woolley, *Analytical and bioanalytical chemistry*, 2015, **407**, 6911–6922.

16 S. T. Sanjay, G. Fu, M. Dou, F. Xu, R. Liu, H. Qi and X. Li, *Analyst*, 2015, **140**, 7062–7081.

17 S. A. Hogan, M. P. Levesque and P. F. Cheng, *Frontiers in Oncology*, 2018, **8**, year.

18 P. A. VanderLaan, D. Rangachari, A. Majid, M. S. Parikh, S. P. Gangadharan, M. S. Kent, D. C. McDonald, M. S. Huberman, S. S. Kobayashi and D. B. Costa, *Lung Cancer*, 2018, **116**, 90–95.

19 T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont and Y. Saeys, *Bioinformatics*, 2009, **26**, 392–398.

20 A. L. Swan, A. Mobasheri, D. Allaway, S. Liddell and J. Bacardit, *Omics: a journal of integrative biology*, 2013, **17**, 595–610.

21 J. Ko, N. Bhagwat, S. S. Yee, N. Ortiz, A. Sahmoud, T. Black, N. M. Aiello, L. McKenzie, M. O'Hara, C. Redlinger *et al.*, *ACS nano*, 2017, **11**, 11182–11193.

22 F. Bianchi, F. Nicassio, M. Marzi, E. Belloni, V. Dall'Olio, L. Bernard, G. Pelosi, P. Maisonneuve, G. Veronesi and P. P. Di Fiore, *EMBO molecular medicine*, 2011, **3**, 495–503.

23 S. Huang, N. Chong, N. E. Lewis, W. Jia, G. Xie and L. X. Garmire, *Genome medicine*, 2016, **8**, 34.

24 C. G. A. Network *et al.*, *Nature*, 2012, **490**, 61.

25 H. G. LaBreche, J. R. Nevins and E. Huang, *BMC medical genomics*, 2011, **4**, 61.

26 M. G. Best, N. Sol, I. Kooi, J. Tannous, B. A. Westerman, F. Rustenburg, P. Schellen, H. Verschueren, E. Post, J. Koster *et al.*, *Cancer cell*, 2015, **28**, 666–676.

27 M. G. Best, N. Sol, S. GJG, A. Vancura, M. Muller, A.-L. N. Niemeijer, A. V. Fejes, L.-A. T. K. Fat, A. E. Huis, C. Leurs *et al.*, *Cancer cell*, 2017, **32**, 238–252.

28 J. Ko, M. Hemphill, Z. Yang, K. Beard, E. Sewell, J. Shallcross, M. Schweizer, D. K. Sandsmark, R. Diaz-Arrastia, J. Kim *et al.*, *Journal of Neurotrauma*, 2019.

29 C. L. Clelland, L. L. Read, L. J. Panek, R. H. Nadrich, C. Bancroft and J. D. Clelland, *PLoS One*, 2013, **8**, e69082.

30 L. L. Koth, O. D. Solberg, J. C. Peng, N. R. Bhakta, C. P. Nguyen

and P. G. Woodruff, *American journal of respiratory and critical care medicine*, 2011, **184**, 1153–1163.

31 D. Chicco, *BioData mining*, 2017, **10**, 35.

32 O. DeMasi, K. Kording and B. Recht, *PLOS ONE*, 2017, **12**, 1–15.

33 S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr and K. P. Kording, *bioRxiv*, 2016.

34 M. Skocik, J. Collins, C. Callahan-Flintoft, H. Bowman and B. Wyble, *bioRxiv*, 2016.

35 M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum and F. Hutter, Advances in neural information processing systems, 2015, pp. 2962–2970.

36 C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold and A. Roth, *Science*, 2015, **349**, 636–638.

37 J. P. Ioannidis, *PLoS medicine*, 2005, **2**, e124.

38 J. P. Simmons, L. D. Nelson and U. Simonsohn, *Psychological science*, 2011, **22**, 1359–1366.

39 A. Gelman and E. Loken, *Chance*, 2014, **27**, 51–56.

40 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, *Journal of machine learning research*, 2011, **12**, 2825–2830.

41 R. Caruana, A. Niculescu-Mizil, G. Crew and A. Ksikes, Proceedings of the twenty-first international conference on Machine learning, 2004, p. 18.

42 M. Feurer, J. T. Springenberg and F. Hutter, Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

43 F. Hutter, H. H. Hoos and K. Leyton-Brown, International conference on learning and intelligent optimization, 2011, pp. 507–523.

44 J. S. Bergstra, R. Bardenet, Y. Bengio and B. Kégl, Advances in neural information processing systems, 2011, pp. 2546–2554.

45 L. Breiman, *Machine learning*, 2001, **45**, 5–32.

46 K. Eggensperger, M. Feurer, F. Hutter, J. Bergstra, J. Snoek, H. Hoos and K. Leyton-Brown, NIPS workshop on Bayesian Optimization in Theory and Practice, 2013, p. 3.

47 P. Brazdil, C. G. Carrier, C. Soares and R. Vilalta, *Metalearning: Applications to data mining*, Springer Science & Business Media, 2008.

48 D. Michie, D. J. Spiegelhalter, C. Taylor *et al.*, *Neural and Statistical Classification*, 1994, **13**, year.

49 A. Kalousis, *PhD thesis*, University of Geneva, 2002.

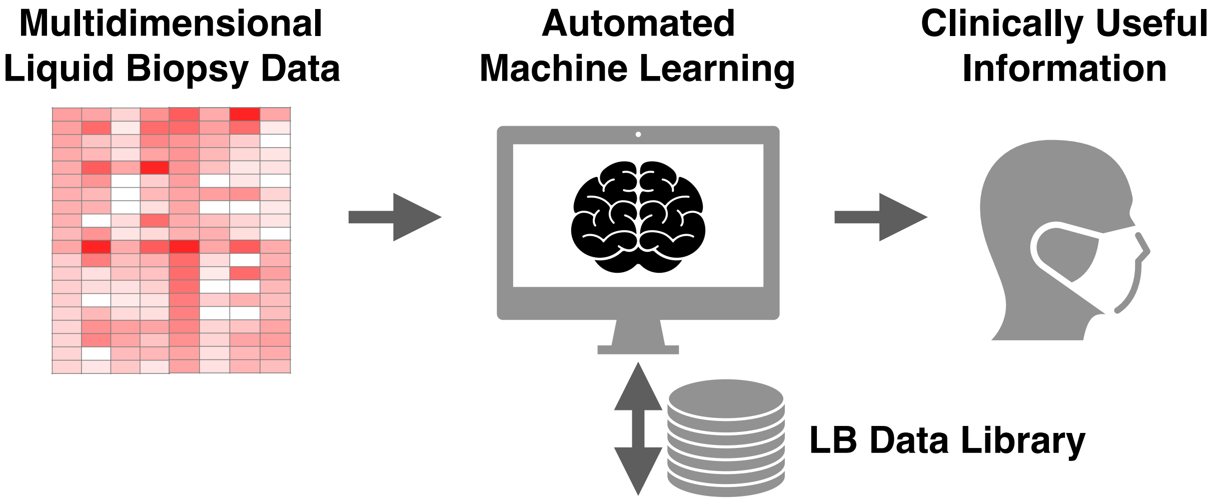
50 J. Vanschoren, J. N. Van Rijn, B. Bischl and L. Torgo, *ACM SIGKDD Explorations Newsletter*, 2014, **15**, 49–60.

51 D. Dua and C. Graff, *UCI Machine Learning Repository*, 2017, `http://archive.ics.uci.edu/ml`.

52 Y. Miyagi, M. Higashiyama, A. Gochi, M. Akaike, T. Ishikawa, T. Miura, N. Saruki, E. Bando, H. Kimura, F. Imamura, M. Moriyama, I. Ikeda, A. Chiba, F. Oshita, A. Imaizumi, H. Yamamoto, H. Miyano, K. Horimoto, O. Tochikubo, T. Mitsushima, M. Yamakado and N. Okamoto, *PLOS ONE*, 2011, **6**, 1–12.

53 Q.-T. Le, P. D. Sutphin, S. Raychaudhuri, S. C. T. Yu, D. J. Terris, H. S. Lin, B. Lum, H. A. Pinto, A. C. Koong and A. J. Giaccia, *Clinical Cancer Research*, 2003, **9**, 59–67.

54 S. Suryawanshi, A. M. Vlad, H.-M. Lin, G. Mantia-Smaldone, R. Laskey, M. Lee, Y. Lin, N. Donnellan, M. Klein-Patel, T. Lee, S. Mansuria, E. Elishaev, R. Budiu, R. P. Edwards and X. Huang, *Clinical Cancer Research*, 2013, **19**, 1213–1224.


Multidimensional Liquid Biopsy Data
Automated Machine Learning
Clinically Useful Information
LB Data Library

We have developed a web-based, self-improving and overfitting-resistant automated machine learning tool tailored specifically for liquid biopsy data, where machine learning models can be built without the user's input.