Lab on a Chip

TUTORIAL REVIEW

View Article Online
View Journal | View Issue

Cite this: *Lab Chip*, 2018, **18**, 395

Machine learning to detect signatures of disease in liquid biopsies – a user's guide

Jina Ko,†[a] Steven N. Baldassano,†[a] Po-Ling Loh,[b] Konrad Kording,[ac] Brian Litt[ad] and David Issadore *[ae]

Received 5th September 2017,
Accepted 23rd November 2017

DOI: 10.1039/c7lc00955k

rsc.li/loc

New technologies that measure sparse molecular biomarkers from easily accessible bodily fluids (*e.g.* blood, urine, and saliva) are revolutionizing disease diagnostics and precision medicine. Microchip devices can measure more disease biomarkers with better sensitivity and specificity each year, but clinical interpretation of these biomarkers remains a challenge. Single biomarkers in 'liquid biopsy' often cannot accurately predict the state of a disease due to heterogeneity in phenotype and disease expression across individuals. To address this challenge, investigators are combining multiplexed measurements of different biomarkers that together define robust signatures for specific disease states. Machine learning is a useful tool to automatically discover and detect these signatures, especially as new technologies output increasing quantities of molecular data. In this paper, we review the state of the field of machine learning applied to molecular diagnostics and provide practical guidance to use this tool effectively and to avoid common pitfalls.

Introduction

Researchers in the field of 'liquid biopsy' are developing technologies to measure sparse molecular biomarkers shed from inaccessible tissue in easily sampled bodily fluids, such as urine, blood, saliva, sweat, feces, and tears.[1–3] The last decade has seen great progress in this field, and there is a growing list of circulating indicators – rare circulating cells, microvesicles, nucleic acids, proteins, and metabolites – that can be detected. The field has focused primarily on developing minimally invasive sensors that are sufficiently sensitive and specific to detect sparse biomarkers against the complex substrate of clinical samples.[4–8] However, as sensor performance has improved, the field has been driven to develop computational tools to decode the complex biomarker information to inform patient treatment.[9–17] This task is made particularly challenging because of variability in biomarker expression across individuals and because many diseases are phenotypically heterogeneous. As a result, it is rare that a single molecular biomarker can accurately diagnose or prognosticate a disease.[9,18–21] Furthermore, healthy individuals can have variable baseline levels of molecular biomarkers because of a variety of unrelated cofactors, such as genetics and diet.[22] To address these challenges, researchers are simultaneously measuring multiple biomarkers in the hope that their combined signatures are conserved across patients and correlate with states of disease. These multidimensional signatures often include measurements of multiple molecular biomarkers taken at several time points.

Rapid technical innovation in the field of liquid biopsy is producing microchip-based technologies that are increasingly sensitive and specific, are able to measure increasing numbers of biomarkers, and are miniaturized and clinically deployable. These diagnostics incorporate advances in microfluidics, microdroplet-based digital sensing, next generation sequencing (NGS), single-cell RNA sequencing, nanopore sensing, nucleic acid microarrays, nanoelectronic sensing, and electrochemical sensing, among others.[21,23–27] One common theme in miniaturized diagnostics is that the marginal cost of measuring additional biomarkers continues to shrink as on-chip multiplexing improves.[21,23–27] There are many excellent, recent reviews that cover the state of the art in microchip development.[2,28–30] The advantage of liquid biopsy compared to traditional biopsy, wherein tissue is excised from the body, is that the invasive procedure to remove the tissue is obviated, enabling more frequent measurements of diseases in tissues that would otherwise be impossible (*e.g.* brain tissue). The machine learning approaches described in this paper can also be applied to measurements on tissues obtained using traditional biopsy, but our focus is on liquid biopsy due to the emerging microchip technology to extract

[a] *Department of Bioengineering, School of Engineering and Applied Sciences, University of Pennsylvania, Philadelphia, Pennsylvania, USA. E-mail: daveissadore@gmail.com*
[b] *Department of Electrical and Computer Engineering, University of Wisconsin – Madison, Madison, Wisconsin, USA*
[c] *Department of Neuroscience, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA*
[d] *Department of Neurology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA*
[e] *Department of Electrical and Systems Engineering, School of Engineering and Applied Sciences, University of Pennsylvania, Philadelphia, Pennsylvania, USA*
† Equal contribution.

increasingly large quantities of molecular data. Moreover, due to the typical lack of knowledge on the mechanisms behind the release and dynamics of blood-based biomarkers, it is often not possible to rationally design a panel of biomarkers that can classify a specific disease state. This lack of knowledge further motivates the use of machine learning approaches to interpret liquid biopsy data.[73]

In addition to their improved performance, lab-on-a-chip systems are becoming increasingly low cost and automated, propelling them from engineering research laboratories into the hands of caregivers and technicians. As this hardware becomes established in new clinical devices, effort is being shifted to the informatics necessary to turn chip outputs into clinically useful information. Automated data processing and thoughtfully engineered user interfaces are becoming increasingly important to deliver clinically capable, next-generation devices. Advances in miniaturized computing and cloud-connected devices have helped facilitate the incorporation of automated data processing and machine learning into point-of-care diagnostic systems.[31,32]

Machine learning encompasses a set of computational techniques widely applied in many fields to reduce large numbers of measurements into lower-dimensional outputs that are more useful.[11,12,15,17,33–38] In recent years, machine learning algorithms have been increasingly applied to liquid biopsy data to aid in disease diagnosis, prediction, and medical decision-making. A growing set of studies use these

Table 1 Liquid biopsy studies that have used various machine learning algorithms to analyze and evaluate the diagnostic performance. SVM (support vector machine), ORD (other respiratory diseases), LOOCV (leave-one-out cross validation), OLK (oral leukoplakia), OSCC (oral squamous cell carcinoma), FLDA (Fisher linear discriminant analysis), NSCLC (non-small cell lung cancer)

Publication	Disease	Method	Sample	Biomarker	No. of subjects	Accuracy	Groups
Pinto *et al.* (2017)	Schizophrenia, bipolar disorder	SVM	Serum	BDNF, CCL11, GPxActivity, GSTActivity, IL6, IL10, *S*-transferase	$N = 20$ per group	72.5% 77.5%	Bipolar disorder *vs.* healthy Schizophrenia *vs.* healthy
Kenny *et al.* (2005)	Pre-eclampsia	Genetic programming	Plasma	3 metabolites	$N = 87$ per group	99%	Pre-eclampsia *vs.* healthy
Jacobs *et al.* (2016)	Tuberculosis (TB)	General discriminant analysis	Plasma	6 host markers	$N = 22$ TB patients $N = 33$ ORD	94% 100%	TB *vs.* ORD HIV infected TB *vs.* ORD
Lugli *et al.* (2015)	Alzheimer disease (AD)	J48 decision trees SVM AdaboostM1	Exosome	7 exosomal miRNAs	$N = 35$ per group	83–89%	AD *vs.* healthy
Agranoff *et al.* (2006)	Tuberculosis	SVM	Serum	20 mass peaks	$N = 179$ TB, $N = 170$ healthy	94%	TB *vs.* healthy
Manterola *et al.* (2014)	Glioblastoma (GBM)	Logistic regression	Exosome	RNU6, miR-320, miR-574-3p	$N = 75$ GBM $N = 55$ healthy	70%	GBM *vs.* healthy
Best *et al.* (2015)	6 cancers	SVM/LOOCV	Tumor-educated platelet	1072 RNAs	$N = 283$	96% 71% 85–95%	Cancer *vs.* healthy 6 cancers and healthy Mutants *vs.* wild type
Kim *et al.* (2016)	Extracapsular prostate cancer	GLMs (generalized linear models)	Urine	34 peptides	$N = 281$	63% 74%	Prostate cancer *vs.* healthy, benign prostatic hyperplasia (BPH) pT2 *vs.* pT3 stage
Banerjee *et al.* (2016)	Oral leukoplakia and cancer	SVM	Oral exfoliative cells	Liquid based exfoliative cytology (LBEC) images	$N = 39$	100%	OLK *vs.* OSCC
Cazzoli *et al.* (2013)	Lung cancer	Logistic regression	Plasma	4–6 exosomal miRNAs	$N = 135$	82.5% 91.4%	Lung adenocarcinoma *vs.* Lung granulomas Lung cancer *vs.* healthy smokers
Roth *et al.* (2011)	Glioblastoma	SVM	Blood	22 miRNAs	$N = 20$ per group	81%	GBM *vs.* healthy
Noerholm *et al.* (2012)	Glioblastoma	Unsupervised clustering	Serum	Microvesicle RNA	$N = 20$	NA	GBM *vs.* healthy
Nebozhyn *et al.* (2006)	Cutaneous T-cell lymphoma (CTCL)	SVM, FLDA	PBMC	STAT4, GATA-3, PLS3, CD1D, and TRAIL	$N = 150$	90%	Sézary syndrome *vs.* healthy
Ponomaryova *et al.* (2013)	Lung cancer	Random Forest	Blood	DNA methylation levels	$N = 92$	83%	NSCLC *vs.* healthy
Honda *et al.* (2005)	Pancreatic cancer	SVM	Plasma	Proteins	$N = 245$	91%	Pancreatic cancer *vs.* healthy
Lodes *et al.* (2009)	5 cancers	Decision tree	Serum	miRNAs	$N = 36$	100%	Cancer *vs.* healthy

approaches to identify signatures in multiple circulating biomarkers for a wide range of applications, including cancer, tuberculosis, dengue fever, heart disease, liver disease, brain disease, and diabetes.[9,10,12,13,15,16,21,39–52] These studies employ a variety of machine learning algorithms, including support vector machines, decision trees, and random forests, that outperform the sensitivity and specificity of individual markers in many applications (Table 1).

Applying machine learning to liquid biopsy

Machine learning algorithms build a model from sample inputs and use that model to make predictions based on subsequent data. Generally, machine learning algorithms fall into two main categories: supervised and unsupervised learning. In supervised learning, the algorithm is provided a set of training data wherein the true state of the data is known, such as which subjects have cancer and which subjects are healthy. Based on this training data, the algorithm generates a model that is deployed to predict the state of subsequent subjects for which the true state is not known. These predictions can take the form of a 'classification problem', identifying a set of discrete states (such as the stage of a patient's cancer), or a 'regression problem', across a set of continuous variables (such as the volume of a developing tumor). In unsupervised learning, on the other hand, algorithms search for patterns in sets of data without labeled states. These algorithms, such as clustering methods, may be used to investigate the structure or distribution of a dataset, discover groups of similar examples within the data, or reduce data dimensionality.

Supervised and unsupervised learning techniques are each useful for specific applications within the field of liquid biopsy. Often, biomarkers are collected to characterize group of subjects such that each subject has a single label (for example, which patients have or do not have the target condition). In these cases, algorithm training is limited by data collection and sample size, making supervised learning techniques most useful. However, machine learning can also be applied in scenarios in which many data points can be collected, but labeling of these data points is resource intensive or unreliable.[78] For example, if the goal is to classify individual cells as either tumor cells or healthy cells, it may be possible to record biomarkers from hundreds or thousands of cells but impractical to manually record the true label for each cell. In these cases, sample labeling is the primary bottleneck for algorithm training and unsupervised learning or active learning methods should be used.[94] By clustering samples with similar features, it may be possible to differentiate between populations without the need for explicit labeling. In this particular application of identifying tumor cells, unsupervised techniques (using a generative mixture model) have been used with performance approaching that of a supervised method (support vector machine) without the need for labelled data.[79] Similar unsupervised approaches have also been used to identify circulating tumor cells based on genetic clustering,[80] identify cell-free DNA of tumor origin,[81] and classify tumor-educated platelets based on RNA expression profiles.[46,82] Unsupervised principal component analysis also offers a method to reduce the dimensionality of complex proteomic[83] or lipidomic[84] spectra sampled with liquid biopsy for patient stratification and classification, even in the absence of per-patient labels.

In this review, we focus on supervised learning applications in which each subject has a single label, the scenario most relevant for diagnostic devices. There is a typical workflow for developing supervised machine learning algorithms (Fig. 1). First, a set of labeled samples is assembled for algorithm training. Each sample in this dataset is described by a set of measured biomarkers. Once this dataset is assembled, it should be partitioned into a 'training dataset', containing labeled samples to be used for algorithm development, and a 'test or validation dataset', containing blinded samples to be used later to assess algorithm performance. All algorithm development and tuning should be restricted to the training dataset. The machine learning algorithm is then fit to the training dataset to generate a model for predicting the labels of subsequent data it has never seen. The performance of this model can be evaluated during development using a technique called cross validation. In cross validation, a subset of the training dataset is purposely left out during model fitting, and the generated model is evaluated by testing its prediction accuracy on that left-out data. Typically this process is performed iteratively, leaving out many possible subsets, or folds, of the data and then averaging performance results. As all testing for each fold is performed on data excluded from model fitting, cross validation provides insight as to how a model will perform on an independent dataset. This technique is a valuable tool for assessing algorithm performance on a specific dataset and for tuning algorithm-specific parameters. Despite the merits of cross validation, a significant issue when developing machine learning algorithms is 'overfitting' the training dataset. Overfitting refers to developing a model that mimics the specific training dataset too closely, such that it performs well on the training data but does not generalize to new data. Such a model may result from excessive experimentation with a single training dataset, due to inadvertent fitting of dataset-specific characteristics. As a result, to properly measure model performance, the trained model should be evaluated on the previously held-out test dataset with even the researcher blinded to the true sample labels. This testing dataset should never be inspected or used in any way during training or algorithm development, and therefore provides an unbiased view of the algorithm's expected performance on independent test data.

Limitations and challenges of applying machine learning to liquid biopsies

Although machine learning offers tremendous potential to improve diagnostics, it is important to consider its

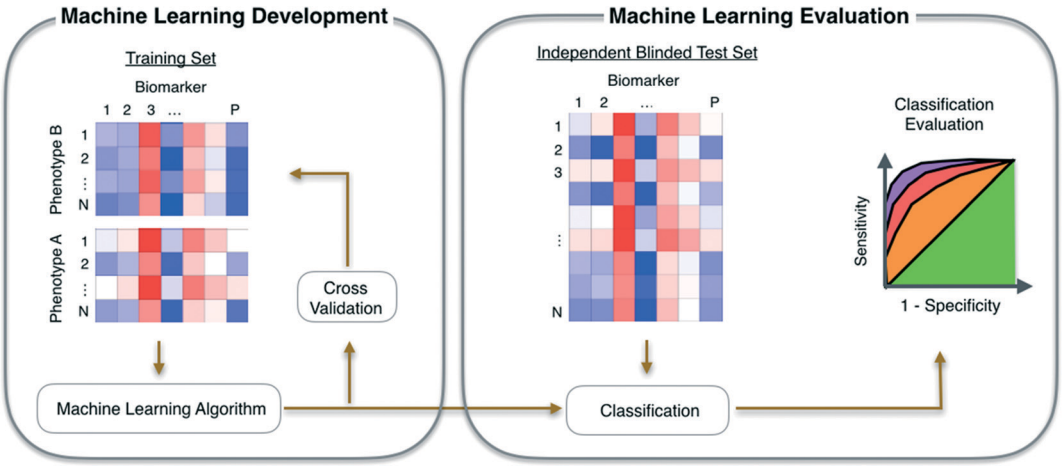

Fig. 1 A generic workflow for developing and evaluating a machine learning based liquid biopsy diagnostic.

limitations. The efficacy of a machine learning algorithm is most significantly determined by the quality, structure, and amount of the underlying data. While complex or computationally-intensive algorithms may provide incremental improvements over simpler models, no level of complexity can compensate for poorly designed, insufficiently sampled, or excessively noisy training data. The critical importance of the training data raises issues that must be considered when developing machine learning algorithms for liquid biopsy applications. Because humans and diseases are so variable, a major challenge in liquid biopsy based studies is to select appropriate cohorts to train and evaluate the system that represent the full range of the disease process. For instance, if one were to develop a diagnostic for tuberculosis (TB), and only trained and evaluated the system for patients that are positive for TB and patients that are healthy, it would be left to chance whether the system could discriminate TB from common alternative diagnoses such as pneumonia. Similarly, because machine learning algorithms are agnostic to which patterns they identify in a dataset, careful thought needs to be taken to design experiments that avoid training on artifacts of sample collection or processing. For example, if all samples from group A are collected and/or measured in one batch and all samples from group B in another batch, it is possible that a discriminative model captures batch-to-batch variation rather than the desired differences between groups. This particular issue can be avoided by including a reference to normalize each batch or by mixing samples from different processing batches across the training and testing sets.

A further challenge to using machine learning in liquid biopsy is the small number of samples typically available to engineers during technology development, due to the time and expense required to collect samples from clinics or animal models. Typical studies that present a new liquid biopsy technology use $n < 50$ samples to train and test the machine learning algorithm.[7,53–55] With such systems, where the number of measured biomarkers p is similar to the number of samples measured n, special consideration needs to be given to mitigate the effects of overfitting, as discussed below. An additional limitation of machine learning is that the generated model is not often easily interpretable.[56] Even if a machine learning algorithm is highly effective, it is typically not possible to understand the fundamental relationship between algorithm structure and the underlying biology, or to glean significant insights to the physical system being modeled. Biomarkers that have important roles classifying different phenotypes in the algorithm do not necessarily have a significant, direct relationship to the biology of the disease. For example, a biomarker may result from an immune response downstream of the disease process of interest. Since it can be difficult to decipher the mechanism driving a machine learning algorithm, caution should be taken when attempting to apply an algorithm to new cohorts sufficiently different from the training subjects, such as translating from an animal model to human clinical samples.

Considerations for choosing a machine learning algorithm

One of the most confusing aspects for newcomers to the field is choosing the correct machine learning algorithm for a particular application. Many machine learning algorithms have exotic sounding names, such as elastic net, least absolute shrinkage and selection operator (LASSO), or random forest, and for a given dataset there are often no obvious criteria for making the best choice. However, careful consideration of a particular data set and the objectives of a study can help to identify a more manageable number of options. Here, we focus on supervised machine learning algorithms, and the methods described below can be applied equally well to both problems that require classification of discrete states and regression of continuous variables. This discussion is limited only to a few key topics for researchers working on liquid biopsy applications with small datasets; a more comprehensive background can be found in Bishop's *Pattern Recognition and Machine Learning*[57] textbook and freely available online

resources including *The Elements of Statistical Learning*[56] by Hastie, Tibshirani, and Friedman, and Andrew Ng's *Machine Learning* course on Coursera.

What is the size of the data?

One of the most important dataset characteristics to consider when choosing a classifier is the size. In machine learning, datasets are typically structured in a matrix of dimensions $n \times p$, where n is the number of observations and p is the number of features. In the case of liquid biopsy, n corresponds to the number of measured samples and p corresponds to the number of biomarkers being measured in each sample. Both dimensions need to be considered when designing a machine learning based diagnostic. The goal of machine learning is to identify and model patterns in the data, and the larger the sample size n, the more clearly these patterns will be represented in the dataset. In general, the more tunable parameters that an algorithm has, the larger the dataset must be to fit an accurate model to it. While machine learning offers the most advantages when analyzing 'big data', it is often untenable to gather thousands of clinical or animal samples when developing a new diagnostic test. In the following section, we suggest specific approaches that should be considered when applying machine learning to 'small data'.

The number of biomarkers p, more generally called features, dictates the dimensionality of the data, which in turn influences algorithm selection. Machine learning datasets are often represented in 'feature space', which is a p-dimensional coordinate space with one dimension representing each feature. Ideally, the number of features must be sufficient to capture the desired trends in the data without too many unnecessary or unhelpful features, which may decrease model generalizability and affect the speed of training and running the algorithm. There are several methods, discussed below, for empirically selecting the best features from the dataset and eliminating those with little or no discriminative value. Therefore, a good first approach is to include all features that may be conceivably useful and to use statistical procedures to choose the best complementary features from the complete set.

What is the application?

When choosing a machine learning algorithm, it is important to first consider the practical application of the algorithm. Will the algorithm be used for offline study of previously collected data or embedded in a device for real-time analysis? If analysis is carried out on a remote system, care must be taken to ensure that all medical data is transmitted and stored in compliance with Health Insurance Portability and Accountability Act (HIPAA) regulations to protect patient privacy. This may entail storing biomarker measurements in a secure database (*e.g.*, an electronic medical record or Research Electronic Data Capture (REDCap) database), and limiting off-board analysis to de-identified patient data. If the algorithm is to be implemented directly in an implanted device, it may be necessary to consider the computational limitations of the system. Processing complexity may need to be curtailed for implanted devices that analyze data continuously due to limitations in on-board processing power, battery life, and acceptable heat generation;[70] however, most liquid biopsy applications use few enough samples and biomarkers that runtime demands are generally not a significant concern.

What kinds of algorithms exist?

While a complete discussion of the breadth of existing machine learning algorithms is outside the scope of this paper, here we provide a view of some of the most popular algorithms likely to be applicable to liquid biopsy. These algorithms are included in standard machine learning libraries and can be quickly implemented in a few lines of code. Commonly used machine learning algorithm packages include R packages (*e.g.* caret, randomForest, e1071, rpart, glmnet), Python libraries (*e.g.* TensorFlow, scikit-learn, Theano, Pylearn2, Pyevolve), and MATLAB Statistics and Machine Leaning Toolbox (*e.g.* SVM, KNN, PCA, ensemble, decision trees). Additionally, these packages include sample data sets that can be used to test the algorithms. The sample data sets can be found in the 'Sample Data Sets' section in the MATLAB & Simulink website, sklearn.datasets package for Python, and the 'Comprehensive R Archive Network (CRAN)' for R.

One of the most basic machine learning algorithms is the naïve Bayes model. This model, traditionally applied to text classification tasks, attempts to classify data by assuming independence among biomarkers and imposing a particular distribution, often a Gaussian distribution, to the data. These strong and generally oversimplified assumptions[85] typically render naïve Bayes the most effective option only in some cases of very small datasets.[86,87]

Another group of algorithms are 'nearest neighbor' methods, such as k-nearest neighbors (KNN). These nonparametric models attempt to classify new data by examining the labels of the nearest training examples in feature space. Nearest neighbor methods do not impose assumptions on the data, but require sufficient sampling, become ineffective with large numbers of features (the 'curse of dimensionality'),[88] and are sensitive to training choices such as the distance metric or number of nearest neighbors to examine.

A commonly used classification algorithm is logistic regression. This model, analogous to standard linear regression, models the probability that data belongs to a certain class using a linear combination of the biomarkers. While logistic regression itself is prone to overfitting, this model can be extended to eliminate less relevant features and favor robustness over maximizing training accuracy (LASSO, ridge regression, and elastic net models).[89]

Another widely applied algorithm is the support vector machine (SVM). This algorithm seeks to separate classes of data by finding the decision boundary that produces the

largest separation, or margin, between classes. In contrast to logistic regression, SVM focuses only on the datapoint at the border between classes, rather than taking obviously correct points into account. SVM can be used to generate linear or nonlinear decision boundaries.

An especially useful class of algorithms for small to medium size datasets is decision-tree-based methods. Decision trees classify data by iteratively splitting the data based on the most informative biomarker, defined in statistical terms as that which produces the largest 'information gain'. Decision trees become particularly useful when grouped into ensembles, such as a random forest, as discussed below. These ensembles can sometimes be trained with 'boosting', a technique designed to assign more weight to samples that are difficult to classify during training. Tree-based models are very popular due to their flexibility, ease of training, and ability to handle correlated or unnecessary features without overfitting.[90,91]

Lastly, deep learning and neural networks are increasingly popular models that rely on layers of interconnected, nonlinear data transformations. While these models can be highly effective in a wide range of tasks,[92] they generally require very large datasets during training to avoid overfitting.[93] This data requirement renders these algorithms inapplicable for most liquid biopsy applications.

Which algorithm should I use?

Even with a clearly defined dataset and objective, it is very rarely possible to predict which algorithm will be most effective. In this case, 'effectiveness' is measured by algorithm performance on a 'held-out' test dataset (*i.e.* a set of data not used to train the algorithm). Sometimes, there may be a benefit to simpler models, such as logistic regression or decision trees. With fewer tunable parameters, these models may be easier to design and, in some cases, are easier to interpret, providing confidence that the model is meaningful and capable of producing insights into the data itself such as which biomarkers are most strongly diagnostic. However, in a typical case with a large number of biomarkers, many of which are weakly diagnostic or correlated with one another, even simple models cannot be interpreted in a useful or reliable manner.

With modern machine learning packages, it can be tempting to apply and compare a large number of algorithms, with a range of tunable parameters for each algorithm, to find the best performer on cross validation. Unfortunately, this approach is prone to overfitting, especially when working with a small training dataset. In most cases, the best model is constructed by combining several algorithms in an ensemble or stack, as is described in more detail in the 'Model averaging and bootstrapping' section. Using a cohort of algorithms mitigates overfitting by any one algorithm, and tends to provide a more effective model than any single method alone. If the individual algorithms have few tunable parameters (as is the case in the previously mentioned algorithms), they can be tuned and evaluated using cross validation without extensive searching or comparison. It is important to remember during training that all such evaluation must be limited to the training dataset to avoid overfitting the test data.

As an example, we compared the performance of various individual algorithms and stacked approaches using real liquid biopsy data from our previous work, Ko *et al. ACS Nano*, 2017, where we measured the expression levels of multiple exosomal mRNAs to diagnose pancreatic cancer patients *versus* healthy controls.[74] We performed 100 iterations of 10-fold cross validation on individual algorithms (SVM, KNN, Logistic regression, Decision tree, Naïve Bayes, Random forest) and on a stack that combines these algorithms. The classification accuracy of individual algorithms ranged from 77.4–92.9%. For the stack, the classification accuracy was 93.7%, and also has the advantage of being less prone to overfitting.

Special consideration for machine learning on 'small data'

Much of the recent progress in the field of machine learning has been driven by a revolution in 'big data'.[58] Machine learning offers a uniquely effective tool for parsing massive quantities of information with millions of data points. However, in the field of liquid biopsy and medical diagnostics, dataset sizes are typically much smaller, on the order of tens to hundreds of observations prior to clinical deployment.[59–61] In these cases, machine learning can still be very useful. With 'small data', we can leverage machine learning by focusing on a particular set of algorithms and techniques designed to constrain model complexity and estimate prediction uncertainty.

The number of observations available for model training dictates the number and complexity of potential models that could be applied to the data. With relatively few data points, there are countless methods that could adequately fit the data, but we lack the capacity to discriminate among their expected performance on yet-to-be-collected data. We are thus limited to only simple hypotheses modeled with algorithms such as linear regression or shallow decision trees, rather than models with nonlinearity or higher-order feature interactions, as the simpler models are more likely to have consistent performance on unseen data.[62] Small datasets inherently preclude optimization of complex models with many tunable weights or parameters or those that rely on stochastic gradient descent learning such as neural networks.

Overfitting the training data is particularly worrisome when working with small datasets. In this case, overfitting stems from applying too complex an algorithm for the limited number of training samples, resulting in a model that closely fits the training data but lacks robust performance on new data. The relationship between model complexity and robustness is demonstrated in Fig. 2. The goal is to develop a model that is complex enough to capture the desired signal but not so complex as to capture noisy data or features in the training set. In the case of liquid biopsy, the goal is to model

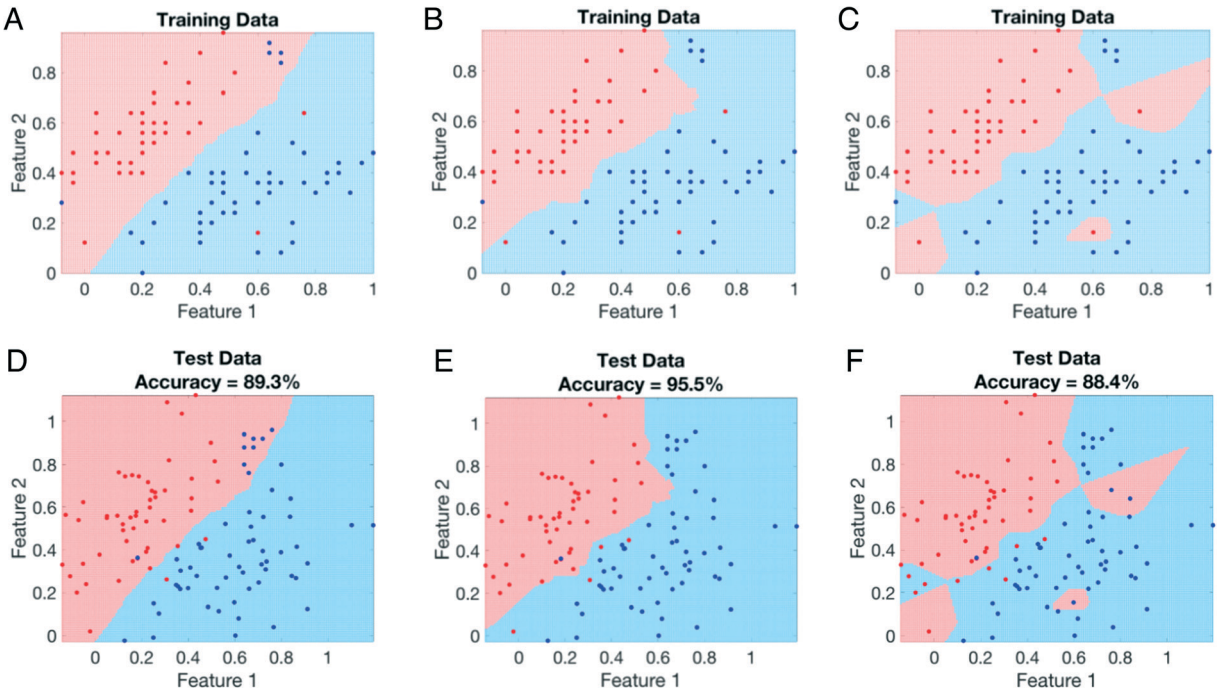

Fig. 2 Model complexity and generalizability. Each panel represents the decision boundaries of a *k*-nearest neighbor classifier with a different level of complexity. A: A low complexity (high bias) model captures the overall trend of the data, but misses some meaningful signal. B: An intermediate complexity model captures the meaningful dataset distribution, and is more likely to perform well on new data. C: A high complexity (high variance) model overfits the dataset. This model fits the training data perfectly, but is not likely to perform well on new data. D–F: A test dataset is drawn from the same distribution as the training data. Superimposed decision boundaries and classification accuracy on the test dataset are shown for the models in A, B, and C, respectively. During model development, model complexity must be optimized to maximize performance on out-of-sample validation data not used for training.

meaningful differences in biomarkers between subjects with and without the target condition, while minimizing modeling effects such as inter-subject noise, inter-measurement noise, or outlier measurements.

One strategy to limit model complexity is to use 'feature selection', a process by which the most useful biomarkers are identified for inclusion in the model. This approach seeks to generate a more robust and generalizable model by eliminating the effects of noisy or irrelevant biomarkers. While feature selection is not always useful or reliable for modern machine learning methods,[71] it is of particularly practical importance in the field of liquid biopsy. For example, many potential biomarkers can be screened during preliminary sequencing studies, but due to practical hardware limitations (*e.g.* cost, simplicity), a much smaller number is used during device implementation. During device design, it is therefore quite meaningful to select the limited set of biomarkers with the best joint performance. It may be possible to reduce the feature space up front using relevant literature and domain expertise, choosing biomarkers with physiologic relevance and eliminating other biomarkers within the same pathway that are likely to be highly correlated. This manually curated feature set can then be further reduced using feature selection methods appropriate for sparse modeling, where the number of features may be significantly greater than the number of observations ($p \gg n$).[72] In this scenario, we recommend applying LASSO (a version of linear regression in which large coefficients are penalized), stepwise regression (in which features are sequentially added to the model if they meet a significance threshold), or LARS (least angle regression; a combination of these two methods)[63] for feature selection. Details on these methods can be found in original papers describing LASSO[64] and LARS[65] or in a recent textbook by the same authors.[66] In some cases, it may be necessary to select a particular number of features to accommodate the diagnostic hardware. This can be accomplished either by tuning the LASSO algorithm such that it produces the desired number of features (at some expense of performance), or by brute force optimization of your classifier by cross validation with all possible feature subsets of the appropriate size. The biomarkers selected by these methods should be taken with the caveat that empiric feature selection with small datasets is inherently noisy; there is no guarantee that these biomarkers have particular biological significance, or that they would consistently be selected if more data were gathered.

During model training, regularization must then be applied to the selected features to further limit model complexity. Regularization refers to a series of techniques designed to prevent overfitting of the training data, such as discouraging large coefficients in a linear model. Appropriate regularization methods for small datasets include L1 (LASSO) or L2 (ridge regression) penalties for linear models, or maximum

depth and pruning requirements for tree-based models. (LASSO is an example of a broader class of regularization methods for feature selection, serving both functions.) To some extent, the choice of regularization strategy is related to insights about the dataset. For instance, L1 regularization can be used if few features are expected to be important, while L2 regularization is more applicable if many features are expected that are weakly predictive. Together, feature selection and regularization produce a simpler, more generalizable model that leverages the most useful biomarkers without overfitting noise in the training data.

Model averaging and bootstrapping

While small datasets limit us to relatively simple models, it is possible to combine many simple models into a single, more powerful algorithm using model averaging.[67,68] Model averaging generates a 'committee' or an 'ensemble' model. If each simple model performs even slightly better than random chance, the average prediction of many such models will lead to a more accurate prediction. One common way to implement model averaging is through 'bagging', in which a series of simple models is generated on randomly sampled, *i.e.* bootstrapped, subsets of the dataset. Each model, or 'weak learner', is sensitive to the noise of its particular training data and has limited predictive accuracy; the average prediction of all the models, however, is much more accurate as this noise will be averaged out across the many data subsets. This approach is often applied to shallow decision trees to generate the popular 'random forest' classifier. Another way to combine multiple models is through Bayesian model averaging.[67] This approach, often applied to linear models, generates a weighted average of model predictions based on each model's level of certainty. Models can also be combined using a slightly different technique known as 'stacking'. In this technique, predictions from multiple models, typically chosen to represent a variety of machine learning methods, are fed into a 'second-level' model that combines this information to generate the final predictions. This second-level model can learn to emphasize each base model where it performs well and discount each base model where it performs poorly. Stacking is especially useful for small datasets since it mitigates model-specific overfitting. So long as each base model overfits the data differently, they can be intelligently combined to cover the feature space. Using stacking, the fear of 'missing out' on the benefit of any particular algorithm is mitigated, as any potentially useful algorithm can simply be added to the stack as another base model. More detail on stacking can be found in *Data Mining* by Witten, *et al.*[58]

An example of ensemble learning is demonstrated in Fig. 3, in which a classification problem is solved using an ensemble of bootstrapped logistic regression models. Each individual logistic regression model is trained on a subset of the data, and produces a linear decision boundary. The decision boundary of the full ensemble is constructed as the median of these individual decision boundaries. When the data are easily separable, the ensemble yields a similar decision boundary as a single model built on all the data (Fig. 3A). However, if a single outlier is introduced, the model trained on all the data is significantly skewed, while the ensemble decision boundary is largely unchanged. (Fig. 3B) This quality is especially important in small datasets, which may be greatly affected by only a few outlying measurements. The distribution of decision boundaries within an ensemble also provides information about the underlying dataset, as greater discrepancies among the models, panel B compared to A, may indicate a higher degree of noise or outlying data.

With small datasets, it is also important to consider inherent statistical limitations during both model optimization and prediction. A small dataset for model training also means a small test set for estimating out-of-sample performance. In this case, it can be easy to overfit the test set with repeated experimentation, negating its value as an independent test set. Small datasets also limit confidence in model parameters. The output of a machine learning algorithm is often reported in terms of a single point estimate, masking a

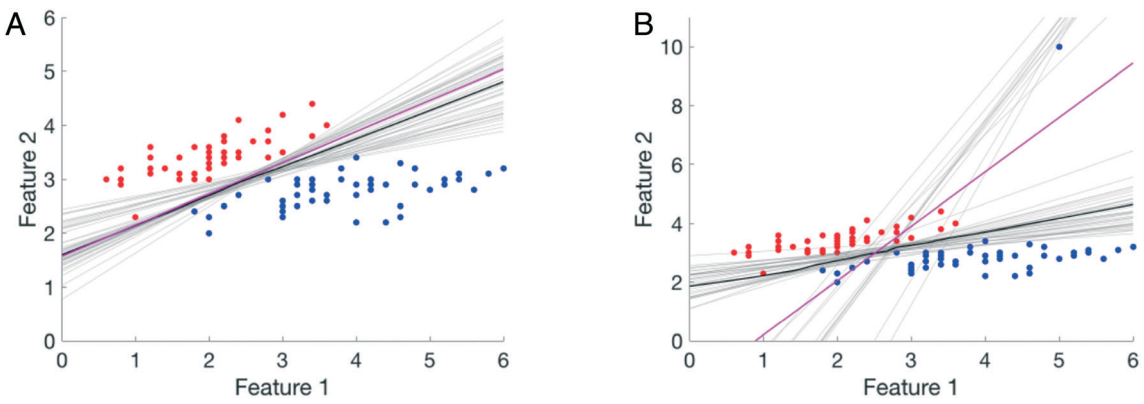

Fig. 3 Using bootstrapping to generate an ensemble classifier. The two-class dataset was modeled with 50 bootstrapped logistic regression models. The decision boundary of each bootstrapped model (*i.e.* the line on which points are equally likely to be red or blue) is represented by a gray line. The median decision boundary of the ensemble is shown in black. The decision boundary of a single logistic regression model built on all of the data is shown in pink. These methods were applied to a dataset without outliers (A) and the same dataset with a single outlier added (B).

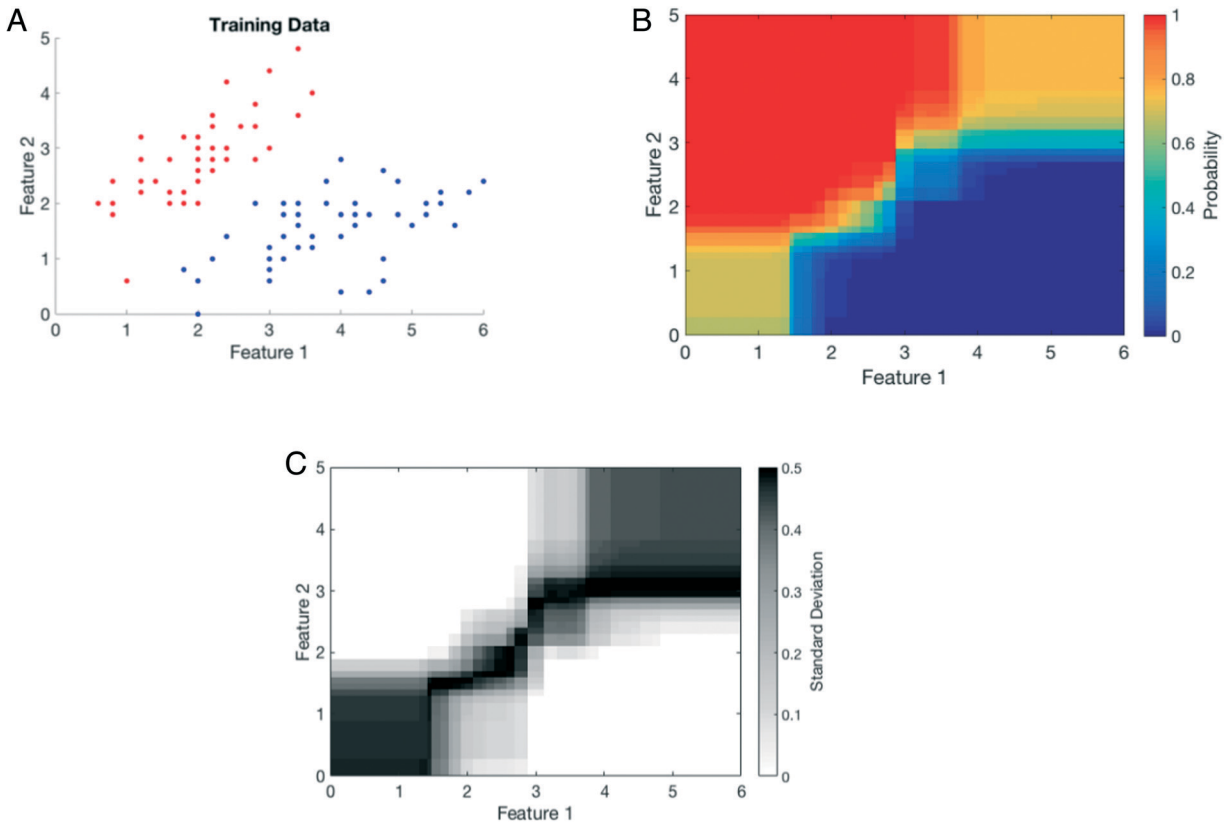

Fig. 4 A. Toy data drawn from two classes (red and blue). A 1000-tree random forest was trained on this dataset for binary classification. B. Heat map showing the ensemble classification score (probability that the test point is red) throughout feature space. C. Heat map showing the standard deviation of the classification score (reflecting classification uncertainty) throughout feature space. Under-sampled areas are associated with higher degrees of classification uncertainty.

potentially high degree of uncertainty that propagates to the model's predictions. One effective way to estimate the uncertainty of model predictions is by analyzing the distribution of predictions from models in a committee. This approach enables generation of approximate confidence intervals, providing insight to overall model performance and allowing identification of areas in which the model performs particularly well or poorly. For example, collective examination of many bootstrapped classification models allows us to compute confidence intervals, or generate receiver operating characteristic (ROC) curves, for classification tasks. An example of this approach is shown in Fig. 4, using a two-class dataset modeled by a random forest classifier with 1000 bootstrapped trees. Each individual tree in the ensemble provides a classification score, in this case the probability that a test point is red, for predicting the label of a given test point. Querying the distribution of classification scores over all trees provides not only an average ensemble classification score (Fig. 4B), but also its uncertainty (Fig. 4C). Therefore, this approach improves on strict binary classification by also providing a level of confidence for the predicted label, further informing decision-making based on test results. Additionally, by comparing the classification uncertainty at different points we can assess which areas in feature space may be under-sampled in the current dataset, leading to predictions with high uncertainty. This methodology is especially valuable in high-dimensionality data, which may be more difficult to visualize than the toy data presented here.

Conclusion

Liquid biopsy based approaches offer many new opportunities to measure molecular biomarkers for the diagnosis, prognosis, and monitoring of disease. Machine learning, and its ability to identify signatures of specific disease states in multiplexed data, will be key to taking advantage of the new molecular information that microchip-based diagnostics can extract. The 'small data' inherent to the development of new liquid biopsy technologies creates challenges in developing machine learning based approaches, but these problems can be overcome through careful study design and thoughtful use of established algorithms. By combining multiple measurements of molecular biomarkers, machine learning has demonstrated that it can improve diagnostic performance relative to a manually chosen biomarker or a set of biomarkers using the same underlying dataset. Furthermore, it is possible in some cases that no individual biomarker has significant predictive value on its own, but that its diagnostic power is derived from the cumulative interpretation of many weak predictors. Machine learning algorithms can evaluate the effects

of many biomarkers simultaneously and can discover higher-order interactions among biomarkers that would not be possible to design manually. The approaches outlined in this paper can also be applied to more conventional diagnostic platforms, including tissue biopsy and imaging techniques.[75–77] Because these technologies are more established platforms that are currently used in the clinic, larger data sets than what is available for liquid biopsies are generally available.

While this paper focused primarily on the 'small data' challenges that are faced by academic labs that develop new diagnostics, companies with budgets many times larger than academic laboratories are entering the field of liquid biopsy and will likely evolve towards 'big data'. These larger data sets, which will include large numbers of patient samples $n > 1000$ and large numbers of features $p > 1000$ (*e.g.* using ultra-broad and ultra-deep sequencing), will enable the use of the emerging, high performance tools such as deep learning.[69]

The trend in liquid biopsy, towards automated microchips that output increasingly large datasets, will be augmented and enhanced by the continued development of machine learning and its continued adoption by those in the liquid biopsy field. Based on these trends, we expect the emergence of a coming next generation of high performance liquid biopsy technologies that will have a significant impact on the improved diagnosis and treatment of patients.

Conflicts of interest

David Issadore is a founder of and holds shares of Chip Diagnostics.

Acknowledgements

Issadore was supported by an American Cancer Society – CEOs Against Cancer – CA Division Research Scholar Grant (RSG-15-227-01-CSM), a grant from The Hartwell Foundation, and NIH R21 5R21CA182336. Litt was supported by NIH 1 UH2 NS095495-01, NIH 5R01NS099348, Mirowski Family Foundation, and Neil and Barbara Smit.

References

1 P. Yager, *et al.*, *Nature*, 2006, **442**, 412–418.
2 J. Ko, E. Carpenter and D. Issadore, *Analyst*, 2016, **141**, 450–460.
3 S. Riethdorf, *et al.*, *Clin. Cancer Res.*, 2007, **13**, 920–928.
4 G. Zheng, F. Patolsky, Y. Cui, W. U. Wang and C. M. Lieber, *Nat. Biotechnol.*, 2005, **23**, 1294–1301.
5 G. Li, *et al.*, *J. Appl. Phys.*, 2003, **93**, 7557–7559.
6 Y. Lu, B. R. Goldsmith, N. J. Kybert and A. T. C. Johnson, *Appl. Phys. Lett.*, 2010, **97**, 083107.
7 D. Issadore, *et al.*, *Sci. Transl. Med.*, 2012, **4**, 141ra92.
8 D. Issadore, *et al.*, *Lab Chip*, 2011, **11**, 2282–2287.
9 D. Agranoff, *et al.*, *Lancet*, 2006, **368**, 1012–1021.
10 S. Banerjee, *et al.*, *Multimodal diagnostic segregation of oral leukoplakia and cancer*, 2016, vol. 4, p. 7.
11 M. Fatima and M. Pasha, *Journal of Intelligent Learning Systems and Applications*, 2017, **9**, 1.
12 L. C. Kenny, *et al.*, *Metabolomics*, 2005, **1**, 227.
13 Y. Kim, *Nat. Commun.*, 2016, **7**, 11906.
14 X. Lai, *et al.*, *Cancer Lett.*, 2017, **393**, 86–93.
15 J. V. Pinto, *et al.*, *Schizophr. Res.*, 2017, 182–184.
16 P. Roth, *et al.*, *J. Neurochem.*, 2011, **118**, 449–457.
17 M. Vidyasagar, *Annu. Rev. Pharmacol. Toxicol.*, 2015, **55**, 15–34.
18 A. Esposito, C. Criscitiello, M. Locatelli, M. Milano and G. Curigliano, *Pharmacol. Ther.*, 2016, **157**, 120–124.
19 R. A. Burrell, N. McGranahan, J. Bartek and C. Swanton, *Nature*, 2013, **501**, 338–345.
20 Q. Fu, F. S. Schoenhoff, W. J. Savage, P. Zhang and J. E. Van, *Proteomics: Clin. Appl.*, 2010, **4**, 271–284.
21 H. Im, *et al.*, *Nat. Biotechnol.*, 2014, **32**, 490–495.
22 J. A. Eastham, *et al.*, *JAMA, J. Am. Med. Assoc.*, 2003, **289**, 2695–2700.
23 S. Sood, *et al.*, *Genome Biol.*, 2015, **16**, 185.
24 V. De Iwijn, *et al.*, *Sci. Transl. Med.*, 2014, **6**, 241ra77.
25 J. Uchida, *et al.*, *Clin. Chem.*, 2015, **61**, 1191–1196.
26 S. K. Arya, C. C. Wong, Y. J. Jeon, T. Bansal and M. K. Park, *Chem. Rev.*, 2015, **115**, 5116–5158.
27 V. R. Yelleswarapu, H.-H. Jeong, S. Yadavali and D. Issadore, *Lab Chip*, 2017, **17**, 1083–1094.
28 E. Samiei, M. Tabrizian and M. Hoorfar, *Lab Chip*, 2016, **16**, 2376–2396.
29 H. Im, K. Lee, R. Weissleder, H. Lee and C. Castro, *Lab Chip*, 2017, 2892–2898.
30 L. G. Biesecker and R. C. Green, *N. Engl. J. Med.*, 2014, **370**, 2418–2425.
31 C. J. Bettinger, *Trends Biotechnol.*, 2015, **33**, 575–585.
32 A. J. Tudos, G. A. J. Besselink and R. B. M. Schasfoort, *Lab Chip*, 2001, **1**, 83–95.
33 S. Dreiseitl, *et al.*, *J. Biomed. Inf.*, 2001, **34**, 28–36.
34 K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis and D. I. Fotiadis, *Comput. Struct. Biotechnol. J.*, 2015, **13**, 8–17.
35 M. W. Libbrecht and W. S. Noble, *Nat. Rev. Genet.*, 2015, **16**, 321–332.
36 S. Singireddy, *et al.*, *Identifying differentially expressed transcripts associated with prostate cancer progression using RNA-Seq and machine learning techniques*, 2015, pp. 1–5.
37 A. L. Swan, A. Mobasheri, D. Allaway, S. Liddell and J. Bacardit, *OMICS*, 2013, **17**, 595–610.
38 L. Wei, Y. Yang, R. M. Nishikawa and Y. Jiang, *IEEE Trans. Med. Imaging*, 2005, **24**, 371–380.
39 E. Crowley, F. Di Nicolantonio, F. Loupakis and A. Bardelli, *Nat. Rev. Clin. Oncol.*, 2013, **10**, 472–484.
40 L. A. Diaz and A. Bardelli, *J. Clin. Oncol.*, 2014, **32**, 579–586.
41 G. Brock, E. Castellanos–Rizaldos, L. Hu, C. Coticchia and J. Skog, *Transl. Cancer Res.*, 2015, **4**, 280–290.
42 L. Manterola, *et al.*, *Neuro-Oncology*, 2014, not218.
43 J. B. Haun, *et al.*, *Sci. Transl. Med.*, 2011, **3**, 71ra16.
44 R. Jacobs, *et al.*, Identification of novel host biomarkers in plasma as candidates for the immunodiagnosis of

tuberculosis disease and monitoring of tuberculosis treatment response, 2016, pp. 57581–57592.
45 G. Lugli, *et al.*, *PLoS One*, 2015, **10**, e0139233.
46 M. G. Best, *et al.*, *Cancer Cell*, 2015, **28**, 666–676.
47 R. Cazzoli, *et al.*, *J. Thorac. Oncol.*, 2013, **8**, 1156–1162.
48 M. Noerholm, *et al.*, *BMC Cancer*, 2012, **12**, 22.
49 M. Nebozhyn, *et al.*, *Blood*, 2006, **107**, 3189–3196.
50 A. A. Ponomaryova, *et al.*, *Lung Cancer*, 2013, **81**, 397–403.
51 K. Honda, *et al.*, *Cancer Res.*, 2005, **65**, 10613–10622.
52 M. J. Lodes, *et al.*, *PLoS One*, 2009, **4**, e6229.
53 A.-E. Saliba, *et al.*, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 14524–14529.
54 H. J. Yoon, *et al.*, *Nat. Nanotechnol.*, 2013, **8**, 735–741.
55 M. T. Deng, *et al.*, *Sci. Rep.*, 2014, **4**, 7261.
56 J. Friedman, T. Hastie and R. Tibshirani, *The elements of statistical learning*, Springer series in statistics New York, 2001.
57 C. M. Bishop, *Pattern recognition and machine learning*, Springer, 2006.
58 I. H. Witten, E. Frank, M. A. Hall and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, 2016.
59 M. A. Mazurowski, *et al.*, *Neural Netw.*, 2008, **21**, 427–436.
60 H.-P. Chan, B. Sahiner and L. Hadjiiski, *Sample size and validation issues on the development of CAD systems*, 2004, vol. 1268, pp. 872–877.
61 A. Onisko, M. J. Druzdzel and H. Wasyluk, *Int. J. Approx. Reason.*, 2001, **27**, 165–182.
62 C. M. Bishop, *Philos. Trans. R. Soc., A*, 2013, **371**, 20120222.
63 V. Stodden, *Breakdown point of model selection when the number of variables exceeds the number of observations*, 2006, pp. 1916–1921.
64 R. Tibshirani, *J. R. Stat. Soc. Series B Stat. Methodol.*, 1996, 267–288.
65 B. Efron, T. Hastie, I. Johnstone, R. Tibshirani and others, *Ann. Stat.*, 2004, **32**, 407–499.
66 T. Hastie, R. Tibshirani and M. Wainwright, *Statistical learning with sparsity: the lasso and generalizations*, CRC press, 2015.
67 J. A. Hoeting, D. Madigan, A. E. Raftery and C. T. Volinsky, *Stat. Sci.*, 1999, 382–401.
68 P. Domingos, *Commun. ACM*, 2012, **55**, 78–87.
69 B. Alipanahi, A. Delong, M. T. Weirauch and B. J. Frey, *Nat. Biotechnol.*, 2015, **33**, 831–838.
70 V. Kremen, *et al.*, *IEEE BioCAS*, 2017, in press.
71 É Perthame, *et al.*, *Stat. Comput.*, 2016, **26**(4), 783–796.
72 L. Wasserman and K. Roeder, *Ann. Stat.*, 2009, **37**(5A), 2178–2201.
73 S. Perakis and M. R. Speicher, *BMC Med.*, 2017, **15**, 75.
74 J. Ko, *et al.*, *ACS Nano*, 2017, DOI: 10.1021/acsnano.7b05503.
75 G. Litjens, *et al.*, *Sci. Rep.*, 2016, **6**, 26286.
76 A. Gertych, *et al.*, *Comput. Med. Imaging Graph.*, 2015, **46**, 197–208.
77 M. Wernick, Y. Yang, J. Brankov, G. Yourganov and S. Strother, *IEEE Signal Process Mag.*, 2010, **27**, 25–38.
78 C. M. Svensson, R. Hübler and M. T. Figge, *J. Immunol. Res.*, 2015, (2015), 573165.
79 C. M. Svensson, C. M. Krusekopf, S. J. Lücke and M. T. Figge, *Cytometry, Part A*, 2014, **85**, 501–511.
80 S. A. Joosse, *et al.*, *Cancer Res.*, 2016, **76**, 14.
81 X. Ma, *et al.*, *PLoS One*, 2017, 12.
82 N. Sol and T. Wurdinger, *Cancer Metastasis Rev.*, 2017, **36**, 263–272.
83 S. Kalantari, *et al.*, *PLoS One*, 2013, 8.
84 Y. S. Ho, *et al.*, *Sci. Rep.*, 2016, **6**, 35110.
85 J. D. M. Rennie, *et al.*, *Proc. Twent. Int. Conf. Mach. Learn.*, 2003, vol. 20, pp. 616–623.
86 D. D. Lewis, *Lecture Notes in Computer Science*, 1998, p. 1398.
87 A. Ng and M. I. Jordan, *Adv. Neural Inf. Process. Syst.*, 2002, **28**, 169–187.
88 K. Beyer, *et al.*, *International Conference on Database Theory*, 1999, pp. 217–235.
89 S. Cessie and J. C. Houwelingen, *Appl. Stat.*, 1992, **41**, 191.
90 A. Liaw and M. Wiener, *R J.*, 2002, **2**, 18–22.
91 R. Díaz-Uriarte and S. Alvarez de Andrés, *BMC Bioinf.*, 2006, **7**, 3.
92 Y. LeCun, *et al.*, *Nature*, 2015, **521**, 436–444.
93 J. Schmidhuber, *Neural Netw.*, 2015, **61**, 85–117.
94 S. Tong and D. Koller, *J. Mach. Learn. Res.*, 2001, 45–66.