Review

For reprint orders, please contact: reprints@futuremedicine.com

Personalized Medicine

Machine learning for biomarker identification in cancer research – developments toward its clinical application

The patterns identified from the systematically collected molecular profiles of patient tumor samples, along with clinical metadata, can assist personalized treatments for effective management of cancer patients with similar molecular subtypes. There is an unmet need to develop computational algorithms for cancer diagnosis, prognosis and therapeutics that can identify complex patterns and help in classifications based on plethora of emerging cancer research outcomes in public domain. Machine learning, a branch of artificial intelligence, holds a great potential for pattern recognition in cryptic cancer datasets, as evident from recent literature survey. In this review, we focus on the current status of machine learning applications in cancer research, highlighting trends and analyzing major achievements, roadblocks and challenges toward its implementation in clinics.

Keywords: cancer biomarkers • genomics • machine learning • personalized medicine • predictive classification models • sequencing

Zeenia Jagga[1] & Dinesh Gupta*[,1]

[1]Bioinformatics Laboratory, Structural & Computational Biology Group, International Centre for Genetic Engineering & Biotechnology (ICGEB), Aruna Asaf Ali Marg, New Delhi 110 067, India

*Author for correspondence:
Tel.: +91 112 674 1358
Fax: +91 112 674 2316
dinesh@icgeb.res.in

Cancer is a worldwide leading cause of morbidity and mortality [1,2]. It is defined as the disease of uncontrolled cell division in any of the cell types or organs in human body, with a capacity to spread to other tissues by invasion and metastasis. Tumor development is accompanied with hallmark deviations from normal cellular functions like – sustained proliferative signaling; evasion for growth suppressors; activation of invasion and metastasis; replicative immortality; induction of angiogenesis; resistance to cell death; deregulation of cellular energetics and evasion of immune destruction [3]. It arises as a cumulative response to somatic alterations at genetic, genomic and epigenetic levels [4]. Canonical cancer diagnosis is based on tumor site and histology, with tumor stage and grade as major prognostic factors. If diagnosed at an early stage, cancer can be treated with better prognosis, using anticancer therapeutics. However, the early-stage diagnosis is often incidental and prognosis as well as therapeutics at later stages is challenging in most cancers [5,6]. Complexity in cancer research is exacerbated due to inter- and intra-tumor heterogeneity. Furthermore, intratumor heterogeneity is complicated by diversity based on tumor genomics, epigenomics and microenvironment [7].

Although all cancer types involve uncontrolled cell division, each cancer patient is unique owing to the heterogeneity at inter-tumor and intra-tumor levels. The concept of personalized medicine takes cognizance of this heterogeneity of genotypic differences between individuals. Personalized medicine is a part of an emerging P4 healthcare concept as 'predictive', 'personalized', 'preventive' and 'participatory' response [8,9]. The advantage of personalized medicine is availability of tailored therapeutic treatments for different molecular subgroups of disease susceptibility, which includes highly specific, low dosage and lesser side effects treatment. Remarkable advances in high throughput data generation technologies have opened up newer avenues in cancer research, providing opportunity to catalogue diverse molecular alterations in the tumors. Genomics-based diagnosis, prognosis and therapeutic implementations in

Future Medicine part of fsg

clinics would be the key to precision medicine for cancer [10]. Altogether this suggests there is an urgent need to identify cryptic patterns of molecular features which can aid early screening and prognosis of cancer patients using pattern recognition and classification techniques like machine learning [11].

Here, we review the current understanding of machine learning and its implementation in cancer research. We review the plethora of available 'omics' data resources, which can be exploited for classification and biomarker identification for cancer.

Omics data in cancer

With the completion of human genome in 2003 and availability of next generation sequencing technologies, there have been significant changes in our knowledge and approach to biomedical research. This also facilitated the molecular profiling of patients tumor [12,13] leading to generation of zeta-bytes of data, also known as 'Big data' [14]. The challenges for data storage, analysis, reproducibility, access and interpretation are coincidental with the emergence of 'Big data' [15]. However, development of crowd analysis, crowd computing and crowd sourcing forge a formidable adaptation to 'Big data' [15,16]. These technological advances are enabling us to integrate all the levels of molecular probing, also referred as 'Panomics'. The molecular levels being probed includes genomics (SNP and copy number alteration by DNA sequencing); transcriptomics (gene expression by RNASeq, microarrays and RT-PCR; miRNA using small RNA sequencing, SAGE, microarray and RT-PCR); proteomics (protein expression by mass spectrometry, antibody microarray and cancer autoantibody); metabolomics (by mass spectrometry, nuclear magnetic resonance and HPLC) [17]; glycomics; lipidomics and so on. Fortunately, most of the high-throughput sequence data are also available in public domain for the cancer research community to analyze and develop predictive computational models.

Data resources availability

Despite technological advancements in sequencing technologies, the barriers to clinical translation of molecular diagnostics have been choice of assay/design, cost, tissue quality, tumor content, analytical validity, clinical laboratory improvements amendment (CLIA) certification, turnaround time and bioinformatics analysis [18,19]. With intent to overcome these challenges and systematically collect molecular data from cancer patients, large-scale international collaborative projects have been established. Public databases in cancer research field deploy foundation to transdisciplinary collaborations for alleviating cancer informatics research. The major cancer dataset resources (summarized in Table 1) include database with curated molecular alterations like Catalog of Somatic Mutation In Cancer (COSMIC); databases of raw data files like Gene Expression Omnibus (GEO), Short Read Archive (NCBI-SRA); international consortiums like The Cancer Genome Atlas (TCGA), International Cancer Genome Consortium (ICGC) and Worldwide Innovative Networking in personalized cancer medicine (WIN) ignites networking between investigators, cancer trials and cancer centers. The projects like Dialogue for Reverse Engineering Assessments and Methods (DREAM) help in bridging experiment with theory by challenging the research community with real data for predictive modeling and inferring cellular networks [20,21].

These large scale initiatives provide common standards at international scale, being used across the consortiums to minimize the redundancy and generate valuable data of cataloged molecular alterations. These comprehensive systematic programs help fostering networks for interdisciplinary research efforts and also reduce the problems of recruiting patients. The next major challenge is to identify biomarkers and develop prediction models for diagnosis, prognosis or targeted therapy. Apart from the systematic cataloging of molecular alterations, research efforts to understand basic tumor biology cannot be undermined. The experiments to determine metastatic proclivity or response to different drugs are indeed consequential to the computational methods. It would require computational biologists to take up challenges for developing algorithms which can work on assumptions of biological systems to generate clinical grade interpretations from the data. Although, the clinical computational biology is still at its infancy, the currently available cancer research resources and computational approaches can lay the foundation for leading computational biology into clinics.

Machine learning

Machine learning (ML), a branch of artificial intelligence, is the panacea for recognizing patterns of cryptic datasets from disparate fields [22]. Since diverse and vast amount of cancer datasets are already in public domain, machine learning has proven potential to address mining cancer-specific patterns in cancer prediction and prognosis [23]; in cancer diagnosis, management [24] and in personalized medicine [25]. Here, we review the applications of machine learning in cancer research along with glimpses of recent developments in the field.

Key terminologies and also the prerequisites for machine learning are: (A) dataset (training and testing); (B) instance; (C) attribute; (D) class label and

Table 1. List of major cancer data project and resources.

Project	Description	Ref.
WIN	Worldwide Innovative Networking in personalized cancer medicine is a networking platform for global collaborations for clinical trials	[65,66]
ICGC	International Cancer Genome Consortium has committed to 73 cancer projects till date with the aim to provide a level controlled comprehensive access to genomic, transcriptomic and epigenomic cancer datasets	[67,68]
TCGA	The Cancer Genome Atlas provides a level controlled comprehensive access to large scale genome analysis datasets of various molecular levels from tumor and tumor matched tissues of various cancer types	[67,69]
DREAM	Dialogue for Reverse Engineering Assessments and Methods is powered by Sage Bionetworks help bridge the gap between experiment and theory. DREAM challenges the research community for cellular network inferences and quantitative model building giving access to real time and simulated datasets	[70]
cBioPortal	cBioportal for Cancer Genomics is a platform for visualization, analysis and downloading high-throughput cancer genomics datasets	[71–73]
NCBI-SRA	Short Read Archive hosts raw data and alignment information from various biological experiments	[74]
GEO	Gene Expression Omnibus is a repository for functional genomics datasets. It also provides tools to query download and curate gene expression profiles	[75]
COSMIC	Catalogue of Somatic Mutations in cancer is a database that compiles somatic gene alterations in cancer genes	[76]
CCLE	Cancer Cell Line Encyclopedia provides access to genomic data from genetic and pharmacological perturbations to human cancer models (approximately 1000 cell lines). Tools for integrated computational analysis and visualization are also available	[77]

This list is not comprehensive but includes the major cancer data resources. The description is based on the access to the mentioned URLs as on 9 December 2014.

(E) cross-validation. Knowledge of the above mentioned terms enables one to understand the methods behind development of machine learning prediction or classification models, a brief description of each is as follows: (A) Dataset is a matrix of instances versus attributes. Training dataset is the dataset subset used to train the learning scheme for developing a classifier. Testing dataset is the subset of the dataset which is not a part of training and helps estimate the error of a classifier to analyze classifier performance on 'unseen' dataset. (B) Instance (also known as case or record or example) is a number of cases in the study used for training or testing the model. With respect to cancer investigations, instances are number of tumor samples or patient samples which have been used in the study. (C) Attribute (also known as feature, variable or field) is a prefixed numeric or categorical quantity describing an instance. Vector of feature set is called as feature vector. A set of feature vectors representing a case is called as a feature space. In relation to cancer biology this could be imaging data, genomic data (copy number alternation and single nucleotide polymorphism), transcriptomics data (mRNA expression and miRNA expression), proteomics, lipidomics, metabolomics data, clinical parameters, which can be used to describe clinical phenotype of the patient. (D) Class label is the class category description of an instance. (E) Cross-validation (CV) is a statistical procedure to estimate errors and generalization capabilities of a classifier. Various CV methods are being employed like k-fold CV, leave one out (LOO), hold-out, bootstrapping and so on. In k-fold CV, the dataset is divided onto k-mutually exclusive subsets now the classifier is trained on k-1 subsets and tested on the 1 subset till each subset has been used as testing dataset. If k is equal to number of instances of the dataset it is called as LOO-CV procedure. Average accuracy of k-folds is the accuracy estimate of the classifier. In bootstrap validation, subsampling is performed with equal replacement from training dataset.

In literature, ML algorithms are generally classified as supervised, unsupervised and semisupervised, based on association of class labels with training instances. Supervised machine-learning algorithms are employed mainly for classification or regression problems where the class label of patient sample is already available.

Unsupervised machine-learning algorithms are implemented for clustering, density estimation or dimensionality reduction and class label information of patient sample is not available. Semisupervised is a class of supervised machine-learning algorithms which can readily utilize both labeled and unlabeled datasets [26]. The models/classifiers generated by machine-learning algorithm can be standalone executable systems to predict clinical phenotype of new instances or patients in clinical decision support. A simplified machine learning workflow for developing supervised machine learning models for cancer diagnostics is illustrated in Figure 1.

Feature selection algorithms

Feature selection is a technique to reduce and optimize feature space dimensionality and determine important features for generating better classifier performance, reducing computational cost. Mining omics data comes with 'curse of dimensionality' where number of instances is always far less than the number of attributes. This problem leads to overfitting of classification models as machine learning techniques for pattern recognition were not initially designed for handling larger number of irrelevant features in the training dataset. In relation to cancer research investigations, feature selection is a solution to determine cancer biomarkers which might as well lead to insightful clues for cancer pathogenesis or progression. Feature selection could also be implemented to decipher molecular signature distinguishing different types of cancer relevant class labels.

Feature selection methods are classified as – filter methods, wrapper methods and embedded methods [27]. Filter methods are independent of the classification algorithms being employed and assess only the intrinsic properties of data. Wrapper methods evaluate learning models with selected feature subset space. Although this accounts for feature dependencies but also increases the risk of overfitting and makes these methods computationally expensive especially for omics grade data. Embedded methods assess optimal feature subset by building feature selection within classification algorithm. This makes the process computationally less expensive as compared with filter methods while retaining the interaction with classification model.

For further details readers may refer to following articles: mathematical explanation and availability of

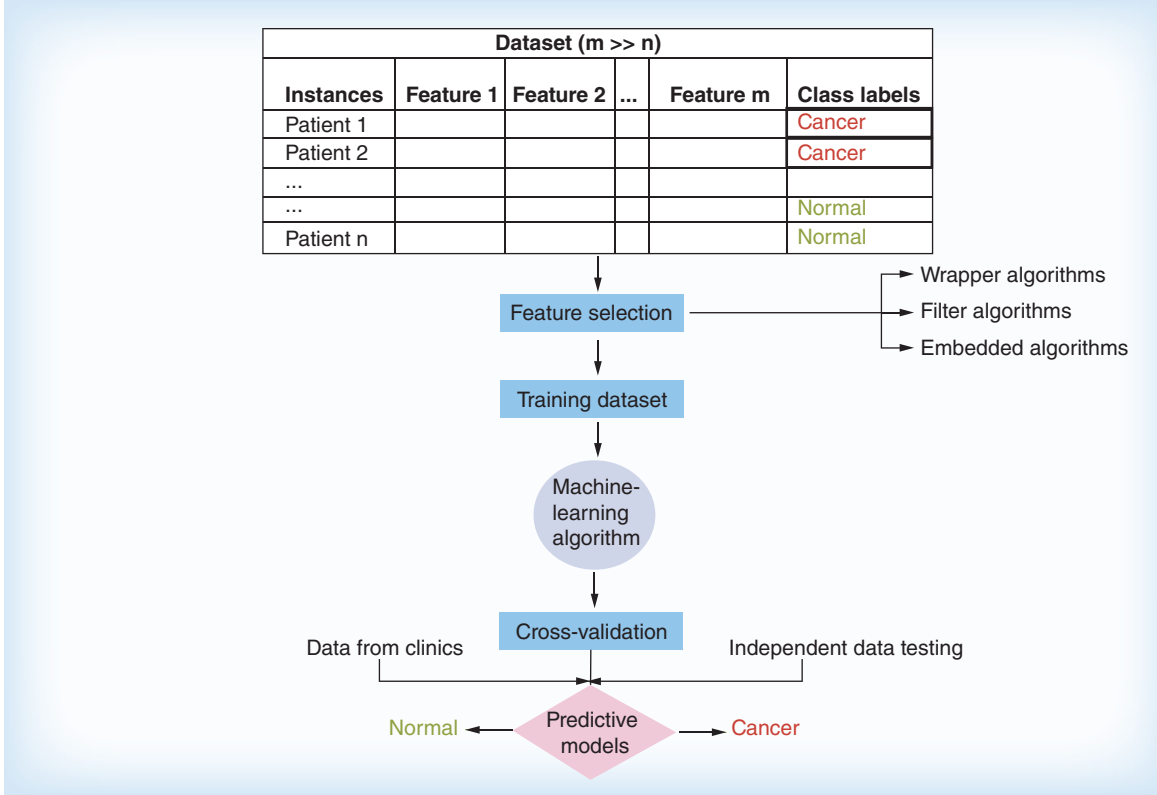

Figure 1. A simplified workflow illustrating supervised machine-learning-based predictive classification models for cancer diagnostics. Dataset is a matrix of n rows (patients/instances) and m columns (features/attributes) with class labels (clinical phenotype). In case of features from 'omics' data m is always larger than n. Feature selection is performed to reduce the data dimensionality. The machine-learning algorithm learns from reduced training dataset. Classification models thus generated are cross-validated, tested on independent datasets. And, the best classification models can aid clinical decisions for cancer diagnostics.

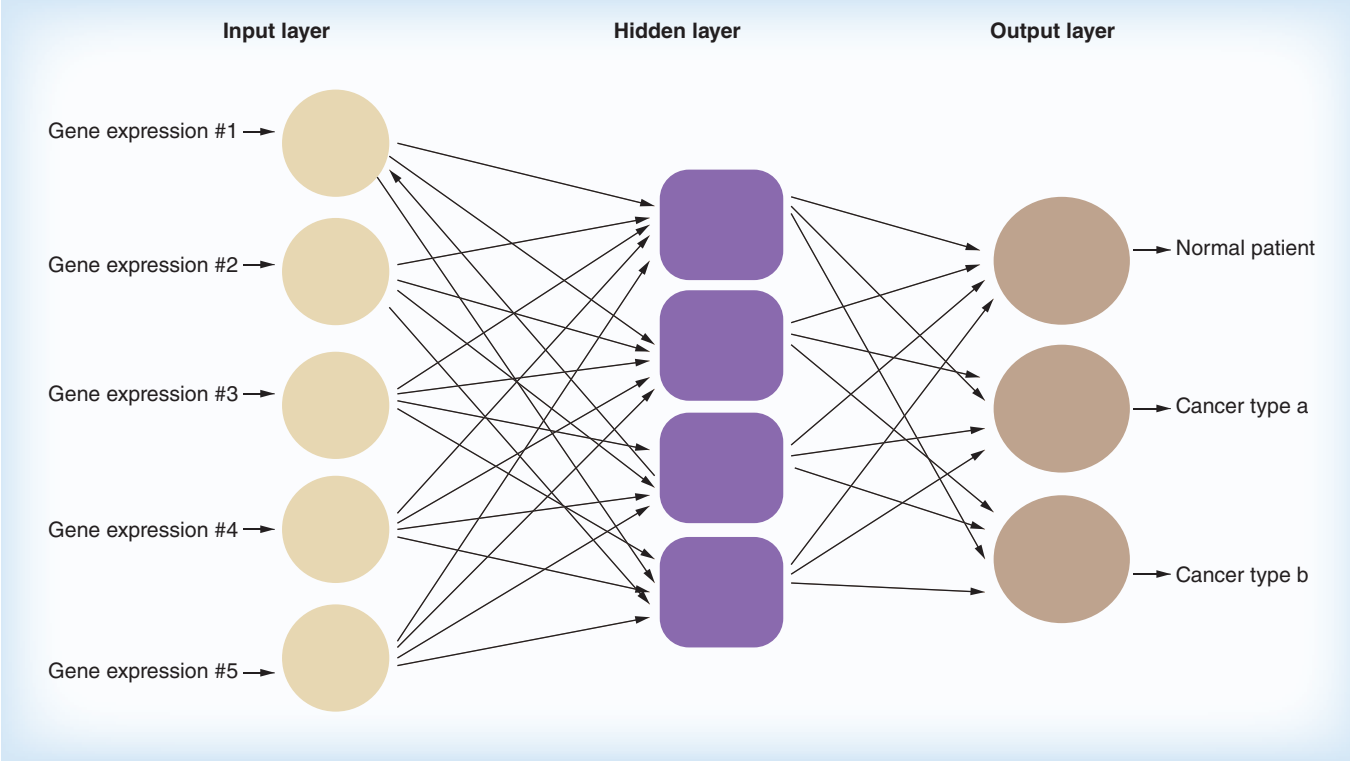

Figure 2. Representation of three layers of artificial neural networks – input, hidden and output layer. In this example, input layer has molecular signature of five gene-expression values which learns in hidden layer and classifies the instances into either one of the three classes in output layer as normal patient, cancer type a, cancer type b profile.

open source feature selection tools [27], feature selection methods in proteomics [28] in cancer diagnosis [29].

Classification algorithms

There are several ML classification algorithms, which have been successfully used for classification problems in cancer research. However, for the sake brevity, discussing all of them is not within the scope of this review. We highlight here the artificial neural network, support vector machine, decision trees and random forest algorithms. These four algorithms have proven to be quite efficient in cancer dataset classifications, as evident from the literature.

Artificial neural networks (ANN) – ANN algorithm is based on statistical principle of empirical risk minimization. This nonparametric algorithm is motivated from biological neural networks and belong to back propagation class of neural networks [30]. ANN has three layers – input, hidden and output layer. Associate weight coefficient with input of each neuron imitates regulation of neural connection in humans. Figure 2 illustrates schematics of ANN-based clinical decision tool classifying normal patients/cancer subtype a/cancer subtype b from gene-expression signature. The advantages like tolerance to noisy inputs and more than one output for instances support implementation of ANN since mid-1990 till date in cancer biology, for development of clinical decision making tools [31–35]. The major limitations of the algorithm are increased computational cost on a complex dataset, occasional overfitting to the datasets and noninterpretability of the predictive models (often referred as 'Black box') [36].

Support vector machine (SVM) – Developed by Vapnik, SVM algorithm is based on structural risk minimization principle. Algorithm supports classification as well as regression problems utilizing multiple continuous and nominal variables. Figure 3 illustrates linearly separable two-class dataset based on SVM algorithm. To classify complex real life datasets like cancer datasets, SVM transforms input space to high dimensional nonlinear feature space. A maximum margin hyperplane classifies data points of different class labels in a multidimensional feature space [37,38]. Since nonlinear transformation is not explicitly performed, kernel functions implement the same effect without expanding feature space. Choice of kernel, kernel parameters and input feature subset can dramatically impact performance of the algorithm. Availability of nonlinear kernel function is particularly suitable for classifying cancer biology data. The advantage of SVM classifiers is its robustness to noisy datasets and offering optimum solution for the quadratic problem [23,39]. The limitation of using SVM is its interpretation, computational cost for larger datasets and SVM essentially being a binary classifier (multi-

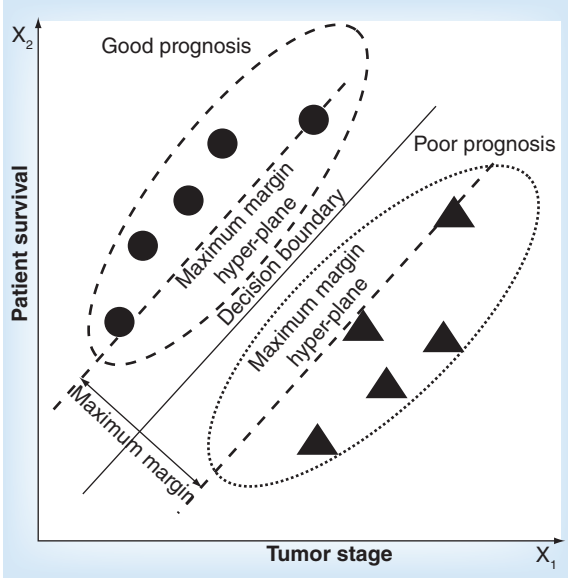

Figure 3. Representation of linear classification by support vector machines algorithm. In this example, prognosis of cancer patient is modeled based in tumor stage and patient survival. The instances close to maximum margin hyperplane are known as support vectors.

class classifications are pairwise implementations of one class against all classes) [23,39].

Decision trees – The decision tree algorithms are based on 'divide and conquer' approach of learning from instances and leading to solution of the problem [40]. Figure 4 represents simplified decision tree for a multiclass decision problem using four attributes. Each node signifies testing an attribute and each new instance following from root to leaf node through the branch based on its attribute values [41]. The order of training instances does not affect training whereas order of attributes selected for final decision tree impacts the training of the classifier [23]. Although decision trees are not computationally complex, however the computational cost increases for high dimensional datasets. Easy interpretability of decision tree models from the perspective of knowledge discovery and input from experts are particularly suitable for mining cancer data.

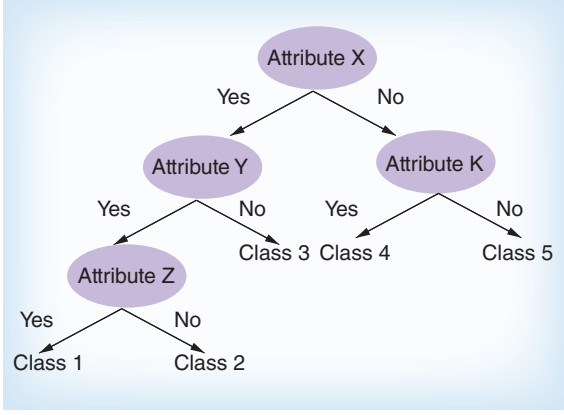

Figure 4. A decision tree structure classifying instances into five classes, based on four attributes. In this example, attribute X is root node; attribute K, Y, Z are internal nodes; classes 1, 2, 3, 4, 5 are leaf nodes and arrows represent branches.

Random forest – Based on ensemble learning, the random forest algorithm was proposed by Breiman in 2001. It is generated by random bootstrap sampling of independent decision trees [42]. The algorithm is suitable for datasets with continuous, categorical, binary data and also missing values. Parameters for optimization are number of trees and features used for splitting each tree. The adaptability of random forests to high dimensional data like – insensitivity to irrelevant features determining correlations, incorporating interactions among the features, little requirement for parameter optimization and applicability to both binary as well as multiclass classification [43,44] – makes it an attractive choice for mining for cancer datasets. The accurate predictions and interpretability of model makes it stand out among other machine learning classification algorithms. Recently we reported that random forest prediction models based on gene expression of clear cell renal carcinoma tumor samples, performed the best for distinguishing early-stage (TNM stage I, II) and late-stage (TNM stage III, IV) tumor samples [45]. Despite all the advantages, random forest should be carefully implemented considering data complexity [46].

Survey of research articles

A systematic literature survey was performed to explore recent trends ML algorithms in cancer research in context with cancer type, study type, data type and data source. Query search for 'machine learning' AND 'cancer ciomarkers' was performed (searched in July, 2014) in electronic literature databases namely Google scholar [47], PubMed Central [48] and PubMed [49]. The query resulted in 819 hits in Google scholar, 134 hits in PubMed Central and five hits in PubMed database. A quick analysis of the outputs revealed the requirement for manual inspection of the full text articles to get into specific details. Hence we proceeded with scrutinization of the 134 hits obtained from PubMed Central, as it gave access to full-text archive of scientific literature. Year-wise trend of the published articles demonstrated a steady increase in the number of published articles (Figure 5A). It was observed that studies pertaining to the application of machine learning for cancer biomarkers research have significantly increased in the recent years.

All the identified articles were downloaded in full text. Based on title and abstract, articles were divided into 4 categories –'relevant research articles,' review

articles, newer methodology articles and not relevant articles. The articles using cell line and animal model datasets were further excluded from the relevant research article category. Finally, 'Relevant research articles' category using machines learning for cancer biomarkers had 28 research articles. Although we understand that we have not achieved complete literature coverage, we believe that our approach has picked up significant and relevant articles for the review.

Our analysis of the identified ML algorithms literature with respect to cancer type, study type, data type and data source categorized on the basis of biological

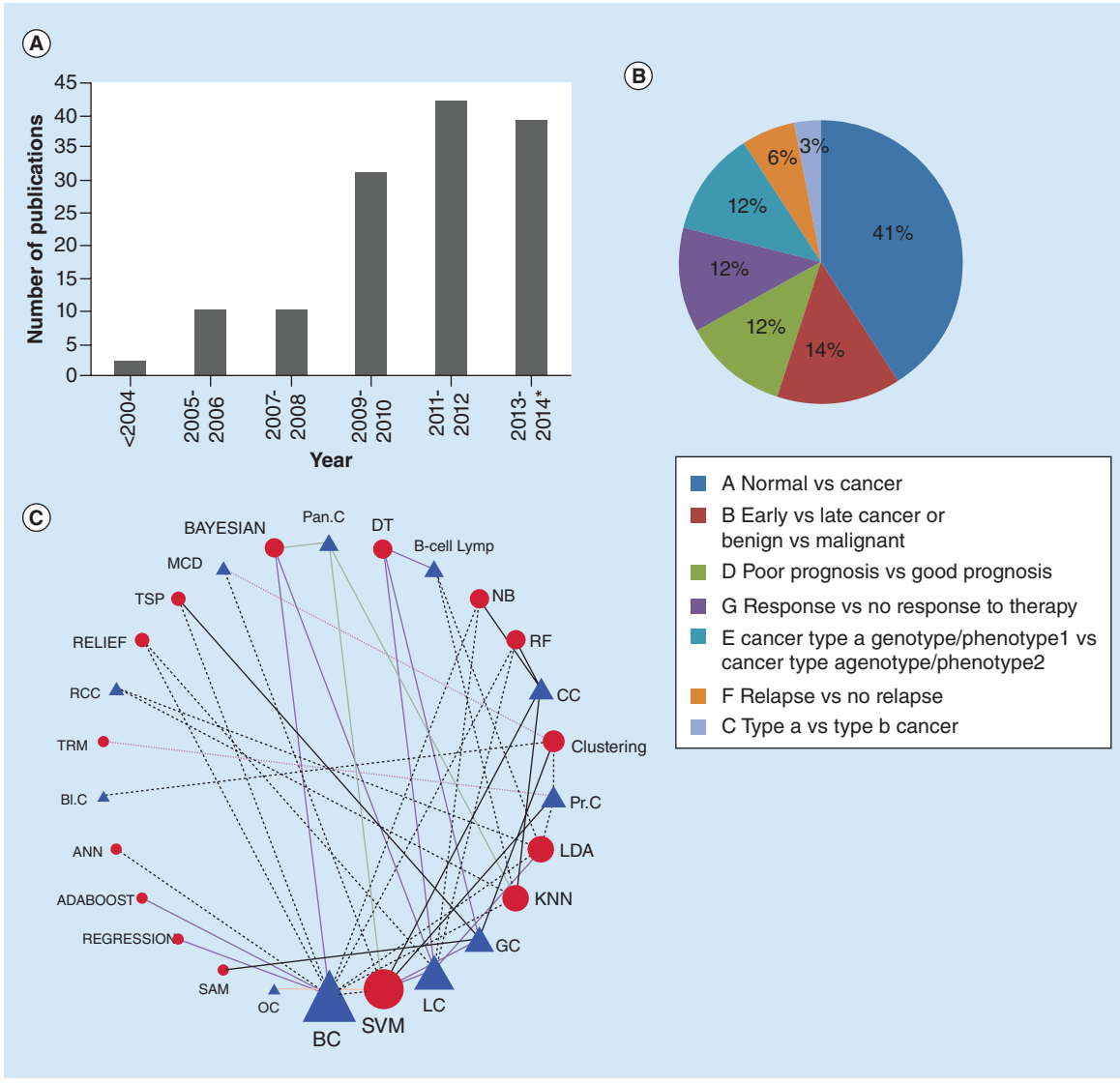

Figure 5. Trends in surveyed articles. **(A)** Year-wise publication profile. The histogram showing steady increase in published literature for machine learning methods used to predict cancer biomarkers. The earliest paper appeared in early 2000s. **(B)** Pie-chart representation for study types. **(C)** Network representation for cancer type, data type, data source and machine-learning algorithms. Nodes represent cancer types in blue triangles and machine-learning algorithm in red circles. For edges, dotted lines solid lines represent data from public domain and laboratory generated, respectively. Colors of edges refer to clinical data (green), transcriptomics (black), proteomics (purple), genomics (magenta) and others (orange). Node size and edge line width are proportional to degree of connectivity.
Blue triangle is BC: Breast cancer; B-cell lymp.: B-cell lymphoma; Bl.C: Bladder cancer; CC: Colorectal cancer; GC: Gastric cancer; LC: Lung cancer; MC: Multiple cancer datasets; OC: Ovarian cancer; Pan.C: Pancreatic cancer; Pr.C: Prostate cancer; RCC: Renal cell carcinoma. Red circle denotes ANN: Artificial neural network; ADABOOST: Adaptive boosting; DT: Decision tree; KNN: K-nearest neighbour; LDA: Linear discriminant analysis; NB: Naive bayes; RF: Random forest; SAM: Significant analysis of microarray; SVM: Support vector machine; TRM: Two-stage Random forest plus multivariate adaptive regression splines; TSP: Top scoring pair.
For color figures please see online at: www.futuremedicine.com/doi/full/10.2217/PME.15.5

Table 2. Summary of investigations surveyed for machine-learning methods used cancer biomarkers studies showing various cancer types, study types, data types, data sources and machine-learning algorithms categorized based on biological sample – tissue, serum or plasma and others.

Cancer	Study type	Biological sample	Data type	Data source	Machine-learning algorithm	Ref.
BC	D	Tissue	Gene expression; clinical information	Microarray†	I-RELIEF algorithm; linear discriminant analysis	[78]
BC	E	Tissue	Gene expression	Microarray†	Top scoring pair; top scoring triplet; naive bayes; k-nearest neighbor; support vector machine; random forest	[79]
BC	E	Tissue	Gene expression	Array express†	Artificial neural network	[80]
BC	G	Tissue	Proteomics peaks	SELDI-TOF-MS‡	Support vector machine - recursive feature elimination; Adaboost	[81]
CC	A	Tissue	Gene expression	Microarray‡	Correlation based feature selection; support vector machine; naive Bayes, k-nearest neighbor	[82]
CC	G	Tissue	Gene expression	Microarray‡	Random forest	[83]
LC	D	Tissue	Gene expression	Director's challenge microarray dataset form caArray†	Univariate cox model; random forest; relief algorithm	[84]
LC	D;G	Tissue	Gene expression	RT PCR‡; microarray†	RELIEF algorithm; naive Bayes	[85]
MCD	A	Tissue -histochemical stained tissue images	Texture feature of images for protein subcellular location	Human protein atlas†	Support vector machine	[86]
MCD	A;B	Tissue	Gene expression-expressed sequence tags	Gene-expression omnibus†; broad institute website†; Tmm database at Columbia University†	Support vector machine	[87]
Pr.C	A	Tissue	mRNA expression; alternatively spliced mRNA isoforms expression	DASL assay‡; array hybridization‡	Support vector machine – recursive feature elimination	[88]
Pr.C	B	Tissue	Gene expression; proteomics	Microarray‡; proteomics data†	Hierarchical agglomerative clustering; linear discriminant analysis	[89]
Pr.C	B	Tissue	Single nucleotide polymorphism	Cancer genetic markers of susceptibility dataset†; Mofitt data†	TRM approach (two-stage random forest plus multivariate adaptive regression splines)	[90]

†Public domain.
‡Laboatory generated.
B-cell Lymp.: B-cell lymphoma; BC: Breast cancer; CC: Colorectal cancer; GC: Gastric cancer; LC: Lung cancer; MC: Multiple cancer dataset; NA: Not available; NV: Not Valid; OC: Ovarian cancer; Pan.C: Pancreatic cancer; Pr.C: Prostate cancer; RCC: Renal cell carcinoma.

Table 2. Summary of investigations surveyed for machine-learning methods used cancer biomarkers studies showing various cancer types, study types, data types, data sources and machine-learning algorithms categorized based on biological sample – tissue, serum or plasma and others (cont.).

Cancer	Study type	Biological sample	Data type	Data source	Machine-learning algorithm	Ref.
RCC	G	Tissue	miRNA	Taq Man Low-densityArrays‡	Logistic regression	[91]
BC	A;B	Serum	Protein expression	ELISA Luminex assay‡	Iterated Bayesian model averaging; support vector machine - recursive feature elimination; least angle regression	[92]
B-cell Lymp.	A;D;F	Serum	Proteomics-peaks	SELDI-TOF-MS‡	Decision tree classification	[93]
GC	A	Serum	Proteomics-peaks	MALDI-TOF‡; immno-based assays‡	Support vector machine; decision trees	[94]
GC, Pr.C	A; C	Serum	Reactomics-chromatic data	Colorimetric reactome‡	Support vector machine	[95]
LC	A	Serum	Auto-antibody expression	Luminex platform using HaloTag technology to detect autoantibodies‡	Compound covariate predictor; diagonal linear discriminant analysis; Bayesian compound covariate predictor	[96]
LC	A	Serum	Protein expression	Multiplexed immunoassay panel‡	Rule learning approach	[97]
OC	B	Serum	Phospholipid levels	Liquid chromatography/ electrospray ionization mass spectrometry‡	Hybrid Huberized support vector machine; support vector machine	[98]
BC	E	Plasma	Proteomics	SELDI TOF-MS‡	Logistic regression model; random forest	[99]
LC	A	Plasma	Monoclonal antibody proteomics	ELISA‡	Support vector machine; sequential minimal optimization	[100]
MCD	E	Others – NV	Single nucleotide polymorphism; genomic region; cancer type	Text mining of published meta-analysis literature†	Clustering	[101]
Pan.C	A	Others – NV	Clinical risk factors	PubMed knowledge†; electronic health record via text mining†	Bayesian network inference; support vector machine; k- nearest neighbor	[102]

†Public domain.
‡Laboatory generated.
B-cell Lymp.: B-cell lymphoma; BC: Breast cancer; CC: Colorectal cancer; GC: Gastric cancer; LC: Lung cancer; MC: Multiple cancer dataset; NA: Not available; NV: Not Valid; OC: Ovarian cancer; Pan.C: Pancreatic cancer; Pr.C: Prostate cancer; RCC: Renal cell carcinoma.

Table 2. Summary of investigations surveyed for machine-learning methods used cancer biomarkers studies showing various cancer types, study types, data types, data sources and machine-learning algorithms categorized based on biological sample – tissue, serum or plasma and others (cont.).

Cancer	Study type	Biological sample	Data type	Data source	Machine-learning algorithm	Ref.
BC, B-cell Lymp.,RCC	F	Others – NA	Gene expression	Gene-expression omnibus†; stanford microarray database†	Linear discriminant analysis; k-nearest neighbor	[103]
GC	A	Others-gastric mucosa	miRNA	miRNA microarray‡; quantitative PCR‡	Significant analysis of microarray; hierarchical clustering by average linkage algorithm; top scoring pair algorithm	[104]
Bl.C	A	Others-bladder wash-exfoliated urothelia	Gene expression	Microarray‡; gene-expression omnibus†	Hierarchical clustering; supervised learning algorithm	[105]

†Public domain.
‡Laboatory generated.
B-cell Lymp.: B-cell lymphoma; BC: Breast cancer; CC: Colorectal cancer; GC: Gastric cancer; LC: Lung cancer; MC: Multiple cancer dataset; NA: Not available; NV: Not Valid; OC: Ovarian cancer; Pan.C: Pancreatic cancer; Pr.C: Prostate cancer; RCC: Renal cell carcinoma.

sample types is summarized in Table 2. The study types are categorized into seven study types A–H, where type A is 'normal versus cancer', B is 'early versus late cancer' or 'benign versus malignant'; C is 'type a versus type b cancer', D is 'poor prognosis versus good prognosis', E is 'cancer type a genotype/phenotype1 versus cancer type a genotype/phenotype2', F is 'relapse versus no relapse', G is 'response versus no response' to therapy and H is continuous response to drug (Table 3). The majority of publications was from study type A, in other words, focus on diagnostic ML models (Figure 5B).

Among 'relevant research articles' most studied cancer type and prediction algorithm was breast cancer and SVM, respectively (Figure 5C). This is in concordance with high global burden of breast cancer as well as indicates SVM as the most commonly implemented algorithm for classification problems in cancer datasets. The analyzed data were either from public domain or generated by the investigators themselves or both. Apart from molecular features, other data types observed were electronic health records and text mining of scientific literature.

Biological samples being investigated were tumor tissues, serum, plasma, urine and so on. The studies were also categorized based on biological samples serum or plasma, tissue samples and others. In the category of serum or plasma samples (nine out of 28 studies, Table 2), a strong inclination (five out of nine studies) was observed for the study type A. The most implemented ML algorithm across different cancer types was SVM. Interestingly, all the datasets were either lab generated proteomics, phospholipids or auto antibody profiles. Studies from tissue samples (14 out of 28 studies, Table 2) and others category (five out of 28 studies, Table 2) relied majorly on transcriptomics or text mining data. Apart from in serum or plasma category, we could not observe any pattern for machine learning or feature selection algorithms implemented on any particular data source or data type.

It was also noted that ML study designs were highly variable in terms of cross-validation procedures, data reporting methods and parameters to compare the classifiers (Supplementary Table 1; see online at www.futuremedicine.com/doi/full/10.2217/PME.15.5). We believe that the lack of data reporting standards as a serious road-block for clinical implementation of ML models. Efforts are needed to lay out the minimal standard guidelines for any ML technique applied to cancer datasets.

Guides for a good machine learning study

Application of ML in cancer research is not new, but its development is following a circuitous path because of nonavailability of quality clinical grade models

Table 3. Study type themes of developing classification and biomarkers for cancer by employing machine learning.

Study type code	Study type detail
A	Normal vs cancer
B	Early cancer vs late cancer; benign vs malignant
C	Cancer type A vs cancer type B
D	Good prognosis vs poor prognosis
E	Cancer type A genotype 1 vs cancer type A genotype 2
F	Relapse vs no relapse
G	Response to therapy vs no response to therapy
H	Prediction of continuous response to drug

that have high sensitivity, specificity and broader implementation. A good ML study should be carefully designed, reproducible, cross comparable, widely implemented and accessible to public domain for testing as well improvements for wider translational implications. Aspects requiring careful consideration for further improvements include good quality data from a robust laboratory technique, reproducible data from different laboratories, dataset size, cross-validation procedures and standard reporting methods.

Although implementation of machine learning is a mature field but adaptations are required based on recent improvements in cancer dataset generation. Due to lack of reporting standards, comparing studies or merging data from various studies becomes a daunting task. A crucial aspect of research is comparison of a proposed approach with the current state of the art. Thereby authors need to be encouraged to record and provide data-cleaning and preprocessing steps in the published literature along with the scripts.

Perhaps, a formal repository that sets minimal guidelines for machine learning investigation with potential clinical implementation could be developed. This repository should emphasize the precise specifications of clinical machine learning tasks and hence motivate the ML community by providing a platform for publishing, exchanging/collection of data sets, benchmarking the statistical evaluators and methods for challenging machine learning problems. Submission of optimized parameters, developed classifiers along with the complete schema and scripts for any particular machine learning study in cancer biology should be emphasized. We believe that this would help to discover biomarkers by attaining classification consensus.

Machine learning may not be the only solution to translate computational algorithms to clinics, however it holds compelling potential. Furthermore, its standard application would help in accelerating translation of ML models into clinical tools for cancer patient management.

Recent improvements in machine-learning algorithms for cancer studies

ML methods were not initially developed for high-throughput biological datasets. Thereby, need of the hour is to develop ML algorithms with assumptions relevant to biological or cancer context (Table 4). Apart from canonical molecular features, secondary hierarchal features like pathway features [50,51] and network based features [52] are reported as stable features for developing classification models.

Newer algorithms for feature selection and classification based on relevant biological assumptions are also being developed. Ren *et al.* have developed ellipsoid Feature Net tool and iPCC for feature selection from gene expression for developing cancer biomarkers [53,54]. The algorithm is based on assumptions relevant to cancer sample heterogeneity. The method can be used for binary as well as multiclass classification. The method was compared with state of the art – mRMR, *t*-test and F-test. The method is available as a MATLAB tool [113]. An embedded approach of regularized random forest regression has been developed by Liu *et al.* for selecting smaller subset of features without compromising prediction performance [55].

Eddy *et al.* reviewed newer classification method of relative expression analysis like top-scoring pair (TSP) classification algorithm based on relative order of gene-expression values which hold great potential for mining high-throughput data [56]. In 2012, Magis *et al.*, generalized the relative expression method as top scoring N algorithm [57]. Recently Czajkowski *et al.* have further improved upon TSP algorithm by combining it with evolutionary algorithm and developed EvoTSP algorithm [58].

Research efforts are being made to benchmark existing machine learning methods on cancer specific dataset. Zervakis *et al.* compared different cross-validation methods for gene selection on public avail-

Table 4. Recent developments in machine-learning algorithms for cancer datasets.

Algorithm	Purpose	Ref.
EvoTSP	For classifying microarray datasets, hybrid of evolutionary and relative expression algorithm; an improvement over relative expression algorithm	[58]
iPCC	Feature extraction from high throughput gene expression data; iterative employment of Pearson's correlation coefficient	[53]
Decision trunks	For cancer gene expression data classification; novel machine-learning algorithm; an improvement over decision tree algorithm	[106]
SPICE	System phenotype-related interplaying components enumerator for feature selection form instance-based and network-based data	[107]
HBSA	Heuristic breadth-first search algorithm; for gene ranking based on the occurrence frequency of genes in the gene subset	[108]
ECD	Extreme class discrimination; for feature selection, determines most discriminative variables	[109]
NPCA	Non-negative principal component analysis; for feature-selection filter-wrapper and NPCA-support vector machine algorithm based for classification of mass spectroscopic serum proteomic patterns	[110]
BRL	Bayesian rule learner; uses Bayesian scores and Bayesian model to build classification models	[111]
ellipsoidFN	Ellipsoid feature net; for feature selection based on ellipsoids; analyzed on gene expression dataset	[54]
SVMRFE + Fisher's method	A novel framework utilizing the advantages of a filtering method as well as an embedded method and redundancy reduction stage was added to address the weakness of the two. Furthermore, the proposed method uses gene ontology	[112]

able microarray datasets highlighting the influence of cross-validation procedures to outcome of gene signature and performance of predictive model [59]. Bartenhagen *et al.* compared unsupervised dimension reduction technique like Kernel PCA, Locally linear embedding, isomap, diffusion maps, laplacian eigenmaps and maximum variance unfolding. Linear embedding and isomap techniques (available as R package RDRToolbox) was proposed as superior alternatives to PCA for microarray data visualization [60].

Yuan *et al.* quantified the variability of prediction performance in developing predictive models while assessing the clinical utility of genomic and proteomics data across four cancers (kidney renal clear cell carcinoma, glioblastoma multiforme, ovarian serous cystadenocarcinoma and lung squamous cell carcinoma) [61]. The top sources of variability reported were reported as cancer type (35.7%), data type (17.4%) and their interactions (11.8%). By contrast, machine-learning algorithms exhibited moderate influence with 5.2%. The reporting standards were benchmark as datasets, scripts as well as workflow have been made available to research community at synapse platform to reproduce and improvise the models.

Active research is also being carried out in the area of clinical research informatics to catalyze collaborative research and develop computational models for cancer diagnosis and prognosis. Tools involving grid-based computations like ImageMiner for tissue microarrays employs SVM for its data analysis components [62], GLORE (Grid binary Logistic Regression) model which can develop shared models without sharing the datasets [63]. Initiatives like GitHub Open source cancer [64], encourages sharing the scripts being used in the field.

Conclusion

The systematic analysis discussed here indicates the trends from the reported studies that have attempted to identify cancer biomarkers using machine learning as a data-mining or classification technique. We have described the characteristics of studies in context of machine-learning algorithms, cancer type, data type, data sources, study type and limitations thereof. We recommend few guidelines for clinically relevant machine learning studies and research, which can potentially lead to better translational impacts. Insights gained from this study provide overview of recent machine learning applications in oncology and provide opportunities to data scientists for analyzing cancer datasets for determination of biomarkers. The study described here also provides introduction of machine learning to experimental oncologists.

Future perspective

The availability of higher resolution molecular data from affordable technologies has rendered impetus to translate the potential of personalized medicine concept. Molecular characterization of tumor cells leads to molecular stratification of the disease with plausible clinical applicability. The future lies in a personalized molecular characterization that will allow appropriate therapeutic combination for each patient as well as at each stage of the disease progression. Given the fact that most cancers are very heterogeneous, there is a need to reclassify them at molecular level rather than based on broad phenotypes. However, ethical, legal and public policy issues will have considerable impact on the implementation and hence shall be given careful due consideration.

Availability of cheaper than ever and affordable sequencing technologies as well as the advances in systems biology is contributing to make personalized medicine a part of standard clinical practices in the not too distant future. ML technologies are central to the idea of determining molecular signatures and classification models. Many studies have been performed on pilot scale demonstrating the applicability of ML in clinical cancer diagnosis and prognosis. However, the applicability of such data at a broader scale is still limited. Nevertheless, this field is evolving, and ML has much to offer toward cancer diagnosis and prognosis.

It is unlikely that we will have a specific drug that can treat a particular cancer for each individual genetic mutation detected. With the advent of molecular profiling technologies we may find an existing approved drug that can address genetic mutations for which it was not developed specifically. Furthermore, it will be pertinent to develop systematic knowledge base to accumulate whole genome sequencing of cancer patients, corresponding treatment outcomes and patient meta data. After two decades, we would not be surprised if as an essential health care routine, we all get our genomes sequenced, and ML algorithms in diagnostic laboratory render a reliable and accurate wellness plan, based on predicted clinical implications.

Financial & competing interest disclosure

The authors thank Department of Biotechnology (DBT, India) grant (BT/BI/25/001/2006) for "Bioinformatics Infrastructure Facility" at ICGEB, New Delhi. Z Jagga acknowledges University Grants Commission (UGC, India) for Senior Research Fellowship. The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript.

No writing assistance was utilized in the production of this manuscript.

Executive summary

- Developing cancer biomarkers and clinical grade classification models is an unmet need in cancer research.

Omics data in cancer

- Advent of high-throughput technologies and ensuing implementation for molecular profiling of tumors at the levels of genomics, transcriptomics, epigenomics, proteomics, lipidomics and metabolomics has generated a lot of cancer related data in public domain.
- International collaborative projects like TCGA and ICGC are comprehensively cataloging molecular alterations in various tumors types and providing opportunities for networking and grounds for interdisciplinary research.

Machine learning

- Machine-learning algorithms are being implemented to analyze cancer omics data from various molecular levels for developing new molecular subtypes, biomarkers and predictive classification models.
- State-of-art classification algorithm like artificial neural networks, support vector machine, decision trees and random forest are being employed for developing predictive classification models.
- Determining cancer biomarker or molecular signature from a molecular feature dataset is essentially a feature selection step in machine learning.

Survey of research articles

- Systematic literature survey for machine learning investigations linked to cancer biomarkers determined the following trends of increased year-wise publications, major investigations in breast cancer and major study type as diagnosis for cancer.
- Recent improvements have been in the fields of incorporating hierarchical features, new feature selection algorithms, classification, benchmarking and clinical informatics.

Future perspective

- Application of machine learning for cancer diagnosis, prognosis or therapeutics in clinics is a fertile research area with potential to alleviate implementation of clinical grade computational models for personalized medicine.

References

Papers of special note have been highlighted as:
• of interest; •• of considerable interest

1 Siegel R, Ma J, Zou Z, Jemal A. Cancer statistics, 2014. *CA Cancer J. Clin.* 64(1), 9–29 (2014).

2 Vineis P, Wild CP. Global cancer patterns: causes and prevention. *Lancet* 383(9916), 549–557 (2014).

3 Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell* 144(5), 646–674 (2011).

4 Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature* 458(7239), 719–724 (2009).

5 Smith RA, Brooks D, Cokkinides V, Saslow D, Brawley OW. Cancer screening in the United States, 2013. *CA Cancer J. Clin.* 63(2), 87–105 (2013).

6 Rabjerg M, Mikkelsen MN, Walter S, Marcussen N. Incidental renal neoplasms: is there a need for routine screening? A Danish single-center epidemiological study. *Apmis* 122(8), 708–714 (2014).

7 Seoane J, De Mattos-Arruda L. The challenge of intratumour heterogeneity in precision medicine. *J. Intern. Med.* 276(1), 41–51 (2014).

8 Garraway LA, Verweij J, Ballman KV. Precision oncology: an overview. *J. Clin. Oncol.* 31(15), 1803–1805 (2013).

9 Shrager J, Tenenbaum JM. Rapid learning for precision oncology. *Nat. Rev. Clin. Oncol.* 11(2), 109–118 (2014).

10 Midorikawa Y, Tsuji S, Takayama T, Aburatani H. Genomic approach towards personalized anticancer drug therapy. *Pharmacogenomics* 13(2), 191–199 (2012).

•• **Comprehensive overview of machine-learning algorithms for prediction of drug efficacy, chemosensitivity and proposes random forest as the best approach for class prediction of chemotherapy.**

11 Syed Z, Rubinfeld I. Personalized risk stratification for adverse surgical outcomes: innovation at the boundaries of medicine and computation. *Per. Med.* 7(6), 695–701 (2010).

12 Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24(3), 133–141 (2008).

13 Meldrum C, Doyle MA, Tothill RW. Next-generation sequencing for cancer diagnostics: a practical perspective. *Clin. Biochem. Rev.* 32(4), 177 (2011).

14 Costa FF. Big data in biomedicine. *Drug Discov. Today* 19(4), 433–440 (2014).

15 Marx V. Biology: the big challenges of Big data. *Nature* 498(7453), 255–260 (2013).

16 Mattmann CA. Computing: a vision for data science. *Nature* 493(7433), 473–475 (2013).

17 Tainsky MA. Genomic and proteomic biomarkers for cancer: a multitude of opportunities. *Biochim. Biophys. Acta* 1796(2), 176–193 (2009).

18 Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. *Nat. Med.* 17(3), 297–303 (2011).

19 Simon R, Roychowdhury S. Implementing personalized cancer genomics in clinical trials. *Nat. Rev. Drug Discov.* 12(5), 358–369 (2013).

20 Margolin AA, Bilal E, Huang E *et al.* Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci. Transl. Med.* 5(181), doi:10.1126/scitranslmed.3006112 (2013) (Epub ahead of print).

21 Cheng W-Y, Yang T-HO, Anastassiou D. Development of a prognostic model for breast cancer survival in an open challenge environment. *Sci. Transl. Med.* 5(181), doi:10.1126/scitranslmed.3005974 (2013) (Epub ahead of print).

22 Mjolsness E, Decoste D. Machine learning for science: state of the art and future prospects. *Science* 293(5537), 2051–2055 (2001).

23 Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform.* 2, 59–77 (2006).

•• **Comprehensive overview of machine-learning applications for cancer prognosis.**

24 Mccarthy JF, Marx KA, Hoffman PE *et al.* Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management. *Ann. NY Acad. Sci.* 1020(1), 239–262 (2004).

• **Comprehensive overview of machine-learning application in cancer detection, diagnosis and management.**

25 Vellido A, Biganzoli E, Lisboa PJ. Machine learning in cancer research: implications for personalised medicine. At: *The 16th European Symposium on Artificial Neural Networks ESANN.* Bruges, Belgium, 23–25 April 2008.

26 Zhu X, Goldberg AB. Introduction to semi-supervised learning. In: *Synthesis Lectures on Artificial Intelligence and Machine Learning.* Brachman R, Cohen WW, Stone P (Eds). Morgan & Claypool Publishers, San Rafael, CA, USA, 1–130 (2009).

27 Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23(19), 2507–2517 (2007).

•• **Comprehensive overview of feature selection methods, application domains and software packages.**

28 Christin C, Hoefsloot HC, Smilde AK *et al.* A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. *Mol. Cell. Proteomics* 12(1), 263–276 (2013).

•• **Comphrensive overview of feature selection methods implemented in clinical proteomics for detemining cancer biomarkers.**

29 Abeel T, Helleputte T, Van De Peer Y, Dupont P, Saeys Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics* 26(3), 392–398 (2010).

30 Basheer IA, Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application. *J. Microbiol. Methods* 43(1), 3–31 (2000).

31 Baxt WG. Application of artificial neural networks to clinical medicine. *Lancet* 346(8983), 1135–1138 (1995).

32 Gerlee P, Kim E, Anderson AR. Bridging scales in cancer progression: mapping genotype tophenotype using neural networks. *Semin. Cancer Biol.* 30, 30–41 (2014).

33 Hu X, Cammann H, Meyer H-A, Miller K, Jung K, Stephan C. Artificial neural networks and prostate cancer – tools for diagnosis and management. *Nat. Rev. Urol.* 10(3), 174–182 (2013).

•• **Comphrensive review of artificial neural network in prostate cancer diagnosis and management.**

34 Khan J, Wei JS, Ringner M *et al.* Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* 7(6), 673–679 (2001).

35 Lisboa PJ, Taktak AF. The use of artificial neural networks in decision support in cancer: a systematic review. *Neur. Netw.* 19(4), 408–415 (2006).

36 Lancashire LJ, Lemetre C, Ball GR. An introduction to artificial neural networks in bioinformatics – application to complex microarray and mass spectrometry datasets in cancer studies. *Brief Bioinform.* 10(3), 315–329 (2009).

37 Noble WS. What is a support vector machine? *Nat. Biotechnol.* 24(12), 1565–1567 (2006).

38 Vapnik VN. *Statistical Learning Theory.* Wiley, NY, USA, 2, 768 (1998).

39 Statnikov A, Wang L, Aliferis CF. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics* doi:10.1186/1471-2105-9-319 (2008) (Epub ahead of print).

40 Witten IH, Frank E, Hall MA. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann Publishers, MA, USA (2011).

41 Kingsford C, Salzberg SL. What are decision trees? *Nat. Biotechnol.* 26(9), 1011–1013 (2008).

42 Breiman L. Random forests. *Machine learning* 45(1), 5–32 (2001).

43 Chen X, Ishwaran H. Random forests for genomic data analysis. *Genomics* 99(6), 323–329 (2012).

•• **Comprehensive overview of implementation of random forest in genome data analysis.**

44 Touw WG, Bayjanov JR, Overmars L *et al.* Data mining in the Life Sciences with random forest: a walk in the park or lost in the jungle? *Brief Bioinform.* 14(3), 315–326 (2013).

45 Jagga Z, Gupta D. Classification models for clear cell renal carcinoma stage progression, based on tumor RNAseq expression trained supervised machine learning algorithms. *BMC Proceedings* 8(Suppl. 6), S2 (2014).

46 Okun O, Priisalu H. Random forest for gene expression based cancer classification: overlooked issues. In: *Pattern Recognition and Image Analysis.* Springer-Verlag Berlin Heidelberg, Germany, 483–490 (2007).

47 Google Scholar. http://scholar.google.co.in/

48 PubMed Central. www.ncbi.nlm.nih.gov/pmc/

49 PubMed. www.ncbi.nlm.nih.gov/pubmed

50 Kim S, Kon M, Delisi C. Pathway-based classification of cancer subtypes. *Biol. Direct* 7, 21 (2012).

51 Gatza ML, Lucas JE, Barry WT *et al.* A pathway-based classification of human breast cancer. *Proc. Natl Acad. Sci USA* 107(15), 6994–6999 (2010).

52 Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* 3, 140 (2007).

53 Ren X, Wang Y, Zhang XS, Jin Q. iPcc: a novel feature extraction method for accurate disease class discovery and prediction. *Nucleic Acids Res.* 41(14), e143 (2013).

54 Ren X, Wang Y, Chen L, Zhang X-S, Jin Q. ellipsoidFN: a tool for identifying a heterogeneous set of cancer biomarkers based on gene expressions. *Nucleic Acids Res.* 41(4), e53 (2013).

55 Liu S, Dissanayake S, Patel S *et al.* Learning accurate and interpretable models based on regularized random forests regression. *BMC Syst. Biol.* 8(Suppl. 3), S5 (2014).

56 Eddy JA, Sung J, Geman D, Price ND. Relative expression analysis for molecular cancer diagnosis and prognosis. *Technol. Cancer Res. Treat.* 9(2), 149–159 (2010).

57 Magis AT, Price ND. The top-scoring 'N' algorithm: a generalized relative expression classification method from small numbers of biomolecules. *BMC Bioinformatics* 13, 227 (2012).

58 Czajkowski M, Kretowski M. Evolutionary approach for relative gene expression algorithms. *ScientificWorldJournal* doi: org/10.1155/2014/593503 (2014) (Epub ahead of print).

59 Zervakis M, Blazadonakis ME, Tsiliki G, Danilatou V, Tsiknakis M, Kafetzopoulos D. Outcome prediction based on microarray analysis: a critical perspective on methods. *BMC Bioinformatics* 10, 53 (2009).

60 Bartenhagen C, Klein HU, Ruckert C, Jiang X, Dugas M. Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data. *BMC Bioinformatics* 11, 567 (2010).

61 Yuan Y, Van Allen EM, Omberg L *et al.* Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat. Biotechnol.* 32(7), 644–652 (2014).

62 Foran DJ, Yang L, Chen W *et al.* ImageMiner: a software system for comparative analysis of tissue microarrays using content-based image retrieval, high-performance computing, and grid technology. *J. Am. Med. Inform. Assoc.* 18(4), 403–415 (2011).

63 Wu Y, Jiang X, Kim J, Ohno-Machado L. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. *J. Am. Med. Inform. Assoc.* 19(5), 758–764 (2012).

64 GitHub. https://github.com/OpenSourceCancer

65 Mendelsohn J, Tursz T, Schilsky RL, Lazar V. WIN Consortium[mdash]challenges and advances. *Nat. Rev. Clin. Oncol.* 8(3), 133–134 (2011).

66 Worldwide Innovative Networking in personalized cancer medicine. www.winconsortium.org/

67 Hudson TJ, Anderson W, Aretz A *et al.* International network of cancer genome projects. *Nature* 464(7291), 993–998 (2010).

68 International Cancer Genome Consortium. https://icgc.org/

69 The Cancer Genome Atlas. http://cancergenome.nih.gov/

70 Dialogue for Reverse Engineering Assesments and Methods. www.the-dream-project.org/

71 Cerami E, Gao J, Dogrusoz U *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2(5), 401–404 (2012).

72 cBioPortal. www.cbioportal.org/

73 Gao J, Aksoy BA, Dogrusoz U *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* 6(269), pl1 (2013).

74 NCBI – Short Read Archive. www.ncbi.nlm.nih.gov/sra

75 Gene Expression Ominibus. www.ncbi.nlm.nih.gov/geo/

76 Catalogue of Somatic Mutations in Cancer. http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/

77 Cancer Cell Line Encyclopedia. www.broadinstitute.org/ccle/

78 Sun Y, Goodison S, Li J, Liu L, Farmerie W. Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics* 23(1), 30–37 (2007).

79 Lin X, Afsari B, Marchionni L *et al.* The ordering of expression among a few genes can provide simple cancer biomarkers and signal *BRCA1* mutations. *BMC Bioinformatics* 10, 256 (2009).

80 Powe DG, Dhondalay GK, Lemetre C *et al.* DACH1: its role as a classifier of long term good prognosis in luminal breast cancer. *PLoS ONE* 9(1), e84428 (2014).

81 Ushijima M, Miyata S, Eguchi S *et al.* Common peak approach using mass spectrometry data sets for predicting the effects of anticancer drugs on breast cancer. *Cancer Inform.* 3, 285–293 (2007).

82 Garcia-Bilbao A, Armananzas R, Ispizua Z *et al.* Identification of a biomarker panel for colorectal cancer diagnosis. *BMC Cancer* 12, 43 (2012).

83 Tsuji S, Midorikawa Y, Takahashi T *et al.* Potential responders to FOLFOX therapy for colorectal cancer by random forests analysis. *Br. J. Cancer* 106(1), 126–132 (2012).

84 Wan YW, Beer DG, Guo NL. Signaling pathway-based identification of extensive prognostic gene signatures for lung adenocarcinoma. *Lung Cancer* 76(1), 98–105 (2012).

85 Wan YW, Sabbagh E, Raese R *et al.* Hybrid models identified a 12-gene signature for lung cancer prognosis and chemoresponse prediction. *PLoS ONE* 5(8), e12222 (2010).

86 Glory E, Newberg J, Murphy RF. Automated comparison of protein subcellular location patterns between images of normal and cancerous tissues. *Proc. IEEE Int. Symp. Biomed. Imaging* 4540993, 304–307 (2008).

87 Campagne F, Skrabanek L. Mining expressed sequence tags identifies cancer markers of clinical interest. *BMC Bioinformatics* 7, 481 (2006).

88 Zhang C, Li HR, Fan JB *et al.* Profiling alternatively spliced mRNA isoforms for prostate cancer classification. *BMC Bioinformatics* 7, 202 (2006).

89 Bismar TA, Demichelis F, Riva A *et al.* Defining aggressive prostate cancer using a 12-gene model. *Neoplasia* 8(1), 59–68 (2006).

90 Lin HY, Amankwah EK, Tseng TS, Qu X, Chen DT, Park JY. SNP-SNP interaction network in angiogenesis genes associated with prostate cancer aggressiveness. *PLoS ONE* 8(4), e59688 (2013).

91 Prior C, Perez-Gracia JL, Garcia-Donas J *et al.* Identification of tissue microRNAs predictive of sunitinib activity in patients with metastatic renal cell carcinoma. *PLoS ONE* 9(1), e86263 (2014).

92 Jesneck JL, Mukherjee S, Yurkovetsky Z *et al.* Do serum biomarkers really measure breast cancer? *BMC Cancer* 9, 164 (2009).

93 Zhang X, Wang B, Zhang XS, Li ZM, Guan ZZ, Jiang WQ. Serum diagnosis of diffuse large B-cell lymphomas and further identification of response to therapy using SELDI-TOF-MS and tree analysis patterning. *BMC Cancer* 7, 235 (2007).

94 Cohen M, Yossef R, Erez T *et al.* Serum apolipoproteins C-I and C-III are reduced in stomach cancer patients: results from MALDI-based peptidome and immuno-based clinical assays. *PLoS ONE* 6(1), e14540 (2011).

95 Kolusheva S, Yossef R, Kugel A *et al.* A novel "reactomics" approach for cancer diagnostics. *Sensors (Basel)* 12(5), 5572–5585 (2012).

96 Jia J, Wang W, Meng W, Ding M, Ma S, Wang X. Development of a multiplex autoantibody test for detection of lung cancer. *PLoS ONE* 9(4), e95444 (2014).

97 Bigbee WL, Gopalakrishnan V, Weissfeld JL *et al.* A multiplexed serum biomarker immunoassay panel discriminates clinical lung cancer patients from high-risk individuals found to be cancer-free by CT screening. *J. Thorac. Oncol.* 7(4), 698–708 (2012).

98 Shan L, Chen YA, Davis L *et al.* Measurement of phospholipids may improve diagnostic accuracy in ovarian cancer. *PLoS ONE* 7(10), e46846 (2012).

99 Washam CL, Byrum SD, Leitzel K *et al.* Identification of PTHrP(12–48) as a plasma biomarker associated with breast cancer bone metastasis. *Cancer Epidemiol. Biomarkers Prev.* 22(5), 972–983 (2013).

100 Guergova-Kuras M, Kurucz I, Hempel W *et al.* Discovery of lung cancer biomarkers by profiling the plasma proteome with monoclonal antibody libraries. *Mol. Cell Proteomics* doi:10.1074/mcp.M111.010298 (2011) (Epub ahead of print).

101 Lanara Z, Giannopoulou E, Fullen M *et al.* Comparative study and meta-analysis of meta-analysis studies for the correlation of genomic markers with early cancer detection. *Hum. Genomics.* doi: 10.1186/1479-7364-7-14 (2013) (Epub ahead of print).

102 Zhao D, Weng C. Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction. *J. Biomed. Inform.* 44(5), 859–868 (2011).

103 Gormley M, Dampier W, Ertel A, Karacali B, Tozeren A. Prediction potential of candidate biomarker sets identified and validated on gene expression data from multiple datasets. *BMC Bioinformatics* 8, 415 (2007).

104 Zheng G, Xiong Y, Xu W *et al.* A two-microRNA signature as a potential biomarker for early gastric cancer. *Oncol. Lett.* 7(3), 679–684 (2014).

105 Rosser CJ, Liu L, Sun Y *et al.* Bladder cancer-associated gene expression signatures identified by profiling of exfoliated urothelia. *Cancer Epidemiol. Biomarkers Prev.* 18(2), 444–453 (2009).

106 Ulfenborg B, Klinga-Levan K, Olsson B. Classification of tumor samples from expression data using decision trunks. *Cancer Inform.* 12, 53–66 (2013).

107 Chen Z, Padmanabhan K, Rocha AM *et al.* SPICE: discovery of phenotype-determining component interplays. *BMC Syst. Biol.* 6, 40 (2012).

108 Wang SL, Li XL, Fang J. Finding minimum gene subsets with heuristic breadth-first search algorithm for robust tumor classification. *BMC Bioinformatics* 13, 178 (2012).

109 Toh SH, Prathipati P, Motakis E, Kwoh CK, Yenamandra SP, Kuznetsov VA. A robust tool for discriminative analysis and feature selection in paired samples impacts the identification of the genes essential for reprogramming lung tissue to adenocarcinoma. *BMC Genomics* 12(Suppl. 3), S24 (2011).

110 Han H. Nonnegative principal component analysis for mass spectral serum profiles and biomarker discovery. *BMC Bioinformatics* 11(Suppl. 1), S1 (2010).

111 Gopalakrishnan V, Lustgarten JL, Visweswaran S, Cooper GF. Bayesian rule learning for biomedical data mining. *Bioinformatics* 26(5), 668–675 (2010).

112 Mohammadi A, Saraee MH, Salehi M. Identification of disease-causing genes using microarray data mining and gene ontology. *BMC Med. Genomics* 4, 12 (2011).

113 MATLAB tool. http://doc.aporc.org/wiki/EllipsoidFN