Discover Artificial Intelligence

RESEARCH Open Access

Biology-based multi-class classification of cancer using genomics and artificial intelligence

Maher Albitar[1*], Hong Zhang[2], James McCloskey[3], David Siegel[3], Martin Gutierrez[3], Jamie Koprivnikar[3], Andrew Pecora[3], Sally Agersborg[1], Ahmad Charifa[1], Michele Donato[3], Andrew Ip[3] and Andre Goy[3]

*Correspondence:
Maher Albitar
malbitar@genomictestingcooperative.com
[1]Genomic Testing Cooperative, 25371 Commercentre Dr., Lake Forest, California 92630, USA
[2]Department of Computer Science, Georgia Southern University, Savannah, GA 31419, USA
[3]John Theurer Cancer Center at Hackensack University Medical Center, Hackensack, NJ 07601, USA

Abstract

Cancer classification traditionally focuses on tissue of origin and histology appearance. Traditional "hard" classification for a specific diagnostic class yields binary yes/no class prediction. However, newer therapeutic approaches have demonstrated that tumors overlap biologically and may respond to a specific therapeutic approach regardless of traditional histological classifications. Instead of histological classification, we propose the use of a "Functional classification" based on machine learning models, leveraging high-dimensional RNA expression data. This approach identifies not only the predicted diagnostic class with the highest probability, but also the second, third, and other possible diagnostic classes. RNA from various types of cancers was sequenced and quantified using next generation sequencing. We used these RNA expression profiles in several functional classification models trained on 3484 tumors across 25 cancer types and tested these models using 1716 tumors. The Extreme Gradient Boosting Classifier showed the greatest accuracy, followed by the Multi-Layer Perceptron and random forest models. We demonstrate that using RNA profiling and machine learning achieve high accuracy in both hard and functional classifications of tumors. However, significantly better accuracy in classification was achieved using functional classification reflecting biological overlapping between tumors. Functional classification shows proximity between tumors and enables exploring treatments used in other close diagnostic class.

Keywords Cancer, Diagnosis, Machine learning, Genomics, RNA, DNA, Functional classification

1 Introduction

Classification of cancer is traditionally based on the site of origin and histological appearance. However, recent developments in molecular pathology have led to attempts to incorporate molecular abnormalities into various classification schemes [1–6]. Furthermore, evolving therapeutic approaches have shown that tumors can be treated and may often respond to therapy based on their specific molecular abnormalities [7–11],

regardless of the pathologic classification as shown by FDA approvals of now several drugs tissue agnostic in immunotherapy and targeted therapy.

Advances in next-generation sequencing (NGS) have enabled the efficient and practical profiling of DNA and RNA changes in tumors, with results often available in a few days. The availability of such extensive molecular profiles offers a valuable opportunity to not only better understand the mechanisms driving the uncontrolled growth and proliferation of tumors but also classify these tumors and determine therapeutic approaches that specifically target the driving genomic abnormalities [12–15]. The recent, rapid expansion of new targeted therapeutic approaches necessitates more reliable approaches for the biological classification of tumors. Therapeutic approaches such as immune therapy, CAR T-cell therapy, antibody-drug conjugates (ADCs), bispecific antibodies, and even small-molecule drugs are all used based mainly on biological findings [16] rather than morphology or site of origin.

Combining DNA and RNA profiling provides valuable information on the biological changes in a tumor as compared to normal tissue. However, the use of artificial intelligence (AI) [17] is becoming essential for better and more comprehensive utilization of such high-dimensional, complex data. AI approaches must incorporate mechanisms that adjust for the many technical and biological complexities typically encountered with RNA and DNA data. For example, the tumor microenvironment may play a significant role in tumor biology, and the spatial relationship of tumor cells to microenvironment cells may also be important for clinical decisions [17]. Although high-dimensional RNA data can provide relevant information, sophisticated algorithms, such as AI models, are needed to decipher the relationship of the tumor with the microenvironment, the downstream effects of mutations and the other biological and technical aspects [18]. This step is crucial to provide information that is practical for use in everyday medical practice. RNA and DNA profiles are generated from biological samples, which can vary significantly in their presentation of the biological processes in the patient. For example, the tumor fraction in the examined tissue may vary significantly from one sample to the next; AI systems must be robust enough to account for such variation.

Traditional statistical and machine learning methods are typically designed to classify input data into one of the predefined, mutually exclusive classes. Multi-class classification problems are often handled with multiple binary classifiers using approaches such as one vs. others and pair-wise classification [19-21]. Although effective, these approaches tend to ignore the overall structure of multiple classes. Processing this type of multi-class, high-dimensional, heterogeneous, and overlapping mixture data from NGS requires special considerations in statistical analysis.

In this paper, we explore different AI approaches that can be used to better understand and utilize RNA and DNA profiling data from cancers. Furthermore, we attempt to utilize this approach to provide objective alternative to diagnosis that is not based on subjective microscopic examination or elaborate and expensive immunohistochemistry or flow cytometry evaluation. We highlight that the goal of an AI approach in genomic data analysis extends beyond distinguishing between historically characterized diseases based on site of tumor or morphology, which we call “hard classification” (Fig. 1). We propose that AI-based classification based on genomic data should involve defining similarities or proximity between diseases and overlapping and calling this “functional classification”. Recognizing such similarities may guide therapeutic approaches and future

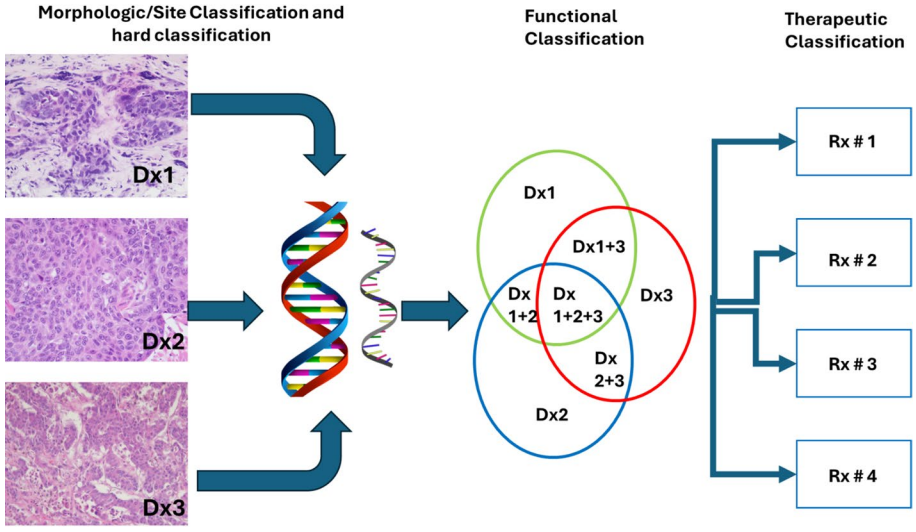

Fig. 1 Diagram representing current classification of cancers based on morphology and site of origin (hard classification) and the proposed classification encompassing hard classification and biologically overlapping or proximal classes (Functional classification) that may lead to therapeutic recommendations based on biology (therapeutic classification)

research and may lead to "therapeutic classification" (Fig. 1). Also this approach may overcome reproducibility issues in the subjective morphologic evaluation and immunohistochemistry staining, particularly in community setting.

Functional classification refers to methods that produce probability estimates for the class prediction, rather than the traditional hard 0/1 class prediction. Functional classification identifies not only the predicted class with the highest probability, but also the second, third, and other possible classes. This approach more closely reflects the biological reality of classification problems. The diagnostic classes predicted with functional classification can be used to explore therapeutic approaches typically used in other diagnostic classes that scored second or third to the hard/conventional classification. This approach may also lead to a classification that reflects response to therapy rather than strictly the site of origin or morphology.

To assess the performance of a functional classifier, we propose a measure that generalizes the standard recall. The generalized recall takes into consideration inter-relationships among the classes. We used a symmetric matrix to describe the class relations and employed an automated process to determine the matrix of class relations by using the simplex formed by the class means. The simplex approach captures the essential structure of the classes in a holistic view.

The advantages of such approach are multiple beyond more properly reflecting biology. This approach helps in expanding the utilization of therapeutic approaches by indicating therapy for tumor types that are biologically overlapping. This approach relies on objective quantitative data rather than morphologic evaluation that varies based on the skills and experience of the pathologist making the diagnosis. Furthermore, neoplasms do not always follow specific classifications and may present in unexpected site and unclassifiable morphology. This is frequently encountered with cancer of unknown primary (CUP). Our approach overcomes this problem. At the same time, the disadvantage of this approach is its relying on extracted RNA from tissue that needs to be standardized and reproducible. Samples processing may affect the quantification of the RNA

and reliability of prediction. Furthermore, the relative fraction of tumors in the analyzed sample plays a role in the cancer-specific RNA and careful standardization of this approach is needed.

The recent advances in targeted therapy, immunotherapy, and cellular therapy demand more concise diagnosis based on biology and not only based on morphologic features. The proposed diagnostic approach provides the tool necessary to better use the new therapeutic approaches. However, more work and research are needed to develop new therapeutic approaches and models utilizing the new proposed diagnostic approach.

Adaptation of the proposed diagnostic approach represents paradigm shift in the practice of pathology and requires educational shift not only for pathologists but also for insurance companies and health care payers.

2 Methods

2.1 Patients and samples

Five-thousand two-hundred consecutive, deidentified fresh bone marrow and formalin-fixed paraffin-embedded (FFPE) cancer samples were included in this retrospective study (Fig. 2). The samples were collected during routine clinical molecular profiling using NGS of DNA and RNA between November 2018 and November 2021. Samples were unique for each patient and no repeat samples were included in this study. The tumor fractions varied between 30% and 80%. These samples presented various solid tumor and hematologic neoplasms, including leukemia, myelodysplasia, lymphoma, lung, colorectal, breast, pancreas, endometrial, ovarian, sarcoma, and others, as well as normal bone marrow Table 1. Tumor diagnosis was confirmed using morphological analysis, flow cytometry, IHC, and molecular profiling of the DNA and RNA. Only RNA expression and DNA mutation profiling were used in testing the various algorithms. No clinical variables are used in any of the machine learning algorithms. Approximately two thirds of patients were used for training various algorithms and one third is used for testing. The training and testing sets were selected randomly covering the entire period of collection. The expression levels of 1408 genes that are commonly involved in oncogenesis are used in the study. Using targeted RNA panel focused on oncogenic genes and eliminated overexpressed housekeeping genes increased the dynamic range of the expression level of various oncogenic genes (data not shown). Special attention is paid to excluding cases in which RNA was contaminated by DNA, which typically increases the expression level of all genes. DNA contamination is expected when the percentage of splice junction read is less than 25% to 30%. Study data was collected under approved IRB (WCG IRB # 1-1476184-1).

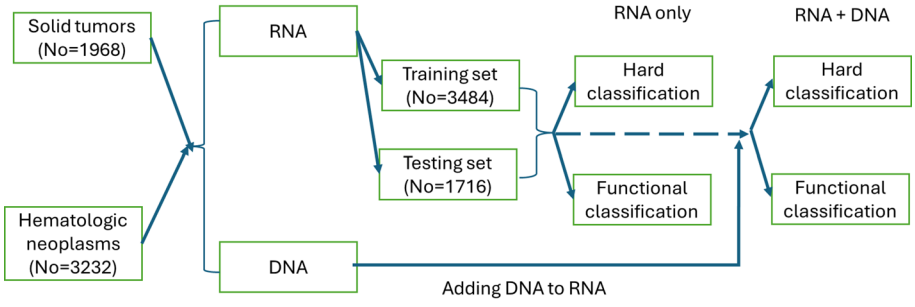

Fig. 2 Schematic presentation of research plan and data analysis

Table 1 Training and test sample sizes for each diagnostic class

Diagnostic class	Training set	Test set
Acute lymphoblastic leukemia	60	32
Acute myeloid leukemia	265	142
Breast cancer	120	60
Carcinoma (tumor of unknow origin)	47	19
Chronic lymphocytic leukemia	140	58
Chronic myelomonocytic leukemia	67	37
Colorectal cancer	188	94
Diffuse large B-cell lymphoma	212	108
EBV-positive diffuse large B-cell lymphoma	48	27
Extra nodal Diffuse large B-cell lymphoma	76	30
Endometrial cancer	55	30
Follicular lymphoma	137	44
Hodgkin lymphoma	44	28
Lung cancer	680	344
Mantle cell lymphoma	78	56
Marginal cell lymphoma	39	26
Myelodysplastic syndrome	224	128
Multiple myeloma	64	37
Normal bone marrow	569	243
Ovarian cancer	90	31
Pancreatic cancer	61	32
Pediatric large B-cell lymphoma	40	26
Plasmablastic lymphoma	53	21
Sarcoma not otherwise specified	79	38
T-cell lymphoma	48	25
Total	3484	1716

The DNA and RNA were extracted from bone marrow samples using an automated Maxwell System platform (Promega). Agencourt FormaPure Total 96-Prep Kits were used to extract both the DNA and RNA from FFPE samples using an automated King-Fisher Flex Purification System (Thermo Fisher Scientific), following the manufacturer's recommendations. The Agencourt FormaPure Kits have provision for a split protocol to extract both the DNA and RNA from the same FFPE lysate. Bone marrow samples were collected in ethylenediaminetetraacetic acid. DNA and RNA were extracted from fresh samples within 72 h of collection. The study protocol was approved by the Western Copernicus Group Institutional Review Board (New England IRB, Aspire IRB, and Midlands IRB) (Number 1-1476184-1). Patient informed consent was waived due to incidental collection and lack of risk. This study was conducted in accordance with the principles of the Declaration of Helsinki and its later amendments.

2.2 RNA and DNA library construction and sequencing

For RNA sequencing, the samples were selectively enriched for 1408 cancer-associated genes using the reagents provided in the Illumina® TruSight® RNA pan-cancer panel (Illumina). cDNA was generated and sequenced as previously reported [17].

For DNA sequencing, 100 ng of DNA was used for library preparation of targeted 434 gene sequencing based on single-primer extension (SPE) chemistry. Target enrichment was performed post-UMI (unique molecular Identifier) assignment to ensure that DNA molecules containing UMIs were sufficiently enriched in the sequenced library. The sequencing was conducted using Illumina NovaSeq 6000 or NextSeq 550 instruments.

2.3 Artificial intelligence models

2.3.1 Functional classification models

Given problems such as data sparsity and increased computational complexity that arise when analyzing data in high-dimensional spaces—sometimes called "the curse of dimensionality"—traditional statistical methods are not readily applicable to NGS data. Statistical learning theory has established mathematical bounds to control the capacity of machine learning systems for optimal performance [19, 20]. To address overfitting and issues with multiple testing stemming from the high dimensionality of NGS data, machine learning techniques have been developed to incorporate strategies such as regularization and dimension reduction [20, 21]. Support vector machine (Table 2) and Gaussian processes employ a kernel as a similarity measure and derive an optimal hyperplane as the classifier. The kernel methods resolve the dimensionality problem by

Table 2 Algorithms used with their strengths and weaknesses

Classifier	Brief Description	Strengths	Weaknesses	References
Random Forest	Ensemble of decision trees	High accuracy, resistance to overfitting	Complexity, high computational need, black box, poor handling of high-dimension	[26]
Extreme Gradient Boosting	Gradient boosted decision trees	High accuracy	Complexity	[26]
Gaussian Naïve Bayes (GNB)	Assumes features are independent given the class	Simplicity, speed, and efficiency with large datasets.	Independence assumption, lower accuracy, underflow problem	[23]
Geometric Mean Naïve Bayesian (GMNB)	Using geometric mean with Naïve Bayesian	Avoids underflow, simplicity, speed, and efficiency with large datasets.	Assumes independence	[22]
Multi-Layer Perceptron	Artificial neural network that consists of multiple layers of interconnected nodes, fundemental model in deep learning.	Can model complex, non-linear relationships in data, versitile, able to establish hierarchical features	Hyperparameter optimization, overfitting	[24, 25]
Adaptive boosting (AdaBoost)	Sequentially combines multiple weak learners using decision trees	Reduce overfitting, good for imbalanced data,	Sensitive to outliers and noisy data, computationally demanding	[25]
k-Nearest Neighbors	classifies a new data point based on the majority class of its k-nearest neighbors in the training data.	Ease of implementation, adaptability to new data, and minimal assumptions about data distribution	Computationally demanding, sensitive to the choice of k and distance metric, poor with high-dimensional data.	[25]
Decision Tree	Flowchart-like tool used for decision-making, visualizing potential choices and their outcomes to guide a choice	easy to understand, good for decision-making, require minimal data preparation.	Decision trees can be prone to overfitting, especially when they grow too deep, and they can be unstable, meaning slight changes in the data can lead to significantly different trees.	[25]
Support vector machine (SVM)	Finds the best hyperplane in an N-dimensional space to separate data points into different classes, maximizing the margin between the closest points of different classes.	Effective in high-dimensional spaces, handle non-linear data, resistant to overfitting.	Computationally demanding, sensitive to outliers, requires careful kernel selection, require data scaling or normalization	[20, 21]

collapsing the input vectors into the Gram matrix of kernel values. Kernel methods can be sensitive to the scaling and noise of the input.

Random forest (Table 2) and other ensemble methods provide a classifier that is relatively simple to construct and show good control of overfitting. Consequently, random forest is a suitable model for direct NGS data analysis. A potential disadvantage of random forest is that, as a black box machine learning model, it is not easily interpretable. Because the output of a random forest is a hard classification, post-processing steps such as a *softmax* function are needed to accommodate overlapping classes in NGS applications.

The Naïve Bayes classifier (Table2) is a simple classifier that naturally facilitates functional classification. However, the standard Naïve Bayes method suffers from severe underflow with high-dimensional data, which makes it essentially a hard classifier. We propose a modified version, Geometric Mean Naïve Bayes (Table2), to eliminate the underflow problem, resulting in a true functional classifier [22].

Traditional classifiers are usually designed for "hard classification." The predictions of such classifiers are specific class labels. Even if certain classifiers produce some intermediate results, such as class probabilities, the intermediate values will be converted to a final hard classification.

Hard classification may not effectively address the multi-class heterogeneity found in NGS expression data analysis. The issue of mixture data in NGS is similar to that found in statistical methods for mixture analysis, as used in chemistry. However, because NGS data in clinical applications do not have fixed mixture ratios, traditional mixture analysis does not apply to NGS data analysis. NGS RNA expression is additive. With a case-wise normalization of the input vector, the input for a mixture will be a weighted sum of the subcomponents. A feasible approach to build a classifier is a machine learning system that produces probability estimates for all the classes, instead of a hard class label. Such a "functional" classification system would be effective for mixture data if it produces higher probability outs for the components of the mixture.

We consider a functional classification paradigm to address problems associated with traditional hard classification. A functional classifier will produce probability estimates for the classes, instead of a one-hot coded answer indicating 1 ("yes") for one class and 0 ("no") for all other classes.

The Naïve Bayesian classifier is based on the Bayes Theorem and the assumption that all attributes are conditionally independent (Table 2).

Let $(x_1, x_2, \ldots, x_d)$ be the input attribute vector and $(C_1, C_2, \ldots, C_k)$ be the classes. By Bayes Theorem

$$P(C_j|x_1, x_2, \ldots, x_d) = \frac{P(C_j)\, P(x_1, x_2, \ldots, x_d|C_j)}{\sum_{i=1}^{k} P(C_i)\, P(x_1, x_2, \ldots, x_d|C_i)}$$

With the assumption of conditional independence, the original Naïve Bayesian classifier decomposes the likelihood into a product of conditional probabilities

$$P(x_1, x_2, \ldots, x_d|C_j) = P(x_1|C_j)\, P(x_2|C_j) \ldots P(x_d|C_j)$$

The probabilities $P(x_i|C_j)$ can be easily estimated from training data.

The Naïve Bayesian algorithm is one of the most elegant machine learning methods. However, when the dimension d is large, the products of the probabilities (the likelihood) will become extremely small, causing underflows.

If each probability value has an average of 1/2, then the likelihood will have the mean

$$E\left[P\left(x_1|C_j\right)P\left(x_2|C_j\right)\ldots\ P\left(x_d|C_j\right)\right]=\ \frac{1}{2^d},$$

which approaches 0 quickly when d is large.

A common method to avoid numerical underflow is to scale all values using the largest probability product during computations. However, with this method, one value often dominates the probability products. As a result, the predicted probability will be 1.0 for one class and 0.0 for all other classes. This effect is not desirable for most applications because it is an artifact of the Naïve Bayesian assumption and usually does not reflect the probabilistic structure of the reality.

The Naïve Bayesian classifier naturally produces a probability estimation for each class, so it can be considered a functional classifier. However, with high dimensional data, a Naïve Bayesian classifier will suffer from underflow and will become essentially a hard classifier. That is, the probability estimates for the classes will be either 0 or 1.

We previously proposed a true functional classifier, the Geometric Mean Naïve Bayesian (GMNB) classifier, to address this problem.

2.3.2 Geometric Mean Naïve Bayesian classifier

The GMNB classifier (Table 2) is an extension to the original Naïve Bayesian method, designed to overcome the underflow problem and produce reasonable probability estimates for high-dimensional data sets.

With the assumption of conditional independence, the original Naïve Bayesian classifier decomposes the likelihood into a product of conditional probabilities

$$P\left(x_1,x_2,\ldots\ ,x_d|C_j\right)=\ P\left(x_1|C_j\right)P\left(x_2|C_j\right)\ldots\ P\left(x_d|C_j\right)$$

The probabilities $P\left(x_i|C_j\right)$ can be easily estimated from training data.

However, when the dimension d is large, the product of the probabilities (the likelihood) will become extremely small, causing underflow. Even with a scaling technique to maintain the highest likelihood at a reasonable value, other likelihood values will still underflow, resulting in estimated class probabilities of 0 or 1 only.

In [1], we propose the GMNB method to address this problem, by applying a multiplicative positive increasing function to the likelihood:

$$P\left(x_1,x_2,\ldots\ ,x_d|C_j\right)=\ \sqrt[d]{\left[P\left(x_1|C_j\right)P\left(x_2|C_j\right)\ldots\ P\left(x_d|C_j\right)\right]}$$

This formula represents the geometric mean of the conditional probabilities. We previously proved that the GMNB method resolves the underflow problem for high dimensional data, by showing that the expected value of such a likelihood will approach $1/e$ when the dimension d approaches infinity. We also proved that such a function is essentially unique up to a constant multiple of the exponent.

The GMNB classifier preserves the elegance of the original Naïve Bayesian classifier, producing the same theoretical ordering of predicted classes as the original method. It

is efficient, effective, and less vulnerable to overfitting than most other machine learning algorithms. All these properties are important for processing high-dimensional data. The GMNB method eliminates underflow and produces reasonable probability estimates for arbitrarily large dimensions. Consequently, the GMNB classifier is a natural approach for analyzing high-dimensional NGS data.

2.3.3 Gaussian Naïve Bayes classifier

This is a Naïve Bayes classifier (Table 2) with a further assumption of Gaussian distributions on the likelihood. This approach assumes a diagonal covariance matrix on features for each class. As a standard Naïve Bayes approach, it is less prone to overfitting. However, the (GNB) is subject to underflow when the dimension is high. Consequently, it is not suitable for functional classification [23].

2.3.4 Multi-Layer Perceptron classifier (MLP)

MLP (Table 2) is the most fundamental type of neural network architecture. A neural network provides a general-purpose machine learning model with a large number of parameters. The successful training of a neural network depends on many factors, such as the problem complexity, the network structure, and regularization for controlling overfitting [24, 25].

2.3.5 Adaptive boosting (AdaBoost) classifier

The adaptive boosting (AdaBoost) (Table 2) classifier is one of the ensemble boosting classifiers that combine multiple weak classifiers to increase accuracy. The basic concept behind AdaBoost is that it assigns higher weights to misclassified training instances, leading to more accurate prediction in the next iteration. The process reiterates until either the training data set is fit without error, or a prespecified number of estimators is reached. AdaBoost can achieve high accuracy in some applications but it is vulnerable to noisy data [25].

2.3.6 k-Nearest Neighbors classifier (KNN)

KNN is a non-parametric, supervised learning classifier that uses proximity to make classifications. It is one of the "lazy learning" models that do not require explicit training. Despite its simplicity, KNN can produce reasonable performance. However, the "curse of dimensionality" and overfitting can be potential problems with KNN [25].

2.3.7 Decision tree classifier

Decision tree (Table 2) is an efficient, intuitive model for classification. The main drawbacks of the decision tree include overfitting and non-robustness [25].

2.3.8 Random forest classifier

Random Forest (Table 2) is an ensemble learning method. Multiple decision trees are trained by randomly sampling the data set and the features. Predictions are based on the voting of decision trees [26].

2.3.9 Extreme gradient boosting classifier

The XGBoost classifier (Table 2) is a gradient-boosted decision tree ensemble learning algorithm. XGBoost is efficient and often produces accurate results on a wide range of applications [26].

2.4 Combining RNA expression functional classifiers with mutation data

The functional classification results from different classifiers, even with very different modalities, can be combined to produce a stronger classifier. Consider the example of incorporating mutation information into the RNA expression data analysis. Supplementary A Table S1 shows the frequencies of certain mutations occurring in different cancers. The NGS data provide the mutation information.

By using the Bayes Theorem, we can calculate the probability of each cancer type given the set of mutations. This will provide a functional classification based on mutations. The functional classifications based on (1) RNA expression and (2) mutation data can be combined to generate a potentially stronger functional classification. Let p_1 and p_2 be the estimated probabilities of a cancer type from the two functional classifiers. It is natural to assume that the complements of probabilities $1-p_1$ and $1-p_2$ satisfy the naïve assumption. Therefore, the probability of not having the disease after the two tests is $(1-p_1)(1-p_2)$, and the combined estimated probability of the cancer type is $1-(1-p_1)(1-p_2)$.

2.5 Performance measures

The confusion matrix is a common way to represent the performance of a (hard) classifier. The (*i*, *j*)-th entry of the matrix is the number of cases of class *i* that are classified into class *j*. Suppliment table S2 shows an example of the confusion matrix for the 4-class problem with the GMNB classifier. A direct generalization of the confusion matrix for functional classification is to replace each entry with the sum the probabilities of class *i* to class *j* classification. Supplement A Table S3 Shows the functional confusion matrix for the 12 breast cancer misclassified cases in S2.

Because the functional classifier spreads the probability estimates instead of a single class, the functional confusion matrix is more diffused than the one for hard classification. The diffused entries do not necessarily indicate noises or classification errors. They could represent actual overlapping and mixture of different classes. To define summary performance measures for functional classification, we consider the generalization of common measures of precision and recall.

For hard classifiers, typical performance measures include precision and recall. For each class i,

$$precision_i = \frac{\#\ of\ instances\ correctly\ predicted\ to\ be\ class\ i}{\#\ of\ instances\ predicted\ to\ be\ class\ i}$$

$$recall_i = \frac{\#\ of\ instances\ correctly\ predicted\ to\ be\ class\ i}{\#\ of\ instances\ in\ class\ i}$$

Other measures can be derived from precision and recall. For example, the F1-score is the harmonic mean of precision and recall. In binary classification, the recall of the positive class is also called sensitivity, and the recall of the negative class is specificity. The

precision of the positive class is the positive predictive value, and the precision of the negative class is the negative predictive value.

To measure the performance of a functional classifier, we propose to generalize the definitions of precision and recall. The classes in a functional classification problem are not independent, because they may have overlap and other correlations. If the number of classes is *m*, the relations between the classes will be described by a symmetric matrix:

$$R = \begin{bmatrix} 1 & a_{12} & \dots & a_{1m} \\ a_{12} & 1 & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \dots & 1 \end{bmatrix},$$

where $0 \leq a_{ij} \leq 1$ indicates the connection or the amount of overlap between class *i* and class *j*.

Let C_i denote the class *i*, y_j the class label of the instance *j*, and $(p_{j1}, p_{j2}, \dots, p_{jm})$ the probability estimates of functional classifier for instance *j*. The generalized precision is defined as:

$$precision_i = \frac{1}{\sum_{j=1}^{n} p_{ji}} \sum_{k=1}^{m} \sum_{j \in C_k} a_{ik} p_{jk}$$

The generalized recall is defined as:

$$recall_i = \frac{1}{|C_i|} \sum_{j \in C_i} \sum_{k=1}^{m} a_{ik} p_{jk}$$

This definition, while retaining the meaning of the original performance measure, will incorporate information on correlation among classes. In the case of hard classification, the generalized precision and recall coincide with the original definitions. If two classes are perfectly correlated, then the generalized recall will give the value as if the two classes are merged.

The measures of precision, recall, and F1-score are defined for each class. The overall performance of a classifier can be defined on the entire data set. The accuracy is the rate of correct classification over all classes. The macro average of the precision, recall, and F1-score are the simple arithmetic averages of these measures of the classes. The weighted average of the measures is the average of these measures weighted by the class sizes.

Our approach for the generation of confusion matrices for hard classifiers and functional classifiers is described in supplement B.

3 Results

In this retrospective study, we compared eight machine learning AI models for classification of cancers. The study sample comprised 5200 consecutive deidentified fresh bone marrow and formalin-fixed paraffin-embedded (FFPE) specimens collected during routine clinical molecular profiling using NGS of DNA and RNA. The specimens encompassed 25 diagnostic classes of hematologic malignancies and solid tumors, including leukemia, myelodysplasia, lymphoma; lung, colorectal, breast, pancreas, endometrium, and ovarian cancers; sarcomas; other tumors; and normal bone marrow (Table 1). Two-thirds of the specimens were used for the training set and one-third for the testing

set. The machine learning models evaluated were the geometric mean Naïve Bayesian (GMNB), Gaussian Naïve Bayes (GNB), Multi-Layer Perceptron (MLP), Adaptive Boosting (AdaBoost), k-nearest neighbors (KNeighbors), decision tree, random forest, and Extreme Gradient Boosting (XGB) classifiers (see Methods).

3.1 Reliability of the classification models for hard classification

As shown in Table 3, the best overall classification accuracy was obtained using XGB (0.83), followed by MLP (0.82) and random forest (0.80). GMNB and GNB classifications both showed an accuracy of 0.6. Decision tree and KNeighbors classifiers had slightly better accuracy (0.7 and 0.71, respectively). AdaBoost showed unacceptable classification accuracy (0.24). However, accuracy varied markedly across diagnostic classes, particularly for diagnoses that overlap with others. For example, random forest classification yielded low F-1 scores for chronic myelomonocytic leukemia (CMML; 0.37) and meylodyplastic syndromes (MDS; 0.61), likely because of the overlap between these two entities and normal tissue. Similarly, the F-1 score for carcinoma of unknown origin was only 0.1; these cases most likely represented a specific other diagnosis, such as lung, breast, or colorectal cancer. In contrast, biologically distinct diagnostic classes, such as mantle cell lymphoma and cancers of the lung, breast, and colorectal region, were classified with high accuracy (F1-score $\geq$ 0.9 with random forest and XGB). These findings suggest that, except for Addaboost, most of the machine learning approaches evaluated are acceptable for developing models for cancer diagnosis based on NGS-generated RNA data. These data also support the reliability of the RNA data generated by NGS.

3.2 Functional classification

Functional classification showed significant improvements in classification for all models. The average F1-score improved by 20% to 30% (Table 4), reaching 0.95 for XGB and 0.89 for random forest. For GNB classification, the average F-1 improved from 0.62 for hard classification to 0.92 for functional classification (Table 4). Looking at individual diagnoses, significant improvements in classification were noted for diseases known to overlap biologically. For example, F1-scores for CMML and MDS classification improved from 0.37 to 0.61 to 0.87 and 0.9, respectively, for the random forest model, and from 0.6 to 0.68 to 0.93 and 0.95, respectively, for the MLP model.

Figure 3 shows the performance of hard and functional classifiers based on recall. For certain classes, hard classification showed poor performance because of the overlap between classes. Functional classification could compensate by offering a second, third, or additional choice. The functional classifier provided much-improved performance on both classes.

We also used confusion matrices for comparing hard classification with functional classification. As shown in supplement B and Fig. 4, more accurate classification is seen with the functional classifications (better precision and recall). To confirm that the functional classification and the true classifications are the same, we applied the generalized McNemar or Stuart-Maxwell test on the confusion matrices to test the null hypothesis and. The P-value of Stuart-Maxwell test was 0.65 for the functional classification and 0.0006 for the hard classification confirming the null hypothesis between functional classification and true classification (supplements C and D). The small p-value in the hard classification indicates a significant difference in distribution of predictions compared

to the true distribution. However, it should be noted that the difference in distributions is not necessarily indicative of a poor classifier. The standard measures for classifier performance include precisions, recalls, and f-1 scores. The similarity between the marginal distributions only reflects one aspect of classifier performance when the prediction accuracy is already high. In addition, a homogeneity test can only give a statistically significant result when the distributions are different. It cannot definitively affirm a good performance by not rejecting the null hypothesis. Figure 4 shows a visualization of one of these classifiers (XGBoost). In this example, the hard classification may misclassify neoplasms known to overlap. For example, MDS is known to overlap with AML and CMML and for this reason, the precision and recall are relatively poor (0.72 and 0.59, respectively). In contrast, the functional classification shows significantly better precision and recall (0.93 and 0.89, respectively). Using functional classification might have clinical implication because therapy of a disease classified as AML will be different from MDS (Fig. 4).

3.3 Examples for functional classification

As an example, we analyzed a sample that was submitted with a pathological diagnosis of ovarian cancer. Random forest classified the sample as endometrial as a first choice, ovarian as a second choice, and breast cancer as a third choice (Fig. 5). Uniform manifold approximation and projection (UMAP) clustering shows significant overlap in clustering between the three diseases, but the sample expression profile showed stronger match with ovarian cancer (Fig. 5A). Similarly, supervised clustering based on expression profile showed better proximity to ovarian cancer but significant overlap with endometrial and breast cancers. Further evaluation of this sample showed overexpression and gene amplification of ERBB2. The functional classification of this cancer as overlapping between ovarian, endometrial and breast cancers helped in guiding the treating physician to incorporate anti-HER2 therapy in the treatment of the patient with this tumor.

It is well established that endometrial and ovarian cancers overlap in many aspects including risk factors, mutation profile and clinical behavior. However, significant differences in tumor histology can be seen between the two neoplasms [27]. Despite the difference, both endometrial and ovarian cancers are commonly treated with platinum- or taxane-based chemotherapies. Furthermore, some ovarian and endometrial cancer share molecular and biological features similar to those seen in breast cancer. The three tumors frequently depend on hormonal factors. The presence of BRCA1 or BRCA2 mutations increases susceptibility to developing cancer in the three organs [28]. Therefore, functional classification of the tumor presented here and the demonstration of similarities and overlapping provide more clinical information and options for therapy to the treating physician.

Another example from hematologic neoplasms is follicular lymphoma that can have indolent clinical behavior or aggressive behavior overlapping with diffuse large B-cell lymphoma (DLBCL). Multiple morphologic subclassifications have been suggested to distinguish between aggressive and less aggressive follicular lymphoma based on the number of large cells under the microscope. This subjective approach is frequently not reproducible. Using expression profiling shows cases of follicular that are clearly distinct from DLBCL. However, some cases overlap with DLBCL (Fig. 6A and B). Individual

Table 3 Results of the machine learning models evaluating the testing set samples for hard classification

	Hard classification											
	GMNB			Kneighbors			GaussianNB			DecisionTree		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
All	0.65	0.81	0.72	0.92	0.75	0.83	0.65	0.81	0.72	0.75	0.56	0.64
AML	0.7	0.3	0.42	0.84	0.54	0.65	0.7	0.3	0.42	0.72	0.66	0.69
Breast	0.93	0.83	0.88	0.91	0.88	0.9	0.93	0.83	0.88	0.88	0.83	0.85
Carcinoma	0.09	0.16	0.12	0.25	0.05	0.09	0.09	0.16	0.12	0.04	0.05	0.04
CLL	0.96	0.78	0.86	0.84	0.91	0.88	0.96	0.78	0.86	0.75	0.83	0.79
CMML	0.23	0.65	0.34	0.7	0.19	0.3	0.23	0.65	0.34	0.31	0.3	0.31
Colorectal	0.88	0.73	0.8	0.76	0.89	0.82	0.88	0.73	0.8	0.76	0.82	0.79
DLBCL	0.69	0.68	0.68	0.72	0.68	0.7	0.69	0.68	0.68	0.75	0.72	0.74
DLBCL.EBV+	0.23	0.63	0.34	0.42	0.56	0.48	0.23	0.63	0.34	0.58	0.41	0.48
DLBCL.EXTRANODAL	0.22	0.17	0.19	0.67	0.6	0.63	0.22	0.17	0.19	0.3	0.33	0.32
Endometrial	0.58	0.5	0.54	0.78	0.6	0.68	0.58	0.5	0.54	0.52	0.47	0.49
Follicular	0.55	0.7	0.62	0.44	0.93	0.59	0.55	0.7	0.62	0.62	0.75	0.68
Hodgkin	0.2	0.43	0.28	0.41	0.54	0.46	0.2	0.43	0.28	0.5	0.5	0.5
Lung	0.92	0.68	0.78	0.89	0.96	0.93	0.92	0.68	0.78	0.87	0.91	0.89
Mantle	0.84	0.46	0.6	0.97	0.52	0.67	0.84	0.46	0.6	0.88	0.82	0.85
Marginal	0.23	0.27	0.25	0.26	0.19	0.22	0.23	0.27	0.25	0.39	0.42	0.41
MDS	0.47	0.3	0.37	0.39	0.37	0.38	0.47	0.3	0.37	0.52	0.48	0.5
MM	0.77	0.46	0.58	0.89	0.43	0.58	0.77	0.46	0.58	0.76	0.59	0.67
Normal	0.68	0.88	0.77	0.62	0.92	0.74	0.68	0.88	0.77	0.75	0.85	0.8
Ovarian	0.57	0.65	0.61	0.67	0.65	0.66	0.57	0.65	0.61	0.5	0.61	0.55
Pancreas	0.23	0.56	0.32	0.64	0.56	0.6	0.23	0.56	0.32	0.63	0.38	0.47
Pedlymphoma	0.53	0.31	0.39	0.64	0.27	0.38	0.53	0.31	0.39	0.41	0.27	0.33
Plasmablasticl	0.69	0.43	0.53	0.78	0.33	0.47	0.69	0.43	0.53	0.5	0.43	0.46
Sarcoma	0.47	0.82	0.6	0.82	0.61	0.7	0.47	0.82	0.6	0.69	0.63	0.66
T-cell	0.14	0.08	0.1	0.7	0.28	0.4	0.14	0.08	0.1	0.35	0.36	0.35
Accuracy	0.6			0.71			0.6			0.7		

Table 3 (continued)

	Hard classification											
	GMNB			Kneighbors			GaussianNB			DecisionTree		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Macro	0.54	0.53	0.51	0.68	0.57	0.59	0.54	0.53	0.51	0.59	0.56	0.57
Weighted	0.68	0.6	0.62	0.72	0.71	0.69	0.68	0.6	0.62	0.7	0.7	0.7
	AdaBoost			RandomForest			MLP			XGB		
All	0	0	0	0.89	0.78	0.83	0.9	0.88	0.89	0.96	0.78	0.86
AML	0.33	0.01	0.01	0.77	0.82	0.79	0.93	0.7	0.8	0.82	0.83	0.83
Breast	0.11	0.73	0.2	0.98	0.88	0.93	0.95	0.87	0.9	0.92	0.93	0.93
Carcinoma	0	0	0	0.5	0.05	0.1	0.33	0.21	0.26	0.6	0.16	0.25
CLL	0	0	0	0.93	0.93	0.93	0.91	0.88	0.89	0.93	0.93	0.93
CMML	0	0	0	0.75	0.24	0.37	0.67	0.54	0.6	0.71	0.41	0.52
Colorectal	0	0	0	0.9	0.9	0.9	0.82	0.95	0.88	0.88	0.93	0.9
DLBCL	0	0	0	0.74	0.92	0.82	0.79	0.81	0.8	0.81	0.92	0.86
DLBCL.EBV+	0	0	0	0.71	0.56	0.63	0.73	0.3	0.42	0.69	0.67	0.68
DLBCL.EXTRANODAL	0	0	0	0.64	0.47	0.54	0.66	0.77	0.71	0.79	0.63	0.7
Endometrial	0.39	0.57	0.46	1	0.5	0.67	0.62	0.87	0.72	0.86	0.6	0.71
Follicular	0	0	0	0.61	0.91	0.73	0.75	0.86	0.8	0.67	0.93	0.78
Hodgkin	0	0	0	0.81	0.79	0.8	0.77	0.71	0.74	0.85	0.79	0.81
Lung	0.54	0.3	0.39	0.83	0.98	0.9	0.95	0.97	0.96	0.93	0.98	0.95
Mantle	0	0	0	1	0.82	0.9	0.96	0.95	0.95	1	0.96	0.98
Marginal	0	0	0	0.8	0.31	0.44	0.52	0.5	0.51	0.73	0.31	0.43
MDS	0	0	0	0.67	0.55	0.61	0.65	0.71	0.68	0.72	0.59	0.65
MM	0	0	0	0.96	0.68	0.79	0.9	0.76	0.82	0.9	0.73	0.81
Normal	0.24	0.98	0.38	0.75	0.95	0.84	0.81	0.94	0.87	0.79	0.95	0.86
Ovarian	0.01	0.03	0.02	0.72	0.74	0.73	0.88	0.68	0.76	0.68	0.87	0.76
Pancreas	0	0	0	0.92	0.34	0.5	0.9	0.59	0.72	0.81	0.53	0.64
Pedlymphoma	0	0	0	0.71	0.19	0.3	0.83	0.38	0.53	0.65	0.42	0.51
Plasmablasticl	0	0	0	0.88	0.67	0.76	0.5	0.86	0.63	0.81	0.81	0.81

Table 3 (continued)

	AdaBoost			RandomForest			MLP			XGB		
Sarcoma	0.2	0.03	0.05	0.71	0.89	0.79	0.87	0.89	0.88	0.78	0.82	0.79
T-cell	0	0	0	0.93	0.52	0.67	0.64	0.72	0.68	0.65	0.68	0.67
Accuracy	0.24			0.8			0.82			0.83		
Macro	0.07	0.11	0.06	0.8	0.66	0.69	0.77	0.73	0.74	0.8	0.73	0.75
Weighted	0.18	0.24	0.15	0.8	0.8	0.78	0.83	0.82	0.82	0.83	0.83	0.82

Table 4 Results of the machine learning models evaluating the testing set samples for functional classification

	GMNB			Kneighbors			GaussianNB			DecisionTree		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
ALL	0.64	0.65	0.64	0.97	0.88	0.92	0.87	0.94	0.91	0.9	0.84	0.87
AML	0.75	0.73	0.74	0.96	0.92	0.94	0.94	0.88	0.91	0.95	0.93	0.94
BREAST	0.77	0.72	0.74	0.97	0.96	0.97	0.98	0.97	0.98	0.98	0.97	0.98
CARCINOMA	0.79	0.67	0.72	0.89	0.77	0.82	0.89	0.76	0.82	0.81	0.79	0.8
CLL	0.62	0.68	0.65	0.9	0.93	0.91	0.98	0.86	0.91	0.85	0.89	0.87
CMML	0.85	0.73	0.78	0.89	0.92	0.91	0.9	0.95	0.93	0.92	0.85	0.89
COLORECTAL	0.78	0.68	0.73	0.93	0.97	0.95	0.97	0.92	0.95	0.93	0.95	0.94
DLBCL	0.64	0.69	0.67	0.93	0.92	0.93	0.94	0.94	0.94	0.94	0.93	0.93
DLBCL.EBV+	0.65	0.69	0.67	0.88	0.85	0.86	0.87	0.93	0.9	0.91	0.84	0.88
DLBCL.EXTRANODAL	0.6	0.72	0.66	0.93	0.9	0.92	0.81	0.86	0.83	0.86	0.79	0.82
ENDOMETRIAL	0.74	0.71	0.73	0.92	0.92	0.92	0.93	0.91	0.92	0.93	0.93	0.93
FOLLICULAR	0.62	0.7	0.66	0.87	0.94	0.9	0.93	0.94	0.94	0.9	0.97	0.93
HODGKIN	0.59	0.71	0.64	0.81	0.88	0.84	0.67	0.9	0.77	0.81	0.86	0.84
LUNG	0.81	0.73	0.77	0.97	0.98	0.98	0.98	0.94	0.96	0.96	0.98	0.97
MANTLE	0.6	0.68	0.64	0.98	0.85	0.91	0.94	0.84	0.89	0.95	0.93	0.94
MARGINAL	0.61	0.68	0.65	0.81	0.8	0.81	0.76	0.8	0.78	0.85	0.79	0.82
MDS	0.81	0.75	0.78	0.93	0.92	0.93	0.95	0.91	0.93	0.9	0.92	0.91
MM	0.7	0.68	0.69	0.93	0.87	0.9	0.91	0.86	0.89	0.93	0.86	0.89
NORMAL	0.86	0.72	0.79	0.91	0.96	0.94	0.94	0.97	0.95	0.95	0.96	0.96
OVARIAN	0.73	0.69	0.71	0.94	0.92	0.93	0.93	0.91	0.92	0.91	0.93	0.92
PANCREAS	0.79	0.69	0.74	0.92	0.88	0.9	0.89	0.86	0.87	0.88	0.87	0.88
PEDLYMPHOMA	0.63	0.75	0.68	0.9	0.87	0.89	0.86	0.88	0.87	0.88	0.82	0.85
PLASMABLASTICL	0.63	0.7	0.66	0.89	0.83	0.86	0.92	0.87	0.9	0.89	0.85	0.87
SARCOMA	0.66	0.7	0.68	0.93	0.89	0.91	0.84	0.95	0.89	0.87	0.9	0.89
T-CELL	0.68	0.73	0.7	0.87	0.82	0.84	0.73	0.79	0.76	0.82	0.79	0.81
Accuracy												
Macro	0.7	0.7	0.7	0.91	0.89	0.9	0.89	0.89	0.89	0.9	0.89	0.89

Table 4 (continued)

	GMNB			Kneighbors			GaussianNB			DecisionTree		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Weighted	0.75	0.71	0.73	0.93	0.93	0.93	0.93	0.92	0.92	0.93	0.93	0.93
	AdaBoost			**RandomForest**			**MLP**			**XGB**		
ALL	0.47	0.47	0.47	0.79	0.78	0.79	0.93	0.94	0.93	0.94	0.9	0.92
AML	0.51	0.53	0.52	0.91	0.89	0.9	0.98	0.95	0.97	0.96	0.97	0.96
BREAST	0.56	0.66	0.61	0.89	0.9	0.9	0.99	0.96	0.98	0.97	0.98	0.98
CARCINOMA	0.59	0.63	0.61	0.82	0.78	0.79	0.84	0.84	0.84	0.86	0.82	0.84
CLL	0.36	0.35	0.36	0.84	0.84	0.84	0.94	0.92	0.93	0.95	0.94	0.94
CMML	0.59	0.41	0.48	0.87	0.88	0.87	0.92	0.94	0.93	0.94	0.92	0.93
COLORECTAL	0.5	0.59	0.54	0.9	0.92	0.91	0.94	0.99	0.96	0.96	0.98	0.97
DLBCL	0.49	0.59	0.53	0.88	0.89	0.89	0.96	0.94	0.95	0.95	0.96	0.96
DLBCL.EBV+	0.49	0.58	0.53	0.77	0.76	0.76	0.93	0.88	0.91	0.92	0.91	0.91
DLBCL.EXTRANODAL	0.53	0.62	0.57	0.84	0.81	0.82	0.91	0.94	0.92	0.93	0.92	0.92
ENDOMETRIAL	0.88	0.85	0.86	0.84	0.85	0.85	0.91	0.97	0.94	0.95	0.93	0.94
FOLLICULAR	0.47	0.55	0.5	0.86	0.9	0.88	0.95	0.97	0.96	0.94	0.98	0.96
HODGKIN	0.5	0.59	0.54	0.79	0.8	0.8	0.92	0.91	0.91	0.94	0.91	0.93
LUNG	0.6	0.65	0.62	0.95	0.94	0.94	0.99	0.99	0.99	0.98	0.99	0.99
MANTLE	0.57	0.61	0.59	0.86	0.83	0.85	0.97	0.97	0.97	0.98	0.97	0.98
MARGINAL	0.52	0.59	0.55	0.78	0.77	0.78	0.88	0.85	0.87	0.88	0.84	0.86
MDS	0.65	0.49	0.55	0.9	0.9	0.9	0.95	0.95	0.95	0.96	0.94	0.95
MM	0.65	0.55	0.6	0.82	0.8	0.81	0.95	0.93	0.94	0.94	0.93	0.93
NORMAL	0.58	0.39	0.46	0.93	0.94	0.94	0.95	0.97	0.96	0.96	0.98	0.97
OVARIAN	0.7	0.8	0.75	0.84	0.85	0.85	0.97	0.92	0.95	0.91	0.94	0.92
PANCREAS	0.54	0.6	0.57	0.84	0.83	0.83	0.96	0.86	0.91	0.93	0.87	0.9
PEDLYMPHOMA	0.51	0.6	0.55	0.77	0.77	0.77	0.96	0.88	0.92	0.9	0.87	0.88
PLASMABLASTICL	0.57	0.63	0.6	0.75	0.74	0.74	0.9	0.96	0.93	0.91	0.92	0.92
SARCOMA	0.53	0.65	0.59	0.8	0.82	0.81	0.94	0.95	0.94	0.91	0.93	0.92
T-CELL	0.66	0.71	0.68	0.78	0.79	0.79	0.88	0.88	0.88	0.89	0.85	0.87

Table 4 (continued)

	AdaBoost			RandomForest			MLP			XGB		
Accuracy												
Macro	0.56	0.59	0.57	0.84	0.84	0.84	0.94	0.93	0.93	0.93	0.93	0.93
Weighted	0.56	0.56	0.56	0.89	0.89	0.89	0.96	0.95	0.95	0.95	0.96	0.95

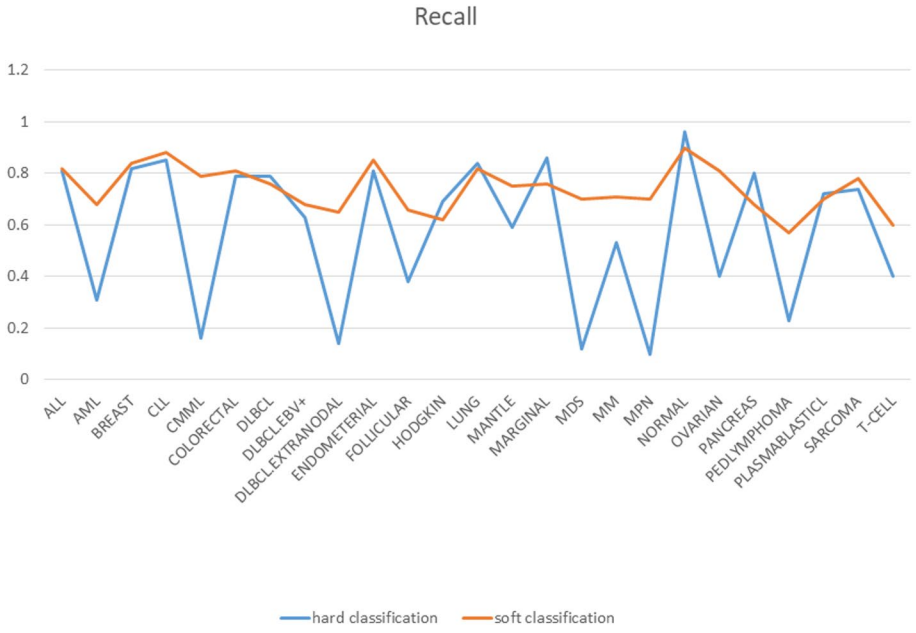

Fig. 3 Performance of hard and functional classifiers based on recall

XGBoost Hard Classification

	ALL	AML	BREAST	CARCINOMA	CLL	CMML	COLORECTAL	DLBCL	DLBCL.EBV+	DLBCL.EXTRANODAL	ENDOMETRIAL	FOLLICULAR	HODGKIN	LUNG	MANTLE	MARGINAL	MDS	MM	NORMAL	OVARIAN	PANCREAS	PEDLYMPHOMA	PLASMABLASTICL	SARCOMA	T-CELL	support	recall
ALL	25	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	1	32	0.78
AML	1	118	0	0	0	2	0	0	0	0	0	0	0	0	0	0	17	0	2	0	0	1	0	1	0	142	0.83
BREAST	0	0	56	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	0	0	1	0	60	0.93
CARCINOMA	0	0	0	3	0	0	0	4	1	1	0	0	0	7	0	0	0	0	1	0	0	0	0	2	0	19	0.16
CLL	0	0	0	0	54	0	0	2	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	58	0.93
CMML	0	7	0	0	0	15	0	0	0	0	0	0	1	0	0	0	0	9	0	5	0	0	0	0	0	0	37	0.41
COLORECTAL	0	0	0	1	0	0	87	0	0	0	0	0	0	3	0	0	0	0	0	1	2	0	0	0	0	94	0.93
DLBCL	0	0	0	0	0	0	0	99	2	0	0	5	0	1	0	1	0	0	0	0	0	0	0	0	0	108	0.92
DLBCL.EBV+	0	0	0	0	0	0	0	3	18	0	0	0	0	1	0	0	0	0	1	0	0	2	1	0	1	27	0.67
DLBCL.EXTRANO	0	1	0	0	0	0	0	3	2	19	0	0	0	0	0	0	0	0	0	1	0	3	1	0	0	30	0.63
ENDOMETRIAL	0	0	2	1	0	0	0	1	0	0	18	0	0	1	0	0	0	0	0	0	6	0	0	0	1	0	30	0.60
FOLLICULAR	0	0	0	0	0	0	0	3	0	0	0	41	0	0	0	0	0	0	0	0	0	0	0	0	0	44	0.93
HODGKIN	0	0	0	0	0	0	0	1	0	0	0	2	22	1	0	0	0	0	0	0	0	0	0	0	2	28	0.79
LUNG	0	0	1	0	0	0	1	0	0	0	2	0	0	336	0	1	0	0	1	0	2	0	0	0	0	344	0.98
MANTLE	0	0	0	0	1	0	0	0	0	0	0	0	0	0	54	0	0	1	0	0	0	0	0	0	0	56	0.96
MARGINAL	0	0	0	0	2	0	1	2	0	0	0	11	0	0	0	8	0	0	1	0	0	0	0	0	1	26	0.31
MDS	0	9	0	0	0	4	0	0	0	0	0	0	0	0	0	0	76	0	37	1	0	0	1	0	0	128	0.59
MM	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	27	7	0	0	0	1	0	0	37	0.73
NORMAL	0	3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	3	1	232	0	0	0	0	1	2	243	0.95
OVARIAN	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	27	0	0	0	1	0	31	0.87
PANCREAS	0	1	1	0	0	0	2	1	0	0	0	0	0	6	0	0	0	0	1	1	17	0	0	2	0	32	0.53
PEDLYMPHOMA	0	1	0	0	0	0	0	6	1	5	0	1	1	0	0	0	0	0	0	0	0	11	0	0	0	26	0.42
PLASMABLASTICL	0	0	0	0	0	0	0	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	17	0	1	21	0.81
SARCOMA	0	0	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	1	0	2	0	0	0	31	1	38	0.82
T-CELL	0	1	1	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	3	0	0	0	0	0	17	25	0.68
sum	26	144	61	5	58	21	99	122	26	24	21	61	26	362	54	11	106	30	294	40	21	17	21	40	26	1716	
precision	0.96	0.82	0.92	0.60	0.93	0.71	0.88	0.81	0.69	0.79	0.86	0.67	0.85	0.93	1.00	0.73	0.72	0.90	0.79	0.68	0.81	0.65	0.81	0.78	0.65		

XGBoost Functional Classification

Cancer type	ALL	AML	BREAST	CARCINOM	CLL	CMML	COLORECT	DLBCL	BCL.EB	EXTRAN	OMETRI	LLICULA	HODGKI	LUNG	MANTLE	ARGINA	MDS	MM	NORMA	OVARIAN	ANCREA	LYMPHO	MABLAS	SARCOM	T-CELL	support	recall
ALL	27.33	0.68	0.01	0.01	0.96	0.03	0.01	0.07	0.05	0.02	0.01	0.08	0.02	0.01	0.02	0.02	0.69	0.07	1.11	0.01	0.01	0.02	0.03	0.02	0.72	32	0.85
AML	0.54	133.04	0.05	0.09	0.25	0.84	0.09	0.17	0.10	0.28	0.10	0.06	0.15	0.20	0.08	0.08	3.26	0.07	1.29	0.07	0.05	0.44	0.09	0.37	0.25	142	0.94
BREAST	0.03	0.02	57.75	0.23	0.02	0.01	0.05	0.02	0.04	0.02	0.09	0.01	0.02	0.65	0.02	0.02	0.04	0.03	0.09	0.22	0.03	0.02	0.02	0.48	0.09	60	0.96
CARCINOMA	0.06	0.06	0.02	18.31	0.05	0.03	1.12	0.55	0.27	0.19	0.03	0.03	0.08	1.11	0.05	0.04	0.03	0.05	0.95	0.05	0.17	0.05	0.13	0.50	0.05	19	0.70
CLL	0.11	0.13	0.04	0.02	53.85	0.04	0.26	1.61	0.02	0.02	0.02	0.23	0.05	0.10	0.05	0.39	0.02	0.04	0.83	0.01	0.03	0.03	0.03	0.02	0.05	58	0.93
CMML	0.10	2.01	0.02	0.02	0.06	30.95	0.01	0.03	0.02	0.02	0.01	0.44	0.08	0.06	0.04	0.04	2.04	0.12	0.70	0.04	0.05	0.02	0.02	0.04	0.05	37	0.84
COLORECTAL	0.03	0.05	0.09	0.51	0.04	0.02	90.29	0.05	0.10	0.12	0.11	0.03	0.06	1.01	0.06	0.06	0.03	0.04	0.02	0.16	0.59	0.06	0.06	0.38	0.04	94	0.96
DLBCL	0.18	0.05	0.06	0.16	0.22	0.05	0.12	100.77	0.86	0.30	0.05	2.57	0.27	0.52	0.10	0.59	0.04	0.20	0.11	0.07	0.08	0.24	0.15	0.10	0.15	108	0.93
DLBCL.EBV+	0.08	0.04	0.03	0.17	0.04	0.05	0.02	0.76	22.21	0.45	0.02	0.04	0.07	0.70	0.04	0.03	0.08	0.19	0.28	0.02	0.03	0.92	0.42	0.10	0.18	27	0.82
DLBCL.EXTRANO	0.08	0.32	0.04	0.12	0.05	0.03	0.03	1.13	0.70	24.38	0.03	0.10	0.06	0.13	0.06	0.05	0.07	0.06	0.05	0.45	0.02	1.18	0.73	0.10	0.04	30	0.81
ENDOMETRIAL	0.03	0.03	0.36	0.38	0.02	0.01	0.55	0.02	0.02	0.05	25.83	0.01	0.02	0.53	0.03	0.02	0.02	0.02	0.01	1.24	0.07	0.03	0.03	0.61	0.04	30	0.86
FOLLICULAR	0.02	0.02	0.01	0.01	0.06	0.01	0.05	1.47	0.02	0.03	0.01	41.89	0.03	0.04	0.01	0.06	0.01	0.02	0.01	0.01	0.01	0.13	0.01	0.01	0.02	44	0.95
HODGKIN	0.11	0.10	0.06	0.08	0.13	0.05	0.14	0.61	0.11	0.08	0.05	0.71	23.51	0.50	0.10	0.40	0.08	0.07	0.21	0.10	0.10	0.05	0.11	0.17	0.38	28	0.84
LUNG	0.09	0.20	0.24	0.08	0.07	0.07	0.55	0.12	0.07	0.10	0.55	0.12	0.16	338.75	0.11	0.50	0.21	0.12	0.65	0.19	0.58	0.09	0.09	0.22	0.08	344	0.98
MANTLE	0.02	0.04	0.02	0.01	0.78	0.01	0.03	0.22	0.01	0.01	0.02	0.12	0.10	0.04	53.67	0.05	0.02	0.33	0.43	0.02	0.01	0.01	0.01	0.01	0.02	56	0.96
MARGINAL	0.07	0.03	0.04	0.03	1.08	0.04	0.65	0.95	0.06	0.03	0.03	2.07	0.11	0.14	0.11	18.71	0.07	0.05	1.20	0.03	0.15	0.03	0.03	0.04	0.25	26	0.72
MDS	0.14	1.88	0.04	0.07	0.13	1.02	0.05	0.05	0.07	0.18	0.04	0.08	0.14	0.15	0.07	0.10	113.93	0.06	7.99	0.88	0.06	0.12	0.49	0.10	0.15	128	0.89
MM	0.05	0.15	0.03	0.04	0.05	0.02	0.03	0.28	0.03	0.06	0.06	0.05	0.07	0.10	0.08	0.04	0.10	32.76	2.35	0.12	0.04	0.03	0.34	0.07	0.04	37	0.89
NORMAL	0.09	0.86	0.03	0.05	0.07	0.12	0.56	0.09	0.14	0.04	0.03	0.11	0.10	0.15	0.06	0.07	1.55	0.47	236.11	0.04	0.38	0.05	0.08	0.84	0.92	243	0.97
OVARIAN	0.04	0.06	0.31	0.02	0.03	0.02	0.19	0.12	0.04	0.04	0.55	0.20	0.23	0.33	0.09	0.12	0.09	0.10	0.04	27.92	0.09	0.04	0.03	0.21	0.07	31	0.90
PANCREAS	0.08	0.64	0.42	0.18	0.05	0.03	0.86	0.38	0.04	0.05	0.13	0.11	0.05	1.30	0.08	0.26	0.22	0.07	0.38	0.58	25.36	0.04	0.06	0.56	0.05	32	0.79
PEDLYMPHOMA	0.12	0.57	0.01	0.02	0.03	0.02	0.06	2.04	1.17	1.89	0.02	0.38	0.37	0.03	0.03	0.07	0.06	0.04	0.05	0.02	0.03	18.49	0.28	0.16	0.04	26	0.71
PLASMABLASTICL	0.02	0.20	0.02	0.07	0.01	0.03	0.02	0.01	0.59	0.04	0.01	0.01	0.09	0.86	0.02	0.01	0.03	0.25	0.06	0.02	0.03	0.12	18.03	0.05	0.39	21	0.86
SARCOMA	0.05	0.04	0.05	0.05	0.12	0.02	0.80	0.07	0.05	0.13	0.24	0.02	0.06	0.44	0.04	0.03	0.02	0.42	0.03	0.51	0.04	0.04	0.20	34.18	0.33	38	0.90
T-CELL	0.10	0.56	0.67	0.09	0.09	0.10	0.16	0.12	0.08	0.08	0.06	0.09	0.42	1.14	0.08	0.04	0.14	0.07	1.89	0.07	0.06	0.05	0.08	0.07	18.69	25	0.75
sum	29.57	141.80	60.41	15.81	58.26	33.62	96.70	111.71	26.88	28.63	28.08	49.60	26.32	348.98	55.09	21.80	122.89	35.72	256.85	32.86	28.06	22.30	21.56	39.43	23.08	1716	
precision	0.92	0.94	0.96	0.84	0.92	0.92	0.93	0.90	0.83	0.85	0.92	0.84	0.89	0.97	0.97	0.86	0.93	0.92	0.92	0.85	0.90	0.83	0.84	0.87	0.81		

Fig. 4 Confusion matrix for hard classification (Top) and functional classification (Bottom) displaying the accuracy of diagnostic classifications using XGBoost. The diagonal in the top matrix represents the rate of correct prediction of each class while the diagonal in the bottom (functional classification) represents the proportion of correct prediction

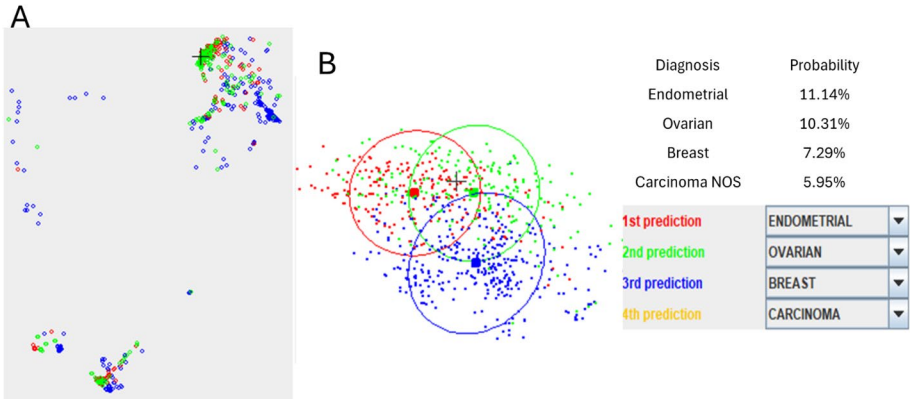

Fig. 5 Ovarian cancer as an example for functional classification. Random forest prediction probability shows endometrial cancer (red) at 11.14%, Ovarian cancer (green) at 10.31%, Breast cancer (blue) at 7.29% and Carcinoma nor otherwise classified (Yellow) at 5.95%. Panel A shows the uniform manifold approximation and projection (UMAP) clustering of the training samples from endometrial(red), Ovarian (green), and Breast (blue) and the tested sample as crosshair. The tested sample (presented as crosshair) is clustering with Ovarian cancer (green). Panel B shows a supervised clustering of the three diagnosis and the tested sample is closely clustering with ovarian cancer. Both supervised and unsupervised clustering show significant overlap between the three cancer types

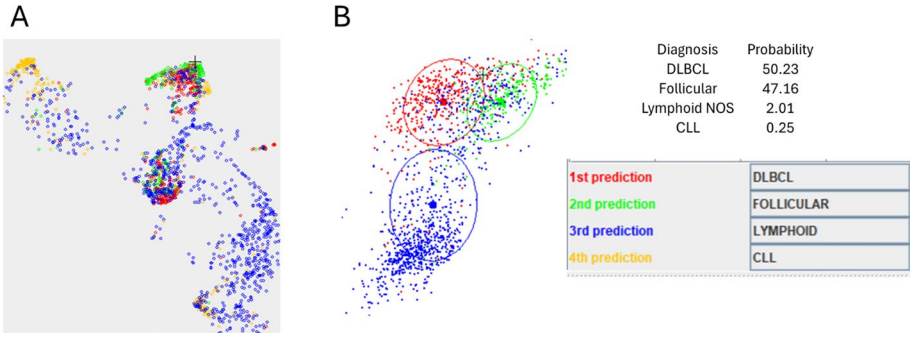

Fig. 6 Follicular lymphoma as an example for functional classification. Random forest prediction probability shows high probability of DLBCL (red) at 50.23% and follicular lymphoma (green) at 47.16%. lymphoid neoplasm, not otherwise specified (NOS) (blue) was distant at probability of 2.01% and chronic lymphocytic leukemia (CLL) at extremely low probability of 0.25%. Panel A shows the uniform manifold approximation and projection (UMAP) clustering of the training samples from the four classes and the tested sample as crosshair. The tested sample (presented as crosshair) appears to cluster with cases overlapping between DLBCL and follicular (red and green). Panel B shows a supervised clustering of the three diagnosis (Follicular, DLBCL and lymphoid, NOS) and the tested sample is closely clustering between DLBCL and follicular

follicular lymphoma can be tested using functional classification to determine proximity to DLBCL which will dictate completely different therapy and different prognosis (Fig. 6).

3.4 Limited overfitting using Bayes models

Comparing the results of training and testing sets showed significant overfitting in the training set for all machine learning models except for GMNB and GNB (Supplement A, Table S1). For GMNB, the average F1-scores for hard and functional classifications were 0.68 and 0.74, respectively, in the training set compared to 0.62 and 0.73 in the testing set. In contrast, with the random forest classifier, F1-scores for hard and functional classification were 1 and 0.96, respectively, for the training set compared to 0.78 and 0.89 for the testing set. These findings suggest that, for small sample sizes, Bayesian statistics represent a reliable and acceptable approach for developing machine learning

prediction. GNB and GMNB avoid overfitting because of their simplicity and because of considering each parameter independently. Therefore, noise generating some features does not influence the overall prediction.

3.5 Overlap between mutation profile and RNA expression profile

Mutation profiles, though not tumor-specific, are frequently helpful for clinical evaluation of tumors. Frequently, the presence of the same mutation in two different tumor types may imply similar responses to targeted therapy. The frequencies of mutations vary between tumors, and the presence of a specific mutation may guide therapy and classification. Supplementary Table S2 shows the frequencies of various mutations by tumor type. In principle, these mutations lead to a change in RNA expression of downstream genes.

We explored whether integrating mutation profiles would improve the clinical value of the diagnostic classification. Focusing on breast, lung, colorectal and ovarian cancers, we first calculated the probability of each diagnostic class, then combined this probability score with the probability from the RNA classification, for both hard and functional classification approaches. As shown in Table 5, the performance of almost all classifications was poor when DNA only was used, and adding DNA data to RNA data in hard classification decreased classification accuracy. The performance of DNA classification improved significantly in functional classification. However, the addition of DNA to RNA did not significantly affect the overall classification of tumors compared to RNA alone. The Confusion matrix of this 4-classes model for GMNB, XGBoost, and MLP classifiers are shown in supplementary Table S3. Table 6 shows the 12 breast cancer cases with hard classification errors using GMNB on RNA. It also shows that with functional classification, the breast class still receives significant probability estimates in most of cases (Table 6).

4 Discussion

The availability of NGS data, especially RNA data, holds great potential for medical applications. NGS data can be quantified reliably and reproducibly, can be used to detect differentially expressed transcripts, and can facilitate further functional/pathway analysis [29]. Furthermore, RNA data from cancer specimens reflect not only the tumor but also the microenvironment. In principle, RNA data can provide a rich medium for developing new biomarkers for predicting not only specific diagnosis, but also response to therapy, the development of side effects, and other clinically relevant indicators [29].

The analysis of RNA data presents unique challenges. In clinical applications, a specimen is usually not homogeneous but a mixture of different cell types, some of which are reactive to the tumor (microenvironment) and some of which reflect normal local tissue. For example, tumor cells may account for 30% of cells in one tumor specimen but 80% in the next. Furthermore, binary classification is frequently inadequate in medical practice. For example, clinical response may need to be classified as complete response, partial response, no response, or progression, rather than simply yes or no. Given this biological complexity, particularly in cancer, proper utilization of the high-dimensional data generated by NGS requires help from machine learning or other forms of AI. Modern advancements in machine learning technology allow the construction of classifiers directly from NGS data, despite their complexity and vast volume. The direct approach

Table 5 Results of integrating DNA data with RNA data in classifying testing set of lung, breast, colorectal and ovarian cancers

RNA	RNA			DNA			RNA+DNA		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Naïve Bayes									
Hard classification									
Breast	0.96	0.80	0.87	0.35	0.56	0.43	0.69	0.77	0.72
Colorectal	0.95	0.79	0.86	0.47	0.63	0.54	0.78	0.83	0.80
Lung	0.90	0.98	0.94	0.80	0.37	0.50	0.90	0.87	0.89
Ovarian	0.88	0.74	0.81	0.09	0.44	0.15	0.63	0.61	0.62
Macro avg	0.92	0.83	0.87	0.43	0.50	0.41	0.75	0.77	0.76
Wght avg	0.92	0.91	0.91	0.64	0.44	0.48	0.84	0.84	0.84
Functional classification									
Breast	0.82	0.85	0.83	0.81	0.83	0.82	0.82	0.86	0.84
Colorectal	0.77	0.81	0.79	0.76	0.75	0.75	0.76	0.82	0.79
Lung	0.91	0.85	0.88	0.92	0.74	0.82	0.92	0.84	0.88
Ovarian	0.75	0.82	0.78	0.71	0.75	0.73	0.74	0.81	0.77
Macro avg	0.81	0.83	0.82	0.80	0.77	0.78	0.81	0.83	0.82
Wght avg	0.87	0.84	0.85	0.87	0.75	0.80	0.87	0.84	0.85
Random forest									
Hard classification									
Breast	0.96	0.87	0.91	0.35	0.56	0.43	0.80	0.88	0.84
Colorectal	0.99	0.90	0.94	0.47	0.63	0.54	0.84	0.90	0.87
Lung	0.94	0.99	0.97	0.80	0.37	0.50	0.95	0.92	0.93
Ovarian	1.00	0.84	0.91	0.09	0.44	0.15	0.72	0.68	0.70
Macro avg	0.97	0.90	0.93	0.43	0.50	0.41	0.83	0.85	0.84
Wght avg	0.96	0.95	0.95	0.64	0.44	0.48	0.90	0.90	0.90
Functional classification									
Breast	0.81	0.85	0.83	0.81	0.83	0.82	0.81	0.86	0.83
Colorectal	0.78	0.84	0.81	0.76	0.75	0.75	0.77	0.84	0.81
Lung	0.94	0.85	0.89	0.92	0.74	0.82	0.94	0.84	0.88
Ovarian	0.75	0.84	0.79	0.71	0.75	0.73	0.74	0.82	0.78
Macro avg	0.82	0.85	0.83	0.80	0.77	0.78	0.81	0.84	0.82
Wght avg	0.89	0.85	0.86	0.87	0.75	0.80	0.88	0.84	0.86
XGBoost									
Hard classification									
Breast	0.98	0.93	0.96	0.35	0.56	0.43	0.95	0.93	0.94
Colorectal	0.99	0.96	0.97	0.47	0.63	0.54	0.99	0.96	0.97
Lung	0.98	0.99	0.98	0.80	0.37	0.50	0.98	0.99	0.98
Ovarian	0.88	0.94	0.91	0.09	0.44	0.15	0.87	0.87	0.87
Macro avg	0.96	0.95	0.95	0.43	0.50	0.41	0.95	0.94	0.94
Wght avg	0.97	0.97	0.97	0.64	0.44	0.48	0.97	0.97	0.97
Functional classification									
Breast	0.99	0.98	0.98	0.81	0.83	0.82	0.94	0.96	0.95
Colorectal	0.99	0.98	0.98	0.76	0.75	0.75	0.94	0.96	0.95
Lung	0.99	1.00	0.99	0.92	0.74	0.82	0.99	0.97	0.98
Ovarian	0.96	0.97	0.96	0.71	0.75	0.73	0.87	0.93	0.90
Macro avg	0.98	0.98	0.98	0.80	0.77	0.78	0.93	0.96	0.94
Wght avg	0.99	0.99	0.99	0.87	0.75	0.80	0.97	0.96	0.96

has the advantage that it is completely data driven and is not dependent on known models and pathways.

In this study, we employed NGS data sets containing RNA expression values of 1408 genes used in analyzing 25 different hematologic and solid tumors. Classes with small sample sizes were removed. The data set was split into training (two-thirds) and testing

Table 6 List of the 12 misclassified breast cancer by hard classification using GMNB

	Breast	Colorectal	Lung	Ovarian
Breast	0.23427	0.201722	0.301188	0.26282
Breast	0.269721	0.207395	0.337244	0.18564
Breast	0.237564	0.190036	0.369199	0.203202
Breast	3.34E-20	0.010225	0.958313	0.031462
Breast	0.254817	0.222705	0.33015	0.192328
Breast	0.267312	0.212982	0.271935	0.247771
Breast	0.432874	0.063355	0.433758	0.070013
Breast	0.312399	0.171057	0.352339	0.164204
Breast	5.94E-04	0.39627	4.50E-06	0.603131
Breast	0.029568	0.017074	0.510057	0.4433
Breast	0.267314	0.212981	0.271934	0.247771
Breast	0.230374	0.271665	0.479387	0.018574

Most of the cases show that breast functional classification still receives significant probability estimates

(one-third) sets. The classification performance was measured on the testing set. We compared eight different machine learning AI models to differentiate among diagnostic classes. Because of the lack of alternatives at this time, we used conventional diagnostic classes, which are based on site of origin and morphologic evaluation. We named this diagnostic approach hard classification. As shown in Table 4, except for the AdaBoost algorithm, all models performed well for hard classification. However, the MLP, random forest, and XGB classifiers performed the best. Bayes-based classifications (GMNB and GNB) allowed us to evaluate the contribution of each of the biomarkers (genes) used in these classifications (data not presented). In contrast, all other classifiers provided limited information on the contributions of the various biomarkers to classification (black box).

As expected, the prediction of diagnosis was poor in diseases known to overlap with others, such as CMML and MDS, extranodal DLBCL and pediatric lymphoma, and carcinoma (tumor of unknown origin). This raises the possibility that such entities are not completely discrete and overlap with other diagnostic classes, supporting the need for functional classification. Functional classification provides probability estimates for each classification, rather than a binary 0 vs. 1 result, and considers first and close classifications as acceptable. Close probability of a class more accurately defines the overlap between classes and, most likely, reflects the biology of the diseases. For example, we know that, in some patients, CMML may represent and have a clinical course similar to that of MDS. The probability score for a specific case reflects how close that case is to MDS or CMML and provides insight on management options. Similarly, for a case classified as carcinoma, the probability estimates for the classifications ranked as second or third should be used to inform management. We illustrate the importance of such overlap by presenting a case in which a tumor diagnosed by pathologic examination as ovarian cancer. Our random forest predicted endometrial cancer as first choice, ovarian as second and breast as third choice (Fig. 5). Unsupervised and supervised clustering showed significant overlapping between the three diagnostic classes, but the sample was in proximity of ovarian cancer (Fig. 5). Further evaluation of the sample showed ERBB2 RNA overexpression and gene amplification. The overlap with breast cancer diagnosis led to a therapeutic approach that incorporates ant-HER2 therapy. Similarly in hematologic neoplasms, overlapping between advanced cases of follicular lymphoma and DLBCL can be very significant with implication for therapy. There is significant

variability between pathologists in classifying such cases. As shown in Fig. 6, functional classification can provide mor objective classification of such cases.

Clearly, the level of difference in probability scores is important in measuring the proximity between two different diagnoses. For example, the similarity between two diagnostic classes for a case with scores of 40% vs. 39% is completely different from two classes with scores of 40% vs. 20%. Therapy using either the first or second diagnostic class might be acceptable in the first example but may not be acceptable for the second example. Clinical trials for treatment and management are needed to determine which level of proximity is acceptable for demonstrating clinical similarity and therapy selection.

Our data also demonstrates that, except for Bayes-based algorithms, almost all machine learning systems tend to overfit in the training set and should not be used without training and testing sets (Table S1).

The approaches described here are reliable based solely on RNA expression profiles and do not need to integrate mutation profiles. Although RNA expression profiles most likely reflect DNA mutation profiles, using DNA mutation profiles for hard or functional classification—alone or in combination with RNA expression profiles–did not add significant biological value (Table 5). Mutations in principle lead to downstream changes in RNA expression, therefore, they are redundant in any algorithm. Furthermore, the same gene, such as in TP53 mutations, may play tumor suppressing function in certain environment and oncogenic function in a different environment. This may compromise its prediction functionality. Our demonstration that DNA mutations are not good for cancer classification suggests that even targeted therapy should be based on RNA expression profiles and not DNA mutation profile. Demonstrating that two tumors share the same mutation(s) is not sufficient to assess whether they might respond to the same targeted treatment; they must show similarity in their overall RNA expression profiles.

Variability in diagnosis of cancer and misdiagnosis is, in part, due to biological overlap that morphology-based is incapable of distinguishing. While our algorithms can provide acceptable Hard classification imitating site and morphology-based classification, the misclassified cases do not necessarily reflect poor performance as much as a reflection of biology. Confusion matrices (Supplement B and Fig. 4) illustrate significantly more accuracy (precision and recall) when functional classification is used (Table 4). This approach is less subjective and more representative of cancer biology. The novelty of this approach is now possible due to the recent advances in AI and in molecular testing and our ability of accurately and reproducibly quantify RNA.

Although the current study indirectly confirm the reliability of our RNA quantification using targeted RNA sequencing, additional studies are needed to explore the interchangeability between the use of different techniques in RNA sequencing and quantification. More importantly, clinical data exploring the clinical value of using functional classification is needed.

In summary, this study shows that RNA profile in combination with appropriate AI algorithms can provide remarkably reliable diagnostic classification, relative to that obtained with conventional classification based on site of origin and morphology. However, this genomic data can also provide information on the similarities and overlaps between diagnostic classes. We propose that determining these similarities can be used clinically to determine therapeutic approaches and management of patients. While the approach described here is used for determining diagnosis, it can also be applied to predict treatment responses and

assess level of response (i.e., determine complete response vs. partial response, no response, or progression) if appropriate response data are available for training. Our study challenges the current World Health Organization (WHO) diagnosis and classification of cancers based on site of origin and morphology, but we believe this study contributes to the current efforts in developing better care and improving outcome for patients with cancer.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s44163-026-01200-8.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

Supplementary Material 4.

Author contributions

MA, HZ, processed, analyzed, and interpreted data. JM, DS, MG, JK, AP, SA, AC, MD, AI, and AG provided data, wrote, edited and critically reviewed manuscript.

Funding

No funding was required.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

The study protocol was approved by the Western Copernicus Group Institutional Review Board (New England IRB, Aspire IRB, and Midlands IRB) (Number 1-1476184-1). Patient informed consent was waived due to incidental collection and lack of risk. This study was conducted in accordance with the principles of the Declaration of Helsinki and its later amendments.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 3 February 2025 / Accepted: 20 March 2026

Published online: 01 July 2026

References

1. Tsang JYS, Tse GM. Molecular classification of breast cancer. Adv Anat Pathol. 2020;27(1):27–35. https://doi.org/10.1097/PAP.0000000000000232.
2. Soomro SH, Ting LR, Qing YY, Ren M. Molecular biology of glioblastoma: Classification and mutational locations. J Pak Med Assoc. 2017;67(9):1410–4.
3. Travis WD, Brambilla E, Noguchi M, Nicholson AG, Geisinger KR, Yatabe Y, Beer DG, Powell CA, Riely GJ, Van Schil PE, Garg K, Austin JH, Asamura H, Rusch VW, Hirsch FR, Scagliotti G, Mitsudomi T, Huber RM, Ishikawa Y, Jett J, Sanchez-Cespedes M, Sculier JP, Takahashi T, Tsuboi M, Vansteenkiste J, Wistuba I, Yang PC, Aberle D, Brambilla C, Flieder D, Franklin W, Gazdar A, Gould M, Hasleton P, Henderson D, Johnson B, Johnson D, Kerr K, Kuriyama K, Lee JS, Miller VA, Petersen I, Roggli V, Rosell R, Saijo N, Thunnissen E, Tsao M, Yankelewitz D. International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. J Thorac Oncol. 2011;6(2):244–85. https://doi.org/10.1097/JTO.0b013e318206a221.
4. Loghavi S, Kanagal-Shamanna R, Khoury JD, Medeiros LJ, Naresh KN, Nejati R, Patnaik MM. WHO 5th Edition Classification Project. Fifth Edition of the World Health Classification of Tumors of the Hematopoietic and Lymphoid Tissue: Myeloid Neoplasms. Mod Pathol. 2024;37(2):100397. https://doi.org/10.1016/j.modpat.2023.100397.
5. Komori T, Komori T. The 2021 WHO Classification of Tumors, 5th edition, Central Nervous System Tumors: A Short Review. Brain Nerve. 2022;74(6):803–9. https://doi.org/10.11477/mf.1416202124.
6. Fanale D, Corsini LR, Scalia R, Brando C, Cucinella A, Madonia G, Dimino A, Filorizzo C, Barraco N, Bono M, Fiorino A, Magrin L, Sciacchitano R, Perez A, Russo TDB, Pantuso G, Russo A, Bazan V. Can the tumor-agnostic evaluation of MSI/MMR status be the common denominator for the immunotherapy treatment of patients with several solid tumors? Crit Rev Oncol Hematol. 2022;170:103597. https://doi.org/10.1016/j.critrevonc.2022.103597.
7. Schram AM, Colombo N, Arrowsmith E, Narayan V, Yonemori K, Scambia G, Zelnak A, Bauer TM, Jin N, Ulahannan SV, Colleoni M, Aftimos P, Donoghue MTA, Rosen E, Rudneva VA, Telli ML, Domchek SM, Galsky MD, Hoyle M, Chappey C, Stewart

R, Blake-Haskins JA, Yap TA. Avelumab Plus Talazoparib in Patients With BRCA1/2- or ATM-Altered Advanced Solid Tumors: Results From JAVELIN BRCA/ATM, an Open-Label, Multicenter, Phase 2b, Tumor-Agnostic Trial. JAMA Oncol. 2023;9(1):29–39. https://doi.org/10.1001/jamaoncol.2022.5218.
8. Gouda MA, Nelson BE, Buschhorn L, Wahida A, Subbiah V. Tumor-Agnostic Precision Medicine from the AACR GENIE Database: Clinical Implications. Clin Cancer Res. 2023;29(15):2753–60. https://doi.org/10.1158/1078-0432.CCR-23-0090.
9. Naito Y, Mishima S, Akagi K, Igarashi A, Ikeda M, Okano S, Kato S, Takano T, Tsuchihara K, Terashima K, Nishihara H, Nishiyama H, Hiyama E, Hirasawa A, Hosoi H, Maeda O, Yatabe Y, Okamoto W, Ono S, Kajiyama H, Nagashima F, Hatanaka Y, Miyachi M, Kodera Y, Yoshino T, Taniguchi H. Japan society of clinical oncology/Japanese society of medical oncology-led clinical recommendations on the diagnosis and use of tropomyosin receptor kinase inhibitors in adult and pediatric patients with neurotrophic receptor tyrosine kinase fusion-positive advanced solid tumors, cooperated by the Japanese society of pediatric hematology/oncology. Int J Clin Oncol. 2020;25(3):403–17. https://doi.org/10.1007/s10147-019-01610-y.
10. Tan O, Shrestha R, Cunich M, Schofield DJ. Application of next-generation sequencing to improve cancer management: A review of the clinical effectiveness and cost-effectiveness. Clin Genet. 2018;93(3):533–44. https://doi.org/10.1111/cge.13199.
11. Ahmad E, Ali A, Nimisha, Kumar Sharma A, Ahmed F, Mehdi Dar G, Mohan Singh A, Apurva, Kumar A, Athar A, Parveen F, Mahajan B, Singh Saluja S. Molecular approaches in cancer. Clin Chim Acta. 2022;537:60–73. https://doi.org/10.1016/j.cca.2022.09.027.
12. Bewicke-Copley F, Arjun Kumar E, Palladino G, Korfi K, Wang J. Applications and analysis of targeted genomic sequencing in cancer studies. Comput Struct Biotechnol J. 2019;17:1348–59. https://doi.org/10.1016/j.csbj.2019.10.004.
13. Tarantino P, Carmagnani Pestana R, Corti C, Modi S, Bardia A, Tolaney SM, Cortes J, Soria JC, Curigliano G. Antibody-drug conjugates: Smart chemotherapy delivery across tumor histologies. CA Cancer J Clin. 2022;72(2):165–82. https://doi.org/10.3322/caac.21705.
14. Pan C, Liu H, Robins E, Song W, Liu D, Li Z, Zheng L. Next-generation immuno-oncology agents: current momentum shifts in cancer immunotherapy. J Hematol Oncol. 2020;13(1):29. https://doi.org/10.1186/s13045-020-00862-w.
15. Zhang H, Qureshi MA, Wahid M, Charifa A, Ehsan A, Ip A, De Dios I, Ma W, Sharma I, McCloskey J, Donato M, Siegel D, Gutierrez M, Pecora A, Goy A, Albitar M. Differential Diagnosis of Hematologic and Solid Tumors Using Targeted Transcriptome and Artificial Intelligence. Am J Pathol. 2023;193(1):51–9. https://doi.org/10.1016/j.ajpath.2022.09.006.
16. Zaitsev A, Chelushkin M, Dyikanov D, Cheremushkin I, Shpak B, Nomie K, Zyrin V, Nuzhdina E, Lozinsky Y, Zotova A, Degryse S, Kotlov N, Baisangurov A, Shatsky V, Afenteva D, Kuznetsov A, Paul SR, Davies DL, Reeves PM, Lanuti M, Goldberg MF, Tazearslan C, Chasse M, Wang I, Abdou M, Aslanian SM, Andrewes S, Hsieh JJ, Ramachandran A, Lyu Y, Galkin I, Svekolkin V, Cerchietti L, Poznansky MC, Ataullakhanov R, Fowler N, Bagaev A. Precise reconstruction of the TME using bulk RNA-seq and a machine learning algorithm trained on artificial transcriptomes. Cancer Cell. 2022;40(8):879–e89416. https://doi.org/10.1016/j.ccell.2022.07.006.
17. Kebschull M, Kroeger AT, Papapanou PN. Differential expression, functional and machine learning analysis of high-throughput -omics data using open-source tools. Methods Mol Biol. 2023;2588:317–51. https://doi.org/10.1007/978-1-0716-2780-8_19.
18. Sampat MP, Patel AC, Wang Y, Gupta S, Kan CW, Bovik AC, Markey MK. Indexes for three-class classification performance assessment–an empirical comparison. IEEE Trans Inf Technol Biomed. 2009;13(3):300–12. https://doi.org/10.1109/TITB.2008.2009440.
19. Altman N, Krzywinski M. The curse(s) of dimensionality. Nat Methods. 2018;15(6):399–400. https://doi.org/10.1038/s41592-018-0019-x.
20. Xu-Monette ZY, Zhang H, Zhu F, Tzankov A, Bhagat G, Visco C, Dybkaer K, Chiu A, Tam W, Zu Y, Hsi ED, You H, Huh J, Ponzoni M, Ferreri AJM, Møller MB, Parsons BM, van Krieken JH, Piris MA, Winter JN, Hagemeister FB, Shahbaba B, De Dios I, Zhang H, Li Y, Xu B, Albitar M, Young KH. A refined cell-of-origin classifier with targeted NGS and artificial intelligence shows robust predictive value in DLBCL. Blood Adv. 2020;4(14):3391–404. https://doi.org/10.1182/bloodadvances.2020001949.
21. Kruschke JK. Bayesian Analysis Reporting Guidelines. Nat Hum Behav. 2021;5(10):1282–91. https://doi.org/10.1038/s41562-021-01177-7.
22. Alharbi F, Vakanski A. Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review. Bioeng (Basel). 2023;10(2):173. https://doi.org/10.3390/bioengineering10020173.
23. Mayr A, Binder H, Gefeller O, Schmid M. The evolution of boosting algorithms. From machine learning to statistical modelling. Methods Inf Med. 2014;53(6):419–27. https://doi.org/10.3414/ME13-01-0122.
24. Abu Alfeilat HA, Hassanat ABA, Lasassmeh O, Tarawneh AS, Alhasanat MB, Eyal Salman HS, Prasath VBS. Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review. Big Data. 2019;7(4):221–48. https://doi.org/10.1089/big.2018.0175.
25. Awaysheh A, Wilcke J, Elvinger F, Rees L, Fan W, Zimmerman KL. Review of Medical Decision Support and Machine-Learning Methods. Vet Pathol. 2019;56(4):512–25. https://doi.org/10.1177/0300985819829524.
26. Munir K, Elahi H, Ayub A, Frezza F, Rizzi A. Cancer Diagnosis Using Deep Learning: A Bibliographic Review. Cancers (Basel). 2019;11(9):1235. https://doi.org/10.3390/cancers11091235.
27. Glubb DM, Thompson DJ, Aben KK, Alsulimani A, Amant F, Annibali D, et al. Cross-Cancer genome-wide association study of endometrial cancer and epithelial ovarian cancer identifies genetic risk regions associated with risk of both cancers. Cancer Epidemiol Biomarkers Prev. 2021;30(1):217–28.
28. Nasioudis D, Gysler S, Latif N, Cory L, Giuntoli RL, Kim SH, Simpkins F, Martin L, Ko EM. Molecular landscape of ERBB2/HER2 gene amplification among patients with gynecologic malignancies; clinical implications and future directions. Gynecol Oncol. 2024;180:1–5. https://doi.org/10.1016/j.ygyno.2023.11.021.
29. Zoabi Y, Shomron N. Processing and analysis of RNA-seq data from public resources. Methods Mol Biol. 2021;2243:81–94. https://doi.org/10.1007/978-1-0716-1103-6_4.

30. Satam H, Joshi K, Mangrolia U, Waghoo S, Zaidi G, Rawool S, Thakare RP, Banday S, Mishra AK, Das G, Malonia SK. Next-Generation Sequencing Technology: Current Trends and Advancements. Biology (Basel). 2023;12(7):997. https://doi.org/10.3390/biology12070997.

Publisher's Note