BrcaDx: Precise identification of breast cancer from expression data using a minimal set of features

Sangeetha Muthamilselvan[1], Ashok Palaniappan[1*]

[1]Department of Bioinformatics, School of Chemical and Biotechnology, SASTRA deemed University, Thanjavur Tamilnadu 613401. India

[*]Corresponding author: apalania@scbt.sastra.edu

Abstract

Background: Breast cancer is the foremost cancer in worldwide incidence, surpassing lung cancer notwithstanding the gender bias. One in four cancer cases among women are attributable to cancers of the breast, which are also the leading cause of death in women. Reliable options for early detection of breast cancer are needed.

Methods: Using public-domain datasets, we screened transcriptomic profiles of breast cancer samples, and identified progression-significant linear and ordinal model genes using stage-informed models. We then applied a sequence of machine learning techniques, namely feature selection, principal components analysis, and k-means clustering, to train a learner to discriminate 'cancer' from 'normal' based on expression levels of identified biomarkers.

Results: Our computational pipeline yielded an optimal set of nine biomarker features for training the learner, namely NEK2, PKMYT1, MMP11, CPA1, COL10A1, HSD17B13, CA4, MYOC, and LYVE1. Validation of the learned model on an internal testset yielded a performance of 99.5% accuracy. Blind validation on an external dataset yielded a balanced accuracy of 95.5%, demonstrating that the model has effectively reduced the dimensionality of the

problem, and learnt the solution. The model was rebuilt using the full dataset, and then deployed as a web app for non-profit purposes at: https://apalania.shinyapps.io/brcadx/ . To our knowledge, this is the best-performing freely available tool for the high-confidence diagnosis of breast cancer, and represents a promising aid to medical diagnosis.

Introduction

Breast cancer is the most commonly diagnosed cancer in the world, with a staggering 2.3 million cases in 2020[1]. It accounts for approximately 24.5% of cancer cases and 15.5% of cancer deaths among women, ranking #1 in both incidence and mortality in most countries. Modelling studies predict an exponential and asymmetric rate of increase in breast cancer incidence among low human development index (HDI) nations relative to high HDI nations, due to an unmitigated increase in risk factors in low HDI nations[2]. In India, for e.g., the age of onset of breast cancer has advanced ten years earlier relative to that in Europe and America. About 29% - 52% of women with breast cancer in India present in the more severe advanced stages, leading to poor prognosis[3]. Low HDI nations are likely to also suffer from problems due to the lack of social awareness and existent taboos, especially in rural areas. Alternative diagnostic methods based on a minimal set of biomarkers are urgently needed to effectively redress the situation[4].

The advent of –omics data has ushered in AI-based approaches to cancer diagnosis. However, contemporary AI-based diagnostic methods are saddled with unreasonable dimensionality of the hypothesis space, and typically require sequencing of hundreds of biomarkers to achieve clinical utility. Dimensionality reduction techniques like principal components (PC) analysis are generally used for extracting optimal feature subsets, especially when linear relationships exist in the dataset. PC analysis has been earlier used to detect multiple cancer types simultaneously, with a costly compromise in accuracy and interpretation[5].

Working in the space of PCs tends to lead to more robust clustering outcomes[6], and k-means clustering is an effective technique for analyzing transformed spaces[7,8]. Building on the above observations, this study has two principal objectives: (i) develop and validate the most efficient integrative computational pipeline for breast cancer classification based on a minimal hypothesis space; and (ii) translate the resulting diagnostic classifier into a web-app service to aid medical decision-making.

Materials and Methods

The overall workflow is summarised in Fig. 1 and discussed in detail below.

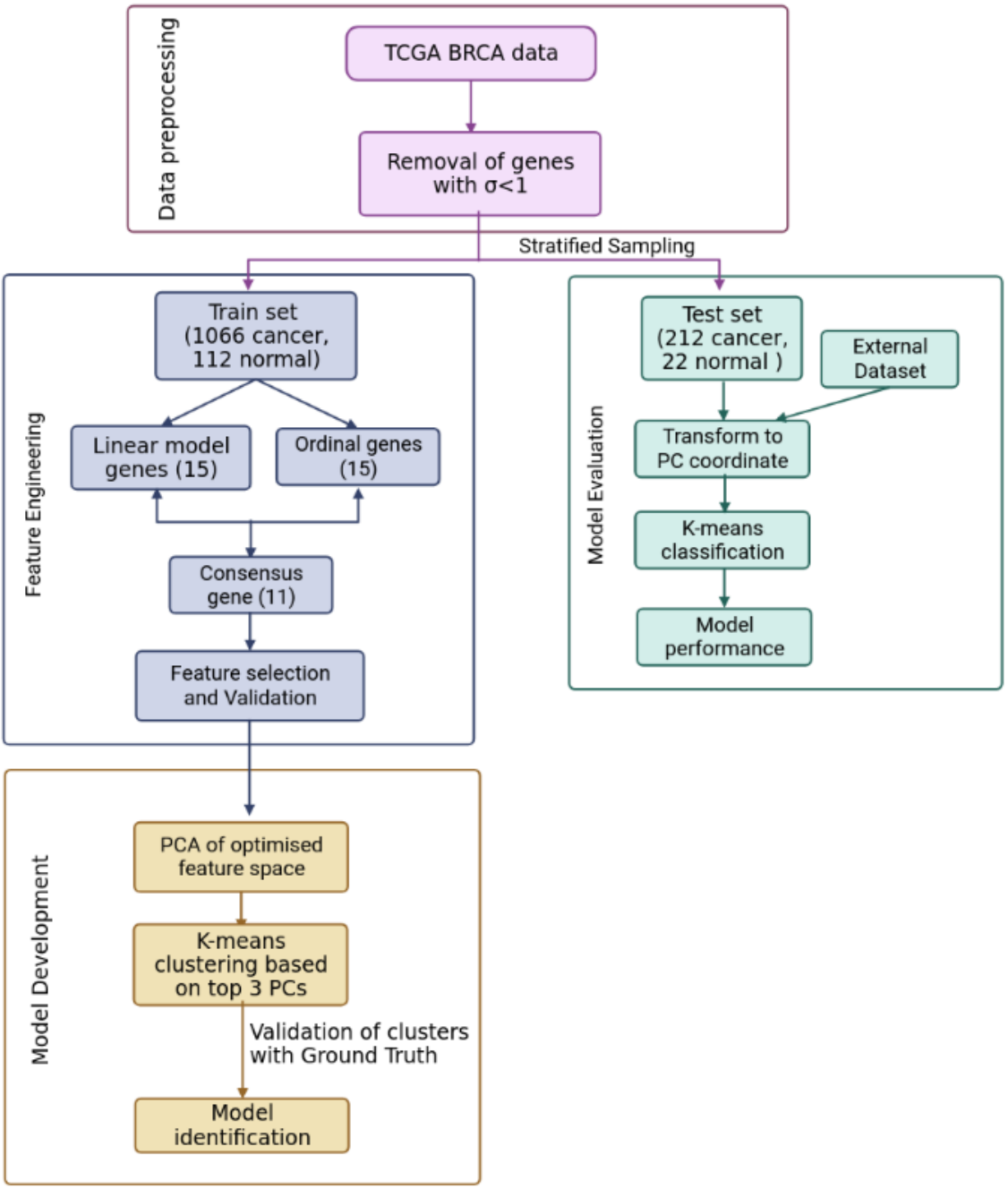

Fig 1. ML pipeline used in the study for the design of a simple, effective and optimal cancer vs normal classifier.

Data Pre-processing:

RSEM-normalised BRCA expression dataset (gdac.broadinstitute.org_BRCA.Merge_rnaseqv2__illuminahiseq_rnaseqv2__unc_edu__Level_3__RSEM_genes_normalized__data.Level_3.2016012800.0.0.tar.gz) was retrieved from the TCGA using firebrowse portal[9] by selecting the

Cohort as 'Breast invasive carcinoma'. The samples were annotated as 'normal' or 'cancer' based on the sample-encoding part in the patient barcode (uuid) in the variable 'Hybridization REF'. The sample stage was extracted from the attribute 'patient.stage_event.pathologic_stage' in the associated clinical metadata file retrieved for the same cohort as gdac.broadinstitute.org_BRCA.Merge_Clinical.Level_1.2016012800.0.0.tar.gz. Genes with minimal variation in expression across the samples were removed if the expression $\sigma < 1$. The resulting data matrix was then processed through voom in limma to prepare for linear modelling[10]. Then it was split into train: test datasets in the ratio 80:20 stratified on the target class. Data pre-processing was done in R (www.r-project.org).

Feature Engineering:

The training dataset was used to identify the features for the problem. Two models were considered to extract potential features:

(1) A linear model of stagewise expression in each gene was performed using R *limma*[10], with the following equation

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \qquad \text{— (1)}$$

where the intercept α is the baseline expression obtained from the controls, the independent variables are indicator variables of the sample's stage, and β_i are the predicted log fold-change (lfc) coefficients relative to controls. Further the model was subjected to empirical Bayes adjustment for obtaining moderated t-statistics[11]. Multiple hypothesis testing was corrected using the Benjamini Hochberg method[12].

(2) An ordinal model of gene expression was also considered. Here the cancer stage is treated as a numeric variable according to the equation:

$$Y = aX + b \qquad \text{— (2)}$$

where X is the cancer stage taking the values 0, 1, 2, 3, and 4, corresponding to Control, Stage-1, Stage-2, Stage-3, and Stage-4, respectively.

Feature space optimization:

Genes from the linear and ordinal expression models were ranked based on the adj. p-value. The consensus set between the top-ranked 15 genes of the linear and ordinal models was determined and then subjected to feature selection using Boruta [13] and Recursive Feature Elimination[14] (RFE). Boruta implements a wrapper algorithm based on Random Forest to select features either strongly or weakly connected to the outcome variable, while RFE implements a backward selection process to identify an optimal set of predictors. Post feature-selection, the retained features were validated using variance inflation analysis, involving regressing each independent variable on all the other independent variables in turn, identifying and removing redundancy till a minimal feature space has been obtained [15]. The variance inflation factor (VIF) score was calculated using:

$$VIF = \frac{1}{1-R^2} \qquad \text{— (3)}$$

where R^2 is the goodness-of-fit of the fitted model. A variable with VIF = 1.0 is perfectly independent of all other variables, whereas any variable with VIF > 2.0 was deemed multicollinear with the other variables and iteratively eliminated.

PCA-based K-Means Clustering:

From the validated set of features, the principal components of the subspace spanned by these features were found, and the optimal number of principal components identified using three different criteria, namely scree plot, Kaiser-Guttmann rule[16], and the proportion of variance explained. K-means clustering with k=2 was performed in the space defined by the optimal principal components, to examine separation between the normal and cancer samples.

Model evaluation:

Classification performance from clustering in the principal components space was evaluated using metrics like accuracy, precision, recall, and F_1-score.

Performance evaluation was done on both the internal testset and an external dataset 'BRCA-KR' retrieved from the ICGC DataPortal (https://dcc.icgc.org/). Since BRCA-KR had just three control samples, it was augmented with 218 control samples from GTEx[17].

Results

BRCA RNA-Seq data retrieved from TCGA consisted of 1212 samples each with the expression values of 20532 genes. Post data pre-processing, we obtained a dataset of 1178 samples, 18880 genes. We performed an 80:20 stratified sampling of the dataset (with 1066 cancer, 112 normal samples) based on the outcome class to obtain the training dataset (with 854 cancer, 90 normal samples), and test dataset (with 212 cancer, 22 normal samples). The training dataset was voom-processed using limma and then subjected to the two modeling protocols. At an adj.p-value threshold of 1E-5, the linear model yielded 8961 significant genes (Supplementary File S1), while the ordinal model yielded 6888 significant (Supplementary File S2). We examined the overlap among the top 15 genes from each model, which produced eleven consensus genes for subsequent analysis.

Application of the Boruta feature selection protocol on the eleven genes yielded a hypothesis space of only nine genes, while application of the RFE feature selection protocol did not yield any reduction in the size of the hypothesis space. A summary of the final nine consensus genes is presented in Table 1. The hypothesis space was subjected to VIF analysis, to ensure absence of multicollinearity among features, and establish a minimal non-redundant set of features (Table 1, last column). We identified the nine principal components (PCs) of this 9-dimensional space (Table 2), and then visualized the training samples using the top PCs from this analysis (Fig. 2). The application of three PCs clearly resolves and separates the cancer and normal samples (Fig. 2b). To decisively identify the optimal number of PCs, we examined the three criteria

outlined in ethods: (i) Kaiser-Guttman criterion yielded top six PCs; (ii) Scree plot showed the first three principal components to be optimal (Fig. 3a); and (iii) the first three PCs explained > 85% variance, passing the proportion of variance explained condition. We reconciled the above findings, and chose the first three principal components to define a 3-dimensional space for applying k-means clustering. Next, we optimized the number of clusters (k) for k-means clustering using the silhouette method[18] (Fig. 3b). A value of k=2 was obtained, which synchronized with the larger objective to partition the structure of the space into cancer and normal signatures.

Table 1. Summary of the consensus features from the two modeling protocols. All features are exceedingly differentially expressed with extreme significance. The largest VIF score does not exceed 1.57, corresponding to a multivariate 'correlation coefficient' < 0.6.

S.No	Feature	lfc	Adj.P.value - linear	Adj.P.value - ordinal	Regulation status	VIF score
1	NEK2	4.57	2.94E-146	6.25E-61	UP	1.05
2	PKMYT1	4.47	1.53E-127	6.14E-53	UP	1.05
3	MMP11	5.99	3.26E-134	2.02E-53	UP	1.00
4	CPA1	-4.20	1.61E-138	2.62E-49	DOWN	1.54
5	COL10A1	7.12	2.04E-137	5.62E-54	UP	1.00
6	HSD17B13	-4.86	5.67E-117	3.71E-51	DOWN	1.22
7	CA4	-6.93	8.41E-127	9.92E-50	DOWN	1.57
8	MYOC	-6.53	3.30E-133	4.03E-57	DOWN	1.34
9	LYVE1	-4.91	3.10E-128	3.21E-47	DOWN	1.02

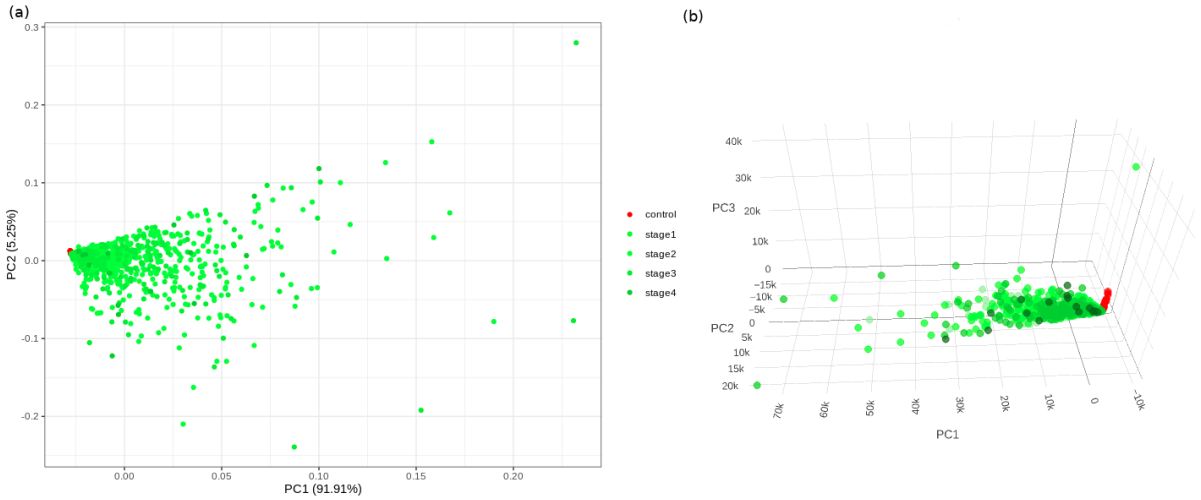

Fig. 2. PC analysis of the biomarker expression space. With (a) top two components; and (b) top three components. It is seen that the use of three components expands the separation between the cancer samples and controls in better-defined sub-spaces.

Table 2. Summary of the nine components from the PC analysis, ranked by associated eigenvalue. Cumulative variance enables the application of the 'proportion of variance explained' criterion.

S.No	PC	Eigenvalue	Variance explained (%)	Cumulative variance explained (%)
1	PC1	34.487	67.24	67.24
2	PC2	7.181	14.00	81.24
3	PC3	2.787	5.43	86.67
4	PC4	2.039	3.97	90.65
5	PC5	1.521	2.97	93.62
6	PC6	1.191	2.32	95.94
7	PC7	0.887	1.73	97.67
8	PC8	0.781	1.52	99.19
9	PC9	0.415	0.81	100

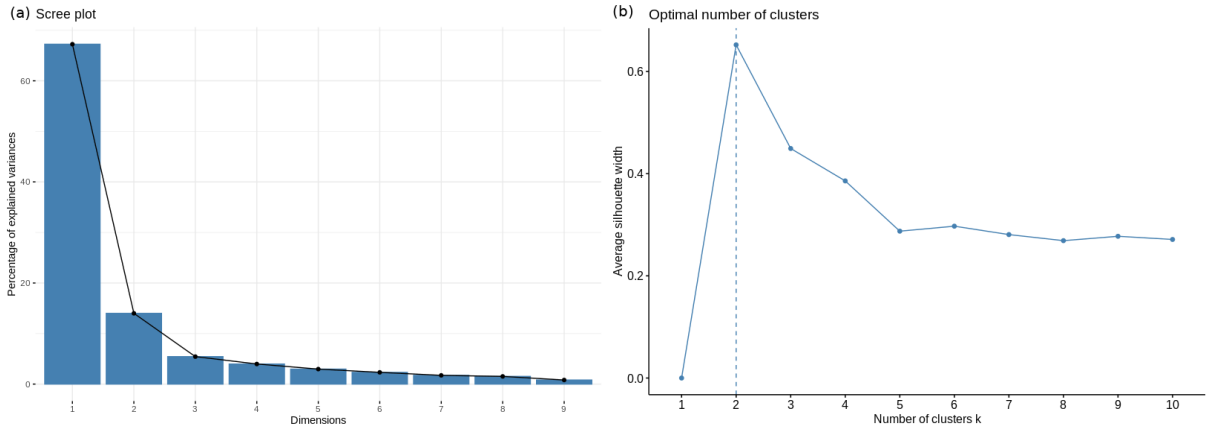

Fig. 3. Model parameterization. (a) Scree plot for determination of the optimal number of principal components. The elbow method yields the first three PCs which have a cumulative variance > 85%. (b) Silhouette plot for ascertaining the optimal number of clusters in the structure of the transformed PC-space. The emergent value, k=2, is in sync with the type of problem at hand: binary classification.

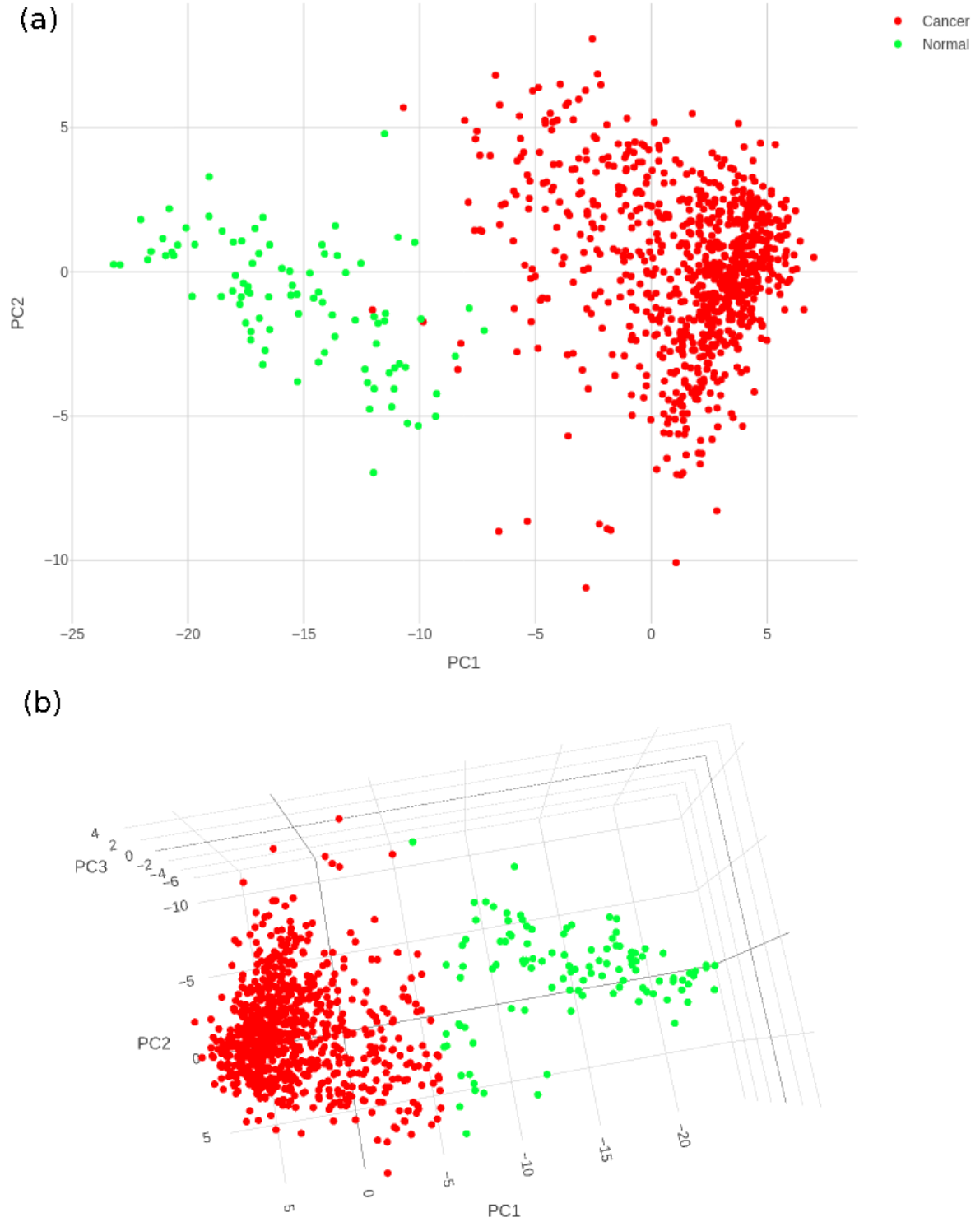

Fig. 4. Cancer (red) and control (green) clusters obtained after training the k-means classifier. (a) Two-dimensional projection onto the first two principal components shows some uncertainty in the boundaries of the two clusters; (b)

Visualization in the three-dimensional space of the PCs satisfactorily resolves the cluster boundaries.

From Fig. 4, it is clear that the k-means classifier in the 3-dimensional PC space of the identified biomarkers effectively partitioned the space into cancer vs normal. The performance of the clustering outcomes assessed against the ground truth labels in the training, test and external datasets is presented in Table 2. It is seen that the model produced by the workflow has yielded balanced accuracies of 99.53% and 95.52% on the internal validation and external validation datasets respectively.

Table 2. Performance metrics of the developed k-means model in the transformed PC space of the identified nine biomarker features. Bal. acc. refers to balanced accuracy, i.e, the average of the accuracies obtained with respect to each class (cancer and normal). F_1-score is defined as the harmonic mean of the precision and recall. Sensitivity is identical to the recall values.

S.No	Dataset	Bal. acc.	Specificity	Precision	Recall	F_1-score
1	Training	98.83	100	100	97.66	98.81
2	Test	99.53	100	100	99.06	99.53
3	**External**	95.52	99.55	97.73	91.49	94.51

Deployment:

To convert the outcomes in effectively classifying cancer vs normal based on the expression of just a handful of features, we have developed an app, BrcaDx, to freely provide the service to the academic community. BrcaDx is deployed at: https://apalania.shinyapps.io/brcadx/ . The model was rebuilt using the full dataset for maximum discriminative performance. Based on an input of the expression of the nine biomarkers, the app carries out a necessary log2 operation of the values, and transforms them into the three-dimensional PC space. The transformed coordinates are fed to the learned k-means clustering model, which locates the sample in either of the two clusters, thus predicting the

class of the sample. The app accepts a single-sample input as well as a batch input (samples x biomarkers), in which case it processes all the samples and returns the corresponding prediction for each sample. To facilitate not-for-profit applications, a video tutorial for using the app has been provided on the landing page. The app was implemented using R-Shiny (https://shiny.rstudio.com/).

Discussion

It is significant to note that some of the biomarkers identified in our study are part of marketed and commercially available signature panels used in the context of breast cancer diagnosis and treatment. Specifically: (i) NEK2 is a constituent of the 11-gene Breast Cancer Index signature used to estimate recurrence[19]; and (ii) MMP11 is a constituent biomarker of the 50-gene Prosigna[20], and 21-gene OncotypeDX[21] signature panels, which are both used in estimating likelihood of recurrence. It is interesting to note that the Prosigna panel is based on the PAM50 signature, which is also used to subtype breast cancer into Luminal-A, Luminal-B, HER2-enriched and Basal-like[22].

The consensus genes used to build our model are known to play key roles in cancers of the breast and other tissues, contributing to breast-cancer specific pathways as well as cancer hallmark processes[23]. The genes NEK2, PKMYT1, and CA4 are known to play indispensable roles in cell cycle progression[24-26]. NEK2 is documented to be overexpressed in breast-cancer tissue relative to normal tissue[27,28], and is required for the growth, maintenance and survival of the transformed cell[29]. PKMYT1 overexpression is known to be significantly correlated with BRCA subtypes, and indicative of poor prognosis[30]. Downregulation of CA4 is associated with poor prognosis in cancers other than that of the breast, notably uveal melanoma, renal cell cancer, glioma, and lung adenocarcinoma[30,31], hinting its role in hallmark processes common to many cancers, and its potential significance in establishing such hallmarks in breast cancer progression. Hypermethylation of the CPA1 gene in breast cancer cells has been earlier demonstrated[32,33], which could lead to its significant

downregulation noted here. Recently, COL10A1 was identified as an overexpressed predictive biomarker for breast cancer coexpressed with LRRC15[34]. COL10A1 protein is a known extracellular matrix molecule released into the blood, and increased levels of circulating COL10A1 protein has been suggested as a diagnostic marker of breast cancer[35]. MYOC has been previously reported as a topranked downregulated gene in breast cancer[36]. MMP11 overexpression in early stages is necessary for cancer progression via inhibition of apoptosis, and promotion of invasion and metastasis[37]. Overexpression of LYVE1 has been suggested as a reliable marker of lymphatic metastasis in breast cancer patients[38]. HSD17B13 is involved in estrogen biosynthesis[39], and its tumor suppressor role in hepatocellular carcinoma has been documented[40], suggesting analogous key roles specific to breast cancer progression.

Due to the substantial heterogeneity in breast cancer, large feature spaces have been necessary for acceptable performance in contemporary classification strategies. Some of these have mandated whole genome sequencing to completely cover the biomarker space of interest. For e.g, Zhao et al identified 817 features and used them to build a model that achieved accuracies of 86.96% and 72.46% in different external validation datasets respectively[41]. Mostavi et al used a feature space of 2090 genes for discriminating cancer vs normal, of which 323 biomarkers were designated for the task of subtyping breast cancer[42]. Convolution-based deep neural networks (CNNs) have been applied to learn from image datasets of mammography, computed tomography (CT), magnetic resonance (MR) and histopathological slides[43-45]. CNNs have been used to extract features from whole-slide tissue-biopsy images, which were subsequently used to train a Support Vector Machine classifier of cancer vs normal, yielding an accuracy of 83.3%[46]. Radiogenomics approaches based on multimodal datasets have also been developed for breast cancer diagnosis[47]. The use of large feature spaces discourages the use of AI-assisted diagnosis in medical decision-making. Very recently Taghizadeh et al have advanced a

solution to the 'cancer' vs. 'normal' problem, proposing a panel of 20 biomarkers for discriminating breast cancer from normal sample[48]. Their study has been validated on an internal test set with a balanced accuracy ~ 86%, but no external validation has been provided. Furthermore their models have not been made available for wider use. It is notable that there is zero overlap between biomarkers identified in their study and identified herein. It is hoped that our study provides a reliable and replicable remedy to the present situation, with a balanced-accuracy performance > 95% on the external validation.

Conclusion

In this work, we set out to negotiate the compromise between model complexity and performance, and develop the simplest possible best-performing model of breast cancer classification. The designed computational pipeline yielded a novel non-redundant hypothesis space of nine biomarkers, which was transformed into a space defined by an optimal number of principal components. A k-means clustering model trained in this transformed space was able to discriminate cancer from normal samples with a high balanced accuracy of 99.5% and 95.5% on the internal and external validation datasets, respectively. At the same time, we note that the model had limited recall (< 92%) on the external validation dataset. The model could be further improved by efforts to predict the subtype of breast cancer as well as its progression to advanced stages or metastasis. The present model has been deployed as a web-service at https://apalania.shinyapps.io/brcadx/ for non-commercial use. The ideas used in our study could be useful in developing elegant, interpretable AI-assisted diagnostic models for many other cancers and disease conditions, fostering effective aid to medical decision-making.

Acknowledgements

We would like to thank the School of Chemical and Biotechnology and CeNTAB, SASTRA Deemed University, for infrastructure and computing

support. This work was partially supported by DST-SERB grant EMR/2017/000470, Government of India.

References

1. Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. **71**, 209-249 (2021).
2. Soerjomataram, I. & Bray, F. J. N. r. C. o. Planning for tomorrow: Global cancer incidence and the role of prevention 2020–2070. **18**, 663-672 (2021).
3. Bhattacharyya, G. S. *et al.* Overview of Breast Cancer and Implications of Overtreatment of Early-Stage Breast Cancer: An Indian Perspective. *JCO global oncology* **6**, 789-798, doi:10.1200/go.20.00033 (2020).
4. Duan, Q. *et al.* L1000CDS2: LINCS L1000 characteristic direction signatures search engine. *NPJ Syst Biol Appl* **2**, 16015, doi:10.1038/npjsba.2016.15 (2016).
5. Fakoor, R., Ladhak, F., Nazi, A. & Huber, M. *Using deep learning to enhance cancer diagnosis and classification*. Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA. JMLR:W&CP **28** (2013).
6. Ding, C. & He, X. K-Means Clustering Via Principal Component Analysis. Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004 1, doi:10.1145/1015330.1015408 (2004).
7. Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques. In: Kogan, J., Nicholas, C., Teboulle, M. (eds) *Grouping Multidimensional Data.* Springer, Berlin, Heidelberg. doi:https://doi.org/10.1007/3-540-28349-8_2
8. Raykov, Y. P., Boukouvalas, A., Baig, F. & Little, M. A. What to Do When K-Means Clustering Fails: A Simple yet Principled Alternative

Algorithm. *PLOS ONE* **11**, e0162259, doi:10.1371/journal.pone.0162259 (2016).

9. Deng M, Brägelmann J, Kryukov I, Saraiva-Agostinho N, Perner S. FirebrowseR: an R client to the Broad Institute's Firehose Pipeline. *Database (Oxford)*. doi: 10.1093/database/baw160. (2017)
10. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47-e47, doi:10.1093/nar/gkv007 (2015).
11. McCarthy, D. J. & Smyth, G. K. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics (Oxford, England)* **25**, 765-771, doi:10.1093/bioinformatics/btp053 (2009).
12. Haynes, W. Benjamini–Hochberg Method. In: Dubitzky, W., Wolkenhauer, O., Cho, KH., Yokota, H. (eds) Encyclopedia of Systems Biology. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-9863-7_1215 (2013).
13. Kursa, M. B. & Rudnicki, W. R. Feature Selection with the Boruta Package. *Journal of Statistical Software* **36**, 1 - 13, doi:10.18637/jss.v036.i11 (2010).
14. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **28**, 1 - 26, doi:10.18637/jss.v028.i05 (2008).
15. Ferré, J. in *Comprehensive Chemometrics* (eds Steven D. Brown, Romá Tauler, & Beata Walczak) 33-89 (Elsevier, 2009).
16. Kaiser, H. F. On Cliff's formula, the Kaiser-Guttman Rule, and the number of factors. *Perceptual and Motor Skills* **74**, 595-598, doi:10.2466/PMS.74.2.595-598 (1992).
17. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of

Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this were obtained from: GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct.gz the GTEx Portal and/or dbGaP accession number phs000424.v8.p2.

18. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53-65, doi:https://doi.org/10.1016/0377-0427(87)90125-7 (1987).
19. Zhang, Y. et al. Breast Cancer Index Identifies Early-Stage Estrogen Receptor–Positive Breast Cancer Patients at Risk for Early- and Late-Distant Recurrence. Clinical Cancer Research, 19, 4196-4205, doi:10.1158/1078-0432.CCR-13-0804 %J Clinical Cancer Research (2013).
20. Parker, J. S. et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 27, 1160-1167, doi:10.1200/jco.2008.18.1370 (2009).
21. Cronin, M. et al. Analytical Validation of the Oncotype DX Genomic Diagnostic Test for Recurrence Prognosis and Therapeutic Response Prediction in Node-Negative, Estrogen Receptor–Positive Breast Cancer. Clinical Chemistry 53, 1084-1091, doi:10.1373/clinchem.2006.076497 %J Clinical Chemistry (2007).
22. Bastien, R. R., Rodríguez-Lescure, Á., Ebbert, M. T., Prat, A., Munárriz, B., Rowe, L., Miller, P., Ruiz-Borrego, M., Anderson, D., Lyons, B., Álvarez, I., Dowell, T., Wall, D., Seguí, M. Á., Barley, L., Boucher, K. M., Alba, E., Pappas, L., Davis, C. A., Aranda, I., … Martín, M. PAM50 breast cancer subtyping by RT-qPCR and concordance with standard clinical molecular markers. BMC medical genomics, 5, 44. doi:10.1186/1755-8794-5-44 2012.

23. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell, 144(5):646–74 doi:10.1016/j.cell.2011.02.013 (2011).

24. Fang, Y. & Zhang, X. Targeting NEK2 as a promising therapeutic approach for cancer treatment. *Cell cycle (Georgetown, Tex.)* **15**, 895-907, doi:10.1080/15384101.2016.1152430 (2016).

25. Mueller, P. R., Coleman, T. R., Kumagai, A. & Dunphy, W. G. J. S. Myt1: a membrane-associated inhibitory kinase that phosphorylates Cdc2 on both threonine-14 and tyrosine-15. **270**, 86-90 (1995).

26. Lagadic-Gossmann, D., Huc, L., Lecureur, V. Alterations of intracellular pH homeostasis in apoptosis: origins and roles. *Cell Death & Differentiation* **11**, 953-961, doi: 10.1038/sj.cdd.4401466 (2004).

27. Hayward, D. G. *et al.* The Centrosomal Kinase Nek2 Displays Elevated Levels of Protein Expression in Human Breast Cancer. *Cancer Research* **64**, 7370-7376, doi:10.1158/0008-5472.CAN-04-0960 (2004).

28. Cappello, P. *et al.* Role of NEK2 on centrosome duplication and aneuploidy in breast cancer cells. *Oncogene.* **33**, 2375-2384 doi: 10.1038/onc.2013.183. (2014).

29. Lee, J. & Gollahon, L. NEK2-targeted ASO or siRNA pretreatment enhances anticancer drug sensitivity in triple-negative breast cancer cells. *International journal of oncology* **42**, 839-847, doi:10.3892/ijo.2013.1788 (2013).

30. Liu, Y. *et al.* Systematic expression analysis of WEE family kinases reveals the importance of PKMYT1 in breast carcinogenesis. *Cell proliferation* **53**, e12741, doi:10.1111/cpr.12741 (2020).

31. Xu, Y. *et al.* Carbonic Anhydrase 4 serves as a Clinicopathological Biomarker for Outcomes and Immune Infiltration in Renal Cell Carcinoma, Lower Grade Glioma, Lung Adenocarcinoma and Uveal Melanoma. *Journal of Cancer* **11**, 6101-6113, doi:10.7150/jca.46902 (2020).

32. Chen, J. *et al.* Downregulation of carbonic anhydrase IV contributes to promotion of cell proliferation and is associated with poor prognosis in non-small cell lung cancer. *Oncol Lett* **14**, 5046-5050, doi:10.3892/ol.2017.6740 (2017).
33. DeVaux, R. S. & Herschkowitz, J. I. Beyond DNA: the Role of Epigenetics in the Premalignant Progression of Breast Cancer. *Journal of Mammary Gland Biology and Neoplasia* **23**, 223-235, doi:10.1007/s10911-018-9414-2 (2018).
34. Fleischer, T. *et al.* Genome-wide DNA methylation profiles in progression to in situ and invasive carcinoma of the breast with impact on gene transcription and prognosis. *Genome biology* **15**, 435, doi:10.1186/preaccept-2333349012841587 (2014).
35. Zhang, M., Chen, H., Wang, M., Bai, F. & Wu, K. Bioinformatics analysis of prognostic significance of COL10A1 in breast cancer. *Bioscience reports* **40**, doi:10.1042/bsr20193286 (2020).
36. Giussani, M. *et al.* Extracellular matrix proteins as diagnostic markers of breast carcinoma. *J Cell Physiol.* **233**, 6280-6290 (2018).
37. Li, X. *et al.* A combined approach with gene-wise normalization improves the analysis of RNA-seq data in human breast cancer subtypes. *PLoS One* **13**, e0201813, doi:10.1371/journal.pone.0201813 (2018).
38. Zhang, Z.-Q. *et al.* Tumor Invasiveness, Not Lymphangiogenesis, Is Correlated with Lymph Node Metastasis and Unfavorable Prognosis in Young Breast Cancer Patients (≤35 Years). *PLOS ONE* **10**, e0144376, doi:10.1371/journal.pone.0144376 (2015).
39. Doan, T. B. *et al.* Breast cancer prognosis predicted by nuclear receptor-coregulator networks. *Molecular oncology* **8**, 998-1013, doi:10.1016/j.molonc.2014.03.017 (2014).

40. Wang. X. *et al.* Identification of prognostic biomarkers for patients with hepatocellular carcinoma after hepatectomy. *Oncol Rep.* **41(3)**1586-1602, doi: 10.3892/or.2019.6953.

41. Zhao, Y. *et al.* CUP-AI-Dx: A tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence. *EBioMedicine* **61**, 103030, doi:https://doi.org/10.1016/j.ebiom.2020.103030 (2020).

42. Mostavi, M., Chiu, Y.-C., Huang, Y. & Chen, Y. Convolutional neural network models for cancer type prediction based on gene expression. *BMC Medical Genomics* **13**, 44, doi:10.1186/s12920-020-0677-2 (2020).

43. Munir, K., Elahi, H., Ayub, A., Frezza, F. & Rizzi, A. Cancer Diagnosis Using Deep Learning: A Bibliographic Review. *Cancers* **11** (2019).

44. Jiang, Y., Yang, M., Wang, S., Li, X. & Sun, Y. Emerging role of deep learning-based artificial intelligence in tumor pathology. *Cancer communications (London, England)* **40**, 154-166, doi:10.1002/cac2.12012 (2020).

45. Saha, A., Harowicz, M.R., Grimm, L.J., Kim, C.E., Ghate, S.V., Walsh, R. and Mazurowski, M.A. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. *British Journal of Cancer*, **119**(4), pp.508-516 (2018).

46. Araújo, T. *et al.* Classification of breast cancer histology images using Convolutional Neural Networks. *PLoS One* **12**, e0177544, doi:10.1371/journal.pone.0177544 (2017).

47. Muduli, D., Dash, R. & Majhi, B. Automated diagnosis of breast cancer using multi-modal datasets: A deep convolution neural network based approach. *Biomedical Signal Processing and Control* **71** 102825, doi:https://doi.org/10.1016/j.bspc.2021.102825 (2022).

48. Taghizadeh, E., Heydarheydari, S., Saberi, A. et al. Breast cancer prediction with transcriptome profiling using feature selection and machine learning methods. BMC Bioinformatics 23, 410 doi:10.1186/s12859-022-04965-8 (2022).