frontiers | Frontiers in Molecular Biosciences

TYPE Mini Review
PUBLISHED 10 November 2023
DOI 10.3389/fmolb.2023.1285795

OPEN ACCESS

EDITED BY
Muhammad Tahir Khan,
University of Lahore, Pakistan

REVIEWED BY
Marcelo Cardoso Dos Reis Melo,
Auburn University, United States
Athar Shafiq,
Shanghai Jiao Tong University, China

*CORRESPONDENCE
Rosa Karlić,
rosa@bioinfo.hr

RECEIVED 30 August 2023
ACCEPTED 30 October 2023
PUBLISHED 10 November 2023

CITATION
Štancl P and Karlić R (2023), Machine learning for pan-cancer classification based on RNA sequencing data.
Front. Mol. Biosci. 10:1285795.
doi: 10.3389/fmolb.2023.1285795

Machine learning for pan-cancer classification based on RNA sequencing data

Paula Štancl and Rosa Karlić*

Bioinformatics Group, Division of Molecular Biology, Department of Biology, Faculty of Science, University of Zagreb, Zagreb, Croatia

Despite recent improvements in cancer diagnostics, 2%-5% of all malignancies are still cancers of unknown primary (CUP), for which the tissue-of-origin (TOO) cannot be determined at the time of presentation. Since the primary site of cancer leads to the choice of optimal treatment, CUP patients pose a significant clinical challenge with limited treatment options. Data produced by large-scale cancer genomics initiatives, which aim to determine the genomic, epigenomic, and transcriptomic characteristics of a large number of individual patients of multiple cancer types, have led to the introduction of various methods that use machine learning to predict the TOO of cancer patients. In this review, we assess the reproducibility, interpretability, and robustness of results obtained by 20 recent studies that utilize different machine learning methods for TOO prediction based on RNA sequencing data, including their reported performance on independent data sets and identification of important features. Our review investigates the strengths and weaknesses of different methods, checks the correspondence of their results, and identifies potential issues with datasets used for model training and testing, assessing their potential usefulness in a clinical setting and suggesting future improvements.

KEYWORDS

cancer of unknown primary, tissue of origin, cancer classification, machine learning, RNA sequencing

1 Introduction

Cancer is the leading cause of death worldwide, and the overall burden of cancer incidence and mortality is expected to increase due to the growing population, aging inhabitants, and changes in the prevalence of risk factors (Sung et al., 2021). Key factors in reducing cancer incidence and improving the survival of cancer patients include prevention, early detection, and the availability of appropriate treatment. Despite the recent advances in cancer diagnostics, cancers of unknown primary (CUP), in which the tissue-of-origin (TOO) cannot be identified at the time of presentation, still constitute approximately 2%–5% of all malignancies. Since the primary site of cancer determines the choice of optimal treatment, CUP patients present a significant clinical challenge with limited treatment options. Although a fraction of CUP patients can be assigned the correct TOO and receive appropriate treatment based on the analysis of clinical, imaging, or histopathological data, this is still not the case for the majority of CUP patients, who then face a less favorable prognosis (Binder et al., 2018; Rassy and Pavlidis, 2020).

The development of array- and next-generation sequencing (NGS)-based whole-genome profiling techniques has enabled the rapid molecular characterization of cells or tissues, inspiring the establishment of several large-scale cancer genomics initiatives, such as ICGC

and TCGA (International Cancer Genome Consortium et al., 2010; Weinstein et al., 2013). These initiatives aim to describe the genomic, epigenomic, and transcriptomic characteristics of a large number of individual patients with multiple types of cancer. The unprecedented volume of data produced by these techniques has led to the introduction of various methods that use machine learning to predict the TOO for cancer patients. This is most frequently done by analyzing somatic alterations, gene expression, microRNA expression, or DNA methylation of cancer samples. Most of the TOO prediction tools currently used in clinical practice are based on qRT-PCR or microarray measurements of different features of preselected genes. The accuracy of such tools typically falls within the range of 54%–100%. While several clinical trials have demonstrated improved overall survival of CUP patients who received tumor-type-specific therapy based on predicted TOO, inconsistencies in the results of randomized and non-randomized trials suggest that there are opportunities for improvement in this area of research (Conway et al., 2019; Rassy et al., 2020).

Various methods that utilize machine learning to predict the TOO based on NGS data have been developed recently, although the clinical use of such methods is still limited (Swanson et al., 2023). Advantages of molecular characterization using NGS approaches, in comparison to array-based techniques, include increased specificity and sensitivity, a broader dynamic range, and whole-genome coverage. Indeed, several recent reviews on the topic of machine learning and deep learning for cancer classification have reported the excellent performance of methods that utilize NGS data (Tufail et al., 2021; Alharbi and Vakanski, 2023). However, the high dimensionality, sparsity, and heterogeneity of input data, as well as dataset imbalance, could lead to issues with overfitting and high variance of the trained models.

In this minireview, we aim to assess the reproducibility, interpretability, and robustness of different NGS-based TOO prediction methods, including reported performance on independent data sets and identification of important features. Since the accuracy of the models depends on the input data type (Conway et al., 2019; Rassy et al., 2020), we have limited our survey to tools developed exclusively on RNA sequencing (RNA-Seq) data. This decision was made considering that they represent the majority of studies. We focused on studies developed on pan-cancer data that include at least nine different cancer types in training data. Our review investigates the strengths and weaknesses of different methods, checks the correspondence of their results, and provides an assessment of their potential usefulness in a clinical setting.

2 Machine learning in cancer of unknown primary classification based on RNA sequencing

We have conducted a literature search to identify studies that used RNA-Seq data to train machine learning models for cancer classification and TOO prediction. Studies that relied exclusively on miRNA sequencing were excluded from further analysis. Based on the aforementioned criteria, we selected 20 recently published studies for analysis.

The majority of studies were based on deep learning, using neural networks of different architectures (Lyu and Haque, 2018; Azarkhalili et al., 2019; De Guia et al., 2019; He et al., 2020b; Mostavi et al., 2020; Zhao et al., 2020; Vibert et al., 2021; Divate et al., 2022; Hong et al., 2022; Jones et al., 2022; Moiso et al., 2022). Several studies utilized ensemble learning methods, in which the final prediction is a combination of multiple predictors (Grewal et al., 2019; He et al., 2020a; Ramroach et al., 2020; Chen et al., 2021; Liu et al., 2021). Bavafaye Haghighi et al. (2019) and Galea et al. (2017) classified different cancer types using hierarchical classification, a 'top-down' classification approach in which classification models are trained at each level of the hierarchy. Additionally, some authors employed simpler machine learning methods, such as k-nearest neighbors (Bagge et al., 2018) or stepwise logistic regression (Wei et al., 2014). Several studies tested their results against additional classification methods (Azarkhalili et al., 2019; De Guia et al., 2019; Grewal et al., 2019; Ramroach et al., 2020; Hong et al., 2022), or compared the performances of various deep learning architectures (Mostavi et al., 2020; Zhao et al., 2020; Vibert et al., 2021). In cases where multiple approaches were used in a single study, we limited our analysis to the best-performing model.

The models were trained on datasets comprising 9 to 40 cancer types (with a median number of 32 cancer types), and the number of sample points used for training ranged from 1,960 to 20,918 (with a median number of 10,116 samples). All of the selected studies trained their models on either TCGA or ICGC data, with some studies including cancer sequencing data produced in-house (Wei et al., 2014) or data from healthy tissues (Azarkhalili et al., 2019; Grewal et al., 2019; Mostavi et al., 2020; Vibert et al., 2021).

3 Performance of models for tissue-of-origin prediction

We have compared the classification accuracy of various models, which we defined as the number of correct predictions divided by the number of total predictions. In the cases where the classification accuracy was not reported, we have calculated it from the results described in the original publication. Cases where no prediction could be made were counted as incorrect predictions. Accuracy was the most commonly used measure of predictive performance in the studies surveyed in this minireview. While this measure can be influenced by class-imbalanced data, the influence of training dataset composition on measures of predictive performance is outside of the scope of this minireview.

In general, the analyzed models achieve high cross-validation prediction accuracy in the range of 73%–99% (with a median cross-validation accuracy of 95.5%; Figure 1A). This accuracy does not seem to depend on the number of training points (Spearman's correlation coefficient $\rho = 0.0591$, p-value = 0.7989). The prediction accuracy varies by tumor type, with some tumor types being more frequently mispredicted. Patterns of more frequent misclassifications among groups of cancers arising from the same organ (e.g., kidney renal clear cell carcinoma, kidney renal papillary cell carcinoma and kidney chromophobe carcinoma, or lung adenocarcinoma and lung squamous cell carcinoma), and/or among cancers represented by a small number of samples in the training set (e.g., cholangiocarcinoma, which is frequently predicted

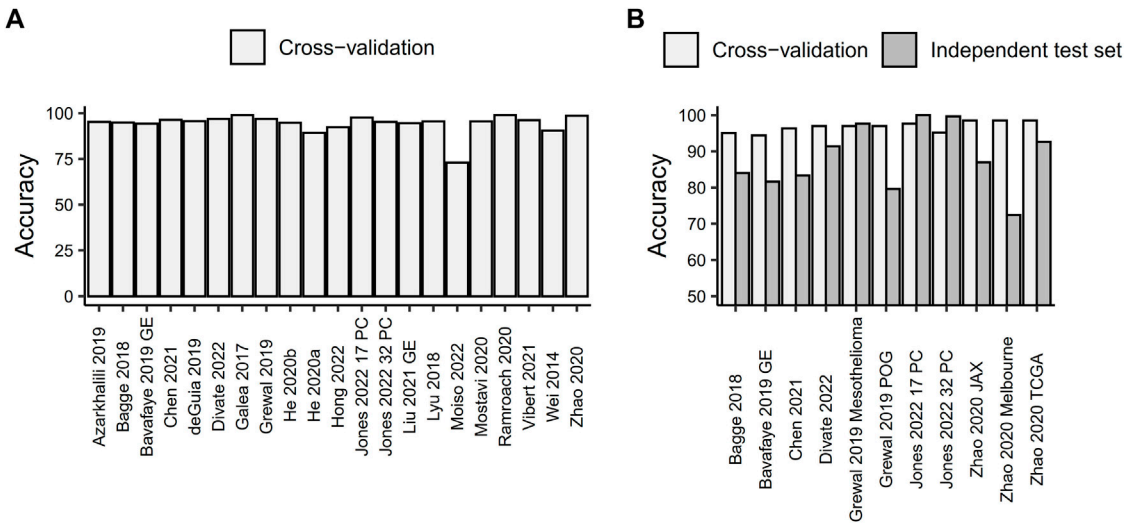

FIGURE 1
Prediction accuracy of machine learning models for tissue-of-origin prediction based on RNA sequencing data. **(A)** Cross-validated prediction accuracy for all models. **(B)** Comparison of cross-validated prediction accuracy and accuracy measured on an independent test set.

as liver hepatocellular carcinoma and *vice versa*), as noted in multiple studies (Bagge et al., 2018; Lyu and Haque, 2018; Bavafaye Haghighi et al., 2019; De Guia et al., 2019; Grewal et al., 2019; Zhao et al., 2020; Vibert et al., 2021; Divate et al., 2022; Jones et al., 2022; Moiso et al., 2022). All of this implies that the distribution of different cancer types in the training set is one of the key factors contributing to the prediction accuracy of the model.

Out of the studies included in this minireview, only seven tested the performance of the developed models on an independent test set. Overall, the predictive accuracy of models on independent test data was lower than the cross-validation accuracy calculated on the data used for training the model, with only two studies attaining a prediction accuracy on an independent test set that was comparable to or higher than the one obtained with cross-validation (Figure 1B). Jones et al. (2022) used 277 primary kidney cancer samples from the CPTAC consortium to test their convolutional neural network-based models, which were trained on 17 or 32 primary cancer types. They achieved prediction accuracies of 100% and 99.63%, respectively. An ensemble of neural networks developed by Grewal et al. (2019) correctly predicted the primary cancer type for 96.73% of 211 samples from the independent Genentech Mesothelioma dataset. However, both of those models were tested on independent data comprising a single primary cancer type, which may not necessarily reflect the putative prediction accuracy on a pan-cancer dataset. In fact, the accuracy of the Grewal et al. (2019) model decreased to 79.60% when tested on an additional independent test set of 201 patients spanning 26 different cancer types. These patients were sequenced as part of the Personalized OncoGenomics project at the BC Cancer Agency and presented with metastatic disease that no longer responds to treatment. The remaining studies used independent test sets consisting of 5–18 cancer types and showed a reduction in prediction accuracy ranging from 5.8% to 26.47% compared to cross-validation (with a mean reduction in accuracy of 13.32%). Out of these, two studies used metastatic samples (Bavafaye Haghighi et al., 2019; Zhao et al., 2020), two used a mixture of primary and metastatic samples (Bagge et al., 2018; Divate et al., 2022) and one did not report the exact source of the independent test set (Chen et al., 2021).

Interestingly, when testing the accuracy of the same model on test sets composed of both primary and metastatic samples, metastatic samples showed a lower prediction accuracy. For example, Divate et al. (2022) reported 88.10% accuracy for metastatic samples compared to 92.13% for primary samples. Bagge et al. (2018) found 53.84% accuracy for metastatic samples compared to 96.67% for patient-derived xenografts of primary cancer and 100% for primary cancer. However, this difference could also be attributed to the distribution of the primary cancer types from which the metastases arose in the test set and their potential underrepresentation in the training data.

Furthermore, Zhao et al. (2020) used three different independent test sets in their research, including two datasets obtained by RNA-Seq of formalin-fixed paraffin-embedded (FFPE) metastatic cancer samples. The FFPE-based datasets showed a lower prediction accuracy (86.96% for the JAX clinical dataset, which included 23 samples across 6 cancer types, and 72.46% for the Melbourne clinical dataset, which encompassed 69 samples across 18 cancer types) compared to RNA-Seq conducted on fresh frozen tissue samples (92.64% for the TCGA, which included 394 samples across 11 cancer types). This suggests that the methods used for tissue processing and storage could impact the results obtained from different sample types.

4 Identification of informative gene sets for cancer classification

Several methods employed feature selection approaches to identify the most informative sets of genes for TOO prediction using different strategies to choose relevant genes. Two studies

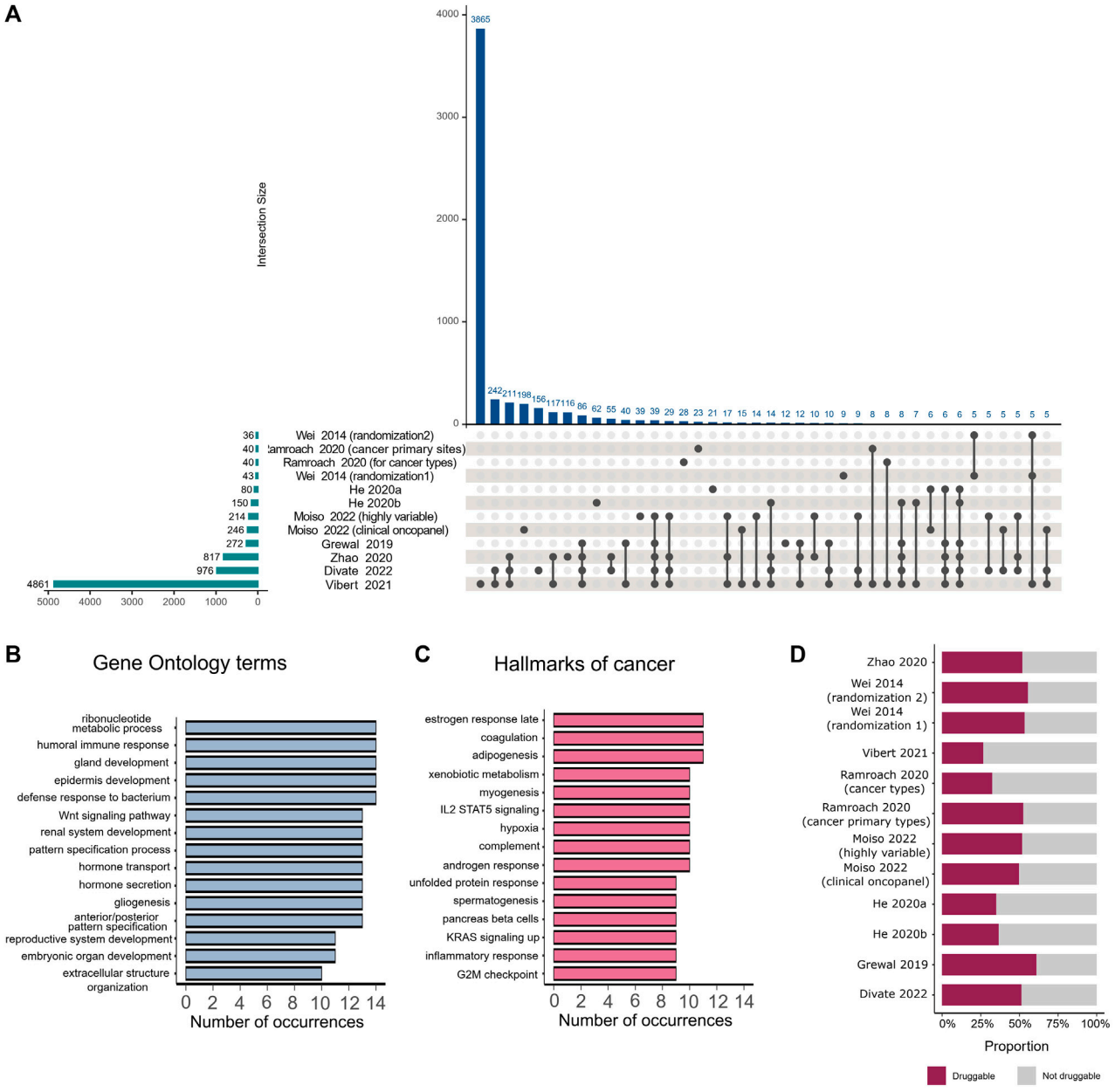

FIGURE 2
Comparison of important genes identified by various feature selection methods. **(A)** The intersection plot shows overlaps between genes detected as important by different studies. Numbers of genes identified by individual studies are displayed as horizontal bars on the lower left corner of the image. Intersection sizes are shown as individual bars on the top of the plot. Specific studies involved in each intersection are identified with connected solid black circles under the vertical bars, with unconnected circles representing genes that are detected exclusively in the corresponding study. **(B)** Number of occurrences of different Gene Ontology terms in studies that employed feature selection. **(C)** Number of occurrences of different hallmarks of cancers in studies that employed feature selection. **(D)** Proportion of druggable genes among all genes that were selected by feature selection in each study.

utilized various subsets of top-selected genes based on the Gini index from Random Forest models (He et al., 2020a; Ramroach et al., 2020). Similarly, neural network models selected a certain number of top-ranked genes, either based on the highest weights for the tumor class (Grewal et al., 2019) or by calculating Shapley additive explanation values in deep neural network models (Divate et al., 2022). Wei et al. (2014) employed univariate transcript analysis with stepwise logistic regression to select the top N transcripts. Their approach achieved high AUC prediction values using only a smaller number of genes. However, the results of two randomizations of the feature selection process show that, although the number of selected genes was similar, the correspondence between the two randomizations was low, with a cosine similarity of 0.53. Another approach was to select highly variable genes across various cancers and either use the top N genes for modeling (Moiso et al., 2022) or perform dimensionality reduction on them before training machine

learning models like variational autoencoders (Vibert et al., 2021). Furthermore, feature selection could be performed by selecting the most important genes in each cancer type. Zhao et al. (2020) did this by selecting the top N differentially expressed genes in each cancer type to create a unique list of genes. A similar approach, based on the Pearson correlation algorithm, was employed by He et al. (2020b). Other approaches involved the selection of genes based on prior knowledge, such as choosing clinical oncopanel genes tested in routine clinical cancer care (Moiso et al., 2022) or a catalog of somatic mutations in cancers (COSMIC) list of genes harboring somatic mutations (Grewal et al., 2019).

To estimate the correspondence of informative gene sets identified by various studies, we extracted gene lists provided by the studies that employed feature selection on the primary dataset. Studies selected between 36 and 4,861 unique genes, with the highest number of overlapping genes found between gene lists identified by Divate et al. (2022), Zhao et al. (2020), and Vibert et al. (2021), which collectively chose the highest number of genes among all the analyzed studies (Figure 2A). No genes were found to be selected by all studies.

We mapped the extracted gene lists to hallmarks of cancer (Dolgalev, 2022) and Gene Ontology terms (Wu et al., 2021) and analyzed the number of occurrences of each hallmark or term among the different studies. The majority of selected genes belonged to specific pathways such as embryonic organ development, gland development, hormone metabolic processes, reproductive system development, and morphogenesis of an epithelium. These pathways were observed in more than 10 studies (Figure 2B). The most frequently occurring hallmarks of cancer, found in the majority of studies, were reproductive system hallmarks such as estrogen late response, androgen response, and spermatogenesis (Figure 2C). We also examined which genes in each study were druggable according to The Drug Gene Interaction Database (Beerenwinkel et al., 2016; Wagner et al., 2016), following the method proposed by Ramroach et al. (2020), and found that all studies selected at least 25% druggable genes (Figure 2D).

We further examined genes proposed by different studies as potential cancer signatures. Well-known prostate cancer signatures, including *KLK3*, a serine protease used as a serum marker in prostate cancer screening and disease monitoring, and *PRAC*, a highly expressed gene in prostate cancer (Edwards et al., 2005), have been selected by 7 and 3 studies, respectively. Another protein selected by 7 studies is the lung biomarker *NAPSA*, an aspartic proteinase expressed in type II pneumocytes. Its expression can be used to distinguish pulmonary lesions originating from primary lung adenocarcinoma or other primaries (Ueno et al., 2003). Additionally, 5 other studies identified *IGFBP1*, a hepatocyte-derived secreted protein required for normal liver regeneration by inhibiting proapoptotic signals (Borlak et al., 2005), as an important feature, particularly for the identification of hepatocellular carcinoma, in which it is overexpressed.

Wei et al. (2014) identified additional potential cancer signatures not previously associated with the cancer types of interest. Some of those genes, such as *DPYS*, were also identified by three more recent studies (Zhao et al., 2020; Vibert et al., 2021; Divate et al., 2022) as important for TOO prediction, especially for kidney cancer. Other genes proposed by Wei et al. (2014), such as the potential new ovarian biomarker *BEST1* and new prostate and gastric biomarker *SI*, were either not found by other studies or were only identified by Vibert et al. (2021), despite having implications in other cancer types.

Ramroach et a. (2020) stated that the majority of the top 40 genes selected by their study belong to the olfactory receptor family, keratin-associated proteins, or the defensin beta family. Interestingly, although it was claimed that the olfactory receptor family plays a significant role in cancer, those genes were not detected by any other study. Genes belonging to keratin-associated proteins were only identified by Vibert et al. (2021). From the defensin beta family, only the *DEFB1* gene was detected by four different studies (Grewal et al., 2019; Zhao et al., 2020; Vibert et al., 2021; Divate et al., 2022), however, this particular gene was not included in the top 40 genes selected by Ramroach et al. (2020).

5 Conclusion and future improvements

In this minireview, we analyzed 20 recent studies that employed machine learning to predict the TOO of cancers based on NGS data. Our goal was to assess their performance, reproducibility, interpretability, and robustness. We found that all of the analyzed methods exhibited very high prediction accuracy, ranging from 73% to 99%. This performance represents an improvement over currently used microarray-based methods, which have a prediction accuracy of 54%–100% (Conway et al., 2019; Rassy et al., 2020). These findings suggest that these machine learning approaches have the potential to bring about significant advancements in the diagnosis and treatment of cancers of unknown primary.

However, while the overall prediction accuracy of the models is high, it varies by tumor type, with tumors originating from the same organ or tumors that are underrepresented in the training set being more frequently mispredicted. This suggests that the accuracy depends more on the composition of the training set than on the method used for training the model. Researchers should, therefore, aim to assemble balanced datasets for model training and include as many samples of rare and underrepresented cancers as possible.

Furthermore, most of the analyzed studies did not employ an independent test set, and the ones that did mostly showed a reduction in accuracy, especially for test sets obtained from metastatic patients or FFPE samples. Since CUP patients are, by definition, metastatic patients and FFPE tissues are still the most commonly available sample type for RNA-Seq, due to the cost-effectiveness of storage (Zhao et al., 2019), both metastatic samples from as many cancer types and FFPE samples should be included in independent test sets to support the claims of potential clinical use of NGS-based approaches to cancer classification. Additional factors, such as data quality and tumor purity, should also be investigated to determine their potential impact on model accuracy.

Identification of features important for prediction, implemented by several of the analyzed studies, could lead to novel biomarker discovery and discovery of genes whose expression is dysregulated in cancer, expand our current knowledge of mechanisms of cancer development and progression, identify potential actionable targets, and inspire novel treatment strategies. Indeed, most of the studies that employed feature selection identified at least 25% of actionable targets among their set of selected genes and showed that some of those genes are already known cancer signature genes. However, the overlap of gene lists provided by different studies is quite low,

indicating that these results should be interpreted with caution. It is important to note that the majority of the analyzed studies employed filter methods, which select relevant features based on their intrinsic characteristics. Some used wrapper methods, where features are added or removed iteratively and scored based on their impact on the machine learning model's performance, or embedded approaches, which combine properties of both filter and wrapper methods.

Filter methods, while computationally less demanding, typically do not consider the subsequent classification model, often resulting in inferior performance compared to wrappers. In contrast, wrappers can be susceptible to overfitting and sensitive to parameter adjustments (Zanella et al., 2022). Recently, various nature-inspired algorithms, such as those based on swarm intelligence and evolutionary principles, have been applied as metaheuristic search methods for wrapper-based feature selection problems and show significant potential in the identification of relevant genes for cancer classification using microarray gene expression measurements (Pham and Raahemi, 2023; Yaqoob et al., 2023).

For example, in a study using ten microarray datasets for cancer classification, Gene Selection Programming (Alanni et al., 2019), a method for selecting relevant genes based on Gene Expression Programming (Ferreira, 2002), demonstrated the highest accuracy and the fewest selected genes in the majority of cases, outperforming swarm-based algorithms and more traditional methods like support vector machines. This suggests that the application of such methods to RNA-Seq datasets could lead to more accurate and robust gene selection for cancer classification. Furthermore, the availability of additional sequencing data and the investigation of possible biases that could influence modeling results could further enhance the clinical applicability of methods described in this minireview and similar tools.

Author contributions

PŠ: Data curation, Formal Analysis, Visualization, Writing–original draft, Writing–review and editing. RK: Conceptualization, Formal Analysis, Funding acquisition, Supervision, Visualization, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. PŠ and RK were funded by grants from the Croatian National Science Foundation Project PREDI-COO (A statistical modeling approach to predict the cell-of-origin and investigate mechanisms of cancer development) (Project number: IP-2019-04-9308).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Alanni, R., Hou, J., Azzawi, H., and Xiang, Y. (2019). A novel gene selection algorithm for cancer classification using microarray datasets. *BMC Med. Genomics* 12, 10. doi:10.1186/s12920-018-0447-6

Alharbi, F., and Vakanski, A. (2023). Machine learning methods for cancer classification using gene expression data: a review. *Bioengineering* 10, 173. doi:10.3390/bioengineering10020173

Azarkhalili, B., Saberi, A., Chitsaz, H., and Sharifi-Zarchi, A. (2019). DeePathology: deep multi-task learning for inferring molecular pathology from cancer transcriptome. *Sci. Rep.* 9, 16526. doi:10.1038/s41598-019-52937-5

Bagge, R. O., Demir, A., Karlsson, J., Alaei-Mahabadi, B., Einarsdottir, B. O., Jespersen, H., et al. (2018). Mutational signature and transcriptomic classification analyses as the decisive diagnostic tools for a cancer of unknown primary. *JCO Precis. Oncol.* 2, 1–25. doi:10.1200/PO.18.00002

Bavafaye Haghighi, E., Knudsen, M., Elmedal Laursen, B., and Besenbacher, S. (2019). Hierarchical classification of cancers of unknown primary using multi-omics data. *Cancer Inf.* 18, 1176935119872163. doi:10.1177/1176935119872163

Beerenwinkel, N., Thurnherr, T., Singer, F., and Stekhoven, D. J. (2016). Genomic variant annotation workflow for clinical applications. *F1000Research* 5, 1963. doi:10.12688/F1000RESEARCH.9357.2

Binder, C., Matthes, K. L., Korol, D., Rohrmann, S., and Moch, H. (2018). Cancer of unknown primary—epidemiological trends and relevance of comprehensive genomic profiling. *Cancer Med.* 7, 4814–4824. doi:10.1002/cam4.1689

Borlak, J., Meier, T., Halter, R., Spanel, R., and Spanel-Borowski, K. (2005). Epidermal growth factor-induced hepatocellular carcinoma: gene expression profiles in precursor lesions, early stage and solitary tumours. *Oncogene* 2411 (24), 1809–1819. doi:10.1038/sj.onc.1208196

Chen, S., Zhou, W., Tu, J., Li, J., Wang, B., Mo, X., et al. (2021). A novel XGBoost method to infer the primary lesion of 20 solid tumor types from gene expression data. *Front. Genet.* 12, 632761. doi:10.3389/fgene.2021.632761

Conway, A.-M., Mitchell, C., Kilgour, E., Brady, G., Dive, C., and Cook, N. (2019). Molecular characterisation and liquid biomarkers in Carcinoma of Unknown Primary (CUP): taking the "U" out of "CUP". *Br. J. Cancer* 120, 141–153. doi:10.1038/s41416-018-0332-2

de Guia, J. M., Devaraj, M., and Leung, C. K. (2019). "DeepGx: deep learning using gene expression for cancer classification," in Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, New York, NY, USA (ACM), 913–920. doi:10.1145/3341161.3343516

Divate, M., Tyagi, A., Richard, D. J., Prasad, P. A., Gowda, H., and Nagaraj, S. H. (2022). Deep learning-based pan-cancer classification model reveals tissue-of-origin specific gene expression signatures. *Cancers (Basel)* 14, 1185. doi:10.3390/cancers14051185

Dolgalev, I. (2022). Msigdbr: MSigDB gene sets for multiple organisms in a tidy data format, R package version 7.5.1.9001. Available at: https://igordot.github.io/msigdbr/ [Accessed August 29, 2023].

Edwards, S., Campbell, C., Flohr, P., Shipley, J., Giddings, I., Te-Poele, R., et al. (2005). Expression analysis onto microarrays of randomly selected cDNA clones highlights HOXB13 as a marker of human prostate cancer. *Br. J. Cancer* 92, 376–381. doi:10.1038/sj.bjc.6602261

Ferreira, C. (2002). "Gene expression programming in problem solving," in *Soft computing and industry* (London: Springer London), 635–653. doi:10.1007/978-1-4471-0123-9_54

Galea, D., Inglese, P., Cammack, L., Strittmatter, N., Rebec, M., Mirnezami, R., et al. (2017). Translational utility of a hierarchical classification strategy in biomolecular data analytics. *Sci. Rep.* 7, 14981. doi:10.1038/s41598-017-14092-7

Grewal, J. K., Tessier-Cloutier, B., Jones, M., Gakkhar, S., Ma, Y., Moore, R., et al. (2019). Application of a neural network whole transcriptome–based pan-cancer method for diagnosis of primary and metastatic cancers. *JAMA Netw. Open* 2, e192597. doi:10.1001/jamanetworkopen.2019.2597

He, B., Lang, J., Wang, B., Liu, X., Lu, Q., He, J., et al. (2020a). TOOme: a novel computational framework to infer cancer tissue-of-origin by integrating both gene mutation and expression. *Front. Bioeng. Biotechnol.* 8, 394. doi:10.3389/fbioe.2020.00394

He, B., Zhang, Y., Zhou, Z., Wang, B., Liang, Y., Lang, J., et al. (2020b). A neural network framework for predicting the tissue-of-origin of 15 common cancer types based on RNA-seq data. *Front. Bioeng. Biotechnol.* 8, 737. doi:10.3389/fbioe.2020.00737

Hong, J., Hachem, L. D., and Fehlings, M. G. (2022). A deep learning model to classify neoplastic state and tissue origin from transcriptomic data. *Sci. Rep.* 12, 9669. doi:10.1038/s41598-022-13665-5

International Cancer Genome Consortium, Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., et al. (2010). International network of cancer genome projects. *Nature* 464, 993–998. doi:10.1038/nature08987

Jones, S., Beyers, M., Shukla, M., Xia, F., Brettin, T., Stevens, R., et al. (2022). TULIP: an RNA-seq-based primary tumor type prediction tool using convolutional neural networks. *Cancer Inf.* 21, 11769351221139491. doi:10.1177/11769351221139491

Liu, H., Qiu, C., Wang, B., Bing, P., Tian, G., Zhang, X., et al. (2021). Evaluating DNA methylation, gene expression, somatic mutation, and their combinations in inferring tumor tissue-of-origin. *Front. Cell Dev. Biol.* 9, 619330. doi:10.3389/fcell.2021.619330

Lyu, B., and Haque, A. (2018). "Deep learning based tumor type classification using gene expression data," in Proceedings of the 2018 ACM International Conference on Bioinformatics Computational Biology, and Health Informatics, New York, NY, USA (ACM), 89–96. doi:10.1145/3233547.3233588

Moiso, E., Farahani, A., Marble, H. D., Hendricks, A., Mildrum, S., Levine, S., et al. (2022). Developmental deconvolution for classification of cancer origin. *Cancer Discov.* 12, 2566–2585. doi:10.1158/2159-8290.CD-21-1443

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

Pham, T. H., and Raahemi, B. (2023). Bio-inspired feature selection algorithms with their applications: a systematic literature review. *IEEE Access* 11, 43733–43758. doi:10.1109/ACCESS.2023.3272556

Ramroach, S., Joshi, A., and John, M. (2020). Optimisation of cancer classification by machine learning generates an enriched list of candidate drug targets and biomarkers. *Mol. Omi.* 16, 113–125. doi:10.1039/C9MO00198K

Rassy, E., Assi, T., and Pavlidis, N. (2020). Exploring the biological hallmarks of cancer of unknown primary: where do we stand today? *Br. J. Cancer* 122, 1124–1132. doi:10.1038/s41416-019-0723-z

Rassy, E., and Pavlidis, N. (2020). Progress in refining the clinical management of cancer of unknown primary in the molecular era. *Nat. Rev. Clin. Oncol.* 17, 541–554. doi:10.1038/s41571-020-0359-1

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Ca. Cancer J. Clin.* 71, 209–249. doi:10.3322/caac.21660

Swanson, K., Wu, E., Zhang, A., Alizadeh, A. A., and Zou, J. (2023). From patterns to patients: advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. *Cell* 186, 1772–1791. doi:10.1016/j.cell.2023.01.035

Tufail, A. B., Ma, Y.-K., Kaabar, M. K. A., Martínez, F., Junejo, A. R., Ullah, I., et al. (2021). Deep learning in cancer diagnosis and prognosis prediction: a minireview on challenges, recent trends, and future directions. *Comput. Math. Methods Med.* 2021, 1–28. doi:10.1155/2021/9025470

Ueno, T., Linder, S., and Elmberger, G. (2003). Aspartic proteinase napsin is a useful marker for diagnosis of primary lung adenocarcinoma. *Br. J. Cancer* 88, 1229–1233. doi:10.1038/sj.bjc.6600879

Vibert, J., Pierron, G., Benoist, C., Gruel, N., Guillemot, D., Vincent-Salomon, A., et al. (2021). Identification of tissue of origin and guided therapeutic applications in cancers of unknown primary using deep learning and RNA sequencing (TransCUPtomics). *J. Mol. Diagn.* 23, 1380–1392. doi:10.1016/j.jmoldx.2021.07.009

Wagner, A. H., Coffman, A. C., Ainscough, B. J., Spies, N. C., Skidmore, Z. L., Campbell, K. M., et al. (2016). DGIdb 2.0: mining clinically relevant drug–gene interactions. *Nucleic Acids Res.* 44, D1036–D1044. doi:10.1093/nar/gkv1165

Wei, I. H., Shi, Y., Jiang, H., Kumar-Sinha, C., and Chinnaiyan, A. M. (2014). RNA-seq accurately identifies cancer biomarker signatures to distinguish tissue of origin. *Neoplasia* 16, 918–927. doi:10.1016/j.neo.2014.09.007

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45, 1113–1120. doi:10.1038/ng.2764

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innov* 2, 100141. doi:10.1016/j.xinn.2021.100141

Yaqoob, A., Aziz, R. M., Verma, N. K., Lalwani, P., Makrariya, A., and Kumar, P. (2023). A review on nature-inspired algorithms for cancer disease prediction and classification. *Mathematics* 11, 1081. doi:10.3390/math11051081

Zanella, L., Facco, P., Bezzo, F., and Cimetta, E. (2022). Feature selection and molecular classification of cancer phenotypes: a comparative study. *Int. J. Mol. Sci.* 23, 9087. doi:10.3390/ijms23169087

Zhao, Y., Mehta, M., Walton, A., Talsania, K., Levin, Y., Shetty, J., et al. (2019). Robustness of RNA sequencing on older formalin-fixed paraffin-embedded tissue from high-grade ovarian serous adenocarcinomas. *PLoS One* 14, e0216050. doi:10.1371/journal.pone.0216050

Zhao, Y., Pan, Z., Namburi, S., Pattison, A., Posner, A., Balachander, S., et al. (2020). 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:10.1016/j.ebiom.2020.103030