Machine Learning for Predicting Cancer Severity

Alex Qin
Research School of Computer Science
Australian National University
Canberra, Australia
alex.qin@anu.edu.au

Md Rakibul Hasan
Department of Electrical and Electronic Engineering
BRAC University
Dhaka, Bangladesh
rakibul.hasan@bracu.ac.bd

Khandaker Asif Ahmed
CSIRO L&W, Black Mountain
Canberra, Australia
khandakerasif.ahmed@csiro.au

Md Zakir Hossain
Biological Data Science Institute, Australian National University
and CSIRO A&F, Black Mountain
Canberra, Australia
zakir.hossain@anu.edu.au

2022 IEEE 10th International Conference on Healthcare Informatics (ICHI) | 978-1-6654-6845-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICHI54592.2022.00098

***Abstract*—Cancer is an extremely heterogeneous disease, and this property becomes increasingly exacerbated as the disease progresses. Its heterogeneity can be reflected by protein signature, called its proteome, which is essentially a dataset indicating which proteins are highly expressed or lowly expressed in a tumour. Whilst no two cancer proteomes are the same, various patterns in the proteome can help distinguish certain cancers from others. There are prominent proteomic patterns that a Machine Learning (ML) technique can pick from different cancer types. However, identifying severity patterns across cancer types is challenging due to major proteomic differences interfering with ML performance. Accordingly, proteomic analyses are rarely performed on datasets consisting of multiple cancer types unless aiming to distinguish between two types. In this study, we tested three ML algorithms in classifying the TCGA (The Cancer Genome Atlas) PanCancer dataset, consisting of 32 different cancer types, into various clinically relevant metrics, such as stage, grade, and treatment response of tumour. On average, we achieved the best accuracies when employing a Support Vector Machine (SVM) classifier with a Radial Basis Function (RBF). The highest accuracies were accomplished when classifying based on pathological stage, suggesting a possible future application as a diagnostic tool, where cancers can be staged based on a quick ML classification rather than a lengthy evaluation by a pathologist.**

***Index Terms*—Cancer, Proteomics, PanCancer Datasets, Machine Learning, Classification**

I. Introduction

Cancer is an extremely heterogeneous disease, differing dramatically among different types/grades and individual patients [1]. This stubborn nature of cancer is especially pronounced in its advanced forms, while early cancer is highly treatable, with a wide range of different therapeutics for different cancers [2]. In general, cancer is classified into four stages and four grades: stage indicates severity, with stage 4 being most severe; grade determines tumour histology and differentiation of cancerous cells from their surrounding cells [3]. Hence, it is crucial to determine the exact stage or grade of a particular cancer with early diagnoses and screenings so that doctors can concentrate their efforts on either curative or palliative treatments.

Currently, numerous biomedical tools or markers are being used for cancer diagnosis and tumour staging [4]. Laboratory tests such as tissue biopsies, blood tests, or imaging-based tests (CT/MRI scans) help practitioners in clinical staging and grading but are costly and time-consuming, requiring lengthy professional examinations by trained pathologists [4]. Ultimately, a method to provide faster feedback would significantly contribute to both patients' well-being and a relief for the burden on the existing health systems.

Machine learning (ML) models can serve as a powerful tool to identify common patterns among different cancers. Several studies have applied ML algorithms to numerical datasets of classical clinical features, such as the size and spread of a specific tumour [5]. Whilst ML models have been applied for single cancer types, they have often been neglected on the relatively more complex task of identifying patterns in proteomic datasets containing multiple cancer types altogether [5]. Complexity arises due to the extreme variations among different cancer types, which often interferes with the model's accuracy [6]. For example, a high variation of estrogen receptor expression between a stage 4 breast cancer and a stage 1 lung cancer can easily be mistaken as a vital feature in an ML model classifying cancers based on severity.

An accurate classifier for multiple cancers could identify common severity markers of several cancers; these markers can then be characterised as potential therapeutic targets for multiple cancers. Way et al. [7] utilised TCGA (The Cancer Genome Atlas) PanCancer dataset of a few cancers to detect anomalies in a specific gene responsible for mutation in cell growth and proliferation. As a result of using PanCancer data, their model can be applied to any tumour. This proves the potential of using the TCGA PanCancer dataset of multiple cancers altogether to provide generalised insights into any cancer type. To the best of our knowledge, no studies have been performed on the PanCancer dataset to categorise it into a metric as broad as pathological staging or grading. In this study, we aim to use three ML classifiers on the TCGA PanCancer dataset to classify cancer grades and severity irrespective of cancer types. Identifying common proteins and traits between cancers would advance numerous treatment options and open up the possibility of global cancer treatments rather than chemotherapy and invasive surgeries.

2575-2634/22/$31.00 ©2022 IEEE
DOI 10.1109/ICHI54592.2022.00098

II. Methods

A. Dataset Information

We used the TCGA clinical data resources [8] and the TCGA PanCancer proteomic dataset [9], which contains protein expression data of 32 different cancer types. A unique ID is provided for each participant, allowing to correlate clinical metadata to the proteomic data. Clinical metadata included data on various clinical metrics for 11,161 cancer patients. The TCGA proteomic data included 259 proteins commonly modulated in cancer from 7695 patients.

B. Data Filtering

Patient IDs found in either one dataset or the other, not in both, were excluded, which resulted in 7,632 samples with both proteomic and clinical annotations. However, proteomic profiles of certain proteins were not elucidated in particular cancers, where a "NaN" value was located. All proteins that contained "NaN" values in any sample were excluded. This resulted in the exclusion of 50 proteins, resulting in 209 proteins being left for analysis. The clinical dataset did not always have available data for all patients for all evaluated clinical metrics. Hence, samples were excluded where data was unavailable, not applicable, unknown, or had a discrepancy for the particular clinical metric.

Finally, the data were divided into different classes depending on the metric. For pathological staging prediction, data were grouped into four classes (S1/S2/S3/S4), where substages were merged into corresponding stages; for example, stages 2A, 2B, and 2C were merged under stage 2. For histological grade prediction, data were again grouped into four classes (G1/2/3/4). For treatment outcome prediction, data were grouped into two classes (responsive or not responsive). The cancer was responsive if the tumour was partially (partial remission) or completely removed (complete remission) following treatment. The cancer was unresponsive if the cancer had grown further despite treatment (progressive disease) or remained the same size despite treatment (stable disease). Due to class imbalances, as there were approximately four times as many responsive cancers as unresponsive, the responsive category was undersampled to 25% of its original size.

C. ML Algorithms

Three different ML algorithms were considered in this study: SVM, RF (Random Forest), and MLP. Using GridSearchCV tuning algorithm from *scikit-learn* python package, several hyperparameters were optimised: "C", "gamma", and "kernel" for SVM; "n_estimators", "max_features", "max_depth", and "criterion" for RF; "hidden_layer_sizes", "activation", "solver", "alpha", and "learning_rate" for MLP.

The number of hidden layers was fixed at 1. The three algorithms were applied individually to the complete dataset. The training and test sets were split into a 70/30 ratio for all algorithms. Where necessary, due to class imbalances—where the sample size of one class was notably larger than another—the class with the higher number of samples was undersampled. Undersampling involved taking a random percentage of the class with higher sample numbers to balance the sample numbers between classes. We conducted 5-fold cross-validation for all ML models. To evaluate the ML models, we used accuracy score and confusion matrices to attain information regarding incorrectly classified classes.

III. Results and Discussion

A. Classifying Cancers—Pathological Stage

First, classification between S1 and S4 yielded a relatively high accuracy of 80.16% (Fig. 1a) by the MLP classifier, suggesting that the proteomics data of a cancer potentially reflects its pathological stage.

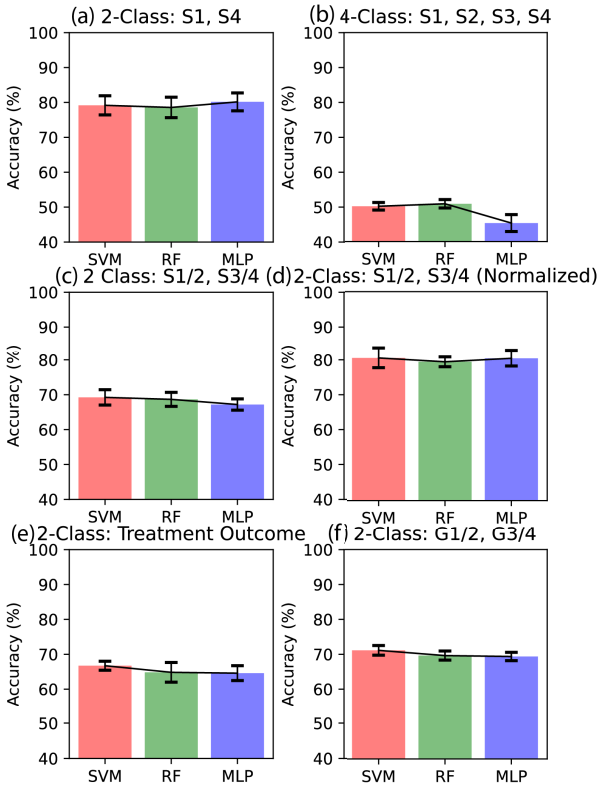

Fig. 1. Accuracies of ML algorithms in categorizing cancers.

Given the promising results, we next categorised all cancers into four stages. Here, the performance was relatively low across all algorithms, with RF achieving the highest accuracy of approximately 50.83% (Fig. 1b).

To test the theory that perhaps there are strong proteomic similarities between the more severe stages (S3 and S4) and the same between the less severe stages (S1 and S2), we split the dataset for a 2-class classification of S1 and S2 combined in one class and S3 and S4 in another. This initially gave relatively poor results, as performance ranged from 60-70% accuracy in all three algorithms (Fig. 1c). As S1/2 had far greater samples than S3/4, S1/2 were undersampled: one-third of the original samples in class S1/2 were randomly

selected to correct for the class imbalances. The algorithms were run again, yielding accuracies of 80.96%, with the best performance achieved by SVM (Fig. 1d). Such high classification performance indicates that proteomic signatures between adjacent stages (S3 and S4, or S1 and S2) may not differ significantly, but they differ significantly between early and severe stages. Further validation of this finding would have notable implications for cancer diagnostics. This would enable us to categorise and stage a cancer as either S1/2 or S3/4, significantly improving diagnosis efficiency by avoiding time-consuming pathologist reports.

B. Classifying Cancers—Treatment Outcome

Despite normalisation of cancer staging and grading, these classification metrics are prone to subjectivity as human errors can occur in pathology and histology reports. We, therefore, applied ML on a more objective metric—cancer's responsiveness to the first course of treatment.

We found no significant improvements in the algorithm performance despite using the more objective classification metric. Instead, performance decreased to 66.72%, with the best accuracy achieved by the SVM classifier with an RBF (Radial Basis Function) kernel (Fig. 1e). This suggests that there are few proteomic differences between responsive and unresponsive cancers. However, several other confounding factors exist, most glaring of which are the lack of information on a patient's exact treatment. The nature of treatment type and course undoubtedly affects treatment outcome but is not reported in the dataset. Furthermore, the mixing of both more severe cancers and less severe cancers may further affect model performance. However, each of these factors is not without remedy—with further cleaning of the dataset, a better performing model could be achieved.

C. Classifying Cancers—Histological Grade

Additionally, we classified cancers based on histological grades. Due to large class imbalances, we merged Grade 1 and 2 (G1 and G2) in one class and Grade 3 and 4 (G3 and G4) into another. Classification performance increased compared to the 4-stage classification but decreased compared to the 2-stage (S1/2 and S3/4) classification. Here, the SVM classifier with an RBF kernel achieved the highest classification of 71.06% (Fig. 1f). Such findings suggest that proteomic data are not necessarily a strong determining factor for the histological grade. Instead, the histological grade may be governed by external factors such as the shear force on the tumour and the tumour micro-environment. Further study should affirm this finding but would be an exciting avenue for future exploration.

IV. Conclusion and Future Work

Given the relative difficulty of categorising a proteomic Pan-Cancer dataset because of its variation (heterogeneity) among cancer types, this study evaluated whether ML algorithms can distinguish common severity patterns. It holds notable implications in cancer diagnosis: a potential future where pathologists and clinicians can have an ML-based tool for diagnosing and predicting any cancer types and its severity levels. We found prominent proteomic differences between pathological stages encompassing all cancer types, reflecting potentially there could be common drug targets for many different cancers. If verified, this could greatly simplify drug discovery—instead of looking for multiple drugs to target multiple cancers, we can focus on a single drug capable of targeting multiple different cancer types. Further polishing the ML training and data pre-cleaning/analysis could eventually make the model an effective tool for predicting treatment outcomes, which easily holds massive potential in helping clinicians to find the best possible treatment for a patient. Feature selection was not performed in this study but is undoubtedly an immediate future direction that could be implemented, potentially further raising ML performance and identifying common proteomic pathways present in multiple cancer types.

References

[1] P. L. Bedard, A. R. Hansen, M. J. Ratain, and L. L. Siu, "Tumour heterogeneity in the clinic," *Nature*, vol. 501, no. 7467, pp. 355–364, Sep. 2013.

[2] M. M. Koo, R. Swann, S. McPhail, G. A. Abel, L. Elliss-Brookes, G. P. Rubin, and G. Lyratzopoulos, "Presenting symptoms of cancer and stage at diagnosis: evidence from a cross-sectional, population-based study," *Lancet Oncol.*, vol. 21, no. 1, pp. 73–79, Jan. 2020.

[3] S. M. Telloni, "Tumor staging and grading: A primer," *Methods Mol. Biol.*, vol. 1606, pp. 1–17, 2017.

[4] C. Pucci, C. Martinelli, and G. Ciofani, "Innovative approaches for cancer treatment: Current perspectives and new challenges," *Ecancermedicalscience*, vol. 13, 2019.

[5] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015.

[6] H. J. Johansson, F. Socciarelli, N. M. Vacanti, M. H. Haugen, Y. Zhu, I. Siavelis, A. Fernandez-Woodbridge, M. R. Aure, B. Sennblad, M. Vesterlund, R. M. Branca, L. M. Orre, M. Huss, E. Fredlund, E. Beraki, Ø. Garred, J. Boekel, T. Sauer, W. Zhao, S. Nord, E. K. Höglander, D. C. Jans, H. Brismar, T. H. Haukaas, T. F. Bathen, E. Schlichting, B. Naume, Consortia Oslo Breast Cancer Research Consortium (OSBREAC), T. Luders, E. Borgen, V. N. Kristensen, H. G. Russnes, O. C. Lingjærde, G. B. Mills, K. K. Sahlberg, A.-L. Børresen-Dale, and J. Lehtiö, "Breast cancer quantitative proteome and proteogenomic landscape," *Nat. Commun.*, vol. 10, no. 1, p. 1600, Apr. 2019.

[7] G. P. Way, F. Sanchez-Vega, K. La, J. Armenia, W. K. Chatila, A. Luna, C. Sander, A. D. Cherniack, M. Mina, G. Ciriello *et al.*, "Machine learning detects pan-cancer ras pathway activation in the cancer genome atlas," *Cell reports*, vol. 23, no. 1, pp. 172–180, 2018.

[8] J. Liu, T. Lichtenberg, K. A. Hoadley, L. M. Poisson, A. J. Lazar, A. D. Cherniack, A. J. Kovatich, C. C. Benz, D. A. Levine, A. V. Lee *et al.*, "An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics," *Cell*, vol. 173, no. 2, pp. 400–416, 2018.

[9] J. Li, Y. Lu, R. Akbani, Z. Ju, P. L. Roebuck, W. Liu, J.-Y. Yang, B. M. Broom, R. G. Verhaak, D. W. Kane *et al.*, "Tcpa: a resource for cancer functional proteomics data," *Nature methods*, vol. 10, no. 11, pp. 1046–1047, 2013.