medical sciences forum

MDPI

Proceeding Paper

Predicting Cancer Prognostics from Tumour Transcriptomics Using an Auto Machine Learning Approach [†]

Ricardo Jorge Pais [1,2,*], Filipa Lopes [3], Inês Parreira [3], Márcia Silva [3], Mariana Silva [3] and Maria Guilhermina Moutinho [2]

[1] Bioenhancer Systems, Office 63 182-184 High Street North, East Ham, London E6 2JA, UK
[2] Egas Moniz Center for Interdisciplinary Research, Egas Moniz School of Health & Science, 2829-511 Almada, Portugal; gmoutinho@egasmoniz.edu.pt
[3] Egas Moniz School of Health & Science, 2929-511 Almada, Portugal; filipa.lopes2000@gmail.com (F.L.); ines.parreira03@gmail.com (I.P.); marciasofyfs@gmail.com (M.S.); marianarodriguessasilva123@gmail.com (M.S.)
[*] Correspondence: rjpais@bioenhancersystems.com

[†] Presented at the 6th International Congress of CiiEM—Immediate and Future Challenges to Foster One Health, Almada, Portugal, 5–7 July 2023.

Abstract: Cancer prognostics using tumour transcriptomics is a promising precision medicine approach for helping decisions during cancer treatment. However, currently used cancer prognostic biomarkers still have low predictive power. This work tested the potential of applying machine learning (ML) algorithms for generating patients' survival prognostics on lung, breast, and kidney tumour transcriptomics datasets. We evaluated the performance of models generated by ML and reported their optimal sensitivity, specificity, accuracy, and computed ROC-AUC. The results support the potential for applying auto ML approaches for the future development of cancer prognostics tools based on transcriptomics data.

Keywords: bioinformatics; cancer prognostics; machine learning; transcriptomics

check for updates

Citation: Pais, R.J.; Lopes, F.; Parreira, I.; Silva, M.; Silva, M.; Moutinho, M.G. Predicting Cancer Prognostics from Tumour Transcriptomics Using an Auto Machine Learning Approach. *Med. Sci. Forum* **2023**, *22*, 6. https://doi.org/10.3390/msf2023022006

Academic Editors: José Brito, Nuno Taveira and Ana I. Fernandes

Published: 8 August 2023

1. Introduction

Cancer is a worldwide flagellum, leading to millions of deaths per year. The success of cancer treatments often depends on the choice of the correct treatment [1]. Treatment success is associated with tumour heterogenicity and genetic factors. Cancer prognostic biomarkers are considered a promising personalized medicine approach for helping decision-making during cancer treatment [2]. Cancer prognostic biomarkers still have low predictive power, explaining only 25% to 75% of the cases [2]. Transcriptomics is an affordable and accurate high-throughput methodology often described as a promising precision medicine approach that enables the quantification of gene expression levels of multiple genes [3]. The application of machine learning (ML) frameworks on transcriptomics data is thought to have the potential to identify biomarker signatures for a binary classification (yes/no) of patient's survival with predictive power [4,5]. However, this approach is still not applied as a solution for treatment prognostics.

Multiple ML algorithms can be applied in bioinformatics to combine biomarkers to improve models' predictive capacity [5,6]. Further, there is an infinite possibility of models that can arise from these algorithms due to all possible parameter combinations making it hard and labour-intensive to find optimal models. Auto ML approaches have proven useful for the optimal model generation with reasonable computational effort and provide a much faster route to achieving better-performing models [7,8]. This work used an auto ML approach to test the potential of ML for generating transcriptomics-based cancer prognostic predictors for lung, breast, and kidney cancers transcriptomics datasets. Here, the model's performance was evaluated and reported their optimal sensitivity, specificity, accuracy, and computed ROC-AUC.

2. Methods

2.1. Data Collection

Transcriptomics data were collected from the 2021 updated Human Protein Atlas database records, which contained mRNA expression (FPKM) of 200 genes from 1075 anonymized cancer patients [9,10]. The TCGA transcriptomics data of breast, lung and renal cancers biopsies were downloaded from these sources. Metadata including patients' age, sex, survival time after biopsy and time of death were also collected from this source.

2.2. Dataset Construction

From collected transcriptomics data, 58 genes were selected. These 58 genes are key components of signalling pathways involved in the regulation of epithelial-to-mesenchymal transition, which plays a critical role in metastasis acquisition [11]. From the collected metadata, we select only the transcriptomes associated with patients who have been reported to survive over 5 years after the diagnostics (good prognostics) or with a reported death within the first 2 years (poor prognostics). The sample numbers of the final curated datasets used in this work are summarized in Table 1.

Table 1. Cancer transcriptomics datasets and their sample numbers.

Tumour Biopsy	Number of Patients	Good Prognostics	Poor Prognostics
Breast	239	199	40
Lung	325	94	231
Kidney	318	210	108

2.3. Auto Machine Learning Framework

Models were generated using the Tree-Based Pipeline Optimization Tool (TPOT). TPOT is an open-source software package developed in Python for an automated generation of ML-derived predictive models [7,8]. TPOT relies on genetic programming to generate predictive models with optimal performance, testing multiple ML algorithms and modelling parameters [8]. The open-source TPOT version 0.11.7 was installed under Python 3.9 anaconda distribution. TPOT auto ML pipeline was implemented and run under the Jupyter notebook environment. All scripts were run on a MacBook pro with a 2.4 GHz 8-Core Intel Core i9 processor.

2.4. Model Generation and Evaluation

The TPOT Classifier method was used for model training, testing and optimization. It was set up to perform 100 generations with a population size of 50 randomly selected ML algorithms. Optimization criterium was set to find the optimal Receiver Operating Curve (ROC) given by the Area Under the Curve (AUC) value. A random selection of 50% of the data were used in training and the remaining for testing [12]. Models' final performance was computed by generating predictions using the selected models on all data selected for training, and then calculating the accuracy, sensitivity, specificity and ROC-AUC [13].

3. Results

We applied the auto ML framework (TPOT) on the curated datasets of breast, lung, and kidney tumour transcriptomics (Table 1) for generating patient survival prognostics. TPOT ran for about 1 h on each dataset, generating and testing an average of 2.6 models/second, evaluating a total of 10,000 different variants that use and combine distinct ML algorithms (e.g., Random Forest, Naïve Bayes, Neural Networks, and many others). The best-performing algorithms selected were substantially distinct among cancer types with different performances (Table 2).

Table 2. Generated optimal predictive models and their associated performances.

Tumour Biopsy	Algorithm Pipeline	SEN *	SPE	AUC	ACC
Breast	Multinomial Naïve Bayes with Random Forest	94%	58%	53%	84%
Lung	KNeigbours with Random Forest	59%	83%	48%	52%
Kidney	Normalized Random Forest	94%	66%	70%	71%

* SEN (sensitivity); SPE (specificity); AUC (Area Under the Curve); and ACC (accuracy).

The results obtained (Table 2) showed that predicted models generated for breast and kidney cancer prognostics performed with very good sensitivity (SEN = 94%). However, these models had poor specificity (SPE < 66%), which indicates a huge tendency to generate false positives if applied to predict the survival of a patient with a tumour [13]. In contrast, the predicted model obtained for lung cancer prognostics had a reasonable specificity (SPE > 83%) but with a poor sensitivity (SEN = 59%), indicating a high tendency for generating false negatives [13]. Further, the obtained ROC-AUCs showed that only the kidney cancer prognostic model has good predictive power (AUC > 70%) with reasonable accuracy (ACC > 70%), indicating that only this model, among all, generates robust predictions not given by chance [13].

The performances obtained from the models generated by TPOT (Table 2) also showed that the currently available ML algorithms are not enough to generate high-performance models on our cancer transcriptomics datasets. This low performance may be explained by cancer heterogenicity, missing key regulatory genes on the dataset, or confounding variables associated with the clinical data (age, gender, ethnicity, death reasons, and treatment choices).

4. Conclusions

This study demonstrated that the auto ML approach is a powerful methodology for the fast and systematic generation of predictive models that can be applied in cancer prognostics from tumour transcriptomics. Here, we illustrated the ML approach application with three types of cancers that showed promising performances, particularly for kidney cancer. Moreover, the results in this work support the idea of technical challenges in this modelling framework that justify future work for improving either the data or the tools to generate predictive models.

Author Contributions: Conceptualization, R.J.P. and M.G.M.; methodology, R.J.P.; formal analysis, R.J.P., F.L., I.P., M.S. (Mariana Silva) and M.S. (Márcia Silva); investigation, F.L., I.P., M.S. (Mariana Silva) and M.S. (Márcia Silva); data curation, F.L., I.P., M.S. (Mariana Silva) and M.S. (Márcia Silva); writing—original draft preparation, R.J.P.; writing—review and editing, M.G.M.; supervision, R.J.P. and M.G.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of Bioenhancer Systems Ltd. on the 21 February 2023.

Informed Consent Statement: All data involving humans used in this study were anonymous and retrieved from public databases, where informed consent was given.

Data Availability Statement: Publicly available datasets on breast, lung and renal cancer transcriptomics were analysed in this study. The original collected data (rna_cancer_sample.tsv.zip) can be found here: https://www.proteinatlas.org/about/download (accessed on 12 September 2021). The curated datasets used in the study can be found here: https://www.bioenhancersystems.com/resources (accessed on 12 September 2021). Medical data are not publicly available due to privacy restrictions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hanahan, D.; Weinberg, R.A. Hallmarks of Cancer: The Next Generation. *Cell* **2011**, *144*, 646–674. [CrossRef] [PubMed]
2. Šutić, M.; Vukić, A.; Baranašić, J.; Försti, A.; Džubur, F.; Samaržija, M.; Jakopović, M.; Brčić, L.; Knežević, J. Diagnostic, Predictive, and Prognostic Biomarkers in Non-Small Cell Lung Cancer (NSCLC) Management. *J. Pers. Med.* **2021**, *11*, 1102. [CrossRef] [PubMed]
3. Van Allen, E.M.; Robinson, D.; Morrissey, C.; Pritchard, C.; Imamovic, A.; Carter, S.; Rosenberg, M.; McKenna, A.; Wu, Y.M.; Cao, X.; et al. A Comparative Assessment of Clinical Whole Exome and Transcriptome Profiling across Sequencing Centers: Implications for Precision Cancer Medicine. *Oncotarget* **2016**, *7*, 52888–52899. [CrossRef] [PubMed]
4. De Ridder, D.; De Ridder, J.; Reinders, M.J.T. Pattern Recognition in Bioinformatics. *Brief. Bioinform.* **2013**, *14*, 633–647. [CrossRef] [PubMed]
5. Mann, M.; Kumar, C.; Zeng, W.F.; Strauss, M.T. Artificial Intelligence for Proteomics and Biomarker Discovery. *Cell Syst.* **2021**, *12*, 759–770. [CrossRef] [PubMed]
6. Pais, R.J. Predictive Modelling in Clinical Bioinformatics: Key Concepts for Startups. *BioTech* **2022**, *11*, 35. [CrossRef] [PubMed]
7. Le, T.T.; Fu, W.; Moore, J.H. Scaling Tree-Based Automated Machine Learning to Biomedical Big Data with a Feature Set Selector. *Bioinformatics* **2020**, *36*, 250–256. [CrossRef] [PubMed]
8. Olson, R.S.; Urbanowicz, R.J.; Andrews, P.C.; Lavender, N.A.; Kidd, L.C.; Moore, J.H. Automating Biomedical Data Science Through Tree-Based Pipeline Optimization. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Cham, Switzerland, 2016; Volume 9597, pp. 123–137.
9. Pontén, F.; Jirström, K.; Uhlen, M. The Human Protein Atlas-a Tool for Pathology. *J. Pathol.* **2008**, *216*, 387–393. [CrossRef] [PubMed]
10. Uhlen, M.; Oksvold, P.; Fagerberg, L.; Lundberg, E.; Jonasson, K.; Forsberg, M.; Zwahlen, M.; Kampf, C.; Wester, K.; Hober, S.; et al. Towards a Knowledge-Based Human Protein Atlas. *Nat. Biotechnol.* **2010**, *28*, 1248–1250. [CrossRef] [PubMed]
11. Pais, R.J. Simulation of Multiple Microenvironments Shows a Pivot Role of RPTPs on the Control of Epithelial-to-Mesenchymal Transition. *Biosystems* **2020**, *198*, 104268. [CrossRef] [PubMed]
12. Swan, A.L.; Mobasheri, A.; Allaway, D.; Liddell, S.; Bacardit, J. Application of Machine Learning to Proteomics Data: Classification and Biomarker Identification in Postgenomics Biology. *Omics J. Integr. Biol.* **2013**, *17*, 595–610. [CrossRef] [PubMed]
13. Dankers, F.J.W.M.; Traverso, A.; Wee, L.; van Kuijk, S.M.J. Prediction Modeling Methodology. In *Fundamentals of Clinical Data Science*; Springer International Publishing: Cham, Switzerland, 2019; pp. 101–120.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.