scientific reports

Check for updates

OPEN

A deep learning model to classify neoplastic state and tissue origin from transcriptomic data

James Hong[1,4], Laureen D. Hachem[1,2,4] & Michael G. Fehlings[1,2,3]✉

Application of deep learning methods to transcriptomic data has the potential to enhance the accuracy and efficiency of tissue classification and cell state identification. Herein, we developed a multitask deep learning model for tissue classification combining publicly available whole transcriptomic (RNA-seq) datasets of non-neoplastic, neoplastic and peri-neoplastic tissue to classify disease state, tissue origin and neoplastic subclass. RNA-seq data from a total of 10,116 patient samples processed through a common pipeline were used for model training and validation. The model achieved 99% accuracy for disease state classification (ROC-AUC of 0.98) and 97% accuracy for tissue origin (ROC-AUC of 0.99). Moreover, the model achieved an accuracy of 92% (ROC-AUC 0.95) for neoplastic subclassification. This is the first multitask deep learning algorithm developed for tissue classification employing a uniform pipeline analysis of transcriptomic data with multiple tissue classifiers. This model serves as a framework for incorporating large transcriptomic datasets across conditions to facilitate clinical diagnosis and cell-based treatment strategies.

Accurate and expeditious tissue classification is central to the practice of clinical medicine and biological research. Disease diagnostics, subclassification and treatment decision-making rely heavily on interpreting the identity and status of patient tissue samples[1]. Moreover, identification of cellular phenotype, stress-state and viability is critical in translating cell-based transplantation strategies to clinical practice. Methods to analyze cellular identity and tissue health have traditionally relied on a finite number of histological markers and imaging characteristics. However, advances in sequencing technology over the last decade have transformed our ability to probe tissue and expanded the availability of transcriptomic data. The cellular transcriptome provides insight into both tissue identity and response to local environmental factors, which offers a more accurate assessment of tissue status as compared to conventional histological measures[2]. While transcriptomic analyses have been employed in select diseases or cell populations[3–7], there remains a significant need to integrate the numerous sequencing datasets available in order to probe multiple features of a tissue sample including disease state, origin, and subclass. Artificial intelligence strategies offer a promising approach to address this need and to integrate this valuable resource into clinical and research practice.

Deep learning approaches have been increasingly incorporated into clinical medicine. Unlike standard machine learning, deep learning offers the ability to train on multiple layers of neural nets, therefore affording greater flexibility and the generation of more accurate models that allow for the identification of complex patterns and granular subtyping[8]. To date, applications of deep learning methods have primarily been employed within specific disease types or in the processing of histological or radiographic images[3,9]. Recently, deep learning methods have begun to be applied to transcriptomic sequencing data in order to better understand heterogeneity in tissue samples[10–13]. Previous attempts to develop a comprehensive model based on transcriptomic data have often used standard machine learning approaches rather than multilayer neural networks[10], and the use of non-uniform pipeline analyses of transcriptomic data processing thus lead to significant confounds in model training and output[14]. Furthermore, many previous models have been restricted to a single classifier of tissue identity[10,13], and therefore do not capture the spectrum of disease states. Specifically, models do not distinguish non-neoplastic tissue from peri-neoplastic tissue despite there being significant differences in gene expression and microenvironment between these sample types[15]. As such, the development of an accurate and efficient transcriptomic deep learning model with multiple classifiers of tissue state is necessary.

[1]Krembil Research Institute, University Health Network, 399 Bathurst Street, Suite 4W-449, Toronto, ON M5T 2S8, Canada. [2]Division of Neurosurgery, Department of Surgery, University of Toronto, 149 College Street, 5th Floor, Toronto, ON M5T 1P5, Canada. [3]Division of Neurosurgery, Krembil Neuroscience Centre, Toronto Western Hospital, University Health Network, 399 Bathurst St., Suite 4W-449, Toronto, ON M5T 2S8, Canada. [4]These authors contributed equally: James Hong and Laureen D. Hachem. ✉email: michael.fehlings@uhn.ca

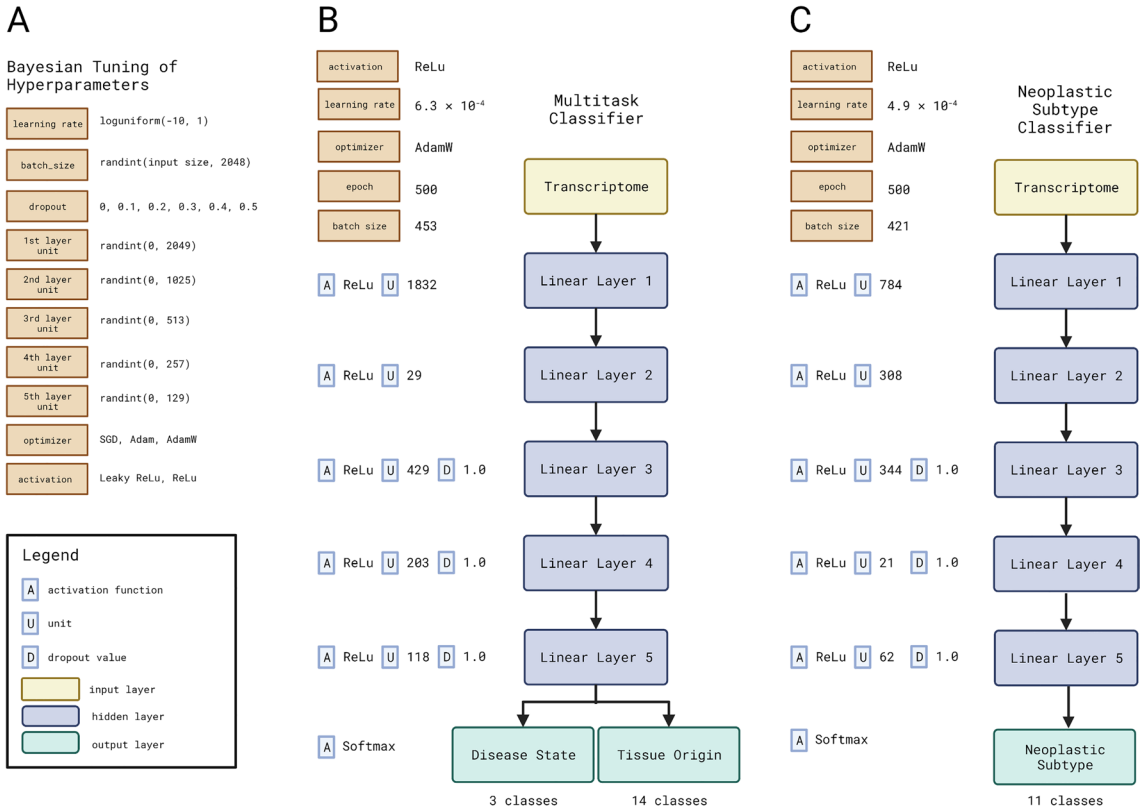

Figure 1. Bayesian Hyperparameter Tuning of Deep Learning Models. (**A**) Search space of hyperparameters for Bayesian tuning; (**B**) Architecture of multitask classifier for disease state and tissue origin along with tuned hyperparameters; (**C**) Architecture of neoplastic subtype classifier along with tuned hyperparameters.

Herein, we developed a multitask deep learning model using publicly available data from the Genotype Tissue Expression (GTEx) Project[16,17] and The Cancer Genome Atlas (TCGA) processed through a uniform analysis pipeline. Specifically, the model contains three classifiers including disease state, tissue origin and neoplastic subclass. Our model achieved high accuracy and performance metrics for all three classifiers and serves as a framework for incorporating large transcriptomic datasets across conditions to facilitate clinical diagnosis and research development.

Methods

RNA sequencing data. RNA sequencing data was obtained from the Genotype Tissue Expression (GTEx) Project[16,17] and The Cancer Genome Atlas (TCGA). As batch differences between different GTEx and TCGA submissions are well-documented, we utilized a common RNA-sequencing analysis pipeline to minimize batch effects[18]. Specifically, all raw reads were imported for alignment against hg19 in STAR, with quality control done in mRIN[19] (mRIN < − 0.11 threshold for sample exclusion), quantification in featureCounts[20] and batch effect correction in SVAseq[21]. In total, 10,116 patient samples were used with 17,993 genes included based on commonality across datasets (Supplementary Table 1). Dimensional reduction was performed using Sklearn package StandardScaler and principal component analysis (PCA), and 2000 principal components were used for model transformation. As a benchmark, 1000 top features selected by Random Forest and all 17,993 features (no PCA) were included in a separate run of the same models.

Deep learning model. Our deep-learning model consists of two models executed in tandem, the first is a multi-tasking model which classifies the type (non-neoplastic, neoplastic or peri-neoplastic) and tissue origin of the tissue. The subsequent subtyping model is primed to be executed only if the sample's tissue of origin has subtyping data available.

Based on prior work in deep learning processing of transcriptomic data and model tuning, the encoders for both models are comprised of 7 fully connected, feed-forward neural network layers (FFNN, Fig. 1B,C). The purpose of the 5 hidden layers is to bring down the dimensionality of the input transcriptomic data. Each of these layers has a Rectified Linear Unit (ReLU) activation function on top of their outputs, which is used to restrict the output of these layers. ReLU was selected over Sigmoid or Tanh due to the lack of vanishing gradient and sparsity, ultimately resulting in faster learning and quicker convergence[22]. Hidden layers 3 through 5 also have dropout layers between their output and the next layer to reduce overfitting. In the output layer, we have task heads, which are represented by layers with a Softmax activation function. These layers map their inputs to the dimension equal to the number of classes for that task. Specifically, for the multi-tasking model, the first

Accuracy	0.9882			
Balanced accuracy	0.9675			
Class	**Precision**	**Recall**	**F1**	**Support**
Non-Neoplastic	1.00	1.00	1.00	414
Neoplastic	0.99	0.99	0.99	1009
Peri-Neoplastic	0.94	0.91	0.93	105
Weighted Avg	0.99	0.99	0.99	1528

Table 1. Disease state classifier.

output head represents the type of tissue (non-neoplastic, neoplastic or normal peri-neoplastic, 3 classes) and the second output head represents the tissue origin (14 classes). Similarly, in the neoplastic subtype model, the output head presents the cancer subtype (11 classes). The Softmax activation function forces these output heads to output a probability distribution over their respective number of classes. All models were trained for 500 epochs.

Bayesian hyperparameter tuning. We performed Bayesian hyperparameter optimization using the hyperopt package[23], using the minimization of the cross-entropy loss as our optimization objective over 25 epochs. For each of the FFNNs, the Cartesian product of the learning rate, batch size, dropout value, unit, optimizer, and activation functions were selected as the search space (Fig. 1A). Instead of arbitrarily setting discrete values within the learning rate, batch size and units, we opted to randomize the range using the randint function. The optimal hyperparameters were then selected after 100 evaluations (Fig. 1A–C).

Benchmarking against other Machine Learning approaches. We compared the balanced accuracy of our proposed deep learning classifiers against other machine learning algorithms in the Sckit-learn package[24], including Decision Tree Classifier (DT), Extra Trees Classifier (ET), Support Vector Machine (SVM), Stochastic Gradient Descent (SGD) classifier, and K-nearest Neighbours Classifier (KNN). In these models, all 17,993 features were used as inputs, and a 70:15:15 ratio was used for train/validation/test splits.

Results

Training dataset. We used data from the Genotype Tissue Expression (GTEx) Project[16,17] and The Cancer Genome Atlas (TCGA) for the training of our model. Specifically, 10,116 patient samples were included and processed through a uniform pipeline[18]. Seventy percent of the data was used for training with the remaining split evenly for validation (15%) and testing sets (15%). The model was trained for 500 epochs. The deep learning model consisted of a multi-tasking model that classifies disease state (non-neoplastic versus peri-neoplastic vs neoplastic) and tissue origin (14 tissue classes). A subsequent subtyping model was primed to be executed only if the sample's tissue of origin had subtyping data available. The performance results reported here are from our proposed deep learning models with PCA dimensional reduction applied (DL PCA).

Multi-task model: disease state and tissue origin classifiers. The multitask portion of the model was trained on disease state (non-neoplastic versus peri-neoplastic vs neoplastic) and tissue origin (14 tissue classes). On the testing set, the disease state classifier achieved an overall accuracy of 0.99, precision of 0.99, recall of 0.99 and f1-score of 0.99 with high performance metrics for each subclass (Table 1). The associated confusion matrix (Fig. 2A) and receiver operating characteristic (ROC) curves (Fig. 2B) demonstrate that the model achieved an area under the curve (AUC) of 0.98 for disease state classification. The top K plot demonstrates that the classifier had excellent predictive accuracy without overfitting (Fig. 2C).

In terms of the tissue origin classifier, the model achieved an accuracy of 0.97, precision of 0.97, recall of 0.97 and f1-score of 0.97 (Table 2). ROC AUCs ranged from 0.97 to 1.00 for individual tissue origins with a macro average of 0.99 (Fig. 3B) with very few misclassifications (Fig. 3A). The top K plot demonstrates that the classifier had excellent predictive accuracy without overfitting (Fig. 3C).

Neoplasm subtype classifier. For tissues with multiple neoplastic subclasses (n = 11), the model was trained on an additional subtype classifier. Here, the model achieved an overall accuracy of 0.92, precision of 0.90, recall of 0.92 and F1-score of 0.91 (Table 3). ROC-AUC ranged from 0.55 to 1.00 for subtypes with a macro-average of 0.95 (Fig. 4B). It should be noted that the majority of subtypes were accurately classified, with the only exception of urterine corpus endothelial carcinoma (ucec) being classified as glioblastomas (gbm) (Fig. 4A,B). Top K plot demonstrates that the classifier had excellent predictive accuracy without overfitting (Fig. 4C).

Benchmark against various feature sets and other machine learning algorithms. We compared the balanced accuracy of our deep learning model (DL PCA) against deep learning models with either the full feature set (DL No PCA) or with the top 1,000 features selected by a Random Forest classifier (DL RF). Except for the neoplastic subtype classifier, DL PCA outperformed all other deep learning models. In both hyperparameter optimization and model training, DL PCA was 6 times more efficient compared to DL No PCA. The marginal gain in accuracy with the full feature set in the neoplastic subtype classifier justifies the use of PCA dimensional

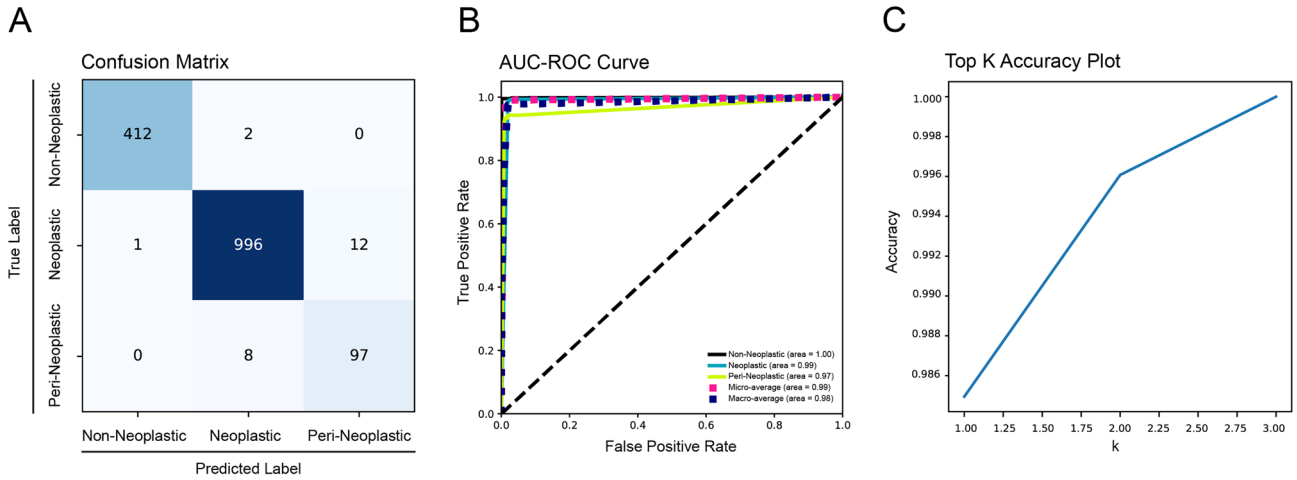

Figure 2. Performance of disease state classifier. (**A**) Confusion matrix of disease state classifier; (**B**) Receiver operating characteristic curve (ROC) with area under the curve (AUC); (**C**) top K accuracy plot.

Accuracy	0.9705			
Balanced accuracy	0.9587			
Class	**Precision**	**Recall**	**F1**	**Support**
Kidney	1.00	0.99	1.00	153
Colon	0.99	1.00	1.00	124
Esophageal	0.94	0.96	0.95	130
Lung	0.95	0.98	0.96	207
Uterus	0.97	0.93	0.95	41
Cervix	0.97	0.88	0.92	40
Liver	0.98	0.98	0.98	58
Thyroid	0.99	0.99	0.99	115
Stomach	0.94	0.96	0.95	98
Brain	1.00	0.99	0.99	141
Breast	1.00	0.99	0.99	202
Bladder	0.87	0.93	0.90	57
Kidney	1.00	1.00	1.00	82
Colon	0.91	0.85	0.88	80
Weighted Avg	0.97	0.97	0.97	1528

Table 2. Tissue origin classifier.

reduction to improve model efficiency. In all classifiers, the DL PCA model outperformed classic machine learning algorithms (DT, ET, SVM, SGD, and KNN; Fig. 5A–C, Supplementary Table 1).

Discussion

In this study, we developed a multitask deep learning model based on the most recent compendium of RNAseq data from GTEx and TCGA. Our model achieves high performance on all metrics across the three tissue classifiers of disease state, tissue identity, and neoplastic subtype. This is the first multitask deep learning algorithm developed for tissue classification employing a uniform pipeline analysis of transcriptomic data with multiple tissue classifiers. This model serves as a foundation for incorporating large transcriptomic datasets to facilitate disease diagnosis, subclassification and treatment decision-making.

While previous models based on transcriptomic data have been attempted for neoplastic classification, these have typically employed a single neural layer and have demonstrated modest performance metrics[25,26]. The architecture of our model incorporates five feed forward layers, which affords increased accuracy and a greater number of classifiers. Indeed, our model achieves overall better accuracy with fewer samples and training epochs than previous models[14]. The ability of our model to perform well even in classes with small sample sizes is of significant value for application to rare diseases or tissue states whereby access to patient samples may be limited. Rare conditions can often pose the greatest clinical diagnostic challenges as standard histological measures may fail to achieve an accurate diagnosis[27]. Furthermore, our model was able to distinguish with high accuracy samples of non-neoplastic tissue versus peri-neoplastic tissue, thus demonstrating its utility in classifying tissue state along a spectrum of disease. The latter is defined by samples collected > 2 cm from the neoplastic margin

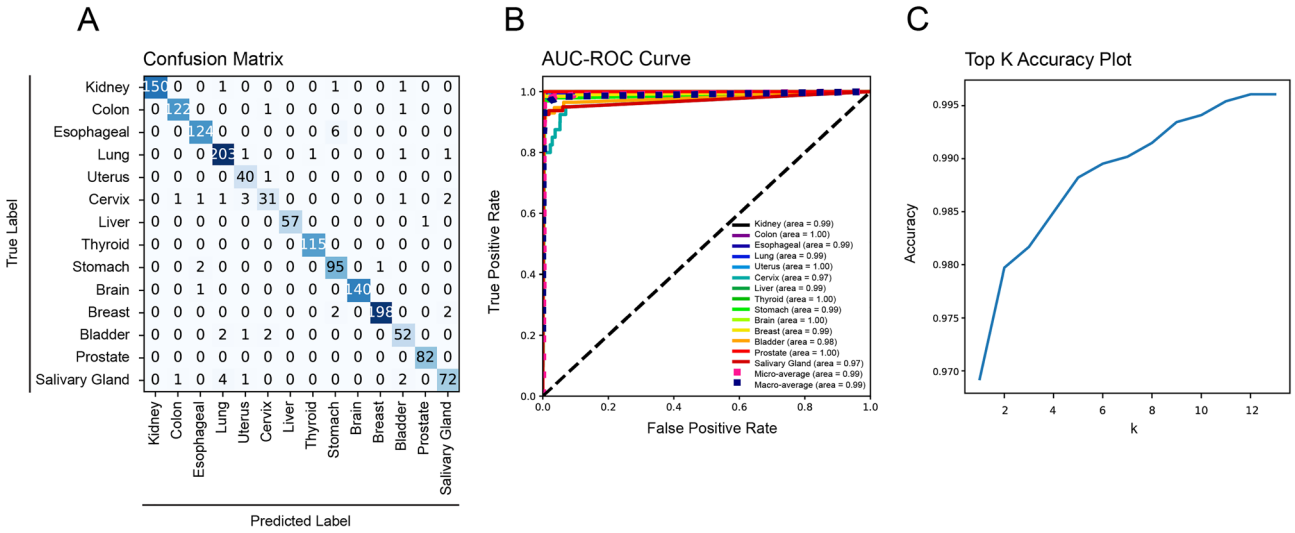

Figure 3. Performance of tissue origin classifier. (**A**) Confusion matrix of tissue origin classifier; (**B**) Receiver operating characteristic curve (ROC) with area under the curve (AUC); (**C**) top K accuracy plot.

Accuracy	0.9229			
Balanced accuracy	0.8548			
Class	**Precision**	**Recall**	**F1**	**Support**
Luad	0.95	1.00	0.97	76
Lgg	0.89	0.94	0.91	79
Coad	0.95	0.90	0.92	82
Ucec	0.00	0.00	0.00	11
Read	0.94	0.96	0.95	70
Gbm	0.80	1.00	0.89	47
Kich	1.00	0.96	0.98	23
Ucs	0.88	0.88	0.88	8
Kirc	0.95	0.88	0.91	24
Kirp	1.00	0.90	0.95	30
Lusc	1.00	1.00	1.00	4
Weighted Avg	0.90	0.92	0.91	454

Table 3. Neoplastic subtype classifier.

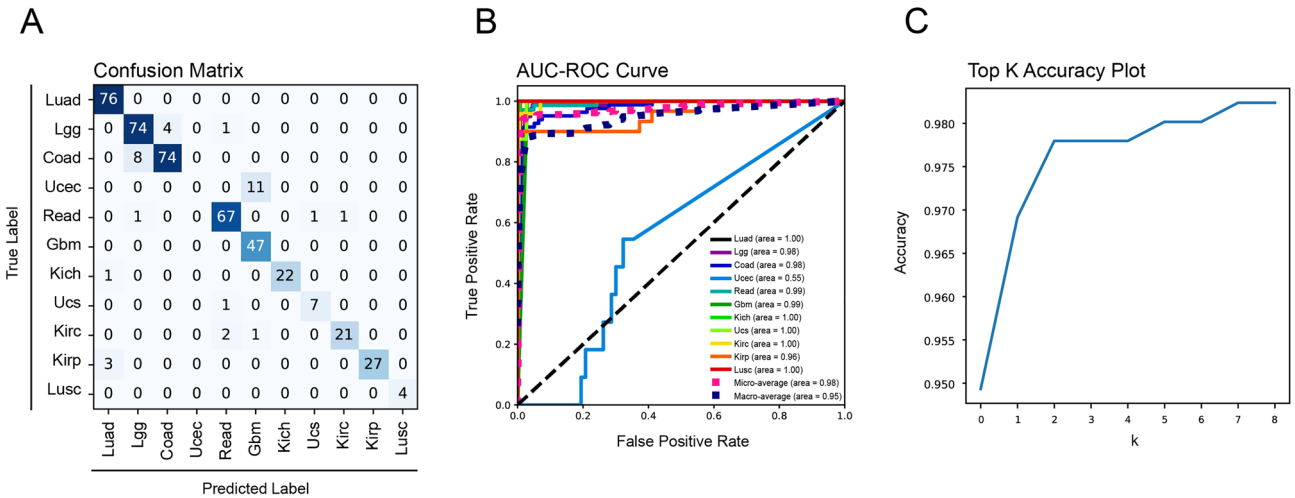

Figure 4. Performance of neoplastic subtype classifier. (**A**) Confusion matrix of tissue origin classifier; (**B**) Receiver operating characteristic curve (ROC) with area under the curve (AUC); (**C**) top K accuracy plot.

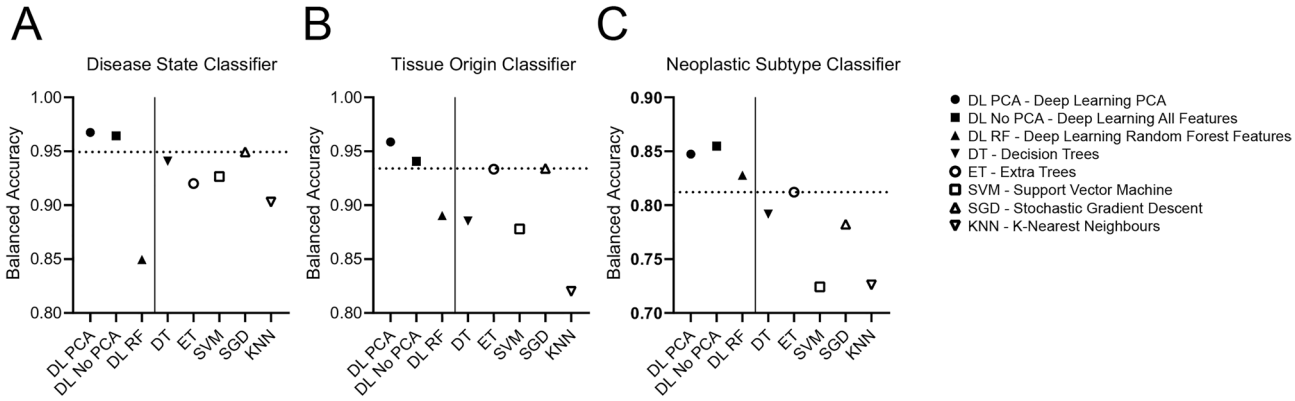

Figure 5. Benchmarking of deep learning classifiers and other machine learning algorithms. (**A**) Disease state classifier benchmarks; (**B**) Tissue origin classifier benchmarks; (**C**) Neoplastic subtype classifier benchmarks. Solid line separates the deep learning models from classic machine learning algorithms, and the dotted line indicates the highest balanced accuracy achieved by machine learning algorithms in each classifier.

with normal histological features and is often used as healthy controls in oncological studies[28,29]. However, normal tissue adjacent to the neoplasm has been shown to have differences in gene expression profiles compared to purely non-neoplastic tissue and as such may represent a distinct entity on the spectrum of neoplastic phenotypes[15]. To date, this distinction has not previously been incorporated into models of tissue typing.

Importantly, we used a uniform pipeline for processing RNA sequencing data from GTEx and TCGA prior to inputting into our model. Specifically, this approach employed mRIN-based exclusion of degraded samples, uniform realignment and expression quantification, along with study-specific bias correction, as previously described[18]. Previous studies have demonstrated that without uniform reprocessing and batch-correction of RNAseq data obtained from various studies, samples of different tissue identity within a single study show stronger similarities than samples of the same tissue type derived from different studies[18]. This underscores the importance of uniform processing of data prior to model training, which is a major pitfall of previous models employing a deep model architecture[14].

Ultimately, our model provides a framework to leverage large datasets of transcriptomic data across diseases and tissue states. Tissue profiling using transcriptomic data may be of particular use in situations of diagnosing cancers of unknown primary whereby standard clinicopathologic investigations do not yield a definitive source[30]. In our model, multiple classifiers including tissue origin and subtype may offer an advantage in the clinical diagnosis of these entities. While the model was trained on a set of non-neoplastic, normal peri-neoplastic and neoplastic tissue, in the future, additional transcriptomic data can be incorporated to include classifiers of tissue stress (e.g. inflammation or oxidative stress) and cellular phenotype (e.g. specific cell lineages), thus expanding the applications of this algorithm. Tissue profiling using this approach may be a valuable tool in determining the health and viability of cell lines used for clinical applications and in comparing cellular responses to injury and disease[31].

Data availability

The datasets and codes for the training and validation of the model are included in the link at: https://drive.google.com/drive/folders/1rSKDasV51ve9tbJjKtnWH9FImDFK9plb?usp=sharing. Details of the uniform pipeline analysis used for pre-processing of the RNAseq data are provided in the following link[18]: https://github.com/mskcc/RNAseqDB.

Received: 15 June 2021; Accepted: 11 April 2022
Published online: 11 June 2022

References

1. Cheung, C. C., Martin, B. R. & Asa, S. L. Defining diagnostic tissue in the era of personalized medicine. *CMAJ* **185**, 135–139. https://doi.org/10.1503/cmaj.120565 (2013).
2. Davidson, E. H. & Erwin, D. H. Gene regulatory networks and the evolution of animal body plans. *Science* **311**, 796–800. https://doi.org/10.1126/science.1113832 (2006).
3. Courtiol, P. *et al.* Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* **25**, 1519–1525. https://doi.org/10.1038/s41591-019-0583-3 (2019).
4. Xu, Q. *et al.* Pan-cancer transcriptome analysis reveals a gene expression signature for the identification of tumor tissue origin. *Mod. Pathol.* **29**, 546–556. https://doi.org/10.1038/modpathol.2016.60 (2016).
5. Burke, E. E. *et al.* Dissecting transcriptomic signatures of neuronal differentiation and maturation using iPSCs. *Nat. Commun.* **11**, 462. https://doi.org/10.1038/s41467-019-14266-z (2020).
6. Sun, C. *et al.* Transcriptome variations among human embryonic stem cell lines are associated with their differentiation propensity. *PLoS ONE* **13**, e0192625. https://doi.org/10.1371/journal.pone.0192625 (2018).
7. Cahan, P. *et al.* Cell Net: Network biology applied to stem cell engineering. *Cell* **158**, 903–915. https://doi.org/10.1016/j.cell.2014.07.020 (2014).
8. Zou, J. *et al.* A primer on deep learning in genomics. *Nat. Genet.* **51**, 12–18. https://doi.org/10.1038/s41588-018-0295-5 (2019).
9. Noorbakhsh, J. *et al.* Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. *Nat. Commun.* **11**, 6367. https://doi.org/10.1038/s41467-020-20030-5 (2020).

10. Yap, M. *et al.* Verifying explainability of a deep learning tissue classifier trained on RNA-seq data. *Sci. Rep.* **11**, 2641. https://doi.org/10.1038/s41598-021-81773-9 (2021).
11. Khorshed, T., Moustafa, M. N. & Rafea, A. Deep learning for multi-tissue cancer classification of gene expressions (GeneXNet). *IEEE Access* **8**, 90615–90629 (2020).
12. Yuan, B., Yang, D., Rothberg, B. E. G., Chang, H. & Xu, T. Unsupervised and supervised learning with neural network for human transcriptome analysis and cancer diagnosis. *Sci. Rep.* **10**, 19106. https://doi.org/10.1038/s41598-020-75715-0 (2020).
13. Jiao, W. *et al.* A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat. Commun.* **11**, 728. https://doi.org/10.1038/s41467-019-13825-8 (2020).
14. Azarkhalili, B., Saberi, A., Chitsaz, H. & Sharifi-Zarchi, A. DeePathology: Deep multi-task learning for inferring molecular pathology from cancer transcriptome. *Sci. Rep.* **9**, 16526. https://doi.org/10.1038/s41598-019-52937-5 (2019).
15. Aran, D. *et al.* Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nat. Commun.* **8**, 1077. https://doi.org/10.1038/s41467-017-01027-z (2017).
16. Consortium, G. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660. https://doi.org/10.1126/science.1262110 (2015).
17. Consortium, G. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585. https://doi.org/10.1038/ng.2653 (2013).
18. Wang, Q. *et al.* Unifying cancer and normal RNA sequencing data from different sources. *Sci. Data* **5**, 180061. https://doi.org/10.1038/sdata.2018.61 (2018).
19. Feng, H., Zhang, X. & Zhang, C. mRIN for direct assessment of genome-wide and gene-specific mRNA integrity from large-scale RNA-sequencing data. *Nat. Commun.* **6**, 7816. https://doi.org/10.1038/ncomms8816 (2015).
20. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930. https://doi.org/10.1093/bioinformatics/btt656 (2014).
21. Leek, J. T. svaseq: Removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.* **42**, e161. https://doi.org/10.1093/nar/gku864 (2014).
22. Nwankpa, C., Ijomah, W., Gachagan, A. & Marshall, S. *Activation Functions: Comparison of Trends in Practice and Research for Deep Learning*. arXiv:1811.03378 (2018).
23. Bergstra, J., Yamins, D. & Cox, D. In *Proceedings of the 30th International Conference on Machine Learning* Vol. 28 (eds Dasgupta Sanjoy & McAllester David) 115--123 (PMLR, Proceedings of Machine Learning Research, 2013).
24. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
25. Liu, X. *et al.* Predicting cancer tissue-of-origin by a machine learning method using DNA somatic mutation data. *Front. Genet.* **11**, 674. https://doi.org/10.3389/fgene.2020.00674 (2020).
26. Grewal, J. K. *et al.* Application of a neural network whole transcriptome-based pan-cancer method for diagnosis of primary and metastatic cancers. *JAMA Netw. Open* **2**, e192597. https://doi.org/10.1001/jamanetworkopen.2019.2597 (2019).
27. Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* https://doi.org/10.1126/scitranslmed.aal5209 (2017).
28. Casbas-Hernandez, P. *et al.* Tumor intrinsic subtype is reflected in cancer-adjacent tissue. *Cancer Epidemiol. Biomark. Prev.* **24**, 406–414. https://doi.org/10.1158/1055-9965.Epi-14-0934 (2015).
29. Koboldt, D. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70. https://doi.org/10.1038/nature11412 (2012).
30. Dermawan, J. K. & Rubin, B. P. The role of molecular profiling in the diagnosis and management of metastatic undifferentiated cancer of unknown primary(☆): Molecular profiling of metastatic cancer of unknown primary. *Semin. Diagn. Pathol.* https://doi.org/10.1053/j.semdp.2020.12.001 (2020).
31. Richards, L. *et al.* Gradient of Developmental and Injury Response transcriptional states defines functional vulnerabilities underpinning glioblastoma heterogeneity. *Nat. Cancer* **2**, 157–173 (2021).

Author contributions

J.H.: conception and design; model creation and development; acquisition, analysis, and interpretation of data; manuscript drafting; critical revision of manuscript; final approval of manuscript. L.D.H.: conception and design; model creation and development; acquisition, analysis, and interpretation of data; manuscript drafting; critical revision of manuscript; final approval of manuscript. M.G.F.: conception and design; critical revision of manuscript; final approval of manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-13665-5.

Correspondence and requests for materials should be addressed to M.G.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.