TULIP: An RNA-seq-based Primary Tumor Type Prediction Tool Using Convolutional Neural Networks

Cancer Informatics
Volume 21: 1–10

DOI: 10.1177/11769351221139491

Sara Jones[1], Matthew Beyers[1], Maulik Shukla[2], Fangfang Xia[2], Thomas Brettin[2], Rick Stevens[2], M Ryan Weil[1] and Satishkumar Ranganathan Ganakammal[1]

[1]Frederick National Laboratory for Cancer Research, Cancer Data Science Initiatives, Cancer Research Technology Program, Rockville, MD, USA. [2]Argonne National Laboratory, Computing, Environment and Life Sciences, Lemont, IL, USA.

ABSTRACT

BACKGROUND: With cancer as one of the leading causes of death worldwide, accurate primary tumor type prediction is critical in identifying genetic factors that can inhibit or slow tumor progression. There have been efforts to categorize primary tumor types with gene expression data using machine learning, and more recently with deep learning, in the last several years.

METHODS In this paper, we developed four 1-dimensional (1D) Convolutional Neural Network (CNN) models to classify RNA-seq count data as one of 17 highly represented primary tumor types or 32 primary tumor types regardless of imbalanced representation. Additionally, we adapted the models to take as input either all Ensembl genes (60,483) or protein coding genes only (19,758). Unlike previous work, we avoided selection bias by not filtering genes based on expression values. RNA-seq count data expressed as FPKM-UQ of 9,025 and 10,940 samples from The Cancer Genome Atlas (TCGA) were downloaded from the Genomic Data Commons (GDC) corresponding to 17 and 32 primary tumor types respectively for training and validating the models.

RESULTS: All 4 1D-CNN models had an overall accuracy of 94.7% to 97.6% on the test dataset. Further evaluation indicates that the models with protein coding genes only as features performed with better accuracy compared to the models with all Ensembl genes for both 17 and 32 primary tumor types. For all models, the accuracy by primary tumor type was above 80% for most primary tumor types.

CONCLUSIONS: We packaged all 4 models as a Python-based deep learning classification tool called TULIP (TUmor CLassIfication Predictor) for performing quality control on primary tumor samples and characterizing cancer samples of unknown tumor type. Further optimization of the models is needed to improve the accuracy of certain primary tumor types.

KEYWORDS: Convolutional neural network, tumor classification, deep learning, The Cancer Genome Atlas (TCGA), RNA-seq

RECEIVED: May 9, 2022. **ACCEPTED:** October 28, 2022.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This has been funded in whole or in part with Federal funding by the NCI-DOE Collaboration established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of the National Institutes of Health, Cancer Moonshot Task Order No. 75N91019F00134 and under Frederick National Laboratory for Cancer Research Contract 75N91019D00024. This work was performed under the auspices of the U.S. Department of Energy by Argonne National Laboratory under Contract DE-AC02-06-CH11357, Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344, Los Alamos National Laboratory under Contract 89233218CNA000001, and Oak Ridge National Laboratory under Contract DE-AC05-00OR22725.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Ranganathan Ganakammal Satishkumar, Cancer Data Science Initiatives, Cancer Research Technology Program, Frederick National Laboratory for Cancer Research, 9605 Medical Center Dr, Rockville, MD 20852, USA. Email: satishkumar.ranganathan@nih.gov

Background

Accounting for nearly 10 million deaths in 2020, cancer is a leading cause of death worldwide and the second leading cause of death in the United States.[1,2] Extensive research has been devoted to improving tools for cancer diagnosis and prognosis and developing targeted cancer therapies. The accumulation of publicly available cancer data has enabled the development of machine learning and deep learning models in several areas of clinical oncology including classification of tumor types and molecular subtyping of cancers.[3]

One such resource is the Genomic Data Commons (GDC), a unified repository and cancer knowledge base that includes several cancer genome programs such as The Cancer Genome Atlas (TCGA).[4,5] By harmonizing data from different programs and incoming submissions from researchers, the GDC provides a robust and growing dataset to enable precision oncology research. As more data enters GDC, it is important to address any data ambiguity that may arise with the clinical and/or sample metadata associated with genomics data. A machine learning or deep learning model that can predict with high accuracy the primary tumor type from RNA-seq data can help identify any misclassified primary tumor types, provide the precise primary tumor type of more generalized or missing primary tumor types, and differentiate any samples that do not express similar expression profiles to the assigned primary tumor type for further analyses.

Several machine learning and deep learning models have been developed for tumor classification of RNA-seq data. For example, Ahn et al.[5] built a fully connected deep neural network (DNN) to differentiate tumor versus normal samples. Similarly, Park et al.[6] constructed PathDeep, a biological function structure based DNN, to discriminate between cancer and normal

tissues. Others such as Li et al.[7] and Lyu and Haque[8] developed models that can perform individual tumor type classification. Li et al. combined a k-nearest neighbors (KNN) algorithm with a genetic algorithm to attain at least 90% accuracy for predicting 31 TCGA cancer types while Lyu et al. utilized a convolutional neural network (CNN) model to achieve 95.59% accuracy for 33 TCGA types. Mostavi et al.[9] also used CNN-based models to produce an accuracy of 93.9% to 95.0% among 34 classes (33 TCGA cancer types and a normal group. Ramirez et al.[10] established 4 models with a graph convolutional neural network (GCNN) to obtain above 94% accuracy for classifying samples as one of 33 TCGA cancer types or as normal, similar to Mostavi et al.

To create a tool that can perform quality control (QC) on RNA-seq samples, we selected the CNN model inspired from the Tumor Classifier 1 (TC1) resource[11] developed under Pilot 1 of the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) program established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of the National Institutes of Health.[12] We first downloaded TCGA RNA-seq data of 32 primary tumors from GDC. Based on the sample distribution of the primary tumors, we developed 2 types of 1D-CNN models that can either classify 17 primary tumor types that had at least 300 samples or all 32 primary tumor types regardless of sample size. In addition, we also experimented with the number of genes or features of the models to determine the accuracy of the models when all 60K genes or all 19K protein coding genes are used. In total, we had 4 different 1D-CNN models that had an overall accuracy of 94.7% to 97.6% on the test dataset. Given the performance values, we created a Python-based deep learning classification tool called TULIP (TUmor CLassIfication Predictor) incorporating our 1D-CNN models to serve as a QC tool for the cancer research community. Lastly, we tested the use of our tool on kidney cancer RNA-seq data from the Clinical Proteomic Tumor Analysis Consortium (CPTAC), also available on GDC. In addition to being a QC tool, TULIP can potentially be used for predicting primary tumor types of samples with unspecified or unknown primary tumor diagnosis.

Methods

Gene expression data collection and preprocessing

We downloaded RNA-seq data expressed as FPKM-UQ, where FPKM-UQ is the upper quartile of the number of fragments per kilobase per million mapped reads for 9,025 and 10,940 samples corresponding to 17 and 32 primary tumor types respectively from GDC (February 2022). Supplemental Table S1 lists the number of samples per primary tumor type. We utilized the gdc-RNA-seq-tool[13] to download and merge individual RNA-seq data files. We then developed an in-house Python script (version 3.7.12) to convert the FPKM-UQ expression values to TPM (transcripts per million) and normalized the TPM values by applying log10 transformation. The scikit-learn package (version 1.0.2)[14] was used to split the data randomly into training (80%), validation (10%) and test (10%) datasets. We encoded the primary tumor types using the OneHotEncoder() function (Supplemental Figure S1).

Since the RNA-seq files from GDC contain all 60,483 genes, we created 2 additional datasets containing only 19,758 protein coding genes for both 17 and 32 primary tumor types. The links to the lists of genes, which are organized alphanumerically based on their Ensembl IDs, for both all the genes and protein coding genes only are provided in Supplemental File 1 along with queries used to obtain the data from GDC.

To test the performance of our models with unknown data, we obtained 277 RNA-seq samples from CPTAC (February 2022) that are associated with kidney cancer. The samples and metadata are listed in Supplemental File 2.

Dimensionality reduction (t-SNE) analysis

To visualize how the samples from different primary tumor types may cluster, we employed t-distribution stochastic neighbor embedding (t-SNE), a non-linear dimensionality reduction technique used for visualizing high dimensional datasets in a low dimensional space.[15] The t-SNE analysis, using default scikit-learn package parameters, was performed on the top 1,000 highly variable genes from the log10 transformed TPM datasets of both the 17 and 32 primary tumor types with all the genes.

CNN model construction and implementation

We created 4 CNN models with the same underlying architecture as that of the TC1 resource[11] mentioned above (Figure 1). The main differences between each model are the number of genes for the input layer and the number of primary tumor types in the output layer. For the sake of simplicity, the names of the models will be referred to as the following:

- 19,758 protein coding genes as input and 17 primary tumor types: *CNN-17-PC*
- 60,483 genes as input and 17 primary tumor types: *CNN-17*
- 19,758 protein coding genes as input and 32 primary tumor types: *CNN-32-PC*
- 60,483 genes as input and 32 primary tumor types: *CNN-32*

The CNN models were implemented using Keras (version 2.4.3).[16] All the models included two 1D convolutional layers, 2 maximum pooling layers of size 10, two fully connected (FC) layers with 200 and 20 nodes respectively, and the output layer. Each convolutional layer contained 128 filters of kernel size 20 and stride of 1. The rectified linear unit (ReLU) activation function was used for all hidden layers, while softmax was used for the output layer. We used categorical cross-entropy as the loss function. To address overfitting of the models, 10% dropout was applied to all the FC layers. The model was trained with a starting learning rate of 0.1, a stochastic gradient descent

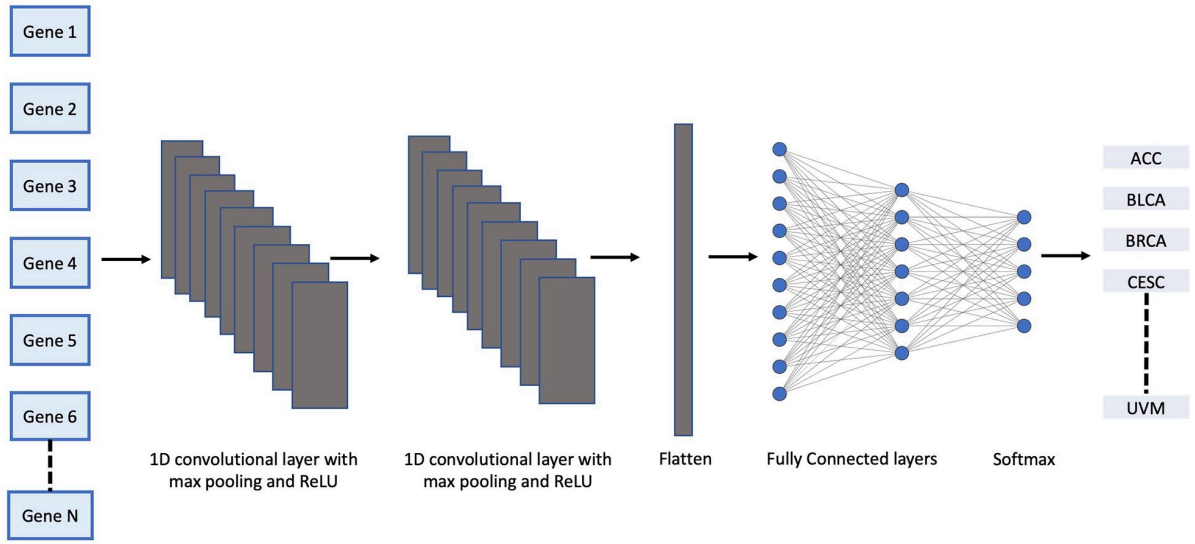

Figure 1. Architecture of 1D convolutional neural network for primary tumor type classification of RNA-seq data.

(SGD) optimizer, and a batch size of 20. We also used "ReduceLROnPlateau" to reduce the learning rate when the cross-entropy loss stops improving after 10 epochs as the model trained for a maximum of 400 epochs. NVIDIA V100 GPUs were used for training the CNN models.

Model performance evaluation metrics

To evaluate the 1D-CNN models during training, we tracked model accuracy and loss value at every epoch to optimize and identify the best performing model. To compare the performance of all 4 models, we used accuracy using Keras' evaluate() function on the training, validation, and test datasets. To find the largest predicted probability, we implemented the argmax() NumPy function on each sample in the test dataset to identify the predicted primary tumor type. We assessed the performance of the models on the test dataset with the weighted average of precision, recall and F1 score to account for class imbalance using the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). The formulas for calculating precision, recall, and F1 score are below.

$$\text{Precision: } TP/(TP + FP)$$

$$\text{Recall: } TP/(TP + FN)$$

$$\text{F1 score: } 2\times(\text{Recall} \times \text{Precision})/(\text{Recall} + \text{Precision})$$

Results

Distribution of primary tumor types in GDC

The histogram (Figure 2) shows the sample distribution of 32 primary tumor types. The number of samples ranges from 45 (cholangiocarcinoma (CHOL)) to 1,220 (breast invasive carcinoma (BRCA)). Due to the class imbalance of the primary tumor types, we created 2 types of models. In one model, we considered all the primary tumor types regardless of the number of samples. For the second model, we selected primary tumor types with high representation in the dataset, with a cut-off of greater than 300 samples, resulting in 17 primary tumor types being represented.

Visualization of RNA-seq data using t-SNE

Next, we used t-SNE to visualize the RNA-seq data of 17 and 32 primary tumor types (Figures 3 and 4, respectively). In both figures, distinct clusters corresponding to many of the primary tumor types can be observed. This indicates that unique gene expression profiles can be used to differentiate primary tumor types; however, there is some overlap of samples associated with certain primary tumor types based on tissue type or cell type. For example, some of the lung squamous cell carcinoma (LUSC) samples can be found within the lung adenocarcinoma (LUAD) cluster (Figures 3 and 4). In Figure 4, there is complete overlap of rectum adenocarcinoma (READ) and colon adenocarcinoma (COAD) samples. Samples for bladder urothelial carcinoma (BLCA), head and neck squamous cell carcinoma (HNSC), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), and LUSC tend to group together based on cell type in both figures. It seems that samples originating from similar tissue types such as the lung for LUSC and LUAD or similar cell types such as carcinoma for BLCA, CESC, HNSC, and LUSC might play a significant factor in the similarity of these samples' gene expression profiles. Even though there are several primary tumor types that may be hard to differentiate with any classifier, the high number of clusters provide a strong level of confidence that most primary tumor types can be classified correctly. Additionally, primary tumor types with small sample sizes, such as adrenocortical cancer (ACC), form their own clusters, indicating that sample size may not be a limiting factor for some primary tumor types to be classified.

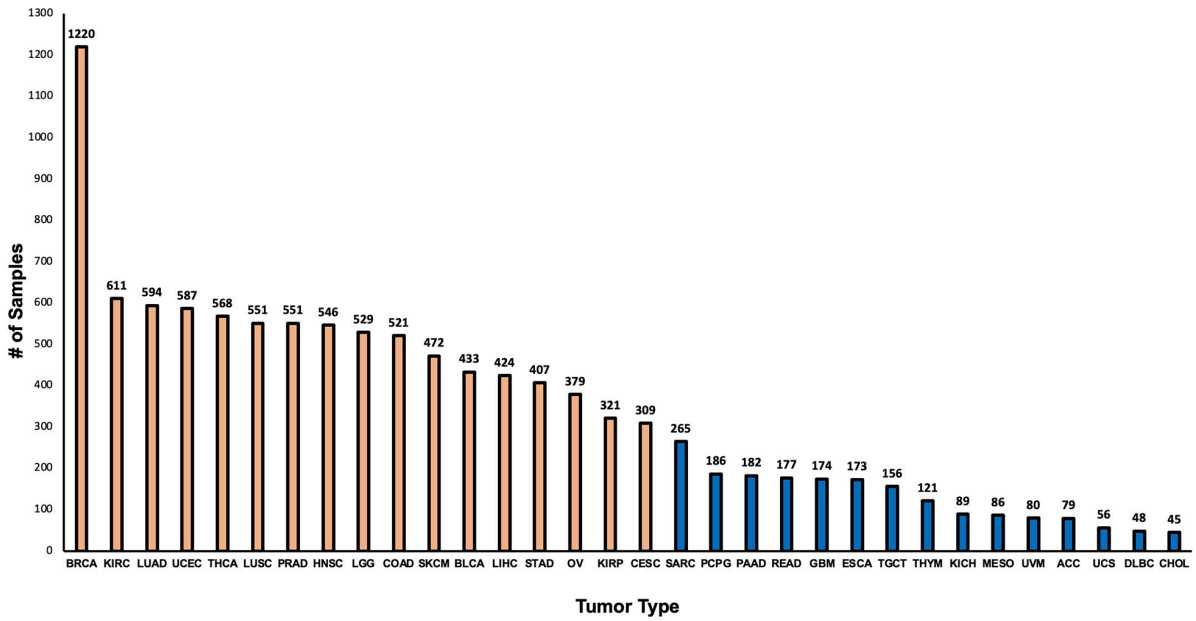

Figure 2. Histogram representation of sample distribution of the data obtained from GDC. The orange bars represent the primary tumor types with number of samples >300, while the blue are <300 samples each.

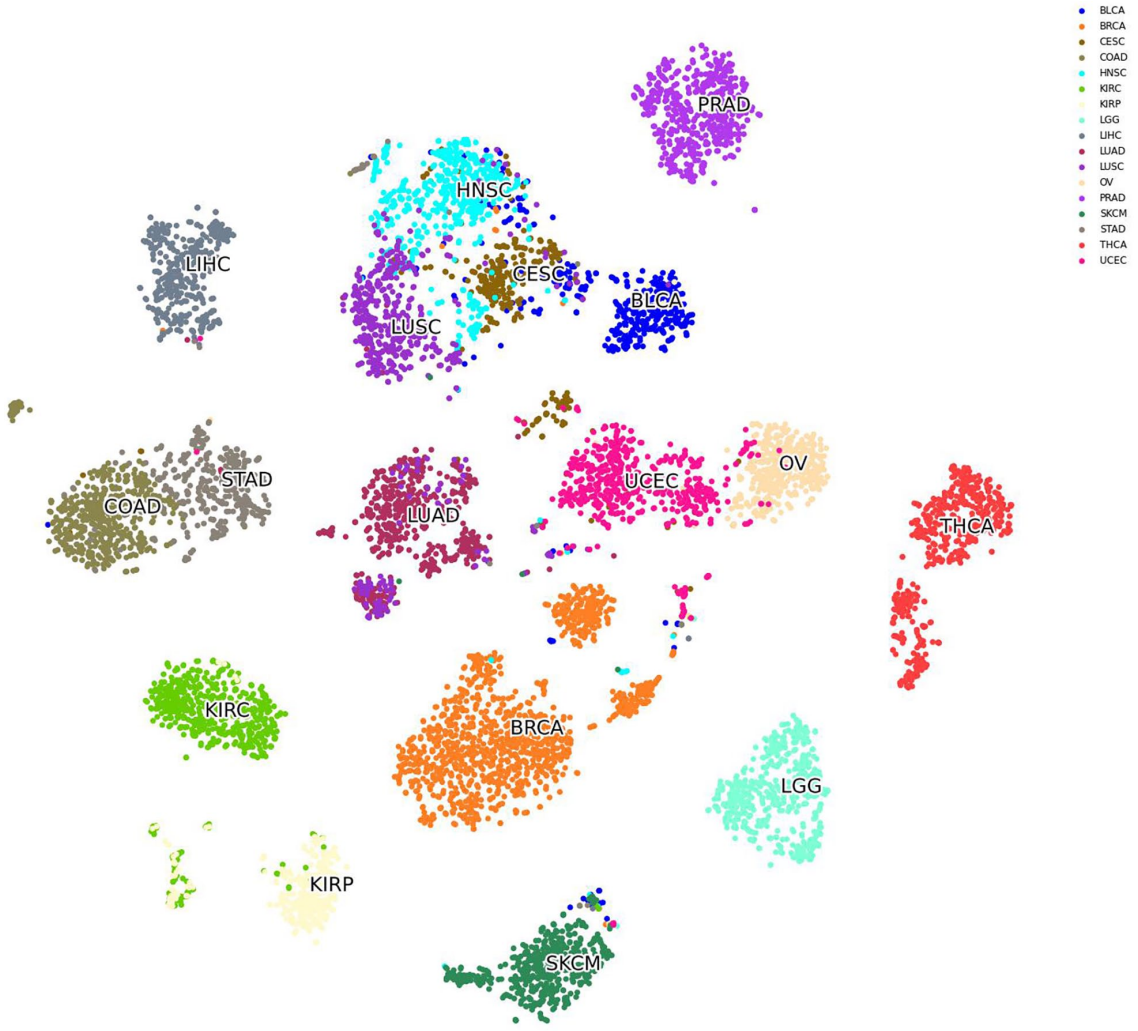

Figure 3. t-SNE of 17 primary tumor types using the top 1,000 highly variable genes.

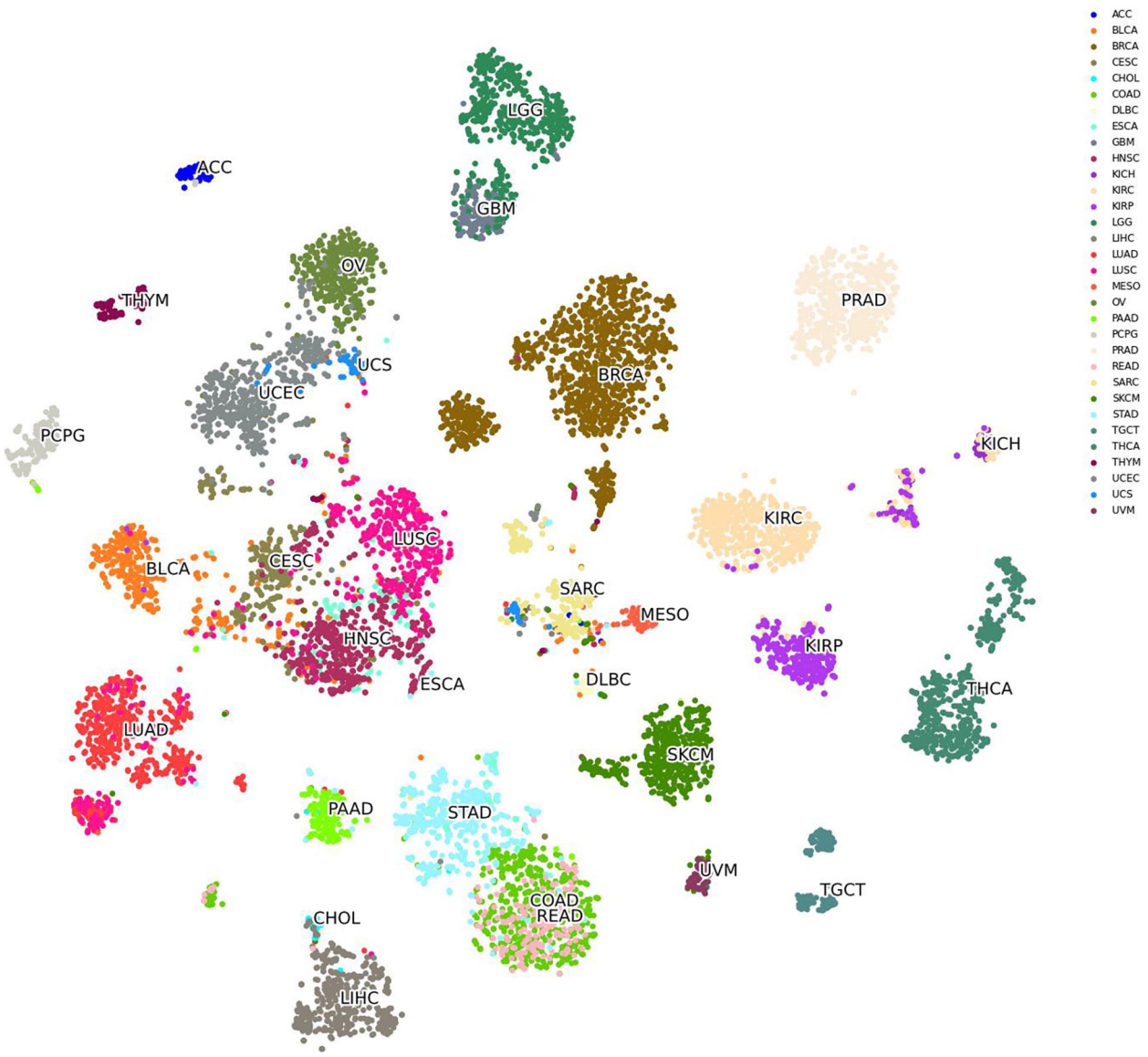

Figure 4. t-SNE of 32 primary tumor types using the top 1,000 highly variable genes.

Performance evaluation of 1D-CNN models

The overall training accuracy of all 4 models ranged from 98.7% to 100%, and the validation accuracy ranged from 95.2% to 97.6% (Table 1). The validation accuracies for the 32 primary tumor type models were slightly lower than the 17 primary tumor type models.

We then examined the performance of the 1D-CNN models on 2 test datasets. The first dataset contained 903 samples corresponding to 17 primary tumor types, and the second dataset contained 1,094 samples corresponding to 32 primary tumor types. The overall test accuracies of the models ranged from 94.7% to 97.6% (Table 2). The weighted averages of precision, recall, and F1 score values were all above 90% for all 4 models. Like the validation accuracies, the test accuracies for the 32 primary tumor type models performed slightly lower than the 17 primary tumor type models. The difference in performance is most likely attributed to the additional primary tumor types with low sample sizes, as well as the greater number of primary tumor types with highly similar expression profiles with other primary tumor types described above. When comparing between models for 17 and 32 primary tumor types,

Table 1. Training and validation accuracy of all 4 models.

MODEL	TRAINING ACCURACY	VALIDATION ACCURACY
17 Primary tumor types		
CNN-17-PC	1.000	0.973
CNN-17	1.000	0.976
32 Primary tumor types		
CNN-32-PC	1.000	0.962
CNN-32	0.987	0.952

Table 2. Accuracy, precision, recall, and F1 score for test dataset.

MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE
17 Primary tumor types				
CNN-17-PC	0.976	0.970	0.967	0.966
CNN-17	0.972	0.954	0.951	0.951
32 Primary tumor types				
CNN-32-PC	0.952	0.921	0.931	0.924
CNN-32	0.947	0.924	0.934	0.927

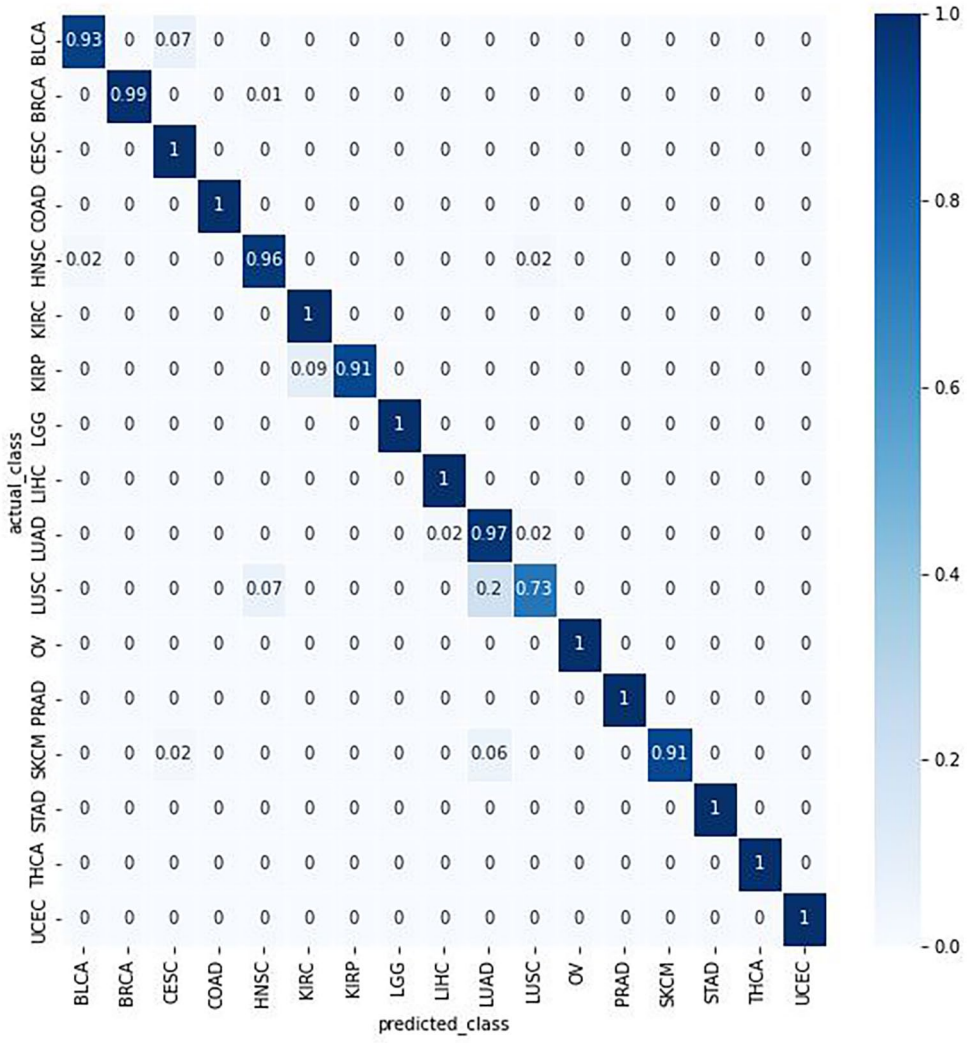

Figure 5. Confusion matrix of accuracy for 17 primary tumor types and protein coding genes only.

the protein coding gene-based models have marginally better accuracy than the all gene-based models.

Accuracy by primary tumor type

Supplemental Table S2 shows the accuracy for each model by primary tumor type while Supplemental Tables S3 and S4 show the precision, recall, and F1 scores. To understand which primary tumor types have a tendency to be misclassified, we generated confusion matrices for each model as well. From Supplemental Table S2 as well as Figure 5 and Supplemental Figure S2, the test accuracy was above 90% for 16 of the 17 primary tumor types for both the *CNN-17* and *CNN-17-PC* models. Only LUSC had an accuracy below 90% with 67% for *CNN-17* and 73% for *CNN-17-PC*. In Figure 5 and Supplemental Figure S2, the most common misclassified

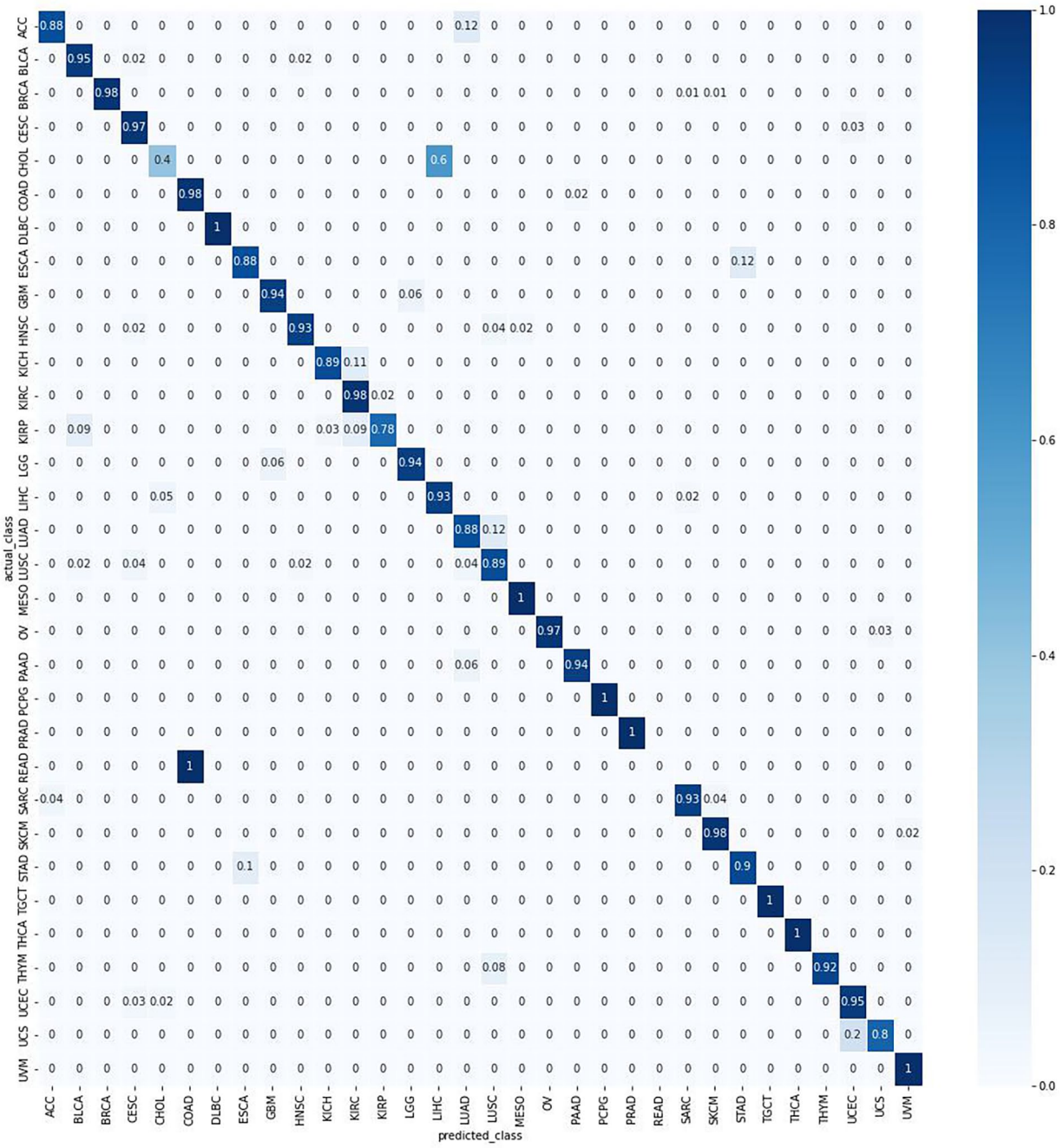

Figure 6. Confusion matrix of accuracy for 32 primary tumor types and protein coding genes only.

primary tumor type for LUSC was LUAD. When looking at the *CNN-32* and *CNN-32-PC* models, there were some slight differences. For the *CNN-32* and *CNN-32-PC* models, the test accuracy was above 80% for 28 and 29 primary tumor types respectively and 90% for 25 primary tumor types (Supplemental Table S2, Supplemental Figure S3, and Figure 6). The primary tumor types with an accuracy of below 80% shared between both models include CHOL at 40% and READ at 0%. The primary tumor type that CHOL was misclassified as was liver hepatocellular carcinoma (LIHC) in the *CNN-32-PC* model while both LIHC and pancreatic adenocarcinoma (PAAD) were the main primary tumor types in the *CNN-32* model. As expected from the t-SNE plot above, all of the test samples for READ in both models were misclassified as COAD. In the *CNN-32* model, the test accuracy for kidney chromophobe (KICH) was much lower than the *CNN-32-PC* model (56% vs 89% respectively). Interestingly, the accuracy for LUSC was slightly better in the 32 primary tumor type models with an accuracy of 89% (*CNN-32-PC*) and 93% (*CNN-32*), perhaps due to random sample selection.

TULIP

For public utility of any of the 4 models, we created a Python-based tool called TULIP (TUmor CLassIfication Predictor). This tool takes as input an RNA-seq count matrix, ideally from

Table 3. Count of predicted primary tumor types of CPTAC kidney cancer RNA-seq data for each model with kidney primary tumor types outlined in bold.

	CNN-17-PC	CNN-17	CNN-32-PC	CNN-32
Kidney renal clear cell carcinoma	274	273	274	220
Kidney renal papillary cell carcinoma	3	1	2	51
Breast invasive carcinoma			1	
Cervical squamous cell carcinoma and endocervical adenocarcinoma				1
Colon adenocarcinoma				1
Lung adenocarcinoma		3		2
Lung squamous cell carcinoma				1
Pancreatic adenocarcinoma				1

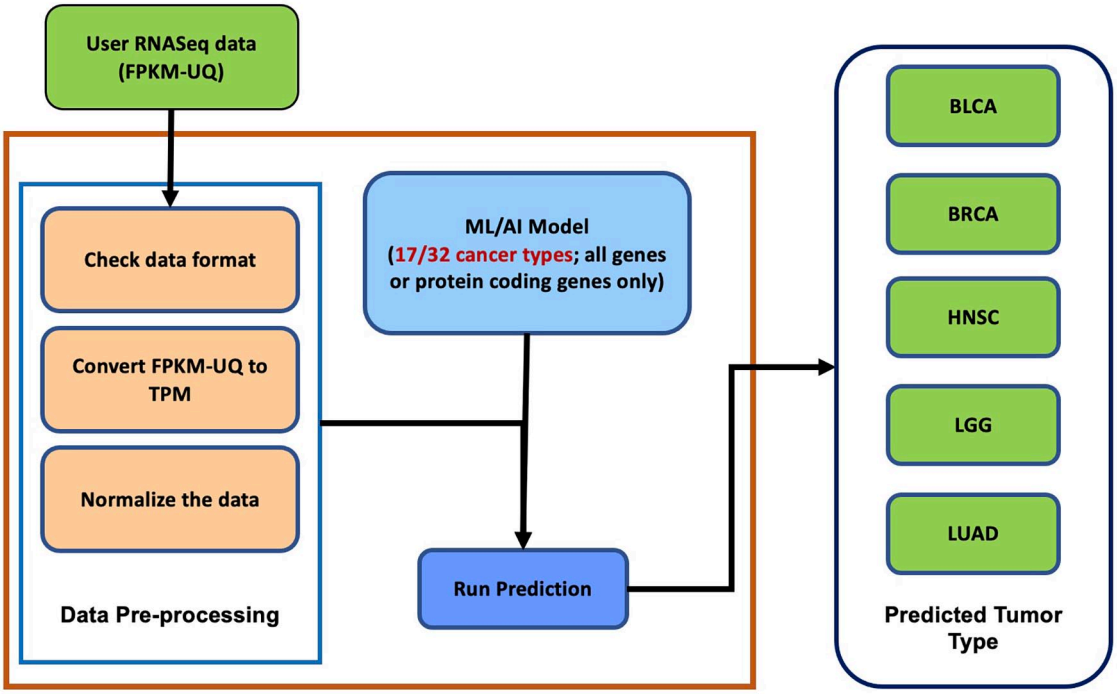

Figure 7. TULIP workflow for preprocessing RNA-seq data and applying one of the provided 1D-CNN models to predict primary tumor type.

GDC, expressed as FPKM-UQ and outputs a file containing the predicted primary tumor types with their probability scores (Figure 7). We provide options for the user to select which model to use and to set the minimum probability score threshold for a sample to be classified as a primary tumor type. Based on our results of higher accuracy with the protein coding models, we have set these as the default parameters. The code for TULIP can be found here: https://github.com/CBIIT/TULIP.

CPTAC kidney cancer prediction

To observe how our tool would perform on non-TCGA data, we downloaded CPTAC RNA-seq data of kidney cancer. We then applied all 4 models of TULIP and achieved the best results with the *CNN-17-PC* model. This model correctly identified kidney cancer for all 277 samples with 274 samples classified as kidney renal clear cell carcinoma (KIRC) and 3 samples classified as kidney renal papillary cell carcinoma (KIRP) (Table 3). Both the *CNN-32-PC* and *CNN-17* models performed similarly with nearly 100% accuracy. Interestingly, the *CNN-32* model, which had 97.8% accuracy, had a different breakdown for KIRC and KIRP. This model classified 220 samples as KIRC and 51 samples as KIRP. Of the samples that were not classified as either KIRC or KIRP by any of the models, the predicted primary tumor types all have carcinoma in common along with KIRC and KIRP. This suggests that cell type may present a challenge for any model to distinguish primary tumor types of the same cell type. The predicted primary tumor types and the probability scores are provided in Supplemental File 2. We highlighted in orange the non-kidney

primary tumor types. Overall, TULIP was able to classify the kidney cancer samples with high accuracy as well as provide the specific type of kidney cancer.

Discussion

The TC1 framework[11] was used to develop the 1D-CNN models with data from TCGA. Our models performed just as well, or better, to predict the primary tumor type from RNA-seq count data, when compared to similar methods. As other published studies have observed using unsupervised clustering methods such as hierarchical clustering or dimension reduction techniques like t-SNE, certain primary tumor types may be difficult to differentiate regardless of the number of samples due to similar tissue or cell types.[17] For example, the expression profiles of READ and COAD completely overlapped in our t-SNE visualization (Figure 4) even though these samples come from different anatomical locations. As expected, our models performed poorly in predicting the correct primary tumor type for READ samples. Instead, these samples were classified as COAD. Due to their homogeneity, COAD and READ have long been considered as a single cancer type, colorectal cancer. Efforts are ongoing to identify other molecular characteristics using omics data such as proteomic data to distinguish them.[18] With only RNA-seq data, we may have to adopt a similar approach to combine these primary tumor types as one.

By not pre-selecting genes to go into the model, contrary to previous studies, we ensure that we include genes that may be important for differentiating primary tumor types as more data is integrated into GDC. This is especially pertinent for primary tumor types with less representation in the database. Even though the high number of features may have added noise to the models, we still obtained an overall test accuracy of 94.7% to 97.6% for all 4 models. In addition, the models had at least 80% test accuracy for most of the primary tumor types in both the 17 and 32 primary tumor type models on our test dataset. However, it is important to note that the ability of the models to predict the primary tumor type is limited by the low number of samples for several primary tumor types in the test dataset. For some of the primary tumor types, we only had 5 samples. As more data becomes available in the future, updating these models with additional data will lend more confidence to the models' prediction accuracy.

To make the models more accessible to the cancer research community, we developed TULIP to take RNA-seq data as input and to generate the predicted primary tumor types with their probability scores. TULIP can be used as a QC tool for identifying any samples that may not have a gene expression profile that aligns with the primary tumor type attached to the sample. Any sample with an incorrect primary tumor type or unknown primary tumor type based on the probability score threshold set by the user can then be further explored to understand how this sample may be different from its assigned primary tumor type. For example, race and sex may lead to differences in RNA expression within individual primary tumor types. Additionally, TULIP can also provide more specific information of the primary tumor type for any sample, such as the CPTAC kidney data, with broad or unknown primary tumor types. Even though TULIP was able to predict the kidney cancer as the primary tumor type with 100% accuracy using the *CNN-17-PC* model, we do not have information that the specific kidney cancer types predicted, KIRC and KIRP, are correct. Having this information would have provided more support in the prediction accuracy of our models.

At present, the scope of TULIP is to provide quality control of tumor tissue type of samples obtained from patients. We plan to update the models on a regular basis to improve the accuracy of the models as more samples become available. We are also interested in incorporating normal versus tumor prediction to TULIP. Previously, our collaborators have developed a normal versus tumor classifier using a 1D-CNN framework.[19] By adding this classifier as a preliminary step before using TULIP or including normal tissue as another class to predict within the 4 models may help to better distinguish genes unrelated to tissue type in classifying primary tumor types. As more data becomes available, we plan to enhance our classification framework to address categorization of various tumor subtypes. We are also interested in identifying gene signatures that are responsible for tumor classification to provide additional information and insights for users of TULIP. Lastly, we hope that common data sharing platforms and other data processing pipelines would adopt TULIP to assist with validating tumor tissue types as part of their genomic data submission workflows.

Conclusions

We have developed 4 1D-CNN models that can perform primary tumor type prediction of high dimensional RNA-seq count data. All 4 models had at least 94.7% prediction accuracy with the best performing model reaching 97.6%. Unlike previous studies that filtered the genes based on expression levels, our models still achieved high prediction accuracy when we kept all the genes for our all-gene-based and protein-coding-based models. To make these models available for the cancer research community, we created TULIP. Our tool can be utilized for performing quality control on primary tumor samples as well as classifying cancer samples of unknown origin. The tool and the source code are publicly available at https://github.com/CBIIT/TULIP

Acknowledgements

This work utilized the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov/). Data repository support for storing models and data was provided by the Predictive Oncology Model and Data Clearinghouse (MoDaC: https://modac.cancer.gov/). Ravichandran Sarangan was instrumental in developing the framework for data processing.

Authors' Contributions

Conception and design: SRG, SEJ, MLB. Collection and assembly of data: SEJ, SRG. Development of original CNN models: FX, MS, TSB, RLS. Training of new CNN models: SEJ, SRG. Data analysis and interpretation: SEJ, SRG. Manuscript writing: SEJ, SRG. Manuscript review and editing: All authors. Final approval of manuscript: All authors.

Availability of Data and Materials

The TCGA and CPTAC RNA-seq datasets are publicly available at https://gdc.cancer.gov/.

The archived manifest file with the FPKM-UQ values can be found at https://docs.gdc.cancer.gov/Data/Release_Notes/gdc_manifest_20211029_data_release_31.0_active.tsv.gz.

The list of samples used in this analysis is available upon request.

The code can be found at https://github.com/CBIIT/TULIP.

Consent for Publication

Not applicable.

Ethics Approval and Consent to Participate

The authors declare no competing financial interests.

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71:209-249.
2. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. *CA Cancer J Clin*. 2021;71:7-33.
3. Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med*. 2021;13:152.
4. Heath AP, Ferretti V, Agrawal S, et al. The NCI genomic data commons. *Nat Genet*. 2021;53:257-262.
5. Ahn T, Goo T, Lee CH, et al. Deep learning-based identification of cancer or normal tissue using gene expression data. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE; 2018:1748-1752. doi:10.1109/BIBM.2018.8621108
6. Park S, Huang E, Ahn T. Classification and functional analysis between cancer and normal tissues using explainable pathway deep learning through RNA-sequencing gene expression. *Int J Mol Sci*. 2021;22:11531.
7. Li Y, Kang K, Krahn JM, et al. A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data. *BMC Genomics*. 2017;18:508.
8. Lyu B, Haque A. 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*. ACM; 2018:89-96. doi:10.1145/3233547.3233588
9. Mostavi M, Chiu YC, Huang Y, Chen Y. Convolutional neural network models for cancer type prediction based on gene expression. *BMC Med Genomics*. 2020;13:44.
10. Ramirez R, Chiu YC, Hererra A, et al. Classification of cancer types using graph convolutional neural networks. *Front Phys*. 2020;8:203.
11. Brettin T, et al. (2021, November 21). *CBIIT/NCI-doe-collab-pilot1-tumor-classifier*. GitHub. Retrieved March 18, 2022, from https://github.com/CBIIT/NCI-DOE-Collab-Pilot1-Tumor-Classifier
12. Hollingsworth P, et al. (2022, May 11). Joint design of advanced computing solutions for cancer (JDACS4C). GitHub. Retrieved March 18, 2022, from https://datascience.cancer.gov/collaborations/joint-design-advanced-computing
13. Reid C. Gdc-rnaseq-tool. GitHub. 2018. Retrieved March 18, 2022, from https://github.com/cpreid2/gdc-rnaseq-tool
14. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825-2830.
15. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9:2579-2605.
16. Chollet F. Keras. GitHub. 2015. Retrieved March 18, 2022, from https://github.com/fchollet/keras
17. Hoadley KA, Yau C, Hinoue T, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*. 2018;173:291-304.e6.
18. Liang JZ, Liang XL, Zhong LY, Wu CT, Zhang J, Wang Y. Comparative proteome identifies complement component 3-mediated immune response as key difference of colon adenocarcinoma and rectal adenocarcinoma. *Front Oncol*. 2021;10:617890.
19. Brettin T, et al. (2021, November 21). Normal-tumor pair classifier model (NT3). GitHub. Retrieved March 18, 2022, from https://github.com/CBIIT/NCI-DOE-Collab-Pilot1-Normal_Tumor_Pair_Classifier