2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

An Ensemble Model for Tumor Type Identification and Cancer Origins Classification

Chenzhao Feng+, Tianyu Xiang+, Zixuan Yi+, Lingzhe Zhao, Sisi He and Kunming Tian*

***Abstract*— Tissue biopsy can be wildly used in cancer diagnosis. However, manually classifying the cancerous status of biopsies and tissue origin of tumors for cancerous ones requires skilled specialists and sophisticated equipment. As a result, a data-based model is urgently needed. In this paper, we propose a data-based ensemble model for tumor type identification and cancer origins classification. Our model is an ensemble model that combines different models based on mRNA groups which serve distinct functions. The experiment on the TCGA dataset exhibits a promising result on both tasks – 98% on tumor type identification and 96.1% on cancer origin classification. We also test our model on external validation datasets, which prove the robustness of our model.**

***Index Terms*— cancers, ensemble learning, RNA-seq, Bioinformatics**

I. INTRODUCTION

Tumor identification and cancer origin classification for tissue biopsy are of great significance for cancer diagnosis and molecular cancer studies. However, cancers must be determined by competent pathologists using multiple equipments and materials, such as X-ray, CT, PET-CT and pathological sections taken from fine-needle aspiration and surgeries. It is also arduous to discriminate precancerous lesions and solid tumors, primary cancers and recurrent cancers, which affect regimens applied to patients. Moreover, this process is also time-consuming. So, a cheap, convenient and fast method is in dire need.

Following computer technical development, researchers recently notice that gene expression data can be utilized in tissue biopsy area. Several data clusters can be found in figure 1 after reducing the RNA data dimension by uniform manifold approximation and projection(UMAP) algorithm[1]. This figure demonstrates the different origins of cancers have different characters of gene expression.

Chenzhao Feng is with School of Basic Medicine, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China. Email:{fengchenzhao@hust.edu.cn

Tianyu Xiang and Lingzhe Zhao are with the Department of Control Science and Engineering, College of Electronics and Information Engineering, Tongji University, Shanghai, China. Email: {1754102, lzzhao}@tongji.edu.cn

Zixuan Yi is with School of Mathematics and Statistic, Wuhan University, Wuhan, China. Email:Yizixuan826@gmail.com

Sisi He is with Department of Oncology, The Second Affiliated Hospital of Zunyi Medical University, Guizhou 563000, P. R. China.Email:nonodinary@163.com

Kunming Tian is with Institute of Reproductive Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China and Department of Preventive Medicine, School of Public Health, Zunyi Medical University, Zunyi, China. Email: nonstandstill@163.com

∗ Corresponding author

+ These authors contributed equally.

Some people try to solve these problems with DNA data. Kang et al.[8] use a probabilistic approach to achieve a 73.5% accuracy for six different types of cancers. Soh et al.[2] model the DNA sequence with SVM and achieve an accuracy of 77.7% for a 28 class prediction problem. Hao et al. [4] notice the tumor and normal tissue can also be classified with DNA data. They build a model based on TCGA DNA methylation to classify four different cancers and two different tumor status together and get an accuracy of 97.1%.

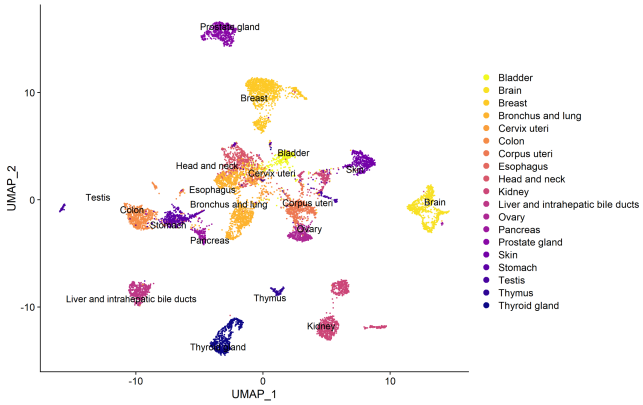

Fig. 1: This figure shows the expression of Gene data after using dimension reducing algorithm UMAP[1]. Twenty-four cancer types from nineteen anatomical sites clusters show a distinct gene expression patterns which provides the theoretical evidence of our paper.

Jung et al. [5], Pal et al.[6] and Wei et al.[23] is the first group of people trying to use RNAs to classify different cancer types. However, most of their methods have to use specific biomarkers from existing. Since the test standard of each dataset is distinct, these methods often fail to deploy on other datasets. Kang et al. [8] build a model based on genetic algorithms and Random Forest to show a high recall (92%) for thirty-two different types of the cancer classification task. Lopez et al. [9] introduce a feature selection method. From all of the 1046 miRNA data, they select only 100 different miRNAs. The few feature model performs a comparably result with the full-feature model on the cancer classification task (only 1.6% decrease on accuracy). In Laplante et al.'s work [10], a deep-learning-based method is proposed to classify twenty different types of cancer which achieve a 96.9% accuracy on TCGA dataset. They also come to an interesting conclusion that following the growth of age, the cancer type of patients will be more difficult to classified.

Sun et al.[11] also build a deep-learning-based model which can predict tumor type and origin of different cancers. Their

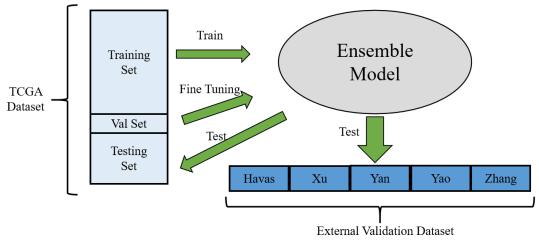

Fig. 2: This figure exhibits our overall pipeline of our model. The left blue part shows the TCGA Dataset used in ensemble models: we train our model with training set, fine tuning the parameters with validation set and test our model with testing set. The grey part is our ensemble model which works with the idea of ensemble learning. Moreover, the dark blue part is the external validation dataset using to varified the rubustness of our model on both identify tumor types and nineteen different origins of cancer tasks.

model achieves not only high performance on the TCGA dataset but also shows promising results on the external validation dataset.

However, few of these methods (both RNA based and DNA based methods) notice different RNA or DNA of different functions may contribute to the overall performance of the model. In this paper, we use the idea of ensemble learning stacking five groups of different mRNA data which serves varied functions to classify tumor types and nineteen different origins of cancer. In the cancer origin classification problem, we also apply a feature selection model selecting important features for cancer origin classification problems. With only half the number of all features, our model shows a comparable result with the full feature model on TCGA dataset.The experiment on TCGA dataset shows our method is state-of-the-art in this area. The external dataset validation shows the robustness of our model.

II. Method

In this section, we will show how our model works step by step.

A. Preprocess

Before training and testing our model, we firstly do a preprocessing step for transformation, grouping and normalizing the raw RNA data.

The first step is to transform raw data from external sources to a standard formation used in TCGA[12]. We download RNA-seq *fastq* files and process them into count matrices following the TCGA mRNA pipeline (`https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/`). Briefly, *fastq* files are aligned to GRCh38.d1.vd1.fa by STAR (version 2.7.6a) to form *bam* files and transcripts are counted from *bam* files using HTseq-count (version 0.13.5). In this way, non-TCGA datasets share a similar distribution with TCGA's. We process count matrices as the TCGA dataset do.

To escalate the overall prediction accuracy and extend biomedical interpretation, we build five sub-models with different gene (feature) lists.Transcription factors are downloaded from AnimalTFDB (`http://bioinfo.life.hust.edu.cn/AnimalTFDB/#!/`). Cancer testis antigen are downloaded from CTDatabase (`http://www.cta.lncc.br/`). Cancer cell metabolism genes are downloaded from ccmGDB (`https://bioinfo.uth.edu/ccmGDB/`). Autophagy gene list is retrieved from the Autophagy database (`http://www.tanpaku.org/autophagy/index.html`). Duplicated genes in different lists.

Some mRNA data, even selected by the pipeline mentioned above, when we check them, we find all of the patients show the same expressions, such as 'PTPRK', 'ACAN', 'ACADs' ... For these mRNAs, we directly remove them from our dataset.

Finally, all the RNA data is normalized within 0 to 1 because most machine learning algorithms perform better with scaled data. For each mRNA:

$$mRNA_i = \frac{mRNA_i - min(mRNA_i)}{max(mRNA_i) - min(mRNA_i)} \quad (1)$$

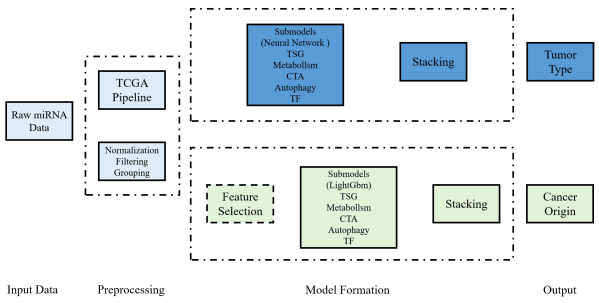

Fig. 3: This figure shows the overall structure of our models. The left blue part, including input data and pre-process module, is shared by tumor type identification and cancer origin classification models. The middle blue part is the tumor type identification part–a neural-network-based ensemble model that stacks five neural networks. The input data of each submodel is a mRNA group serving different functions. The green part is the cancer origins classification model. The general idea is similar to the tumor type identification model, but it has two unique parts: the base model is lightgbm, the other is a feature selection module that chooses the important mRNA. The feature selection part is optional. If we deploy this part in our model, only half of the features will be used, and the model performs comparably good results with the full feature model.

B. Feature Selection

After data normalizing and classification, we use Linear SVM and L1 regularization to select significant mRNAs which showed very high correlations with 19 anatomical sites on training set. For each anatomical site situation, We determine which mRNA features are related to the position by obtaining the L1 regularization minimum solution of the loss function:

$$\min \sum_{i=1}^{N} [1 - y_i(\omega^T x_i + b)]_+ + \lambda ||\omega||_1 \tag{2}$$

Where $N = 1090$, y_i represents the ith sample's tumor situation in this site, and $\mathbf{x}_i$ represents the ith sample's all mRNA features. ω, b are the parameters of the loss function. In this equation, we apply L1 penalty to obtain sparse solutions of ω. Then, mRNAs, whose coefficient in this least solution was more than 0 were chosen. Following this principle, 1693 significant mRNAs are selected from the original 3239 features.

Linear models using L1 norm as penalty term will get sparse solution–the coefficients of most features are 0, so we can have a more reduced model with less feature. Moreover, the regularization method reduces the risk of overfitting.

C. Submodel

In this part, we will discuss the Mathematical form of each submodel. As we build two models for two different tasks – tumor identification and cancer origins classification.

We will introduce each model separately. The general form of each model is quite similar – with five different submodels digging five different groups of RNA which serve a specific function. Since the tasks are different, the submodel of each task is distinct. For tumor identification, the submodel is a two-layer neural network; for cancer origins classification, the submodel is lightgbm. The final step of these two tasks are similar – the output of each submodel are given different weight and combined with the final output of our model.

For a specific task, the structure of each model is the same. We denote submodel i ($i = 1, 2, 3, 4, 5, 6$) for each submodel and will discuss the mathematics formation of each submodel.

1) Neural Network: We first build a Neural Network model trying to identify tumour types.

The formation of each sub-model is the same, which is a two-layer neural network. The formation can be written as follows:

$$\hat{y}_i = f_2(f_1(w_{i2} f_1(w_{i1} input_i + b_{i1}) + b_{i2})) \tag{3}$$

$\hat{y}_i$ is predicted by submodel i ($i = 1, 2, 3, 4, 5, 6$). w_{i1} and w_{i2} are weights of the first and the second layers of submodel i; b_{i1} and b_{i2} are bias of the first and the second layers of submodel i; f_1 and f_2 are activation functions. f_1 is relu:

$$f_1 = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases} \tag{4}$$

f_2 is sigmoid:

$$f_2 = \frac{1}{1 + e^{-x}} \tag{5}$$

Since the range of the sigmoid function is between 0 and 1, we can then treat the output of this function as the possibility of whether the tissue is normal.

As the normal samples are much less than tumor samples(about $1 : 16$), we introduce focal loss[13] to deal with the extremely unbalanced data distribution. The loss function is written as follows:

$$L = \begin{cases} -\alpha(1 - \hat{y})^{\lambda} log\hat{y} & y = 1 \\ -(1 - \alpha)\hat{y}^{\lambda} log(1 - \hat{y}) & y = 0 \end{cases} \tag{6}$$

2) Lightgbm: The lightgbm[14] model is built to identify the origins of different types of cancers.

Lightgbm is one of Gradient Boost Decision Tree(GBDT) algorithms. Among all of the GBDT methods, lightgbm is one of the most effective algorithms, because it introduces some tricks like Gradient-based One-Side Sampling(GOSS) and EFB(Exclusive Feature Bundling) to solve the defect that too much time consumption when the dimension of input is large.

For traditional GBDT methods like XGboost[15], the original objective function is expanded as the second-order of Taylor expansion.

$$L \simeq \Sigma_{i=1}^{n} [g_i f(x_i) + \frac{1}{2} h_i f^2(x_i)] + \omega(f) \tag{7}$$

L is the loss function of GBDT model; g_i and h_i are the first and second derivatives of the loss function; x_i is the ith dimension of input and ω is regularization terms for preventing over-fitting. From this function, we can find that the computing cost will expansion for the second-order of derivatives is tough to compute, when the dimension of input data grows.

Lightgbm solves this problem by introducing a novel GOSS mechanism which can maintain a balance between the accuracy of model and time consumption of training the model. Based on this idea, all the features are divided into two groups – one is large gradient samples and another is small gradient. The model will retain all the large gradient features and randomly choose some of the small gradient ones. For instance, top $a\%$ of the gradient of features are selected as a large gradient(LG) group and $b\%$ of rest features are used as a small gradient(SG) group. Only these two groups of data are used in training. The idea can be found in follows:

$$G = \Sigma_{x_i \in LG} g_i + \frac{1 - a\%}{b\%} \Sigma_{x_i \in SG} g_i \tag{8}$$

EFB helps to deal with high-dimensional sparse features by binding mutually exclusive features into a single feature. Also, leaf-wise and parallel tricks ensure lightgbm efficiency.

Since this model is used to classify nineteen different types of cancer origins, we use multiple cross entropy as the loss function:

$$L = -\frac{1}{N}\Sigma_{i=0}^{N-1}\Sigma_{k=0}^{K-1} y_{i,k} log p_{i,k} \quad (9)$$

While N is total samples, the label of $y_{i,k}$ means the ith sample in group k and $p_{i,k}$ is predicted by model for the possibility of ith sample in group k.

D. Model Integration

Finally, the submodels will be integrated into an ensemble model for final prediction. As both models for different tasks share the same idea, they will be illustrated together.

$$P = \Sigma_{i=1}^{5} \alpha_i p_i \quad (10)$$

P denotes the final output of our model, p_i is the prediction of the ith submodel and α_i is the weight of each model. In the training step, we divide the whole dataset into the training set, testing set and validation set. The performance of each model on the validation set is used as the weight of each model.

III. Experiments

This section will show the training details of our models. And then, the simulation results on TCGA[12] and other validation datasets will be discussed.

A. Training Details

This work is implemented by TensorFlow[16](first model) and lightgbm[14](second model) framework.

For the first model, α is set to 0.2 and λ is set to 2. To optimize this neural network, we applied Adam[17] optimizer and the learning rate of all of the submodels is set to 0.01. The weights of networks are initialized by glorot[18] uniform distribution.

The second model is based on lightgbm[14]. Since there are many hyper-parameters for lightgbm, we fine-tuning the model with an effective hyper-parameter optimization framework called optuna[20]. The specific hyperparameter can be found on our project website `https://github.com/GARYXTY/TCGA-Project`.

B. Dataset Description

We download level 3 RNA-seq raw count matrices of 24 cancer types in TCGA from UCSC Xena (`https://xenabrowser.net/datapages/`). These 24 types of cancers originate from 19 anatomical sites. Raw counts are normalized into TPM values which are more comparable between samples. ENSEMBL-IDs are annotated as gene symbols using gencode.v22.annotation.gtf (`https://www.gencodegenes.org/`). When encounter with duplicated gene symbols corresponding to different transcripts IDs, transcripts that owns the highest median values would represent the genes. We train and cross-validated models in the TCGA cohort.

We split TCGA dataset into three different cohorts – training set(67.5%), testing set(25%) and validation set(7.5%). We train the model with the training set and fine-tuning it with the validation set. The best model on the validation set will be applied to the testing set to evaluate our model's final performance.

We enroll five external validation cohorts to further demonstrate our models' accuracy, two of which are breast cancers (The detailed information can be found on `https://github.com/GARYXTY/TCGA-Project`). The remaining datasets are cervical cancer, clear cell renal cell cancer and gastric cancer. We download RNA-seq *fastq* files and process them into count matrices following the TCGA mRNA pipeline (`https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/`). Briefly, *fastq* files are aligned to GRCh38.d1.vd1.fa by STAR (version 2.7.6a) to form *bam* files and transcripts were counted from *bam* files using HTseq-count (version 0.13.5). In this way, non-TCGA datasets share a similar distribution with TCGA's. We process count matrices as described above.

C. Result on Tumor Classification

Our approach achieves an accuracy of 98.3% on the TCGA dataset. The comparison of our methods and other state-of-the-art on this task can be found in table I. Even using the same dataset, the samples and raw data are different among these methods. So the accuracy can only be used as a reference index. From this table, we can find our model also show outstanding performance.

Study	Approach	Accuracy
Sun et al.[11]	Deep Learning	98.2%
Peng et al. [21] et al	Unsupervised clustering	92%
Ours	Ensemble Deep Learning	**98.3%**

TABLE I
Tumor Type identification result on TCGA Data set: from this table, we can find that our model shows a state-of-the-result on the TCGA dataset.

D. Result on Cancer Origin Classification

Our approach achieves an accuracy of 96.1% for the task cancer origin classification among 19 different types. The F1 score of different types of cancer can be found in figure 5. We can find that even if the fewer feature model uses only half of the RNA features in the full model, its performance does not decrease so much. Among all of these cancers, the F1 score of Stomach and Esophagus is relatively poor. This phenomena can also be found in figure 4. What's more, we can found that nearly half of Esophagus cancer samples are mistakenly recognized as Stomach samples. This might be attributed to the fact that esophagus and stomach own similar and continuous epithelial tissues. Besides, esophageal adenocarcinomas resemble a subset of gastric cancers, which

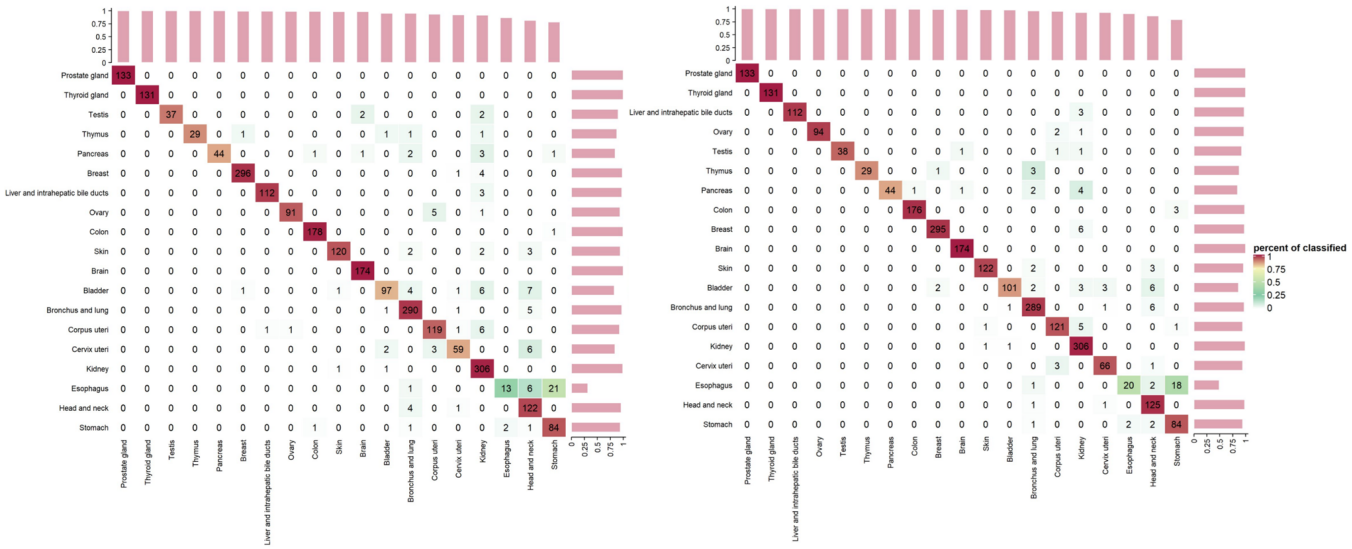

Fig. 4: This figure shows the confusion matrix of both few feature model and full feature model. The left figure is the model with selected features and the right figure is the model with full features. It can be found that most types of cancer show a plausible result on the TCGA dataset. Moreover, both of selected model and full feature model show similar performance – 100% accuracy on Prostate gland and Thyroid gland, relatively low performance on Stomach and Esophagus.

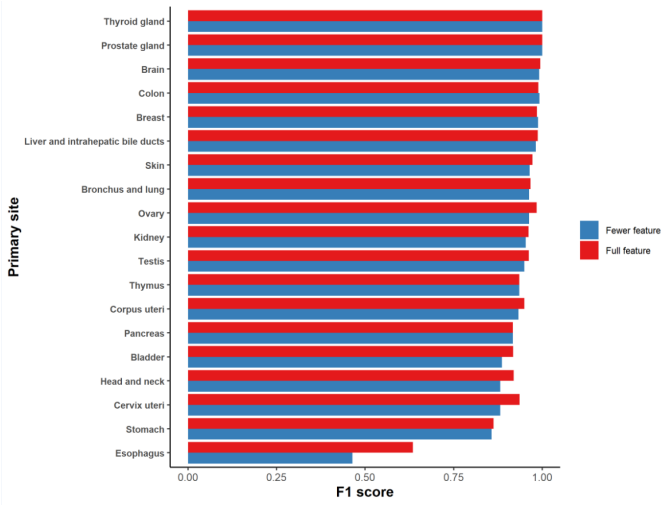

Fig. 5: This figure shows the F1 score of our model on nineteen different cancer origins classification tasks. The red bar is the model trained with all of the features and the blue bar shows the result of half of the features. From this figure, we can see fewer features model show comparable results with the full feature model. The overall performance of both models is plausible – eighteen different types of cancers have an F1 score over 0.8.

suggest that these two cancers could be considered a single disease entity as previously reported[19].

We also compare our model with other state-of-the-art methods in this task(see table II). However, as the number of samples and the classification types of these models are different, these models' performance can not be reflected by the accuracy.

Study	Method	ACC	types
Sun et al.[11]	Deep Learning	98.2%	11
Li et al. [22]	K nearest neighbors	95.6%	31
Wei et al [23].	Logistic regression	90.5 %	26
Tang et al [24].	Random forest	96.3 %	14
Laplante et al. [10]	Deep Learning	96.9%	20
Ours(few)	Ensemble Lightgbm	94.7%	19
Ours(full)	Ensemble Lightgbm	96.1%	19

TABLE II
Cancer origins classification result on TCGA dataset: in this table, we can compare our model and other state-of-the-art methods of cancer origins classification task on the TCGA dataset. This table shows our method state-of-the-art in this area.

E. Performance on External Validation Dataset

To evaluate the robustness of our model, we test our model in five different external validation datasets. All of these models are only trained on a TCGA training set and tested directly with this data. It should be noticed that the raw data of these five datasets are preprocessed with TCGA steps. The detailed reuslt can be found `https://github.com/GARYXTY/TCGA-Project`.

From tumor identification problem, the average performance is relatively good with a 92.6% accuracy of all of the datasets. For a cancer origin classification problem, the performance is not so good as the result on the TCGA dataset. However, from table III, we can find that the result is similar with TCGA testing dataset – cancers like cervix uteri(rank 17 in F1 score) and Stomach (rank 18 in F1

score)'s performances are relatively bad, Kidney(rank 10 in F1 score) and breast(rank 5 in F1 score) show impressive results.

Dataset	Samples(n/type)	TC	CTC(full)	CTC(few)
Havas et al.	14(Breast)	100%	92.86%	92.86%
Xu et al.	12(Breast)	58.33%	100%	100%
Zhang et al.	68(Cervix uteri)	100 %	85.29 %	70.59%
Yao et al.	20(Kidney)	75%	100 %	100%
Yan et al	21(Stomach)	100 %	52.38 %	52.38%
All	135	92.6%	84.44%	77.4%

TABLE III
Performance of Our Model on External Dataset: TC means Tumor Classification, CTC means cancer type classification, full and few mean if we use all of the features to train the model.

IV. CONCLUSIONS

In this work, we proposed an assemble learning model dealing with tumor classification and cancer classification problems. For a tumor identification problem, we deploy a deep-learning-based assemble method. It achieves 98.2% accuracy on the TCGA task and 92.6% accuracy in this task. For the cancer origins classification problem, we deploy a lightgbm based assemble method. We use both full features and selected features in this model. Even the model-based selected features are only half of the full features based model, it comparably high performance on TCGA data. Also, validation experience on the external datasest shows the robustness of our model.

In the future, we will continue elevating the performance of this model. Even though our model's overall performance is state-of-the-art on TCGA dataset, there are still some drawbacks of our model. For instance, our model can hardly identify gastric and esophageal cancers. The input gene expression data are limited to RNA-seq data on Illumina platforms, which means array-based transcript quantification methods could not be directly used. We enrolled five groups of genes to infer sample characteristics, however, we do not uncover specific biological links between genes and outcomes.

V. ACKNOWLEDGEMENT

This work is supported by the Young Elite Scientist Sponsorship Program by CAST (2018QNRC001) and China postdoctoral science foundation (2020M672363).

REFERENCES

[1] McInnes, L., Healy, J., Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

[2] Soh, K. P., Szczurek, E., Sakoparnig, T., Beerenwinkel, N. (2017). Predicting cancer type from tumour DNA signatures. Genome medicine, 9(1), 1-11.

[3] Kang, S., Li, Q., Chen, Q., Zhou, Y., Park, S., Lee, G., ... Zhou, X. J. (2017). CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. Genome biology, 18(1), 1-12.

[4] Hao, X., Luo, H., Krawczyk, M., Wei, W., Wang, W., Wang, J., ... Zhang, K. (2017). DNA methylation markers for diagnosis and prognosis of common cancers. Proceedings of the National Academy of Sciences, 114(28), 7414-7419.

[5] Jung, S., Bi, Y., Davuluri, R. V. (2015). Evaluation of data discretization methods to derive platform independent isoform expression signatures for multi-class tumor subtyping. BMC genomics, 16(11), 1-10.

[6] Pal, S., Bi, Y., Macyszyn, L., Showe, L. C., O'Rourke, D. M., Davuluri, R. V. (2014). Isoform-level gene signature improves prognostic stratification and accurately classifies glioblastoma subtypes. Nucleic acids research, 42(8), e64-e64.

[7] Wei, I. H., Shi, Y., Jiang, H., Kumar-Sinha, C., Chinnaiyan, A. M. (2014). RNA-Seq accurately identifies cancer biomarker signatures to distinguish tissue of origin. Neoplasia, 16(11), 918-927.

[8] Kang, S., Li, Q., Chen, Q., Zhou, Y., Park, S., Lee, G., ... Zhou, X. J. (2017). CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. Genome biology, 18(1), 1-12.

[9] Lopez-Rincon, A., Martinez-Archundia, M., Martinez-Ruiz, G. U., Schoenhuth, A., Tonda, A. (2019). Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection. BMC bioinformatics, 20(1), 1-17.

[10] Laplante, J. F., Akhloufi, M. A. (2020, July). Predicting Cancer Types From miRNA Stem-loops Using Deep Learning. In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC) (pp. 5312-5315). IEEE.

[11] Sun, K., Wang, J., Wang, H., Sun, H. (2018). GeneCT: a generalizable cancerous status and tissue origin classifier for pan-cancer biopsies. Bioinformatics, 34(23), 4129-4130.

[12] Tomczak, K., Czerwińska, P., Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary oncology, 19(1A), A68.

[13] Lin, T. Y., Goyal, P., Girshick, R., He, K., Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

[14] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30, 3146-3154.

[15] Chen, T., Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[16] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Ghemawat, S. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.

[17] Kingma, D. P., Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[18] Glorot, X., Bengio, Y. (2010, March). Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics (pp. 249-256). JMLR Workshop and Conference Proceedings.

[19] Cancer Genome Atlas Research Network. (2017). Integrated genomic characterization of oesophageal carcinoma. Nature, 541(7636), 169.

[20] Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery data mining (pp. 2623-2631).

[21] Peng, L., Bian, X. W., Xu, C., Wang, G. M., Xia, Q. Y., Xiong, Q. (2015). Large-scale RNA-Seq transcriptome analysis of 4043 cancers and 548 normal tissue controls across 12 TCGA cancer types. Scientific reports, 5(1), 1-18.

[22] Li, Y., Kang, K., Krahn, J. M., Croutwater, N., Lee, K., Umbach, D. M., Li, L. (2017). A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. BMC genomics, 18(1), 1-13.

[23] Wei, I. H., Shi, Y., Jiang, H., Kumar-Sinha, C., Chinnaiyan, A. M. (2014). RNA-Seq accurately identifies cancer biomarker signatures to distinguish tissue of origin. Neoplasia, 16(11), 918-927.

[24] Tang, W., Wan, S., Yang, Z., Teschendorff, A. E., Zou, Q. (2018). Tumor origin detection with tissue-specific miRNA and DNA methylation markers. Bioinformatics, 34(3), 398-406.