Towards Cancer Patients Classification Using Liquid Biopsy *

Sebastian Cygert[1][0000−0002−4763−8381], Franciszek Górski[1], Piotr Juszczyk[1], Sebastian Lewalski[1], Krzysztof Pastuszak[1,2][0000−0002−5562−0114], Andrzej Czyżewski[1][0000−0001−9159−8658], and Anna Supernat[1,2][0000−0002−3113−0540]

[1] Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Poland
[2] Intercollegiate Faculty of Biotechnology, University of Gdańsk and Medical University of Gdańsk, Poland

Abstract. Liquid biopsy is a useful, minimally invasive diagnostic and monitoring tool for cancer disease. Yet, developing accurate methods, given the potentially large number of input features, and usually small datasets size remains very challenging.
Recently, a novel feature parameterization based on the RNA-sequenced platelet data which uses the biological knowledge from the Kyoto Encyclopedia of Genes and Genomes, combined with a classifier based on the Convolutional Neural Network (CNN), allowed significantly improving the classification accuracy. In this work, we take a closer look at this approach and find that similar results can be obtained using significantly smaller models. Additionally, competitive results were achieved using gradient boosting. Since it has another advantage of adding interpretability to the model, we further analyze it in this work.

Keywords: Image-based classification · Tumor Educated Platelets · RNA sequencing · liquid biopsy.

1 Introduction

Liquid biopsies offer a minimally invasive sample collection instead of tissue biopsies of solid tumors, traditionally used in cancer evaluation. The most common material for this type of analysis is blood: the source of circulating tumor DNA, circulating tumor cells, miRNAs, exosomes and, lately, tumor-educated platelets

* Corresponding author : sebcyg@multimed.org. This work has been partially supported by Statutory Funds of Electronics, Telecommunications and Informatics Faculty, Gdansk University of Technology. This work was supported in part through the European Regional Development Fund as part of the Project entitled: Academy of Innovative Applications of Digital Technologies under Grant The Operational Programme "Digital Poland" 2014-2020 number POPC.03.02.00-00-0001/20-00. This research was supported by the SONATA grant of the National Science Centre (2018/31/D/NZ5/01263) and Medical University of Gdańsk statutory work (ST-23, 02-0023/07).

(TEPs). The introduction of high-throughput sequencing techniques allowed for the unprecedented resolution of the analysis. However, generated data complexity and a multitude of features created the need for more advanced approaches than assuming a simple cut-off for final result interpretation. The utility of Support Vector Machine (SVM) and Particle Swarm Optimization-enhanced SVM, known as throbmoSeq classifier, applied to sequenced RNA of tumor educated platelets has already been demonstrated for cancer detection (e.g., non-small cell lung cancer, breast cancer, sarcoma) [3].

Recent work [20] further improved the classification accuracy by implementing biological knowledge on the sequenced RNA molecules from the Kyoto Encyclopedia of Genes and Genomes [10]. Features obtained from an RNA-sequenced platelet, were converted into images and classified by custom-built CNN architecture, resulting in a significant improvement. Their approach introduced two main novelties:

1. using novel feature extraction step.
2. using the CNN-based model with a custom architecture for classification.

However, from the paper, it is unknown whether the improvements come from the new feature extraction step, using a CNN-based model, or using custom architecture. Therefore, we take a step-by-step approach in this work, starting with the standard CNN architectures and detailed data analysis. To improve CNN classification accuracy, standard techniques such as ImageNet pretraining, Dropout [27] and mixup data augmentation [28] are applied.

Finally, other machine learning approaches such as k-nearest neighbors (kNN) and gradient boosting [11] were applied, to compare their accuracies with a CNN-based approach. To sum up, the contributions of this work are as follows:

- We performed an ablation study on the CNN-based classifier using different architectures and regularization strategies to improve model accuracy.
- It was shown that the CNN models are not crucial to the final model performance and similar accuracy can be obtained by using gradient tree boosting. It has another advantage of adding interpretability to the model, which is briefly analyzed in our work.

2 Method

2.1 Parameterization

Briefly, raw RNA-sequencing data encoded in FASTQfiles were subjected to a standardized RNA-sequencing alignment pipeline, as described in Thromboseq protocol [2]. The expression data for each sample were then normalized using DESeq2 package [17] with Variance Stabilizing Transformation [14]. Gencode v19 GRCh37 annotation [10] was used for annotation. Transcripts that could not be mapped to a transcript with Gencode status "known" were excluded.

Filtered expression profiles were then used to build images. Each row corresponds to a signaling pathway from the KEGG database [10]. Each pixel in

a row corresponds to the expression level of a single transcript from the pathway. Pathways from the KEGG database corresponding to three aspects: cancer, metabolism and signaling processes, were selected. R package gage was used to gather KEGG pathway data [18]. As a result, a feature vector is a two-dimensional array with 345 rows (number of signaling pathways) and 243 columns (length of the longest pathway).

In [20] such parameterized data is then directly fed into the CNN. However, because each row contains a different number of values, more than half of the values in the array are empty (filled with 0s). As such, we experiment with another input variant where all values are put into a square of minimal size (by simply removing empty values in the original arrays). Only the last few values in the last row are empty. However, now each row may contain data from different signaling pathways. It allowed us to reduce the size of the array to 143 rows by 143 columns and reduce input dimensionality from 83835 to 20449 values, which could make the task easier for the classifier (especially given the limited data). In the experiment section, this variant is named *reduced*.

2.2 Methods

In this section our methods for liquid biopsy data classification are described.

While CNNs were originally developed for computer vision, they were used in a much larger set of applications, including analysis of EEG signals [5], calling genetic variants [23], or text classification [7]. It is because CNNs are parameter-efficient algorithms exploiting local feature patterns. However, for many applications labelled data are scarce, which makes the training very challenging. In such a scenario it was shown that usually smaller architectures are already very efficient [24]. In this work, standard CNN architecture, namely ResNet [12] is used for liquid biopsy classification, and due to the small data regime, smaller variants of the ResNet architecture are used. Additionally, it was tested whether ImageNet pretraining helps in our scenario.

A standard approach to prevent model overfitting is to use some form of data augmentation. While many forms of data augmentation exist for images (e.g., rotations, translations, changes in color), none of the standard forms of data augmentation can be applied to liquid biopsy data. As such, in this work a data-agnostic *mixup* [28] augmentation routine is applied to the data. It creates new data samples by means of linear interpolation between existing data:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j$$
$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j$$

where (x_i, y_i) and (x_j, y_j) are randomly selected training pairs of input vectors and the corresponding label, and $\lambda \in [0, 1]$ is the interpolating factor. In the original paper λ is drawn from a symmetric Beta distribution and its α value is a hyperparameter. Despite its simplicity, *mixup* is a powerful technique that works as a strong regularizer on the model. Apart from *mixup*, other experiments were

conducted using another regularization technique, namely Dropout [27], which works by randomly zeroing output from some of the neurons during training.

Another popular machine learning technique is a gradient tree boosting [11], which often performs very well on diverse applications such as ranking problems [4], recent COVID-19 patient deterioration prediction [25] and many others [19]. A great advantage of gradient boosting algorithms is their interpretability, especially important for medical applications. Additionally, a number of well-maintained open-source libraries are available (e.g., XGBoost[6], LightGBM [16]. A popular XGBoost library is used for data processing in this work, and standard hyperparameter search is applied over selected parameters.

Finally, a simple k-nearest neighbors classifier was applied to the problem. While we do not expect it to perform better than previous methods, it is an important baseline for comparing results to. It could be possible that the classification improvements in [20] were possible mainly due to the novel parameterization, and in such a case even a simple k-nearest neighbors algorithm would perform well.

3 Experiments

3.1 Evaluation

Table 1. Datasets used for experimentation.

Name	train samples	test samples	imbalance ratio	details
OC [20]	158	104	8.36	ovarian cancer
NSCLC [3]	157	447	1.96	non-small cell lung cancer
Sarcoma [13]	118	56	1.8	sarcoma

Three publicly available datasets were used to test the classifier (Table 1): 401 non-small cell lung cancer patients (NSCLC) and 203 healthy controls [3], 62 sarcoma and 37 former sarcoma patients who recovered at least 5 years ealier, now treated as healthy) and 75 healthy controls [13] and the original imPlatelet dataset consisting of 204 healthy controls and patients with ovarian cancer (28) or benign gynaecological conditions (30) [20].

For evaluation, the same setting as in [20] was followed, i.e., the model is evaluated on exactly the same held-out test set. Then, the rest of the data is split into train and validation parts and a 5-fold stratified validation is run.

Balanced accuracy is used to measure performance, which is defined as:

$$\text{Balanced accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}$$

where sensitivity is the true positive rate and the specificity is the true negative rate.

3.2 CNN model

For the CNN approach, standard ResNet backbones are used for classification. Since the amount of data is limited, we focused on smaller variants of the ResNet, namely ResNet-18 and ResNet-34 from the PyTorch library [21] were tested. First, a detailed study of the influence of different factors is performed on the ovarian cancer dataset. To find the best model a grid search is executed, where the learning rate is sampled from the range [0.01, 0.1], Dropout rate $\in \{0.2, 0.5\}$ and L1-weights regularization $\in \{0.001, 0.0001\}$. During training standard cross-entropy classification loss it optimized. The model is being trained for 100 epochs and the model for testing is chosen based on the balanced accuracy. As there are still minor differences between the same runs, each experiment is repeated 3 times, and mean accuracy is reported.

In the first experiment, two parameterization approaches for the CNN model, as described in section 2.1, were compared (Table 2). In general all models perform very well on the validation set (balanced accuracy ranging from 88% to 92%), and as expected, have a few percent accuracy drops on the test set. In general, there is no clear winner for one specific architecture or data parameterization. Note, that the model with the highest validation accuracy obtain the lowest score on the test set. For the next round of experiments, it was decided to utilize the *reduced* parameterization, as it performs similarly to the original parameterization, while using a much smaller input size.

Table 2. Accuracy of different backbones and parameterization on ovarian cancer classification. Balanced accuracy reported.

Backbone	**Validation acc.**	**Test acc.**
Standard parameterization [20]		
ResNet-18	0.9080	**0.8958**
ResNet-34	0.8793	0.8317
Reduced parameterization		
ResNet-18	0.8938	0.8563
ResNet-34	**0.9218**	0.8255

Further, it was tested whether using ImageNet pretraining can improve the final accuracy (Table 3). As it can be noticed, using ImageNet pretraining and mixup improved the validation and testing accuracy. Also the variance in classification accuracy was reduced, which shows that using the above methods helped to stabilize trained models. Balanced accuracy reported.

Finally, evaluation was performed on NSCLC and Sarcoma datasets. For the NSCLC dataset the best model obtained a balanced accuracy of 86,52% on the test set. The Sarcoma dataset turned out to be challenging, which might be because of the limited dataset size. When doing standard 5-fold validation, it turned out that the accuracy on the balanced dataset is poorly correlated with the accuracy on the test set. As such, the models were trained for 100 epochs and

Table 3. Effects of ImageNet pretraining and *mixup* data augmentation on the ovarian cancer classification.

Model	**Validation. acc.**	**Test. acc.**	**Test std.**
ResNet-18	0.8938	0.8563	0.0614
ResNet-34	0.9218	0.8255	0.0738
ImageNet pretraining			
ResNet-18	0.9236	0.8952	**0.0056**
ResNet-34	0.9236	0.8652	0.0229
mixup			
ResNet-18	**0.9379**	0.8798	0.0242
ResNet-34	0.9042	0.8221	0.0477
mixup + ImageNet pretraining			
ResNet-18	0.9343	**0.9043**	0.0328
ResNet-34	0.9343	0.8782	0.0185

simply the model from the last epoch was used for testing. It was possible since no signs of overfitting were noticed, which might be due to the used regularization techniques, i.e., dropout. In such a setting, the Sarcoma balanced accuracy was 94,09%.

3.3 Other algorithms

In this section experiments with tree gradient boosting and kNN model are presented. For the XGBoost model the following parameters were used in the hyperparameter search: maximal depth of a tree $\in \{1, 2, 3, 4\}$, number of boosting stages $\in \{50, 100, 300, 500\}$ and learning rate $\in \{0.1, 0.01\}$, following insights from [19].

In the case of kNN algorithm, given the large dimensionality of the input space, first, a Principal Component Analysis (PCA) is performed using *scikit* library [22]. Then the grid search for the kNN algorithm is applied, searching for the number of neighbours ($n \in \{1, 2, 3\}$) and number of principal components (explained variance in $\{0.6, 0.7, 0.8, 0.9, 0.95, 0.99, 0.999\}$. For evaluation, the same procedure is applied as in the CNN model. Stratified 5-fold cross-validation is used for model selection (PCA is separately computed for each fold), and models with the best validation accuracy are used for testing (Table 4).

As expected the kNN algorithm is the worst performing algorithm. However, on the OC dataset it reached 69.06% of balanced accuracy which is a fair result. When comparing gradient boosting and CNN classifier, CNN scores similar on the OC dataset and better on remaining datasets, and CNN accuracy is the most stable across datasets. However, note that the CNN model is the only one that used data augmentation, so it is very likely that gradient boosting would benefit from it, especially on the Sarcoma dataset, on which the model is overfitting (large difference between validation and test accurracy).

Table 4. Accuracy comparison of different classification methods on all datasets. Balanced accuracy reported.

Aggregation method	**Validation acc.**	**Test acc.**
OC		
CNN	0.9343	**0.9043**
Boosting	1.0	0.8991
kNN	0.7556	0.6906
NSCLC		
CNN	0.9129	**0.8652**
Boosting	0.76	0.7343
kNN	0.61	0.5299
Sarcoma		
CNN	1.00	**0.9409**
Boosting	0.9818	0.6316
kNN	0.4522	0.3592

3.4 Discussion

In this work various machine learning approaches were evaluated on the task of cancer patient classification using liquid biopsy. It was found that both standard CNN-based models and gradient boosting algorithms perform very well. However, all of the models are sensitive to the hyperparameters. It is because of the limited size of datasets and a high number of input features. At the same time, it is expected that ensembling results from different models will further increase and stabilize the performance.

Compared to the recent work [20] it was found that standard CNN backbones (i.e. ResNet architecture) can perform very well on the task (as opposed to the custom architecture). Further, it was shown that other models (i.e., gradient boosting) could also perform very well on the task, given the novel parameterization proposed in [20]. Our models performed better on the Sarcoma dataset and worse on the OC dataset and NSCLC datasets. At the same time, the CNN model used in our work is significantly smaller in terms of a number of parameters.

Various regularization techniques were tested for the CNN model (Dropout, *mixup* data augmentation, ImageNet pretraining, L1-weight regularization). At the same time, there is no combination of methods and hyperparameters that work the best on all datasets; in general, those methods allowed to improve and stabilize the performance.

The importance of features was calculated using XGboost built-in functionality and depended on the number of splits a particular feature was involved in. In the NSCLC dataset, RPL7A showed the highest importance. This gene has been studied only in relation to osteosarcoma [29]. Four genes demonstrated consistently high importance in the detection of ovarian cancer - SH3GL2 involved in breast [15] and lung cancer [8], PRPF6, which is involved in tumor

growth in colon cancer [1], HLA-DRA, which expression levels are known to affect the prognosis of a number of malignancies [9] and UGT2B7, which mutations are known to increase the risk of breast and colorectal cancer [26]. UGT2B7 has not been studied in relation to ovarian cancer and the research on PRFR6 and SH3GL2 has been minimal. Since the performance of gradient boosting on sarcoma dataset was poor, no analysis of feature importance was performed. The analysis of feature importance may provide additional targets for further research on the biological background of studied cancers.

4 Conclusions

In this work an analysis of different machine learning approaches to patients classification using liquid biopsy data, was presented. It was found out that given the novel parameterization presented in [20], standard CNN-based models and gradient boosting methods are very effective.. However, because of the limited datasets size, and significant size of the input space, different regularization techniques (such as dropout, mixup data augmentation) are crucial to the final performance of the model and its stability.

Gradient boosting allowed us to add interpretability to the model. Using data augmentation for the gradient boosting model to improve and stabilize its performance is essential for future work. It will also allow us to perform a more detailed analysis of the importance of the features returned by the model.

Acknowledgments

Authors would like to thank Tomasz Bączek, Myron Best, Jacek Bigda, Peter Grešner, Jacek Jassem, Tomasz Stokowy, Thomas Würdinger, Anna Żaczek for constant support.

References

1. Adler, A.S., et al.: An integrative analysis of colon cancer identifies an essential function for PRPF6 in tumor growth. Genes Dev **28**(10), 1068–1084 (2014)
2. Best, M.G., In 't Veld, S.G.J.G., Sol, N., Wurdinger, T.: RNA sequencing and swarm intelligence-enhanced classification algorithm development for blood-based disease diagnostics using spliced blood platelet RNA. Nature protocols **14**(4), 1206–1234 (2019)
3. Best, M.G., et al.: Swarm intelligence-enhanced detection of non-small-cell lung cancer using tumor-educated platelets. Cancer cell **32**(2), 238–252 (2017)
4. Burges, C.J.: From ranknet to lambdarank to lambdamart: An overview. Tech. Rep. 23-581 (2010)
5. Cecotti, H., Graser, A.: Convolutional neural networks for p300 detection with application to brain-computer interfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(3), 433–445 (2010)

6. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794. ACM (2016)
7. Conneau, A., Schwenk, H., Barrault, L., LeCun, Y.: Very deep convolutional networks for text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017
8. Dasgupta, S., et al.: SH3GL2 is frequently deleted in non-small cell lung cancer and downregulates tumor growth by modulating EGFR signaling. Journal of Molecular Medicine **91**(3), 381–393 (2013)
9. Dunne, M.R., et al.: HLA-DR expression in tumor epithelium is an independent prognostic indicator in esophageal adenocarcinoma patients. Cancer Immunology, Immunotherapy **66**(7), 841–850 (2017)
10. Frankish, A., et al.: GENCODE reference annotation for the human and mouse genomes. Nucleic acids research **47**(D1), D766–D773 (2019)
11. Friedman, J.H.: Greedy function approximation: A gradient boosting machine. Annals of statistics pp. 1189–1232 (2001)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016)
13. Heinhuis, K.M., et al.: Rna-sequencing of tumor-educated platelets, a novel biomarker for blood-based sarcoma diagnostics. Cancers **12**(6) (2020)
14. Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., Vingron, M.: Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics **18**(suppl_1), S96–S104 (2002)
15. Kannan, A., et al.: Mitochondrial Reprogramming Regulates Breast Cancer Progression. Clinical cancer research **22**(13), 3348–3360 (2016)
16. Ke, G., et al.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems **30**, 3146–3154 (2017)
17. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biology **15**(12), 1–21 (2014)
18. Luo, W., Friedman, M.S., Shedden, K., Hankenson, K.D., Woolf, P.J.: Gage: generally applicable gene set enrichment for pathway analysis. BMC bioinformatics **10**(1), 1–17 (2009)
19. Olson, R.S., Cava, W.G.L., Mustahsan, Z., Varik, A., Moore, J.H.: Data-driven advice for applying machine learning to bioinformatics problems. In: Biocomputing 2018: Proceedings of the Pacific Symposium. pp. 192–203 (2018)
20. Pastuszak, K., et al.: implatelet classifier: image-converted rna biomarker profiles enable blood-based cancer diagnostics. Molecular Oncology (2021)
21. Paszke, A., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32 (2019)
22. Pedregosa, F., et al.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
23. Poplin, R., et al.: A universal snp and small-indel variant caller using deep neural networks. Nature biotechnology **36**(10), 983–987 (2018)
24. Raghu, M., Zhang, C., Kleinberg, J.M., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. In: Annual Conference on Neural Information Processing Systems, NeurIPS 2019
25. Shamout, F.E., et al.: An artificial intelligence system for predicting the deterioration of COVID-19 patients in the emergency department. arXiv preprint (2020), https://arxiv.org/abs/2008.01774

26. Shen, M.L., Xiao, A., Yin, S.J., Wang, P., Lin, X.Q., Yu, C.B., He, G.H.: Associations between UGT2B7 polymorphisms and cancer susceptibility: A meta-analysis. Gene **706**, 115–123 (2019)
27. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)
28. Zhang, H., Cissé, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: 6th International Conference on Learning Representations, ICLR 2018
29. Zheng, S.E., Yao, Y., Dong, Y., Lin, F., Zhao, H., Shen, Z., Sun, Y.J., Tang, L.N.: Down-regulation of ribosomal protein L7A in human osteosarcoma. J Cancer Res Clin Oncol **135**(8), 1025–1031 (2009)