Artificial Intelligence based nodal metastasis prediction

Authors: Fahad Shabbir Ahmed, MD[1]; Furqan Bin Irfan MD, PhD[2]
[1]Wayne State University, Detroit Michigan, USA; [2] College of Osteopathic Medicine, Michigan State University, Lansing, Michigan
Abstract: 1986

Background:

Nodal metastatic spread of cancer is a sign of late progression of the disease process. The use of artificial intelligence in clinical sciences has shown great promise in accurately predicting difficult-to-predict clinical outcomes. In our current study, we hypothesize that RNA transcription data can be used as a biomarker to predict nodal metastasis in a number of different cancers. Machine learning has been used in other clinical settings to predict clinical outcomes[1, 2].

Methods:

The Cancer Genome Atlas (TCGA) database was utilized to identify differentially expressed genes (DEGs) and corresponding clinicopathological characteristics for all types of cancers. Replicate cases were generated for these experiments. In two experiments we used: 1) complete gene expression data and 2) 199 selected genes involved in multiple cancer pathways to predict nodal metastasis All data were downloaded from the TCGA, compiled, coupled using key columns, next we developed a deep-learning algorithm that would predict nodal metastasis. For the data we applied a split of 80/20 for training and test sets, we used SMOTE to fix the uneven distribution of the outcome in the dataset. The accuracy of the models was assessed by measuring sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the receiver-operator curve (AUROC).

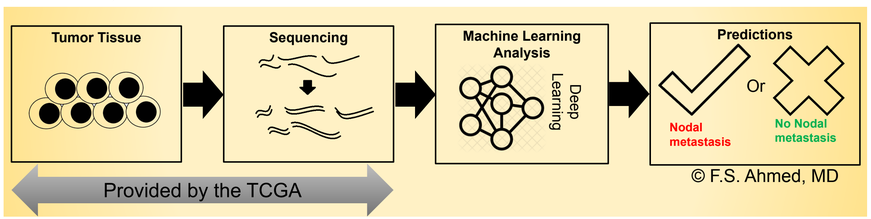

Figure 1. Machine learning analysis

Author contact:
Fahad S. Ahmed, MD, fahadshabbirahmed@gmail.com, fahadahmed@wayne.edu

The targeted gene *deep learning (199 genes) model shows better performance than the complete gene expression model.* Deep learning can predict nodal metastasis with good accuracy.

Future Directions for Research:

- Further validation of these finding on an independent external dataset.
- Evaluation of Quartile cutoff for low vs high expression in training the future algorithms.
- Exploration of the tumor microenvironment regarding the immune-landscape in the pan-cancer model between nodal and non-nodal metastasis tumors.
- Exploration of new algorithm for individual cancers types and evaluation of there immune-landscape and tumor microenvironment.

Results:

A total of 5507 RNASeq samples from 2869 patient data were analyzed in the TCGA database. Patient data was used after alignment for the development of a machine learning model. The model was a constructed and trained multi-layer deep-learning network. The complete gene expression profile produced a sensitivity of 59.3%, specificity of 64.9%, PPV of 63.9%, NPV of 60.3%, and an AUROC of 0.97(95%CI 0.03-0.05). While the selected gene expression panel showed a sensitivity of 92.6%, specificity of 92.7%, PPV of 92.5% and NPV of 92.8%, and AUROC of 0.98 (95%CI 0.05-0.09).

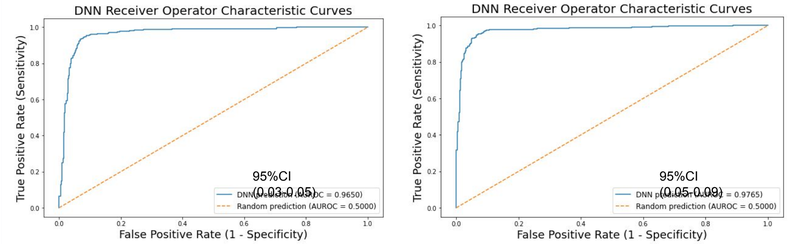

Figure 2. Area Under the Receiver Operating curves for the machine learning analysis

Experiments	Test-set accuracy	Training-set accuracy	Sensitivity	Specificity	Positive Predictive Values	Negative Predictive Values	False positive rate
Pan-Cancer, Pan-transcriptomics	93.02	98.70	59.26	64.94	63.90	60.34	35.07
Pan-Cancer 199 genes transcriptomics	92.67	96.24	92.63	92.70	92.49	92.84	7.30

Table 1. Machine learning analysis

References:

1. Ahmed, F.S., Ali, L., Raza-Ul-Mustafa *et al.* A hybrid machine learning framework to predict mortality in paralytic ileus patients using electronic health records (EHRs). *J Ambient Intell Human Comput* **12,** 3283–3293 (2021). https://doi.org/10.1007/s12652-020-02456-3
2. Ahmed, F.S., Ali, L., Joseph, B.A., MD, Ikram, A., Ul Mustafa, R., Bukhari, S.A.C. A statistically rigorous deep neural network approach to predict mortality in trauma patients admitted to the intensive care unit, Journal of Trauma and Acute Care Surgery: October 2020 - Volume 89 - Issue 4 - p 736-742 doi: 10.1097/TA.0000000000002888