Improving Early Cancer Detection Based on Liquid Biopsy Using Machine Learning

Mansour Abou Shaar, Karl Al Skaff, Karl Deek, Farid Eid El Beyrouthy, Joe Wakim, Frederic Zein

Department of Electrical and Computer Engineering, MSFEA AUB

mma239, kga09, ked03, fne07, jgw02, fhz01@mail.aub.edu

I. Project Proposal

Cancer cases are widespread around the world (Torre et al., 2015). Millions of deaths are recorded to be caused by various types of cancer each year (Chen et al., 2016) and yet, the numbers are projected to increase in all countries, even developed countries (Rahib et al., 2014). Early detection of cancer cases is a critical key to reducing these numbers. Effectively, the majority of identified cancers can be cured by single surgery when detected in their early stages, without any therapy [4]. Consequently, it is crucial to diagnose any cancer case promptly. However, traditional detection methods are costly and occurs usually after a patient suffers symptoms, thus detecting at a late stage as explained in [5]. Therefore, one of the major goals in cancer research is cancer detection before reaching an advanced stage like metastasis. Cancer detection based on blood tests could be a breakthrough, fast, and efficient solution for the detection of cancer in the early stages. Effectively, this non-evasive method could increase the testing rates as it is relatively easy and simple from the patient side.

In this context, Liu et al. (2021) conducted a survey to offer an overview of machine learning concepts and their applications in the context of early-stage cancer detection based on liquid biopsy (mainly blood). The authors also presented code templates for the multiple approaches. Machine learning algorithms are presented as an important tool by analyzing and identifying regularities in data and predicting based on training. As explained in [10], for machine learning, cancer detection is observed as a supervised problem, called a classification task. The research concludes that simple machine learning algorithms like linear regression models can result in a quality liquid biopsy-based diagnosis for multiple types of common cancer types.

Recently, a team of researchers has proposed a brand-new test that can, using only blood samples, screen multiple cancer types. The team presented in [6] CancerSEEK, a test that at once inspect the existence of 8 different types of cancer, including liver, lung, breast cancers, which are responsible for approximately 60% of cancer deaths in the US (Foster 2018). In this research, scientists took blood samples from thousands of people having one of the eight different cancer types studied. It was found that CancerSEEK was able to identify consistently the tumors in 70% of the cases. The success rate was different among cancer types: 98% in people with ovarian tumors to only 33% in people with breast cancer [7]. In order to investigate whether the CancerSEEK test can help identify the cancer type, the team used **supervised machine learning** to predict the cancer type of patients with positive CancerSEEK tests. From another perspective, the work done in [8] centered around a binary detector (cancer/normal) by considering eight input features (protein markers) in addition to the OmegaScore (DNA mutation score). The second part of the research emphasized determining the tissue targeted by cancer (i.e., Ovary, Breast, Pancreas, etc.). While logical regression and random forest were implemented in [6] (to determine the presence of cancer and classify it), the authors in [8] selected multiple multiclass supervised learning algorithms like AODE, deep learning, decision tree, and NB (naïve Bayes). After selecting the parameter settings, the work presented better results than in [6], primarily because the authors opted for generative modeling rather than discriminative. Furthermore, Kourou et al. (2014) discuss the accurate predictive performance of SVMs (Support Vector Machines). In fact, it constitutes a more recent approach in cancer prediction/prognosis, being widely due to its accurate predictive performance. Mainly, the authors in [11] argue that the choice of the best algorithm depends on multiple parameters including the type of data collected, the size of the data samples, and finally, the wanted type of prediction outcome wanted.

Other studies have been conducted but specific to one type of cancer. For example, the work presented in [9] aimed to predict ovarian cancer based on blood markers including circular tumor cells (CTCs). 156 ovarian cancer patients participated in the study randomly and were split between two cohorts (training and validating). This study involved 8 machine learning algorithms (classifier), including Random Forest (RF), Support Vector Machine, Gradient Boosting Machine, Conditional RF, Neural Network, Naive Bayes, Elastic Net, and Logistic Regression (Ma et al., 2021) that considered 11 blood parameters as features. Additionally, Goryński et al. (2014) targeted lung cancer early detection using an Artificial Neural Network based on multiple proteins markers present in the blood. Moreover, a sensibility analysis was conducted to determine the usefulness of the variable considered in the prediction. Adopting discriminative modeling, the prediction accuracy was very promising (determined 47 out of 48 cases) as the authors were targeting a single type of cancer.

In this context, our research and project will focus on predicting if a person has cancer as a first step and identifying the type of cancer by training an AI model based on available datasets as a second step. Additionally, we will try to improve the prediction accuracy regarding the various cancer type studied in [6]. The markers' concentration will constitute the main features used to evaluate the presence of cancer in addition to other information like Sex, and finally predicting its type. As features identification has been tackled in previous research, the main features used in the models are already known, thus solving a major challenge. We will be trying to improve the prediction score of the cancer type especially since it recorded a very low prediction score for some types of cancers in [6], by selecting a generative modeling algorithm rather than a discriminative one, improving the results of [6] while determining the type of cancer (classification issue). We will be working on a public data set stored as a Microsoft Excel Worksheet that includes the concentration of multiple proteins markers and other features like Sex, counting up to 31 features for around 1800 patients, stating their health status (healthy/cancer), and the respective cancer type for cancer patients as shown in the below figure.

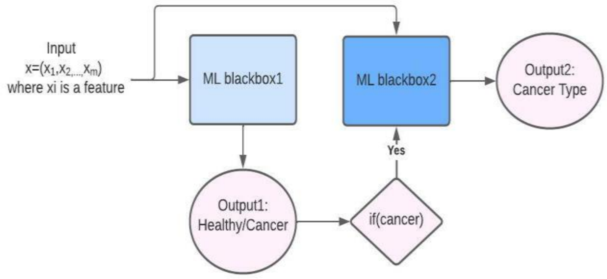

Fig. 1. **Diagram illustrating the solution to the problem proposed**

APPENDIX

Members' participation	
Name	Surveyed Paper
Mansour Abou Shaar	[11] Machine learning applications in cancer prognosis and prediction.
Karl Al Skaff	[10] Machine Learning Protocols in Early Cancer Detection Based on Liquid Biopsy: A Survey.
Karl Deek	[12] Artificial neural networks approach to early lung cancer detection.
Farid Eid El Beyrouthy	[6] Detection and localization of surgically resectable cancers with a multi-analyte blood test.
Joe Wakim	[8] Cancer Detection from Multianalyte Blood Test Results
Frederic Zein	[9] Artificial Intelligence Based on Blood Biomarkers Including CTCs Predicts Outcomes in Epithelial Ovarian Cancer: A Prospective Study.

REFERENCES

[1] Torre L.A., Bray F., Siegel R.L., Ferlay J., Lortet-Tieulent J., Jemal A. Global cancer statistics, 2012. *CA Cancer J. Clin.* 2015;65:87–108.

[2] Chen W., Zheng R., Baade P.D., Zhang S., Zeng H., Bray F., Jemal A., Yu X.Q., He J. Cancer statistics in China, 2015. *CA Cancer J. Clin.* 2016;66:115–132

[3] Rahib, L., Smith, B. D., Aizenberg, R., Rosenzweig, A. B., Fleshman, J. M., and Matrisian, L. M. (2014). Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. Cancer research, 74(11):2913-2921.

[4] Semrad TJ, Fahrni AR, Gong IY, Khatri VP. *Ann Surg Oncol.* 2015;22(suppl 3):S855–S862.

[5] Al-Azri, Mohammed H. "Delay in Cancer Diagnosis: Causes and Possible Solutions." *Oman medical journal* vol. 31,5 (2016): 325-6. doi:10.5001/omj.2016.65

[6] Cohen, J. D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., Douville, C., Javed, A. A., Wong, F., Mattox, A., Hruban, R. H., Wolfgang, C. L., Goggins, M. G., Dal Molin, M., Wang, T. L., Roden, R., Klein, A. P., Ptak, J., Dobbyn, L., Schaefer, J., … Papadopoulos, N. (2018). Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science (New York, N.Y.)*, *359*(6378), 926–930. https://doi.org/10.1126/science.aar3247

[7] Forster, V. (2018, January 19). *A new $500 blood test could detect cancer before symptoms develop*. Forbes. Retrieved February 28, 2022, from https://www.forbes.com/sites/victoriaforster/2018/01/18/a-new-500-blood-test-could-detect-cancer-before-symptoms-develop/?sh=7ec1ecca7dd4

[8] Wong, K. C., Chen, J., Zhang, J., Lin, J., Yan, S., Zhang, S., Li, X., Liang, C., Peng, C., Lin, Q., Kwong, S., & Yu, J. (2019). Early Cancer Detection from Multianalyte Blood Test Results. *iScience*, *15*, 332–341. https://doi.org/10.1016/j.isci.2019.04.035

[9] Ma, J., Yang, J., Jin, Y., Cheng, S., Huang, S., Zhang, N., & Wang, Y. (2021). Artificial Intelligence Based on Blood Biomarkers Including CTCs Predicts Outcomes in Epithelial Ovarian Cancer: A Prospective Study. *OncoTargets and therapy*, *14*, 3267–3280. https://doi.org/10.2147/OTT.S307546

[10] Liu, L., Chen, X., Petinrin, O. O., Zhang, W., Rahaman, S., Tang, Z. R., & Wong, K. C. (2021). Machine Learning Protocols in Early Cancer Detection Based on Liquid Biopsy: A Survey. *Life (Basel, Switzerland)*, *11*(7), 638. https://doi.org/10.3390/life11070638

[11] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2014). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, *13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

[12] Goryński, K., Safian, I., Grądzki, W., Marszałł, M., Krysiński, J., Goryński, S., Bitner, A., Romaszko, J. & Buciński, A. (2014). Artificial neural networks approach to early lung cancer detection. *Open Medicine*, *9*(5), 632-641. https://doi.org/10.2478/s11536-013-0327-6