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

***Abstract*—Cancer is the most common cause of death and kills more than 600,000 people in the US annually. Most deaths are caused by late cancer detection as cancer in late stages is incurable. Early cancer detection is crucial to reducing death rates by treating patients in their early stages. Unfortunately, most patients diagnose cancer late after experiencing related symptoms as most cancer-detecting techniques are invasive, slow, and costly, discouraging most people to test routinely for cancer. In this context, fast, cheap, and non-invasive detection techniques should be developed to address this dangerous challenge. We propose in this paper a cancer detection technique based on a blood test using Machine Learning. The detection technique, centered around measuring the concentration of multiple protein biomarkers in blood, can predict cancer presence with very low false-negative rates and its type (Liver, Lung, etc.) with the maximum achievable precision. We present in this paper and discuss two proposed supervised ML prototypes. Moreover, the prototypes will be tested using different classification ML models as predicting cancer presence (yes/no) and cancer type are both classification problems. Finally, we will be choosing the best prototype along with the classification models that will be used based on a set of previously defined metrics.**

Keywords—Cancer detection, Machine Learning, Classification models.

I. Introduction

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 for 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. Previous work has been done in this context in [6] scoring a low detection rate (60%) which should be by far higher.

II. Contribution

In this context, our research and project will focus on predicting cancer presence and type-based multiple protein biomarkers' concentration in blood, and other information like Sex. We will improve the results of previous blood-based cancer tests by increasing the detection rate of cancer while improving the precision of predicting its type. Previous work offered a low detection rate for a certain type of cancer which should be improved rigorously. Effectively, predicting that someone is cancer-free while being cancer positive is a dangerous issue. Solving this problem would be a revolutionary approach to testing for cancer as it would provide a fast, reliable, and accurate method to diagnose cancer. Our contribution emphasizes predicting most cancer cases while predicting the type accurately. It is important to note that we aim to increase the recall score as we wish to diagnose all cancer cases, but this will be at the cost of precision which will decrease. The main challenge resides in balancing between recall and precision to be able to diagnose all cancer cases while maintaining a low false-positive rate that will push patients to undergo invasive, costly, and slow traditional cancer detection methods. However, in this balancing problem, we should always be biased to increase the recall function over precision to the simple fact that predicting a false positive (a result of increased increasing recall) is relatively better compared to predicting a false negative (a result of increase precision).

III. Dataset and Data Analysis

The dataset was obtained from a published paper on the National Library of Medicine (Cohen et al., 2018) under Associated Data rubric, stored in Excel type. As a first step, we investigated a dataset stored in EXCEL that contained multiple sheets. We were mostly interested in Sheet #6 that states the concentration of around 40 proteins in the blood with the cancer diagnostic result with the type of cancer for non-healthy patients. The dataset was complete without redundancies with around 1800 rows, but the concentration of proteins (float type) needed cleaning as it contained non-digits characters. Moreover, scaling the data was an important step in the process as some protein concentrations were in the order of 10^4 while others were in the order of 10^{-1}.

The dataset was observed to be not large enough for training classification models, but it was observed that after training models with a 70%-30% train-test split, the models were able to generalize well on the unseen data, but we had some problems with increasing the precision of cancer type prediction as we had only 200 breast cancer cases, 300 colorectum cases, etc.

Before tackling solving the problem, some data analysis was made to offer a general view of the dataset and to what extent we can extract information from it. 2D graphs presenting the concentration of a specific protein versus another one were plotted for healthy and cancerous patients to discover some patterns shown in Figure 1:

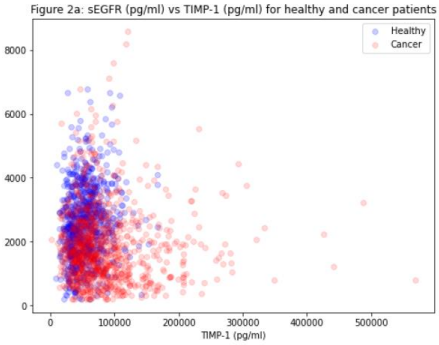

Figure 1. sEGFR concentration vs TIMP-1 concentration in healthy and cancerous patients

As shown in the plot above, people with high sEGFR concentration are more likely to have cancer while people with high TIMP-1 concentration are more likely to be healthy. After getting some insight from multiple plots, we calculated the correlation factor between the concentration of all protein biomarkers being healthy or not. We noticed the importance of some proteins like OPN, Prolactin, TIMP-1, and sEGFR. For example, sEGFR has a high correlation factor (-0.3) relative to CD44 (-0.014) which reflects the importance of some proteins over others.

As the aspect of the problem is related to classification, we made some clustering analyses. KNN clustering was used to cluster the data into multiple clusters to identify the ability to classify our data in our future steps. As a first step, we decided to cluster the data into two clusters. It was clear that most healthy patients were present in the first cluster which supposes that classifying healthy/cancer will not be challenging as seen in Figure 2. Moreover, the majority in cluster 2 are cancerous patients. We should note that 9 identify healthy patients while 1→8 denote multiple cancers by type. As numerous cancer patients are with healthy patients in the same cluster, a decreased threshold will surely shift the majority of cancer patients in the second cluster when classifying in the following sections.

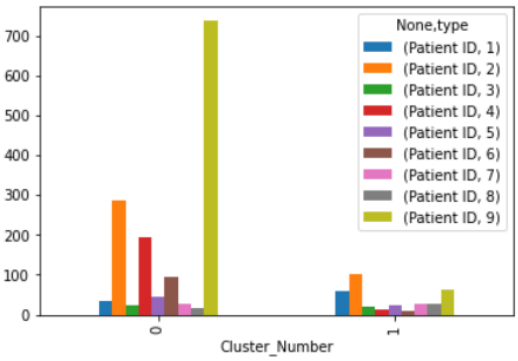

Figure 2. Clustering the data into two clusters

Finally, the classification potential can be observed after operating PCA reduction which allowed us to observe how cancer types (healthy included) were clustered based on the first and second components after reduction.

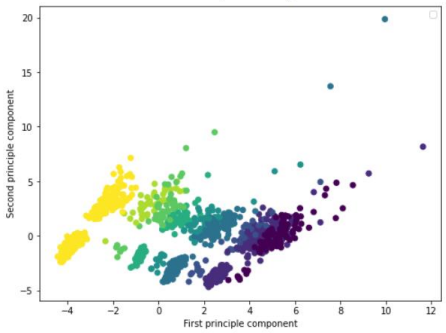

Figure 3. PCA plot

The yellow color represents the healthy type while all other distinguished colors represent the other different types based on 2D components of PCA. The work done after clustering and PCA reduction and plotting helped us assess to what extent classification is possible in the following sections as we will be classifying cancer presence along with cancer type.

IV. Proposed Solution

i. Prototype 1

In this paper, we are proposing two prototypes to predict cancer presence and type based on classification supervised machine learning models. In our first prototype, the prediction is done at two levels as depicted in Figure 3:

1. Predict cancer presence (binary yes/no)
2. Predict cancer type (multi-output value)

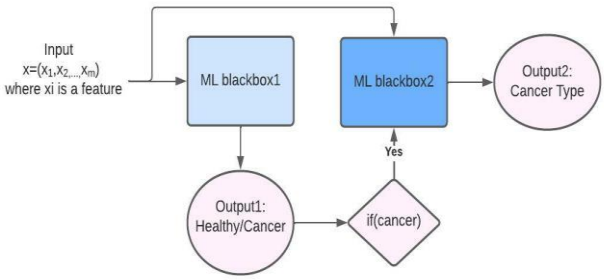

Figure 4. Diagram illustrating prototype 1

In the first level of the prototype, a binary classification algorithm will be implemented to detect cancer presence. At this stage, it is important to obtain a high recall score for cancer detection because we should be predicting most cancer cases. Accuracy is not our primary concern as we could be predicting that no one is cancerous and still score 95% accuracy (around 5% of the US population has cancer). We can think to reduce the prediction threshold as we could obtain higher recall but at the cost of precision. The best model at this stage should have a very high recall while maintaining an acceptable precision. However, if we were to set threshold = 0 obtaining a maximum recall (=1), everyone will be diagnosed with cancer and we would be obtaining a very low precision score which is inefficient. In the second level, we will predict cancer type for all people who were diagnosed with cancer. In this step, the most important metric is precision because we don't have 'the fear' of obtaining false-negative results. The classification models will be multi-output given the type of cancer (1 or 0) for each type resulting in an array of length 8 containing only one 1.

The first level of the prototype is based on binary classification. We used multiple classification algorithms like Logistic Regression, SVM, Neural Networks, and Random Forest Classifier. The metrics obtained for each classifier model training and tested were as follows:

Table 1. Recall and Precision for cancer presence detection

Model	LG	SVM (poly)	SVM (rbf)	Neural Network (custom)	MLP	RF
Recall	0.83	0.44	0.86	0.94	0.91	0.97
Precision	0.92	0.99	0.92	0.90	0.88	0.98

We can observe that LG and SVM using RBF as kernel offered low recall while maintaining a high precision rate. Moreover, SVM using polynomial kernel offered extremely low recall but with a perfect precision, which suggests that all predictions made by SVM (poly) were made with high confidence. In our case, we are ready to sacrifice some precision to increase the recall, which is observed in the Neural Network Models. Finally, the best model in our case is the RF classifier as it offers a very high recall and precision at the same time. In terms of increasing the recall for some classifiers by decreasing the threshold, it is not possible to obtain a high recall while having the same precision as for the Random Forest classifier. In fact, we plotted the graph showing Precision-Recall levels versus the threshold for each classifier in the following figures.

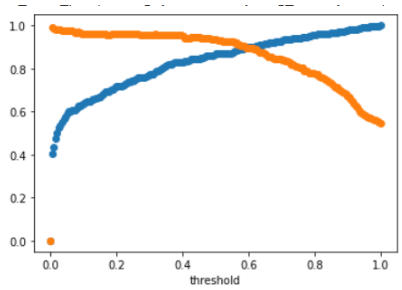

Figure 5. Precision (blue) and Recall (orange) versus threshold – Logistic Regression

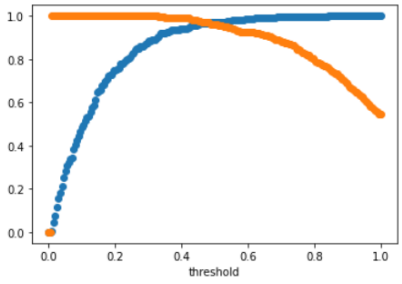

Figure 6. Precision and Recall versus threshold – RF

By comparing the two graphs for LG and Random Forest, we can observe that the stability region of RF is very narrow. If we want to increase recall by 1% (go left, reduce threshold), the precision will decrease abruptly compared to Logistic Regression or other models.

We investigated ways to improve SVM naturally (so that SVM 'learns' more - without reducing threshold) like Grid Search optimization to find the best parameters, setting the score target equal to recall and not accuracy. However, after finding the best parameters, the recall didn't improve but scored even lower results. The only explanation possible is that the best parameters found were pushing the model to overfit the training data even after using k-fold validation, scoring high recall over training data, and reducing the ability of the model to generalize over unseen data. In this context, to avoid the Neural Network built to overfit the training data, we implemented an early stop callback to monitor the validation loss, stopping the training when the val_loss is stagnant and not decreasing even after an increase number of epochs.

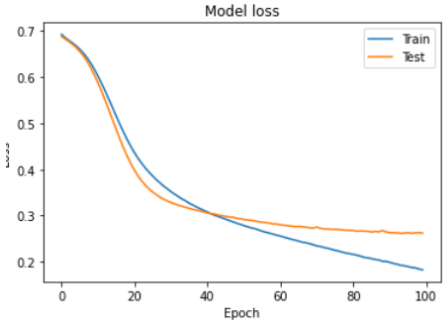

Figure 7. Testing & Validation loss vs Epochs – Neural Network Classifier

After predicting the cancer presence using the previous classifiers, **the second level** of prototype 1 is based on predicting cancer type. The same models will be used with multi-outputs. As logistic regression doesn't support directly multi-outputs, we used OneVsRest() classifier that will split the multi-output problem into multiple binary classification problems.

Cancer type	Code
Esophagus	1
Liver	2
Lung	3
Ovary	4
Pancreas	5
Stomach	6
Colorectum	7
Breast	8

Figure 8. Cancer types with their respective codes

After training LG, we were able to obtain acceptable results in terms of precision. The model was able to predict Liver and Esophagus most precisely with a respective precision of 0.90 and 0.83. It was interesting to note that most people who were falsely predicted as cancer positive (they are not cancerous) were mostly diagnosed with Breast Cancer, even males as shown in the following figure.

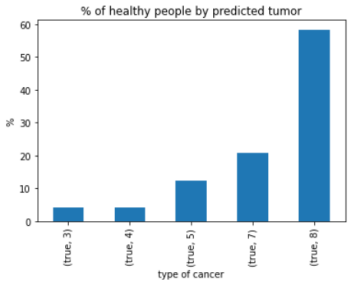

Figure 9. Percentage of healthy people by predicted tumor

Other models were used as SVM, Neural Networks (MLP), and Random Forest. They offered different precisions for different cancer types. For example, Breast cancer was best predicted in terms of precision by SVM (linear), while Random Forest was the best in predicting Liver cancer. Each classifier would handle predicting a type of cancer differently based on the shape of boundaries drawn among classes. It is interesting to note that SVM is yielding good results relatively to RF, LG, and NN at this stage (level 2) compared to the mediocre results in binary classification (level 1 prediction). Moreover, as RF offered excellent results in the previous stage, it offers good results for some cancer types, but it cannot detect Esophagus

cancer, scoring a precision of 0. The following tables summarize the type of cancer in terms of the most precise classifier.

Table 2. Best Classifier for each cancer type

Type	Esophagus	Liver	Lung	Ovary	Pancreas	Stomach	Colorectum	Breast
Best classifier	LG	RF	LG	RF	RF	RF	SVM	SVM
Precision	**0.83**	**1.00**	**0.72**	**0.83**	**0.72**	**1.00**	**0.63**	**0.55**

Prototype 1, as discussed, is composed of two levels or stages. The first stage that emphasizes on increasing the recall metric will employ RF for cancer detection as it scores incredibly high recall for also high precision. The second stage that places an emphasis on the precision metric will use different classifier models. As RF cannot detect Esophagus cancer and detect Liver cancer with a precision of 1.00 (after predicting cancer presence, don't forget we are in stage 2), we thought to implement an ensemble method formed by the classifiers employed above. The voting weight of each classifier is equal to the average precision of the cancer type voted regarding this classifier. Consider RF predicts Liver and SVM predict Breast for a given patient. In this context, RF's weight (precision of RF regarding liver = 1) is higher than SVM's weight for Breast (precision = 0.55) which results in predicting LIVER as cancer type following this ensemble method.

ii. **Prototype 2** (*Alternative Design*)

After proposing prototype 1 in the previous section, we will present prototype 2 which is based on a more direct method. Effectively, prototype 2 aims to predict cancer presence and type in one prediction stage or round by using a multi-output classifier. The output will range from 1 to 9 using the previous cancer type–code pairs in addition to 9 representing a healthy condition. The prediction is done using one stage as depicted in the following flow chart:

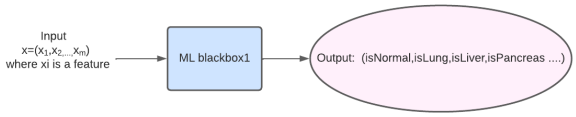

Figure 10. Fig. 2. Diagram illustrating the alternative design proposed

We will be using the same classifier models as used in stage 2 of prototype 1 aiming to detect cancer presence and cancer type. The important metrics here to take into consideration are the recall and precision at the same time, as we wish to predict most cancer cases and predict the type accurately. Assessing the F1 score should be a great idea as it reflects the combination of precision and recall at the same time. However, we should always place an emphasis on an overall high recall rate for cancer detection even at the cost of some precision. After training LG, the classification model offered a very low recall score (0.74) and average precision for predicting cancer type for each type. It is interesting to note that most healthy patients that were diagnosed with a cancer type were diagnosed with Colorectum, in contrast with prototype 1, where most healthy patients were diagnosed with Breast cancer. In stage 2 of prototype 1, SVM offered good results with a multi-output classification which was reflected when used in prototype 2. SVM (linear) offered good precision results on average, being able to detect all cancer types but suffered from a low recall. As expected from the previous testing, the highest recall score obtained was by training RF with 0.89 but wasn't able to predict Esophagus cancer as in stage 2 of prototype 1. All classifiers are not able to score high recall rates as they are "overwhelmed" by maintaining at the same time a relatively high precision for both cancer detection (yes/no) and cancer type. The following table summarizes the cancer types in the function of the best classifiers to use based on the f1 score, as we wish to detect a specific cancer type and detect most cases of it.

Type	Esophagus	Liver	Lung	Ovary	Pancreas	Stomach	Colorectum	Breast
Best classifier	MLP	SVM (poly)	LG	RF	RF	SVM (linear)	RF	RF
f1-score	0.19	0.55	0.57	0.62	0.74	0.22	0.67	0.47

iii. Prototypes comparison

After dressing up the prototypes' designs and results, we decided to select the best prototype for best use. It is clear that prototype 1 is by far the best in terms of recall score, thus detecting cancer presence, which is the most crucial goal. Although prototype 2 offered better results in terms of precision of the type of cancer, predicting cancer type is of secondary importance, yet important but not of the same degree as cancer detection. Prototype 1 suffers from lower precision in the second stage as healthy patients diagnosed as cancerous are obliged to be part of a cancer type group, thus lowering precision. Moreover, the recall score is lower for prototype 2 which will predict cancer for most confident cases, which will make it easier to diagnose patients in terms of cancer type, increasing precision.

In this context, it is important to note that prototype 1, elected as the best prototype regarding our targets, outperformed many blood testing techniques discussed in the introduction of the paper, especially the CANCERSEEK method proposed in [6], scoring a recall score of 62%, 35 points way from prototype 1 in term of recall score. CANCERSEEK detected Breast cancer with 33% in terms of precision for cancerous patients, in contrast with our prototype which scored 55% in terms of precision.

V. Conclusion

In this paper, we discussed the main problem to solve which is to offer an easy, fast, and accurate method to detect cancer to encourage people to test for cancer routinely as traditional methods are invasive, costly, and slow. We proposed two prototypes based on machine learning classification models that were compared using predefined metrics, mostly recall, precision, and f1-score. Prototype 1 was mainly targeting the main goals better than Prototype 2 and was elected as our final design with known experimental results. Compared to the literature, our machine learning design was able to outperform CANCERSEEK, an established blood testing method to diagnose cancer. The main gap in our project was that we didn't assess the importance of each protein (biomarkers concentration in blood) in detecting cancer presence and predicting its type aiming to reduce the cost induced by the blood test as it requires testing for around 39 proteins in the blood. In future work, this important idea should be tackled to address reducing the cost of this type of test, as its performance was assessed in this paper.

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