Machine Learning Methods for Gastric Cancer Diagnosis

Thomas Seli

*Quantitative and Computational Biology Department*

*Dornsife College of Letters Arts and Sciences*

*University of Southern California* seli@usc.edu

# Abstract

In this report, I will explore various machine-learning techniques that I used to create effective models for gastric cancer detection. This study is based on a research article published by *Nature Communications* in February 2024 titled ‘Metabolomic machine learning predictor for diagnosis and prognosis of gastric cancer’. This paper explores the use of machine learning to diagnose gastric cancer based on data collected from metabolomic analysis of 702 patient plasma samples, where 389 patients had gastric cancer and the remaining 313 did not have gastric cancer. The researchers developed a random forest model trained on 10 key metabolites that were identified and it was found that the 10-DM model outperformed conventional methods for gastric cancer diagnosis. My intention with this project was to take the metabolomic data provided by the authors and study different machine-learning techniques to replicate and improve the performance of their model. I focused primarily on the diagnostic model and did not attempt to work with the prognostic model they developed. By studying the behavior of various models, I was able to find ways to improve upon the baseline performance of the random forest classifier developed in the paper. Some methods used to improve baseline performance included data rescaling, PCA feature reduction, and modifying the size of the training dataset. Overall, the most effective models for diagnosing gastric cancer appear to be neural networks that leverage PCA.

## Introduction

Gastric cancer is a significant factor in global cancer mortality, thus it is important to develop early-diagnosis strategies to detect gastric cancer quickly and effectively. In fact, according to the National Cancer Institute, the 5-year relative survival rate of gastric cancer is only 36%. These numbers are shockingly low considering that other common types of cancer, such as breast cancer, have 5-year survival rates of up to 91.2%. This is likely because gastric cancer is often already advanced when it is diagnosed and thus can be treated but rarely cured at that late stage. In addition, the most commonly used method for diagnosing gastric cancer is endoscopic examination, which is both invasive and costly (Chen et al. 1). Therefore, my objective is to enhance the diagnostic model for gastric cancer discussed in this paper and explore even better-performing methods of gastric cancer diagnosis, aiming to contribute to this field of research. For reference, when replicating the random forest model from the research paper, I achieved 91.5% accuracy on the main testing set and 85.2% accuracy for the external testing set. The performance metrics are visualized in Figure 1 below. I also wanted to analyze the distribution of these datasets, so I created histograms for each dataset to visualize the balance of the labels. The results can be found in Figure 2 below. Personally, I was surprised when seeing these charts as I thought that the data could have been balanced more evenly across the datasets. This led me to hypothesize that perhaps I could incorporate more training examples of gastric cancer to get better performance from the models. I believe that by leveraging different methods of feature selection, parameter hypertuning, data splitting, and deep learning we can further improve upon the diagnostic model discussed in the paper.

A group of graphs showing different colors

Description automatically generated with medium confidence

***Figure 1:*** *Baseline Performance of 10-DM Diagnostic Model*

A group of bar charts

Description automatically generated with medium confidence

***Figure 2:*** *Dataset Label Distributions*

## Methods

The researchers collected data related to 147 metabolites from the patient samples and were able to identify the 10 most important metabolites for gastric cancer diagnosis. To conduct a preliminary exploratory analysis of these ten metabolites, I generated a correlation heatmap to examine the relationships between their measurements. As illustrated in Figure 3, the majority of these metabolites exhibit weak correlations with each other. This suggests that the variances among these features are largely independent, which is a good indicator of the robustness of the diagnostic models. However, I also wanted to experiment with different methods of feature selection so I also used PCA to reduce the dimensionality of the data and attempt to improve the performance of the models. As seen in Figure 4 below, PCA and t-SNE do start to create some clusters of healthy and non-healthy samples which was another reassuring sign that the diagnostic models would perform well.

A screenshot of a computer

Description automatically generated A group of red and blue dots

Description automatically generated

***Figure 3:*** *Correlation Matrix of Top 10 Metabolites* ***Figure 4:*** *PCA and t-SNE Scatterplots*

After performing this preliminary analysis and getting baseline performance metrics from the 10-DM model in the paper, I began to experiment with different techniques to attempt to improve performance and see how different models fit the data. The traditional models that I experimented with included random forests, logistic regressions, support vector machines, k-nearest neighbors, and gradient boosting. In addition to using the top ten metabolites from the research I also used methods such as rescaling of data, grid search for parameter tuning, PCA for feature selection, and adjustments of class weights. Interestingly, each model seemed to behave differently when using these methods. For instance, random forests seem to perform quite well on the top ten metabolites as seen in the research paper itself, but did not respond to PCA well. On the other hand, I found that support vector machines do not perform well on the ten metabolites, but saw a great boost in performance when using PCA and some parameter tuning. Overall, I found that random forests, logistic regressions, and support vector machines seemed to be the best models for gastric cancer diagnosis. In fact, some of these models already outperformed the original 10-DM model in certain aspects. For example, logistic regression with just 19 principle components was 93.7% accurate on the external testing dataset in comparison to the 85.2% accuracy from the baseline model. While I was able to get some impressive results from some of these models I still wanted to find some way to get clearly superior performance across the board, as a boost in performance for one testing set usually came at the cost of performance on the other testing set. So, I opted to approach this from a deep learning angle as well.

First, I created a separate notebook called ‘Neural\_Network\_Parameter\_Tuning.ipynb’ to help me find the optimal structure and parameters for the neural networks. By looping through different numbers of neurons at each subsequent layer, I was able to efficiently build deep learning models with solid performance. I began by building neural networks on the original training data and seeing the performance on both testing datasets to try and strike the best balance possible. However, using neural networks on the ten metabolites was not giving me the performance I was looking for so I opted to rescale the data and use PCA for the feature selection similar to the more traditional models I experimented with earlier. This changed the game, as I was able to get 4% more accuracy than the baseline performance on both testing sets. While doing this, I was curious how the deep learning models would perform if I were to increase the size of the training data by merging it with one of the testing datasets and evaluating the performance on the primary testing data. To balance the loss of the external dataset I doubled the validation split for the neural network training from 10% to 20% for more robust validation scores for the new training dataset.

Using these techniques, I successfully enhanced both the accuracy and AUC scores beyond the initial baseline metrics. However, accuracy alone does not provide a complete assessment of the diagnostic models’ effectiveness. As previously mentioned, it is crucial to develop methods that not only diagnose gastric cancer with high accuracy but also detect the disease at its earliest stages. So, using the figures in the paper as a reference, I developed a function that calculates the accuracy of the model at these early stages using the predicted probabilities from the chosen model. At this stage, it was important to keep track of the type of training data used for each model as the PCA models need different data for performance evaluation than the models trained on the ten metabolites.

## Results and Discussion

Now, that the methods have been discussed it is time to get into the results. In Table 1 below, we can see the results of different models that I developed using the various methods discussed in the previous section. These results present some interesting insights into the behavior of the various models and highlight why accuracy is not always the best metric for model validation. For instance, looking at the first two rows we can see that I was able to tune the parameters of the original random forest and get slightly better performance in terms of sheer accuracy. However, when looking at the effectiveness of the model for detecting gastric cancer at the early stages it actually performed worse overall. This introduces an interesting discussion of what metrics we want to consider as the most important. When looking at the more traditional machine-learning models in the table we can see that some models certainly perform better than others in certain areas, but I was actually unable to find a model that performs better than the baseline across the board. Despite this, it is interesting to see how different models perform better at different tasks. We can see that the PCA logistic regression performs great on the external dataset, but shows worse performance for detecting early stages of cancer across the board compared to the baseline. In general, we can see that a boost in performance on one testing dataset usually means a decrease in the performance on another dataset, and this is likely due to the differences in the label distributions in these two datasets, which we can refer back to Figure 2 and see. So, while there were certainly areas of improved performance for the traditional models it is difficult to say if any of these are ‘better’ than the baseline. However, this all changes when we move down the table and look at the deep learning models.

Firstly, we can see that the neural networks actually do not perform better then the random forests when they are trained on the top ten metabolites. Even when increasing the size of the training data, the neural network trained on the ten metabolites still does not match the performance of the baseline model which was surprising for me. Personally, I expected neural networks to simply perform better than traditional methods across the board, but this is not always the case. Additionally, we can see further evidence that sheer accuracy does not tell the whole story of the model performance when looking at the PCA neural network that was trained on the original training data. While this neural network is more accurate overall on both testing datasets then the baseline, it actually does not detect stage 1 and 2 gastric cancer as well. Nevertheless, the last two rows of the table show that when using a combination of all the methods discussed earlier, we can actually find a model that beats the 10-DM random forest. By training neural networks on the new training data and not doing any feature selection we can get significantly boosted performance across the board on the testing data. Now, one drawback of this model is that it is significantly more complex and computationally expensive to train than the other models, but I believe that the performance boost warrants this tradeoff. Finally, the model that performs the best across the board is the PCA neural network that is trained on the new training data. It reaches almost 96% accuracy on the primary testing data and is also great at detecting gastric cancer at the earliest stages. It also saw great performance on the validation splits at each training epoch reaching upwards of 93% accuracy to further validate the performance. We can see some visualizations of the model performance in Figure 5 to further support the robustness of the neural network. Furthermore, in Figure 6, I created a plot of the predicted probability given by the model against the first principal component of the data, which captures about 11% of the total variance. This figure illustrates just how much more confident the PCA neural network is in diagnosing gastric cancer compared to the baseline model. Now, one drawback of this model is that it is much less interpretable than the other models due to the PCA transformation applied to the features. However, I believe that the sheer increase in performance across the board overshadows this.

***Table 1:*** *Results Table*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Primary Test Accuracy** | **External Test Accuracy** | **Primary Test AUC Score** | **External Test AUC Score** | **Stage 1A Accuracy**  **(Primary Test)** | **Stage 1B Accuracy (Primary Test)** | **Stage 1 & 2 Accuracy (Primary Test)** | **Stage 1A Accuracy**  **(External Test)** | **Stage 1B Accuracy (External Test)** | **Stage 1 & 2 Accuracy (External Test)** |
| Baseline 10-DM random forest | 0.915 | 0.853 | 0.971 | 0.922 | 0.875 | 1.0 | 0.8 | 0.909 | 0.75 | 0.931 |
| Hypertuned random forest | 0.922 | 0.853 | 0.970 | 0.917 | 0.875 | 1.0 | 0.8 | 0.818 | 0.75 | 0.897 |
| Logistic regression with 19 principal components | 0.873 | 0.937 | 0.944 | 0.975 | 0.75 | 1.0 | 0.76 | 0.909 | 0.75 | 0.90 |
| SVM with 75 principal components | 0.937 | 0.821 | 0.975 | 0.930 | 0.9375 | 1.0 | 0.92 | 0.818 | 1.0 | 0.897 |
| Neural network with 60 principal components (original training data) | 0.922 | 0.895 | 0.970 | 0.944 | 0.8125 | 1.0 | 0.84 | 0.818 | 0.75 | 0.862 |
| Neural Network with top-ten metabolites (new training data) | 0.894 | N/A | 0.962 | N/A | 0.875 | 1.0 | 0.84 | N/A | N/A | N/A |
| Neural network with all 147 metabolites (new training data) | 0.958 | N/A | 0.982 | N/A | 0.9375 | 1.0 | 0.92 | N/A | N/A | N/A |
| Neural network with 60 principal components (new training data) | 0.958 | N/A | 0.988 | N/A | 1.0 | 1.0 | 0.96 | N/A | N/A | N/A |

A close-up of a graph

Description automatically generated

***Figure 5:*** *Performance of the PCA Neural Network*

*A graph of colored dots

Description automatically generated with medium confidence A diagram of a model performance prediction test

Description automatically generated*

***Figure 6:*** *Confidence of Baseline Model (Left) vs. PCA Neural Network (Right)*

# Conclusions

My objective with this report was to improve upon the original 10-DM diagnostic model discussed in the research paper. Improving the accuracy and AUC scores is obviously important, however I am most interested in how effective the model is at detecting gastric cancer at its early stages. Gastric cancer is a particularly deadly type of cancer because it is difficult to detect early with conventional methods and is often diagnosed too late. However, I believe that through my study of various machine-learning models, we have further evidence that non-invasive methods for gastric cancer detection can outperform traditional methods. Furthermore, other research is being done in this field including the use of deep learning methods on images for early gastric cancer detection (Gao et al. 2). The research being done in this field is showing great promise for the future. I believe that my results and more importantly the work being done by researchers around the globe show that the future of early gastric cancer detection will be based on deep learning methods, which could help to save thousands of lives.

# References

Chen, Yangzi, et al. “Metabolomic machine learning predictor for diagnosis and prognosis of gastric cancer.” *Nature Communications*, vol. 15, no. 1, 23 Feb. 2024, https://doi.org/10.1038/s41467-024-46043-y.

Gao, Zhihong, et al. “Development of a deep learning model for early gastric cancer diagnosis using preoperative computed tomography images.” *Frontiers in Oncology*, vol. 13, 6 Oct. 2023, https://doi.org/10.3389/fonc.2023.1265366.

“Stomach Cancer Survival Rates and Statistics.” *Stomach Cancer Survival Rates and Statistics - NCI*, 31 May 2023, www.cancer.gov/types/stomach/survival.