Machine Learning Prediction for 50 Anti-Cancer Food Molecules from 968 Anti-Cancer Drugs

Simiao Zhao[1], Xuanyue Mao[2], Hanghong Lin[3], Hao Yin[4], Peixuan Xu[5]

[1]University of Edinburgh, Edinburgh, UK
[2]Maranatha High School, Pasadena, CA, USA
[3]Shanghai Shangde Experimental School, Shanghai, China
[4]Tabor Academy, Marion, MA, USA
[5]Wellington College International Shanghai, Shanghai, China
Email: Zhaosimiao1@gmail.com, abudou100@gmail.com, Yw020416@gmail.com, hyin22@taboracademy.org, jenniferxu0330@icloud.com

How to cite this paper: Zhao, S.M., Mao, X.Y., Lin, H.H., Yin, H. and Xu, P.X. (2020) Machine Learning Prediction for 50 Anti-Cancer Food Molecules from 968 Anti-Cancer Drugs. *International Journal of Intelligence Science*, **10**, 1-8.
https://doi.org/10.4236/ijis.2020.101001

Received: November 27, 2019
Accepted: January 28, 2020
Published: January 31, 2020

Abstract

Cancer-beating molecules (CBMs) are abundant in many types of food and potentially anti-cancer therapeutic agents. In the previous work, researchers introduced a network-based machine learning platform to identify the cancer-beating molecules, for example, comparing the similarities in the molecular network between approved anticancer drug and food molecules. Herein, we aim to build on this work to enhance the accuracy of predicting food molecules. In this project, we improve supervised learning approaches by applying Soft Voting algorithm to seven machine learning algorithms: Support Vector Machine with Radial Basis Function (SVM with RBF kernel), multilayer perceptron neural network (MLP), Random forest, Decision trees, Gaussian Naive Bayes, Adaboosting, and Bagging. As a result, the accuracy in the dataset of 50 food molecules utilized increased from 82% to 87%, achieving a significant improvement in the precision of predicting anti-cancer molecules.

Keywords

Machine Learning, Food, Anti-Cancer, Optimization

1. Introduction

1.1. Background

Given the aging global population, individuals have to undertake a more substantial burden of unsustainable healthcare and significant expenses on medi-

cines. Fortunately, research has shown that some dietary food can potentially replace medications in those fatal cancers cases, with plant-based diets rich in cancer-beating molecules; such as polyphenols, flavonoids, terpenoids, and botanical polysaccharides. Not only do these foods contain high nutritional values, but also, they are affordable and highly accessible to patients. To better identify CBMs from foods, identifying key ingredients followed by modifying genetics in food via technology is the turning point in this cancer—beating field.

Although some cancers that are complex in molecular interactions cannot be predicted precisely by interpreting the single gene defect, by specialized molecule functions and mediate the molecules more optimally, the previous paper has demonstrated that the innovative combination of machine learning sorting and traversing is successful in implementing into gene sequence, with mapping hyperfood CBMs earning a splendid result of 30% - 40% of cancer curing rate and 85% of classification accuracy.

1.2. Related Works

Considering the ageing global population, individuals need to undergo a more sub-substantial burden of inadequate healthcare and high medium-cine expenses. Luckily, work has shown that certain dietary foods in these fatal cancer cases may potentially substitute drugs with plant-based diets rich in cancer-beating molecules; such as polyphenols, flavonoids, terpenoids and botanical polysaccharides [1]-[7]. Such products not only contain high nutritional values, they are also inexpensive and highly patient-accessible. To better identify CBMs from foods, the turning point in this cancer-beating area is the identification of key ingredients followed by genetic modification in foods through technology.

Using a network-based machine learning method, people have shown that plant-based foods such as tea, carrot, celery, orange, grape, coriander, cabbage and dill contain the largest number of molecules with high anti-cancer likeness through exerting influence on molecular networks in a similar fashion to existing therapeutics. The large scale computational analysis further demonstrates more cancer-beating potential of certain foods calling for more tailored nutritional strategies. However, it has some limitations of the methodology used before; firstly, concentrations of bioactive molecules are not taken into account and it is unclear they would be present in sufficient enough concentration to exert their beneficial biological activity. Furthermore, the proposed methodology only accounts for interactions between bioactive food compounds and cancer-related molecular networks, without explicit regard for directionality of these relationships. In addition, the methods described here do not take into account specific cancer molecular phenotypic characteristics. Finally, the accuracy of the method requires further improvement [8].

1.3. Our Aims and Proposal

The amazing previous work has provided us a feasible access to figure out and

analyzing CBMs; however, the accuracy and F-value is not high enough. This misclassification may lead to clinical puzzles and public fallacy of CMBs. Therefore, our team aims to improve the accuracy by applying a voting algorithm on eight different classifiers from 85% to 90%.

1) Preprocessing the input data

For a limited tool, particularly in the number of unlabeled food compounds, the original compound data is excessive, preventing us from building a well-trained machine. With the primary purpose of research being to enhance the predictive reliability of supervised learning, we need to run algorithm to minimize the unlabeled data from thousands to tens without losing relations and information of data.

2) Selecting and comparing existing supervising algorithm

As for improving the classification accuracy, selecting a well-performed algorithm specifically on our data is super crucial. We need to make comparison on the existing mature supervising algorithm to get the clue for further advanced improvement by running each of them on our food compounds data.

3) Optimizing parameters

For getting the best results for classifying anti-cancer molecules, we must search a method to optimize the weights and parameters for every algorithm. After that, we need to make further improvement by evolving the existing algorithms.

4) Improvement on algorithm

By comparing and contrasting the existing algorithm and then visualizing their performance, we find a way to improve the accuracy by adding weight to let each algorithm vote but using the vote concept. By adding weights corresponding to the performance of select algorithms, each of them can contribute to the classification label for instances with different weight. By this method, some discriminative and deterministic algorithms such as decision tree can be smoothed and transformed into a probabilistically model and it can deal outliers and skew data with its flexibility. These can largely increase the accuracy of the classification for Cancer Beating Molecules.

2. Methods and Materials

2.1. Selecting Algorithms

We downloaded the code of Linear SVM and extract the CSV file of compounds from the original paper. By modifying the learning packages of scikit-learn, we built a platform to run our planned 8 different methods, later partitioning the labeled data into three parts: 80% for training, 15% for testing, and 5% for five-cross fold validation. After optimizing the parameters and running each classifier, we compared the performance by categorizing results and evaluated reliability to adjust the voting weight.

We chose seven machine learning classifiers for testing in the improvement of accuracy:

1) SVM with RBF Kernel;
2) MLP Neural Network;
3) Decision Tree;
4) Random Forest;
5) Gaussian Naïve Bayes;
6) Adaboosting;
7) Bagging.

Although classifiers such as Extreme Learning Machine and several neural networks are viable options, we excluded them due to their incompatibility to our data. After implementing these algorithms, we analyzed their performances with a combination of Soft Voting Algorithm using weights.

2.2. Voting Algorithm

By adding weights corresponding to the performance of select algorithms, each of them can contribute to the classification label for instances with different weights. By this method, some discriminative and deterministic algorithms such as decision tree can be smoothed and transformed into a probabilistically model and it can deal outliers and skew data with its flexibility. These can largely increase the accuracy of the classification for Cancer Beating Molecules.

2.3. Preprocessing the Data

The original compound data is in excess for a limited device to run, especially in the number of unlabeled food compounds, preventing us from building a well-trained machine. With the primary research purpose being improving the prediction accuracy of supervised learning, we cut down the unlabeled data from thousands to tens. By running tests on reducing data, we found that unlabeled data would not influence the accuracy of the eventual operation.

Moreover, we trimmed down some training data proportionally to the labels (CBMs from 199 to 88, non-CBMs from 1950 to 880), as it does not affect the accuracy of prediction. However, we kept the complete internal connections and mapping data of genes-proteins and compounds-gene connections, due to the potential risks of cutting down these ambiguous internal links.

2.4. Optimizing Parameters

Our team experimented with plenty of methods to optimize the parameters of each model, including evolutionary Search CV algorithm Random Search and manual settings. Finally, we found GridSearchCV the best parameter search tool for modeling.

To visualize such a vast difference between GridSearchCV and other parameters optimizer, we compared the accuracy between GridSearchCV and setting parameters manually. After more than twenty times experiments, we obtained the accuracy of anti-cancer, non anti-cancer and F-score for the Voting All algorithm (0.74, 0.78, 0.76 respectively), find that the accuracy dropped 10%

compared to GridSearchCV parameter optimizer with the same algorithm applied.

3. Results and Discussion

3.1. Optimizing Parameter Strategy for Training Model

Among all the optimizing parameter algorithms, we applied GridSearchCV algorithm for our training model. One of the advantages is that GridSearchCV hardly falls in local optimization. Due to GridSearchCV's high expense when meeting a vast dataset, we selected a partial dataset from the original one. More-over, since GridSearchCV is compatible with every supervised learning algorithm, we applied it on the seven supervised machine learning we chose. Utilizing GridSearchCV, we found out the best parameter for voting algorithm (**Figure 1**).

3.2. Supervised Learning Combination Strategy for Anti-Cancer Food Molecule Prediction

The graphs elucidate the accuracy of supervised machine learning on predicting anti-cancer molecules, with Linear SVM being the best. Previously, a group of researchers introduced the Support Vector Machine (SVM) as the best-prediction supervised machine learning [8]. However, SVM is a non-probabilistic binary linear classifier proved to be overly absolute and unreliable on classifying the complex high-dimensional dataset. Therefore, we introduced seven supervised learning machines to the ranking model. Despite that SVM is performing an above-average level in classifying anti-cancer molecules (ac score 0.82, non score 0.83) (**Figure 2**), the anti-cancer score of Radial SVM (0.83) is marginally higher than the SVM's. Hence, we developed a technique by combining machine learning strategies—"soft voting".

To reinforce the performance of supervised machine learning, by obtaining the voting from the confidence level of the seven supervised machines (**Figure 3**). After over thirty times of experimentations, we obtained the best anti-cancer.

4. Conclusions and Future Works

4.1. Comparison with Original Proposal

We used the code from the original paper to extract a CSV file from the Linear

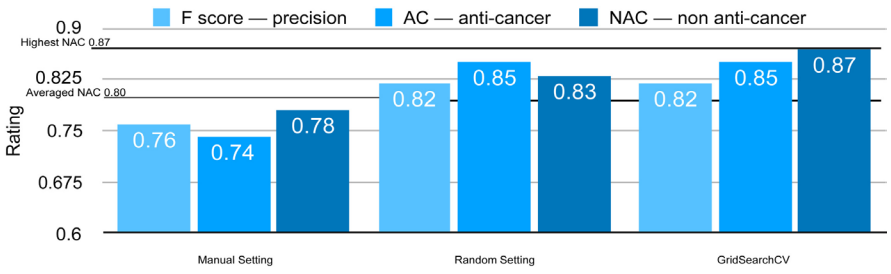

Figure 1. The accuracy improvement of using manual setting, random setting, and GridSearchCV methods.

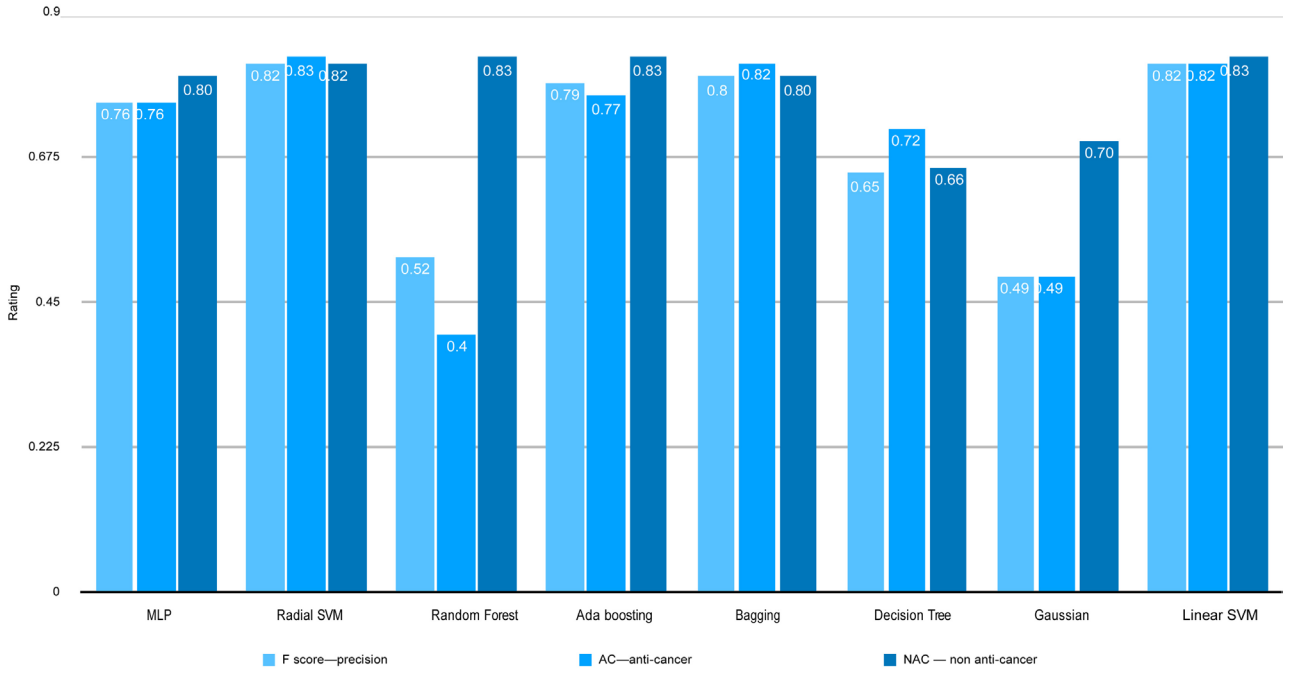

Figure 2. Accuracy rating for 8 different methods.

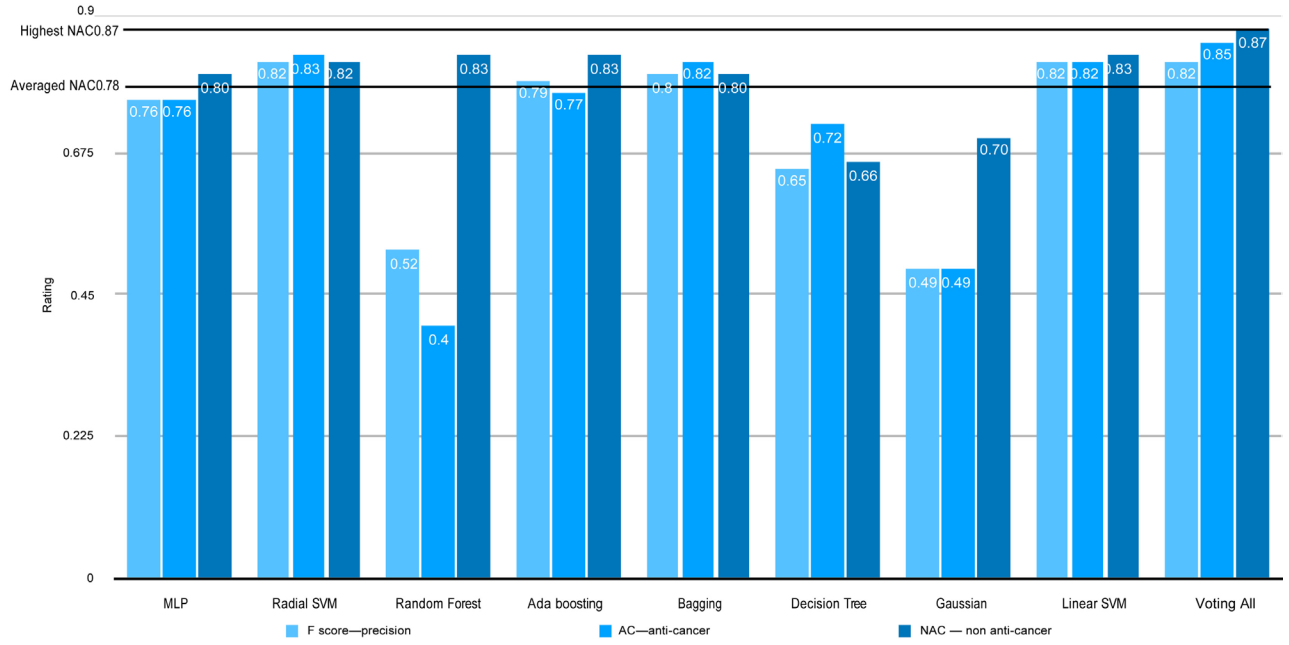

Figure 3. Comparison figure of voting all methods with previous 8 methods.

SVM. By modifying the learning packages of scikit-learn, we built a platform to run our planned methods, later partitioning the labeled data into three parts: 80% for training, 15% for testing, and 5% for five-cross fold validation. After optimizing the parameters and running each classifier, we compared the performance by categorizing results and evaluated reliability to adjust the voting weight.

Our initial goal is to enhance the prediction accuracy of anti-cancer food mo-

lecules of the original paper. Also, we planned to obtain results from the best of the seven supervised learning approaches, with an expected improvement of 85% to over 90% accuracy.

Compared to the original proposal, the final results of the research were as expectation, and the non-anti-cancer (NAC) score of the algorithms adopted improved from 85% to 90%. Additionally, our team replaced Extreme Computer Algorithm with Reinforcement Learning Approaches to weight the supervised learning algorithms, thus obtaining more reliable and stable results. We also attempted to expand our application field and transfer the dataset of anti-cancer to the area of Alzheimer disease. The idea is potentially feasible but halted due to time limitations.

4.2. Commentary on Experience

Despite the significant achievements, we could have avoided a lot of existing problems. One of the most critical issues is that it is exceptionally time-consuming to collect datasets and combine algorithms from available resources in the primary stage. Instead, we should have utilized the time in improving the efficiency of collected algorithms, which could potentially give rise to the accuracy of the code. Secondly, the dataset could not suffice the requirement. By using the data-set from the previous research coupled with a handful of foods, it could not provide sufficient reference to predict CBMs in foods accurately.

Overall, all of the group members performed well, with everyone specializing in what they do best, and collaborating seamlessly throughout the project. With contributions from everyone, we were able to achieve such accomplishment at last.

4.3. Future Goals

There are two primary goals for our project: resolvable and accessible in further experimentation. Based on our progress on supervised learning method, one of our future aims is to design a specifically-matched supervised machine learning algorithm for data analysis and anti-cancer hyperfood molecules labeling, in order to reach or higher accuracy of unsupervised modeling. The other objective is to advance the unsupervised learning part of modeling drug-gene, gene-gene interaction, as so to optimize the true accuracy of the anti-cancer prediction.

Acknowledgements

Special thanks to MIT professor Manolis Kellis for his outstanding contribution to the guidance and integrity of the paper.

Conflicts of Interest

Simiao Zhao participated in code writing and improvement, result analyzing, and composing the method part of this paper. Hanghong Lin participated in re-viewing code, analyzing results, and composing the result and discussion part

of this paper. Peixuan Xu participated in research, data analysis, graph modeling and the revision of this paper. Xuanyue Mao participated in research, data collecting, data analysis, code collecting, and modifying the frame of this paper. HaoYin participated in research, paper writing, results analysis, and graphrevision.

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Prince, M.J., *et al*. (2015) The Burden of Disease in Older People and Implications for Health Policy and Practice. *The Lancet*, **385**, 549-562. https://doi.org/10.1016/S0140-6736(14)61347-7

[2] Drewnowski, A. and Popkin, B.M. (1997) The Nutrition Transition: New Trends in the Global Diet. *Nutrition Reviews*, **55**, 31-43. https://doi.org/10.1111/j.1753-4887.1997.tb01593.x

[3] Tilman, D. and Clark, M. (2014) Global Diets Link Environmental Sustainability and Human Health. *Nature*, **515**, 518-522. https://doi.org/10.1038/nature13959

[4] Dewar, S.L. and Porter, J. (2018) The Effect of Evidence-Based Nutrition Clinical Care Pathways on Nutrition Outcomes in Adult Patients Receiving Non-Surgical Cancer Treatment: A Systematic Review. *Nutrition and Cancer*, **70**, 404-412. https://doi.org/10.1080/01635581.2018.1445768

[5] Donaldson, M.S. (2004) Nutrition and Cancer: A Review of the Evidence for An anti-Cancer Diet. *Nutrition Journal*, **3**, Article No. 19. https://doi.org/10.1186/1475-2891-3-19

[6] Kotecha, R., Takami, A. and Espinoza, J.L. (2016) Dietary Phytochemicals and Cancer Chemoprevention: A Review of the Clinical Evidence. *Oncotarget*, **7**, 52517-52529. https://doi.org/10.18632/oncotarget.9593

[7] Baena Ruiz, R. and Salinas Hernandez, P. (2016) Cancer Chemoprevention by Dietary Phytochemicals: Epidemiological Evidence. *Maturitas*, **94**, 13-19. https://doi.org/10.1016/j.maturitas.2016.08.004

[8] Veselkov, K., Gonzalez, G., Aljifri, S., Galea, D., Mirnezami, R., Youssef, J., Bronstein, M. and Laponogov, I. (2019) HyperFoods: Machine Intelligent Mapping of Cancer-Beating Molecules in Foods. *Scientific Reports*, **9**, Article No. 9237. https://doi.org/10.1038/s41598-019-45349-y