HIGH SCHOOL EDITION
Journal of Student Research
Volume 14 Issue 1 (2025)

From Structure to Function: Biological Activity Prediction of Phytochemicals Using Molecular Fingerprints with Convolutional Neural Networks

Rachel Choi[1] and Sinae Kim[#]

[1]Gangnam International School, Republic of Korea
[#]Advisor

ABSTRACT

Phytochemicals, naturally occurring compounds in plants, offer significant potential for drug development due to their diverse structures and biological activities. They exhibit antioxidant, anti-inflammatory, antimicrobial, anticancer, cardiovascular, and neuroprotective properties, making them beneficial for treating various health conditions. Advantages of phytochemicals include their natural origin, better safety profiles, multi-targeted actions, synergistic effects with other compounds, and sustainability. However, the traditional knowledge-based approach to analizing phytochemicals is limited by its reliance on well-known, locally available plants, resulting in a narrow scope and frequent redundancy. This approach often depends on anecdotal and subjective evidence, faces the risk of knowledge loss, and encounters ethical and legal issues. To address this issue, I propose a convolutional neural network-based systematic approach to predict potential biological activities from molecular structure inputs. The proposed system converts molecular structures into one-hot vector representations using SMILES notation and molecular fingerprint algorithms. These vectors are then fed into a biological activity prediction network to estimate possible biological activities. Through comprehensive experiments, I have demonstrated that applying a convolutional neural network-based machine learning approach yields promising results by achieving an accuracy of 87.8%.

Introduction

The total number of plant species existing on earth is approximately 390,000. Among these, the number of plants being used for medicinal purposes is only about 17,000, indicating that the utilization rate of plant resources is low (Willis 2017).

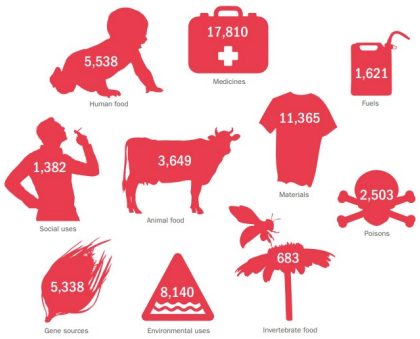

Figure 1. Number of plant species in each use category (Royal Botanic Gardens, Kew. (2016). State of the World's Plants (2016)).

The current approach to developing health foods and pharmaceuticals relies on established knowledge and experience, advancing gradually through stable research. This methodology makes it challenging to explore lesser-known plants due to the high risks involved. To enhance the utilization of plant resources, it is essential to develop advanced plant analysis technologies. By improving these technologies, we can better understand and harness the potential of a wider variety of plant species, thereby expanding the scope of their applications in various fields.

Phytochemical

Phytochemicals are naturally occurring compounds in plants that offer significant health benefits, including antioxidant, anti-inflammatory, antimicrobial, anticancer, cardiovascular, and neuroprotective effects. They are diverse in structure and mechanism, providing a rich source for new drug discovery. Phytochemicals are typically safer and more biocompatible than synthetic drugs, often acting on multiple targets and enhancing the effects of other compounds. They are also more sustainable to produce.

Phytochemicals such as curcumin, resveratrol, quercetin, and EGCG are naturally occurring compounds in plants with significant health benefits, including antioxidant, anti-inflammatory, and anticancer properties. Examples of natural products derived from these compounds include aspirin from willow bark, penicillin from mold, morphine from the opium poppy, and artemisinin from sweet wormwood.

Proposed Methodology

The fundamentals of the proposed system are to list the phytochemicals of various plants, collect their molecular structural formulas, and analyze their main effects on the human body via machine learning systems. The system could be used to develop insights for studies on future food and medical products, predicting potential effects.

The proposed method initiates by collecting phytochemical datasets. The system stores the molecular structure form of the dataset in the SMILES (Simplified Molecular Input Line Entry System) notation (Weininger 1988), and to prepare the data as an input form of machine learning, it goes through preprocessing to be expressed as a one-hot vector (Muegge and Mukherjee 2016).

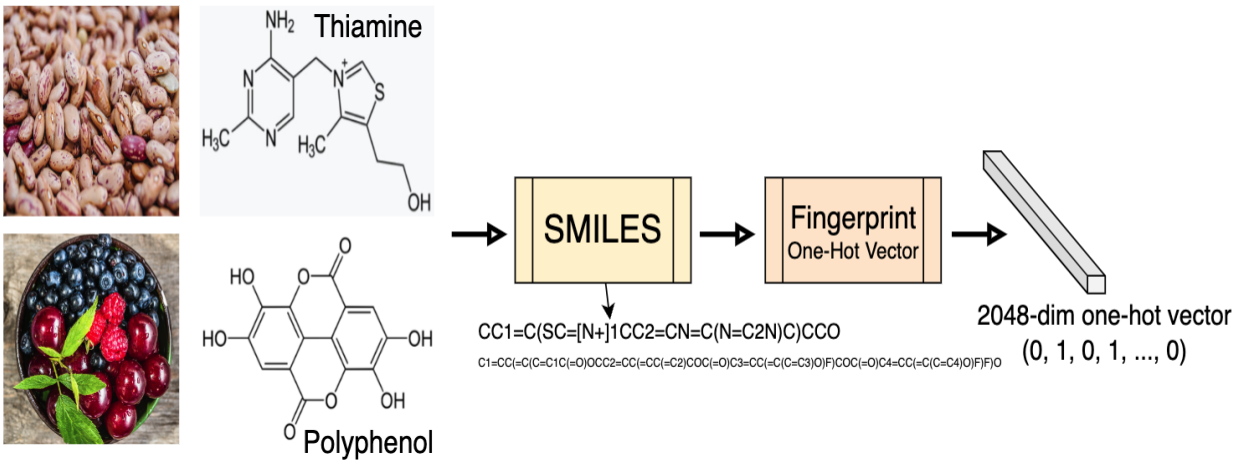

Figure 2. Preprocessing of the research shows the process of changing notations from SMILES to one-hot vector.

After the dataset is converted, it is analyzed by machine learning. Through the analysis, multi-label classification is performed through neural network prediction and binary classification. In this study, each label symbolizes anticancer effect, antioxidant effect, toxicity, and lipid metabolism involvement. Through neural network prediction, the percentage of each label to be represented is shown, and by binary classification, the result for true or false to have a potential effect is verified.

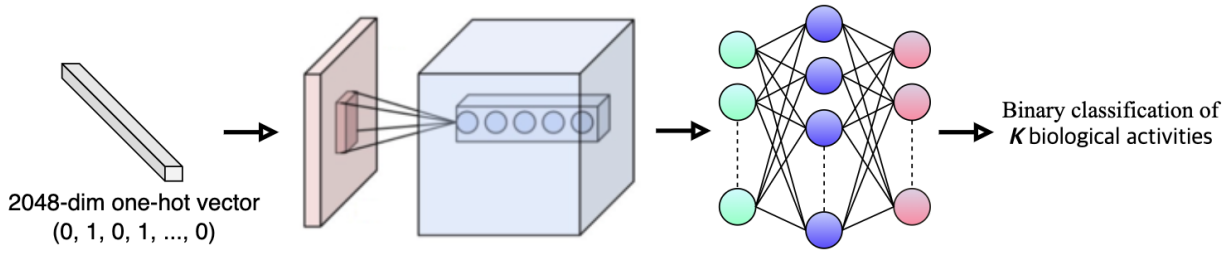

Figure 3. Architecture of the proposed bioactivity prediction network

Equation 1: Binary cross entropy loss function

$$L = -\frac{1}{K}\sum_{k=1}^{K} gt_c \times ln\left(\widehat{pred_c}\right) + (1 - gt_c) \times \ln\left(1 - \widehat{pred_c}\right)$$

This equation represents the cross-entropy loss function that is used for the system. This type of loss is employed in machine learning models to quantify the discrepancy between predicted probabilities and actual class labels for binary classification tasks.

The loss function calculates a penalty for each data point by considering the predicted probability from machine learning, belonging to a specific class, and the corresponding true label. The calculated penalties are then averaged across all data points to obtain the final loss value. The weight parameter, gt, represented by the binary classification, potentially adjusts the contribution of each term based on the target value.

Experimental Results

Dataset

LThe dataset used in this study for predicting the biological activity of phytochemicals consists of SMILES codes sourced from the 'Biological Activity Dataset' provided by AI Hub (AI Hub 2023). SMILES codes represent the molecular structure of phytochemicals, which are then converted into molecular fingerprints. These fingerprints serve as input features for convolutional neural networks (CNNs) by capturing essential structural information of the molecules.

Additionally, the dataset includes labels indicating the biological activity (active or inactive) of each phytochemical. The data also utilized supplementary data focused on metabolite bioactivity classification (Learning data for classification of bioactivity of metabolites), which provides further annotations on the physiological effects of the compounds. This data covers the biological activities of phytochemicals—in the case of this experiment and data, lipid metabolism (60 items, 10.2%), antioxidant activity (169 items, 28.8%), anti-inflammatory and immunity (160 items, 27.3%), and neural toxicity (197 items, 33.6%)—that supports the development of models that can predict types.

Evaluation Protocol

To access the performance of the proposed network, I utilized four evaluation metrics: accuracy, precision, recall, and F1-score (Hossin and Sulaiman 2015). The following metrics were selected to provide assessments of the model's predicting abilities.

Equation 1: Accuracy

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Accuracy is the general measure of ranges when the model is correct. It is defined as the ratio of correctly predicted predictions—which means true positive and true negative—to the total number of predictions in a dataset.

Equation 2: Precision

$$Precision = \frac{TP}{TP+FP}$$

Precision reflects the model's ability to identify correctly for the relevant predictions. It is the ratio of true positive predictions to all positive predictions.

Equation 3: Recall

$$Recall = \frac{TP}{TP+FN}$$

Recall reflects the model's ability to identify all relevant instances.

Equation 4: F1-Score

$$F1score = 2 \times \frac{Precion \times Recall}{Precision + Recall}$$

The F1 score is the harmonic mean of precision and recall, which is a metric that balances both sides. For all experiments, a K-fold validation was tested to evaluate the evaluations. In this approach, the dataset was structured as 4 subsets (folds), and models are trained and validated. The experiment was evaluated four times, collecting outcomes as a validation set and the rest as training.

Performance Comparison

For the presetting of the experiment, an experiment was conducted to compare performance differences according to model layer depth changes. As a result, layers of 32 were found to be the most effective, showing the highest rate of all accuracy, recall, precision, and F1-score. As the number of layers got bigger than 32, an overfitting problem occurred in which results were biased toward some specific data we used for training.

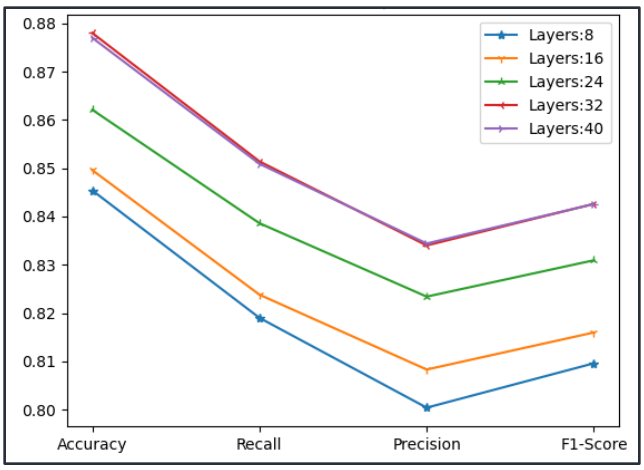

Figure 4. Evaluation result for different layer depth settings

As a result of checking the error value using the results of K-fold cross-validation, there were not many differences in the values of validation results for each training data. It was found that the model results could be used. Accuracy could be confirmed by seeing that the true positive and false negative values were higher than the others in all four categories.

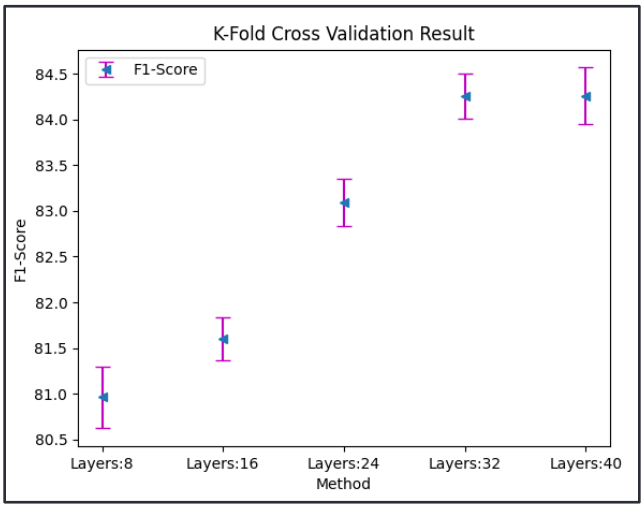

Figure 5. K-fold cross validation result

(a) (b)

(c) (d)

Figure 6. Confusion matrix
(a): toxicity, (b): lipid metabolism, (c): antioxidant, and (d): anti-inflammatory and immune

Conclusion

In this research, I proposed a model to predict the potential bioactive effects of previously unknown plants by organizing the molecular structures and functions of phytochemicals using a machine learning model. Initially, for finding the most effective number of layers, results were analyzed by training networks with various layer-depths. When there were 32 layers, it achieved an accuracy of 87.8, which was the highest among. By this evidence, modeling was conducted with 32 layers, and as a result of the confusion matrix experiment, true positive was achieved at about 89.3% average, indicating successful validation by the model. The proposed system would be effectively used to introduce the phytochemicals of not-well-known plants and will evaluate their development potential for those who want to develop new drugs for specific physiological functions. Afterwards, to develop better functions for the model, more machine learning could be conducted by adding more bioactivity categories. Also, in order to use one hot vector, the three-dimensional molecular structures were changed to the one-dimensional SMILES code. This method could be devised to convert the preprocessing process into a method that can be better evaluated by using the biochemical structure for higher accuracy and recognition.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

AI Hub (2023, Dec 14). “Plant functionality prediction genomic data”: AI Hub. https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71316

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, 5(2), 1.

Muegge, I., & Mukherjee, P. (2016). An overview of molecular fingerprint similarity search in virtual screening. Expert opinion on drug discovery, 11(2), 137-148.

Pathania, S., Ramakrishnan, S. M., & Bagler, G. (2015). Phytochemica: a platform to explore phytochemicals of medicinal plants. Database, 2015, bav075.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of chemical information and computer sciences, 28(1), 31-36.

Willis, K. (2017). State of the world's plants 2017. Royal Botanics Gardens Kew.