HIGH SCHOOL EDITION
Journal of Student Research

Volume 14 Issue 1 (2025)

Machine Learning-Based Prediction of Biological Activity in Natural Product Phytochemicals

Jean Lee[1] and Sue Jung Kim[#]

[1]Sunny Hills High School, USA
[#]Advisor

ABSTRACT

Natural product phytochemicals hold significant promise for drug discovery due to their diverse biological activities and potential to offer novel therapeutic solutions. These compounds are derived from plants, fungi, and other natural sources, providing a rich source of chemical diversity that is often not found in synthetic libraries. Their complex structures and unique mechanisms of action can lead to the development of drugs with new and innovative therapeutic properties. However, discovering and developing new drugs from these compounds presents challenges due to the reliance on knowledge-based methods, which can be time-consuming and inefficient. To address this issue, we propose a novel bioactivity classification network designed to predict the biological activities of phytochemical samples. This study utilized a dataset categorized into four groups: Antioxidant, Toxicity, Anti-inflammatory and Immune, and Lipid Metabolism. The performance of the network was assessed using key metrics including Accuracy, Recall, Precision, and F1-Score. The results revealed that a model with a depth of 4 layers achieved the highest performance. The proposed network achieved an accuracy of 0.7926 and an F1-Score of 0.7248 on the public dataset.

Introduction

The drug discovery process begins with target identification and validation. Scientists first identify a molecule, often a protein, that is associated with a disease and can be targeted by a drug. Once identified, they validate that affecting this target will indeed alter the disease's course. Following this, researchers move on to hit identification. This involves screening natural products, chemical libraries, or synthetic compounds to find potential drugs (hits) that affect the target. Once potential hits are identified, the process advances to the hit-to-lead stage, where these hits are optimized for increased effectiveness, selectivity, and safety through medicinal chemistry and iterative testing (Rao and Srinivas 2011).

Natural product-based drugs offer several advantages in this process (Ben-Shabat et al. 2020). They often possess unique and complex structures that are not easily replicated by synthetic chemistry. These characteristics allow for novel interactions with biological targets which potentially leads to more effective drugs. Natural products have evolved to interact with biological systems, which can translate to better compatibility and fewer side effects in humans. Examples of successful natural product-based drugs include aspirin, derived from salicin found in willow bark; penicillin, discovered from the mold Penicillium notatum; taxol (paclitaxel), originally extracted from the Pacific yew tree and used in cancer treatment; and artemisinin, derived from the sweet wormwood plant and used to treat malaria. These examples illustrate how natural products can inspire synthetic analogs or derivatives with improved pharmacological properties.

Utilizing natural products in drug discovery presents several challenges. One of the difficulties is the complexity and diversity of natural compounds. While this diversity can be beneficial, it also makes the isolation, identification, and characterization of bioactive compounds more challenging. Natural products often exist in minute quantities in their natural sources which requires extensive and sometimes destructive harvesting methods to obtain sufficient material for study. Furthermore, while traditional knowledge provides valuable insights into the therapeutic uses of certain natural products, it is often limited to well-known and historically used substances. This knowledge-based approach can overlook the vast potential of lesser-known or undiscovered natural products. Exploring these new sources requires extensive bioprospecting, which involves searching diverse ecosystems for novel bioactive compounds. This process is time-consuming, and expensive, and often requires collaboration with local communities and adherence to biodiversity and conservation laws.

To address these challenges, I propose a machine learning-based biological activity prediction system. This system takes the molecular structures of phytochemicals from natural products as input and predicts their potential biological activities. To develop this system, I utilized fully connected neural networks, which are highly effective for analyzing complex non-linear patterns. Additionally, I introduced preprocessing steps to appropriately handle the 3-D structures of the molecular inputs for the proposed network.

The remainder of this research chapter is structured as follows: Chapter 2 provides an overview of phytochemicals. Chapter 3 explains the detailed steps in developing the proposed system, including the design and training of the neural networks. Chapter 4 reviews the experimental results which demonstrates the system's effectiveness and accuracy. Finally, Chapter 5 presents the conclusions drawn from this research and suggests directions for future work.

Phytochemicals

Phytochemicals are naturally occurring compounds found in plants that have various biological activities and health benefits (Yadav and Agarwala 2011). These compounds are responsible for the color, flavor, and aroma of plants. Phytochemicals can be classified into several categories based on their chemical structures and functional properties, including flavonoids, alkaloids, terpenoids, and phenolic acids.

For example, tomatoes contain lycopene, a powerful antioxidant belonging to the carotenoid family. Lycopene is responsible for the red color of tomatoes and is known for its potential health benefits. Green tea, derived from the camellia sinensis plant, is rich in catechins, particularly epigallocatechin gallate. These flavonoids possess potent antioxidant properties that help neutralize free radicals, potentially reducing the risk of chronic diseases such as cardiovascular disease and cancer. Turmeric, from the Curcuma longa plant, contains curcumin, a polyphenol known for its strong anti-inflammatory and antioxidant effects. Curcumin can help reduce inflammation, making it beneficial for managing chronic inflammatory conditions like arthritis and inflammatory bowel disease. Additionally, its antioxidant properties protect cells from damage caused by free radicals and may offer neuroprotective benefits, potentially aiding in the prevention of neurodegenerative diseases such as Alzheimer's disease.

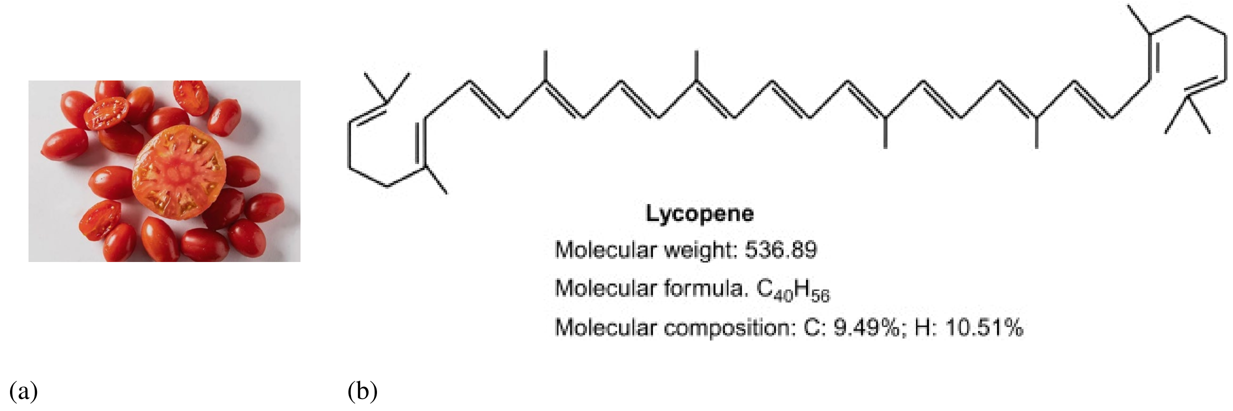

Lycopene
Molecular weight: 536.89
Molecular formula. $C_{40}H_{56}$
Molecular composition: C: 9.49%; H: 10.51%
(a)
(b)

Figure 1. Example of phytochemical (Lycopene)
(a): Tomatoes are one of the best dietary sources of lycopene, and (b): molecular structure of lycopene

Despite the significant advantages of natural products, the current utilization focuses on those that are already well-known. To broaden the application of natural products globally, this research proposes a high-throughput biological activity prediction system specifically designed to leverage phytochemicals. By utilizing advanced machine learning techniques, the system aims to predict the biological activities of a wide range of phytochemicals, including those not yet explored.

Biological Activity Prediction System

The proposed biological activity prediction system consists of two main modules: a preprocessing module for molecular structures and a biological activity prediction module. The preprocessing module takes the molecular structures of phytochemicals and converts the data format into one-hot encoded vectors suitable for neural network input. These vectors mathematically represent the characteristics of the input molecules. The biological activity prediction module then inputs these converted vectors to predict possible biological activities, such as antioxidant, lipid metabolism, toxicity, anti-inflammatory, and immune responses.

Preprocessing of Molecular Structures

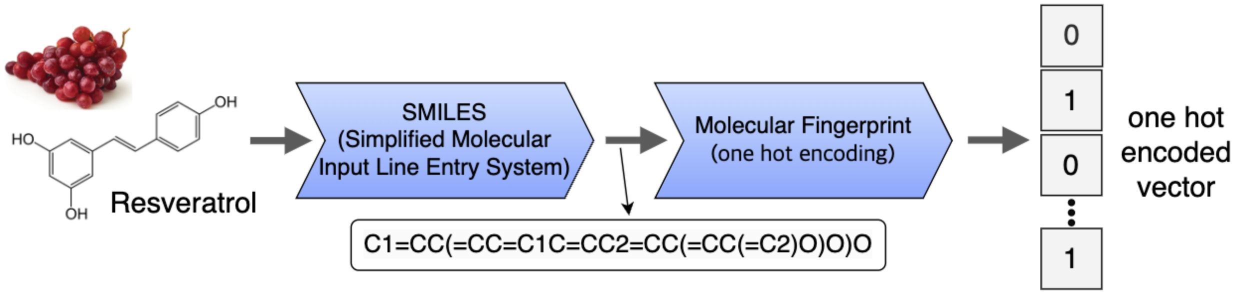


Figure 2. Flow chart of the preprocessing of molecular structures

Figure 2 illustrates the proposed preprocessing of molecular structures. Each molecular structure of a phytochemical sample is converted into SMILES format (Achary et al. 2019), representing the molecular structure as a single string. This representation is then converted into a binary one-hot encoded vector using a molecular fingerprint algorithm (Cereto-Massagué et al. 215). These two steps are commonly used in bioinformatics to preprocess data for machine learning techniques.

Biological Activity Prediction

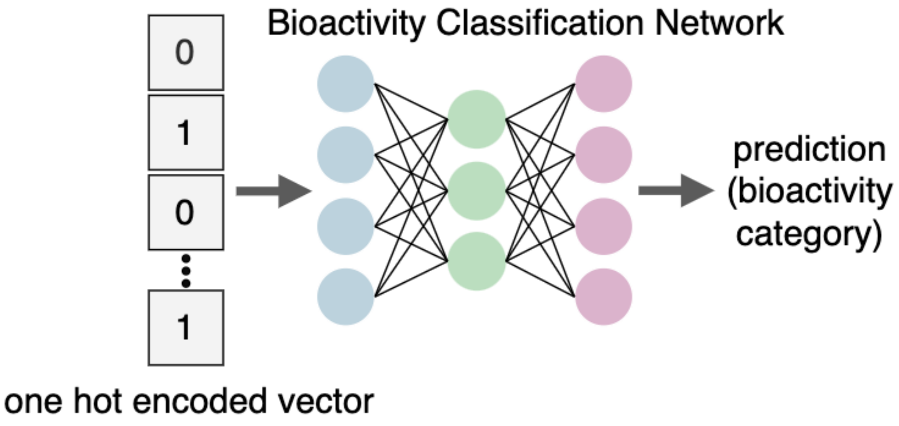

Bioactivity Classification Network
0
1
0
1
one hot encoded vector
prediction (bioactivity category)

Figure 3. Architecture of the proposed biological activity prediction network

Figure 3 depicts the architecture of the proposed bioactivity classification network. The network takes the one-hot encoded vector as input and maps it into four different categories: antioxidant, lipid metabolism, toxicity, and anti-inflammatory/immune responses in a multi-label manner. The output prediction vector consists of four numbers, each representing the probability of a specific bioactivity. For example, if the output prediction vector $\hat{y}$ is $\hat{y}$ = {0.7, 0.8, 0.4, 0.2, it indicates a 70% probability of having the first bioactivity, an 80% probability of having the second bioactivity, a 40% probability of having the third bioactivity, and a 20% probability of having the fourth bioactivity. To measure the error of the prediction vector, the binary cross-entropy loss function is used, as explained in Equation 1.

Equation 1: Binary cross entropy loss function for multi-label classification

$$Loss = -\sum_{b=1}^{B}(y_b \times \ln(\hat{y}_b) + (1 - y_b) \times ln(1 - \hat{y}_b)$$

In Equation 1, y denotes the prediction vector's corresponding ground truth vector, with y_b representing the *b-th* element of this ground truth vector. B denotes the total number of biological activities, which is set to 4 in this paper. The loss function accumulates the loss values for all four categories and is then used to train the proposed network using the gradient descent algorithm, which is the most popular algorithm for training machine learning models.

Experimental Results

Phytochemical Dataset

The dataset used in this study comprises phytochemical samples categorized into four distinct biological activities: antioxidant (169 samples, 28.84%), toxicity (197 samples, 33.62%), anti-inflammatory and immune (160 samples, 27.30%), and lipid metabolism (60 samples, 10.24%). The antioxidant category includes samples known for neutralizing free radicals, while the toxicity category contains samples that exhibit potential toxicity, crucial for safety assessments. The anti-inflammatory and immune category comprises samples with properties important for treating inflammatory diseases and boosting immune responses. The lipid metabolism category includes samples that influence lipid metabolism.

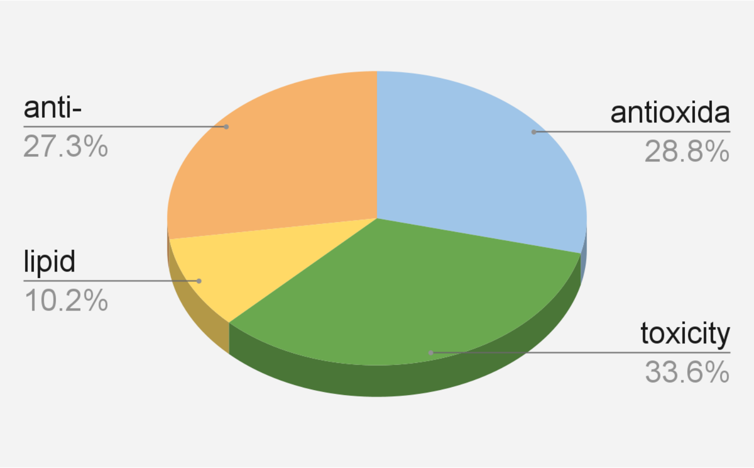


Figure 4. Sample distribution of dataset used in this research

Inference Metrics

To assess the performance of the proposed method, I utilized four evaluation metrics commonly used in many classification tasks: Accuracy, Recall, Precision, and F1-Score (Hossin and Sulaiman 2015). Accuracy is the ratio of correctly predicted samples to the total number of samples which provides an overall measure of the model's performance. Recall, also known as sensitivity, measures the ability of the model to correctly identify positive samples. Precision evaluates the proportion of true positive predictions among all positive predictions made by the model. F1-Score is the harmonic mean of Precision and Recall which offers a balanced measure that accounts for both false positives and false negatives. Equations 2-5 explain the aforementioned four evaluation metrics.

Equation 2: Accuracy metric

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

Equation 3: Recall metric

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

Equation 4: Precision metric

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

Equation 5: F1-Score metric

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

Here, *TP* and *TN* represent true positive and true negative predictions, respectively, which are correct predictions, while *FP* and *FN* denote false positive and false negative predictions, respectively, which are incorrect predictions.

Evaluation

Table 1. Evaluation results

Architectures	Accuracy	Recall	Precision	F1-Score
# of layers: 2	0.7785	0.7284	0.7004	0.7141
# of layers: 4	0.7926	0.7486	0.7148	0.7248
# of layers: 6	0.7915	0.7472	0.7134	0.7299

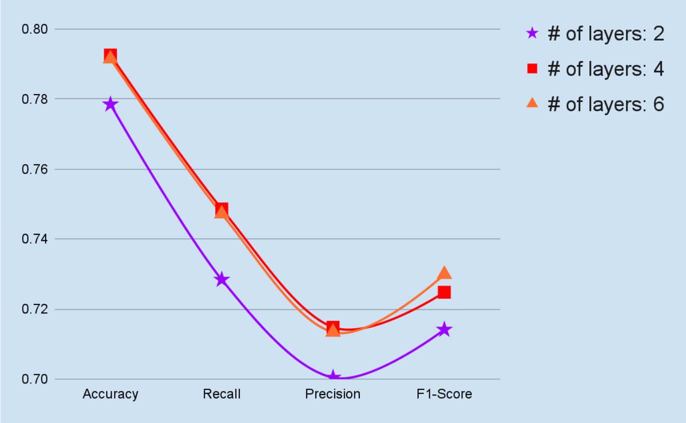
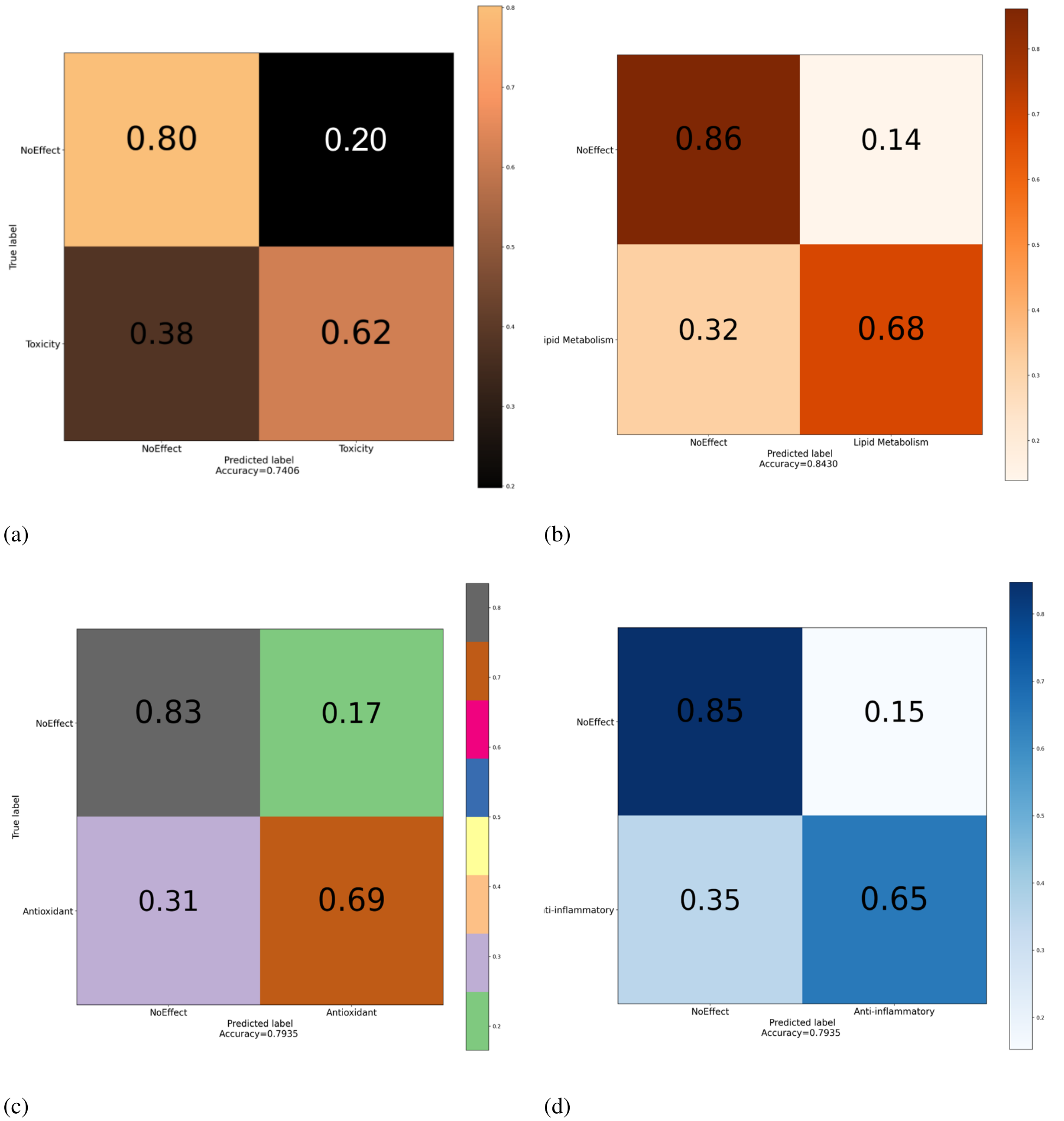


Figure 5. Visual illustration of Table 1

Table 1 and Figure 5 summarize the evaluation results. I trained three different models with varying layer depths from 2 to 6. Generally, increasing the layer depth improves the model's accuracy; however, I found that a depth of 4 layers yields the best results. Increasing the depth beyond 4 layers does not lead to significant improvements in performance.



(a)

(b)

(c)

(d)

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

Additionally, I evaluated the confusion matrix for each category to assess the true positive (TP) and true negative (TN) ratios, as shown in Figures 6 (a)-(d). The results indicate that the model performs well across all categories: for Toxicity, TP is 0.80 and TN is 0.62; for Lipid Metabolism, TP is 0.86 and TN is 0.68; for Antioxidant, TP is 0.83 and TN is 0.69; and for Anti-inflammatory and Immune, TP is 0.85 and TN is 0.65. These results reveal that while Lipid Metabolism has the highest TP ratio, Toxicity has the lowest TN ratio.

Conclusion

In this study, a bioactivity classification network was proposed to predict the biological activities of phytochemical samples. The dataset comprised four categories: Antioxidant, Toxicity, Anti-inflammatory and Immune, and Lipid Metabolism. The model's performance was evaluated using four metrics: Accuracy, Recall, Precision, and F1-Score. The results demonstrated that increasing the layer depth of the model generally improved accuracy, with a depth of 4 layers yielding the best performance. Further increases in depth did not result in significant gains. Confusion matrix analysis for each category revealed varying levels of performance. The model showed strong true positive (TP) ratios across all categories, with the highest being for Lipid Metabolism and the lowest true negative (TN) ratio for Toxicity. In the future, I plan to apply different machine learning architectures, such as sequence models, to enhance accuracy.

Acknowledgments

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

References

Achary, P. G. R., Toropova, A. P., & Toropov, A. A. (2019). Combinations of graph invariants and attributes of simplified molecular input-line entry system (SMILES) to build up models for sweetness. Food Research International, 122, 40-46.

Ben-Shabat, S., Yarmolinsky, L., Porat, D., & Dahan, A. (2020). Antiviral effect of phytochemicals from medicinal plants: Applications and drug delivery strategies. Drug delivery and translational research, 10, 354-367.

Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., & Pujadas, G. (2015). Molecular fingerprint similarity search in virtual screening. Methods, 71, 58-63.

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.

Rao, V. S., & Srinivas, K. (2011). Modern drug discovery process: An in silico approach. Journal of bioinformatics and sequence analysis, 2(5), 89-94.

Yadav, R. N. S., & Agarwala, M. (2011). Phytochemical analysis of some medicinal plants. Journal of phytology, 3(12).