HIGH SCHOOL EDITION
Journal of Student Research

Volume 14 Issue 1 (2025)

BioActNet: A Machine Learning Approach to Biological Activity Prediction Using Molecular Fingerprints

Sydney Choi[1] and Hangil Song[#]

[1]Jericho High School, USA
[#]Advisor

ABSTRACT

Phytochemical-based drug development offers several advantages, including the discovery of novel therapeutic agents derived from natural sources, often with fewer side effects compared to synthetic drugs. These compounds, found in a variety of plants, have evolved to serve protective functions which makes them great candidates for pharmacological applications. However, the traditional knowledge-based approach to phytochemical drug development has significant limitations. It relies heavily on well-documented plants, which restricts the scope of exploration to already known phytochemicals, thereby potentially overlooking a vast array of unexplored natural compounds with therapeutic potential. To overcome this limitation, we propose a novel approach that leverages advanced computational techniques to predict biological activity from input phytochemicals. The proposed system consists of two modules: the phytochemical preprocessing module and the bioactivity prediction module. The preprocessing module takes the molecular structures of phytochemicals and converts them into molecular fingerprint representations to be fed into the subsequent machine learning-based biological activity prediction network. This prediction network then takes these molecular fingerprints as input and outputs the probability of various biological activities. To enhance the accuracy of the system, a vector shift technique is introduced, which can be easily applied to the prediction module without altering its network architecture. Comprehensive experiments demonstrated that the proposed system achieved state-of-the-art performance, with an accuracy of 90.80% on a public phytochemical dataset.

Introduction

Phytochemicals are naturally produced, bioactive chemical compounds found in various plant species. These naturally derived compounds often exhibit a wide range of bioactivities which provides a rich source of potential therapeutic agents. This natural origin can result in fewer side effects and better biocompatibility compared to synthetic drugs (Chen and Kirchmair 2020). Many phytochemicals have evolved to protect plants from various pathogens, making them potent candidates for antimicrobial, antiviral, and anti-inflammatory treatments. Additionally, the structural diversity of phytochemicals allows for the exploration of novel mechanisms of action, potentially leading to the development of drugs with unique therapeutic properties (Najmi et al. 2022).
However, the traditional knowledge-based approach to phytochemical research has significant limitations. It relies heavily on well-documented plants, which restricts the scope of exploration to already known phytochemicals. This dependence on previously studied plants not only narrows the potential for discovering novel compounds but also overlooks the vast diversity of lesser-known species. Additionally, this approach is time-consuming, as researchers must painstakingly evaluate each phytochemical one by one. Humans possess knowledge of only a small portion of phytochemicals, as merely 5% of the 650,000 possible plant species have been studied and documented. This knowledge-based barrier significantly limits the potential for developing new drugs to improve human health, hindering the discovery of cures for currently incurable diseases.

To break this limiting barrier, I propose a machine learning system which will compare brand new molecular structures with past knowledge to compute potential uses and benefits that the inputted plant can achieve. The proposed system is composed of two main modules: a phytochemical preprocessing module and a bioactivity prediction module. Initially, the preprocessing module processes the molecular structures of phytochemicals and converts them into molecular fingerprint representations. These fingerprints are then fed into a machine learning-based network designed to predict biological activity. The prediction network analyzes these molecular fingerprints and outputs the probability of various biological activities. To improve system accuracy, a vector shift technique is incorporated, which can be seamlessly integrated into the prediction module without modifying its network architecture.

Background Knowledge

Phytochemical

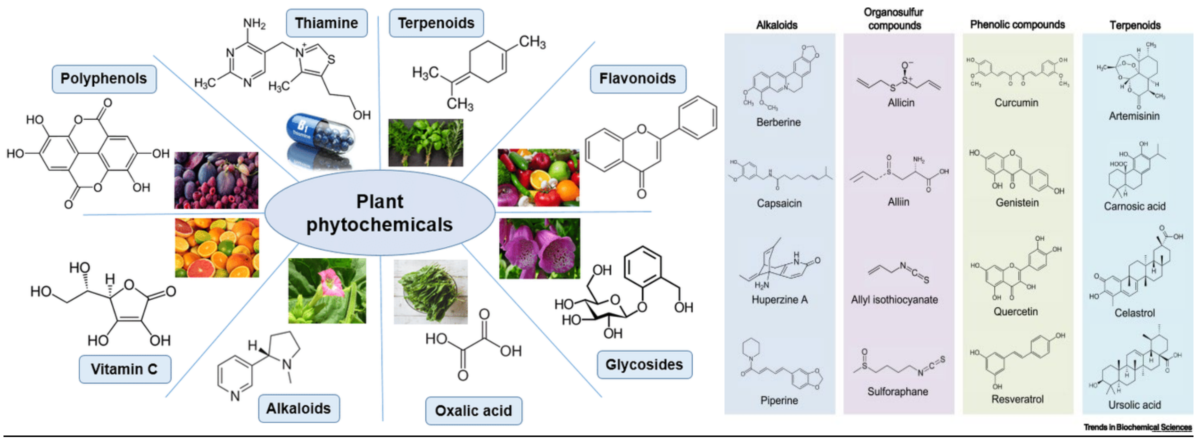


Figure 1. Example of phytochemicals (Soltys et al. 2021) (Martel et al. 2020)

Phytochemicals are natural chemical compounds found in all plants. These compounds contribute to the plant's color, taste, and disease resistance as they play a role in the immune system, protecting the plant from viruses, bacteria, fungi, and parasites. Phytochemicals are biologically active, and while they are not necessary to basic human nutrition, they can provide many health benefits when consumed. There are many different phytochemicals, each with potential antioxidant, anti-inflammatory, antimicrobial, anticancer and other medically beneficial properties.

For instance, quercetin in apples and onions has antioxidant and anti-inflammatory properties, while curcumin in turmeric is known for its potent anti-inflammatory and anticancer effects. Resveratrol, found in grapes and red wine, is celebrated for its antioxidant and cardioprotective benefits. Green tea contains catechins and epigallocatechin gallate, which support heart health and cancer prevention. Lycopene in tomatoes is linked to a reduced risk of certain cancers, and anthocyanins in berries provide antioxidant and anti-inflammatory benefits. Garlic's allicin has antimicrobial properties, and sulforaphane in broccoli offers detoxifying and anticancer effects. Lastly, gingerol in ginger is effective for its anti-inflammatory and digestive health benefits.

CNN-Based Classification

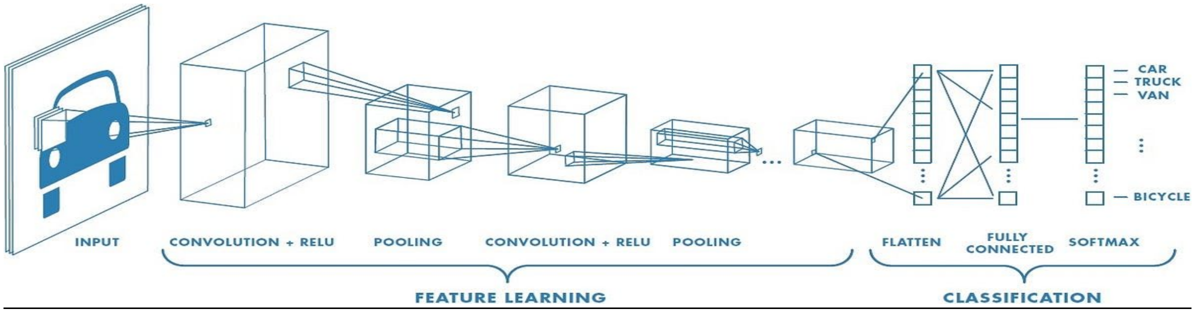


Figure 2. Example of convolutional neural network-based classification system (MathWorks 2024)

A Convolutional Neural Network (CNN)-based classification system is a deep learning model designed for processing and classifying visual data. Its architecture consists of multiple layers, including convolutional layers that apply filters to the input image to extract features, pooling layers that downsample the feature maps to reduce spatial dimensions and computational load, and fully connected layers that perform the final classification based on the extracted features (Li et al. 2021). During training, the CNN adjusts its weights and biases to minimize the difference between its predictions and the actual labels, allowing it to learn and extract complex features crucial for accurate classification.

In a practical example, such as classifying images of handwritten digits using the MNIST dataset (Kadam et al. 2020), the CNN processes grayscale images of digits (0-9) through its layers. Convolutional layers extract features, pooling layers reduce the data size, and fully connected layers produce a probability distribution over the digit classes. For instance, when presented with an image of the digit '7', the trained CNN extracts relevant features and classifies it as a '7' with high probability, recognizing patterns characteristic of the digit.

In this research, I consider biological activity prediction as a multi-label classification task, where the input is a matrix representation of the molecular structure of phytochemicals and the output is the probability of various biological activity categories. A detailed explanation of the proposed method is provided in Chapter 3.

Proposed Bioactivity Prediction Network

Preprocessing

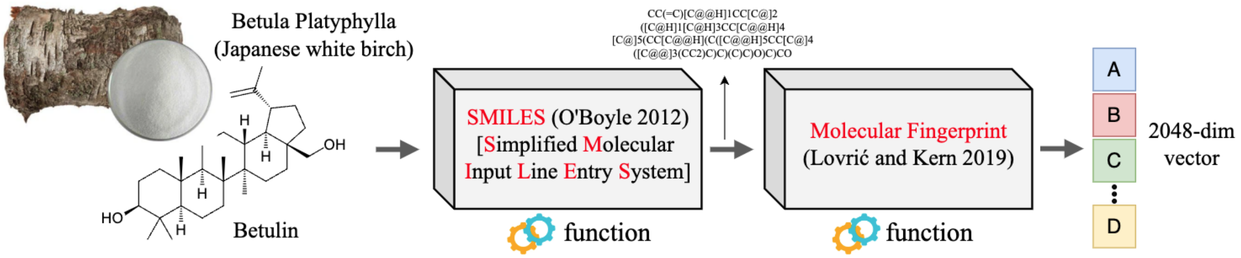


Figure 3. Illustrating converting a 3D structure to 2048 Dimensional Vector

Figure 3 demonstrates the preprocessing procedure, where a Betulin, a complicated 3D structure, is converted into a CNN-applicable 2048 dimensional vector. The 3D structure is first put into the Simplified Molecular Input Entry

System (SMILES), where it is converted into a simple string of letters and symbols. This string is then transformed into a 2048-dimensional vector consisting of just numbers, through the Molecular Fingerprint system, which allows the input information to be used in the machine learning program. In order to input such a complicated structure into a machine learning network, this preprocessing is a necessary step, because without it, it would not be possible to enter a 3D structure into the program to create the needed output.

Bioactivity Prediction Network

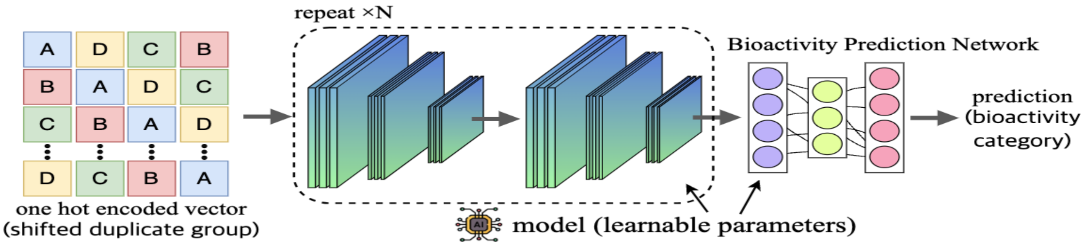


Figure 4. Architecture of the proposed bioactivity prediction network (BioActNet)

The 1 dimensional preprocessing data is duplicated and shifted to create the 2 dimensional input, one hot encoded vector. This is to ensure our data is not separated and that local information stays together even with a small filter size, ensuring accurate information extraction. While this problem could be solved by enlarging the filter size, this would make the whole process computationally inefficient as it would take weeks to process the information in the convolutional neural network. Once our input is efficiently processed with the CNN, feature maps are created and flattened continuously repeated by N. The flattened maps are then inserted into the Bioactivity Prediction Network, a multi-label classifying machine learning program, which creates our final output of the bioactivity category percentages of the inputted phytochemical structure. To train the proposed network, I utilize the binary cross entropy loss function as explained in Equation 1.

Equation 1: Binary Cross Entropy Loss Function

$$Cost = -\frac{1}{B}\sum_{b=1}^{B} y^b \times log_e\left(BioActNet\left(x^b; w, b\right)\right) + (1 - y^b) \times log_e\left(1 - BioActNet\left(x^b; w, b\right)\right)$$

Here, B denotes the number of biological activities while $\boldsymbol{x}$ represents the input. *BioActNet* acts as the function and represents the multi-label classification results, and $\boldsymbol{y}$ conveys the final ground truth.

Model Update

Equation 2: Gradient Descent with momentum

$$\theta_{t+1} \leftarrow \theta_t - lr \times v_t$$
$$v_t = rho \times v_{t-1} + \frac{\partial Cost}{\partial \theta_t}$$

Equation 3: Parameters of the proposed model

$$\theta = \{w, b\}$$

To update the parameters of the proposed network ($\theta = \{w, b\}$), I applied a gradient descent algorithm (Baldi 1995) with momentum (Liu et al. 2020) to minimize loss and allow the network to be more accurate. The gradient descent algorithm utilizes the loss function from equation 1 and optimizes it by picking a randomized $\boldsymbol{w}$ value from the parameter and iteratively stepping in the direction of the negative gradient of the function, minimizing loss. Through this procedure, the parameters of the model, equation 3, are optimized as the learning-rate, *lr,* multiplied by the current momentum, v_t, is subtracted from the current parameter, θ_t, to create θ_{t+1}, which is the new, lower position on the loss function.

The momentum gets optimized as well, using the equation $v_t = rho \times v_{t-1} + \frac{\partial Cost}{\partial \theta_t}$. Momentum is applied so that the algorithm continues to function when met with a local minimum, or saddle point, when the value in the function meets zero and stops descending, even if it hasn't reached the lowest loss amount. The addition of momentum that inserts velocity, or inertia, is represented by v_t. The idea of inertia is that an object will continue its current motion until some force causes its speed or direction to change, and it allows the algorithm to continue decreasing loss, even when met by a local minimum because the velocity allows it to continue moving. The momentum is continuously updated along side the parameters through the process of multiplying the previous momentum (v_{t-1}) by approximately 0.9 (*rho*) and adding the current gradient which is symbolized by $\frac{\partial Cost}{\partial \theta_t}$.

Experimental Results

Dataset

To train and evaluate the proposed method, I utilized a phytochemical dataset from the Korean government Ai hub. There are four different biological activity categories: anti-oxidant, toxicity, anti-inflammatory, and lipid metabolism. Anti-oxidy protects body cells from free radicals, which are unstable molecules produced by the body through metabolism and oxidation processes, that can cause damage. Toxicity signals whether or not the phytochemical brings harm to a living organism. Anti-inflammatory properties reduce swelling, redness, and pain. Lipid Metabolism refers to the phytochemical's ability to aid in the process of breaking down or synthesizing lipids to be utilized in the body. Each category has a different number of samples; anti-oxidant has 169, toxicity has 197, anti-inflammatory has 160, and lipid metabolism has 60. The dataset samples are not very balanced, especially the low number of samples in the Lipid Metabolism category. Due to this imbalance, there is a limitation on how accurate our machine system can be, which is why I collected additional phytochemical samples that fit into our four categories, to further balance and broaden the spectrum of data available for use.

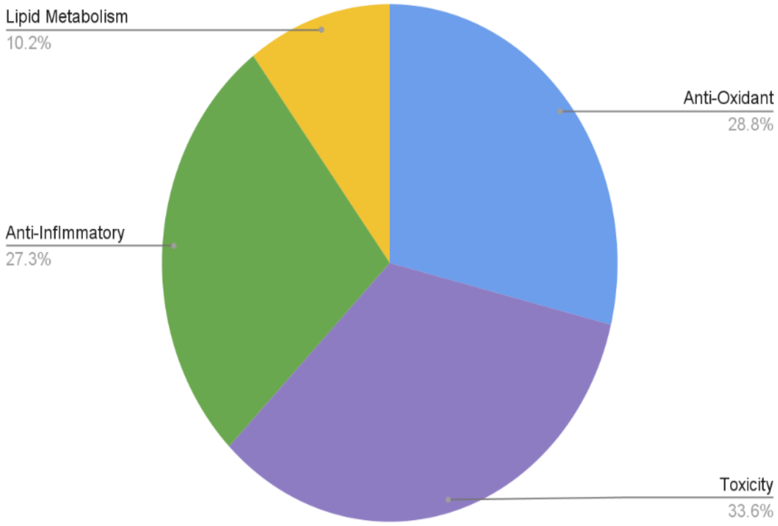


Figure 5. Category distribution of the dataset used in this paper

Evaluation Metric

To evaluate the proposed system, I used four evaluation metrics: accuracy, recall, precision, and F1-Score. The accuracy measures the proportion of correct predictions made by the machine program based on the given dataset, evaluating how many of the classified datasets were in the right classifications. The recall shows the percentage of correct predictions over the total number of actual positive results for the particular classification (ex. out of 10 positive results, the machine correctly detected 7 and classified 3 as negative results). Precision represents how many out of the positively classified results were actually positive. The F1-Score combines the Precision and Recall results to represent the middle value between the two metrics, called the Harmonic mean.

Performance Comparison

To assess whether or not our proposed vector shift does demonstrate increased precision and accuracy in machine learning programs, I used 2 CNN architectures (depth-50, depth-101) and 2 trials for each, resulting in a total of 4 experiments. Each trial was done by inserting the phytochemical data of the machine and running it to evaluate the results. This process was then repeated amongst our 4 machine samples, where each architecture got run once without vector shift (X) and another with vector shift (O), to produce our 4 experimental results, which were organized into Table 1, Figure 6, and Figure 7. Table 1 and Figure 6 both exhibit the four evaluation metrics used and the outcomes per each architecture, where Table 1 shows it as an organized table, and figure size as a line graph.

Table 1. Performance comparison (ablation study)

Architecture	Vector Shift	Accuracy	Recall	Precision	F1-Score
ResNet-50(B)	X	82.66	82.50	82.47	82.49
ResNet-50(P)	O	87.26	87.31	87.25	87.42
ResNet-101(B)	X	84.29	84.14	84.06	84.27
ResNet-101(P)	O	90.80	90.81	90.78	90.80

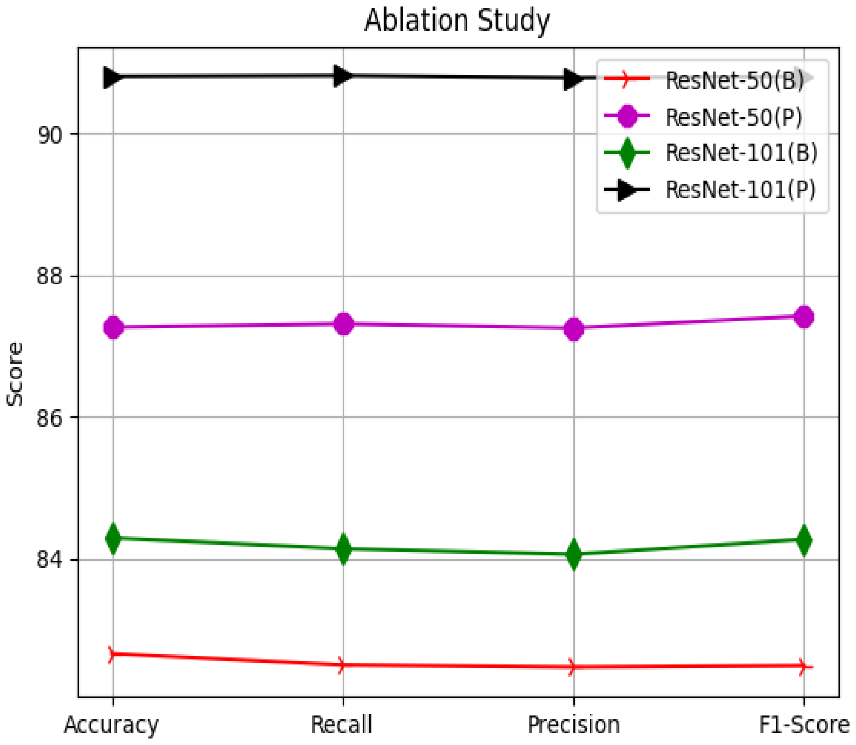
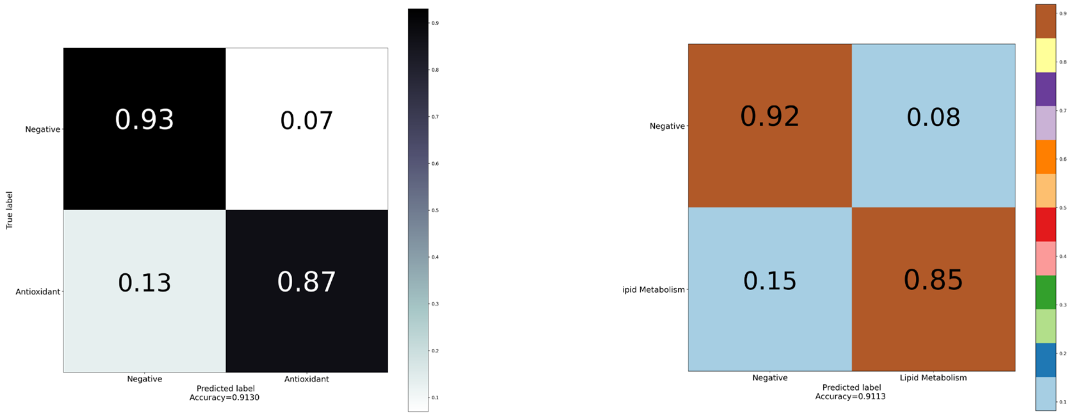


Figure 6. Performance comparison (ablation study)

Figure 7 depicts the evaluation of the confusion matrix for our data, displaying the correct predictions. It is divided into each of the four classifications: (a): Antioxidant, (b): Lipid Metabolism, (c): Anti-inflammatory, and (d): Toxicity. The confusion matrix summarizes the overall performance of the classification model, displaying the percentage of correct classifications as well as misclassifications. The classification with the highest, most consistent rate is the Lipid Metabolism category, while the poorest performance was shown in the Toxicity category. After analyzing the results, we determined the effectiveness of our proposed vector shift approach and confirmed that it does increase the accuracy of the machine network. This can be seen prominently in Table 1 and Figure 6 where the accuracy and precision is much higher in the architectures containing the vector shift, further demonstrating the efficiency of our proposed procedure.



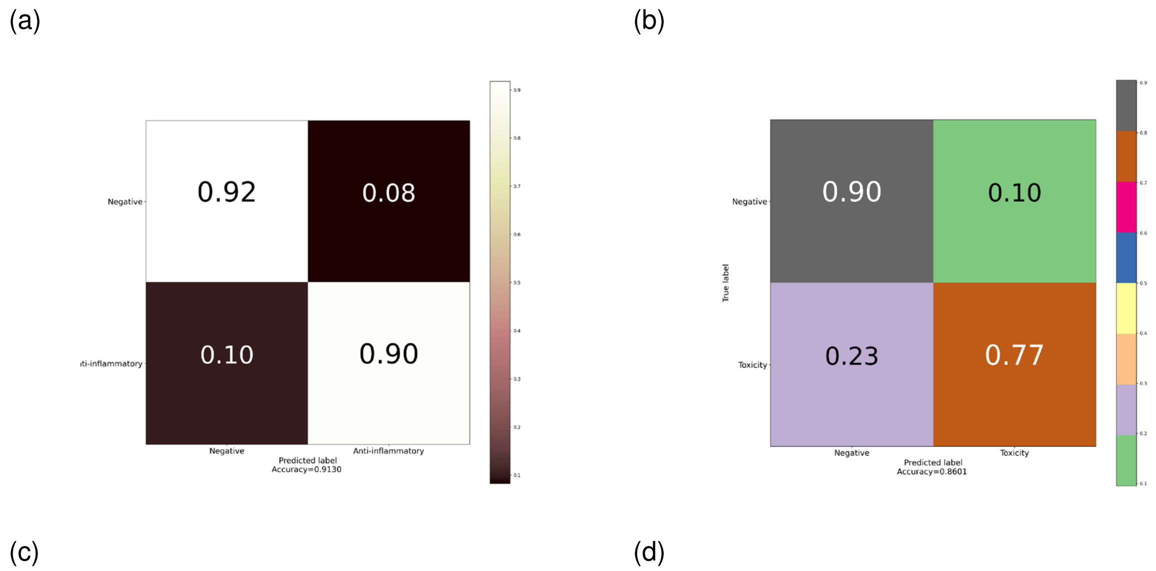


Figure 7. Confusion matrix evaluation result
(a): Antioxidant, (b): Lipid Metabolism, (c): Anti-inflammatory, and (d): Toxicity

Conclusion

In this study, we proposed a machine learning system to identify and classify different phytochemicals found in the molecular structure of various plants. By designing a CNN network with the addition of our proposed vector shift, we created an ai classifier that has the potential to break through the present knowledge barrier of phytomedicine. By preprocessing plant molecular structures in simpler, machine input-able number sequences, the classifier would analyze the structure and by using the dataset of already discovered phytochemical bio-active structures, compare and hypothesize the potential activity that the new plant could bring. Through this, we can broaden the spectrum of plants that can be used for medicine and science, and allow for the creation of better health drugs and scientific breakthroughs. We tested the classifier's abilities by inputting a phytochemical data set from the Korean government Ai hub and allowed the machine to analyze and classify the various molecular structures, and the results proved that our machine is able to classify correct biological activity of phytochemical structures, safe for a few errors.While the results aren't perfect, this ai model demonstrates that this project is feasible and with more training and larger data sets, it can be an incredible tool for the future. Additionally, alongside studying the system, we also experimented with our vector shift proposition. By comparing our CNN networks with and without the shift, it was concluded that the vector shift does increase accuracy and precision of Ai machine systems. This proposal isn't exclusive to our machine system and can be effective in other ai classifiers, allowing for better, more accurate results. For the future, I plan to expand the available dataset for my machine learning system, allowing it to become smarter and more knowledgeable on the biological activities of phytochemicals, which in turn makes the system much more credible to use for plants it doesn't know. By doing so, I hope to create a classifier that can be used without worry in future medical or scientific fields to effectively break through the current knowledge-based barriers and help contribute to new medicinal and botanical discoveries.

Acknowledgments

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

References

Baldi, P. (1995). Gradient descent learning algorithm overview: A general dynamical systems perspective. IEEE Transactions on neural networks, 6(1), 182-195.

Chen, Y., & Kirchmair, J. (2020). Cheminformatics in natural product-based drug discovery. Molecular Informatics, 39(12), 2000171.

Chihomvu, P., Ganesan, A., Gibbons, S., Woollard, K., & Hayes, M. A. (2024). Phytochemicals in drug discovery—a confluence of tradition and innovation. International journal of molecular sciences, 25(16), 8792.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Kadam, S. S., Adamuthe, A. C., & Patil, A. B. (2020). CNN model for image classification on MNIST and fashion-MNIST dataset. Journal of scientific research, 64(2), 374-384.

Martel, J., Ojcius, D. M., Ko, Y. F., & Young, J. D. (2020). Phytochemicals as prebiotics and biological stress inducers. Trends in biochemical sciences, 45(6), 462-471.

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems, 33(12), 6999-7019.

Liu, Y., Gao, Y., & Yin, W. (2020). An improved analysis of stochastic gradient descent with momentum. Advances in Neural Information Processing Systems, 33, 18261-18271.

MathWorks. (2024, Dec 5). "What Are Convolutional Neural Networks? | Introduction to Deep Learning": MathWorks.
https://www.mathworks.com/videos/introduction-to-deep-learning-what-are-convolutional-neural-networks--1489512765771.html

Najmi, A., Javed, S. A., Al Bratty, M., & Alhazmi, H. A. (2022). Modern approaches in the discovery and development of plant-based natural products and their analogues as potential therapeutic agents. Molecules, 27(2), 349.

Soltys, L., Olkhovyy, O., Tatarchuk, T., & Naushad, M. (2021). Green synthesis of metal and metal oxide nanoparticles: Principles of green chemistry and raw materials. Magnetochemistry, 7(11), 145.

Tiwari, N., Gedda, M. R., Tiwari, V. K., Singh, S. P., & Singh, R. K. (2018). Limitations of current therapeutic options, possible drug targets and scope of natural products in control of leishmaniasis. Mini Reviews in Medicinal Chemistry, 18(1), 26-41.

UCLA Health, (2023, May 10). "What are phytochemicals? (And why should you eat more of them?)": UCLA Health.
https://www.uclahealth.org/news/article/what-are-phytochemicals-and-why-should-you-eat-more-them