

Molecular Computational Drug Design Algorithms Through Machine Learning for Phytocompounds from *Dimocarpus Longan*

Asita Elengoe¹, Prasenjit Pal², Sandeep Poddar^{3,*}

¹ Department of Biotechnology, Faculty of Applied Science, Lincoln University College, 47301 Petaling Jaya, Selangor, Malaysia; asitaelengoe@yahoo.com; (A.E.);

² Department of Fisheries Extension, Economics and Statistics, College of Fisheries, CAU (I), Lembucherra, Tripura, India; prasenjit3agstat@gmail.com; (P.P);

³ Lincoln University College, Petaling Jaya, Selangor Darul Ehsan, Malaysia; sandeepoddar@lincoln.edu.my; (S.P.);

* Correspondence: asitaelengoe@yahoo.com; (A.E.);

Scopus Author ID 56118646500

Received: 28.06.2023; Accepted: 23.11.2023; Published: 26.07.2024

Abstract: Since ancient times, plants have been used as a source of medicine. Today, increasingly commercially significant medications have their roots in plants. Medicinal plants have proven to treat various diseases with low or no side effects. This study analyzed the suitability of the drug design of the plant compounds from the *Dimocarpus Longan* (longan) plant for cancer treatment. Three-dimensional (3-D) of thirty selected bioactive compounds from the Longan plant were retrieved from the PubChem database. Retrieved plant compounds are screened using Lipinski's Rule of Five, which provides a standardized requirement or criteria that a ligand should pass to be suitable for drug design. It establishes criteria for drug-like qualities and focuses on medication bioavailability. The Naïve Bayes Machine Learning algorithm was applied to the dataset to classify the plant compounds into two groups after screening using Lipinski's Rule of Five. After classifying the plant compounds into two groups with optimal accuracy, the most influencing plant compounds were identified using principal component Analysis (PCA) techniques. Using Cross-Validation with k-fold=10 shows the accuracy produced by the Naïve Bayes Classifier, which reaches 93.33%. This study suggests that α -terpineol (PubChem ID: 442501) may be the safest and most effective cancer treatment. This *in silico* analysis of α -terpineol's anticancer properties will aid in creating a new and effective drug for cancer therapy.

Keywords: *Dimocarpus Longan*; Lipinski's Rule of Five; Naïve Bayes machine learning algorithm; principal component analysis.

© 2024 by the authors. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Cancer is the world's second-largest cause of death, with an estimated 10 million deaths in 2020 [1]. One in 5 men and one in 6 women worldwide develop cancer during their lifetime, and one in 8 men and one in 11 women die from the disease [2]. Worldwide, the overall number of people living within 5 years of being diagnosed with cancer, called the 5-year prevalence, is estimated at 43.8 million [3]. Around one-third of cancer deaths are due to high body mass index, poor consumption of fruits and vegetables, lack of physical activity, and use of tobacco and alcohol. Tobacco use is the most significant risk factor for cancer and is responsible for about 22% of deaths from cancer [1]. The increasing cancer burden is due to several factors, including population growth and aging, as well as the changing prevalence of certain causes of

cancer linked to social and economic development. The economic effect of cancer is significant and increasing. This is especially true in fast-growing economies, where poverty-related cancers change, and lifestyle-related cancer infections are more common in developed countries. Total annual cancer economic costs were estimated at around US\$ 1.16 trillion in 2010 [4, 5].

Global findings indicate that almost half of the new cases and more than half of the world's cancer deaths in 2018 are expected to occur in Asia for men and women combined, partially because the country has nearly 60 percent of the global population. Europe accounts for 23.4% of global cancer cases and 20.3% of cancer deaths, though it has just 9.0% of the world's population [6]. The Americas account for 13.3 percent of the world's population, 21.0 percent of the incidence, and 14.4 percent of mortality worldwide [7]. Unlike other regions of the world, the proportion of cancer deaths in Asia and Africa (57.3% and 7.3%, respectively) is higher than the proportion of accidents (48.4% and 5.8%, respectively), as these regions have a higher incidence of some forms of cancer associated with lower prognosis and higher mortality rates [8].

Cancer refers to a large number of diseases characterized by the production of abnormal cells that uncontrollably divide and are capable of invasion and destruction of normal body tissues [9]. Such changes are the product of the interaction between the genetic factors of an individual and three groups of outside agents, including chemical carcinogens, such as asbestos, components of tobacco smoke, aflatoxin (a food contaminant), and arsenic (a drinking water contaminant), physical carcinogens, such as ultraviolet and ionizing radiation, and biological carcinogens, such as infections from certain viruses, bacteria, or parasites [10]. The most common cancers are lung (2.09 million cases), breast (2.09 million cases), prostate (1.28 million cases), skin cancer (non-melanoma) (1.04 million cases) and stomach (1.03 million cases) [11].

Medicinal plants have long been recognized for their potential therapeutic benefits and anticancer properties. Many plants contain bioactive compounds that can exert anticancer effects by various mechanisms, such as inducing apoptosis (programmed cell death), inhibiting cell proliferation, and reducing inflammation. Several plants have shown promise as potential anticancer agents [12]. For example, curcumin, the active compound in turmeric (*Curcuma longa*), has been extensively studied for its anticancer properties. It exhibits antioxidant, anti-inflammatory, and anti-proliferative effects and has shown the potential to inhibit the growth of various cancer cells [13,14]. Green tea contains polyphenols, particularly epigallocatechin gallate (EGCG), which has been studied for its anticancer effects. Green tea polyphenols have been shown to inhibit tumor cell growth, induce cell death, and prevent angiogenesis (formation of new blood vessels to supply tumors) [15].

Dimocarpus longan, commonly known as longan, is a tropical fruit tree native to Southeast Asia. It belongs to the Sapindaceae family, including other fruit trees like lychee and rambutan. The longan fruit is highly valued for its sweet and aromatic flavor. It contains various antioxidants, such as flavonoids and phenolic compounds, which help protect the body against oxidative stress caused by free radicals. Antioxidants may help reduce the risk of chronic diseases and support overall health. It also contains vitamin C, which supports immune system function. Vitamin C acts as an antioxidant and plays a crucial role in producing white blood cells, which are essential for fighting off infections. It is often used in traditional Chinese medicine as a natural remedy for anxiety and to promote relaxation. It is believed to have a calming effect on the nervous system and may help reduce stress and promote better sleep. In

traditional Chinese medicine, it is considered a ‘blood tonic’. It is believed to nourish and strengthen the blood, potentially benefiting individuals with anemia or those recovering from illness. It is rich in dietary fiber, which promotes healthy digestion and can help prevent constipation. Fiber-rich foods like longan may support gut health and maintain regular bowel movements. It possesses anti-inflammatory properties. These properties could help reduce inflammation in the body and relieve inflammatory conditions [16-18].

Computer-aided drug design (CADD) is a field of computational biology that involves using computational tools and techniques to design and discover new drugs [19]. One of the approaches used in CADD is structure-based drug design, which involves identifying and optimizing small molecules that can bind to a target protein and modulate its function [20]. The process of structure-based drug design (SBDD) typically involves the following steps. The first step is to identify a target protein that plays a key role in a particular disease. This target protein could be an enzyme, receptor, or other molecule in disease progression. Once the target protein is identified, its three-dimensional structure must be determined. Experimental techniques such as X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy can be used for this purpose. If the experimental structure is unavailable, homology modeling can be used to predict the protein structure based on its similarity to known structures. Virtual screening is used to identify potential drug candidates from large chemical libraries. This can be done using docking algorithms that predict the binding affinity of small molecules to the target protein. The docking algorithms generate a set of ligand poses and rank them based on their binding scores. After the virtual screening, the identified drug candidates, also known as leads, undergo a process of optimization. This involves modifying the chemical structure of the leads to improve their binding affinity, selectivity, and pharmacokinetic properties. Computational methods such as molecular dynamics simulations and quantitative structure-activity relationship (QSAR) analysis can be used to guide lead optimization. Absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties of lead compounds are crucial for their success as drugs. Computational tools can predict various ADMET properties, such as solubility, permeability, metabolic stability, and toxicity, which aid in selecting promising drug candidates. Finally, the selected drug candidates are synthesized and tested *in vitro* and *in vivo* to validate their activity and safety [21, 22]. The computational predictions are compared with experimental results to assess the accuracy and reliability of the computational methods used. The iterative process of designing, synthesizing, and testing compounds continues until a suitable drug candidate with desired properties is identified for further development and clinical trials.

The study aims to verify whether the plant compounds are correctly classified into two groups (Yes=Suitable for drug designing, No= Not Suitable for Drug designing), which were screened using Lipinski’s Rule of Five, and finding suitable phytochemical that can act as a drug for cancer treatment.

2. Materials and Methods

2.1. Plant compounds.

Phytochemicals were used as ligands that act as drugs for cancer treatment. They were identified through a literature review search. The literature review was carried out using electronic databases such as Google Scholar, Science Direct, Elsevier, etc. The bioactive compounds were retrieved based on their medicinal activities in humans. The retrieved thirty

plant compounds (neohesperidin, hesperetin 5-O-glucoside, nobiletin, diosmin, avicularin, nicotiflorin, isotrifoliin, biorobin, spiraeoside, Lepicatechin, piperidine, α -terpineol, lysopc 18:1, o-phosphocholine, betaine, ellagic acid, procyanidin A2, L-glutamic acid, L-aspartic acid, citric acid, kaempferol, quercetin, flavogallonic acid, p-coumaric acid, corilagin, vanillic acid, gallic acid, isoscopoletin, tannin and 4-methylcatechol) in sdf format from the PubChem database [23].

2.2. Screening of plant compounds using Lipinski's Rule.

Lipinski's Rule of Five, which outlines a set of standards that a ligand must meet to be appropriate for drug design, is used to screen the retrieved ligands. It provides standards for drug-like characteristics and concentrates on drug bioavailability [24-26]. The molecular weight (MW) must be equal to or less than 500 daltons (MW 500 daltons), the number of hydrogen bond donors must be equal to or less than five (HBD 5), the number of hydrogen bond acceptors must be equal to or less than ten (HBA 10), the number of rotatable bonds must be equal to or less than ten (RB 10), and the log P value must be equal or less than 5 ($\text{LogP} \leq 5$) and polar surface area ($\text{PSA} \leq 140 \text{ \AA}^2$) for a ligand to pass the requirement to be the suitable drugs for cancer treatment [27-30].

2.3. Toxicity prediction.

The prediction of toxicity was carried out using the Admet SAR 2.0 web-based server [31]. Human Etherà-go-go-Related Gene (hERG) toxicity, AMES toxicity, carcinogenicity (CGT), hepatotoxicity, lethal dose LD_{50} , plasma protein binding (PPB), respiratory toxicity, reproductive toxicity, mitochondrial toxicity, nephrotoxicity, and skin sensitization were among the metrics calculated by this descriptor [32, 33].

2.4. Naïve Bayes machine learning algorithm.

Naïve Bayes Classifier is one of the simplest and most effective Classification algorithms that help build fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts based on the probability of an object. They are fast and easy to implement, but their disadvantage is the requirement for predictors to be independent [34].

2.5. Principal component analysis (PCA).

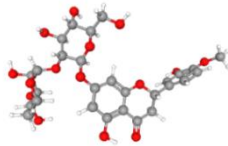
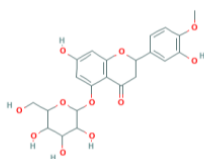
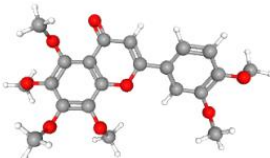
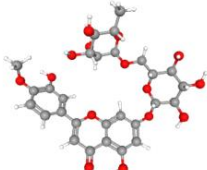
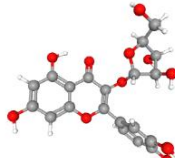
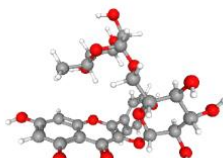
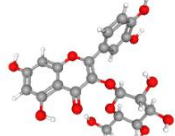
The principal component analysis aims to derive a small number of linear combinations (principal components) of a set of variables that retain as much information in the original variables as possible. Often, a few principal components can be used instead of the original variables for plotting, regression, clustering, and so on. This technique transformed the original set of variables into a new set of uncorrelated random variables. These new variables were linear combinations of the original variables and were derived in decreasing order of importance so that the first principal component accounts for as much of the variation in the original data as possible.

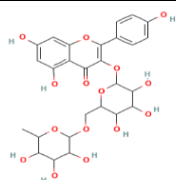
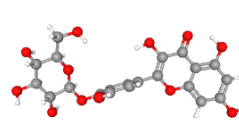
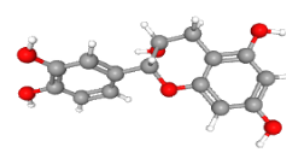
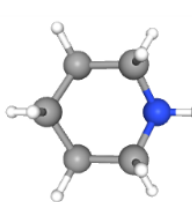
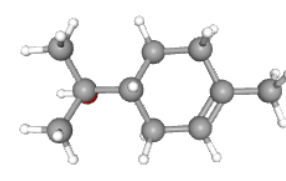
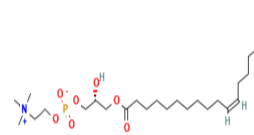
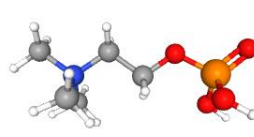
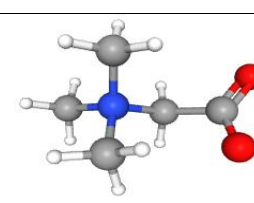
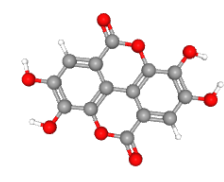
3. Results and Discussion


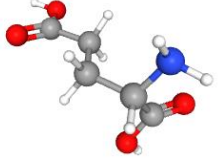
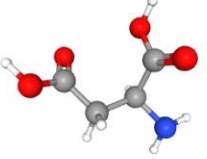
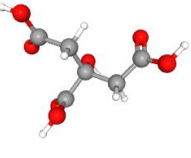
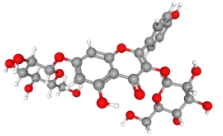
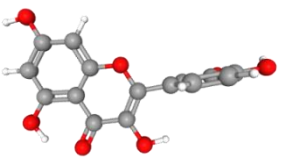
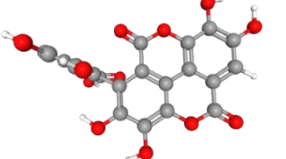
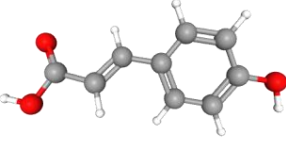
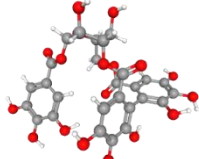
3.1. Retrieval of natural compounds of *Dimocarpus Longan*.

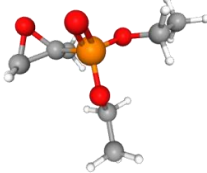
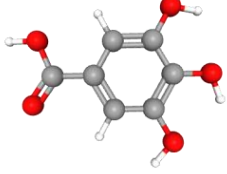
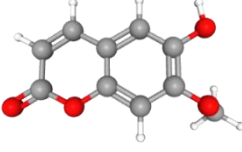
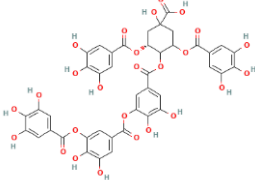
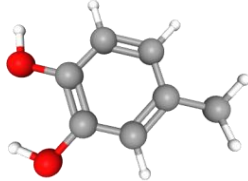
The bioactive compounds of *Dimocarpus Longan* obtained from the PubChem database are displayed in Table 1. They were saved in three-dimensional (3-D) format.

Table 1. Plant compounds of *Dimocarpus Longan* retrieved from the PubChem database.

No	PubChem ID	Bioactive Compound	3-D Structure
1	442439	Neohesperidin	
2	18625123	Hesperetin 5-O-glucoside	
3	72344	Nobiletin	
4	5281613	Diosmin	
5	5490064	Avicularin	
6	5318767	Nicotiflorin	
7	5280804	Isotrifoliin	

No	PubChem ID	Bioactive Compound	3-D Structure
8	12313332	Biorobin	
9	5320844	Spiraeoside	
10	72276	L-Epicatechin	
11	8082	Piperidine	
12	442501	α -Terpineol	
13	53480465	LysoPC 18:1	
14	1014	O-Phosphocholine	
15	247	Betaine	
16	5281855	Ellagic acid	

No	PubChem ID	Bioactive Compound	3-D Structure
17	124025	Procyanidin A2	
18	33032	L-Glutamic acid	
19	5960	L-Aspartic acid	
20	311	Citric acid	
21	6325460	Kaempferol	
22	5280343	Quercetin	
23	14503023	Flavogallonic acid	
24	637542	p-Coumaric acid	
25	73568	Corilagin	

No	PubChem ID	Bioactive Compound	3-D Structure
26	277423	Vanillic acid	
27	370	Gallic acid	
28	69894	Isoscooletin	
29	44144428	Tannin	
30	9958	4-Methylcatechol	

3.2. Screening of plant compounds using Lipinski's Rule of Five.

Selected phytochemicals (ligands) obtained from the PubChem database were screened using Lipinski's Rule of 5. In drug discovery, Lipinski's Rule of 5 can predict the ability and strength of absorption and permeation [35, 36]. According to the Rule of 5, poor absorption and permeation are more likely when there are more than 5 hydrogen bond donors ($HBD \leq 5$), 10 hydrogen bond acceptors ($HBA \leq 10$), the molecular weight is greater than 500 ($MW \leq 500$ daltons), and if the calculated Log P is greater than 5 ($Log P \leq 5$).

Out of the thirty ligands that were screened, seventeen ligands (nobiletin, Lepicatechin, piperidine, α -terpineol, O-phosphocholine, betaine, ellagic acid, L-glutamic acid, L-aspartic acid, and citric acid, kaempferol, quercetin, p-coumaric acid, vanillic acid, gallic acid, isoscooletin, and 4-methylcatechol) pass the evaluation test (Table 2). These ligands show no violation of Lipinski's Rule of 5 where these ligands have less than 5 hydrogen bond donors, less than 10 hydrogen bond acceptors, a molecular weight of less than 500, and calculated Log P is less than 5, which indicates that these compounds have good absorption and permeation which possess the chemical and physical properties to be orally active drugs and can proceed for Naïve Bayes Machine Learning algorithm and Principal Component Analysis.

3.3. Prediction toxicity of plant compounds.

An *in silico* toxicity test was conducted using the admet SAR 2.0 web server to identify the harmful effects of the phytochemicals. Table 3 shows the human Etherà-go-go-Related Gene (hERG) toxicity, AMES toxicity, carcinogenicity (CGT), hepatotoxicity (HT), lethal dose LD₅₀, and plasma protein binding (PPB), respiratory toxicity (RT), reproductive toxicity (RPT), mitochondrial toxicity (MT), nephrotoxicity (NT) and skin sensitization (SS) results. The results show that ellagic acid, nobiletin, kaempferol, quercetin, gallic acid, and 4-methylcatechol had hepatotoxicity. Phloroglucinol, kaempferol, isobutyl isothiocyanate, taurine, and apigenin are a few examples of hepatotoxic flavonoid compounds. Numerous studies on these phyto-compounds' medicinal effects have previously been published [37]. Furthermore, O-phosphocholine possesses properties of carcinogenicity. Nobiletin may cause cardiac side effects due to its hERG toxicity.

Table 3. Toxicity test on selected natural compounds.

Plant Compounds	hERG Toxicity	AMES Toxicity	CGT	HT	SS	RT	RPT	MT	NT	PPB	RAT (LD ₅₀)
Betaine	NO	NO	NO	NO	NO	YES	NO	YES	YES	0.376	0.6371
Citric Acid	NO	NO	NO	NO	NO	NO	NO	NO	NO	0.220	0.8407
Ellagic Acid	NO	NO	NO	YES	NO	YES	YES	YES	NO	0.993	0.6020
L-Aspartic Acid	NO	NO	NO	NO	NO	NO	NO	NO	NO	0.194	0.5911
L-Epicatechin	NO	YES	NO	NO	NO	YES	YES	YES	NO	1.034	0.6433
L-Glutamic Acid	NO	NO	NO	NO	NO	YES	NO	YES	NO	0.104	0.6349
Nobiletin	YES	NO	NO	YES	NO	NO	YES	NO	NO	1.055	0.6245
O-Phosphocholine	NO	NO	YES	NO	NO	YES	NO	YES	YES	0.658	0.5117
Piperidine	NO	NO	NO	NO	NO	NO	NO	YES	YES	0.655	0.7407
α-Terpineol	NO	NO	NO	NO	YES	NO	NO	NO	NO	0.652	0.6381
Kaempferol	NO	YES	NO	YES	NO	YES	YES	YES	NO	1.094	0.6238
Quercetin	NO	YES	NO	YES	NO	YES	YES	YES	NO	1.164	0.7348
p-Coumaric acid	NO	NO	NO	NO	YES	NO	YES	NO	YES	0.509	0.4898
Vanillic acid	NO	NO	NO	NO	NO	NO	YES	NO	NO	0.584	0.4923
Gallic acid	NO	NO	NO	YES	YES	NO	YES	YES	NO	0.506	0.6904
Isoscopoletin	NO	NO	NO	NO	NO	NO	YES	NO	NO	0.779	0.8059
4-Methylcatechol	NO	NO	NO	YES	YES	NO	YES	NO	NO	0.626	0.5777

3.4. Naïve Bayes machine learning algorithm analysis.

Retrieved plant compounds are screened using Lipinski's Rule of Five, which provides standardized criteria that a ligand should pass to be suitable for drug design. It establishes criteria for drug-like qualities and focuses on medication bioavailability. The Naïve Bayes Machine Learning algorithm was applied to the dataset to classify the plant compounds into two groups (Yes=Suitable for drug designing, No= Not Suitable for Drug designing), which are screened using Lipinski's Rule of Five [38-40]. The Naïve Bayes applications allow all attributes to contribute to the final decision equally. This simplicity is equivalent to computational efficiency, which makes the Naïve Bayes technique attractive and suitable for various fields. Table 4 describes the detailed summary of the Naïve Bayes Classifier, and Table 5 provides the Confusion matrix. Using Cross-Validation with k-fold=10 shows the best accuracy produced by the Naïve Bayes Classifier, which reaches 93.33 %. Table 6 shows the classification results consisting of the value of accuracy, error, precision, and recall from the Naïve Bayes Classifier. In this study, it was also concluded that the Naïve Bayes Classifier algorithm could classify the plant compounds suitable for drug design or not with the optimal value of accuracy.

3.5. Principal component analysis (PCA).

After classifying the plant compounds into two groups with optimal accuracy, the most influencing plant compounds were identified using principal component Analysis (PCA) techniques. A PCA was performed with all 30 plant compounds with all the variables, and the loadings plot and the component matrix were used to interpret the dataset. Principal component analysis (PCA) contributed 99.90% of the total variance. The loading plot (Figure 1) shows the relationship between the PCs and the original variables.

Table 2. List of pharmacokinetics properties, molecular weight (MW), hydrogen bond donor (HBD), hydrogen bond acceptor (HBA), partitioning coefficient (LogP), number of rotatable bonds (RB), polar surface area (PSA), synthetic accessibility (SA), gastrointestinal (GI) absorption, and Lipinski's Rule of 5 of all plant compounds.

Bioactive Compound	PubChem ID	Molecular Weight (≤500)	Hydrogen bond donor (≤5)	Hydrogen bond acceptor (≤10)	LogP (≤5)	Rotatable bond (≤10)	Polar surfaces area (<140Å ²)	Synthetic accessibility	Log S	Gastrointestinal absorption	Lipinski Rule
Neohesperidin	442439	610.6	8	15	2.57	7	234.29	Moderate (6.36)	-3.07 (Soluble)	Low	NO
Hesperetin 5-O-glucoside	18625123	464.4	6	11	1.88	5	175.37	Moderate (95.25)	-2.81 (Soluble)	Low	NO
Nobiletin	72344	402.4	0	8	3	7	85.59	Easy (3.90)	-4.18 (Moderate soluble)	High	YES
Diosmin	5281613	608.5	8	15	3.05	7	238.20	Moderate (6.48)	-3.51 (Soluble)	Low	NO
Avicularin	5490064	434.3	7	11	1.86	4	190.28	Moderate (5.04)	-3.27 (Soluble)	Low	NO
Nicotiflorin	5318767	594.5	9	15	2.79	6	249.20	Moderate (6.48)	-3.42 (Moderate soluble)	Low	NO
Isotrifoliin	5280804	464.4	9	12	0.94	4	210.51	Moderate (5.32)	-3.04 (Soluble)	Low	NO
Biorobin	12313332	594.5	9	15	2.79	6	249.20	Moderate (6.48)	-3.42 (Moderate soluble)	Low	NO
Spiraeoside	5320844	464.4	8	12	1.45	4	210.51	Moderate (5.23)	-3.64 (Soluble)	Low	NO
L-Epicatechin	72276	290.27	5	6	1.47	1	110.38	Easy (3.50)	-2.22 (Soluble)	High	YES
Piperidine	8082	85.15	1	1	1.70	0	12.03	Easy (1.00)	-0.90 (Soluble)	Low	YES
α-Terpineol	442501	154.25	1	1	2.51	1	20.23	Easy (3.24)	-2.87 (Soluble)	High	YES
LysoPC 18:1	53480465	521.7	1	7	0.59	25	114.93	Moderate (5.94)	-5.07 (Moderate soluble)	Low	NO
O-Phosphocholine	1014	184.15	2	4	-2.54	4	76.57	Easy (2.94)	0.23 (Highly soluble)	High	YES
Betaine	247	117.15	0	2	-2.19	2	40.13	Easy (1.00)	-0.35 (Very soluble)	Low	YES
Ellagic Acid	5281855	302.19	4	8	0.79	0	141.34	Easy (3.17)	-2.94 (Soluble)	High	YES

Bioactive Compound	PubChem ID	Molecular Weight (≤ 500)	Hydrogen bond donor (≤ 5)	Hydrogen bond acceptor (≤ 10)	LogP (≤ 5)	Rotatable bond (≤ 10)	Polar surfaces area ($< 140\text{\AA}^2$)	Synthetic accessibility	Log S	Gastrointestinal absorption	Lipinski Rule
Procyanidin A2	124025	576.5	9	12	1.80	2	209.76	Moderate (5.85)	-5.21 (Moderate soluble)	Low	NO
L-Glutamic Acid	33032	147.13	3	5	0.41	4	100.62	Easy (1.81)	1.84 (Highly soluble)	High	YES
L-Aspartic Acid	5960	133.1	3	5	-0.14	3	100.62	Easy (1.80)	1.98 (Highly soluble)	High	YES
Citric Acid	311	192.12	4	7	-1.49	5	132.13	Easy (2.18)	0.38 (Highly soluble)	Low	YES
Kaempferol	6325460	610.52	10	16	2.34	7	269.43	Moderate (6.56)	-3.31 (Soluble)	Low	YES
Quercetin	5280343	302.24	5	7	1.63	1	131.36	3.23 (Easy)	-3.16 (Soluble)	High	YES
Flavogallonic acid	14503023	470.30	8	13	0.58	2	239.33	3.73 (Easy)	-3.85 (Soluble)	Low	NO
p-Coumaric acid	637542	164.16	2	3	0.95	2	57.53	1.61 (Easy)	-2.02 (Soluble)	High	YES
Corilagin	73568	634.45	11	18	2.03	3	310.66	6.66 (Moderate)	-3.92 (Soluble)	Low	No
Vanillic acid	277423	168.15	2	4	1.40	2	667.76	1.42 (Easy)	-2.02 (Soluble)	High	Yes
Gallic acid	370	170.12	4	5	0.21	1	97.99	1.22 (Easy)	-1.64 (Very soluble)	High	Yes
Isoscopoletin	69894	192.17	1	4	1.95	1	59.67	2.65 (Easy)	-2.36 (soluble)	High	Yes
Tannin	44144428	952.69	15	26	0.33	16	452.02	6.46 (Moderate)	-1.97 (Very soluble)	Low	NO
4-Methylcatechol	9958	124.14	2	2	1.39	0	40.46	1.00 (Easy)	-1.97 (Very soluble)	High	Yes

4. Conclusions

The results of the testing analysis showed naïve Bayes had a stable accuracy after being tested with an accuracy value of 93.33%. It helps to classify the plant compounds suitable for drug design or not precisely. The principal Component Analysis method identifies the most influencing plant compounds. According to this study, α -terpineol (PubChem ID: 442501) may be the safest and most efficient drug for treating cancer. This *in silico* analysis of the anticancer properties of α -terpineol will contribute to developing a novel and potent drug for cancer treatment.

Funding

This research received no external funding.

Acknowledgments

This work was supported by the Biotechnology Department, Faculty of Applied Science, Lincoln University College, Malaysia.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. World Health Organization (WHO). **2022**. Available: <https://www.who.int/news-room/fact-sheets/detail/cancer> (accessed on 21 May 2022).
2. World Health Organization (WHO). **2020**. Available: <https://www.iarc.who.int/news-events/latest-global-cancer-data-cancer-burden-rises-to-19-3-million-new-cases-and-10-0-million-cancer-deaths-in-2020/>. (accessed on 20 May 2022).
3. World Health Organization (WHO). **2018a**. Available: https://www.iarc.who.int/wp-content/uploads/2018/09/pr263_E.pdf. (accessed on 20 May 2022).
4. Watkins, E.J. Overview of breast cancer. *J. Am. Acad. Phys. Assistants* **2019**, *32*, 13-17, <https://doi.org/10.1097/01.JAA.0000580524.95733.3d>.
5. Cheikh, A.; El Majjaoui, S.; Ismaili, N.; Cheikh, Z.; Bouajaj, J.; Nejjari, C.; El Hassani, A.; Cherrah, Y.; Benjaafar, N. Evaluation of the cost of cervical cancer at the National Institute of Oncology, Rabat. *Pan. Afr. Med. J.* **2016**, *23*, <https://doi.org/10.11604/pamj.2016.23.209.7750>.
6. Cao, W.; Chen, H.-D.; Yu, Y.-W.; Li, N.; Chen, W.-Q. Changing profiles of cancer burden worldwide and in China: a secondary analysis of the global cancer statistics 2020. *Chin. Med. J.* **2021**, *134*, 783-791, <https://doi.org/10.1097/CM9.0000000000001474>.
7. World Health Organization (WHO). **2018b**. Available: <https://apps.who.int/iris/bitstream/handle/10665/274603/9789241565639-eng.pdf>. (accessed on 16 May 2022).
8. American Medical Association. **2018**. Available: https://journalofethics.ama-assn.org/sites/journalofethics.ama-assn.org/files/2020-01/joe-2002_3.pdf. (accessed on 17 May 2022).
9. Jiang, W.G.; Sanders, A.J.; Katoh, M.; Ungefroren, H.; Gieseler, F.; Prince, M.; Thompson, S.K.; Zollo, M.; Spano, D.; Dhawan, P.; Sliva, D.; Subbarayan, P.R.; Sarkar, M.; Honoki, K.; Fujii, H.; Georgakilas, A.G.; Amedei, A.; Niccolai, E.; Amin, A.; Ashraf, S.S.; Ye, L.; Helferich, W.G.; Yang, X.; Boosani, C.S.; Guha, G.; Ciriolo, M.R.; Aquilano, K.; Chen, S.; Azmi, A.S.; Keith, W.N.; Bilsland, A.; Bhakta, D.; Halicka, D.; Nowsheen, S.; Pantano, F.; Santini, D. Tissue invasion and metastasis: Molecular, biological and clinical perspectives. *Semin. Cancer Biol.* **2015**, *35*, S244-S275, <https://doi.org/10.1016/j.semcancer.2015.03.008>.
10. Parsa, N. Environmental Factors Inducing Human Cancers. *Iran J. Public Health* **2012**, *41*, 1-9.
11. Saini, A.; Kumar, M.; Bhatt, S.; Saini, V.; Malik, A. Cancer causes and treatments. *Int. J. Pharm. Sci. Res.* **2020**, *11*, 3121-3134, [https://doi.org/10.13040/IJPSR.0975-8232.11\(7\).3121-34](https://doi.org/10.13040/IJPSR.0975-8232.11(7).3121-34).

12. Chinnadurai, R.K.; Khan, N.; Meghwanshi, G.K.; Ponne, S.; Althobiti, M.; Kumar, R. Current research status of anticancer peptides: Mechanism of action, production, and clinical applications. *Biomed. Pharmacother.* **2023**, *164*, 114996, <https://doi.org/10.1016/j.biopha.2023.114996>.
13. Garg, P.; Awasthi, S.; Horne, D.; Salgia, R.; Singhal, S.S. The innate effects of plant secondary metabolites in preclusion of gynecologic cancers: Inflammatory response and therapeutic action. *Biochim. Biophys. Acta Rev. Cancer* **2023**, *1878*, 188929, <https://doi.org/10.1016/j.bbcan.2023.188929>.
14. Elengoe, A.; Sundramoorthy, N.D. Molecular Docking of Curcumin With Breast Cancer Cell Line Proteins. *Pharm. Biomed. Res.* **2020**, *6*, 27-36, <http://doi.org/10.18502/pbr.v6i1.3425>.
15. Mokra, D.; Joskova, M.; Mokry, J. Therapeutic Effects of Green Tea Polyphenol (–)-Epigallocatechin-3-Gallate (EGCG) in Relation to Molecular Pathways Controlling Inflammation, Oxidative Stress, and Apoptosis. *Int. J. Mol. Sci.* **2023**, *24*, 340, <https://doi.org/10.3390/ijms24010340>.
16. Alam, A.; Biswas, M.; Zahid, A.; Ahmed, T.; Kundu, G.K.; Biswas, B.; Hasan, M.K. Effects of Different Solvents and their Purity on the Extraction of Total Phenolic Content, Total Flavonoid Content and Antioxidant Activity from the Peels of Lotkon (*Baccaurea Motleyana* Müll. Arg.) and Longan (*Dimocarpus Longan* Lour.). *Asian J. Food Res. Nutr.* **2023**, *2*, 14-21.
17. Elengoe, A.; Suhaibun, S.R. Dimocarpus longan phytochemicals possess anticancer activity by specifically targeting breast cancer biomarkers via computational biology tools. *Int. J. Health Sci.* **2022**, *6*, 14389-14409, <https://doi.org/10.53730/ijhs.v6nS2.8780>.
18. ke, Z.; Tan, S.; Shi, S. Physicochemical characteristics, polyphenols and antioxidant activities of *Dimocarpus longan* grown in different geographical locations. *Anal. Sci.* **2023**, *39*, 1405-1412, <https://doi.org/10.1007/s44211-023-00352-2>.
19. Kohli, R.K.; Santoshi, S.; Yadav, S.S.; Chauhan, V. Applications of AI in computer-Aided Drug Discovery. In Applying AI-Based IoT Systems to Simulation-Based Information Retrieval, IGI Global, **2023**, 77-89.
20. Yasir, M.; Tripathi, A.S.; Tripathi, M.K.; Shukla, P.; Maurya R.K. Approaches and Antiviral Drug Discovery. In CADD and Informatics in Drug Discovery. Interdisciplinary Biotechnological Advances, Rudrapal, M.; Khan, J., Eds.; Springer, Singapore. **2023**, 313-334 https://doi.org/10.1007/978-981-99-1316-9_13.
21. Choudhuri, S.; Yendluri, M.; Poddar, S.; Li, A, Mallick, K.; Mallik, S.; Ghosh, B. Recent Advancements in Computational Drug Design Algorithms through Machine Learning and Optimization. *Kinases Phosphatases* **2023**, *1*, 117-140, <https://doi.org/10.3390/kinasesphosphatases1020008>.
22. Bassani, D.; Moro, S. Past, Present, and Future Perspectives on Computer-Aided Drug Design Methodologies. *Molecules* **2023**, *28*, 3906, <https://doi.org/10.3390/molecules28093906>.
23. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E.E. PubChem 2023 update. *Nucleic Acids Res.* **2023**, *51*, D1373-D1380, <https://doi.org/10.1093/nar/gkac956>.
24. Lipinski, C.A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today: Technol.* **2004**, *1*, 337-341, <https://doi.org/10.1016/j.ddtec.2004.11.007>.
25. Veber, D.F.; Johnson, S.R.; Cheng, H.-Y.; Smith, B.R.; Ward, K.W.; Kopple, K.D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45*, 2615-2623, <https://doi.org/10.1021/jm020017n>.
26. Jagtap, N.M.; Yadav, A.R.; Mohite, S.K. Synthesis, Molecular Docking Studies and Anticancer Activity of 1,3,4-Oxadiazole-3(2H)-thione Derivatives. *J. University Shanghai Sci. Technol.* **2020**, *22*, 535-550.
27. Rodrigues, J.; Hullatti, K.; Jalalpure, S.; Khanal, P. In-vitro Cytotoxicity and in silico Molecular Docking of Alkaloids from *Tiliacora acuminata*. *Indian J. Pharm. Educ. Res.* **2020**, *54*, s295-s300, <https://doi.org/10.5530/ijper.54.2s.86>.
28. Tantawy, E.S.; Amer, A.M.; Mohamed, E.K.; Alla, M.M.A.; Nafie, M.S. Synthesis, characterization of some pyrazine derivatives as anticancer agents: Invitro and in silico approaches. *J. Mol. Struct.* **2020**, *1210*, 128013, <https://doi.org/10.1016/j.molstruc.2020.128013>.
29. Grosdidier, A.; Zoete, V.; Michielin, O. SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Res.* **2011**, *39*, W270-W277, <https://doi.org/10.1093/nar/gkr366>.
30. Tkiwiriyah, A.A.D.; Elengoe, A.; Saqar, F.K. MOLECULAR DOCKING OF BIOACTIVE COMPOUNDS AGAINST p38, EGFR AND BCL-2 PROTEINS. *Eur. Chem. Bull.* **2023**, *12*, <https://doi.org/10.48047/ecb/2023.12.7.158>.
31. Yang, H.; Lou, C.; Sun, L.; Li, J., Cai, Y.; Wang, Z.; Li, W.; Liu, G.; Tang, Y. admetSAR 2.0: web-service for prediction and optimization of chemical ADMET properties. *Bioinformatics* **2019**, *35*, 1067-1069. <https://doi.org/10.1093/bioinformatics/bty707>.

32. Tian, S.; Wang, J.; Li, Y.; Li, D.; Xu, L.; Hou, T. The application of *in silico* drug-likeness predictions in pharmaceutical research. *Adv. Drug Deliv. Rev.* **2015**, *86*, 2-10, <https://doi.org/10.1016/j.addr.2015.01.009>.
33. Usha, T.; Goyal, A.K.; Lubna, S.; Prashanth, H.; Mohan, T.M.; Pande, V.; Middha, S.K. Identification of anticancer targets of eco-friendly waste Punica granatum peel by dual reverse virtual screening and binding analysis. *Asian Pac. J. Cancer Prev.* **2014**, *15*, 10345-10350, <https://doi.org/10.7314/apjcp.2014.15.23.10345>.
34. Xhemali, D.; Hinde, C.J.; Stone, R. Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages. *Int. J. Comput. Sci.* **2009**, *4*, 16-23.
35. Austin, A.; Colyson, J.H.; Rohmatulloh, F.G.; Destriani, W.; Rosyati, M.M. Potential of Indonesian Medicinal Plant Biodiversity as CHK1 Inhibitor Agent for Cancer Treatment by Bioinformatics and Computational Chemistry. *Indones. J. Comput. Biol.* **2023**, *2*, 1-10, <https://doi.org/10.24198/ijcb.v2i1.45150>.
36. Tariq, S.; Malik, A.; Landry, K.B.; Malik, M.; Ajnum, H.; Latief, N.; Malik, K.; Ijaz, B. *IN SILICO SCREENING OF COMPOUNDS DERIVED FROM TECTONA GRANDIS LEAVES AGAINST COVID-19 NSP12 AND NSP15 THROUGH MOLECULAR DOCKING APPROACH*. *Biol. Clin. Sci. Res. J.* **2023**, *1*, 285, <https://doi.org/10.54112/bcsrj.v2023i1.285>.
37. Suhaibun, S.R.; Elengoe, A.; Poddar, R. Technology Advance in Drug Design Using Computational Biology Tool. *Malaysian J. Med. Health Sci.* **2020**, *16*, 18-24.
38. Ramamurthy, K.; Thekkath, R.D.; Batra, S.; Chattopadhyay, S. A novel deep learning architecture for disease classification in Arabica coffee plants. *Concurr. Comput.: Pract. Exp.* **2023**, *35*, e7625, <https://doi.org/10.1002/cpe.7625>.
39. Carrillo, J.K.; Durán, C.M.; Cáceres, J.M.; Cuastumal, C.A.; Ferreira, J.; Ramos, J.; Bahder, B.; Oates, M.; Ruiz, A. Assessment of E-Senses Performance through Machine Learning Models for Colombian Herbal Teas Classification. *Chemosensors* **2023**, *11*, 354, <https://doi.org/10.3390/chemosensors11070354>.
40. Dalvi, P.; Kalbande, D.R. A Comprehensive Review of Plant Recognition Approaches: Techniques, Challenges, and Future Direction. *SSGM J. Sci. Eng.* **2023**, *1*, 1-5.