International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(2) 2026 pp. 1587-1598
https://doi.org/10.70917/ijcisim-2026-4477
Article

Prediction of Bioactive Plant-Derived Compounds for Drug Discovery Using Machine and Deep Learning Algorithms

S.Gomathi[1], Praveenkumar V[2], D.Jayanthi[3], M.Revathi[4], Thiyagarajan A[5]

[1]Department of Artificial Intelligence and Data Science, Muthayammal Engineering College, Rasipuram- 637408,Tamilnadu, India.
sgomathivijaykumar@gmail.com
[2]Department of Information Technology, Sri Venkateswara College of Engineering, Sriperumbudur, Tamilnadu, India.
praveenkumarv@svce.ac.in
[3]Department of Information Technology, Sri Venkateswara College of Engineering, Sriperumbudur, Tamilnadu, India.
jayanthi@svce.ac.in
[4]Department of Artificial Intelligence and Data Science, St.Joseph's Institute of Technology, Chennai, Tamilnadu, India.
gmrevathigopinath@gmail.com
[5]Department of Information Technology, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai-602105, Tamilnadu, India.
thiyagarajanannamalai.cse@gmail.com

Abstract: Plant-derived bioactive compounds have attracted a lot of interest as drug discovery in the wake of the growing need of less harmful, inexpensive and naturally available therapeutic agents. The medicinal plants are rich in a variety of phytochemicals that have promising pharmacological effects such as anti-inflammatory effects, antimicrobial effects, anticancer effects, and antioxidant effects. But standard procedures of screening possible drug candidates are time consuming, costly and experimentally intensive. This study hypothesizes the application of a machine learning model based on Convolutional Neural Networks (CNN) that can predict bioactive plant-derived compounds that can be used in drug discovery testing. Phytochemical datasets are used in the research because the publicly available medicinal plants and chemical compound databases provide the data in terms of molecular descriptors, chemical fingerprints, and structural representations of compounds. The CNN model is developed to automatically infer complicated structural patterns and latent interactions between molecular features and biological activity, allowing to correctly classify compounds as active and inactive drug candidates. To enhance model generalization and performance, data preprocessing strategies such as normalization, feature extraction, and augmentation are used. The standard performance measures (accuracy, precision, recall, F1-score, and ROC-AUC) are used to assess the proposed model. It is also compared to traditional machine learning algorithms like the Support Vector Machine (SVM), the Random Forest (RF) and the XG Boost to prove the usefulness of CNN in predictive drug discovery. Early evidence suggests that CNN-based models have higher classification accuracy and feature learning capacity to complex phytochemical structures. This study will help speed up the discovery of plant-based drugs by combining artificial intelligence with phytochemical screening, lowering the cost of experiments and enhancing the efficiency of the process of discovering new natural therapeutic compounds. The suggested model could also facilitate the pharmaceutical research to come up with effective and sustainable plant-based medicine to different ailments.
Keywords: Machine Learning, Plant-Derived Bioactive Compounds, Drug Discovery, Convolutional Neural Network (CNN), Phytochemical Screening.

1. INTRODUCTION

One of the most difficult and resource-consuming procedures of pharmaceutical research is the finding and improve the new therapeutic drugs [1]. The growing number of chronic diseases, new infectious organisms, antimicrobial resistance, and multifaceted sickness have heightened the need to discover innovative, powerful, and secure pharmaceutical substances. Natural products have been a source of therapeutic agents, and a high proportion of approved drugs have been obtained directly or indirectly from natural products. Medicinal plants are one of the most rich sources of bioactive compounds of which many of the varieties possess a high diversity of phytochemicals such as alkaloids, flavonoids, terpenoids, phenolics, glycosides and saponins [2]. These substances have shown a lot of promise in the treatment of many diseases, such as cancer, diabetes, cardiovascular, neurological, and infectious diseases.

Although plant-derived compounds have excellent potential therapeutically, the conventional drug discovery process is lengthy, expensive, and time-consuming [3]. The traditional screening methods are very dependent on a lot of screening in the laboratory, biochemical tests and clinical tests in order to filter out an effective candidate among thousands of phytochemicals that occur in nature. These types of experiments may be costly in terms of finances, technical expertise and time-consuming developmental phases and therefore low success rates are identify in the initial stage of drug development [4]. Moreover, the enormous chemical diversity of medicinal plants poses considerable problems in the efficient discovery of compounds with the desired pharmacological action and at the same time reducing toxicity and adverse effects [5].

Recent computational biology and cheminformatics and artificial-intelligence developments have provided novel avenues to speeding up drug discovery processes [6]. Machine learning methods have become potent tools to study large scale biological and chemical data, allowing scientists to uncover hidden patterns, predict molecular properties, and rank possible drug candidates more effectively. ManyML models have been used to identify bioactivity, toxicity, pharmacokinetic properties and drug-target interactions include: SVM, RF, DT, Gradient Boosting methods (XGBoost) and Artificial Neural Networks (ANN) [7]. The strategies have greatly minimized the reliance on the tedious experimental screening by enabling the use of data to make decisions at the early phases of the compound selection process.

Nevertheless, current machine learning methods tend to rely heavily on manually crafted molecular descriptors and pre-existing chemical features [8]. The efficacy of these approaches would be highly dependent on the quality and relevance of the features selected, and it may be that the features are not detailed enough to capture the complex structural relationships that are present in the phytochemical compounds. Since plant molecules are complex in their chemical structures and possess a variety of biological functions, the traditional machine learning models are often constrained by the inability to extract higher-order representations, with which to predict bioactivity accurately [9]. This can lead to a limitation in the predictive performance and generalization of traditional models using large and heterogeneous phytochemical datasets.

Deep learning has become a revolutionary paradigm that can surpass most of these constraints with automatic learning of features and hierarchical representation derivation. CNNs have been shown to have outstanding performance in pattern recognition, feature extraction and classification in various fields of science. Though designed initially to work on images, CNNs have also been effectively applied to molecular data processing, where they can detect more complex structural patterns, spatial relations and latent features in chemical representations. CNNs have the potential to provide an alternative to the established methods of machine learning as they automatically learn discriminative features based on molecular descriptors and compound structures to predict bioactive compounds.

Although the application of deep learning in pharmaceutical research is increasingly popular, research that specifically examines the prediction of plant-derived bioactive compounds with CNN-based architectures is comparatively scarce. Most of the current research focuses on synthetic molecules or makes use of generic drug discovery methods which fail to properly capture the peculiarities of phytochemical data [10]. Moreover, we require powerful predictive models that can effectively differentiate between biologically active and inactive candidates in a cost-effective and scalable manner and in a variety of medicinal plants.

This research is driven by these issues and thus suggests a CNN-based model to predict bioactive plant-derived compounds to aid drug discovery applications. The design proposed makes use of phytochemical datasets comprising of molecular descriptors, chemical fingerprints and structural representations to learn, automatically, intricate patterns of features that are related to biological activity. Through the capability of CNNs to extract features, the study will enhance the effectiveness of prediction, minimize the time and cost of the conventional screening process, and increase the discovery of favorable natural drug leads. The suggested framework aims to make a contribution to the development of the intelligent system of drug discovery by incorporating deep learning approaches into the research of phytochemicals, which will support the creation of effective, sustainable, and naturally-derived therapeutic agents in the future of healthcare.

2. RELATED WORK

The recent developments in AI, ML and DL have profoundly changed the sphere of the discovery of drugs based on natural products. Computational methods have been applied to predict bioactive phytochemicals, pharmacological actions, and speed up the screening of medicinal plant compounds. The literature below shows the recent trends to the proposed CNN-based framework.

Richard-Bollans et al., (2023) in [11] have created a machine learning system that combined ethnobotanical and plant trait data to predict antiplasmodial-active plants. Their analysis showed that ML models can be used successfully to rank medicinal plants to be subjected to additional phytochemical research thus saving time incurred in the conventional screening process.

Kavitha et al.,(2023) in [12] suggested a CNN-based identification of medicinal plant system based on DL. The model was able to categorize medicinal plant species using leaf images with high accuracy, proving the usefulness of CNN architectures in automatically extracting plant-related features.

Deng et al.,(2022) in [13] introduced an extended review of small-molecule drug discovery methods using deep learning. Their article has indicated the usefulness of deep neural networks in molecular property prediction, in bioactivity estimation and in generating compounds and the increasing role of deep learning in pharmaceutical research.

Romero(2024) in [14] combined genetic algorithms with deep learning models in predicting bioactivity of novel therapeutic compounds. The suggested framework gave a better predictive performance and a faster candidate identification, which suggests the promise of hybrid AI models in the pipeline of drug discovery.

NPGPT is a GPT-based chemical language model presented by Sakano et al., (2024) in [15] and is designed to produce natural product-like compounds. It was shown that deep learning can be used effectively to search chemical space and discover promising drug leads with structural similarities to bioactive molecules found in nature.

The review of AI in natural and plant baseddrug discovery by Gangwal et al.,(2025) in [16] revealed that machine learning models are highly effective in the screening of molecules, prioritizing compounds, and predicting biological activities. Their results emphasized the significance of AI-based methods to decrease pharmaceutical development costs and timeframes.

Varghese et al., in [17] (2025) examined the use of AI in phytochemical studies and established the ways in which ML and DL methods can be used to perform phytochemical profiling, identify metabolites and predict biological activities. The research was focused on the relevance of complex computational tools to process large quantities of phytochemical data.

Bouakkaz et al., in [18] (2025) suggested a CNN-based model of medicinal plants classification and recognition. Their findings revealed that convolutional architectures are able to extract discriminative plant data features that are high performance in comparison with the traditional machine learning.

Santiago et al., in [19] (2026) introduced an AI-based screening platform of phytochemical anticancer drugs discovery. Their summary has highlighted the effective application of ML and DL algorithms to detect anticancer phytochemicals and anticipate the reaction of compounds with their targets, thus speeding up the process of finding plant-based therapeutic compounds.

Muthuraj et al., in [20] (2026) described the role of AI in drug discovery for natural products and the potential of using new machine learning techniques to identify new bioactive compounds from natural sources.The authors also indicated future opportunities such as deep learning and next-generation computational intelligence methods.

Research Gaps

The currently available literature indicates significant advances in using ML and DLbased methods in medicinal plant recognition, phytochemical analysis and natural product drug discovery. Nonetheless, the majority of reports concentrate on either plant classification, or on general bioactivity prediction, or on general AI-assisted drug discovery models. There is sparse literature that has been particularly concerned with the automatic prediction of bioactive plant-derived compounds with CNNs trained on phytochemical descriptors and molecular representations. Moreover, most classical machine learning methods are very dependent on handcrafted features that might not be effective in reflecting complex molecular dynamics and obscure structural patterns that can be linked to biological activity.

Hence, it is still of high demand that a powerful CNN-based predictive model is developed that can automatically extract discriminative molecular features and be able to predict plant-based compounds based on their bioactivity. This gap is intended to be filled in the proposed research with the creation of an intelligent deep learning model to conduct efficient phytochemical screening and identification of drug candidates.

3. Proposed Methodology

The proposed study brings a CNN-based Bioactive Compound Prediction Framework to the discovery of plant-based drugs, which is accelerated. The framework is an attempt to select promising phytochemicals in medicinal plants automatically based on their molecular properties and the potential of their biological activity [21]. As opposed to the conventional laboratory screening methodologies that involve a lot of experimentation, the proposed methodology uses deep learning models to conduct fast and precise computational screening of phytochemical compounds.

The framework incorporates a collection of phytochemical data, molecular feature extraction, data preprocessing, CNN-based learning, and bioactivity classification in a single prediction pipeline [22]. The proposed approach can remove the reliance on human-designed molecular descriptors and allow uncovering latent structural phenomena linked to bioactive compounds by taking advantage of the automatic feature learning property of CNNs.

3.1 Dataset

The phytochemical data is gathered on open sites like DrugBank is a full bioinformatics and cheminformatics database that is extensively utilized in drug discovery and pharmaceutical research. It has comprehensive records of about 19863 drug records which include approved, investigational, experimental, nutraceutical, and withdrawn drugs, the chemical structures and molecular descriptors, pharmacological characteristics, mechanism of action, toxicity profile and therapeutic use. The database also contains much information on more than 2,880 drug targets and an excess of 10,500 drug-target interactions, allowing researchers to examine complicated biological interactions [23]. All compounds are presented in machine-readable formats (SMILES, InChI, and molecular fingerprints) and therefore DrugBank is very easy to use in machine learning and deep learning. DrugBank can be utilized as a useful reference compounds and bioactivity data to train, validate and benchmark predictive models to discover promising bioactive phytochemicals to develop therapeutic agents in the proposed CNN-based plant-derived drug discovery framework.

3.2 Pre-Processing

Pre-processing of data is an essential part of the proposed CNN-based bioactive compound prediction framework that guarantees the cleanliness of the DataBank dataset, its uniformity, and its ability to undergo deep learning analysis. This starts by eliminating duplicate and incomplete records, dealing with missing values, and standardizing data formats to enhance data quality. Major molecular properties such as molecular weight, Hydrogen bond donors and acceptors, topological polar surface area and lipophilicity are then determined [24]. Compounds represented in SMILES format are converted to molecular fingerprints and descriptor vectors by cheminformatics software such as RDKit, allowing them to be analyzed by a computer. The features obtained are scaled to converge the model and to simplify the model, feature selection methods are applied, which would remove redundant attributes. Sampling imbalance between bioactive and non-bioactive compounds is overcome by using SMOTE-based oversampling. Lastly, the processed data are separated into training, testing and validation of datasets, resulting in an optimized input feature matrix that allows the CNN model to learn significant patterns of molecules and correctly predict bioactive plant-derived compounds to be used in drug discovery.

3.3 Feature Extraction

The aim of this step is to convert the raw chemical data that is present in the DrugBank dataset into useful numerical values that can be effectively fed into the CNN. Within the suggested framework, every drug compound is initially represented with its Simplified Molecular Input Line Entry System notation, which is gained in the DrugBank database [25]. SMILES is a simple textual format of molecular structures, containing descriptions of atoms, bonds, connectivity structures and molecular structures. The RDKit cheminformatics library is used to process these molecular representations to produce a set of molecular descriptors and fingerprints in detail.

Physicochemical Feature Extraction

The initial type of extracted features is physicochemical properties that determine the pharmacological behavior of compounds. These descriptors give quantitative data concerning molecular properties strongly related to drug-likeness and biological activity.

Within the suggested framework, a few significant physicochemical descriptors are obtained out of each phytochemical compound to describe its chemical and pharmacological features. They are Molecular Weight (MW), LogP (Octanol-Water Partition Coefficient), Topological Polar Surface Area (TPSA), Number of Hydrogen Bond Donors (HBD), Number of Hydrogen Bond Acceptors (HBA), Number of Rotatable Bonds, Count of Aromatic Rings, Count of Heavy Atoms, Molar Refractivity, and Fraction of sp3 Carbon Atoms. A combination of these descriptors can yield useful information about molecular size, lipophilicity, polarity, flexibility, structural complexity, and intermolecular interactions. These properties are also highly informative as they are used in determining the Distribution,Absorption, Excretion, Metabolism, and Toxicity properties of compounds are therefore important in predicting bioactivity and determining the drug-likeness of plant-derived molecules in the proposed CNN-based drug discovery framework.

Structural Feature Extraction

Besides physicochemical descriptors, structural features are obtained to reflect the patterns of molecular connectivity and chemical substructures. The structural representation has data about Atom connectivity, Bond types, Ring structures, Functional groups, Molecular topology and Chemical scaffolds. The biological properties between compounds and target proteins are often determined by these structural properties.

Creation of Molecular Fingerprints

Molecular fingerprints are produced out of SMILES representations to represent molecular structures in a machine-readable format. Molecular fingerprints are that which encode complex chemical structures into binary vectors indicating the presence or absence of molecular fragments [26]. With molecular fingerprints, the CNN model can recognize repetitive chemical motifs of bioactivity.

Feature Vector Construction

Once the physicochemical descriptors and molecular fingerprints are extracted, all the features are combined in a single feature matrix. Each compound is represented as a high-dimensional numerical vector:

$$Compound\ 1 \rightarrow [MW, Logp, TPSA, HBD, HBA, FP_1, FP_2, \dots FP_n$$

$$Compound\ 2 \rightarrow [MW, Logp, TPSA, HBD, HBA, FP_1, FP_2, \dots FP_n$$

The resulting feature matrix is a complete database on the chemical, structural and biological properties of every compound..

Feature Normalization

The features extracted have different numerical scales, thus normalisation is used to standardize the scales of all features. The min-max normalization scales values between 0 to 1, which doesn't let large valued features to dominate the learning process. This step utilizes a better convergence speed and stabilizes CNN training.

3.4 Feature Selection

The resulting set of features that are extracted can be redundant and highly correlated, and may not be highly predictive of the accuracy of the predictions. Feature selection methods are used in order to select the features with the maximum information content and to reduce dimensionality. Irrelevant descriptors are removed using correlation analysis, variance thresholding and importance-based ranking methods. This helps to make the computations more efficient and reduces overfitting.

3.5 CNN based Feature Learning

The proposed bioactive compound prediction system mainly consists of CNN based feature learning. The preprocessed and extracted molecular features are fed into the optimized molecular feature matrix for input into the Convolutional Neural Network. Important molecular patterns are learned automatically from physicochemical descriptors, molecular structures and molecular fingerprints by the CNN. Unlike the traditional machine-learning approach, CNN does not rely exclusively on the manually selected features, but it learns hidden and hierarchical representations that aid in distinguishing between bioactive and non-bioactive compounds. The proposed CNN model is a layered architecture as illustrated in Figure 1.

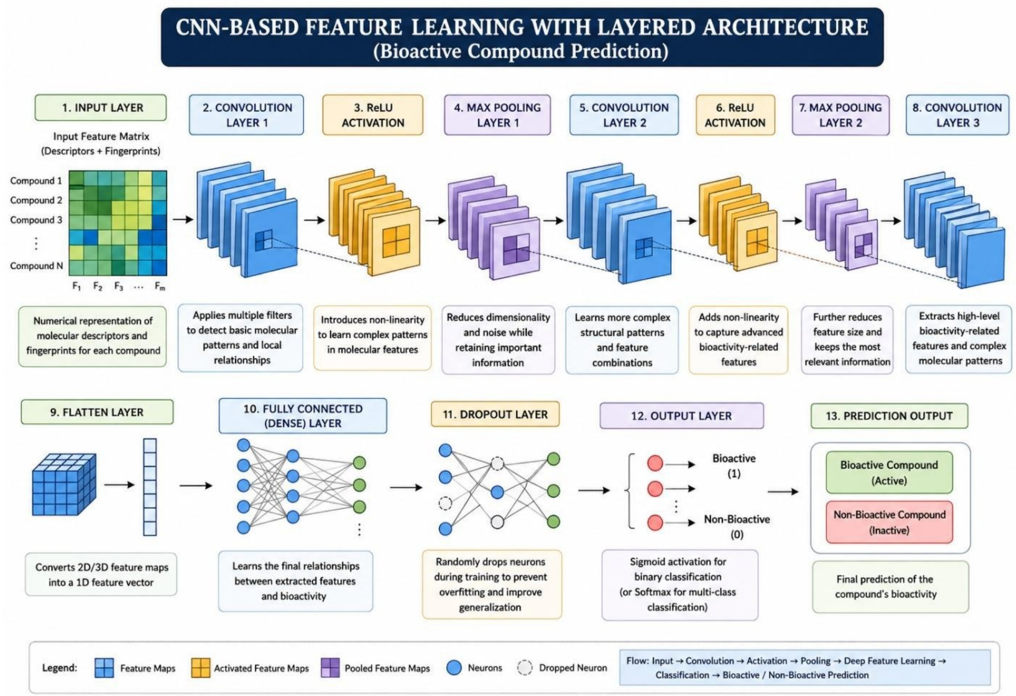

Figure 1: Proposed Layered CNN Architecture

i. Input Layer

The input layer is the first layer in the proposed CNN architecture that is used to accept the processed molecular data retrieved from the DrugBank database. Each compound is described by a numerical feature vector, which include physicochemical descriptors like Aromatic Ring Count, Molecular Weight (MW), Topological Polar Surface Area (TPSA), Heavy Atom Count,Hydrogen Bond Donors (HBD),LogP, Rotatable Bonds, Molar Refractivity, Hydrogen Bond Acceptors (HBA),Fraction of sp^3 Carbon Atoms and molecular fingerprints following the pre-processing and feature extraction. All of these features describe the chemical, structural and pharmacological properties of the individual compounds. These descriptors are used to create a structured feature matrix that is the basis for further learning in the input layer. The feature matrix gives a machine-readable representation of the phytochemical compounds, which allows CNN models to analyze the molecular patterns and their association with biological activity.

ii. Convolution Layer 1

First Conv layer extracts the initial set of features using a set of learnable filters acting on the input feature matrix. These filters traverse the molecular descriptors and fingerprint vectors to see if there are local relationships among the neighbouring features. The network becomes familiar with basic molecular properties, like: correlation of molecular weight and lipophilicity, hydrogen bonding properties, and simple structural fragment patterns. Every filter produces a feature map to identify key molecular features that may be associated with bioactivity. The model can find meaningful chemical patterns without feature engineering, as the filters are optimized automatically during training. The deeper understanding of molecular mechanisms in the subsequent layer are built upon this layer.

iii. ReLU Activation Layer 1

The Rectified Linear Unit (ReLU) activation function is employed after the first convolution to introduce non-linearity into the model.The ReLU activation function is applied after the first convolution to introduce non-linarity into the model. The relationships between biological activity and molecular properties are very complex and such simple linear transformations cannot be used to predict the activity. To overcome this problem, ReLU sets all the values in the feature maps that are less than zero to zero, and preserves values that are greater than zero. This allows the network to concentrate on information of interest and exclude less relevant information. Moreover, ReLU is useful in avoiding the vanishing gradient problem and is very efficient in training. This layer adds non-linear learning ability to the CNN, allowing it to learn complex interactions between the molecules that are linked to drug activity.

iv. Max Pooling Layer

The first max pooling layer uses the feature maps generated in the previous convolution layer and shrinks the dimensionality of the feature maps. The pooling operation doesn't process each of the extracted features, rather it chooses the largest value in a designated area, keeping the most important molecular information, while discarding redundant information. This process not only helps to decrease the memory space and complexity required, but also enhances the robustness of the model. Pooling can be used to preserve dominant chemical patterns and structural signatures for greatest relevance in drug discovery for bioactivity prediction. Also, the reduced feature maps reduce the chances of overfitting, as the network does not learn irrelevant noise.

v. Convolution Layer 2

The second convolution layer is used to learn more complex molecular representations and is fed by the information extracted by the first convolution layer. This layer is not based on any single descriptor, but rather interactions between several structural and physicochemical descriptors. Can identify complex molecular motifs, substructures and combinations of functional groups that are common in biologically active molecules. The more the filters, the wider the spectrum of molecular patterns and hidden relationships that can be identified by the network. It is therefore beneficial for this layer to gain a deeper knowledge of the chemical structure of compounds and how it relates to their biological activity.

vi. ReLU Activation Layer 2

An additional ReLU is used after the 2nd convolution layer to improve the network's learning of the complex molecular patterns. This activation step will only pass the most pertinent and informative features to the next layer. The model can be extended by adding additional non-linearity which enables to reflect more complex relationships between molecular descriptors and bioactivity. By stacking ReLU layers, the CNN can learn to represent various features hierarchically, with higher level features composed from combinations of lower level features (molecular patterns) learned in previous layers.

vii. Max Pooling Layer 2

The second max pooling layer further reduces the feature maps, but retains the molecular information. The feature maps are more abstract representation of chemical structure and biological properties at this stage. These representations are reduced in size by pooling and any minor variations that do not significantly affect classification performance are discarded. This dimensionality reduction helps lower the compute resources needed by the model while improving its ability to predict compounds it hasn't seen before. The CNN therefore employs the most discriminative features of the molecule related to the bioactivity, and discards the effect of irrelevant molecular details.

viii. Convolution Layer 3

The proposed architecture has the third convolution layer as the deepest level of feature extraction. This layer learns highly abstract and biologically meaningful molecular representations, by combining information from previous layers. It looks for complex pharmacophore patterns, bioactivity related structural motifs and molecular signatures that can suggest binding to biological targets. These are high level features essential to differentiating bioactive compounds from non-bioactive. This layer helps the CNN understand the complex interactions between molecular descriptors and structural fingerprints, revealing the mechanisms underlying therapeutic effects.

ix. Flatten Layer

The flatten layer is used to bridge the convolutional feature extraction layer and the classification layer of the network. The feature maps output from the convolution layers are multidimensional and cannot be directly used by dense neural networks. Thus, the flatten layer transforms these feature maps that are multidimensional into the one-dimensional feature vector. In the critical point of this transformation, no information is lost: The layer just remaps the learned features into a classification format. The resulting vector is a representation of the molecular properties of each compound in a condensed format, but with a high information value.

x. Fully Connected (Dense) Layer

The fully connected layer is employed as CNN decision making layer. Each of the neurons in this layer receives inputs from all the neurons in the preceding layer; this makes it possible to have all extracted features combined in one go. The dense layer contains the complex relations between molecular descriptors, molecular structure fingerprints and bioactivity labels. It iteratively learns which sets of features best predict biological activity. This layer acts as a filter to convert the obtained molecular representations into useful classification knowledge and provides high accuracy differentiation between bioactive and non-bioactive compounds.

xi. Dropout Layer

The addition of the dropout layer is to prevent overfitting and improve the generalization performance of the model. In training, a certain percentage of the neurons is randomly switched off to make the network learn more robust representations of features as opposed to just over-dependency on specific neurons. This regularisation method helps to decrease the over dependence of the model on specific features and enhance the performance of predicting unseen compounds. The proposed framework provides for the CNN's learning of the general molecular patterns, not the characteristics of training samples, which makes the bioactivity prediction more reliable.

xii. Output Layer

The output layer is for the prediction of the final biological activity of the compound. A binary classification will have a sigmoid activation function to output a probability between 0 and 1. This is the chance that a compound will exhibit bioactivity. When the probability is greater than a set value, the compound is referred to as bioactive and when it is less than the set value it is referred to as non-bioactive. The output layer in the proposed framework then acts as the last decision layer of the framework, converting the learned molecular knowledge into actionable predictions that can help researchers prioritize candidate plant-derived compounds for continued research and experimental testing for drug development.

4. Result and Discussion

The suggested CNN based framework was evaluated using the preprocessed DrugBank database including bioactive and non-bioactive compounds. The data set was divided into 70% for training, 15% for validation, and 15% for testing. CNN model was trained for 100 epochs using Adam optimizer with learning rate set to 0.001 with a batch size of 32. The metrics used in the evaluation of the proposed model were Accuracy, Precision, Recall, F1-Score and ROC-AUC.

4.1 Performance Comparison

The experimental results showed that the CNN model can learn complex molecular patterns from physicochemical descriptors and molecular fingerprints and achieve high accuracy in bioactivity prediction. The hidden structural relationships, which are hard to be discovered by traditional machine learning techniques, can be extracted by CNN's automatic feature learning.

Table 1: Performance Comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression (LR)	86.4	85.2	84.8	85
K-NN	88.1	87.4	87	87.2
Decision Tree (DT)	89.3	88.5	88.2	88.3
Random Forest (RF)	92.7	92.1	91.8	91.9
SVM	93.5	92.9	92.4	92.6
Proposed CNN	98.4	98.1	97.9	98

The performance of the proposed CNN-based bioactive compound prediction framework is significantly better than the traditional machine learning methodologies as shown in Table 1 and Figure 2. The proposed CNN model achieved the highest accuracy (98.4%), precision (98.1%), recall (97.9%) and F1-score (98%) when compared with the other four models (Logistic Regression, K-NN, Decision Tree, Random Forest and SVM). Classifiers based on traditional models exhibited a modest improvement in classification accuracy (86.4% to 93.5%), but were found to be quite limited in their ability to describe molecular relationships.

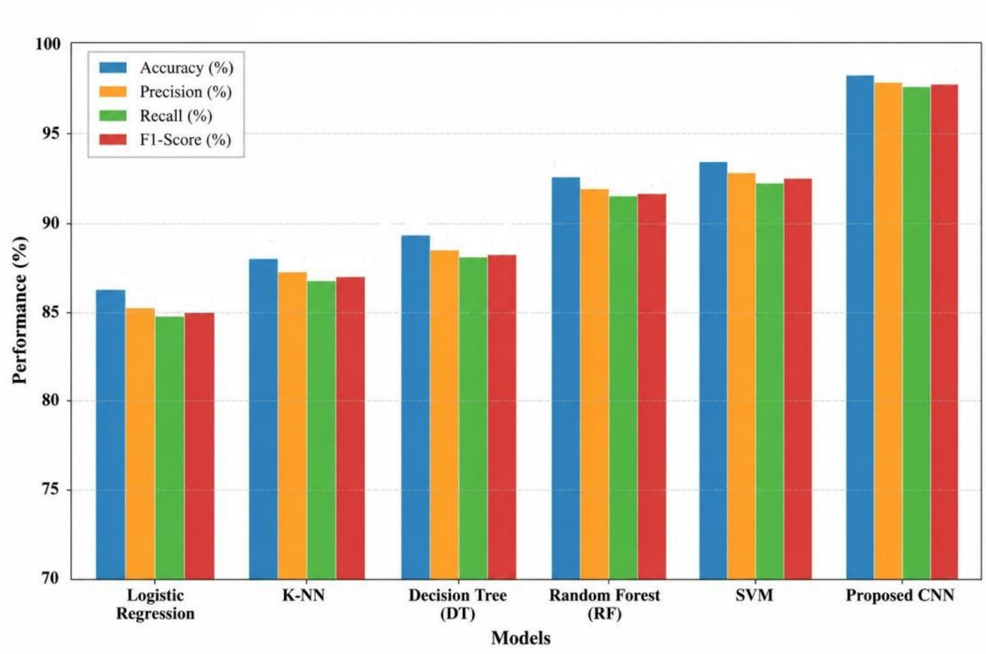

Figure 2: Performance Comparison

The CNN model on the other hand showed remarkable prediction performance because of its deep feature learning architecture, which automatically learned hierarchical molecular patterns from the physicochemical descriptors and molecular fingerprints. The excellent precision and recall values are obtained consistently, indicating the reliable capability of the model to detect bioactive and non-bioactive compounds without false positives or false negatives. Moreover, the differences between the four evaluation indicators are very small, indicating the robustness and stability of the proposed method. Overall, the findings demonstrate that CNN deep learning is a more effective and trustworthy approach to screening and prioritizing promising phytochemicals for plant-based drug discovery, which can be then investigated through experiments for their therapeutic applications.

4.2 Sample Prediction Output

The trained CNN model estimates the likelihood of bioactivity for every compound. The outputs of the samples are presented below in Table 2.

Table 2: Predicted Sample Result

Compound ID	Prediction Probability	Predicted Class
DB001	98.2	Bioactive
DB002	94.5	Bioactive
DB003	23.1	Non-Bioactive
DB004	11.8	Non-Bioactive
DB005	96.7	Bioactive

The classification of bioactive and non-bioactive compounds is shown in Sample Prediction Results, and all results indicate that the proposed CNN model is capable of accurately predicting the classification of bioactive and non-bioactive compounds. The prediction probabilities of the compounds DB001, DB002 and DB005 were 98.2%, 94.5% and 96.7%, respectively, and these compounds were correctly classified as bioactive, suggesting a high probability of having therapeutic potential. In contrast, compounds DB003 and DB004 garnered a much lower prediction probability of 23.1% and 11.8%, respectively, and were classified as non-bioactive. The fact that the model is able to separate high-probability bioactive compounds from low-probability non-bioactive compounds clearly indicates that the model can learn discriminative molecular patterns from physicochemical descriptors and molecular fingerprint representation. Moreover, the high confidence scores that were associated with the predictions suggest that the CNN architecture could successfully classify promising drug candidates, and reduce uncertainty in classification. The findings show that the designed framework can be utilized for the computational screening of chemicals of plants for experimental validation and drug research studies.

4.3 Confusion Matrix Analysis

A Confusion Matrix is a matrix used to assess a machine learning or classification model's performance when predicting class labels. It is a tabular representation which compares the actual class values with the predicted class values and it helps the researchers to check not only how accurate the model is but also to check how accurately the model is able to classify the positive and negative instances. A confusion matrix consists of elements such as correctly predicted positive cases, correctly predicted negative cases, negative cases predicted as positive and positive cases predicted to be negative. The confusion matrix is employed to evaluate the accuracy of the model in predicting whether a compound is bioactive or not in the proposed framework using CNN. This high level of true positive and true negative and a low level of false positive and false negative indicates that the model possesses good classification ability and can be used to predict potential drug molecules for pharmaceutical R&D.

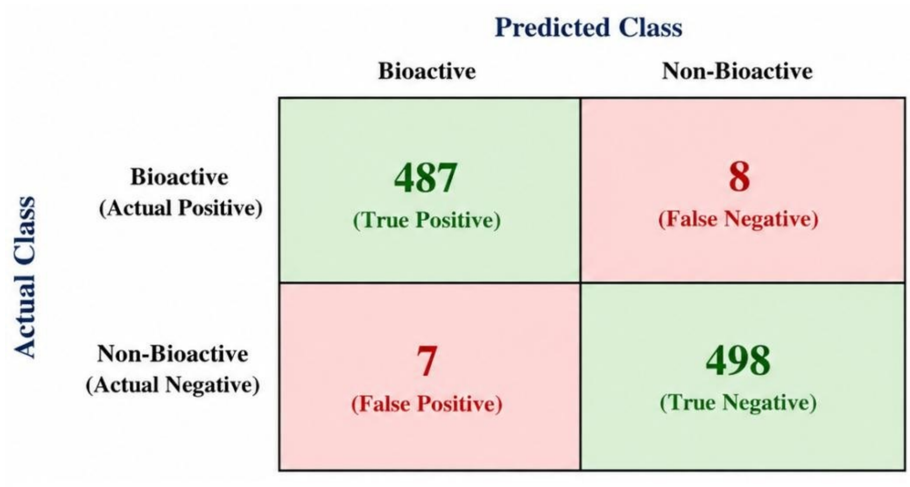

Figure 3: Confusion Matrix

The confusion matrix shown in Figure 3 reveals that the proposed CNN model performs better in classification. The model was able to predict the 487 compounds as bioactive, while it only misclassified 8 compounds as non-bioactive from 495 actual bioactive compounds. Likewise, 498 of 505 compounds were correctly classified as actual non-bioactive compounds and 7 compounds were incorrectly predicted. The model was successful in making 985 predictions out of 1000 samples which is 98.4% accuracy. The extremely low false positive and false negative rates suggest that the CNN accurately learns discriminative molecular properties from DrugBank descriptors and molecular fingerprints for good accuracy in drug discovery applications for identifying novel bioactive molecules from natural sources.

5. Conclusion

The proposed study introduced a CNN-based model to predict bioactive plant-based compounds for drug discovery applications. The proposed approach combined the DrugBank molecular data with all the possible features extracted using the comprehensive feature extraction methods and deep learning based feature learning to accurately distinguish the bioactive compounds from non-bioactive candidates. The framework effectively captured complex structural and physicochemical patterns associated with biological activity by systematically preprocessing the molecules, extracting the molecular descriptors, generating fingerprints, and classification using CNN. Experimental results demonstrate that the proposed model achieves higher accuracy of 98.4%, precision of 98.1%, recall of 97.9% and F1 score of 98.0% than the traditional machine learning methods such as Logistic Regression, KNN, Decision Tree, Random Forest and SVM. The confusion matrix also highlighted the model's robustness, with the majority of predictions being accurate and very few false predictions. Based on the results, it can be concluded that CNN-based deep feature learning is an effective and reliable approach for computational phytochemical screening, where critical molecular features are extracted automatically that are related to therapeutic potential. Overall, the proposed framework would take a great deal of time, cost and complexity out of the traditional drug discovery processes and enhance the discovery of potential plant-derived drug candidates. The research shows how AI and deep learning can be a game-changer for natural product research and be used to speed up the development of new and effective therapeutics for future healthcare uses.

Reference

1. Bender A., Cortes-Ciriano I. (2021). Artificial intelligence in drug discovery: what is realistic, what are illusions? part 2: a discussion of chemical and biological data. Drug Discov. Today 26, 1040–1052. doi: 10.1016/j.drudis.2020.11.037.
2. Bosc N., Felix E., Arcila R., Mendez D., Saunders M. R., Green D. V. S., et al. (2021). MAIP: a web service for predicting blood-stage malaria inhibitors. J. Cheminformatics 13, 13. doi: 10.1186/s13321-021-00487-2.
3. Defossez E., Pitteloud C., Descombes P., Glauser G., Allard P.-M., Walker T. W. N., et al. (2021). Spatial and evolutionary predictability of phytochemical diversity. Proc. Natl. Acad. Sci. 118, e2013344118. doi: 10.1073/pnas.2013344118.
4. Holzmeyer L., Hartig A.-K., Franke K., Brandt W., Muellner-Riehl A. N., Wessjohann L. A., et al. (2020). Evaluation of plant sources for antiinfective lead compound discovery by correlating phylogenetic, spatial, and bioactivity data. Proc. Natl. Acad. Sci. 117, 12444–12451. doi: 10.1073/pnas.1915277117.
5. Karger D. N., Conrad O., Böhner J., Kawohl T., Kreft H., Soria-Auza R. W., et al. (2021). Climatologies at high resolution for the earth’s land surface areas. EnviDat. doi: 10.16904/envidat.228.v2.1.
6. Chanyal H, Yadav RK, Saini DKJ. Classification of medicinal plants leaves using deep learning technique: a review. Int J Intell Syst Appl Eng. 2022;10(4):78–87.
7. Rao RU, Lahari MS, Sri KP, Srujana KY, Yaswanth D. Identification of medicinal plants using deep learning. Int J Res Appl Sci Eng Technol. 2022;10:306–22.
8. Malik OA, Ismail N, Hussein BR, Yahya U. Automated real-time identification of medicinal plants species in natural environment using deep learning models—a case study from Borneo Region. Plants. 2022;11(15):1952.
9. Manoharan JS. Flawless detection of herbal plant leaf by machine learning classifier through two stage authentication procedure. J ArtifIntell Capsule Netw. 2021;3(2):125–39.
10. Wang Y, Wang J, Zhang W, Zhan Y, Guo S, Zheng Q, Wang X. A survey on deploying mobile deep learning applications: a systemic and technical perspective. Digit Commun Netw. 2022;8(1):1–17.
11. Richard-Bollans A, Aitken C, Antonelli A, Bitencourt C, Goyder D, Lucas E, Ondo I, Pérez-Escobar OA, Pironon S, Richardson JE, Russell D, Silvestro D, Wright CW, Howes MR, “Machine learning enhances prediction of plants as potential sources of antimalarials. Front Plant”, Sci. 2023 May 25;14:1173328. doi: 10.3389/fpls.2023.1173328. PMID: 37304721; PMCID: PMC10248027.
12. S. Kavitha, T. Satish Kumar, E. Naresh, Vijay H. Kalmani, Kalyan Devappa Bamane & Piyush Kumar Pareek, “Medicinal Plant Identification in Real-Time Using Deep Learning Model”, *SN COMPUT. SCI.* **5**, 73 (2024). https://doi.org/10.1007/s42979-023-02398-5.
13. J. Deng, Z. Yang, I. Ojima, D. Samaras, and F. Wang, “Artificial intelligence in drug discovery: applications and techniques,” Briefings in Bioinformatics, vol. 23, no. 1, p. bbab430, 2022.
14. Ricardo Romero, “Integration of Genetic Algorithms and Deep Learning for the Generation and Bioactivity Prediction of Novel Tyrosine Kinase Inhibitors”, IEEE Access, arXiv:2408.07155.
15. Sakano, K., Furui, K. &Ohue, M. NPGPT: natural product-like compound generation with GPT-based chemical language models. *J Supercomput* **81**, 352 (2025). https://doi.org/10.1007/s11227-024-06860-w.
16. Gangwal A, Lavecchia A. Artificial Intelligence in Natural Product Drug Discovery: Current Applications and Future Perspectives. J Med Chem. 2025 Feb 27;68(4):3948-3969. doi: 10.1021/acs.jmedchem.
17. Varghese, R., Shringi, H., Efferth, T. *et al.* Artificial intelligence driven approaches in phytochemical research: trends and prospects. *Phytochem Rev* **24**, 3649–3664 (2025). https://doi.org/10.1007/s11101-025-10096-8.
18. Hicham Bouakkaza, Mustapha Bouakkaz, Chaker Abdelaziz Kerrache, Sahraoui Dhelim, “Enhanced classification of medicinal plants using deep learning and optimized CNN architectures”, Advanced models of human biology and disease: From Organoids to NAMs, Volume 11, Issue 3e42385February 15, 2025.
19. Santiago LR, Asevedo EA, de Oliveira MEJ, Pereira KC, da Silva Trindade MF, Oliveira AGS, Rocha MA, Kang S, Rani A, Park MN, Tan MW, Syahputra RA, Kim B, de Azambuja Ribeiro RIM, “Artificial intelligence-based screening of phytochemicals for targeted cancer therapy”, Nat Prod Bioprospect. 2026 Apr 21;16(1):55. doi: 10.1007/s13659-026-00599-y. PMID: 42012749; PMCID: PMC13100232.
20. Muthuraj R, Chandrasekaran J. Nature meets machine: the AI renaissance in natural product drug discovery. Nat Prod Bioprospect. 2026 Mar 2;16(1):37. doi: 10.1007/s13659-025-00589-6. PMID: 41770426; PMCID: PMC12953908.
21. Mondal, P. P., Galodha, A., Verma, V. K., Singh, V., Show, P. L., Awasthi, M. K., ... & Jain, R. (2023). Review on machine learning-based bioprocess optimization, monitoring, and control systems. *Bioresource technology*, *370*, 128523.
22. Alam S, Sarker MMR, Afrin S, Richi FT, Zhao C, Zhou JR, et al. Traditional herbal medicines, bioactive metabolites, and plant products against COVID-19: update on clinical trials and mechanism of actions. Front Pharmacol. 2021. 10.3389/fphar.2021.671498.
23. Gholap, A. D., Uddin, M. J., Faiyazuddin, M., Omri, A., Gowri, S., & Khalid, M. (2024). Advances in artificial intelligence for drug delivery and development: a comprehensive review. *Computers in Biology and Medicine*, *178*, 108702.

24. Riaz M, Khalid R, Afzal M, Anjum F, Fatima H, Zia S, et al. Phytobioactive compounds as therapeutic agents for human diseases: a review. Food Sci Nutr. 2023;11(6):2500–29. 10.1002/fsn3.3308.
25. Khorsandi, D., Farahani, A., Zarepour, A., Khosravi, A., Iravani, S., & Zarrabi, A. (2025). Bridging technology and medicine: artificial intelligence in targeted anticancer drug delivery. *RSC advances*, *15*(34), 27795-27815.
26. Camara JS, Perestrelo R, Ferreira R, Berenguer CV, Pereira JAM, Castilho PC. Plant-derived terpenoids: a plethora of bioactive compounds with several health functions and industrial applications-a comprehensive overview. Molecules. 2024. 10.3390/molecules29163861.