ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov1074

Machine Learning-Based Screening and Anticancer Activity Classification of Quinoline-Containing Natural Products Using RDKit

Prabeena Jelin A. P.[1,2]; Abbs Fen Reji T. F.[2*]

[1]Research Scholar (Reg. No.: 23213112032005), Affiliated to Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli 627012, Tamil Nadu, India.
[2]Department of Chemistry, Nesamony Memorial Christian College, Marthandam, Kanyakumari District, Tamil Nadu 629165, India.

Corresponding Author: Abbs Fen Reji T. F.[*]

Publication Date: 2026/11/26

Abstract: Natural products are important sources of bioactive molecules and potential anticancer agents. This study aimed to investigate the anticancer potential of quinoline-containing natural products using cheminformatics and machine-learning approaches. Initially, 52 compounds were collected from the PubChem database, and their molecular structures, SMILES notations, molecular weights and reported bioactivity data were retrieved. Experimentally reported IC_{50} values were used to classify compounds as active (≤10 µM) or inactive (>10 µM). Following data curation, nine compounds with consistent activity labels were selected, comprising four active and five inactive compounds. Molecular structures were converted into 1024-bit Morgan fingerprints using RDKit. Decision Tree, Random Forest and Logistic Regression algorithms were developed for activity classification. All three models reproduced the training-set labels, although independent validation was not performed. Random Forest feature-importance analysis identified potentially relevant molecular fingerprint features. Physicochemical properties, including molecular weight, LogP, topological polar surface area, hydrogen bond donors, hydrogen bond acceptors and rotatable bonds, were also compared. Nitidine showed the lowest reported IC_{50} value (0.40 µM) against A549 and PC-3 cells, while kokusaginine exhibited 4.62 µM activity against MCF-7/ADR cells. Overall, this study demonstrates a preliminary cheminformatics workflow for screening quinoline-containing natural products and highlights the need for larger datasets and independent model validation.

Keywords: *Quinoline-Containing Natural Products, Anticancer Activity, ML, RDkit, Morgan Molecular Fingerprints.*

How to Cite: Prabeena Jelin A. P.; Abbs Fen Reji T. F. (2026) Machine Learning-Based Screening and Anticancer Activity Classification of Quinoline-Containing Natural Products Using RDKit. *International Journal of Innovative Science and Research Technology,* 11(11), 1730-1744. https://doi.org/10.38124/ijisrt/26nov1074

I. INTRODUCTION

Natural products are an important source of biologically active compounds and have played a significant role in drug discovery and development. Many natural products contain heterocyclic ring systems, which contribute to their diverse pharmacological properties. Among these, quinoline-containing natural products have attracted considerable attention because of their reported anticancer, antimicrobial, antioxidant and other biological activities. The identification of potential anticancer compounds through conventional experimental methods can be time-consuming and expensive. Therefore, computational approaches such as cheminformatics and machine learning (ML) have become valuable tools for preliminary drug discovery and biological activity screening. Machine learning techniques can identify patterns between molecular structures and biological activities by analysing molecular descriptors and molecular fingerprints. Molecular fingerprints encode structural information into numerical representations that can be processed by machine learning algorithms. In this study, RDKit, an open-source cheminformatics toolkit, was used to generate Morgan fingerprints of selected quinoline-containing natural products. A dataset of 52 quinoline-containing natural products was initially collected from the PubChem database. Information including compound names, SMILES notations, molecular weights and reported biological activity values (IC_{50}) was collected where available. Experimentally reported cancer-cell IC_{50} values were used to classify compounds into Active and Inactive categories based on a provisional threshold of 10 µM. Three machine learning algorithms, namely Decision Tree, Random Forest and Logistic Regression, were employed to investigate the relationship between molecular structures and their

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov1074

activity classifications. Furthermore, Random Forest feature-importance analysis was performed to identify influential molecular fingerprint features. The corresponding structural fragments were examined using RDKit. Physicochemical properties, including molecular weight, lipophilicity (LogP), topological polar surface area (TPSA), hydrogen bond donors, hydrogen bond acceptors and rotatable bonds, were also compared between the Active and Inactive groups.

This investigation provides a preliminary computational screening approach for identifying promising quinoline-containing natural products with potential anticancer activity.

➢ *Objectives*

The main objectives of this study are:

- *Dataset Collection:*
  To collect quinoline-containing natural products from the PubChem database and retrieve their molecular information, including compound names, SMILES notations and molecular weights.

- *Bioactivity Data Collection:*
  To identify experimentally reported IC_{50} values for selected natural products from cancer-cell-based biological assays.

- *Molecular Fingerprint Generation:*
  To generate molecular fingerprints using RDKit, particularly Morgan fingerprints, for numerical representation of molecular structures.

- *Activity Classification:*
  To classify natural products into Active and Inactive categories based on their reported IC_{50} values using a provisional threshold of 10 µM.

- *Machine Learning Model Development:*
  To implement Decision Tree, Random Forest and Logistic Regression algorithms for preliminary classification of natural products.

- *Feature-Importance Analysis:*
  To identify important molecular fingerprint features influencing the classification results using the Random Forest algorithm.

- *Structural Interpretation:*
  To investigate the chemical fragments associated with important fingerprint features and explore possible structural patterns among active compounds.

- *Physicochemical Property Analysis:*
  To compare molecular weight, LogP, TPSA, HBD, HBA and rotatable bonds between Active and Inactive compounds.

- *Promising Compound Identification:*
  To identify two promising natural products based on their experimentally reported anticancer IC_{50} values and preliminary computational analysis.

II. MATERIALS AND METHODS

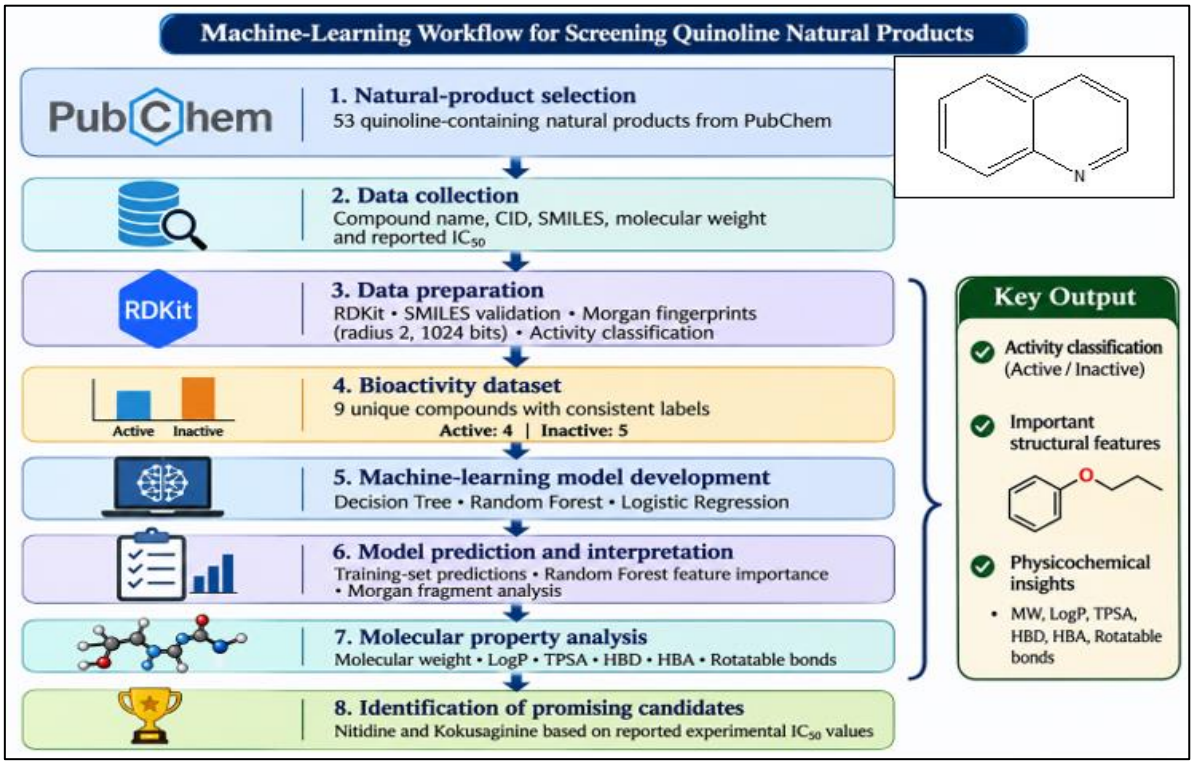


Fig 1 Machine-Learning Workflow for Screening Quinoline Natural Products

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov1074

➢ *Data Collection*

A dataset of 52 quinoline-containing natural products was initially selected for this study from the PubChem database (https://pubchem.ncbi.nlm.nih.gov/). The selected compounds were chosen based on the presence of a quinoline or related quinoline-containing structural framework.

For each compound, the following molecular information was collected:

- Compound name
- PubChem Compound Identification Number (CID)
- Canonical and isomeric SMILES
- Molecular formula
- Molecular weight
- Reported biological activity (IC_{50}), where available

The PubChem database was also explored to retrieve experimentally reported bioactivity information. Cancer-cell-based IC_{50} records were selected for preliminary anticancer activity classification.

➢ *Bioactivity Data Preparation*

The collected bioactivity records were examined to identify numerical IC_{50} values expressed in micromolar (μM). Records associated with cancer-cell-based assays were selected, while unrelated biological assays were excluded. Camptothecin was excluded from the natural-product training dataset because it is an established anticancer reference compound and its large number of assay records could dominate the dataset. After filtering, 20 cancer-cell IC_{50} observations representing 10 unique natural products were retained. These included Dictamnine, Flindersiamine, Gamma-Fagarine, Graveolinine, Kokusaginine, Maculine, Nitidine, Quinine, Skimmianine and Streptonigrin. The IC_{50} values were converted into binary activity labels using a provisional threshold of 10 μM:

- Active: $IC_{50} \leq 10$ μM (label = 1).
- Inactive: IC_{50} greater than 10 μM (label = 0).

The activity classification was initially performed at the observation level. Compounds exhibiting inconsistent classifications across different cancer cell lines were excluded from the initial compound-level machine-learning training dataset. Consequently, nine unique compounds with consistent activity labels were retained for the exploratory machine-learning analysis.

➢ *Molecular Fingerprint Generation Using RDKit*

RDKit, an open-source cheminformatics toolkit, was used to convert the molecular structures into numerical representations suitable for machine-learning analysis. The Canonical SMILES of each compound was converted into an RDKit molecular object. Morgan circular fingerprints were subsequently generated using the following parameters:

Table 1 Parameters Used for the Generation of Morgan Molecular Fingerprints Using RDKit.

Parameter	Value
Fingerprint type	Morgan fingerprint
Radius	2
Fingerprint size	1024 bits
Software	RDKit
Representation	Binary fingerprint

Morgan fingerprints encode local molecular environments around individual atoms. The resulting 1024-bit fingerprints were converted into numerical feature vectors and used as input variables for the machine-learning algorithms. The final feature matrix contained nine compounds and 1024 fingerprint features.

➢ *Machine-Learning Model Development*

Three supervised machine-learning classification algorithms were implemented using the Scikit-learn library in Python.

- *Decision Tree Classifier*

The Decision Tree algorithm was employed to classify natural products into Active and Inactive categories based on their molecular fingerprints. The model was configured with a maximum tree depth of 2 and a random state of 42.

- *Random Forest Classifier*

Random Forest is an ensemble learning algorithm that combines multiple decision trees to generate classification predictions. In this study, a Random Forest classifier containing 100 decision trees was developed. The maximum tree depth was restricted to 2, and the random state was set to 42. The algorithm was also used to estimate the relative importance of molecular fingerprint features for the classification task.

- *Logistic Regression*

Logistic Regression was employed as a binary classification algorithm to distinguish Active and Inactive compounds. Feature scaling was performed using the StandardScaler function before applying Logistic Regression. The maximum number of iterations was set to 1000 to facilitate model convergence.

➢ *Model Prediction and Feature-Importance Analysis*

The trained models were used to predict the activity classifications of the nine compounds in the prepared dataset. The predicted classifications were compared with the assigned experimental activity labels to examine the agreement between the three algorithms. Random Forest feature-importance scores were extracted to identify the most influential Morgan fingerprint bits. The ten highest-ranking fingerprint features were selected for further investigation. RDKit was subsequently used to map the selected fingerprint

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov1074

bits to their corresponding molecular fragments. This analysis provided preliminary structural insights into the molecular environments represented by the important fingerprint features.

➢ *Physicochemical Property Analysis*

The physicochemical properties of the selected natural products were evaluated to investigate possible relationships between molecular characteristics and activity classification. The following six molecular properties were considered:

Table 2 Physicochemical and Molecular Descriptor Parameters Used for Molecular Characterization of the Compounds.

Property	Description
Molecular Weight (MW)	Molecular mass of the compound
LogP	Partition coefficient representing lipophilicity
TPSA	Topological polar surface area
HBD	Number of hydrogen bond donors
HBA	Number of hydrogen bond acceptors
Rotatable Bonds	Number of rotatable single bonds

The average values of these properties were calculated separately for Active and Inactive compounds. Bar charts were generated using Matplotlib to visualize the differences between the two groups.

➢ *Identification of Promising Natural Products*

Promising natural products were identified primarily based on their experimentally reported cancer-cell IC_{50} values. Compounds exhibiting lower IC_{50} values were considered to have greater potency within the reported assay conditions. The preliminary computational classifications and molecular fingerprint analysis were also considered as supporting information. Nitidine and Kokusaginine were selected as two promising candidates based on their reported IC_{50} values of 0.40 µM and 4.62 µM, respectively. These values were obtained from different cancer-cell assay contexts and should not be interpreted as a direct comparative potency measurement.

➢ *Software and Computational Tools*

The computational investigation was performed using the following software and resources:

Table 3 Software Tools, Databases, and Computational Libraries Employed for Molecular Data Collection, Processing, Machine-Learning Model Development, and Visualization.

Software / Database	Application
PubChem	Natural-product and bioactivity data collection
Google Colab	Python-based computational environment
Python	Data processing and machine learning
RDKit	Molecular structure processing and fingerprint generation
Scikit-learn	Machine-learning model development
Pandas	Dataset processing and analysis
NumPy	Numerical operations
Matplotlib	Graphical representation of molecular properties

➢ *Important Methodological Limitations*

The machine-learning dataset was limited to nine unique compounds, which is insufficient for establishing a statistically reliable predictive classification model. The models were fitted and their predictions were examined using the same compounds used for training. Therefore, the observed agreement between predicted and assigned activity labels represents training-set agreement rather than independent predictive accuracy. Furthermore, the activity threshold of 10 µM was used as a provisional classification criterion, and differences in cancer cell lines and experimental assay conditions may influence the reported IC_{50} values. The results should therefore be considered exploratory.

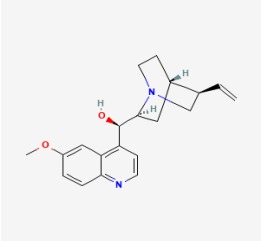
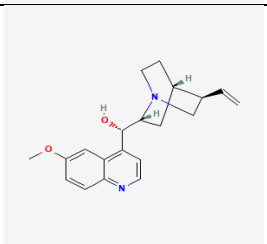
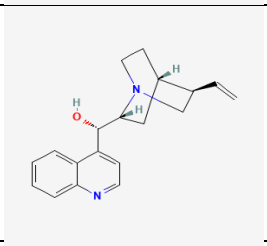
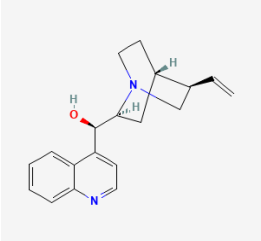
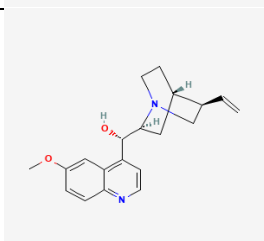
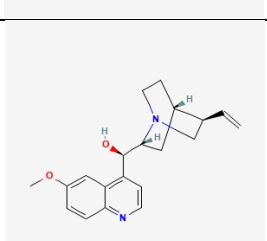
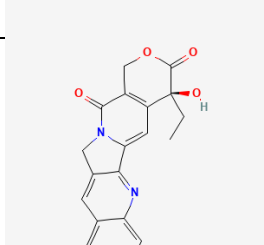
III. RESULTS AND DISCUSSION

➢ *Collection and Preparation of the Dataset*

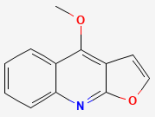
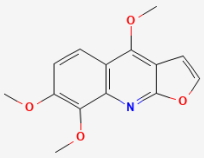
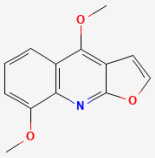
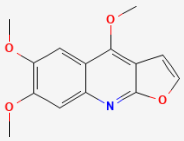
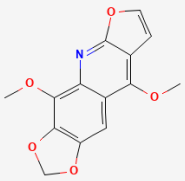
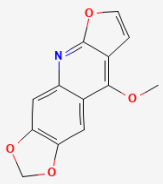
A preliminary library of 52 quinoline-containing natural products was compiled using PubChem. The dataset included compound names, PubChem identifiers, canonical SMILES, and molecular weights. Reported IC_{50} values were retrieved from PubChem bioassay records to identify compounds with experimentally reported anticancer activity. After filtering the assay records for cancer-cell-related IC_{50} measurements and excluding camptothecin, 20 experimental observations representing 10 unique natural products were obtained. The IC_{50} values were expressed in µM. Compounds with IC_{50} values less than or equal to 10 µM were classified as active, while compounds with values greater than 10 µM were classified as inactive. Kokusaginine showed different activity classifications across different cancer cell lines. Therefore, it was excluded from the initial structure-based model training dataset to maintain consistent compound-level labels. The final dataset used for model training consisted of nine unique natural products, comprising four active and five inactive compounds. The original list contained 50 candidate names, and additional compounds were retrieved during dataset expansion. I can recover most of the list, but I don't want to present an unverified list as the exact 52-compound dataset. Here are the 50 names from the original candidate list:

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov1074

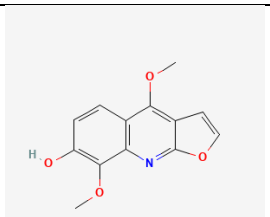
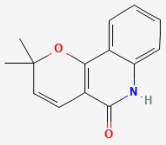
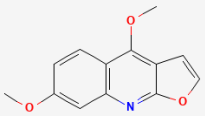
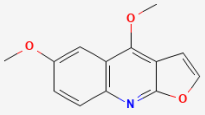
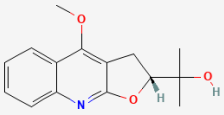
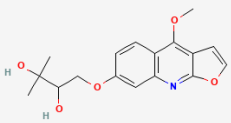
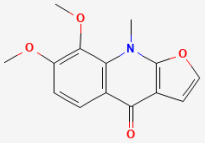
Table 4 Chemical Structures and Molecular Identification Details of the Selected Natural Compounds Retrieved from the PubChem Database.

No.	Compound Name	2D Chemical Structure	PubChem CID	Molecular Weight
1	Quinine		3034034	324.4
2	Quinidine		441074	324.4
3	Cinchonine		90454	294.4
4	Cinchonidine		101744	294.4
5	Epiquinine		10448938	324.4
6	Epiquinidine		94175	324.4
7	Camptothecin		24360	348.4

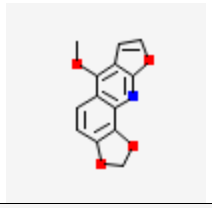
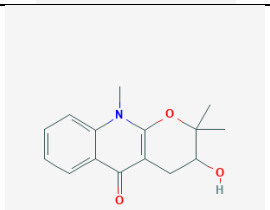
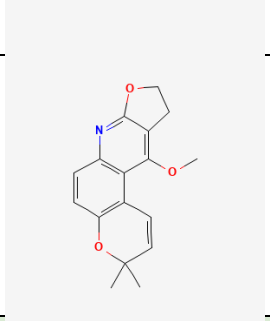
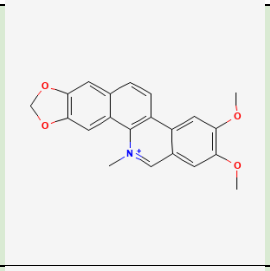
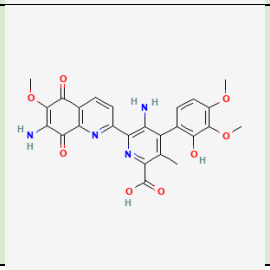
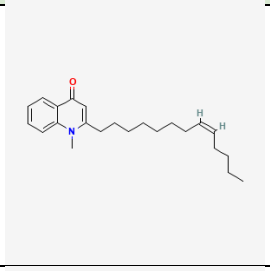
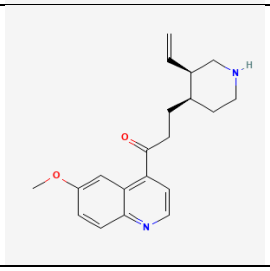
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov1074

8	Dictamnine		68085	199.20
9	Skimmianine		6760	259.26
10	Gamma-Fagarine		107936	229.23
11	Kokusaginine		10227	259.26
12	Flindersiamine		68221	273.24
13	Maculine		68232	243.21
14	Robustine		164950	215.20

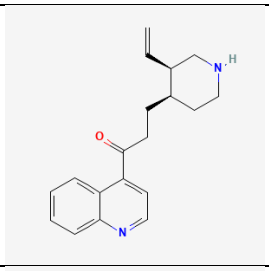
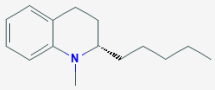
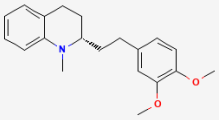
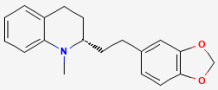
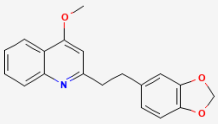
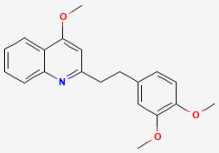
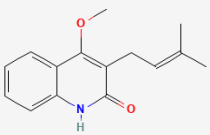
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov1074

15	Haplopine		5281846	245.23
16	Flindersine		68230	227.26
17	Evolitrine		196980	229.23
18	Pteleine		159650	229.23
19	Platydesmine		6451457	259.30
20	Evellerine		177139	317.34
21	Isoskimmianine		621199	259.26

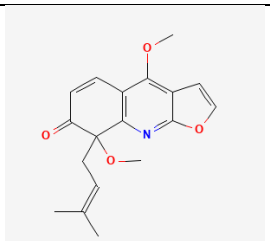
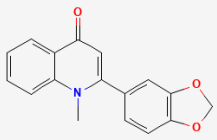
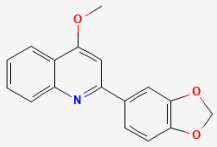
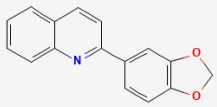
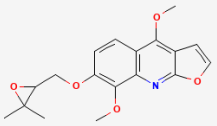
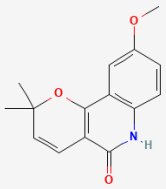
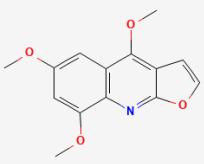
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov1074

22	kokusagine		5318829	243.21
23	Ribalinine		356086	259.30
24	Medicosmine		44559265	283.32
25	Nitidine		4501	348.4
26	Streptonigrin		5298	506.5
27	Evocarpine		5317303	339.5
28	Quinicine		65753	324.4

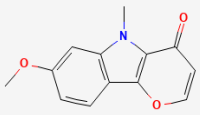
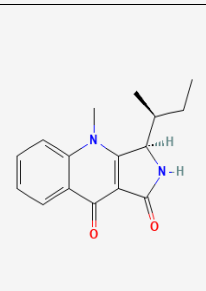
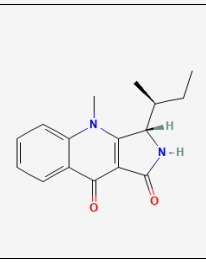
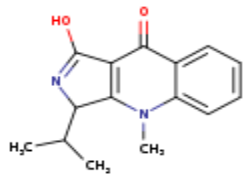
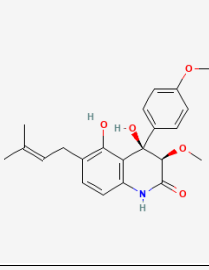
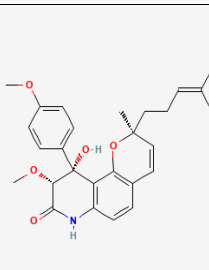
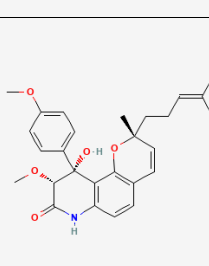
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov1074

29	Cinchonicine		3053309	294.4
30	Angustureine		15606170	217.35
31	Cuspareine		44138453	311.4
32	Galipinine		91749530	295.4
33	Cusparine		442893	307.3
34	Galipine		68235	323.4
35	Atanine		10014667	243.30

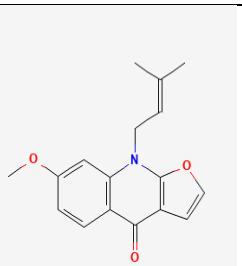
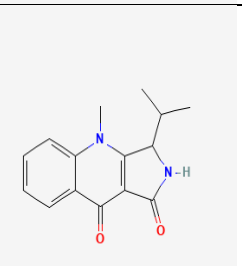
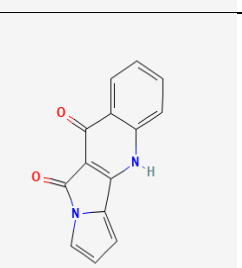
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov1074

36	Perfamine		3110227	313.3
37	Graveoline		353825	279.29
38	Graveolinine		11044132	279.29
39	Dubamine		360322	249.26
40	Anhydroevoxine		161441	329.3
41	Haplamine		648601	257.28
42	Maculosidine		68222	259.26

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov1074

43	Koniamborine		11345225	229.23
44	Quinolactacin A1		26202199	270.33
45	Quinolactacin A2		10061684	270.33
46	Quinolactacin B1	 S	10220728	256.3
47	Peniprequinolone		16681748	383.4
48	Yaequinolone J1		11655315	449.5
49	Yaequinolone J2		11669563	449.5

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov1074

50	Acrophylline		87840	283.32
51	Quinolactacin B		10220728	256.30
52	Quinolactacide		11481760	236.22

Chemical structures retrieved from PubChem. Selected compounds are highlighted according to their role in the present study.

- *Green:*
  Selected for ML training (8 compounds).
- *Yellow:*
  Kokusaginine – considered for further analysis, but excluded from ML training due to inconsistent labels.

Table 5 Summary of Dataset Collection, Curation, and Final Compound Selection for Machine-Learning Model Development.

Dataset Stage	**Number**
Initially collected quinoline natural products	52
Compounds with cancer-cell IC_{50} records	10
Total experimental IC_{50} observations after excluding camptothecin	20
Compounds excluded due to inconsistent activity labels	1
Final compounds used for ML training	9
Active compounds	4
Inactive compounds	5

➢ *Molecular Fingerprint Generation*

The molecular structures were converted into numerical representations using RDKit. Morgan fingerprints with a radius of 2 and a fingerprint size of 1024 bits were generated for each compound.The resulting feature matrix contained nine compounds and 1024 molecular fingerprint features.Feature matrix: 9 × 1024.These fingerprints represent the local chemical environments surrounding atoms in each molecule. They allow machine-learning algorithms to identify structural patterns that may be associated with the observed biological activity.

➢ *Performance of Machine-Learning Models*

Three machine-learning algorithms were developed: Decision Tree, Random Forest, and Logistic Regression. All three models were successfully trained using the nine-compound dataset.The predictions obtained from the trained models agreed with the assigned activity labels for all nine compounds in the training dataset.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov1074

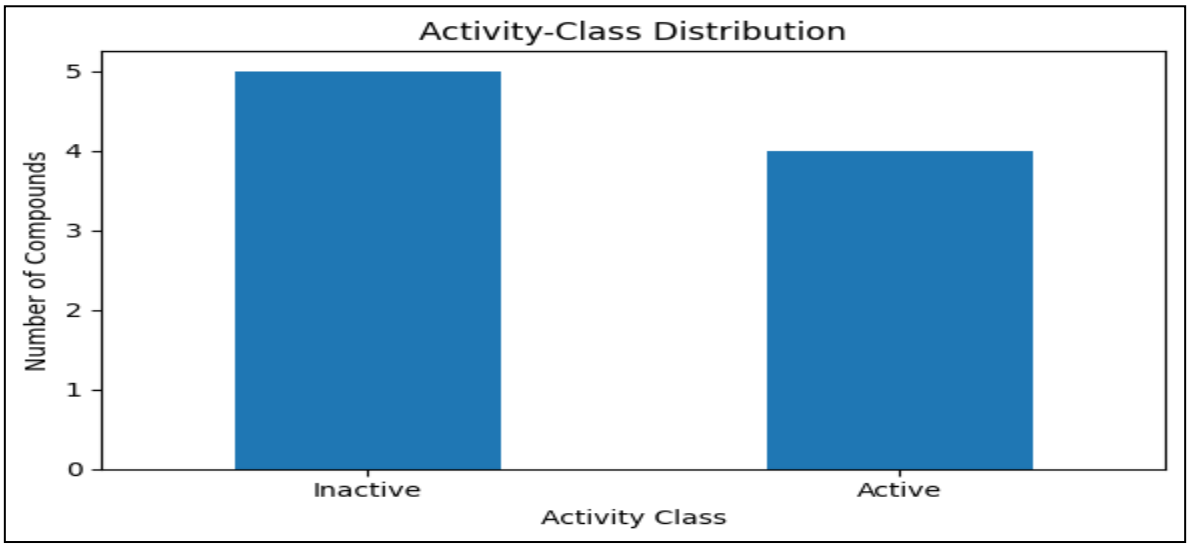


Fig 1 Distribution of Active and Inactive Compounds in the Curated Quinoline Dataset Used for Machine-Learning Model Development.

Table 6 Comparison of Experimentally Determined and Machine-Learning-Predicted Anticancer Activity Classifications of Selected Quinoline Natural Products Using Decision Tree, Random Forest and Logistic Regression Models.

Compound	Experimental Class	Decision Tree	Random Forest	Logistic Regression
Quinine	Inactive	Inactive	Inactive	Inactive
Dictamnine	Inactive	Inactive	Inactive	Inactive
Skimmianine	Active	Active	Active	Active
Gamma-Fagarine	Inactive	Inactive	Inactive	Inactive
Flindersiamine	Active	Active	Active	Active
Maculine	Active	Active	Active	Active
Nitidine	Active	Active	Active	Active
Streptonigrin	Inactive	Inactive	Inactive	Inactive
Graveolinine	Inactive	Inactive	Inactive	Inactive

All three models showed 9/9 agreement with the training labels (100% training-set agreement).Important interpretation: This result represents agreement on the same data used to train the models. It does not establish independent predictive accuracy or model generalisation. The small dataset requires further experimental data and independent validation before the models can be considered reliable screening tools.

➢ *Analysis of Important Molecular Fingerprint Features*

Random Forest feature importance was used to identify fingerprint bits that contributed most to the model's classification decisions. The ten highest-ranked fingerprint features were:

Table 7 Top 10 ranked Morgan fingerprint features identified as important contributors to the machine-learning model predictions.

Rank	Fingerprint Feature
1	FP_389
2	FP_64
3	FP_997
4	FP_709
5	FP_881
6	FP_166
7	FP_441
8	FP_549
9	FP_980
10	FP_613

The fingerprint-to-fragment analysis indicated the presence of several recurring structural environments, including aromatic carbon frameworks, oxygen-substituted aromatic rings, ether or methoxy-containing groups, and nitrogen-containing environments.Some fingerprint features, particularly FP_997, FP_881 and FP_441, were observed in Flindersiamine, Maculine and Nitidine, which were classified as active in the dataset. This observation suggests a possible

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov1074

association between these structural environments and the activity labels.However, Morgan fingerprints are hashed representations, and a single fingerprint bit may correspond to different molecular environments in different compounds. Therefore, these features should be interpreted as preliminary structural associations rather than definitive pharmacophoric groups or proven causes of anticancer activity.

➢ *Comparison of Physicochemical Properties*

The average physicochemical properties of active and inactive compounds were compared to investigate possible relationships between molecular characteristics and biological activity.

Table 8 Comparative Analysis of Physicochemical and Molecular Descriptor Properties of Active and Inactive Quinoline Compounds.

Molecular Property	Active Compounds	Inactive Compounds
Molecular weight (g/mol)	281.03	307.72
LogP	3.04	3.02
TPSA (Å²)	52.80	72.62
H-bond donors (HBD)	0.0	1.0
H-bond acceptors (HBA)	5.0	5.2
Rotatable bonds	2.0	3.0

The active compounds showed lower average molecular weight, topological polar surface area, hydrogen-bond donor count and number of rotatable bonds compared with the inactive compounds. The average LogP and hydrogen-bond acceptor values were relatively similar between the two groups. These observations suggest that, within this limited dataset, compounds with lower molecular weight, lower polarity and fewer rotatable bonds were more frequently associated with the active classification. Such properties may influence molecular permeability and interactions with biological targets. Nevertheless, these trends are exploratory because the number of compounds is very small. They should not be considered statistically established relationships.

➢ *Compound-Wise Molecular Property Heatmap*

A compound-wise molecular property heatmap was generated to visualize the physicochemical characteristics of the nine quinoline-containing natural products used in the machine-learning analysis. Six molecular descriptors, namely molecular weight (MW), lipophilicity (LogP), topological polar surface area (TPSA), hydrogen bond donors (HBD), hydrogen bond acceptors (HBA), and rotatable bonds, were included in the analysis. Before plotting, the molecular property values were standardized using the Z-score method to enable comparison between descriptors with different units and numerical ranges. In the heatmap, each row represents an individual natural product, while each column represents a molecular descriptor. Positive Z-scores indicate values above the dataset average, whereas negative Z-scores indicate values below the average. The heatmap provides a visual comparison of the molecular properties of individual compounds and helps identify differences in molecular size, lipophilicity, polarity, hydrogen-bonding capacity, and molecular flexibility. These characteristics may contribute to differences in biological activity and can support the interpretation of the machine-learning results. However, the heatmap represents descriptive molecular property patterns and does not independently establish a relationship between a particular descriptor and anticancer activity.

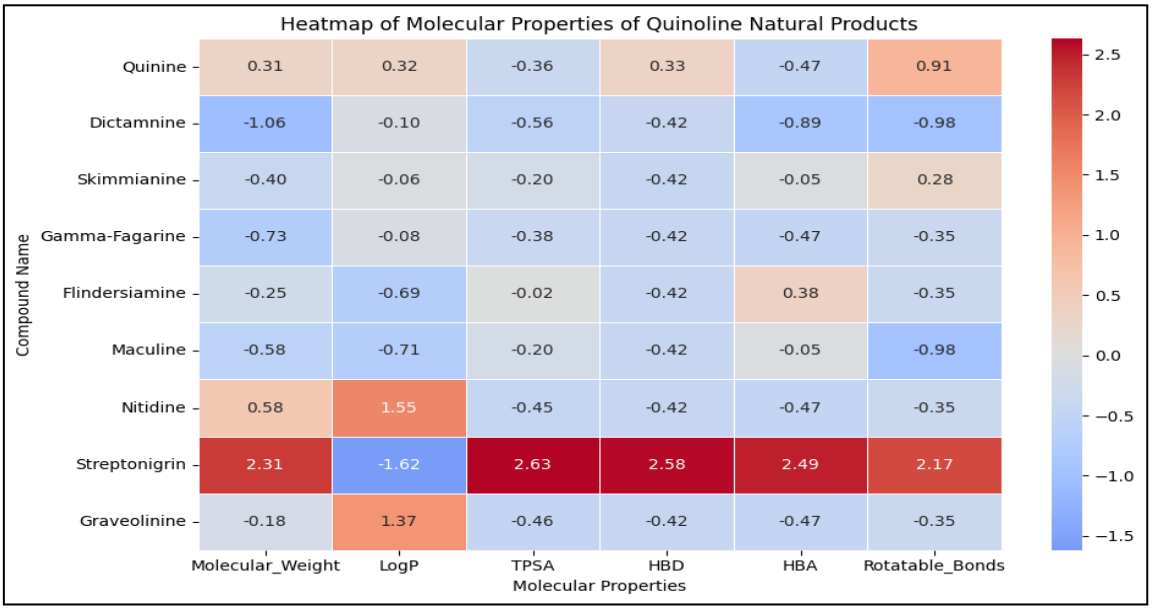


Fig 2 Heatmap Illustrating the Distribution and Comparative Patterns of Physicochemical Properties Among the Selected Quinoline Compounds.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov1074

➢ *Identification of Promising Natural Products*

Based on the experimentally reported IC_{50} values, Nitidine and Kokusaginine were identified as two promising natural products for further investigation.

Nitidine chloride (Standard) | Anti-malarial Agent | MedChemExpress

- *Nitidine*

Highest reported potency in this dataset

✓ IC_{50}: 0.40 μM
✓ Cell lines: A549 and PC-3
✓ PubChem assay records showed activity in both cell lines.

Nitidine exhibited the lowest reported IC_{50} value among the selected compounds, indicating strong in-vitro cytotoxic activity in the reported assays.

Kokusaginine - Chemical Compound

- *Kokusaginine*

✓ IC_{50}: 4.62 μM in MCF-7/ADR
✓ IC_{50}: 7.23 μM in MCF-7
✓ IC_{50}: 9.01 μM in MDA-MB-231
✓ IC_{50}: 10.19 μM in MDA-MB-231/ADR

Kokusaginine showed activity across several breast cancer cell-line records. However, its activity classification varied between cell lines, so it was excluded from the initial machine-learning training dataset. Nitidine was the most potent compound based on the reported IC_{50} values. Kokusaginine was also considered promising because of its activity across multiple breast cancer cell lines. Differences in assay conditions and cell lines should be considered when comparing their IC_{50} values.

IV. CONCLUSION

The study demonstrates a basic cheminformatics and machine-learning workflow for the preliminary screening of quinoline-containing natural products. PubChem was used to collect molecular structures and experimentally reported bioactivity data. RDKit Morgan fingerprints provided numerical molecular representations, while Decision Tree, Random Forest and Logistic Regression algorithms were used for activity classification.The analysis also identified potentially relevant structural fingerprint features and differences in physicochemical properties between active and inactive compounds. Nitidine and Kokusaginine emerged as promising candidates based on their reported experimental IC_{50} values. Despite these findings, the study has important limitations. The final model-training dataset contained only nine unique compounds, which is insufficient for developing robust predictive models using 1024 fingerprint features. Furthermore, the models were evaluated on their training data rather than an independent test set. Differences in cancer cell lines and experimental assay conditions may also influence the reported IC_{50} values. Therefore, the present work should be considered a preliminary machine-learning screening exercise. Expanding the dataset, incorporating additional molecular descriptors, accounting for cell-line information, and performing independent validation would improve the reliability of future models.

One important point before submitting: Your assignment asks for the top two compounds identified through ML screening. The two compounds above were selected based on experimental IC_{50} values, not a separate prediction on an unseen candidate library. In the report, describe them as promising compounds identified from the experimental bioactivity dataset, rather than claiming the ML models independently discovered them.

REFERENCES

[1]. Kim S, Chen J, Cheng T, et al. PubChem 2025 update. Nucleic Acids Res. 2025;53(D1):D1516–D1525. doi:10.1093/nar/gkae1059.
[2]. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50(5):742–754. doi:10.1021/ci100050t.
[3]. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–2830.
[4]. Organisation for Economic Co-operation and Development (OECD). Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models. OECD Series on Testing and Assessment No. 69. Paris: OECD Publishing; 2014. doi:10.1787/9789264085442-en.
[5]. Kim S, Chen J, Cheng T, et al. PubChem 2023 update. Nucleic Acids Res. 2023;51(D1):D1373–D1380. doi:10.1093/nar/gkac956.