Korean J Physiol Pharmacol 2024;28(6):527-537
https://doi.org/10.4196/kjpp.2024.28.6.527

KJPP

Original Article

Predicting antioxidant activity of compounds based on chemical structure using machine learning methods

Jinwoo Jung[1,2], Jeon-Ok Moon[1], Song Ih Ahn[2,*], and Haeseung Lee[1,*]

[1]*Department of Pharmacy, College of Pharmacy and Research Institute for Drug Development,* [2]*School of Mechanical Engineering, Pusan National University, Busan 46241, Korea*

ARTICLE INFO

Received May 7, 2024
Revised July 12, 2024
Accepted July 12, 2024

***Correspondence**
Song Ih Ahn
E-mail: songihahn@pusan.ac.kr
Haeseung Lee
E-mail: haeseung@pusan.ac.kr

Key Words
Antioxidants
Artificial intelligence
Data mining
Machine learning
Quantitative structure-activity relationship

ABSTRACT Oxidative stress is a well-established risk factor for numerous chronic diseases, emphasizing the need for efficient identification of potent antioxidants. Conventional methods for assessing antioxidant properties are often time-consuming and resource-intensive, typically relying on laborious biochemical assays. In this study, we investigated the applicability of machine learning (ML) algorithms for predicting the antioxidant activity of compounds based solely on their molecular structure. We evaluated the performance of five ML algorithms, Support Vector Machine (SVM), Logistic Regression (LR), XGBoost, Random Forest (RF), and Deep Neural Network (DNN), using a dataset of over 1,900 compounds with experimentally determined antioxidant activity. Both RF and SVM achieved the best overall performance, exhibiting high accuracy (> 0.9) and effectively distinguishing active and inactive compounds with high structural similarity. External validation using natural product data from the BATMAN database confirmed the generalizability of the RF and SVM models. Our results suggest that ML models serve as powerful tools to expedite the discovery of novel antioxidant candidates, potentially streamlining the development of future therapeutic interventions.

INTRODUCTION

Oxidative stress is a recognized consequence of an imbalance between free radical generation and the body's antioxidant defenses, ultimately leading to cellular and tissue damage [1]. This imbalance has been implicated in the development and progression of various diseases, including cardiovascular disease, neurodegenerative disease, and cancers. Therefore, the identification of novel antioxidant substances is imperative for advancing therapeutic strategies aimed at mitigating these health issues [2,3]. Traditionally, the assessment of antioxidant capacity has relied on *in vitro* biochemical assays, such as the 2,2-diphenyl-1-picrylhydrazyl (DPPH) and 2,2'-azino-bis(3-ethylbenzothiazoline-6-sulfonic acid) (ABTS) tests [4,5]. While the DPPH and ABTS assays are effective, they are often labor-intensive and time-consuming, requiring substantial amounts of sample material. Moreover, the complexity of these methods limits their scalability and efficiency and hinders the high-throughput screening of compounds, particularly in early-stage drug discovery processes.

The emergence of machine learning (ML) technologies offers promising alternative approaches to enhance the efficiency of antioxidant identification [6,7]. ML has the potential to overcome the limitations of traditional methods by facilitating rapid, cost-effective *in silico* screening of vast chemical libraries for antioxidant activity [8,9]. By leveraging accumulated histological data sets of chemical bioassays and advanced algorithmic models, ML can predict the antioxidant potential of compounds solely based on their chemical structures. This capability expedites the discovery process and enriches our understanding of structure-activity relationships within antioxidant compounds [10].

Author contributions: J.J., Investigation, Data curation, Methodology, Visualization, Writing - Original Draft; J.O.M., Conceptualization, Writing - Original Draft; S.I.A., Supervision; H.L., Conceptualization, Supervision, Funding acquisition, Writing - Review & Editing.

This study investigated the applicability of various ML algorithms for predicting the antioxidant activity of chemical compounds. We curated a dataset of chemical structures annotated with experimentally determined antioxidant activities and employed well-established ML algorithms to develop predictive models. We demonstrated the efficacy of these models in predicting antioxidant activities and assessed the reliability of prediction as a preliminary screening tool before extensive *in vitro* validation.

METHODS

Acquisition of antioxidant activity data

Publicly available antioxidant activity data for a diverse range of compounds was retrieved from the PubChem database. To control methodological consistency within the dataset, data was exclusively sourced from well-established assays: ABTS and DPPH. ABTS and DPPH assays quantify a compound's free radical scavenging (ABTS) or reduction (DPPH) capacity, reflected by a measurable color change [4,5]. Selection of these assays was based on their widespread adoption in rapid antioxidant potential screening due to their simplicity and effectiveness. PubChem searches employing the keywords "DPPH" and "ABTS" identified a total of 1,651 DPPH and 366 ABTS assays encompassing 19,454 compounds.

Data preprocessing

To eliminate redundancy arising from compounds with identical structures but varying identifiers, the International Union of Pure and Applied Chemistry (IUPAC) InChIKeys was adopted as the unique identifier. The RDKit Python package was utilized to convert the Simplified Molecular Input Line Entry System (SMILES) strings for each compound into IUPAC InChIKeys. This process refined the initial 19,454 PubChem CIDs into 10,053 unique compounds. Subsequently, Extended-Connectivity Fingerprints (ECFP) with a radius of 4 (ECFP-4) were generated using the RDKit Python package to represent chemical structures based on each compound's SMILES string. Compounds were categorized into four activity groups for each assay: 'Active', 'Inactive', 'Unspecified', or 'Inconclusive', and those with solely 'Inconclusive' and 'Unspecified' designations across all assays were excluded. In addition, compounds exhibiting inconsistent activity results between the two assays were eliminated. To enhance the dataset's comprehensiveness, 24 well-known antioxidant compounds documented in prior studies were incorporated and designated as the 'Active' set [11].

Structural similarity calculation

The Tanimoto coefficient [12] between each pair of all compounds was computed using their ECFP-4 fingerprint *via* the DataStructs Python package. ECFP-4 fingerprint is a binary vector of 2,048 bits in length, where each bit represents the presence (set to 1) or absence (set to 0) of specific circular substructures within a molecule. The Tanimoto coefficient itself is a value between 0 and 1, representing the degree of structural similarity between two molecules. A higher coefficient indicates a greater degree of structural similarity. The calculation is as follows:

$$T_c = \frac{N_{ab}}{N_a + N_b - N_{ab}}$$

where N_{ab} is the number of common features between the two compounds, N_a is the total number of features in compound A, and N_b is the total number of features in compound B. To visualize the relationships between the compounds based on the calculated Tanimoto coefficients, a compound-compound network was constructed. In this network, each compound is represented by a node, and edges connect nodes that exhibit a Tanimoto coefficient exceeding a predefined threshold (set to 0.7 in this case). Cytoscape 3 (version 10.3.1), an open-source software platform for network visualization, was used to generate this network representation.

ML model training and evaluation

Five ML algorithms were chosen for their capability to model complex relationships between molecular structure and antioxidant activity: Support Vector Machine (SVM), Logistic Regression (LR), XGBoost (XGB), Random Forest (RF), and Deep Neural Network (DNN). Each model was trained *via* functions from the scikit-learn Python package with default parameter settings to prevent overfitting and ensure generalizability (Table 1, Supplementary Fig. 1). For robust and unbiased evaluation, five-fold cross-validation was conducted using two data splitting strategies: (i) random splitting and (ii) scaffold splitting. To achieve scaffold splitting, all compounds were classified into scaffold groups using a Scaffold Network Generator [13] implemented in the RDKit Python package. This tool organized compounds into a hierarchical tree structure, with groups ranging from single-ring to a maximum of 15-ring structures. Consequently, 1,931 compounds were categorized into 778 scaffold groups. The training data was then split based on scaffold membership, ensuring the testing set contained compounds with unseen scaffold structures. Using both splitting strategies, the five-fold cross-validation was repeated 100 times. Performance metrics including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUROC) were computed for each iteration across all five models using the scikit-learn Python package.

Table 1. Machine learning models and hyperparameter settings

Method	Class (Package)	Parameter
SVM	SVC (scikit-learn)	C: 1.0, kernel: rbf, degree: 3, gamma: scale, coef: 0.0, shrinking: True, probability: True, tol: 0.001, class weight: False, max_iter: 1, decision function shape: ovr, break ties: False
LR	LogisticRegression (scikit-learn)	penalty: l2, dual: False, tol: 0.0001, C: 1.0, fit intercept: True, intercept scaling: 1, class weight: None, solver: lbfgs, max iter: 100, multi class: auto, wart start: False
XGB	XGBClassifier (xgboost)	booster: gbtree, learning rate: 0.3, gamma: 0, max depth: 6, min child weight: 1, max delta step: 0, subsample: 1, sampling method: uniform, colsample bytree: 1, colsample bylevel: 1, colsample bynode: 1, lambda: 1, alpha: 0, tree method: auto, scale pos weight: 1, refresh leaf: 1, max leaves: 0, max bin: 256, num parallel tree: 1
RF	RandomForestClassifier (scikit-learn)	n estimator: 100, criterion: gini, max depth: None, min samples split: 2, min samples leaf: 1, min weight fraction leaf: 0, max features: sqrt, max leaf nodes: None, min impurity decrease: 0, bootstrap: True, oob score: False, warm strat: False, class weight: None, ccp alpha: 0, max samples: None, monotonic: None
DNN	Model (TensorFlow)	k: 5, input shape: 2048, layers: 2048, 1024, 512, 256, train size: 0.6, validation size: 0.2, test size: 0.2l2 regularization: null, batch normalization: False, activation function: relu, loss function: BinaryCrossentropy, learning rate: 0.001, optimizer: Adam, metric: BinaryAccuracy, AUC, early stop monitor: val loss, early stop patience: 10, class weight: False, batch size: 256, epochs: 1000, seed: 42

SVM, Support Vector Machine; LR, Logistic Regression; XGB, XGBoost; RF, Random Forest; DNN, Deep Neural Network.

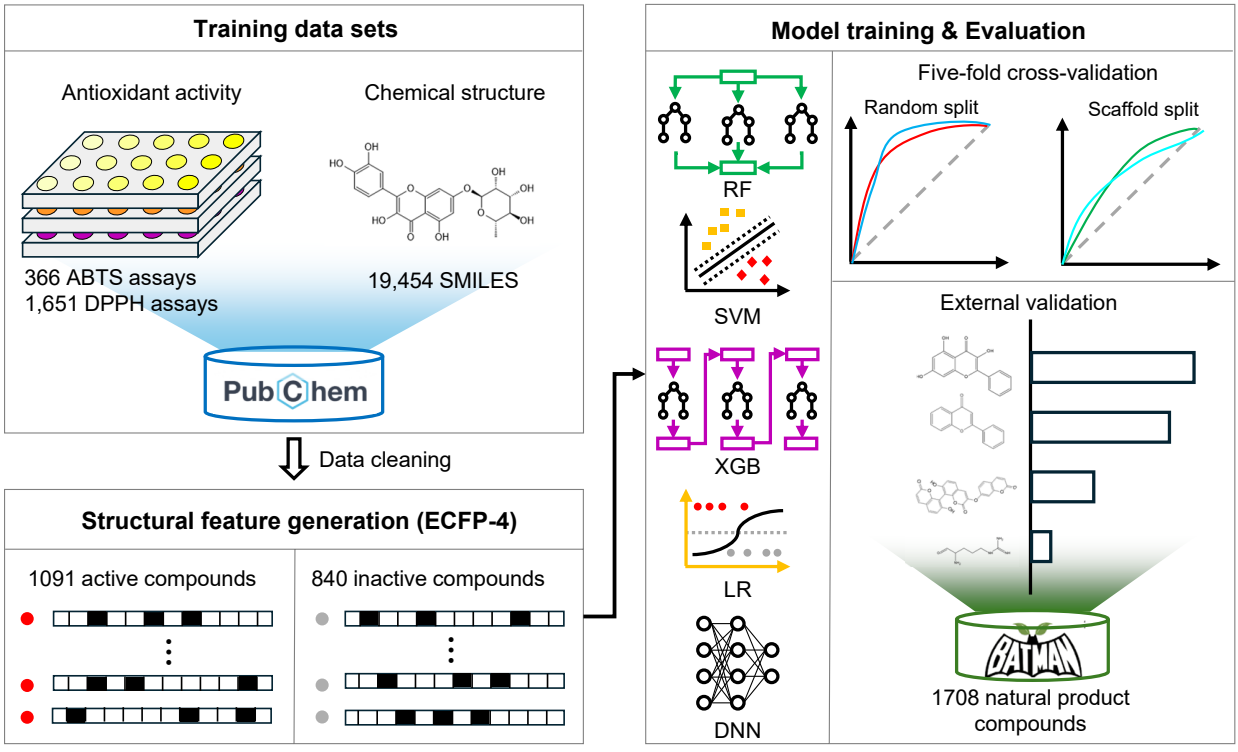


Fig. 1. Schematic illustration of the analytic workflow for constructing antioxidant compound prediction models employing five ML algorithms. ML, machine learning; ABTS, 2,2′-azino-bis(3-ethylbenzothiazoline-6-sulfonic acid); DPPH, 2,2-diphenyl-1-picrylhydrazyl; SMILES, Simplified Molecular Input Line Entry System; ECFP-4, Extended-Connectivity Fingerprints with a radius of 4; RF, Random Forest; SVM, Support Vector Machine; XGB, XGBoost; LR, Logistic Regression; DNN, Deep Neural Network.

RESULTS

Overall analytic process

This study evaluated well-established ML algorithms for predicting the antioxidant activity of compounds based solely on their molecular structure represented by SMILES strings (Fig. 1). A dataset of 19,454 compounds with antioxidant activity data and corresponding SMILES information was collected from PubChem and literature searches. Following a rigorous data-cleaning

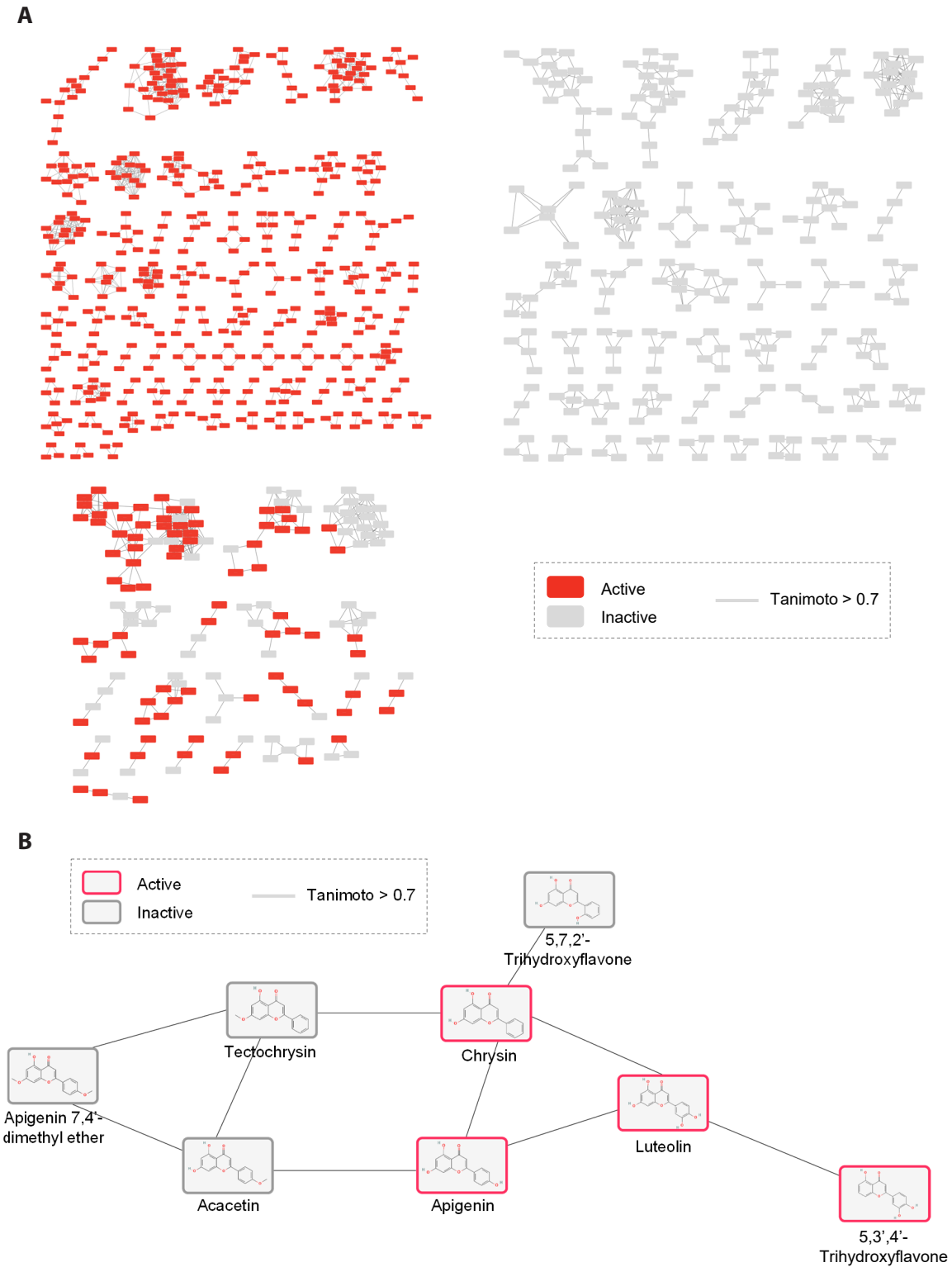


Fig. 2. Compound structural diversity and its association with antioxidant activity. (A) Network visualization of compound relationships, where nodes represent individual compounds and edges connecting the nodes indicate high structural similarity between compounds (Tanimoto coefficient exceeding 0.7). Nodes are colored according to their experimentally determined antioxidant activity (gray for inactive, red for active). The network is segregated into three sub-panels: (i) a network containing only interconnected active compounds, (ii) a network with solely interconnected inactive compounds, and (iii) a network with a mixture of active and inactive interconnected compounds. (B) A representative example of a network module containing both active and inactive compounds. This highlights the potential for structurally similar compounds to exhibit diverse antioxidant properties. The border color of each node corresponds to its experimentally determined antioxidant activity (gray for inactive, red for active).

process, a final set of 1,931 compounds (1,092 active, 839 inactive) was used to develop antioxidant activity prediction models. The ECFP-4 fingerprints, encoding the chemical environment of each molecule, were used as input features for the five ML algorithms (SVM, LR, XGB, RF, and DNN). A cross-validation scheme was implemented to identify the most suitable model for predicting antioxidant activity. The generalizability of the chosen model was further evaluated using external datasets of natural product compounds.

Structural diversity of compounds and its relationship to antioxidant activity

To quantify the chemical diversity within the dataset and explore potential structural relationships associated with antioxidant activities, pairwise Tanimoto similarity coefficients [12] were calculated for all compounds. A compound-compound network, where nodes represented individual compounds, with edges connecting nodes that shared a Tanimoto coefficient exceeding a threshold of 0.7, was constructed to explore structurally similar compounds in the collected compounds (Fig. 2A). This network revealed that active and inactive compounds formed separate clusters, suggesting a positive correlation between structural similarity and antioxidant activity. However, an intriguing exception was observed within a cluster enriched with flavonoids structurally related to chrysin, a potent antioxidant flavone (Fig. 2B). While all compounds within this cluster shared a flavone backbone similar to the potent antioxidant chrysin, their antioxidant capacities diverged. This highlights the importance of the specific arrangement of functional groups within the flavone structure. Chrysin, with hydroxyl groups at the 5 and 7 positions of the A ring, exhibits strong antioxidant activity due to the well-documented radical scavenging properties of these groups [14]. The addition of hydroxyl group at the 3' position of the B ring (as in apigenin) or at both the 3' and 4' positions (as in luteolin) enhances antioxidative activity. However, the addition of a hydroxyl group at the 2' position of the B ring (as in 5,7,2'-trihydroxyflavone) does not result in antioxidative effects. Substitution of

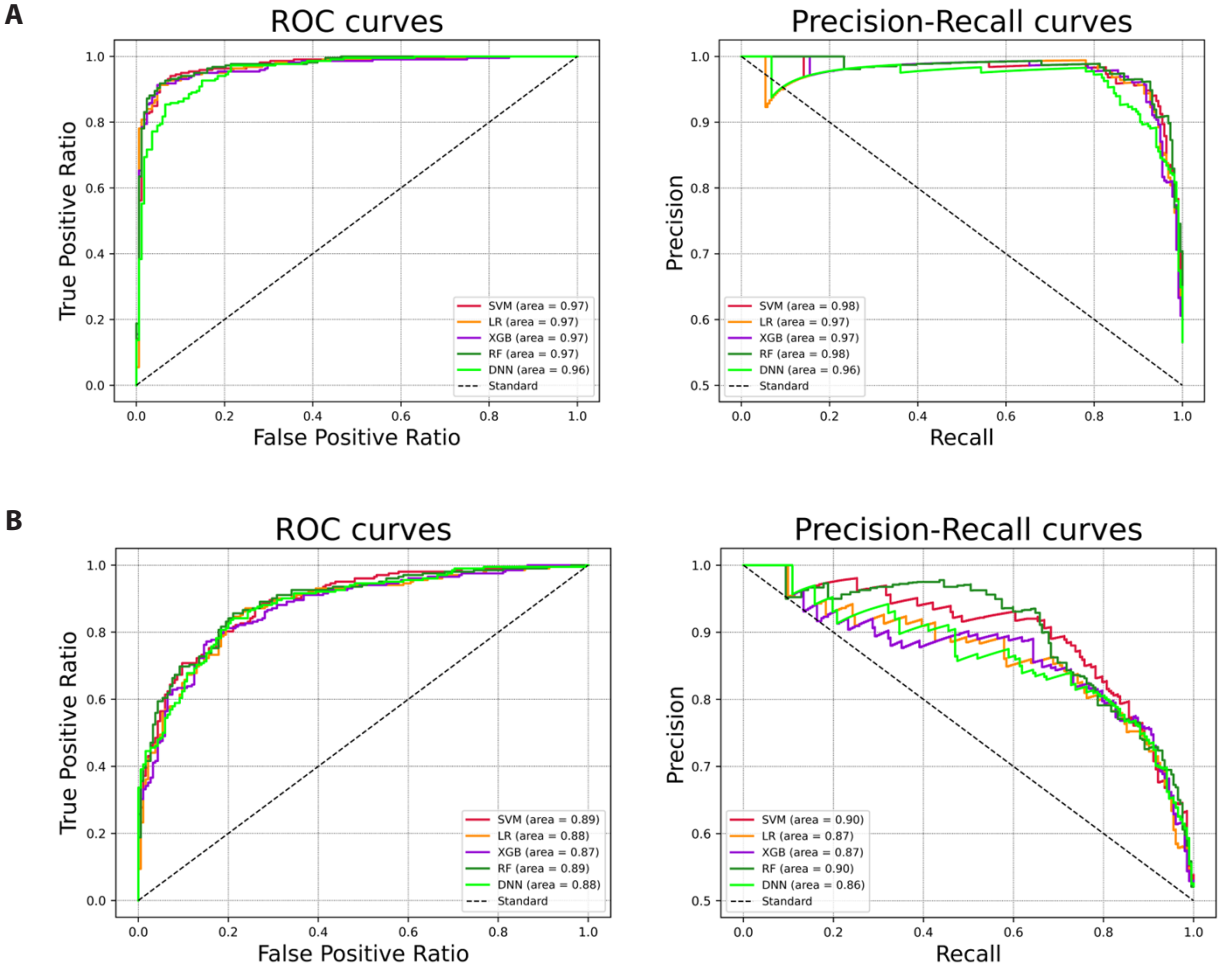


Fig. 3. Performance comparison of five ML models. (A, B) Representative receiver operating characteristic (ROC) curves and precision-recall curves were obtained from a single iteration of 5-fold cross-validation with (A) random splitting and (B) scaffold splitting. ML, machine learning; SVM, Support Vector Machine; LR, Logistic Regression; XGB, XGBoost; RF, Random Forest; DNN, Deep Neural Network.

hydroxyl groups with methyl groups diminishes activity; for example, tectochrysin and acacetin (methylated at the 7 position of the A ring and 3' position of the B ring, respectively) and apigenin 7,4'-dimethyl ether (methylated at the 7 position of the A ring and 4' position of the B ring) both lose their antioxidative properties. These observations highlight that the hydroxyl groups at the 5 and 7 positions on the A ring, with an unblocked 3' hydroxyl group on the B ring, are crucial for the antioxidant activity of flavones. This data suggests that while overall structural features are informative in determining antioxidant activity, it is the precise arrangement of functional groups within the molecule that ultimately dictates its efficacy.

Model performance and evaluation

Five ML models, including SVM, LR, XGB, RF, and DNN, were employed to learn these subtle yet important structural features for predicting antioxidant activity based on ECFP-4 fingerprints. Model performance was evaluated using two five-fold cross-validation (5-fold CV) schemes: random splitting and scaffold splitting (Fig. 3). Random splitting divided the training data into five equal folds, where each fold is used for testing once while the remaining four are used for training. On the other hand, scaffold splitting grouped structurally similar compounds together in each fold. This approach evaluates the model's ability to predict the activity of compounds with novel scaffold structures not present in the training data, a crucial capability for discovering novel antioxidants with distinct chemical backbones (known as scaffold hopping).

All models achieved commendable performance on the random splitting CV (Fig. 3A and Table 2). RF outperformed other models in all metrics, including accuracy (0.908 ± 0.004), precision (0.912 ± 0.004), recall (0.927 ± 0.005), F1 score (0.919 ± 0.003), and AUROC (0.968 ± 0.002). SVM and XGB showed competitive performance with accuracies exceeding 0.900 and AUROC above 0.955. LR performed similarly but with slightly lower recall and F1 scores. DNN, while achieving a respectable accuracy (0.877 ± 0.015), exhibited higher variability in its metrics, suggesting potential overfitting or sensitivity to the composition of training data.

Notably, all models maintained good performance on scaffold-splitting CVs, albeit their scores were slightly lower compared to random-splitting (Fig. 3B and Table 3). All models except DNN maintained high accuracies above 0.900, with SVM and XGB recording the highest (0.906 ± 0.007). Precision was notably higher for LR (over 0.918), while RF and SVM demonstrated superior recall rates (both exceeding 0.950). F1 scores remained consistent and high for SVM, LR, XGB, and RF. However, DNN's performance significantly declined, suggesting its lower robustness to scaffold-based splits. Similarly, AUROC scores for SVM, LR, XGB, and RF remained high, demonstrating their ability to distinguish classes across diverse data segmentation.

Overall, the results suggest that SVM and RF are well-suited for predicting antioxidant activity based on ECFP-4 fingerprints, exhibiting both high accuracy and generalizability across splitting methodologies.

Table 2. Model performance obtained from 5-fold CV with random-splitting

	SVM	LR	XGB	RF	DNN
Accuracy	0.903 ± 0.003	0.898 ± 0.004	0.9 ± 0.005	0.908 ± 0.004	0.877 ± 0.015
Precision	0.907 ± 0.003	0.908 ± 0.004	0.907 ± 0.005	0.912 ± 0.004	0.889 ± 0.02
Recall	0.923 ± 0.004	0.913 ± 0.005	0.918 ± 0.006	0.927 ± 0.005	0.889 ± 0.025
F1 score	0.915 ± 0.003	0.91 ± 0.003	0.912 ± 0.005	0.919 ± 0.003	0.891 ± 0.014
AUROC	0.959 ± 0.001	0.955 ± 0.002	0.955 ± 0.003	0.968 ± 0.002	0.945 ± 0.012

Performance metrics were obtained across 100 iterations of 5-fold CV with random-splitting (average ± standard deviation). CV, cross-validation; SVM, Support Vector Machine; LR, Logistic Regression; XGB, XGBoost; RF, Random Forest; DNN, Deep Neural Network, AUROC, area under the receiver operating characteristic curve.

Table 3. Model performance from 5-fold CV with scaffold-splitting

	SVM	LR	XGB	RF	DNN
Accuracy	0.906 ± 0.005	0.905 ± 0.007	0.906 ± 0.007	0.904 ± 0.006	0.801 ± 0.03
Precision	0.902 ± 0.005	0.918 ± 0.007	0.917 ± 0.007	0.9 ± 0.006	0.834 ± 0.041
Recall	0.958 ± 0.005	0.935 ± 0.006	0.939 ± 0.008	0.958 ± 0.006	0.811 ± 0.064
F1 score	0.929 ± 0.004	0.926 ± 0.005	0.927 ± 0.006	0.927 ± 0.004	0.818 ± 0.036
AUROC	0.968 ± 0.003	0.965 ± 0.003	0.964 ± 0.004	0.968 ± 0.003	0.886 ± 0.028

Performance metrics were obtained across 100 iterations of 5-fold CV with scaffold-splitting (average ± standard deviation). CV, cross-validation; SVM, Support Vector Machine; LR, Logistic Regression; XGB, XGBoost; RF, Random Forest; DNN, Deep Neural Network, AUROC, area under the receiver operating characteristic curve.

Discriminative power of RF and SVM on the antioxidant activity of structurally similar compounds

To assess the ability of RF and SVM models to capture subtle structural features that are important for predicting antioxidant activity, we investigated their performance in differentiating between active and inactive compounds with high structural similarity. We focused on three network modules, each containing reference active compounds (chrysin, eriophorin A, and 4-(1H-indol-2-yl)aniline) (Fig. 4). Within the chrysin module (Fig. 2B),

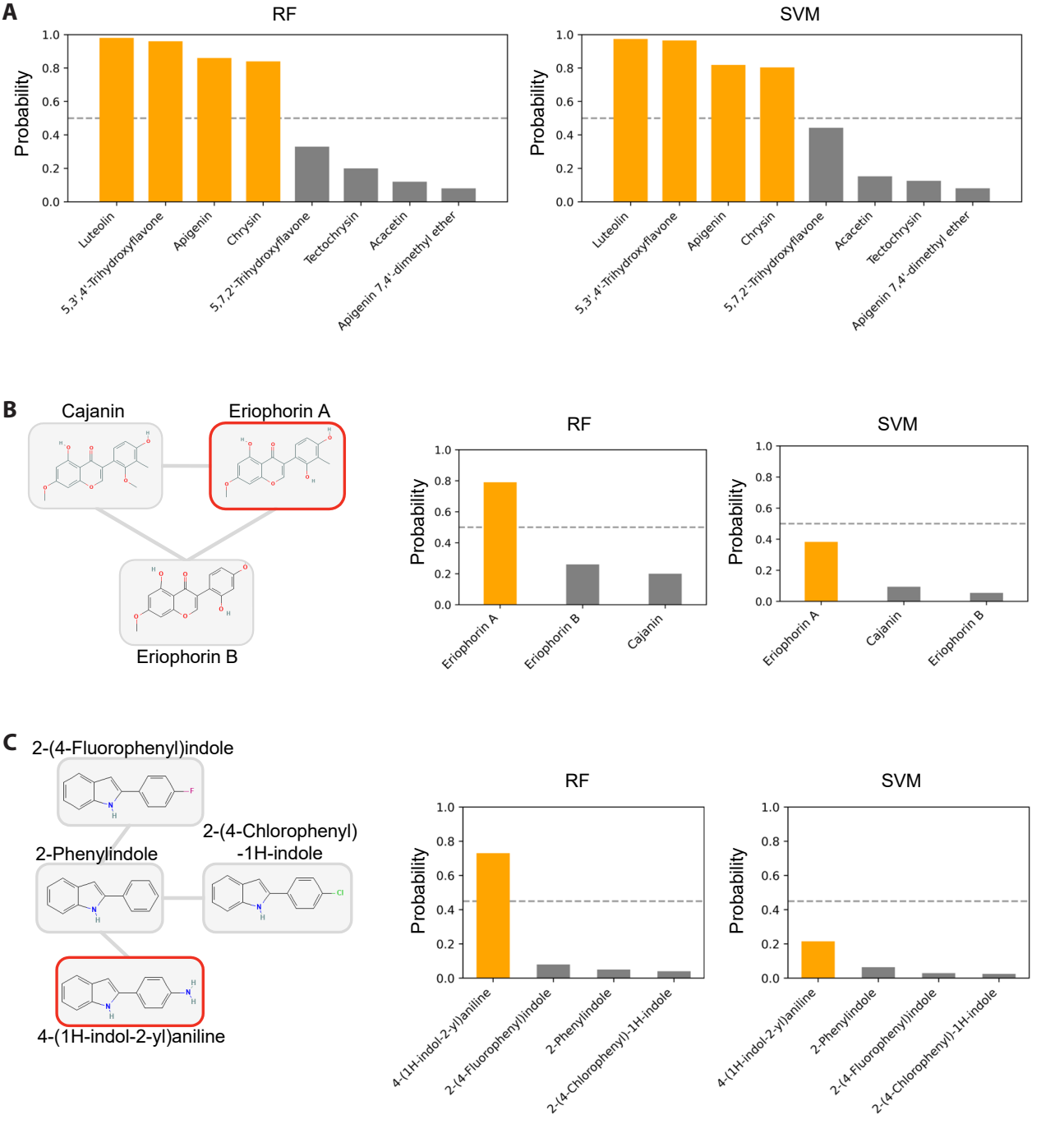
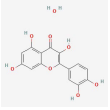
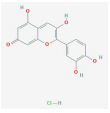
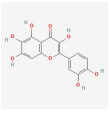
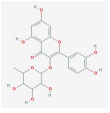
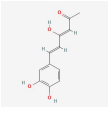
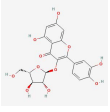
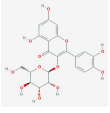
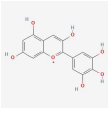
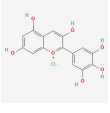
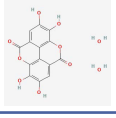


Fig. 4. Comparison of predicted antioxidant activity scores for structurally similar compounds. (A) Predicted activity scores of compounds within the chrysin network module, as determined by RF and SVM models. (B, C) Network modules for (B) eriophorin A and (C) 4-(1H-indol-2-yl)aniline. The left panel shows the network where nodes represent compounds and edges represent high structural similarity (Tanimoto coefficient exceeding 0.7). The right panel displays the corresponding predicted activity scores obtained by the RF and SVM models for each compound within the respective network module. The border color of each node indicates its experimentally determined antioxidant activity (gray for inactive, red for active). RF, Random Forest; SVM, Support Vector Machine.

both models accurately classified chrysin and luteolin as active and apigenin as inactive, despite their high Tanimoto coefficient (> 0.7) (Fig. 4A). However, their performance diverged in other modules. Specifically, the RF model exhibited superior discriminatory power in the eriophorin A module (Fig. 4B), correctly classifying eriophorin A as active and eriophorin B and cajanin as inactive, even with their high structural similarity. Conversely, the SVM model, while differentiating active from inactive compounds, underestimated eriophorin A's activity, classifying it as inactive. A similar trend emerged in the 4-(1H-indol-2-yl)aniline module (Fig. 4C). The RF model accurately predicted active compounds, while the SVM model struggled. These findings demonstrate that, while both models can discriminate between some highly similar compounds, the RF model exhibits a stronger ability to distinguish active and inactive compounds with a high degree of structural similarity across diverse scaffold compounds. This suggests that RF models may be better suited to capture subtle structural variations that significantly influence antioxidant

Table 4. Top ten predicted antioxidant compounds from the BATMAN database

Compound	PubChem CID	Structure	SVM	RF	No. of reference	Reference reporting antioxidant activity
Quercetin hydrate	16212154		1.00	1.00	29	[16]
Cyanidin chloride	68247		0.99	0.99	3,860	[17]
Quercetagetin	5281680		1.00	0.99	679	[18]
Quercetin 3-O-rhamnoside	5353915		0.99	1	6	[19]
Hispolon	10082188		0.99	1.00	86	[20]
Avicularin	5490064		0.98	1	1,056	[21]
Isoquercitrin	51402807		0.98	1	2,246	[22]
Delphinidin chloride	128853		1.00	0.99	1,003	[23]
Delphinidin	68245		1.00	0.99	2,624	[17]
Ellagic acid dihydrate	16760409		0.99	0.99	15	-

BATMAN, Bioinformatics Analysis Tool for Molecular mechANism of Traditional Chinese Medicine; SVM, Support Vector Machine; RF, Random Forest.

activity.

Prediction of antioxidant activity of natural products

To evaluate the generalizability of SVM and RF models, we performed external validation using a dataset of natural product compounds from the BATMAN (Bioinformatics Analysis Tool for Molecular mechANism of Traditional Chinese Medicine) database [15]. BATMAN offers a comprehensive resource for bioactive compounds found in traditional medicine and other natural products. We retrieved chemical structure data for 1,708 well-defined ingredient compounds extracted from 8,404 medicinal plants. A subset of 1,594 compounds not included in the training set was selected for unbiased evaluation. These natural compounds were then subjected to the SVM and RF models to predict their potential antioxidant activity (Supplementary Table 1). Subsequently, the top candidates with the highest average scores obtained from both models were shortlisted for further investigation. Notably, this shortlist was significantly enriched for highly hydroxylated flavonoids (Table 4). These specific classes of compounds are recognized for exhibiting antioxidant activity through free radical scavenging, metal chelation, and involvement in redox reactions [16-23].

To explore natural compounds with novel antioxidant bioactivities, a literature search was conducted using SciFinder [24]. This search yielded publication counts for each compound, providing an indicator of their prior investigation in the context of biological activity. Most of the top 10 compounds are well-studied (median publication count = 841), suggesting their high potential for bioactivity. Among them, we focused on two particularly intriguing compounds, ellagic acid dihydrate and strictinin, which had less than 30 references each, potentially indicating a lack of previous exploration regarding their antioxidant properties (Fig. 5). Ellagic acid dihydrate is a crystalline form of ellagic acid, a well-known polyphenol found in various fruits and nuts, containing two water molecules within its structure. While ellagic acid itself exhibits potent antioxidant activity, exceeding established antioxidants like butylated hydroxytoluene and vitamin E in

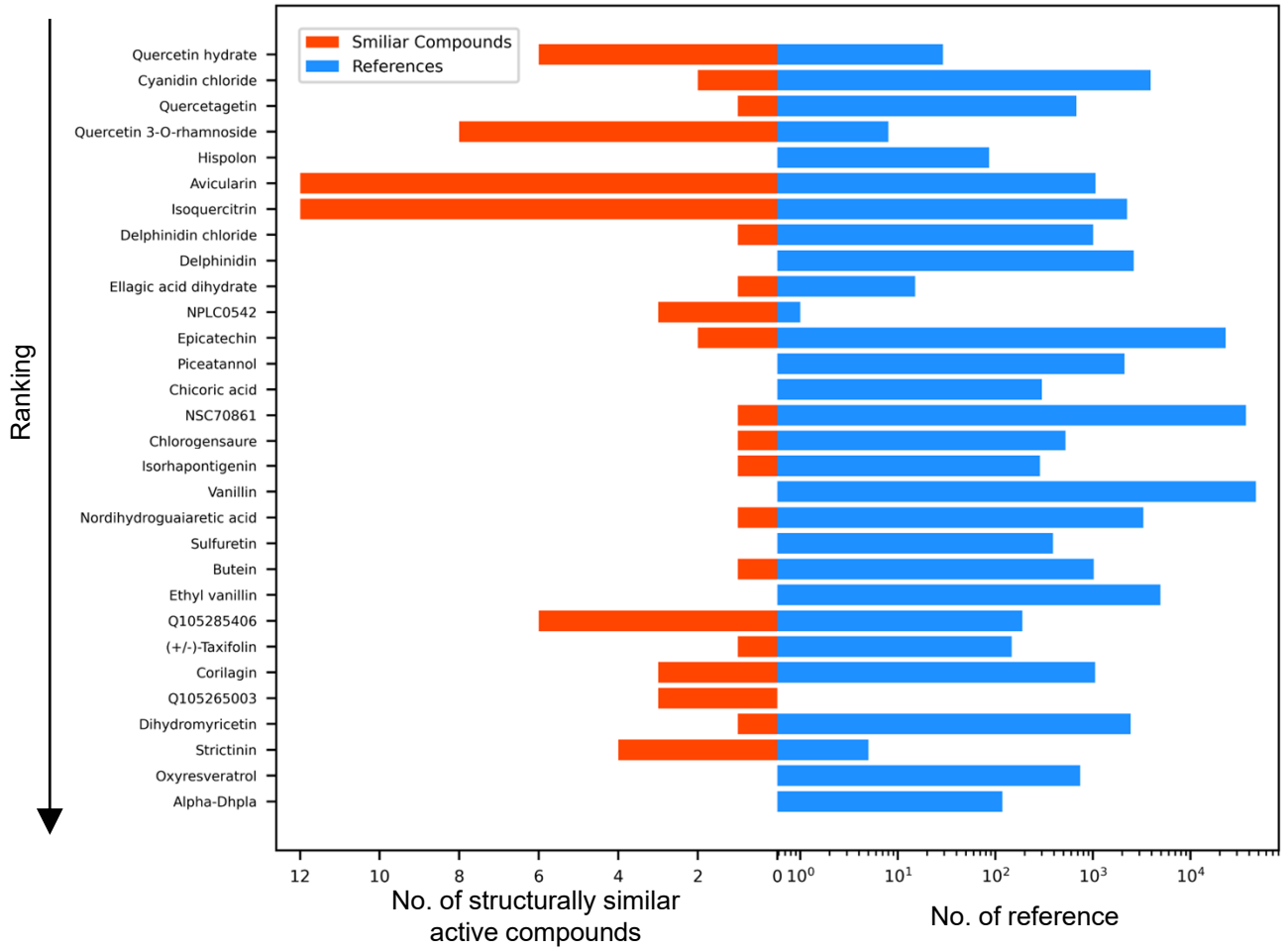


Fig. 5. Distribution of publication counts and number of structurally similar active compounds for top-scoring compounds. Blue bars represent the number of references associated with the top 20 compounds identified through RF and SVM models. The red bars represent the number of structurally similar active compounds present within the training datasets for these same top-ranked compounds. The compounds were ranked based on their average scores obtained from both models. RF, Random Forest; SVM, Support Vector Machine.

inhibiting lipid peroxidation [25], the specific activity of its dihydrate form remains unexplored. Given the established bioactivity of ellagic acid, its dihydrate form is a promising candidate for further investigation with a high probability of exhibiting antioxidant properties. Strictinin, a hydrolyzable ellagitannin, has also been demonstrated to possess significant antioxidant properties [26]. Multiple studies have demonstrated that strictinin possesses potent antioxidant properties that can inhibit lipid peroxidation and scavenge free radicals [27,28]. Collectively, our ML approach effectively prioritizes promising candidates for further investigation of their potential antioxidant bioactivities.

DISCUSSION

This study evaluated the applicability of various ML algorithms for predicting the antioxidant activity of compounds solely based on their chemical structure information. In the current study, we aimed to investigate a baseline performance for these models within a standardized framework. To facilitate a controlled initial assessment, all models were evaluated using their default hyperparameter settings. Under these conditions, RF and SVM demonstrated superior performance compared to other algorithms LR, XGB, and DNN. Notably, DNN displayed the lowest performance among the five models. This finding aligns with the known susceptibility of DNNs to overfitting on datasets with limited sample sizes. The relatively small size of our dataset likely contributed to its underperformance, emphasizing the critical role of data availability and characteristics in model selection. For robust model comparisons, future research should incorporate a rigorous hyperparameter tuning process to optimize the potential of each algorithm.

To simply focus on the applicability of ML in predicting antioxidant activity based on chemical structure, we utilized structural features, particularly ECFP-4 fingerprints (a widely employed representation of compound structure). These fingerprints effectively captured subtle yet critical structural features associated with antioxidant activity. Future research should incorporate feature importance analysis to identify and interpret the most significant features influencing the models' predictions in the context of antioxidant activity. Expanding the feature space to include additional data sources, such as chemical descriptors and chemical-induced transcriptomic data, alongside ECFP-4 fingerprints, could be explored to enhance model generalizability and provide a more comprehensive understanding of the structure-function relationship in antioxidant activity.

While *in silico* approaches offer significant promise, experimental validation remains an essential step. Compounds predicted by the ML models to have high antioxidant activity should be subjected to biochemical assays to confirm their activity. This step is critical for bridging the gap between computational predictions and practical applications, ensuring that predictions translate reliably into real-world benefits. We propose that our models can serve as a preliminary screening tool, facilitating the selection of candidate compounds for subsequent *in vitro* validation.

FUNDING

This work was supported by a 2-Year Research Grant of Pusan National University.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to Professor Minhye Yang for her generous support throughout this project. We are also grateful to Dr. Changyong Lee for his invaluable supervision during the formal validation experiments.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

SUPPLEMENTARY MATERIALS

Supplementary data including one figure and one table can be found with this article online at https://doi.org/10.4196/kjpp.2024.28.6.527

REFERENCES

1. Lobo V, Patil A, Phatak A, Chandra N. Free radicals, antioxidants and functional foods: impact on human health. *Pharmacogn Rev*. 2010;4:118-126.
2. Ashok A, Andrabi SS, Mansoor S, Kuang Y, Kwon BK, Labhasetwar V. Antioxidant therapy in oxidative stress-induced neurodegenerative diseases: role of nanoparticle-based drug delivery systems in clinical translation. *Antioxidants (Basel)*. 2022;11:408.
3. Kim J, Lee C, Noh SG, Kim S, Chung HY, Lee H, Moon JO. Integrative transcriptomic analysis reveals upregulated apoptotic signaling in wound-healing pathway in rat liver fibrosis models. *Antioxidants (Basel)*. 2023;12:1588.
4. Kedare SB, Singh RP. Genesis and development of DPPH method of antioxidant assay. *J Food Sci Technol*. 2011;48:412-422.
5. Re R, Pellegrini N, Proteggente A, Pannala A, Yang M, Rice-Evans C. Antioxidant activity applying an improved ABTS radical cation decolorization assay. *Free Radic Biol Med*. 1999;26:1231-1237.
6. Du Z, Wang D, Li Y. Comprehensive evaluation and comparison of machine learning methods in QSAR modeling of antioxidant tripeptides. *ACS Omega*. 2022;7:25760-25771.
7. Shao L, Gao H, Liu Z, Feng J, Tang L, Lin H. Identification of antioxidant proteins with deep learning from sequence information.

Front Pharmacol. 2018;9:1036.
8. Jeličić ML, Kovačić J, Cvetnić M, Mornar A, Amidžić Klarić D. Antioxidant activity of pharmaceuticals: predictive QSAR modeling for potential therapeutic strategy. *Pharmaceuticals (Basel).* 2022;15:791.
9. Wiriyarattanakul A, Xie W, Toopradab B, Wiriyarattanakul S, Shi L, Rungrotmongkol T, Maitarad P. Comparative study of machine learning-based QSAR modeling of anti-inflammatory compounds from durian extraction. *ACS Omega*. 2024;9:7817-7826.
10. Mao J, Akhtar J, Zhang X, Sun L, Guan S, Li X, Chen G, Liu J, Jeon HN, Kim MS, No KT, Wang G. Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models. *iScience*. 2021;24:103052.
11. Carocho M, Ferreira IC. A review on antioxidants, prooxidants and related controversy: natural and synthetic compounds, screening and analysis methodologies and future perspectives. *Food Chem Toxicol*. 2013;51:15-25.
12. Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminform*. 2015;7:20.
13. Kruger F, Stiefl N, Landrum GA. rdScaffoldNetwork: the scaffold network implementation in RDKit. *J Chem Inf Model*. 2020;60:3331-3335.
14. Sordon S, Popłoński J, Milczarek M, Stachowicz M, Tronina T, Kucharska AZ, Wietrzyk J, Huszcza E. Structure-antioxidant-antiproliferative activity relationships of natural C7 and C7-C8 hydroxylated flavones and flavanones. *Antioxidants (Basel).* 2019;8:210.
15. Kong X, Liu C, Zhang Z, Cheng M, Mei Z, Li X, Liu P, Diao L, Ma Y, Jiang P, Kong X, Nie S, Guo Y, Wang Z, Zhang X, Wang Y, Tang L, Guo S, Liu Z, Li D. BATMAN-TCM 2.0: an enhanced integrative database for known and predicted interactions between traditional Chinese medicine ingredients and target proteins. *Nucleic Acids Res*. 2024;52:D1110-D1120.
16. Sakurai S, Kawakami Y, Kuroki M, Gotoh H. Structure–antioxidant activity (oxygen radical absorbance capacity) relationships of phenolic compounds. *Struct Chem*. 2022;33:1055-1062.
17. Noda Y, Kaneyuki T, Mori A, Packer L. Antioxidant activities of pomegranate fruit extract and its anthocyanidins: delphinidin, cyanidin, and pelargonidin. *J Agric Food Chem*. 2002;50:166-171.
18. Wang W, Xu H, Chen H, Tai K, Liu F, Gao Y. In vitro antioxidant, anti-diabetic and antilipemic potentials of quercetagetin extracted from marigold (Tagetes erecta L.) inflorescence residues. *J Food Sci Technol*. 2016;53:2614-2624.
19. Materska M, Konopacka M, Rogoliński J, Ślosarek K. Antioxidant activity and protective effects against oxidative damage of human cells induced by X-radiation of phenolic glycosides isolated from pepper fruits Capsicum annuum L. *Food Chem*. 2015;168:546-553.
20. Yousfi M, Djeridane A, Bombarda I; Chahrazed-Hamia; Duhem B, Gaydou EM. Isolation and characterization of a new hispolone derivative from antioxidant extracts of Pistacia atlantica. *Phytother Res*. 2009;23:1237-1242.
21. Kim SM, Kang K, Jho EH, Jung YJ, Nho CW, Um BH, Pan CH. Hepatoprotective effect of flavonoid glycosides from Lespedeza cuneata against oxidative stress induced by tert-butyl hyperoxide. *Phytother Res*. 2011;25:1011-1017.
22. Ding L, Zhang X, Zhang J. Antioxidant activity in vitro guided screening and identification of flavonoids antioxidants in the extract from *Tetrastigma hemsleyanum* Diels et Gilg. *Int J Anal Chem*. 2021;2021:7195125.
23. Estévez L, Mosquera RA. Molecular structure and antioxidant properties of delphinidin. *J Phys Chem A*. 2008;112:10614-10623.
24. Wagner AB. SciFinder Scholar 2006: an empirical analysis of research topic query processing. *J Chem Inf Model*. 2006;46:767-774.
25. Kilic I, Yeşiloğlu Y, Bayrak Y. Spectroscopic studies on the antioxidant activity of ellagic acid. *Spectrochim Acta A Mol Biomol Spectrosc*. 2014;130:447-452.
26. Tzen JTC. Strictinin: a key ingredient of tea. *Molecules*. 2023;28:3961.
27. Hossain H, Rahman SE, Akbar PN, Khan TA, Rahman MM, Jahan IA. HPLC profiling, antioxidant and in vivo anti-inflammatory activity of the ethanol extract of Syzygium jambos available in Bangladesh. *BMC Res Notes*. 2016;9:191.
28. Tu EC, Hsu WL, Tzen JTC. Strictinin, a major ingredient in Yunnan Kucha tea possessing inhibitory activity on the infection of mouse hepatitis virus to mouse L cells. *Molecules*. 2023;28:1080.