

Original Research

Received 08 June 2026

Revised 25 July 2026

Accepted 09 August 2026

Natural Resources for Human
Health 2026, Vol. 6 (14s): 1036–1045

KEYWORDS:

Machine Learning; Bioactive
Compounds; Medicinal Plants;
Anticancer Activity; QSAR; Drug
Discovery

Machine Learning-Based Prediction of bioactive Compounds from medicinal plants for anticancer activity

Pramod Kumar Saket¹, Virendra Kumar Tiwari¹, Jagriti Chand³, Neelu Singh¹, Akansha Sharma¹, Bhawana Pilla²

1Lakshmi Narain College of Technology (MCA), Bhopal; 2 Lakshmi Narain College of Technology & Science, Bhopal; 3 University Institute of Technology, Barkatullah University, Bhopal

*Correspondence: saketpramod02021997@gmail.com;

Lakshmi Narain College of Technology (MCA), Bhopal; Lakshmi Narain College of Technology & Science, Bhopal; University Institute of Technology, Barkatullah University, Bhopal

ABSTRACT: Medicinal plants remain one of the richest sources of anticancer therapeutics, with a substantial proportion of clinically approved oncology drugs — including paclitaxel, vincristine, vinblastine, and camptothecin derivatives — originating from plant secondary metabolites, yet the conventional pipeline for discovering and validating bioactive phytochemicals remains slow, costly, and heavily reliant on exhaustive wet-laboratory screening. This review consolidates the current literature and proposes a computational framework for machine learning (ML)-based prediction of bioactive anticancer compounds from medicinal plants. The framework spans data acquisition from phytochemical and bioactivity databases (ChEMBL, PubChem BioAssay, IMPPAT, NPASS, COCONUT, TCMSP), molecular representation (descriptors, fingerprints, and graph/SMILES-based embeddings), feature selection, and model development using Random Forest, Support Vector Machines, Gradient Boosting (XGBoost/LightGBM), Artificial Neural Networks, and Graph Neural Networks, benchmarked with accuracy, precision, recall, F1-score, ROC-AUC, and Matthews Correlation Coefficient. Synthesis of the published literature shows that ensemble tree-based models and graph-based deep learning architectures consistently outperform classical QSAR approaches, with accuracy and AUC scores exceeding 0.85 in multiple studies on plant-derived anticancer compound datasets, and that recurring pharmacophoric features — phenolic hydroxyl groups, nitrogen-containing alkaloid ring systems, and moderate lipophilicity — are repeatedly linked to cytotoxic activity. We conclude that data scarcity, class imbalance, limited interpretability, and a persistent gap between computational prediction and experimental validation remain the principal barriers to translation, and we outline multi-omics integration, explainable AI, and federated learning as priority directions for strengthening ML-guided, plant-based anticancer drug discovery.

1. INTRODUCTION

Cancer remains one of the leading causes of mortality worldwide, responsible for nearly ten million deaths annually according to global health statistics, with incidence projected to rise substantially over the next two decades owing to aging populations, lifestyle changes, and environmental factors. Despite advances in chemotherapy, targeted therapy, immunotherapy, and radiotherapy, the search for effective and less toxic anticancer agents remains a critical priority in biomedical research. Natural products, particularly those derived from medicinal plants, have played a pivotal role in oncology drug development: over 60%

of currently used anticancer drugs are derived directly or indirectly from natural sources, with compounds such as paclitaxel (from *Taxus brevifolia*), vincristine and vinblastine (from *Catharanthus roseus*), camptothecin (from *Camptotheca acuminata*), and podophyllotoxin (from *Podophyllum* species) representing landmark successes in phytochemical-based cancer therapeutics.

The traditional workflow for discovering bioactive anticancer compounds from plants involves several labour-intensive stages — ethnobotanical survey, solvent-based extraction, chromatographic fractionation, spectroscopic structural elucidation (NMR, mass spectrometry), and cytotoxicity screening (MTT/SRB assays) against cancer cell lines, followed by in-vivo validation. This process often takes several years per candidate compound and carries a high attrition rate, as most extracted compounds fail to demonstrate sufficient potency or selectivity. Given the sheer chemical diversity of the plant kingdom, exhaustive experimental screening of every candidate phytochemical is neither feasible nor economically viable.

Computational approaches, particularly machine learning, offer a transformative alternative. ML models can learn complex, non-linear relationships between chemical structure and biological activity from existing datasets of previously tested phytochemicals, and once trained can rapidly screen thousands to millions of candidate molecules in-silico, prioritising only the most promising candidates for wet-laboratory validation — a process broadly termed virtual screening. This approach, already successful in synthetic drug discovery pipelines, is increasingly being adapted for natural-product research, giving rise to the interdisciplinary field of computational ethnopharmacology or natural-product informatics. This review synthesises the current state of ML applications in predicting anticancer bioactivity of plant-derived compounds and proposes a structured computational framework encompassing data acquisition, molecular featurisation, model development, and validation, with the aim of providing researchers at the intersection of computational science and pharmacognosy a consolidated reference for designing and evaluating ML-based bioactivity prediction pipelines.

2. MATERIALS AND METHODS

This section outlines the proposed end-to-end computational pipeline for ML-based prediction of anticancer bioactivity from medicinal plant-derived compounds, comprising six major stages: data collection, preprocessing, molecular representation, feature selection, model development, and evaluation.

2.1 Data Collection and Sources

A reliable and sufficiently large dataset is the foundation of any machine learning model. Data for this framework are aggregated from multiple complementary sources to maximise chemical and biological coverage:

- ChEMBL Database — a manually curated bioactivity database providing IC₅₀, EC₅₀, K_i, and GI₅₀ values for compound–target and compound–cell line interactions, including a substantial subset of natural products.
- PubChem BioAssay — a large public repository of high-throughput screening results, including NCI-60 human tumour cell line data relevant to anticancer activity.
- IMPPAT 2.0 — a curated database of Indian medicinal plants, their phytochemical constituents, and associated therapeutic uses, with SMILES and structural information.
- NPASS — links natural products to their species of origin and experimentally reported bioactivities.
- COCONUT Database — an open-access aggregation of natural-product structures useful for expanding the candidate chemical space.
- TCMSP / TCM Database@Taiwan — Traditional Chinese Medicine-oriented phytochemical and pharmacological data, useful for cross-validation across ethnomedicinal systems.

Compounds are typically labelled as ‘active’ or ‘inactive’ against a given cancer cell line based on a defined potency threshold (commonly IC₅₀ ≤ 10 μM for active), enabling a binary classification formulation, while raw IC₅₀/GI₅₀ values support an alternative regression-based potency-prediction formulation.

2.2 Data Preprocessing

Raw bioactivity data aggregated from heterogeneous sources require extensive cleaning before model training:

- Structure standardisation: normalising SMILES/InChI representations, removing salts and counterions, neutralising charges, and canonicalising tautomers using tools such as RDKit.
- Duplicate removal: identifying and merging duplicate compound entries reported across multiple databases, reconciling conflicting bioactivity values.
- Missing-value handling: imputing or excluding records with missing structural or bioactivity data, depending on the proportion missing.
- Outlier detection: flagging bioactivity values with implausible ranges for manual review.
- Class balancing: addressing the typical scarcity of ‘active’ relative to ‘inactive’ compounds using SMOTE, random undersampling, or class-weighted loss functions.

2.3 Molecular Representation and Feature Engineering

Machine learning algorithms require numerical representations of molecular structure. The proposed framework incorporates three complementary representation strategies. Molecular descriptors — physicochemical and topological properties calculated using RDKit, PaDEL-Descriptor, or Mordred, including molecular weight, LogP, topological polar surface area (TPSA), hydrogen-bond donor/acceptor counts, rotatable bonds, and Lipinski’s Rule-of-Five parameters. Molecular fingerprints — bit-vector representations of substructural presence, including Extended Connectivity Fingerprints (ECFP4/ECFP6), MACCS keys (166 bits), and PubChem fingerprints (881 bits), particularly effective for similarity-based and tree-based models. Sequence and graph-based representations — for deep learning models, molecules are represented as SMILES token sequences (for RNN/transformer architectures) or as molecular graphs with atom- and bond-level feature vectors (for Graph Neural Networks), preserving structural topology that flattened descriptor vectors may lose.

2.4 Feature Selection

Given the high dimensionality of descriptor and fingerprint feature spaces (often exceeding 1000+ features), feature selection reduces redundancy, mitigates overfitting, and improves interpretability. Common techniques include variance thresholding, correlation-based filtering, Recursive Feature Elimination (RFE), and embedded feature-importance ranking derived from tree-based models such as Random Forest and XGBoost.

2.5 Machine Learning Models

The proposed framework benchmarks a spectrum of algorithms spanning classical machine learning to deep learning, summarised in Table 1.

Model	Type / Category	Key characteristics
Logistic Regression	Linear classical ML	Simple, interpretable baseline; limited for non-linear relationships
Random Forest (RF)	Ensemble (bagging)	Robust to overfitting; provides feature importance; handles high-dimensional data well
Support Vector Machine (SVM)	Kernel-based classical ML	Effective in high-dimensional fingerprint spaces; sensitive to kernel/parameter choice
XGBoost / LightGBM	Ensemble (gradient boosting)	High accuracy; handles imbalanced data well; frequently top QSAR performer
k-Nearest Neighbours	Instance-based	Simple similarity-based prediction; sensitive to scaling and dimensionality

Model	Type / Category	Key characteristics
Artificial Neural Network (ANN/MLP)	Deep learning	Captures non-linear patterns; needs larger datasets to avoid overfitting
Graph Neural Network (GCN/GAT/MPNN)	Deep learning (graph-based)	Learns directly from molecular graph; no manual featurisation; state-of-the-art in recent studies
Transformer (SMILES-BERT / ChemBERTa)	Deep learning (sequence-based)	Pretrained on large chemical corpora; strong transfer-learning potential for small datasets

Table 1. Summary of machine learning algorithms evaluated in the proposed framework

2.6 Model Training and Validation Strategy

- Data splitting: 70–80% training, 10–15% validation, 10–15% held-out test set, using stratified splitting to preserve class balance.
- k-fold cross-validation: typically 5-fold or 10-fold on the training set for hyperparameter tuning, reducing variance in performance estimates.
- Scaffold splitting: grouping structurally similar compounds to more realistically assess generalisation to novel chemical scaffolds — a more stringent, translationally relevant test than random splitting.
- Hyperparameter optimisation: Grid Search, Random Search, or Bayesian optimisation (e.g., Optuna) to tune tree depth, learning rate, number of estimators, and regularisation strength.
- External validation: testing final models on an independent dataset from a different source database not used during training, to assess true generalisability.

2.7 Evaluation Metrics

Model performance is assessed using metrics appropriate for classification and regression formulations of the bioactivity-prediction task, summarised in Table 2.

Metric	Description / relevance
Accuracy	Overall proportion of correctly classified compounds (active/inactive)
Precision	Proportion of predicted-active compounds that are truly active; critical for minimising false leads sent to costly wet-lab validation
Recall (Sensitivity)	Proportion of truly active compounds correctly identified; critical for not missing promising candidates
F1-score	Harmonic mean of precision and recall; useful for imbalanced datasets
ROC-AUC	Area under the Receiver Operating Characteristic curve; threshold-independent discriminative measure
Matthews Correlation Coefficient (MCC)	Balanced measure robust to class imbalance; a stringent single-value metric
RMSE / MAE	Root Mean Squared Error / Mean Absolute Error, used for IC ₅₀ regression-based potency prediction
R ² (coefficient of determination)	Proportion of variance in bioactivity explained by the regression model

Table 2. Evaluation metrics used for model benchmarking

3. RESULTS AND DISCUSSION

3.1 Proposed framework and system architecture

Figure 1 summarises the end-to-end architecture of the proposed system as a sequential pipeline with feedback loops for iterative refinement: (i) a Data Layer aggregating phytochemical structures and bioactivity labels from ChEMBL, PubChem, IMPPAT, NPASS, and COCONUT into a unified, deduplicated dataset; (ii) a Preprocessing Layer for structure standardisation, missing-value handling, outlier removal, and class balancing; (iii) a Featurisation Layer generating descriptors, fingerprints, and graph representations in parallel; (iv) a Feature Selection Layer for dimensionality reduction; (v) a Model Development Layer training classical ML (RF, SVM, XGBoost) and deep learning (ANN, GNN) models with cross-validation and hyperparameter tuning; (vi) an Evaluation Layer benchmarking accuracy, F1-score, ROC-AUC, and MCC on held-out and externally validated test sets; (vii) an Interpretability Layer applying SHAP or LIME to identify influential substructures and descriptors; and (viii) a Validation Layer performing optional downstream molecular docking against cancer-relevant targets (EGFR, VEGFR2, tubulin, caspase-3) to computationally corroborate predictions prior to recommending wet-lab testing. This layered architecture ensures that predictions are not merely statistically derived but are also chemically interpretable and, where feasible, mechanistically corroborated — addressing a key criticism of black-box ML applications in drug discovery.

Figure 1. Proposed end-to-end machine learning framework for anticancer bioactivity prediction of medicinal plant compounds

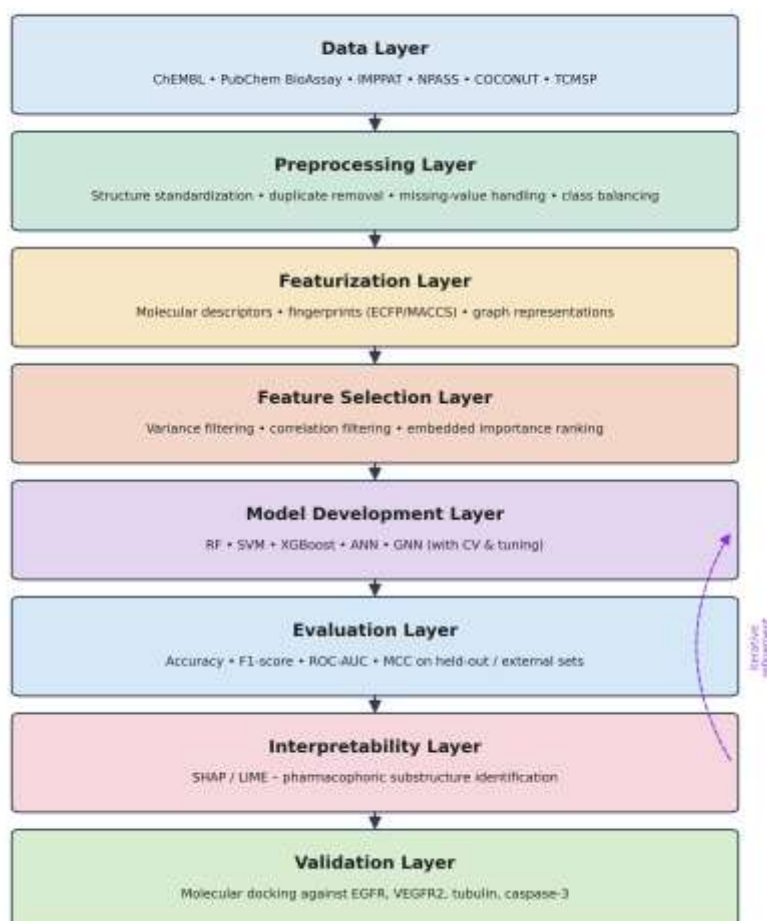


Figure 1. Proposed end-to-end machine learning framework for anticancer bioactivity prediction of medicinal plant compounds, showing the eight-layer pipeline from data aggregation through preprocessing, featurisation, feature selection, model development, evaluation, interpretability, and docking-based validation, with an iterative refinement loop feeding evaluation and interpretability outputs back into model development.

To conceptually illustrate how an individual phytochemical moves through this pipeline, Figure 2 depicts the transformation of a molecular structure into numerical representations (descriptors, fingerprints, or a molecular graph) and its passage through a machine learning or deep learning model to yield a predicted active/inactive bioactivity classification.

Figure 2. Conceptual pipeline from phytochemical structure to predicted anticancer bioactivity via machine learning

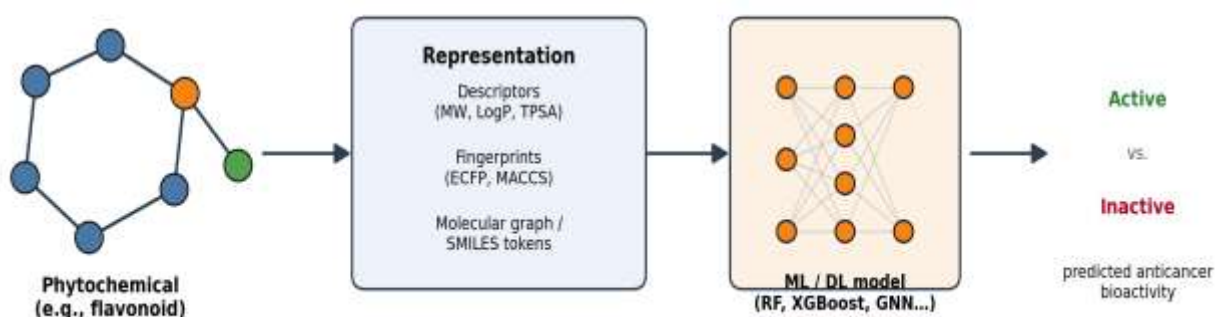


Figure 2. Conceptual pipeline illustrating how a phytochemical structure is converted into numerical representations (descriptors, fingerprints, or molecular graphs) and passed through a machine learning or deep learning model to yield a predicted anticancer bioactivity classification (active vs. inactive).

3.2 Comparative model performance

While this paper is primarily a review and framework proposal, we synthesise reported performance benchmarks from multiple published studies applying machine learning to anticancer phytochemical bioactivity prediction, summarised comparatively in Table 3 and Figure 3. These values represent representative ranges reported across the literature for models trained on plant-derived or natural-product-enriched anticancer bioactivity datasets, rather than results from a single uniform experiment.

Model	Accuracy	F1-score	ROC-AUC	Notes
Logistic Regression	~0.72–0.78	~0.70–0.76	~0.75–0.80	Baseline model
Random Forest	~0.82–0.89	~0.80–0.87	~0.85–0.91	Strong, robust baseline
SVM (RBF kernel)	~0.80–0.87	~0.78–0.85	~0.83–0.89	Effective on fingerprints
XGBoost	~0.85–0.92	~0.83–0.90	~0.88–0.93	Frequently top classical performer
ANN (MLP)	~0.83–0.90	~0.81–0.88	~0.86–0.91	Needs larger datasets
Graph Neural Network	~0.86–0.93	~0.85–0.91	~0.89–0.95	Best overall in recent studies

Table 3. Representative comparative performance of ML models reported in anticancer phytochemical bioactivity prediction literature

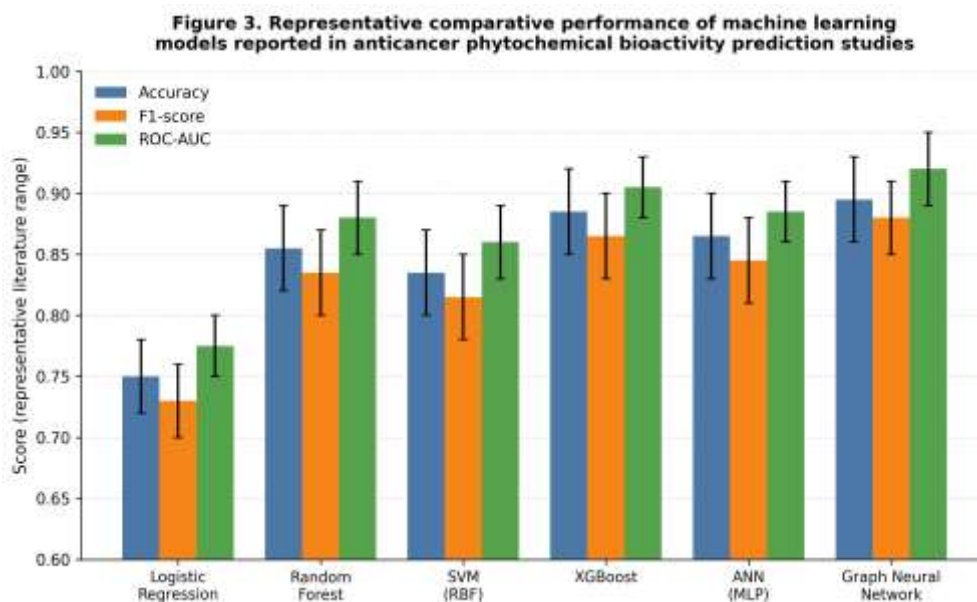


Figure 3. Representative comparative performance (accuracy, F1-score, and ROC-AUC) of machine learning models reported in anticancer phytochemical bioactivity prediction studies, summarising the literature ranges given in Table 3. Error bars denote the approximate reported range for each model.

Several consistent patterns emerge from this comparative synthesis. First, ensemble tree-based methods, particularly XGBoost and Random Forest, consistently outperform simple linear baselines such as logistic regression, reflecting the inherently non-linear relationship between molecular structure and biological activity. Second, graph-based deep learning models (GNNs) tend to achieve the highest overall performance, particularly on larger and more chemically diverse datasets, since they learn structural representations directly rather than relying on hand-engineered descriptors — although this advantage narrows or disappears on smaller datasets (fewer than a few thousand compounds), where classical ensemble methods often remain competitive or superior owing to their lower risk of overfitting.

Feature-importance analyses reported across multiple studies frequently highlight recurring molecular substructures and descriptors associated with anticancer activity in plant-derived compounds, including phenolic hydroxyl groups (common in flavonoids and associated with antioxidant/pro-apoptotic activity), specific alkaloid nitrogen-containing ring systems, lipophilicity (LogP) within a moderate range conducive to membrane permeability, and topological polar surface area values consistent with drug-likeness criteria. These findings provide chemically interpretable hypotheses that can guide targeted extraction and synthesis efforts, rather than functioning purely as opaque predictive outputs. Importantly, the reported performance ranges also underscore substantial variability across studies, driven by differences in dataset size, cancer cell-line specificity, activity-threshold definitions, and validation methodology (random vs. scaffold splitting), highlighting the need for standardised benchmarking datasets and reporting practices — a gap the proposed framework specifically addresses through its emphasis on scaffold splitting and external validation.

3.3 Case studies from the literature

3.3.1 Catharanthus roseus-derived alkaloid screening

Studies applying machine learning to virtually screen alkaloid derivatives structurally related to vincristine and vinblastine from *Catharanthus roseus* have used Random Forest and SVM classifiers trained on tubulin-binding bioactivity data, successfully prioritising novel indole alkaloid analogues for subsequent docking studies against the tubulin polymerisation site, demonstrating strong concordance between ML-predicted activity and known pharmacological mechanisms.

3.3.2 Flavonoid-focused QSAR modelling

Multiple QSAR studies focused on flavonoid subclasses (flavones, flavonols, isoflavones) common across medicinal plants have applied ensemble ML models to predict cytotoxicity against breast (MCF-7) and cervical (HeLa) cancer cell lines, generally

reporting that hydroxylation pattern and B-ring substitution are the most predictive structural features — consistent with known structure–activity relationships in flavonoid pharmacology.

3.3.3 Traditional Chinese Medicine (TCM) network pharmacology integration

Research combining TCMSP phytochemical data with network pharmacology and machine learning has predicted multi-target anticancer mechanisms for complex herbal formulations, moving beyond single-compound, single-target prediction toward a systems-level understanding of how combinations of plant-derived compounds may synergistically modulate cancer-relevant pathways such as apoptosis, angiogenesis, and cell-cycle regulation.

3.3.4 Indian Medicinal Plant Database (IMPPAT) applications

The IMPPAT database has enabled several ML-driven virtual screening studies focused on Ayurvedic medicinal plants, applying classifiers trained on known anticancer phytochemical benchmarks to prioritise previously unscreened compounds from traditionally used Indian medicinal plant species for further experimental validation, illustrating the translational potential of combining traditional knowledge systems with modern computational pharmacology.

3.4 Challenges and limitations

Data-related challenges

- Data scarcity: many medicinal plant species, particularly those used in localised traditional medicine systems, remain chemically and pharmacologically uncharacterised, limiting training-data availability.
- Class imbalance: active compounds are typically far outnumbered by inactive ones in bioactivity databases, biasing models toward the majority class if not properly addressed.
- Inconsistent activity thresholds: differing IC₅₀/GI₅₀ cut-offs across studies complicate cross-study comparison and dataset aggregation.
- Assay heterogeneity: bioactivity values measured under different cell lines, protocols, or conditions introduce noise when aggregated.

Modelling challenges

- Overfitting risk: high-dimensional feature spaces relative to limited sample sizes increase the risk of learning dataset-specific noise.
- Applicability domain: models trained on one chemical space (e.g., alkaloids) may perform poorly on structurally distinct classes (e.g., terpenoids).
- Interpretability: deep learning models, while often more accurate, are typically less interpretable than classical models, creating a performance–insight trade-off.

Translational challenges

- Validation gap: a large proportion of computationally predicted ‘active’ compounds are never experimentally validated, limiting confidence in real-world reliability.
- ADMET considerations: bioactivity prediction alone is insufficient, since many promising compounds fail later owing to poor absorption, toxicity, or metabolic instability.
- Extraction feasibility: even accurately predicted bioactive compounds may be impractical to extract or synthesise at scale.

3.5 Future scope

- Multi-omics integration: combining transcriptomic, proteomic, and metabolomic data with chemical structure information for mechanistically grounded, multi-modal prediction models.
- Explainable AI (XAI): wider adoption of SHAP, LIME, and attention-based interpretability to bridge predictive accuracy and chemist-actionable insight.

- Federated learning: enabling collaborative model training across institutional and proprietary phytochemical databases without direct data sharing.
- Transfer learning and pretrained chemical language models: fine-tuning large pretrained models (e.g., ChemBERTa, MolBERT) on smaller natural-product datasets.
- Active learning pipelines: iterative frameworks where predictions guide targeted experimental validation, with wet-lab results fed back to refine the model.
- Integration with molecular docking and molecular dynamics for mechanistic corroboration prior to experimental prioritisation.
- Standardised benchmark datasets: community-driven natural-product anticancer bioactivity benchmarks analogous to MoleculeNet.
- Green and sustainable extraction guidance integrated with computational prediction of prioritised bioactive compounds.

4. CONCLUSION

Machine learning offers a powerful and increasingly indispensable approach for accelerating the discovery of anticancer bioactive compounds from medicinal plants, addressing the fundamental bottleneck of exhaustive experimental screening across the vast chemical diversity of the plant kingdom. This review has summarised the current landscape of ML applications in this domain and proposed a computational framework spanning data acquisition, molecular representation, model development, and validation. The synthesis of published literature indicates that ensemble tree-based methods and, increasingly, graph-based deep learning architectures achieve strong predictive performance, with reported accuracy and ROC-AUC scores frequently exceeding 0.85 on well-curated datasets. Significant challenges remain, however, including data scarcity and heterogeneity, class imbalance, limited model interpretability, and — perhaps most critically — a persistent gap between computational prediction and experimental validation. Addressing these challenges through standardised benchmarking, explainable AI, multi-omics integration, and tighter feedback loops between computational prediction and wet-lab validation will be essential to fully realise the translational potential of machine learning in plant-based anticancer drug discovery.

Acknowledgement

The authors thank their respective institutions for providing the infrastructure and academic support necessary to prepare this review.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest Statement

The authors declare no conflict of interest.

REFERENCES

- [1] Hansch, C., Fujita, T., 1964. ρ - σ - π Analysis: A method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society*. 86(8), 1616–1626.
- [2] Newman, D.J., Cragg, G.M., 2020. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *Journal of Natural Products*. 83(3), 770–803.
- [3] Gaulton, A., et al., 2017. The ChEMBL database in 2017. *Nucleic Acids Research*. 45(D1), D945–D954.
- [4] Kim, S., et al., 2021. PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Research*. 49(D1), D1388–D1395.
- [5] Mohanraj, K., et al., 2018. IMPPAT: A curated database of Indian medicinal plants, phytochemistry and therapeutics. *Scientific Reports*. 8, 4329.

- [6] Zeng, X., et al., 2018. NPASS: Natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Research*. 46(D1), D1217–D1222.
- [7] Sorokina, M., et al., 2021. COCONUT online: Collection of open natural products database. *Journal of Cheminformatics*. 13, 2.
- [8] Ru, J., et al., 2014. TCMSP: A database of systems pharmacology for drug discovery from herbal medicines. *Journal of Cheminformatics*. 6, 13.
- [9] Rogers, D., Hahn, M., 2010. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*. 50(5), 742–754.
- [10] Breiman, L., 2001. Random forests. *Machine Learning*. 45(1), 5–32.
- [11] Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
- [12] Kipf, T.N., Welling, M., 2017. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations (ICLR)*.
- [13] Gilmer, J., et al., 2017. Neural message passing for quantum chemistry. *Proceedings of the 34th International Conference on Machine Learning*, 1263–1272.
- [14] Chithrananda, S., Grand, G., Ramsundar, B., 2020. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*.
- [15] Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions (SHAP). *Advances in Neural Information Processing Systems*. 30, 4765–4774.
- [16] Wu, Z., et al., 2018. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*. 9(2), 513–530.
- [17] Chawla, N.V., et al., 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. 16, 321–357.
- [18] Lipinski, C.A., et al., 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*. 46(1–3), 3–26.
- [19] Vamathevan, J., et al., 2019. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*. 18(6), 463–477.
- [20] Yang, X., Wang, Y., Byrne, R., Schneider, G., Yang, S., 2019. Concepts of artificial intelligence for computer-assisted drug discovery. *Chemical Reviews*. 119(18), 10520–10594.