

# Deep Learning for Intelligent Classification of Medicinal Plants Using Cell Wall Component and Metabolomic Features

Merk Heyes

Department of Computer Science, University of New Hampshire, Durham, NH, USA.

markhayes@unh.edu

Lucas M. Steele

Department of Electrical Engineering and Computer Science, University of Missouri,

Columbia, MO, USA.

lucasmsteele@missouri.edu

Zhengjin Gu

Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.

zhengmail@buffalo.edu

## Abstract

The reliable classification of medicinal plants is essential for pharmaceutical quality control, botanical safety, and regulatory compliance. Traditional morphological identification is increasingly complemented by biochemical profiling based on cell wall components and metabolomic signatures. However, these data sources are heterogeneous, high dimensional, and often noisy, which limits the effectiveness of classical statistical classifiers. Deep learning offers a powerful framework for discovering multilayered representations across such multimodal chemical and structural features. This paper examines the design of intelligent classification systems for medicinal plants from a systems-oriented perspective. It addresses the integration of cell wall component profiles and metabolomic data, the selection of appropriate deep learning architectures, and the infrastructure required for reproducible model development and deployment. The discussion emphasizes structural trade-offs among model capacity, interpretability, data heterogeneity, computational cost, and regulatory readiness. Rather than focusing narrowly on algorithmic performance, the paper analyzes the broader socio-technical context, including data governance, fairness across underrepresented plant taxa, robustness under environmental and instrumental variation, and the sustainability of classification infrastructures. The paper further considers deployment pathways in botanical research, herbal supply chains, and pharmacovigilance. It argues that intelligent medicinal plant classification should be designed as a governed, auditable, and multimodal learning system rather than as an isolated pattern recognition task. Future directions include federated model training, standardized spectral and metabolomic exchange formats, uncertainty-aware inference, and closer alignment with regulatory frameworks for botanical identification and quality assurance.

## Keywords

medicinal plant classification; deep learning; metabolomics; cell wall components; multimodal fusion; system architecture; governance; quality assurance.

## 1. Introduction

Medicinal plants support a substantial share of global healthcare and pharmaceutical discovery, yet their authentication remains a persistent challenge. Misidentification, adulteration, and substitution can compromise therapeutic efficacy and introduce safety risks. Classical botanical approaches rely on morphological keys, histological examination, and increasingly on genetic markers, but these methods do not always reflect the biochemical state or medicinal quality of a sample. Cell wall components and metabolomic profiles can provide rich chemical fingerprints that vary with genetic identity, tissue type, cultivation conditions, harvest time, and postharvest handling. As analytical instruments become more accessible, these phenotypic data streams are growing in volume and complexity. Deep learning has emerged as a transformative computational approach for extracting patterns from large, high-dimensional datasets across many scientific domains [1]. When applied to plant biochemistry, deep learning can combine structural and metabolic information in ways that improve classification accuracy and reveal latent biochemical relationships. However, the construction of such systems requires more than model training. It involves careful attention to data acquisition, feature engineering, system architecture, governance, fairness, and downstream deployment. This paper develops a systems-level analysis of deep learning for intelligent classification of medicinal plants using cell wall component and metabolomic features. It does not propose a single optimal algorithm; instead, it examines structural and organizational considerations that determine whether such systems can be robust, interpretable, and institutionally credible.

## **2. Background and Related Work**

Metabolomics has matured into a central discipline for linking chemical phenotypes to genetic and environmental variation. Early definitions emphasized the comprehensive measurement of small-molecule metabolites as a functional readout of biological systems [2]. In plant science, metabolomic approaches have been used to compare species, tissues, developmental stages, and stress responses, offering a chemical complement to transcriptomic and proteomic data [3]. The analytical foundation typically combines chromatographic separation with mass spectrometry or nuclear magnetic resonance spectroscopy, generating large spectral datasets that require sophisticated processing and statistical modeling [4]. At the same time, cell wall chemistry provides a complementary layer of information. The relative abundance of cellulose, hemicellulose, pectin, lignin, and characteristic monosaccharides can differ across species and may contribute taxonomically relevant markers. Unlike metabolomic profiles that can be highly dynamic, cell wall composition can provide relatively stable structural signatures, making it valuable for authentication. Combining these two data modalities is biologically appealing because metabolic state and structural polysaccharide composition reflect partially overlapping but not identical aspects of plant identity.

Recent advances in deep learning have expanded the analytical toolbox for such complex data. Convolutional neural networks have proven effective for grid-like data such as spectral matrices and chemical images [5]. Deeper architectures with residual connections address optimization difficulties and enable the learning of hierarchical representations from complex inputs [6]. Recurrent neural networks, especially long short-term memory networks, are suitable for sequential forms of analytical data such as chromatographic elution profiles [7]. Attention mechanisms further enhance the ability of models to focus on informative spectral regions or feature interactions across heterogeneous inputs [8]. These techniques provide the computational language for integrating cell wall and metabolomic data, but their successful use in medicinal plant classification demands more than algorithmic selection. It requires a

coherent system design that aligns data representation, model capacity, and institutional requirements.

### **3. Data Acquisition and Feature Engineering for Medicinal Plant Characterization**

The construction of a robust classification system begins with data acquisition and feature engineering. Medicinal plant specimens may be collected from wild populations, cultivated fields, herbarium collections, or commercial supply chains. Each source introduces distinct sources of variation. For cell wall characterization, samples are typically subjected to fractionation procedures that isolate polysaccharides and lignin, followed by hydrolysis and chromatographic quantification of monosaccharide residues. For metabolomic profiling, extraction protocols, derivatization, separation, and detection parameters must be standardized to reduce technical drift. Spectral alignment, baseline correction, peak detection, and normalization are essential preprocessing stages. Because cell wall and metabolomic data have different units, dynamic ranges, and noise structures, naive concatenation can harm model learning. Dimensionality reduction tools such as t-distributed stochastic neighbor embedding provide exploratory evidence about whether biologically meaningful clusters emerge before supervised training [9]. Standard machine learning libraries offer reproducible pipelines for scaling, splitting, and evaluating such high-dimensional feature sets [10].

A deeper problem is that data quality can vary across laboratories, instruments, and ecological contexts. Batch effects are common in metabolomic studies and can dominate biological signal if not modeled carefully. Data management practices should therefore follow the FAIR principles of findability, accessibility, interoperability, and reusability [14]. This is especially important for medicinal plant datasets that may need to support regulatory decisions. In the study of *Salvia miltiorrhiza* and its close relatives, cell wall component characteristics have been proposed as identification markers, demonstrating that structural polysaccharide profiles can carry taxonomically useful information [15]. Such findings support the integration of cell wall features with metabolomic profiles, but they also highlight the need for standardized reference data and well-characterized plant vouchers. Feature engineering should preserve provenance, instrument metadata, and sample preparation details because these contextual variables shape the distribution of the learned representations. Without such contextual discipline, deep learning models may learn laboratory-specific artifacts instead of biologically meaningful differences.

### **4. Deep Learning Architectures for Multimodal Phytochemical Classification**

The choice of deep learning architecture for medicinal plant classification depends on the structure of each data modality and the degree to which the modalities should interact. One common design uses separate encoders for cell wall features and metabolomic features, followed by a fusion layer that combines the learned representations for final classification. This approach allows each encoder to specialize in its own feature type before integration. One-dimensional convolutional layers can process spectral sequences, while fully connected or residual blocks can process tabular cell wall composition vectors. Residual learning helps stabilize training when the model contains many layers [6]. Attention mechanisms are particularly useful for metabolomic data because they allow the network to emphasize informative retention times or mass-to-charge regions [8]. Recurrent networks remain relevant for chromatographic profiles with sequential structure [7]. For tabular features with modest dimensionality, deep fully connected networks may offer strong performance, but they can also be prone to overfitting when training data are limited.

Optimization and software infrastructure also shape model development. Adaptive moment estimation is widely used for training deep networks because it combines momentum and per-parameter learning rate adjustments [11]. Production-scale experimentation benefits from frameworks that support automatic differentiation, distributed training, and model serialization. TensorFlow provides a mature ecosystem for serving and mobile deployment [12]. PyTorch offers a more imperative programming style and has become a dominant platform for research prototyping [13]. Contemporary engineering treatments emphasize iterative prototyping, rigorous evaluation, and deployment-ready model development [22]. In multimodal systems, a key structural question is whether fusion should occur early, late, or through multiple gated interactions. Early fusion can capture fine-grained cross-modal correlations, but it is highly sensitive to missing data and scaling differences. Late fusion is modular and easier to maintain, but it may miss interactions that are only visible when cell wall and metabolomic information are considered jointly. Attention-based fusion offers a middle path by allowing the network to weight contributions from each modality dynamically. The trade-off is that attention mechanisms increase architectural complexity and may reduce interpretability. A systems perspective therefore considers not only classification accuracy but also maintainability, transparency, and the ease of validating the model for regulatory use.

## **5. System Architecture and Infrastructure Design**

Deploying deep learning for medicinal plant classification at scale requires a carefully designed infrastructure that supports data ingestion, model training, evaluation, and inference. A reference architecture often includes a data lake or distributed storage layer for raw analytical files, a metadata catalog for sample provenance, a feature store for standardized training features, and a model registry for tracking trained models. This design enables reproducibility because every model can be linked to the exact data version, feature transformation, and training configuration used to create it. Containerized workflows support portability across cloud, laboratory, and edge environments. Data lineage and provenance tracking are especially important in botanical applications because the ultimate decision may involve safety-critical products such as herbal medicines. The FAIR principles provide guidance for structuring data and metadata so that they remain useful across institutional boundaries [14].

A major structural trade-off in system design is the division of labor between central cloud infrastructure and edge devices. Central training can use large batches and distributed accelerators, but field classification may require real-time decisions with limited connectivity. Edge deployment reduces data transfer and enables immediate feedback in field surveys or quality inspection points, but it demands compressed models, efficient inference runtimes, and careful power management. Hybrid architectures can separate model development from inference, with periodic synchronization of updated models to remote devices. Model monitoring is also essential because analytical instruments drift over time, and supplier sources can introduce new adulterants or closely related species. Performance metrics computed during development may not reflect performance after deployment. Therefore, a robust system should include drift detection, scheduled retraining, and a mechanism for escalating uncertain predictions to expert review. These infrastructural concerns are not peripheral; they determine whether the classification system remains trustworthy beyond a single laboratory demonstration.

## **6. Interpretability, Robustness and Fairness**

Interpretability is a central requirement for intelligent medicinal plant classification because the outputs may inform quality control decisions or regulatory action. Deep networks can be interrogated through gradient-based visualization methods that highlight the input regions most responsible for a prediction. Grad-CAM, originally developed for image classification, has been adapted to spectral and temporal inputs, enabling researchers to visualize which metabolomic regions or cell wall features drive a classification [16]. Model cards offer a structured reporting mechanism for documenting intended use, training data, evaluation metrics, and known limitations [17]. Such documentation is particularly valuable when models are transferred across institutions or used by non-specialists.

Robustness and fairness are also critical. A model trained predominantly on common cultivated species may perform poorly on wild relatives or rare medicinal plants, leading to biased performance across taxa. Fairness in machine learning for health-related applications requires deliberate attention to subgroup performance, because average accuracy can mask important disparities [19]. For botanical classification, fairness includes ensuring that underrepresented medicinal plants from specific ecological or cultural contexts are not systematically misclassified. Environmental variation, harvest season, and storage conditions can shift biomolecular profiles, and models must be evaluated under such distributional shifts. Adversarial robustness, while less commonly discussed in phytochemistry, is relevant when economic incentives exist for adulteration. A system should therefore be tested against plausible perturbations that mimic realistic sample variation rather than only against random noise. Ethical frameworks for artificial intelligence emphasize beneficence, non-maleficence, autonomy, justice, and explicability [18]. These principles apply to medicinal plant classification because incorrect identification can have direct health consequences. Interpretability, robustness, and fairness must therefore be designed into the system rather than treated as post hoc checks.

## **7. Governance, Deployment and Policy Implications**

Governance arrangements shape how intelligent classification systems are used in botanical research, herbal supply chains, and regulatory science. A model that classifies medicinal plants should not be treated as a standalone diagnostic device. It should be embedded within a quality management system that specifies training data requirements, validation protocols, change control, and human oversight. High-performance artificial intelligence in medicine has shown that algorithmic tools must be integrated into clinical and organizational workflows to produce reliable outcomes [20]. Similar principles apply to botanical quality assurance. Machine learning applications increasingly require attention to algorithmic selection, data pipelines, and operational monitoring across domains [21]. In the medicinal plant sector, this means that botanists, chemists, informaticians, and regulators must jointly define the criteria for acceptable evidence.

Policy implications extend beyond technical validation. Many medicinal plant species are sourced from biodiversity-rich regions, and their data may carry cultural and economic significance. The development of large-scale classification systems should respect data sovereignty and ensure that benefits are shared equitably with source communities. International standards for analytical data exchange, voucher specimen reporting, and model documentation would support interoperability among laboratories and regulatory bodies. From a sustainability perspective, deep learning systems should be evaluated not only according to predictive performance but also according to energy consumption, data storage costs, and the environmental footprint of redundant large-scale experiments. Lightweight

models and federated learning can reduce centralization risks and allow institutions to collaborate without transferring sensitive raw data. Deployment in pharmacopeial contexts will require uncertainty quantification, explainable outputs, and audit trails that connect each classification to its supporting analytical evidence. The long-term value of intelligent medicinal plant classification lies in its capacity to strengthen quality assurance while remaining responsive to ecological, institutional, and ethical constraints.

## 8. Conclusion

Deep learning offers a promising pathway for integrating cell wall component and metabolomic features into intelligent medicinal plant classification systems. The combination of stable structural polysaccharide markers with dynamic metabolomic profiles can produce richer biochemical representations than either modality alone. However, the design of such systems requires a broad interdisciplinary perspective. Attention must be given to data acquisition, feature engineering, multimodal fusion, interpretability, infrastructure, fairness, and regulatory alignment. A purely performance-oriented approach risks producing models that fail outside controlled laboratory settings. Instead, medicinal plant classification should be treated as a governed socio-technical system in which analytical chemistry, deep learning, software engineering, and regulatory science are mutually constitutive. Future research should develop standardized multimodal benchmark datasets, uncertainty-aware inference methods, federated learning frameworks for cross-institutional collaboration, and lightweight models suitable for field deployment. By addressing these structural and policy dimensions, the field can move toward classification systems that are not only accurate but also transparent, robust, sustainable, and socially accountable.

## References

1. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
2. Fiehn, O. (2002). Metabolomics—the link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1–2), 155–171.
3. Sumner, L. W., Mendes, P., & Dixon, R. A. (2003). Plant metabolomics: Large-scale phytochemistry in the functional genomics era. *Phytochemistry*, 62(6), 817–836.
4. Wolfender, J.-L., Rudaz, S., Choi, Y. H., & Kim, H. K. (2013). Plant metabolomics: From holistic data to relevant biomarkers. *Current Medicinal Chemistry*, 20(8), 1056–1090.
5. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097–1105.
6. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).
7. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
9. van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605.

10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
11. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.
12. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., & Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation* (pp. 265–283).
13. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 8024–8035.
14. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S. A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018.
15. Zhao, K., Li, G., Li, C., Jiang, T., Wang, W., & Yang, Z. (2026). Identification Markers for *Salvia miltiorrhiza* and Its Close Relatives Based on Cell Wall Component Characteristics. *Engineered Science*, 40, 2122.
16. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626).
17. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229).
18. Floridi, L., & Cows, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1).
19. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169(12), 866–872.
20. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56.

21. Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN Computer Science, 2(3), 160.
22. Chollet, F. (2021). Deep learning with Python (2nd ed.). Manning Publications.