

AI-Assisted Chemotaxonomic Identification of *Salvia* Species Based on Cell Wall and Phytochemical Profiles

Claudio Richards

Department of Computer Science, Binghamton University, Binghamton, NY, USA.

claudiomail@binghamton.edu

Chetan M. Yadav

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

yadavchetan@unh.edu

Christopher M. Bowman

Department of Computer Science, University of Houston, Houston, TX, USA.

christopher.bowman256@uh.edu

Abstract

The identification of *Salvia* species is a central problem in medicinal plant authentication, biodiversity monitoring, and phytochemical standardization. Traditional morphological approaches are often insufficient for closely related taxa, while chemical profiling offers diagnostic information that can be difficult to interpret at scale. This paper examines AI-assisted chemotaxonomic identification based on integrated cell wall component profiles and phytochemical fingerprints. Rather than focusing solely on classifier performance, the discussion adopts a system-level perspective that encompasses data acquisition, multimodal fusion, model architecture, technical debt, governance, robustness, fairness, deployment, sustainability, and policy alignment. The paper argues that reliable identification depends not only on algorithmic capacity but also on the design of surrounding infrastructure for data provenance, model auditing, and human oversight. It reviews architectural trade-offs between centralized and distributed deployment, the role of interpretability methods in supporting domain expert judgment, and the risks of bias arising from unrepresentative training collections. Cross-domain lessons from responsible machine learning in health care and ecological monitoring are used to evaluate governance strategies for botanical identification systems. The analysis highlights the need for voucher-linked digital reference libraries, structured dataset documentation, prospective validation, and policy frameworks that respect international biodiversity governance obligations. The paper concludes that AI-assisted chemotaxonomic identification can be scientifically valuable and institutionally sustainable when technical design is integrated with governance arrangements and long-term capacity building.

Keywords

Salvia identification, chemotaxonomy, cell wall profiling, phytochemical fingerprinting, artificial intelligence, machine learning governance, explainability, system architecture.

1. Introduction

The reliable identification of medicinal plant species is a foundational problem in pharmacognosy, biodiversity informatics, and ecological conservation. Traditional taxonomic identification has depended on morphological characters, reproductive structures, and geographic provenance, but these criteria are often insufficient for closely related species with overlapping macroscopic features. Statistical learning approaches offer formal tools for building classification systems that can integrate heterogeneous evidence into probabilistic decisions [1]. Deep learning, in particular, has transformed pattern recognition across domains by learning hierarchical representations from high-dimensional inputs [2]. Ensemble methods such as random forests have further provided robust decision boundaries in settings where the number of samples is moderate relative to feature dimensionality [3]. Attention-based architectures have also made it possible to align and integrate multiple feature streams without hand-crafted fusion rules [4]. These advances suggest that artificial intelligence can support chemotaxonomic identification when the underlying biological measurements are appropriately curated and modeled.

Chemotaxonomy exploits chemical variation as a systematic character, and *Salvia* species are especially suitable candidates because they produce diverse phenolic acids, diterpenoids, essential oils, and polysaccharide-rich cell walls. Cell wall component characteristics have been shown to offer identification markers for *Salvia miltiorrhiza* and its close relatives [5]. Phytochemical fingerprints obtained through chromatographic and spectroscopic methods provide complementary signals, but both data types pose challenges of high dimensionality, batch variability, and semantic granularity. The system-level problem is therefore not merely the selection of a classifier; it is the design of an integrated infrastructure that supports data acquisition, feature harmonization, model training, auditing, and deployment across institutional boundaries. At the same time, the use of artificial intelligence in biological identification raises governance questions concerning data provenance, reproducibility, fairness, and the responsible allocation of computational resources [6].

This paper examines the AI-assisted chemotaxonomic identification of *Salvia* species from a systems perspective. Rather than presenting a single experimental pipeline, the discussion emphasizes structural trade-offs across the machine learning life cycle, including multimodal data fusion, model interpretability, technical debt, regulatory alignment, and long-term sustainability. The analysis draws on cross-domain lessons from applied machine learning in medicine, environmental monitoring, and public sector automation. The remainder of the paper is organized as follows. Section 2 reviews chemotaxonomic data foundations. Section 3 analyzes system architecture and technical debt. Section 4 addresses data governance and provenance infrastructure. Section 5 discusses robustness, fairness, and explainability. Section 6 examines deployment, sustainability, and policy implications. Section 7 identifies forward-looking research directions, and Section 8 concludes the paper.

2. Chemotaxonomic Data Foundations and Acquisition

The information substrate for automated *Salvia* identification consists of two distinct but complementary families of measurements. The first family includes cell wall polysaccharide composition, monosaccharide ratios, lignin content, pectin structure, and related cell wall architectural traits. These traits are relatively stable within taxa and can discriminate species that are otherwise morphologically similar. The second family includes phytochemical profiles derived from liquid chromatography, gas chromatography, mass spectrometry, or vibrational spectroscopy. Each data stream captures a different dimension of biosynthetic capacity and ecological adaptation. However, the integration of these streams is non-trivial

because cell wall data are often compositional and expressed as relative percentages, whereas phytochemical data may be represented as peak intensities with varying signal-to-noise ratios and retention time drift. A robust system must therefore normalize data without destroying the compositional constraints that contain diagnostic information.

A central concern in data acquisition is the representativeness of the training corpus. Herbarium specimens, cultivated accessions, and wild-collected samples may differ systematically in age, storage conditions, extraction protocols, and geographic origin. If the training distribution overrepresents a small number of cultivated varieties, the resulting model may perform poorly on wild populations and may exhibit systematic misclassification across lineages [7]. The system design should treat sample provenance and preparation protocol as first-class metadata rather than as auxiliary annotations. This perspective aligns with broader recognition that machine learning bias frequently originates in data collection practices, not solely in model optimization. A chemotaxonomic identification system must therefore include rigorous sample inclusion criteria, population stratification, and documentation of batch effects.

In addition to raw chemical measurements, feature engineering plays an important role. Unsupervised dimensionality reduction can reveal latent structure, while supervised feature selection can identify diagnostic markers that are interpretable to taxonomists. However, an overreliance on automated feature selection without botanical domain review risks producing models that exploit spurious correlations, such as instrument drift or source location. Therefore, the curation workflow should include domain experts who evaluate whether selected features correspond to biologically plausible chemotaxonomic characters. This human-in-the-loop design is a governance choice as much as an analytical choice, and it affects the portability of the model across laboratories and field settings.

3. System Architecture for AI-Assisted Identification

A production-oriented AI-assisted identification system must balance several architectural requirements: modularity, reproducibility, interpretability, and operational efficiency. The core pipeline generally includes data ingestion, quality control, feature extraction, model training, explanation generation, and decision support. Within this pipeline, explanation modules are not ancillary; they provide the evidentiary trail that allows taxonomists and regulators to evaluate why a particular species assignment was made. Post hoc attribution methods such as Shapley additive explanations can quantify feature contributions for individual predictions [8]. Gradient-based localization techniques, originally developed for convolutional image analysis, can be adapted to spectral and chromatographic inputs to highlight regions of the chemical profile that most strongly influence classification [9]. These methods increase the transparency of otherwise opaque neural models, but they also introduce additional computational dependencies that must be maintained under version control.

The architectural choice between a centralized cloud service and a distributed edge deployment is a structural trade-off with direct implications for data sovereignty and operational resilience. Centralized architectures simplify model update and quality assurance but require reliable connectivity and may raise concerns about the movement of sensitive biodiversity data across jurisdictions. Edge architectures can support field identification in low-connectivity environments, yet they constrain model size and complicate the collection of monitoring data. Hybrid designs often adopt federated learning or model distillation approaches, but these alternatives add synchronization overhead and governance complexity.

Technical debt is a pervasive risk in machine learning systems because the interactions among data processing code, feature extraction routines, model versioning, and explanation layers can become tightly coupled [10]. In chemotaxonomic identification, a change in chromatographic column chemistry or sample preparation can alter the input distribution and invalidate previously trained models. Consequently, the architecture must treat data pipeline code, model weights, feature schemas, and evaluation datasets as versioned artifacts. Continuous monitoring of data drift and periodic revalidation against reference specimens are essential. The cost of these practices should be recognized at the design stage rather than retrofitted after field failures occur.

4. Data Governance, Provenance, and Infrastructure

Governance for AI-assisted botanical identification extends beyond algorithmic fairness to include the legal, ethical, and institutional dimensions of data stewardship. Lessons from responsible machine learning in health care are directly relevant because both domains handle heterogeneous biological data, class imbalance, and high-stakes decisions [11]. Clinical machine learning research has emphasized the need for prospective validation, transparent reporting, and the involvement of domain experts in model development [12]. These principles transfer naturally to chemotaxonomic systems, where misidentification can affect herbal product safety, conservation planning, and regulatory enforcement.

A foundational governance mechanism is the use of structured documentation for datasets. Datasheets for datasets provide a template for recording motivation, composition, collection procedures, preprocessing, recommended uses, and limitations [13]. In the context of *Salvia* identification, such documentation should include voucher specimen identifiers, geographic coordinates at an appropriate resolution, extraction protocols, instrument parameters, and known taxonomic revisions. Without this metadata, independent auditors cannot determine whether a model has been evaluated on a representative sample or whether its performance estimates are optimistic due to data leakage.

Infrastructure design also influences governance outcomes. A decentralized repository with strong provenance tracking may increase transparency, but it raises maintenance costs and requires common data models across institutions. Conversely, a centralized data lake can enforce standardized curation but may become a bottleneck and a single point of failure. The selected infrastructure must support access controls that respect the rights of source countries and Indigenous communities under biodiversity governance frameworks. Data sovereignty, benefit sharing, and prior informed consent are therefore not external policy considerations but structural requirements that shape system architecture.

5. Robustness, Fairness, and Explainability

The robustness of chemotaxonomic classifiers depends on their ability to maintain reliable performance under distribution shift, label noise, and accidental or adversarial perturbations. Ethical analysis of algorithmic systems has highlighted the importance of anticipating disparate impacts and of auditing models beyond aggregate accuracy [14]. In species identification, disparate impacts may occur when a model performs well for economically important taxa but poorly for rare, wild, or less-studied relatives. Formal fairness criteria, such as awareness-based notions of individual fairness, require that similar inputs receive similar outputs, which is particularly challenging when chemical profiles vary continuously within a species [15]. These formal frameworks encourage system designers to measure

performance across taxonomic or geographic partitions rather than relying solely on overall confusion matrices.

Robustness can be improved through architectural choices that promote stable representation learning. Deep residual networks introduced skip connections that mitigate optimization difficulties in very deep models and support better generalization in high-dimensional visual tasks [16]. Although botanical chemical profiles are not natural images, the principle of learning residual transformations is broadly useful when stacked nonlinear layers must capture subtle differences among closely related species. Foundational deep learning treatments further clarify the trade-offs among capacity, regularization, and generalization [17]. However, capacity alone does not guarantee robustness; it must be paired with data augmentation, spectral alignment, and domain-informed constraints.

Explainability in this context must be calibrated to the needs of multiple audiences. A taxonomist may require chemical rationales, a regulator may require auditability, and a field technician may require simple confidence indicators. Post hoc explanations can be useful, but they may not capture causal relationships. The distinction between statistical association and causal explanation is critical in biological identification [18]. Causally informed models would require experimental or observational data on biosynthetic pathways and environmental influences, which are rarely available at scale. Nevertheless, explanation interfaces should report uncertainty, highlight influential spectral regions, and provide access to reference specimens so that human experts can exercise final judgment.

6. Deployment, Sustainability, and Policy Implications

Deployment of AI-assisted identification systems for *Salvia* species involves practical challenges related to hardware, connectivity, regulatory acceptance, and institutional capacity. Algorithmic accountability scholarship has demonstrated that opaque predictive models can embed structural inequalities and obscure the political choices behind their design [19]. In botanical identification, oversimplified deployment models may impose standardized classification thresholds that do not account for local taxonomic uncertainty or traditional knowledge systems. Policy frameworks should therefore require impact assessments before systems are used in enforcement, trade regulation, or conservation prioritization.

Sustainability has at least two dimensions: environmental and institutional. The computational cost of training large models can be substantial, and the marginal benefit of increasingly complex architectures must be weighed against the carbon footprint and hardware lifecycle impacts. Critical appraisals of deep learning have cautioned against assuming that larger models always yield proportionally better real-world performance [20]. A sustainable deployment strategy may favor smaller, interpretable models for routine identification and reserve more complex models for ambiguous cases. Institutional sustainability also requires training local staff, maintaining reference collections, and developing open but governed model registries.

Policy implications extend to model registration, adverse event reporting, and international data sharing. The regulatory perception of artificial intelligence in scientific decision support has evolved as black-box systems have become more common [21]. For plant identification, transparent model cards, independent audits, and clear appeals processes can increase trust among botanists, customs agencies, and pharmaceutical manufacturers. International coordination is necessary because *Salvia* species are distributed across multiple biogeographic regions and are subject to different legal regimes governing access to genetic resources and

associated data. A well-designed policy architecture should align technical standards with biodiversity governance principles rather than treating them as separate domains.

7. Cross-Domain Lessons and Future Directions

The chemotaxonomic identification problem shares structural features with other high-dimensional biological classification tasks. Comprehensive reviews of deep learning architectures have identified a recurring pattern: high performance on curated benchmarks does not automatically translate to robust field deployment, especially when the input domain is heterogeneous and data collection is expensive [22]. The experience of medical machine learning, where prospective validation and regulatory oversight are now standard expectations, offers a template for botanical applications. Cross-domain comparisons also reveal that the most successful systems combine automated feature extraction with structured domain knowledge and clear human review.

Future systems may move toward integrated multi-omic identification that combines cell wall profiles, phytochemical fingerprints, genomic markers, and environmental metadata. Such integration will require more sophisticated alignment mechanisms to resolve differences in scale, missingness, and biological meaning. Active learning can reduce annotation costs by selecting specimens whose identification would most improve the model, but active learning strategies must be evaluated for fairness and representativeness. Foundation models pretrained on large chemical or spectral datasets could provide transferable representations, though their training data and licensing arrangements must be transparent.

Governance maturity will become increasingly important as AI-assisted identification systems move from research prototypes to routine environmental regulation. The development of common data models, standardized evaluation protocols, and independent audit frameworks can support interoperability while preserving local institutional autonomy. Future research should also investigate long-term model drift, the maintenance of voucher-linked digital reference libraries, and mechanisms for accountable human override. These efforts will require sustained collaboration among taxonomists, analytical chemists, computer scientists, ethicists, and policy scholars.

8. Conclusion

AI-assisted chemotaxonomic identification of *Salvia* species represents a compelling case study in the design of trustworthy scientific machine learning systems. The combination of cell wall component profiles and phytochemical fingerprints can provide richly diagnostic information, but only when embedded in an infrastructure that respects data provenance, model interpretability, and institutional governance. The technical challenge is not simply to train a classifier with high accuracy; it is to create a socio-technical system in which algorithmic outputs are interpretable, auditable, and accountable to botanical expertise and regulatory standards. A classifier that performs well on a curated laboratory dataset may fail when deployed in a herbarium with different specimen preparation conventions, or when applied to a wild population from a region underrepresented in the training corpus. The architecture must therefore support continuous monitoring of data drift, versioned model updates, and structured channels for expert override. In this respect, the identification problem becomes a governance problem as much as an analytical problem. The choices made about data storage, access control, model transparency, and auditability determine whether the system can earn the trust of taxonomists, regulators, and local communities. The most sophisticated deep

learning model will not be sustainable unless it is embedded in an institutional environment that treats identification as a collaborative human-machine process rather than a purely automated decision.

In summary, this paper has argued that AI-assisted chemotaxonomic identification of *Salvia* species should be designed as an integrated system of data provenance, multimodal feature engineering, interpretable modeling, and accountable deployment. The structural trade-offs between centralized and distributed architectures, between opaque high-capacity models and transparent simpler models, and between automated feature selection and domain-guided marker curation are not purely technical. They reflect deeper commitments about scientific rigor, fairness, and the distribution of decision-making authority. Robustness requires attention to distribution shift, label noise, and the ecological and geographic diversity of training specimens. Fairness requires disaggregated evaluation across taxa and populations, while explainability must be adapted to the distinct needs of taxonomists, enforcement agencies, and field practitioners. The policy environment must recognize that botanical identification systems can influence trade, conservation, and public health, and therefore demand transparency, independent audit, and clear accountability mechanisms. By learning from cross-domain experiences in medical machine learning and environmental monitoring, researchers and institutions can develop AI-assisted identification systems that are scientifically useful, operationally sustainable, and normatively defensible.

References

1. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
2. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
3. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
5. 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.
6. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 1–21.
7. Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1521–1528.
8. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
9. 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. *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.
10. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 28, 2503–2511.

11. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347–1358.
12. Vollmer, S., Mateen, B. A., Bohner, G., Király, F. J., Ghani, R., Jonsson, P., Cumbers, S., Jonas, A., McAllister, K. S. L., Myles, P., Granger, D., Birse, M., Branson, R., Moons, K. G. M., Collins, G. S., Ioannidis, J. P. A., Holmes, C., & Hemingway, H. (2020). Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ*, 368, 16927.
13. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92.
14. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104, 671–732.
15. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.
16. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
17. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
18. Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.
19. Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.
20. Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.
21. Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36–43.
22. Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8, 53.