

Spectroscopic Machine Learning for Rapid Discrimination and Quality Assessment of Salvia Medicinal Resources

Petri A. Bailey

Department of Computer Science, Colorado State University, Fort Collins, CO, USA.

petri.work@colostate.edu

Abstract

The rapid authentication and quality evaluation of Salvia medicinal resources remain persistent challenges because morphological similarity, variable cultivation conditions, post-harvest processing, and complex chemical composition can obscure botanical identity and therapeutic consistency. Spectroscopic platforms generate information-dense signals that are suitable for machine learning, but the translation of raw spectral responses into reliable regulatory-grade decisions requires careful attention to data architecture, algorithmic selection, robustness, and governance. This paper presents a systems-oriented analysis of spectroscopic machine learning for *Salvia miltiorrhiza* and related taxa. It examines spectral preprocessing, representation learning, data partitioning, model comparison, deployment infrastructure, and fairness across geographical and agronomic contexts. The discussion emphasizes that no single algorithm provides universal superiority; rather, performance emerges from the alignment between spectroscopic measurement physics, data curation strategy, and institutional data governance. The paper further considers how cell wall component markers can complement spectral models for resolving closely related *Salvia* species. It concludes by identifying policy and standardization pathways for integrating machine learning into the quality assessment of medicinal plant resources, with particular attention to reproducibility, auditability, and sustainable supply chains.

Keywords

Salvia miltiorrhiza; spectroscopic machine learning; quality assessment; chemometrics; botanical authentication; data governance; sustainability.

1. Introduction

Machine learning has become a foundational capability for interpreting complex analytical data. Support vector machines [1], random forests [2], and deep neural networks [3] have enabled substantive advances in spectral classification because they learn decision boundaries and hierarchical representations from high-dimensional data without requiring full physical models of every chemical interaction. In medicinal plant research, this shift is particularly significant for *Salvia* resources, where the differentiation of closely related species and the assessment of quality cannot rely solely on visual inspection or sparse chromatographic markers. The genus *Salvia* contains numerous species with overlapping morphological traits, and the medicinal preparation *Salvia miltiorrhiza* is subject to substitution and adulteration pressures across global markets. Spectroscopic methods offer rapid and non-destructive measurements, but their effective use depends on an integrated view of signal processing, model selection, and institutional accountability.

The chemometric tradition has long recognized that spectral data demand multivariate reasoning. Partial least squares [4] and multivariate calibration [5] established a rigorous framework for relating spectral matrices to chemical properties when predictor variables are numerous and highly correlated. These methods treat the spectrum as a coherent multivariate object rather than a collection of isolated peaks, which is essential for complex botanical matrices. In *Salvia* quality assessment, the relationship between spectral bands and active constituents is rarely expressed through a small number of isolated wavelengths. Instead, meaningful variance is dispersed across overlapping vibrational and electronic responses. The systems-level challenge therefore lies not only in fitting a model but also in constructing an analytical pipeline that remains stable when samples originate from different cultivation regions, harvest years, drying conditions, and instrumentation environments.

Before any learning algorithm can be usefully applied, the spectral signal must be stabilized and represented. Savitzky-Golay smoothing and differentiation [6], standard normal variate transformation [7], and systematic sample partitioning through experimental design procedures such as Kennard-Stone selection [8] provide important safeguards against baseline drift, multiplicative scatter, and uneven calibration coverage. These techniques are not merely technical details; they shape the distribution of training data and therefore influence the generalization behavior of downstream models. In *Salvia* discrimination, failure to correct for particle size variation, moisture effects, or detector response drift can produce models that perform well on internal validation but fail under field conditions. The overall system must be designed with this preprocessing burden in mind.

2. Background and Related System Structures

Support vector machines have been widely examined in analytical chemistry as a supervised classification approach for spectral data [9]. Their structural value lies in the use of kernel representations that can accommodate nonlinear boundaries while retaining a principled optimization objective. In medicinal plant authentication, this provides a useful middle ground between simple linear scoring and heavily parameterized deep learning. However, support vector machines also require careful tuning of the penalty and kernel parameters, and their interpretability can be limited when the feature space is complex. At a systems level, they are valuable when the number of training samples is modest and when spectral classes are separable under an appropriate transformation. For *Salvia* resources collected from multiple agroecological zones, the choice of kernel and the standardization of spectral inputs become central architectural decisions.

Partial least squares discriminant analysis and related projection methods have also been applied extensively in metabolomics and botanical authentication [10]. These approaches emphasize latent variable extraction and are especially relevant when the spectral matrix contains many collinear variables and when chemical reference measurements are available. The projection to a lower-dimensional subspace can make the spectral variance more interpretable and can expose sources of batch variation. For *Salvia*, such methods help connect spectral variation to known classes such as species identity, geographic origin, or processing grade. Yet projection-based models can be sensitive to class imbalance and to the presence of outliers. A systems perspective therefore treats them as components of a broader quality architecture rather than as standalone solutions.

Salvia miltiorrhiza and its preparations have been characterized through chromatographic and mass spectrometric fingerprints that establish chemical differences among related materials [11]. Near-infrared spectroscopy has gained acceptance in natural products analysis because it

enables rapid screening of bulk plant material [12]. Mid-infrared spectroscopy has similarly been used to assess traditional Chinese medicines and can reveal diagnostically useful functional group information [13]. These measurement modalities differ in penetration depth, scattering behavior, sensitivity to moisture, and instrumental cost. In a spectroscopic machine learning system for *Salvia*, the selection among near-infrared, mid-infrared, or combined sensor regimes is not a purely technical choice. It affects data sharing, model portability, calibration transfer, and the level of regulatory scrutiny required when a model is deployed across heterogeneous supply chains.

3. Data Architecture and Representation Learning

Data architecture for medicinal plant spectroscopy must reconcile biological variability with the need for reproducible model training. Random forest methods have been shown to be effective for compound classification and quantitative structure-activity modeling because they can handle mixed data types and provide variable importance estimates [14]. In *Salvia* quality assessment, such ensemble approaches can be used to identify spectral regions associated with tanshinones, salvianolic acids, moisture content, and other quality attributes. However, variable importance from correlated spectral data can be unstable, and the resulting models may rely on regions that are not chemically interpretable. Data architecture should therefore include multiple complementary views of the same samples, including raw spectra, preprocessed spectra, and known chemical or botanical metadata.

Metabolomic profiling has demonstrated that global biochemical fingerprints can be acquired and interpreted through statistical workflows [15]. In the specific case of *Salvia miltiorrhiza* and its close relatives, markers derived from cell wall component characteristics can provide additional taxonomic resolution [16]. These structural markers are not always visible in conventional small-molecule chromatographic profiles, which makes them valuable for distinguishing closely related species that share common secondary metabolites. When integrated with spectroscopic data, such markers can serve as auxiliary labels or weak supervision signals that guide model learning. The architectural implication is that a well-designed *Salvia* assessment system should not treat spectral data in isolation. Instead, it should maintain a curated reference database in which spectroscopic measurements are linked to authenticated botanical vouchers, chemical profiles, and cell wall marker information where available.

Supervised pattern recognition has a long history in food and botanical analysis, and its methodological expectations have been codified through comparative studies [17]. Preprocessing remains a decisive stage in near-infrared spectral analysis, with multiple transformations available for baseline correction, scatter removal, and derivative calculation [18]. In *Salvia* applications, the preprocessing sequence must be selected with attention to the physical origin of unwanted variance. For example, diffuse reflectance from powdered root material is strongly influenced by particle size distribution and packing density. A preprocessing protocol that is appropriate for one instrument may not transfer directly to another instrument. The system architecture should therefore store preprocessing parameters alongside model artifacts, making it possible to audit how a prediction was produced and to reproduce the exact data lineage from raw scan to final label.

4. Model Selection and Evaluation Trade-Offs

The chemometric literature emphasizes that partial least squares regression remains a basic tool because of its balance between predictive power and interpretability [19]. In *Salvia*

quality assessment, interpretability is not a secondary concern. Regulatory authorities, quality control laboratories, and supply chain auditors often ask why a model decided that a particular batch belongs to a given species or quality grade. A model that simply outputs a class probability without explanation may be difficult to integrate into current pharmacopoeial frameworks. Projection-based models and tree ensembles can provide interpretable diagnostics such as variance explained, variable importance, or nearest class distances. Deep neural networks can achieve higher predictive accuracy in some settings, but their opacity may require additional explanation layers before deployment in regulated environments.

Deep residual networks introduced architectural innovations that allow very deep neural networks to be trained more reliably by mitigating vanishing gradient phenomena [20]. In spectral classification, residual structures can be adapted to one-dimensional signals or two-dimensional time-frequency representations. For *Salvia*, a residual network might learn hierarchical spectral features, from narrow absorption patterns to broader chemical fingerprints. Nevertheless, the value of such architectures depends on the availability of sufficiently large and diverse spectral libraries. In many medicinal plant applications, high-quality authenticated samples are expensive to obtain, and class labels may be uncertain. This scarcity shifts the design trade-off toward models that are less data-hungry and toward training strategies that transfer knowledge from related spectral tasks.

Attention mechanisms have reshaped sequence modeling by allowing models to attend to informative positions without recurrent processing [21]. In spectroscopy, attention can help identify spectral intervals that matter for discrimination and can align samples measured under different instrument configurations. Data augmentation has also emerged as a practical strategy for improving generalization in image classification and can be adapted to spectral data by simulating noise, baseline shifts, and small wavelength shifts [22]. Architectures based on depthwise separable convolutions, such as Xception, illustrate how parameter efficiency can be improved while preserving representational capacity [25]. These advances do not automatically resolve *Salvia* authentication problems, but they provide a rich design space for building compact, portable models that can operate on edge devices in field settings.

5. Deployment Infrastructure and Governance

Deployment of spectroscopic machine learning for medicinal resources must operate within existing quality assurance frameworks. The World Health Organization has published guidelines for assessing herbal medicines with attention to contaminants, residues, and botanical identity [23]. A machine learning model that flags a *Salvia* batch as adulterated is only useful when its output can be connected to appropriate sampling procedures, reference methods, and enforcement actions. This requires governance mechanisms that define acceptable false positive and false negative rates, specify model update cycles, and mandate periodic validation against independent reference data. Without such mechanisms, predictive models may drift silently as new production regions, climate conditions, or adulteration practices emerge.

Regional regulatory guidance also shapes the practical use of spectroscopic models in herbal medicinal products [24]. Pharmaceutical quality systems expect analytical methods to be validated with respect to specificity, linearity, accuracy, precision, and robustness. Machine learning models do not fit easily into these categories because their performance is data-dependent and may change with retraining. A governance framework for *Salvia* assessment should therefore distinguish between model qualification, software validation, and operational monitoring. Model qualification establishes that the algorithm can discriminate species and

quality grades under specified conditions. Software validation ensures that the implementation is free from unintended changes. Operational monitoring tracks prediction stability and drift over time. These layers together can create an audit trail that is acceptable to regulators and supply chain participants.

In addition to regulatory alignment, deployment infrastructure must address practical constraints in field laboratories, customs checkpoints, and production facilities. Portable spectrometers may have lower signal-to-noise ratios than laboratory instruments, and network connectivity may be intermittent. Edge computing architectures can reduce dependence on centralized cloud services, but they also introduce challenges in software updating, data synchronization, and model version control. A well-governed deployment strategy should therefore include standardized model containers, clear version identifiers, and secure logging mechanisms. This allows a spectral prediction to be traced from the original instrument scan to the model output and to any subsequent regulatory action.

6. Robustness, Fairness, and Sustainability

Robustness in Salvia spectroscopic models depends on how well the training data represent the range of biological and commercial variation. A model trained only on samples from a single province may attribute regional signals to species identity, producing misleading results when applied to samples from other regions. Fairness in this context refers to equitable performance across different cultivation origins, processing practices, and producer scales. Smallholder farmers and large agricultural enterprises may use different drying equipment and storage conditions, which can affect spectral baselines. A system that is accurate for industrialized supply chains but unreliable for smallholder producers could exacerbate market exclusion. To avoid this, data collection should be geographically distributed, and equity analyses should be built into model reporting.

Sustainability also emerges as a system-level consideration. Spectroscopic screening can reduce destructive testing, minimize solvent use, and enable faster decisions about harvest timing and storage. However, the computational infrastructure behind machine learning also consumes energy, and frequent retraining of large models may not be justified when simpler approaches achieve comparable results. The ecological footprint of cloud-based inference and the hardware lifecycle of portable spectrometers should be considered when designing deployment strategies. For Salvia medicinal resources, which are connected to broader biodiversity and conservation concerns, a sustainable machine learning system should support traceability, discourage overharvesting of wild populations, and reward quality differentiation in ways that align with long-term ecological health.

A further robustness concern is seasonal and climatic variability. Salvia plants accumulate secondary metabolites in response to temperature, water availability, and soil conditions. If these environmental effects are correlated with species labels in the training data, the model may learn ecological signals rather than identity signals. Cross-season validation and multi-year sampling are therefore important safeguards. Similarly, spectral models should be tested under different sample preparation protocols, including variations in grinding time, sieving, and moisture equilibration. Only through such deliberate stress testing can the system reveal whether its decision boundaries are truly linked to botanical identity and quality attributes rather than to incidental acquisition conditions.

7. Policy Implications and Future Directions

The integration of spectroscopic machine learning into Salvia quality assessment raises policy questions about data ownership, reference standards, and cross-border harmonization. Spectral libraries developed by one institution may not be accessible to others because of proprietary or privacy concerns. Yet the robustness of any model depends on shared reference data that represent authentic botanical materials. Policymakers and scientific communities should consider creating open reference repositories with clear data quality levels, including authenticated specimens, instrumental metadata, and standardized preprocessing histories. Such repositories would reduce duplication, improve reproducibility, and support the development of models that are auditable by multiple stakeholders.

Future research should examine the fusion of spectroscopic signals with genomic, metabolomic, and cell wall marker data to build more resilient identity systems. Attention should also be directed to calibration transfer methods that allow a model trained on one instrument to be used on another without extensive recalibration. In parallel, the regulatory community should develop guidance for the lifecycle management of machine learning methods in herbal quality control. This guidance might include requirements for training data documentation, performance reports across demographic and geographic subgroups, and automated monitoring of prediction drift. If these elements are aligned, spectroscopic machine learning can become a legitimate and sustainable component of the global governance of medicinal plant resources.

Another important direction is the development of formal interlaboratory studies that compare spectroscopic machine learning systems across instruments, algorithms, and botanical reference collections. Such studies would help establish realistic performance expectations and identify the most transferable preprocessing and modeling choices. They would also clarify whether certain Salvia species can be reliably distinguished by spectral means alone or whether complementary molecular and structural markers are necessary. The field would benefit from consensus reporting guidelines that specify sample sizes, validation protocols, and metadata requirements. These efforts can build confidence among regulators, producers, and consumers while preserving scientific rigor.

8. Conclusion

Spectroscopic machine learning offers a promising pathway for rapid discrimination and quality assessment of Salvia medicinal resources, but its transformative potential depends on robust integration with authenticated reference data, transparent preprocessing and reporting standards, and governance structures that ensure equitable access and accountability across the medicinal plant supply chain. The preceding sections have shown that the value of spectral models arises not from algorithmic novelty alone but from the careful construction of data pipelines, the alignment of measurement physics with representation choices, and the embedding of predictive systems within broader regulatory and sustainability frameworks.

A central conclusion from this analysis is that Salvia authentication should be understood as a socio-technical infrastructure rather than a discrete modeling task. The infrastructure includes botanical voucher repositories, spectral instrument networks, preprocessing libraries, machine learning registries, and audit mechanisms that connect model outputs to laboratory decisions. When any component is weak, the overall reliability of the system declines. For instance, a high-performing classifier cannot compensate for an unrepresentative reference collection, and a carefully curated spectral library cannot be exploited if instrument calibration is not

maintained. Future investments should therefore prioritize end-to-end traceability, allowing each prediction to be reconstructed and reviewed.

From a regulatory perspective, the introduction of machine learning into Salvia quality assessment should proceed through iterative validation rather than abrupt replacement of established pharmacopoeial methods. Spectroscopic models can serve as screening tools that reduce the burden on confirmatory chromatographic or microscopic methods. Over time, accumulated validation evidence may support expanded uses, including quantitative prediction of key quality markers and detection of previously unseen adulterants. However, such expansion requires ongoing interlaboratory comparisons, open reporting of model errors, and a commitment to updating reference data as new cultivation regions and processing practices emerge.

Finally, the long-term sustainability of spectroscopic machine learning for medicinal resources will depend on interdisciplinary collaboration among analytical chemists, data scientists, regulatory scientists, and botanical taxonomists. This collaboration should extend to the development of shared data standards, the publication of negative results and failure modes, and the creation of educational pathways that enable quality control laboratories in resource-limited settings to participate fully. If these conditions are met, spectroscopic machine learning can support a more transparent, efficient, and ecologically responsible Salvia supply chain, while contributing broader lessons for the digitalization of natural product authentication.

References

1. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.
2. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
3. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.
4. Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2), 109-130.
5. Martens, H., & Naes, T. (1989). *Multivariate calibration*. Wiley.
6. Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8), 1627-1639.
7. Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, 43(5), 772-777.
8. Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics*, 11(1), 137-148.
9. Brereton, R. G., & Lloyd, G. R. (2010). Support vector machines for classification and regression. *Analyst*, 135(2), 230-267.
10. Gromski, P. S., Muhamadali, H., Ellis, D. I., Xu, Y., Correa, E., Turner, M. L., & Goodacre, R. (2015). A tutorial review: Metabolomics and partial least squares-discriminant analysis. *Analytica Chimica Acta*, 879, 10-23.

11. Li, Y. G., Song, L., Liu, M., Hu, Z. B., & Wang, Z. T. (2009). Advancement in analysis of *Salviae miltiorrhizae Radix et Rhizoma* (Danshen). *Journal of Chromatography A*, 1216(11), 1941-1953.
12. Cozzolino, D. (2014). An overview of the use of near infrared spectroscopy and chemometrics in food analysis. *Food Research International*, 61, 105-112.
13. Sun, S., Chen, J., Zhou, Q., Lu, W., & Zhang, H. (2010). Application of mid-infrared spectroscopy in the quality control of traditional Chinese medicine and its related products. *Vibrational Spectroscopy*, 52(2), 145-150.
14. Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, 43(6), 1947-1958.
15. Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W.-M., Fiehn, O., Goodacre, R., Griffin, J. L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A. N., Lindon, J. C., Marriott, P., Nicholls, A. W., Reily, M. D., Thaden, J. J., & Viant, M. R. (2007). Proposed minimum reporting standards for chemical analysis. *Metabolomics*, 3(3), 211-221.
16. 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.
17. Berrueta, L. A., Alonso-Salces, R. M., & Héberger, K. (2007). Supervised pattern recognition in food analysis. *Journal of Chromatography A*, 1158(1-2), 196-214.
18. Rinnan, Å., van den Berg, F., & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, 28(10), 1201-1222.
19. Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, 185, 1-17.
20. 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.
21. 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.
22. Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 60.
23. World Health Organization. (2007). WHO guidelines for assessing quality of herbal medicines with reference to contaminants and residues. World Health Organization.
24. European Medicines Agency. (2011). Guideline on quality of herbal medicinal products/traditional herbal medicinal products. EMA/HMPC/2011/6/2005 Rev. 2. Committee on Herbal Medicinal Products.

25. Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1251-1258.