

Explainable AI for Tracing Botanical Origin and Quality Markers of Traditional Herbal Medicines

Ishaan Mendoza

Department of Computer Science, University of North Texas, Denton, TX, USA.

ishaan.work@unt.edu

Pierre Burns

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

contactpierre@buffalo.edu

Guoran He

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

Columbia, MO, USA.

guoranmail@missouri.edu

Abstract

Traditional herbal medicines present distinctive scientific and regulatory challenges because their identity, therapeutic value, and safety depend on botanical origin, cultivation conditions, processing history, and the integrity of complex chemical profiles. Machine learning increasingly supports the authentication of herbal materials and the detection of quality markers, yet many high-performing models remain opaque to regulators, practitioners, and supply chain stakeholders. This paper develops a system-level perspective on explainable artificial intelligence for tracing botanical origin and quality markers in traditional herbal medicines. It examines how explanation methods can be integrated into the broader analytical pipeline, spanning spectral data acquisition, metabolomic profiling, feature engineering, classification, and decision support. Rather than treating explainability as a localized post hoc technique, the paper argues that robust explanation requires an architectural approach that coordinates data provenance, model modularity, uncertainty reporting, and interpretable feature representation. The discussion focuses on structural trade-offs between predictive performance and interpretability, the governance of distributed data infrastructure, the deployment of explainable systems across heterogeneous laboratories and regulatory environments, and the implications of fairness and robustness for public health. The paper also compares developments in medical artificial intelligence, chemometrics, and natural product authentication to identify transferable design principles. It emphasizes that sustainable deployment depends not only on algorithmic transparency but also on institutional mechanisms for validation, audit, and continuous monitoring. The analysis concludes with policy considerations and forward-looking perspectives on building accountable, interpretable, and scientifically credible artificial intelligence systems for herbal medicine quality assurance.

Keywords

explainable artificial intelligence; botanical origin tracing; quality markers; herbal medicines; systems architecture; metabolomics; governance; interpretability; robust machine learning.

1. Introduction

The authentication of traditional herbal medicines has become a significant interdisciplinary problem because commercial supply chains increasingly involve multiple cultivation regions, processing facilities, and distribution intermediaries. Botanical materials that appear morphologically similar may differ substantially in their chemical composition, pharmacological activity, and safety profile. Regulatory agencies and manufacturers therefore require reliable methods for tracing geographic origin and verifying the presence of quality markers. In this context, machine learning offers attractive capabilities for modeling complex relationships between analytical measurements and botanical identity. However, the adoption of machine learning in quality assurance is constrained by concerns about transparency, accountability, and scientific interpretability. Explainable artificial intelligence has emerged as a response to these concerns, with methods that seek to clarify how a model arrives at its predictions [1,2]. These methods range from local approximations that explain individual decisions to visualizations that highlight influential input regions. While such techniques can increase user confidence, their integration into herbal medicine authentication demands more than algorithmic explanations. It requires a coherent system architecture that aligns data acquisition, feature representation, model selection, explanation generation, and regulatory reporting.

A central tension in this field is the relationship between predictive performance and interpretability. High-capacity models such as deep neural networks can capture subtle patterns in complex spectral or chromatographic data, but their internal representations are often difficult to map onto chemically meaningful markers. Simpler models may be more interpretable, but they can sacrifice accuracy when the underlying relationships are nonlinear and highly interactive. This trade-off has been widely discussed in the broader explainable artificial intelligence literature, where scholars have argued that high-stakes applications should favor inherently interpretable models when possible [3,4]. In herbal medicine, the stakes include public health, consumer protection, and the economic viability of legitimate producers. The selection of an appropriate modeling strategy therefore cannot be reduced to a purely technical optimization. It must reflect institutional priorities, regulatory expectations, and the available analytical infrastructure. A robust explanation of botanical origin should not merely identify that a sample was classified as a particular species or region; it should also reveal which chemical features, spectral bands, or preprocessing decisions supported that conclusion.

The system-level view adopted in this paper treats explainability as an emergent property of the entire socio-technical infrastructure, rather than as a property of a single model. This perspective recognizes that explanations are consumed by diverse stakeholders, including laboratory scientists, quality control managers, regulators, clinicians, and supply chain auditors. Each group may require different forms of evidence, different levels of granularity, and different degrees of mathematical detail. Consequently, a single explanation interface is unlikely to satisfy all demands. Instead, modular architectures are needed in which explanation services can be composed according to the user context and the analytical task. This paper develops that argument by examining the structural trade-offs, data infrastructure, interpretability methods, robustness requirements, governance frameworks, and deployment pathways relevant to explainable artificial intelligence for traditional herbal medicines.

2. Related Work and Analytical Context

Research on explainable artificial intelligence has produced a diverse family of techniques for making complex models more transparent. Gradient-based localization methods, for example,

have been used extensively in computer vision to highlight image regions that most strongly influence a classification decision [5]. These methods can be adapted to spectral or chromatographic data by treating two-dimensional representations of analytical signals as images, although such adaptation requires careful attention to the physical meaning of the input dimensions. In parallel, studies on botanical authentication have explored the use of cell wall component features to distinguish species that are morphologically similar but chemically distinct. For example, a recent investigation into *Salvia miltiorrhiza* and its close relatives demonstrated that cell wall component characteristics can serve as identification markers [6]. That work illustrates how chemical and structural features may complement conventional metabolomic profiles in origin tracing and taxonomic discrimination.

Metabolomic approaches have become central to the authentication of natural products because they capture broad chemical variation across samples. Global metabolite profiling can reveal differences associated with cultivation region, harvest season, processing method, and storage conditions [7]. However, the interpretation of metabolomic data is complicated by the large number of measured features, the presence of noise, and the influence of uncontrolled environmental factors. Analytical chemists and data scientists have therefore emphasized the importance of standardization and transparent reporting in metabolomic studies. Minimum reporting standards help ensure that data can be compared across laboratories and that machine learning models trained on one instrumental platform can be evaluated on another [8]. The complexity of natural extracts also means that no single analytical technique is sufficient to capture all relevant chemical dimensions. Chromatographic separation coupled with mass spectrometric detection can offer high sensitivity, but sample preparation and ionization conditions may introduce systematic biases [9]. These analytical realities shape the design space for explainable artificial intelligence systems, because explanations derived from biased or incomplete data cannot be considered reliable evidence of botanical origin.

The convergence of these fields suggests that explainability in herbal medicine authentication should be understood as a multi-layered problem. At the lowest layer, explanations must reflect the chemical and physical basis of the measurement. At an intermediate layer, they must clarify how the model transformed those measurements into feature representations and predictions. At the highest layer, they must communicate the practical significance of the result to decision makers. Existing work in medical artificial intelligence has highlighted the limitations of purely technical explanations when they are disconnected from clinical workflows and regulatory accountability structures [3]. A similar caution applies to herbal medicine, where the decision to accept or reject a batch may have immediate commercial and safety consequences. Therefore, related work from both interpretable machine learning and analytical chemistry must be integrated into a broader systems framework.

3. System Architecture for Explainable Botanical Origin Tracing

Developing an explainable artificial intelligence system for botanical origin tracing requires an architectural strategy that separates concerns while maintaining traceability across the analytical pipeline. Taxonomies of explainable artificial intelligence often distinguish between local and global explanations, model-specific and model-agnostic approaches, and ante hoc versus post hoc interpretability [10]. These distinctions are useful but insufficient for designing a deployable system. A more productive starting point is to structure the system into layers responsible for data ingestion, quality assessment, feature computation, classification, explanation generation, and decision reporting. Each layer should expose well-

defined interfaces so that changes in an analytical instrument or a modeling algorithm do not propagate unpredictably through the entire system. This modularity also enables independent validation of critical components, which is essential when the system must be audited by external parties.

One major architectural trade-off concerns the placement of interpretability. Some system designs embed interpretability directly into the model by constraining its function form or by using self-explaining mechanisms. Others apply post hoc explanation methods to a trained black-box model. The former can produce explanations that are more faithful to the model's computation, but it may limit predictive flexibility. The latter can preserve flexibility, but the explanation may only approximate the model's behavior for selected instances. Interpretability scholars have noted that the term interpretability itself is ambiguous, and that different definitions lead to different evaluation criteria [11,12]. A rigorous approach therefore requires the system to specify what kind of explanation is being provided, for whom, and under what conditions it is valid [13]. For botanical origin tracing, local explanations that identify influential spectral regions can be valuable for laboratory scientists, while global explanations that summarize marker distributions across regions may be more useful for regulatory decision making. Both can be supported by visualizing the marginal effect of feature variables on model predictions [14].

A well-designed architecture also separates the classification objective from the explanation objective. The classifier may prioritize robustness to noise and generalization across cultivation regions, while the explanation layer may prioritize sparsity, chemical plausibility, and alignment with known markers. These objectives can conflict. A model may achieve high accuracy by combining many weak spectral signals that are difficult to interpret chemically, whereas a simpler model may rely on a small number of well-characterized markers but miss subtle adulteration patterns. The architectural response is to treat these as distinct components with explicit interfaces and to provide multiple explanation views. For example, a system might include a high-performance ensemble classifier for final decision support and a separate interpretable model for generating candidate biomarker explanations. The outputs of both components can be compared to assess consistency. When the explanations diverge, the system should flag the result for expert review rather than silently presenting a potentially misleading explanation.

The architecture must also account for temporal dynamics. Botanical materials are biological products whose chemical profiles can shift with climate, soil conditions, and post-harvest treatment. A static model trained on historical samples may become unreliable as new cultivation practices emerge or as adulteration strategies change. Therefore, the system should include monitoring mechanisms that track data drift, model performance, and explanation stability. Monitoring is not a peripheral concern; it is a core architectural requirement for maintaining the validity of both predictions and explanations over time. In addition, the architecture should support provenance tracking so that every prediction can be linked to raw data files, preprocessing scripts, model versions, and calibration records. Such traceability is particularly important when analytical results are challenged in regulatory proceedings or commercial disputes.

4. Data Infrastructure and Quality Marker Representation

The quality of an explainable artificial intelligence system depends fundamentally on the quality and representativeness of the underlying data infrastructure. Traditional herbal medicine authentication draws on diverse data sources, including botanical voucher

specimens, genomic information, chromatographic profiles, mass spectra, and nuclear magnetic resonance measurements. Pharmacognosy has long emphasized the importance of authentic reference materials and documented collection practices [15]. Without reliable reference data, machine learning models can learn spurious correlations that do not reflect true botanical or geographic distinctions. Data infrastructure must therefore include rigorous sample accessioning, metadata standards, and quality control procedures. The application of chemometric methods to spectral data, for instance, requires careful preprocessing to correct baseline drift, alignment shifts, and scaling differences [16]. These preprocessing decisions can alter the apparent importance of specific features and thereby influence downstream explanations.

A critical challenge is the representation of quality markers. Analytical platforms generate high-dimensional data in which many measured variables are redundant or correlated. Dimensionality reduction and visualization techniques such as t-distributed stochastic neighbor embedding can support exploratory analysis by revealing clusters associated with origin or processing method [17]. However, such embeddings are not always easy to interpret because the reduced dimensions do not correspond directly to chemical compounds. Supervised learning methods based on random forests or regularized linear models can provide more direct feature importance estimates [18]. These methods are widely available in open-source scientific computing environments, which facilitates their integration into laboratory workflows [19]. Yet feature importance alone does not establish that a particular compound is a robust quality marker. It must be confirmed through chemical identification, concentration measurements, and correlation with known biological activities.

The data infrastructure should also support harmonization across analytical platforms. A model trained on liquid chromatography-mass spectrometry data from one laboratory may not transfer directly to another laboratory using a different instrument or protocol. System-level design must therefore include strategies for calibration transfer, batch correction, and domain adaptation. These strategies are not merely technical conveniences; they are necessary for multi-site validation and regulatory acceptance. Moreover, the data infrastructure should be organized to capture negative examples, including adulterated samples, closely related species, and samples from regions outside the model's training distribution. The absence of such negative examples can produce optimistic performance estimates and explanations that overlook important failure modes. In botanical origin tracing, the ability to explain why a sample belongs to a particular region is incomplete without the ability to explain why it does not belong to plausible alternatives.

A forward-looking data infrastructure should also embrace community data sharing while respecting intellectual property and source confidentiality. Herbal material datasets often involve commercial sensitivities because they can reveal supply chain relationships or proprietary production methods. Distributed federated learning architectures can allow multiple institutions to train models without pooling raw data, but they introduce additional governance and technical trade-offs. For example, federated training may complicate explanation because the model's parameter updates are aggregated across sites, and the contribution of individual samples may be harder to reconstruct. These challenges highlight the need to align technical design with institutional policies and data-sharing agreements from the outset.

5. Explainability Methods and Interpretability Trade-offs

The selection of explainability methods in herbal medicine authentication requires a careful consideration of the relationship between mathematical fidelity, chemical plausibility, and user comprehension. Local model-agnostic explanations can approximate the behavior of a complex classifier around a specific sample by perturbing input features and observing changes in predictions. Such methods are flexible and can be applied to many model classes, but their explanations depend on the perturbation strategy and the definition of locality [20]. In high-dimensional spectral data, the perturbation of individual wavelengths may not correspond to physically meaningful changes because spectral intensities are often correlated. Consequently, explanations based on isolated feature perturbations may mislead users by attributing importance to a wavelength that is merely correlated with the true causal marker. A chemically informed explanation method would instead perturb groups of features that correspond to known peaks or chromatographic regions, thereby preserving the integrity of the analytical signal.

Global explanation methods offer a complementary perspective by summarizing model behavior across many samples. Global surrogate models can approximate a complex model with a simpler, more interpretable model, but the fidelity of the surrogate is often uneven across the input space. Partial dependence plots and individual conditional expectation curves can reveal how model predictions vary with specific feature values, but they may obscure interactions among multiple features [14]. For botanical origin tracing, interactions are important because the diagnostic value of a compound may depend on the presence of other compounds, the processing method, or the detection limit of the instrument. Explanations that ignore these interactions can overstate the confidence of a classification result. The system should therefore enable analysts to examine both local and global explanations and to compare them against domain knowledge.

Another trade-off concerns the use of inherently interpretable models versus post hoc explanations for complex models. In high-stakes applications, some researchers argue that opaque models should be avoided when interpretable alternatives achieve comparable performance [4]. This argument has practical force in herbal medicine, where the number of well-characterized markers is sometimes limited and where a transparent scoring rule may be easier to validate. However, the chemical diversity of herbal materials and the sophistication of adulteration practices may require models that capture complex nonlinear interactions. In such cases, self-explaining neural networks and other architectures that integrate explanations into the learning process may provide a middle path [21]. These models can be designed to produce feature attributions that are constrained by chemical priors or sparsity penalties. Yet their use also raises new questions about how to validate the explanations themselves, especially when the underlying model remains difficult to inspect independently.

It is also essential to recognize the limitations of current explanation methods in health-related domains. Recent critical assessments have warned that popular explanation techniques may create a false sense of transparency when their outputs are not rigorously evaluated against clinical or chemical ground truth [22]. In herbal medicine authentication, the ground truth may be botanical identity established by multiple independent methods, or the presence of a known marker compound confirmed by reference standards. Explanation evaluations should therefore include both quantitative measures of fidelity and qualitative assessments by domain experts. A system that provides a neatly highlighted spectral region is not sufficient if the highlighted region cannot be connected to a defensible chemical interpretation. Sustainable

deployment requires continuous evaluation of explanation quality under conditions that reflect real operational variability.

6. Robustness, Fairness, and Validation

Robustness is a central concern in the design of explainable artificial intelligence systems for botanical origin tracing. The system must maintain acceptable performance when analytical conditions change, when new sample types appear, or when adversarial actors attempt to evade detection. Explanations must also remain stable under realistic perturbations of the input data. If a small change in baseline correction or chromatographic alignment reverses the explanation of a sample, the system cannot be considered reliable for regulatory use. Robustness should therefore be assessed across multiple dimensions, including instrumental noise, sample preparation variability, batch effects, and distributional shifts. Validation protocols should reserve independent test sets that are collected under conditions deliberately different from the training data, so that performance claims do not depend on narrow laboratory consistency.

Fairness in this context does not refer primarily to individual demographic fairness, but rather to equitable treatment across botanical sources, geographic regions, and producer communities. A model that performs well for well-studied species but poorly for rare or regional varieties may disadvantage smaller producers and reduce the diversity of the herbal medicine supply. Similarly, a system trained mostly on samples from a particular region may produce miscalibrated explanations for samples from underrepresented regions. The governance of data infrastructure must therefore consider how sampling strategies, data contributions, and model updates affect the distribution of benefits and risks. Fairness also extends to the accessibility of explanations. If only highly resourced laboratories can use sophisticated explanation tools, smaller producers may face barriers to demonstrating the quality of their products. Simplified but scientifically sound explanation interfaces can help bridge this gap without compromising analytical rigor.

Validation of explainable models requires a combination of computational, chemical, and organizational evidence. Computational validation may include perturbation-based fidelity checks, sensitivity analysis, and comparisons between explanation outputs and known marker compounds. Chemical validation involves confirmatory analysis of the highlighted features using reference standards, orthogonal analytical techniques, or targeted quantitative methods. Organizational validation concerns the institutional processes through which model updates are approved, deviations are reviewed, and explanations are integrated into batch release decisions. These validation layers should be documented in a manner that supports external audit. In many jurisdictions, traditional herbal medicine quality control is governed by pharmacopoeial standards and good manufacturing practice requirements. Machine learning outputs that inform release decisions must therefore be embedded within a quality management system rather than treated as standalone computational results.

The robustness of explanations is also related to uncertainty quantification. A point prediction alone can obscure the possibility that a sample is near a decision boundary or that the model has not encountered similar samples during training. Explanations should report not only which features influenced the prediction but also how confident the system is in those attributions. Uncertainty can arise from measurement error, missing data, model ambiguity, and incomplete reference libraries. System design should propagate uncertainty estimates through the explanation layer so that end users can distinguish between a high-confidence marker attribution and a tentative suggestion. This is particularly important when the model is

used to identify adulteration, because an incorrect high-confidence attribution may have serious public health consequences.

7. Governance, Deployment, and Policy Implications

The deployment of explainable artificial intelligence in herbal medicine authentication is not solely a technical undertaking; it requires governance frameworks that connect computational outputs to regulatory decision making and industry practice. Governance mechanisms should define the responsibilities of data providers, model developers, laboratory users, and regulatory reviewers. They should also establish criteria for model acceptance, including minimum performance thresholds, explanation requirements, and documentation standards. A governance framework that is too rigid may discourage innovation and make it difficult to update models when new adulteration methods appear. A framework that is too flexible may allow models to enter use without adequate validation. The appropriate balance will vary across jurisdictions, but common principles can be drawn from existing quality assurance systems in pharmaceutical manufacturing and from emerging guidance on artificial intelligence in medical devices [3,22].

One policy challenge is the need to reconcile the proprietary nature of some herbal material data with the public interest in robust authentication. Regulatory authorities may require access to training data or model documentation to assess whether a system is fit for purpose. Producers and technology vendors may resist such access because it exposes trade secrets or commercial strategies. Federated learning and secure computation can partially address this tension, but they also complicate the audit process. Policymakers should therefore consider mechanisms such as trusted third-party auditors, standardized model cards, and regulatory sandboxes for explainable authentication systems. These mechanisms can create a space for iterative development while ensuring that public health protections remain intact.

Deployment also raises questions about the division of labor between automated systems and human experts. Explainable artificial intelligence should support expert judgment rather than replace it. In practice, this means that decision workflows should be designed so that model outputs are reviewed by qualified analysts, particularly when the model's prediction conflicts with other evidence or when the sample falls outside the validated domain. The role of explanation is to make this human review more efficient and more directed, not to eliminate it. Effective deployment therefore requires investment in training, user interface design, and standard operating procedures that specify when and how explanations should be consulted. Without such organizational infrastructure, explanation tools may be ignored or misused.

Sustainability is another important consideration. Long-term operation of an explainable authentication system requires maintenance of reference databases, recalibration of analytical instruments, periodic model retraining, and ongoing evaluation of explanation quality. These activities are resource intensive and may not be financially sustainable for small laboratories or regulatory agencies in all regions. International collaboration and open-source tool development can reduce costs, but they also require coordination and shared standards. The broader scientific community can contribute by publishing benchmark datasets, developing open evaluation protocols, and reporting negative results that reveal the limitations of current explanation methods. Such collective infrastructure can help prevent the fragmentation of effort and improve the credibility of machine learning approaches in herbal medicine.

Finally, policy should encourage the development of domain-specific explanation standards that go beyond generic transparency requirements. A generic statement that a model is

interpretable is of limited regulatory value. More useful is a specification of the explanation type, its scope, its validation evidence, and its limitations. For botanical origin tracing, such a specification might require explanations to identify candidate marker compounds, to quantify their contribution to the classification, to report uncertainty, and to indicate whether the sample falls within the model's applicability domain. Aligning these technical specifications with pharmacopoeial expectations and good manufacturing practice can facilitate regulatory acceptance. In the longer term, harmonized standards could support mutual recognition of authentication results across markets, thereby reducing trade barriers and strengthening the integrity of the global herbal medicine supply chain.

8. Conclusion

Explainable artificial intelligence holds considerable promise for tracing botanical origin and quality markers of traditional herbal medicines, but its successful application depends on more than the adoption of explanation algorithms. This paper has argued that a system-level perspective is necessary to address the structural trade-offs between predictive accuracy and interpretability, the demands of heterogeneous data infrastructure, the complexities of feature representation, and the need for robust validation. Explanation methods must be integrated into modular architectures that preserve traceability from raw analytical data to final decision support. They must be evaluated not only by computational fidelity but also by chemical plausibility and usefulness to regulatory and laboratory stakeholders. The governance and deployment of such systems require mechanisms for audit, uncertainty reporting, and continued monitoring under changing operational conditions. Fairness considerations demand attention to underrepresented botanical sources and producer communities, while policy frameworks should promote standardized explanation requirements and sustainable data-sharing arrangements. By treating explainability as a property of the broader socio-technical system, researchers and practitioners can build authentication tools that are not only technically sophisticated but also accountable, interpretable, and credible in real-world public health contexts.

References

1. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM.
2. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates.
3. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115.
4. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
5. 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). IEEE.

6. 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.
7. Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G., & Kell, D. B. (2004). Metabolomics by numbers: Acquiring and understanding global metabolite data. *Trends in Biotechnology*, 22(5), 245–252.
8. 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., ... Viant, M. R. (2007). Proposed minimum reporting standards for chemical analysis. *Metabolomics*, 3(3), 211–221.
9. Wolfender, J.-L., Marti, G., Thomas, A., & Bertrand, S. (2015). Current approaches and challenges for the metabolite profiling of complex natural extracts. *Journal of Chromatography A*, 1382, 136–164.
10. Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11), 4793–4813.
11. Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *ITU Journal: ICT Discoveries*, 1(1), 39–48.
12. Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36–43.
13. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
14. Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4), 1059–1086.
15. Heinrich, M., Barnes, J., Prieto-Garcia, J., Gibbons, S., & Williamson, E. M. (2018). *Fundamentals of Pharmacognosy and Phytotherapy* (3rd ed.). Elsevier.
16. Ebrahimi, P., Viereck, N., Bro, R., & Engelsen, S. B. (2017). Chemometric analysis of NMR spectra. In *Modern Magnetic Resonance* (pp. 1–20). Springer.
17. Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605.
18. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
19. 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.

20. Biecek, P., & Burzykowski, T. (2021). *Explanatory Model Analysis*. Chapman and Hall/CRC.
21. Alvarez-Melis, D., & Jaakkola, T. S. (2018). Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems* 31 (pp. 7775–7784).
22. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11), e745–e750.