Below is a concise list of five sources you can refer to, with short notes on their relevance to your project. These papers cover both theory and applied solutions (particularly in agriculture) for black-box interpretability in neural networks.

1. **Lundberg, S. M., & Lee, S.-I. (2017).**  
   *A Unified Approach to Interpreting Model Predictions.* NeurIPS.
   * **Key Point:** Introduces SHAP, a method that uses game theory to explain how features influence a prediction in complex models.
2. **Radenovic, F., Dubey, A., & Mahajan, D. (2022).**  
   *Neural Basis Models for Interpretability.* NeurIPS.
   * **Key Point:** Proposes a neural network architecture (NBM) that’s inherently interpretable, building on ideas from Generalized Additive Models.
3. **Feng, J., & Xu, X. (2024).**  
   *Deciphering Plant Seedlings: Enhancing Classification and Interpretability with Vision Transformers.* IEEE Conference (CVIDL).
   * **Key Point:** Applies explainable Vision Transformers in agriculture for plant seedling classification, visualizing *why* the model makes each decision.
4. **Mahmud, T., et al. (2023).**  
   *Explainable AI for Tomato Leaf Disease Detection: Insights into Model Interpretability.* ICCIT, IEEE.
   * **Key Point:** Uses CNNs with saliency maps/Grad-CAM to help farmers see exactly which part of the leaf image triggered a disease classification.
5. **Das, S., & Chatterjee, S. (2023).**  
   *Explainable Machine Learning for Crop Recommendation from Agriculture Sensor Data – A New Paradigm.* Preprint (arXiv/ResearchGate).
   * **Key Point:** Integrates sensor data for crop recommendation and uses methods like LIME/SHAP so that farmers understand *why* a given crop is recommended.