Available online www.jsaer.com

***Journal of Scientific and Engineering Research*, 2022, 9(10):95-102**

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Explainable AI Models for Risk Assessment in Banking

Abhishek Shivanna

abkshvn@gmail.com

Abstract: In the evolving landscape of banking, effective risk assessment has traditionally relied on linear models and tree-based models to analyze data and guide decision-making. However, as the complexity of financial data and the demand for more accurate predictions increase, banks are transitioning toward deep neural networks, which offer greater predictive power. This shift to deep learning has raised concerns about transparency, as these models often function as "black boxes," making it difficult to interpret their decisions. This paper explores the role of explainable deep learning models in modernizing risk assessment processes, bridging the gap between model complexity and the need for interpretability. By comparing traditional methods with deep neural networks, we highlight the importance of explainability in enhancing trust, regulatory compliance, and decision-making. Furthermore, we examine the challenges posed by black-box models and discuss how explainable deep learning can provide solutions to meet the stringent demands of regulators while ensuring fairness and accountability in risk assessments.

Keywords: explainable ai, risk assessment, deep neural networks, banking, transparency, interpretability, machine learning, regulatory compliance, credit risk, fairness

1. Introduction

The banking sector is inherently risk-averse [1], where accurate and reliable risk assessment is paramount to maintaining financial stability and regulatory compliance. Traditionally, risk assessment models in banking have relied on linear models, such as logistic regression, or tree-based models, such as decision trees and random forests [2]. These models are favored for their simplicity, interpretability, and ability to provide insights into how various factors contribute to credit risk, default probabilities, and other critical financial metrics. However, with the increasing complexity of financial markets and the growing volume and diversity of data available for analysis, traditional models are being stretched to their limits in delivering accurate predictions [3].

One of the most significant shifts driving this transition is the rising importance of unstructured data, such as text, images, and transactional histories, which have the potential to offer deeper insights into customer behavior, market trends, and risk factors [4]. Traditional models are limited in their ability to process and learn from unstructured data, making it difficult for banks to fully leverage these valuable resources. This limitation has led to the adoption of deep learning models, particularly deep neural networks, which excel at handling large, complex datasets that include both structured and unstructured information. Deep learning's ability to capture intricate patterns in these diverse data sources provides a significant advantage in improving the accuracy and robustness of risk assessments.

However, while deep learning models offer powerful predictive capabilities, they introduce a major challenge: the lack of transparency. Often referred to as "black boxes," these models make it difficult to understand their decision-making processes, raising concerns around trust, regulatory compliance, and accountability. This is especially important in banking, where clear and interpretable risk assessments are necessary for building confidence with stakeholders and adhering to stringent regulatory frameworks.

In response to these concerns, explainable AI (XAI)[5] has emerged as a vital area of research. Explainable deep learning models seek to maintain the high predictive accuracy of neural networks while enhancing transparency and interpretability, enabling banks to extract insights from unstructured data without sacrificing the clarity

required for risk management. This paper explores the role of explainable deep learning models in modernizing risk assessment. By comparing traditional, interpretable methods with more advanced but opaque neural networks, we examine how explainability tools can bridge the gap between predictive power and transparency. Furthermore, we discuss the challenges posed by black-box models, particularly in terms of fairness, trust, and regulatory compliance, and consider how explainable deep learning can provide solutions that meet the demands of modern banking, ensuring both accuracy and accountability.

2. Traditional Risk Assessment Models

Historically, traditional risk assessment models have played a pivotal role in the banking sector by helping institutions evaluate creditworthiness, assess market risk, and make informed lending decisions. These models, which are often built on structured financial data, rely heavily on statistical techniques that are interpretable and easy to implement. Among the most widely used approaches are linear models, such as logistic regression, and tree-based models, like decision trees and random forests. These models have been favored for their simplicity, transparency, and ability to generate straightforward explanations that are essential for regulatory reporting and decision-making processes.

Linear Models

Logistic regression and other linear models have long been the backbone of risk assessment in banking. These models assume a linear relationship between input variables (e.g., income, credit score, loan amount) and the target outcome (e.g., probability of default). The strength of linear models lies in their interpretability — the relationship between each predictor and the outcome is explicit, making it easy to identify which factors contribute to increased risk. This interpretability is critical for regulatory compliance, as it allows banks to justify their decisions in a clear and understandable manner. However, linear models often struggle with capturing complex, non-linear relationships in data, limiting their predictive power in more intricate risk assessment scenarios.

An example of using linear model (classifier) in risk assessment is classifying loan default based on a set of borrower characteristics, as seen in Fig 1 [13] which allows for a very explainable model based on the features we are drawing a decision boundary.

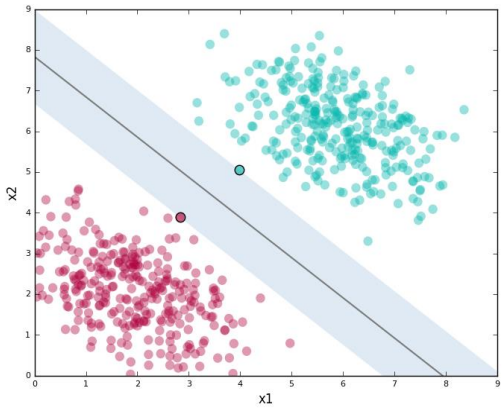

Figure 1. Linear Models

Interpretability: One of the major strengths of using linear regression is that it allows the bank to easily interpret the effect of each feature on the likelihood of default. For example:

- A negative coefficient for **income** would suggest that higher-income borrowers are less likely to default.
- A positive coefficient for **loan amount** would indicate that larger loans increase the probability of default.
- A low **credit score** with a large positive coefficient would imply a significant risk factor for default.

This kind of model is not only easy to interpret, but it also provides transparency, which is crucial when explaining risk decisions to regulators or internal stakeholders. However, as mentioned before, the limitations of linear regression emerge when non-linear relationships or interactions between variables exist, or when unstructured data needs to be considered.

Tree-Based Models

Decision trees and their more advanced versions, such as random forests and gradient-boosted trees, are another set of traditional models widely used in risk assessment. These models work by recursively splitting the dataset based on the most significant predictors, creating a tree-like structure of decisions. Tree-based models are generally more flexible than linear models, as they can capture non-linear relationships and interactions between variables [6]. They are also relatively easy to interpret, as the decision paths can be visualized and understood by non-technical stakeholders. Random forests (as seen in Fig 2 [14]) and gradient boosting further enhance predictive performance by combining multiple decision trees to reduce overfitting and improve accuracy.

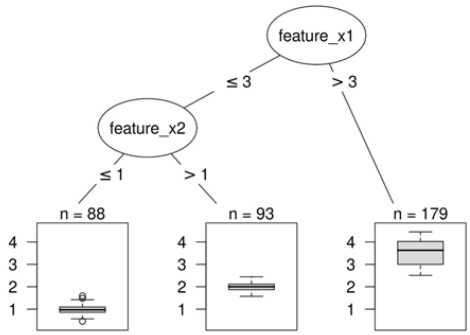

Figure 2. Decision Tree Based Model

Advantages of Traditional Models

The simplicity and interpretability of these traditional models make them well-suited for risk assessment, particularly in highly regulated environments like banking. Linear and tree-based models provide clear insights into the factors driving risk, making it easier for banks to explain their decision-making to regulators, auditors, and stakeholders. Additionally, traditional models are computationally efficient and can be deployed quickly, making them ideal for real-time risk evaluation in scenarios where transparency and speed are critical.

Limitations of Traditional Models

Despite their advantages, traditional models are constrained by their inability to handle unstructured data and complex, non-linear patterns. As the volume and variety of financial data grow, including unstructured data sources such as customer reviews, emails, and social media activity, traditional models struggle to fully capture the predictive signals embedded in these datasets. Moreover, the rigid structure of linear models limits their effectiveness in identifying interactions between variables that are critical for accurate risk assessment. Tree-based models, while more flexible, can also become difficult to interpret when used in ensemble methods like random forests, where the output is an aggregation of multiple decision trees.

As financial institutions face increasing pressure to improve predictive accuracy and leverage the growing wealth of available data, the limitations of traditional risk assessment models have prompted a shift toward more sophisticated approaches. The need to incorporate unstructured data and model complex relationships has spurred the adoption of deep learning techniques, which promise enhanced predictive capabilities but come with challenges of their own, particularly around explainability and transparency.

3. Emergence of Deep Neural Networks

The rise of deep neural networks (DNNs) in banking is driven by several key factors, reflecting both the growing complexity of financial data and the limitations of traditional risk models. Historically, banks have relied on linear models, such as logistic regression, and tree-based models, like decision trees and random forests, for risk assessment tasks. While these models are interpretable and computationally efficient, they struggle to capture non-linear relationships and fail to fully leverage the high-dimensional, unstructured data now abundant in modern banking. Financial institutions handle increasingly diverse data sources—ranging from transactional histories, customer demographics, and credit reports to real-time behavioral data and alternative data sources such as social media or geolocation data. This data complexity requires more sophisticated models capable of recognizing hidden patterns and correlations.

Deep neural networks, with their ability to model complex, non-linear relationships, have emerged as powerful tools for addressing these challenges. By employing multiple layers of interconnected neurons, DNNs can automatically learn feature representations from raw data, making them highly effective in processing large volumes of both structured and unstructured data. This ability is particularly beneficial in banking, where subtle patterns in customer behavior or transaction flows can have significant implications for credit scoring, fraud detection, and other risk assessment activities. For example, DNNs can improve credit risk models by learning nuanced representations of customer repayment behavior, which might not be evident through traditional statistical models.

Despite these advantages, the rise of DNNs in banking brings with it several critical implications. One of the most pressing concerns is the "black box" nature of deep neural networks, which makes it difficult for banks, regulators, and even the model's developers to understand how the model arrives at a specific prediction. This lack of interpretability poses significant risks in a heavily regulated industry where transparency and accountability are paramount. Regulatory bodies, such as the Basel Committee on Banking Supervision [7] and the European Union's General Data Protection Regulation (GDPR)[8], require that financial institutions can explain their models' decisions, especially when these decisions affect individuals' financial well-being.

Moreover, the shift toward DNNs may create challenges related to bias and fairness. As DNNs learn from historical data, they may inadvertently perpetuate biases present in the data, leading to unfair treatment of certain customer groups. For instance, if historical lending decisions were biased against specific demographics, DNNs might reinforce these biases in their risk predictions. Ensuring fairness and addressing bias within deep learning models is an ongoing area of research and concern for banks adopting these advanced technologies.

The adoption of DNNs in risk assessment also has operational and technical implications. Training and deploying deep neural networks require significant computational resources and expertise in machine learning, which may be a barrier for some institutions. Additionally, DNNs often require large amounts of labeled data to achieve optimal performance, which can be a challenge in some risk assessment contexts where high-quality data is not readily available.

In summary, the emergence of deep neural networks in banking risk assessment represents both an opportunity and a challenge. While these models offer improved predictive accuracy and the ability to handle complex data, they also raise concerns around interpretability, fairness, and operational complexity. Banks must navigate these challenges carefully, ensuring that they can leverage the power of deep learning without compromising transparency, regulatory compliance, or customer trust.

4. Addressing Drawbacks and Enhancing Transparency

As deep neural networks (DNNs) become more prevalent in banking risk assessment, the need for explainability has become critical to addressing several key drawbacks associated with their use. DNNs are often criticized for being "black boxes," meaning that their decision-making processes are opaque and difficult to interpret. This lack of transparency is a significant issue in banking, where regulatory requirements, customer trust, and accountability demand that models provide not only accurate predictions but also understandable reasoning. To mitigate these challenges, explainable AI (XAI) techniques have emerged as vital tools for making deep neural networks more transparent and interpretable while preserving their powerful predictive capabilities.

Key Explainable AI Techniques for Deep Neural Networks

Several techniques have been developed to provide insight into how deep neural networks arrive at their predictions. These methods can be categorized into post-hoc explainability methods, which generate explanations after the model has made a prediction, and inherent interpretability methods, which modify the model's structure to make it more understandable.

1. LIME (Local Interpretable Model-agnostic Explanations): LIME [9] is a post-hoc explainability technique that creates simple, interpretable models (such as linear models) around the predictions of a complex DNN (example in Fig 3). It does so by perturbing the input data and observing how changes affect the output, thus identifying which features were most important in making a specific prediction. In a banking context, LIME can be used to explain why a deep learning model flagged a particular transaction as high-risk or why a customer was denied a loan. By offering localized explanations, LIME allows stakeholders to understand the model's behavior in individual cases, improving transparency and trust.

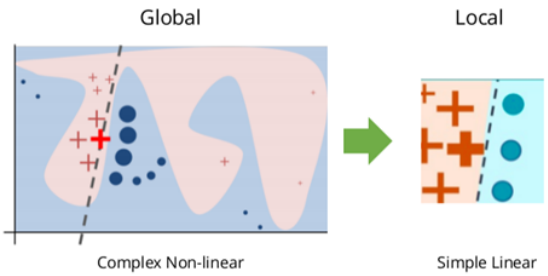

Figure 3. Local Interpretable Model-agnostic Explanations [9]

2. SHAP (SHapley Additive exPlanations): SHAP [10] is another widely used explainability technique that leverages concepts from cooperative game theory to assign an importance score (Shapley value) to each feature in a model. It quantifies how much each feature contributes, positively or negatively, to a particular prediction (as seen in Fig 4 [15]). SHAP values are consistent and accurate in representing feature importance across different model types, including deep neural networks. In risk assessment, SHAP can help explain how various factors, such as a customer's credit history, income, or spending patterns, influenced the predicted risk score. SHAP is particularly valuable because it provides both local and global interpretability, helping explain individual predictions as well as broader model behavior.

Figure 4. SHapley Additive exPlanations (SHAP)

3. Counterfactual Explanations: Counterfactual explanations provide a "what if" scenario, showing how slight changes in the input data could lead to a different outcome. For example, if a customer is denied a loan based on a DNN's prediction, a counterfactual explanation [11] could show that if the customer's income had been $5,000 higher, the loan would have been approved. This approach not only explains the decision but also gives actionable insights into how outcomes could be changed, which is crucial in a customer-facing industry like banking.

4. DeepLIFT (Deep Learning Important FeaTures): DeepLIFT [12] is a method specifically designed for deep neural networks. It works by attributing each neuron's contribution to the final prediction based on a reference point (usually the average behavior of the data). It decomposes the output prediction into the contributions of each input feature, similar to backpropagation, but with more transparency. In a banking context, DeepLIFT can help explain the factors behind a model's assessment of a customer's risk profile, making it clear how various inputs (e.g., credit score, transaction history) influenced the final decision.

5. Model-Agnostic Feature Importance: This technique evaluates feature importance by altering or removing specific features and observing how it impacts the model's predictions. It is model-agnostic and works well with deep neural networks. This approach is useful for understanding which features the DNN deems most important for its predictions, even though the model itself remains complex. In a banking scenario, it could reveal whether income, repayment history, or account activity are the primary factors driving the model's risk assessments.

Implications For Banking

By incorporating these explainability techniques, banks can address several key drawbacks of deep neural networks. First, they help meet regulatory demands for transparency by providing clear, interpretable reasons for the model's decisions. Regulatory bodies often require institutions to justify their risk models' decisions, and

XAI ensures that deep learning models can offer understandable explanations, even for complex, high-stakes predictions.

Second, XAI improves trust and fairness. By providing insight into how models make decisions, explainable AI helps banks ensure that their models are not perpetuating bias or making unfair predictions. If a model is found to be biased against certain demographics or customer segments, explainability tools like SHAP and counterfactual explanations can help identify and mitigate these biases. This is especially important in the context of loan approvals, credit scoring, and fraud detection, where biased or opaque decisions can have significant financial and reputational consequences.

Finally, explainable AI enhances decision-making by making deep learning models more actionable. When stakeholders—including risk managers, auditors, and customers—can understand how and why a model made a particular decision, they are more likely to trust the model and use its outputs in their decision-making processes. Explainability also aids in model debugging and optimization, allowing data scientists to refine models by understanding their strengths and weaknesses in specific scenarios.

In summary, explainable AI plays a crucial role in mitigating the drawbacks of deep neural networks in banking risk assessment. By using techniques such as LIME, SHAP, saliency maps, counterfactuals, and DeepLIFT, banks can ensure their models are not only powerful but also transparent, fair, and compliant with regulatory requirements. This combination of accuracy and interpretability is essential for the responsible use of deep learning in financial services.

5. Future Trends

As deep neural networks (DNNs) continue to transform banking and risk assessment, the field of explainability is also rapidly evolving to address the unique challenges posed by these complex models. Ensuring that DNNs can be both highly accurate and interpretable is at the forefront of research, as financial institutions increasingly rely on these models for critical decision-making tasks. Several emerging trends and challenges are shaping the future of explainable AI (XAI) in deep learning, with a focus on balancing model performance, fairness, and transparency.

End-to-End Explainability

A growing trend in research is the development of end-to-end explainable deep learning models, where interpretability is built into the model architecture from the start, rather than relying solely on post-hoc methods like LIME or SHAP. One promising approach involves integrating inherently interpretable layers or modules within DNN architectures, such as attention mechanisms or self-explanatory neural networks (SENN). These models aim to provide explanations at each layer, offering insight into how data flows through the network and how decisions are made. In banking, this could allow for more transparent credit scoring models or fraud detection systems that are both interpretable and powerful, addressing concerns about the opaque nature of deep learning.

Causal Interpretability

Traditional XAI techniques focus on correlation-based explanations, which can be limited in their ability to provide true causal insights. Researchers are increasingly exploring methods that combine deep learning with causal inference to offer more robust explanations of model behavior. By identifying causal relationships between input features and outcomes, these models can provide more actionable insights into why a certain decision was made, which is crucial in risk assessment scenarios. For example, causal interpretability could help a bank understand not only which factors influence a customer's loan approval but also how changes to specific variables would alter the outcome in a meaningful way.

Explainability for Unstructured Data

As banking increasingly incorporates unstructured data—such as text from customer reviews, social media, and transaction histories—explaining deep learning models that process this data poses new challenges. Cutting-edge research is focusing on enhancing explainability for natural language processing (NLP) and computer vision models. Techniques such as attention-based mechanisms in NLP and feature visualization in convolutional neural networks (CNNs) are being developed to explain how models handle unstructured data. For instance, in fraud detection, an explainable NLP model could highlight specific phrases or transaction details that triggered a high-risk alert, offering more transparency in decision-making.

Explainability in Federated and Distributed Learning

As banks and financial institutions adopt federated learning and distributed systems to build models collaboratively across multiple institutions while preserving data privacy, new explainability challenges arise. In these settings, models are trained on decentralized data without central access to raw data. Researchers are exploring how to ensure that explanations for DNNs trained in federated environments are as transparent and interpretable as those trained on centralized data. Techniques like federated SHAP and federated counterfactual explanations are emerging to address these challenges, ensuring that models trained across different data silos can still provide clear, interpretable decisions.

Ethical and Fairness Concerns in Explainable AI

As banks increasingly rely on DNNs for high-stakes decisions, ensuring fairness and ethical decision-making is becoming a critical area of research. Bias in deep learning models, especially in financial services, can have severe consequences, leading to discriminatory outcomes in loan approvals, credit scoring, or fraud detection. Researchers are working on explainability techniques that not only provide insight into model behavior but also flag potential biases in the data or decision-making process. Methods like fairness-aware explainability, which highlight discriminatory patterns within models, are being developed to ensure that banking institutions can trust their models to make fair decisions across different customer demographics. These techniques are essential for banks to comply with regulatory requirements and ethical standards while building trust with their customers.

Challenges in Scaling Explainable AI for Banking

While significant progress is being made in explainable AI, several challenges remain in scaling these techniques across large financial institutions. One major challenge is the trade-off between model complexity and interpretability. As models become more complex to improve predictive performance, ensuring they remain interpretable becomes increasingly difficult. Banks must carefully balance the need for highly accurate models with the ability to explain their decisions to regulators, customers, and internal stakeholders.

Another challenge is the operationalization of explainable AI in real-time environments. Banking applications often require real-time decision-making, such as in fraud detection or credit approval, where speed is critical. Many XAI techniques, however, are computationally intensive, making it difficult to deploy them at scale without sacrificing performance. Ongoing research is focused on developing more efficient, scalable explainability methods that can be integrated into production systems without compromising speed or accuracy.

Finally, the legal and regulatory landscape continues to evolve, with increasing demands for model transparency and accountability. As regulations around AI in financial services become stricter, banks will need to ensure that their deep learning models meet these standards, even as the complexity of these models grows. Future XAI research will need to focus on aligning cutting-edge techniques with these evolving regulatory requirements, ensuring that explainability is not only a technical challenge but also a strategic priority for the banking industry.

In conclusion, the future of explainable AI in deep neural networks is full of exciting possibilities and challenges. With ongoing advancements in end-to-end interpretability, causal explanations, and fairness-aware AI, banks will be able to build more transparent, accountable, and powerful risk assessment models. However, significant work remains in scaling these innovations to meet the operational, ethical, and regulatory demands of the modern banking environment.

6. Conclusion

In conclusion, the integration of deep neural networks into banking risk assessment offers significant potential to improve accuracy, efficiency, and the ability to process complex financial data. However, the black-box nature of these models presents critical challenges around transparency, accountability, and regulatory compliance. Explainable AI techniques provide a pathway to address these challenges by making deep learning models more interpretable and ensuring that decisions can be understood and trusted by regulators, bank executives, and customers alike. As the field of explainable AI continues to advance, cutting-edge methods such as causal inference, model compression, and federated explainability will become essential for banks to navigate the growing complexity of their risk models. Balancing the power of deep neural networks with the need for transparency, fairness, and regulatory alignment is not just a technical imperative but a strategic one, shaping the future of risk assessment in the financial industry.

References

[1]. Nishiyama, Y. (2007). Are Banks Risk-averse?. Eastern Economic Journal, 33, 471-490. https://doi.org/10.1057/EEJ.2007.36.

[2]. Chang, Y., Chang, K., & Wu, G. (2018). Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. Appl. Soft Comput., 73, 914-920. https://doi.org/10.1016/j.asoc.2018.09.029.

[3]. Mercep, A., Mrčela, L., Birov, M., & Kostanjčar, Z. (2020). Deep Neural Networks for Behavioral Credit Rating. Entropy, 23. https://doi.org/10.3390/e23010027.

[4]. Tsui, E., Wang, W., Cai, L., Cheung, C., & Lee, W. (2014). Knowledge-based extraction of intellectual capital-related information from unstructured data. Expert Syst. Appl., 41, 1315-1325. https://doi.org/10.1016/j.eswa.2013.08.029.

[5]. Saeed, W., & Omlin, C. (2021). Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities. Knowl. Based Syst., 263, 110273. https://doi.org/10.1016/j.knosys.2023.110273.

[6]. Lundberg, S., Erion, G., Chen, H., DeGrave, A., Prutkin, J., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2, 56 - 67. https://doi.org/10.1038/s42256-019-0138-9.

[7]. Rost, B. (2010). Basel Committee On Banking Supervision., 319-328. https://doi.org/10.1163/EJ.9789004163300.I-1081.238.

[8]. Tikkinen-Piri, C., Rohunen, A., & Markkula, J. (2017). EU General Data Protection Regulation: Changes and implications for personal data collecting companies. Comput. Law Secur. Rev., 34, 134-153. https://doi.org/10.1016/J.CLSR.2017.05.015.

[9]. Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672.2939778.

[10]. Murdoch, W., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116, 22071 - 22080. https://doi.org/10.1073/pnas.1900654116.

[11]. Feder, A., Oved, N., Shalit, U., & Reichart, R. (2020). CausaLM: Causal Model Explanation Through Counterfactual Language Models. Computational Linguistics, 47, 333-386. https://doi.org/10.1162/coli_a_00404.

[12]. Shrikumar, A., Greenside, P., Shcherbina, A., & Kundaje, A. (2016). Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. ArXiv, abs/1605.01713.

[13]. https://medium.com/@paarthbir/image-classification-with-a-linear-classifier-cab02f7f8a30

[14]. https://christophm.github.io/interpretable-ml-book/tree.html

[15]. https://shap.readthedocs.io/en/latest/