Towards Transparency in Black-Box Models: A Brief Review of Methods in Explainability for AI Systems

Saleh Alkhalifa*

Northeastern University

Abstract

In recent years, the integration of AI models within life sciences, biopharmaceuticals, and other regulated industries has significantly advanced, influencing domains such as regulatory affairs, manufacturing, process development, and quality control. However, as discriminative and generative models become increasingly sophisticated, their decision-making processes often remain opaque, giving rise to the so-called "black-box problem." This paper explores the most recent and commonly used approaches for developing and evaluating distinct AI models, with a focus on both regression and classification tasks within discriminative AI, generative AI, and deep learning frameworks. By applying explainability techniques such as SHAP, LIME, Integrated Gradients, and Attribution, we offer practical insight into interpreting these models. The overarching goal is to improve transparency, foster trust, and ensure ethical compliance in AI systems deployed in sensitive and regulated environments.

1. Introduction

Artificial Intelligence (AI) and Machine Learning (ML) technologies have rapidly emerged as transformative tools across various industries, including life sciences, biotechnology, finance, and healthcare. These technologies excel in automating complex decision-making processes, enabling accelerated progress in areas such as drug discovery, disease diagnosis, and financial forecasting. However, the increasing reliance on AI has spotlighted a critical challenge: the "black box" problem. Many advanced and widely used models, such as neural networks and large language models, generate highly accurate predictions but lack transparency in their decision-making processes. This opacity poses significant challenges in high-stakes scenarios, particularly in domains where ethical compliance, trustworthiness, and interpretability are paramount. As highlighted in [1], achieving interpretable machine learning is no longer optional but an essential requirement to ensure that AI systems adhere to rigorous ethical and regulatory standards.

The research community has developed numerous explainability methods to address the black-box problem. These methods range from explanation methods, such as SHAP [1] and LIME [2], to intrinsic techniques like self-explaining deep learning neural networks [13]. SHAP and LIME frameworks have gained traction due to their versatility across various datasets and models.

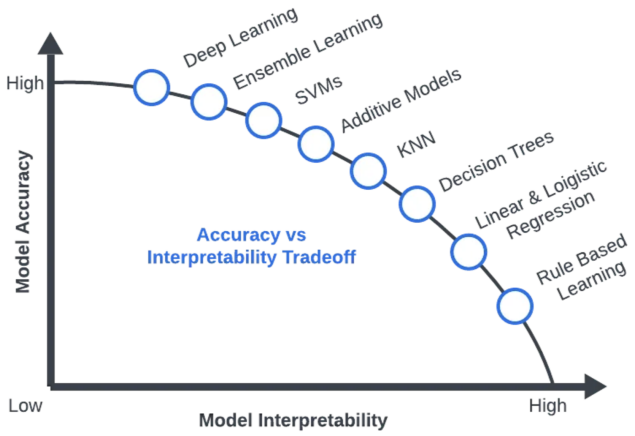

Figure 1: A visualization depicting the tradeoff between accuracy and interpretability.

While SHAP uses Shapley values can help to quantify feature importance, LIME approximates model behavior. This can be done specifically locally by perturbing input data. Both methods excel in structured datasets, providing insights into feature contributions at global and local levels. However, as noted in [6], these methods are not without limitations. For example LIME can sometimes generate inconsistent explanations due to its reliance on local approximations, highlighting the need for a more robust evaluation of interpretability techniques.

Explainability techniques for unstructured data, such as text and images, have followed a different path way. Integrated Gradients [3] and Grad-CAM [11] have become foundational tools and methods for understanding complex neural networks. Integrated Gradients attribute predictions to input features by integrating gradients along a path from a baseline, making it suitable for tasks like text classification and image recognition. On the other hand, Grad-CAM provides visual explanations by highlighting regions in an image that contribute most to a model's prediction. These methods address the challenges in high-dimensional data but also face limitations, such as sensitivity to a given baseline selection in Integrated Gradients. Recent work in [11] shows how these techniques can complement each other providing a more generalized understanding of deep learning models.

Generative models, particularly popular large language models like GPT-4, introduce additional challenges and opportunities in explainability. While their proprietary status often limits and prohibits direct inspection, token-level importance methods and sensitivity analyses provide valuable insights into their decision-making. Research in [5] and [4] has explored multiscale attention visualizations to demystify transformers, revealing how attention mechanisms can prioritize specific tokens or phrases. These methods align with findings in [13], which argue that self-explaining architectures can improve the interpretability of complex systems without sacrificing performance. However, as noted in [7] and [8], a significant gap remains in quantifying the overall robustness and reliability of explainability algorithms for generative models. These findings are overall consistent in both methods and results.

Comparative analyses reveal strengths and weaknesses across these explainability approaches. SHAP and LIME excel in structured data due to their straightforward interpretability, but they struggle with scalability in high-dimensional tasks. In this paper, we will explore both sides of this fact. Additionally, methods like Integrated Gradients and Grad-CAM are better suited for unstructured data but require careful parameter tuning and baseline selection. Counterfactual explanations [6] offer a unique perspective by identifying minimal changes to input data that alter predictions, bridging the gap between actionable insights and interpretability. Despite their complementary strengths, there is a pressing need for unified frameworks, as emphasized in [12], to standardize the evaluation and application of these methods across diverse domains.

2. Data

Multiple datasets were utilized to explore explainability across different types of models and tasks:

- **California Housing Dataset (Regression):** Used to predict median house values, offering a continuous target variable and structured input features ideal for regression tasks.

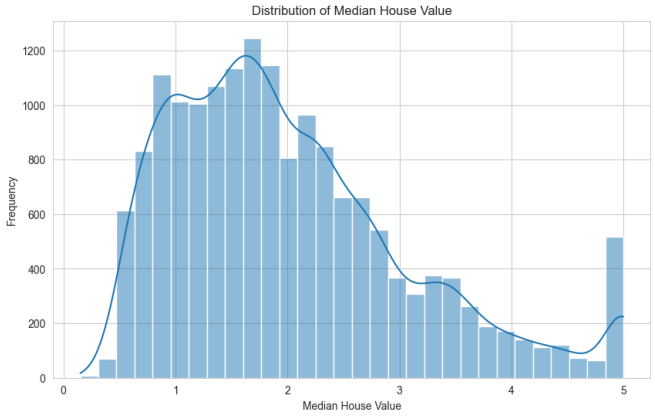

Figure 2: A visualization depicting the distribution of the California Housing Dataset.

- **Breast Cancer Dataset (Classification):** We used this dataset for binary classification, distinguishing malignant from benign tumors based on diagnostic features.

- **AG News Dataset (Open Source Generative AI):** A text classification task used to evaluate token-level attributions in a transformer-based language model.

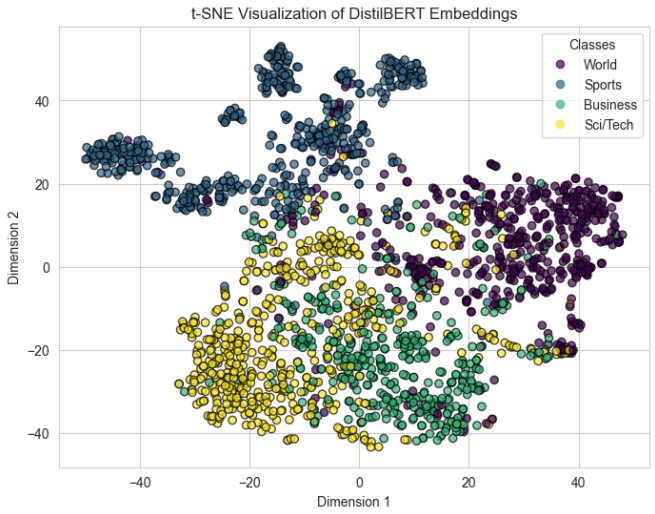

Figure 3: A visualization of the AG News Dataset represented via t-SNE.

- **Sentiment Analysis Data (Closed Source Generative AI):** Sentiment classification scenarios where the model assigns positive, negative, or neutral sentiments to input text.
- **MNIST Dataset (Deep Learning):** Handwritten digit classification ideal for testing pixel-level attribution methods in image-based tasks.

These datasets represent a broad range of input modalities (tabular, textual, and image data) and prediction tasks (regression, classification, sentiment analysis), providing a comprehensive environment for investigating explainability techniques.

Data Preprocessing and Visualization

Before training the models, each dataset underwent careful preprocessing to ensure data quality and integrity:

- **Outlier Removal:** For numerical datasets (California Housing), we implemented outlier detection using the Interquartile Range (IQR) method. Data points falling outside this generally typical range were removed to reduce the influence of extreme or incorrect values on the models. This step was important for improving model stability. This is because certain explainability techniques assume that the input space is consistent and representative of normal conditions.
- **Normalization and Scaling:** Features in tabular datasets were scaled (StandardScaler) to ensure that no single feature strongly dominated the model training process. This scaling improves both the model performance and the coherence of feature attributions from methods like SHAP and LIME.
- **Text Tokenization and Truncation:** For textual data (AG News and custom GPT-4o prompts), inputs texts were first were tokenized using appropriate tokenizers (DistilBERT's tokenizer) and then truncated to fixed lengths. This preprocessing step standardized input sizes, making it much easier to attribute model predictions to specific tokens.
- **Image Normalization:** For MNIST, images were normalized to standard ranges to ensure consistency and better gradient-based explanations. Integrated Gradients for example generally benefits quite a bit from well-conditioned input domains.

Data Visualization and Insight Generation

Visualizing the data before modeling helped guide our choice of models and explainability techniques. Although much of the visualizations focused on past work from CS-6140 through the use of these datasets, the knowledge gained from those investigations were implemented here.

- **Distribution Analysis:** Histograms and KDE plots were used to examine target distributions in the regression tasks (housing prices) and class balance in classification tasks (malignant vs. benign). Understanding the data distribution allows us to anticipate where explainability methods might be most informative for us, like identifying boundary cases or outlier-driven predictions.
- **Feature Correlations:** SNS Heatmaps and pair plots identified feature correlations guiding model selection, fitting, and interpretation. For example, strong correlation between features in the California Housing dataset suggested that Ridge Regression's regularization would be beneficial and that SHAP values would highlight which correlated features were very impactful and useful.
- **Textual Inspection:** For text classification tasks, sampling and reading raw examples from AG News and the GPT-4o sentiment prompts helped us hypothesize which tokens might be most influential. Visualizing the token distributions

and sentence lengths provided context for understanding Integrated Gradients and token-level attributions.

- **Image Examination:** Inspecting the MNIST samples allowed us to determine which digits and strokes might be vital to classification setting the stage for comparing Integrated Gradients explanations with our expectations.

These datasets represent a wide range of input modalities (tabular, textual, and image data) and prediction tasks (regression, classification, sentiment analysis), providing a comprehensive environment for investigating various explainability techniques.

3. Implementations

We implemented five separate models across discriminative, generative, and deep learning domains, each paired with distinct datasets to capture a wide range of data modalities, including tabular (Structured), textual, and image-based data (Unstructured). This selection allowed us to assess model performance and interpretability under different conditions, such as Ridge Regression for linear trends in the California Housing dataset, Random Forest for non-linear classification of the Breast Cancer dataset, and DistilBERT and GPT-4 for textual tasks.

To ensure reliable evaluations and minimize overfitting or underfitting, we used k-fold cross-validation, iteratively training and testing on different subsets of the data to validate generalizability. This approach provided a robust foundation for identifying patterns in model behavior and performance across varying data splits.

In addition we conducted a detailed ablation study testing various configurations by changing model parameters, input preprocessing strategies, and explainability settings. This iterative analysis revealed the factors that significantly influenced both performance and interpretability, such as tree depth in Random Forests or baseline selection in Integrated Gradients. We also analyzed edge errors to uncover instances where models failed, using techniques like SHAP, LIME, and token-level attributions to explain misclassifications.

These insights not only highlighted model limitations but also informed improvements in design and preprocessing, demonstrating how rigorous evaluation strategies combined with explainability methods can lead to more transparent, robust, and trustworthy AI systems.

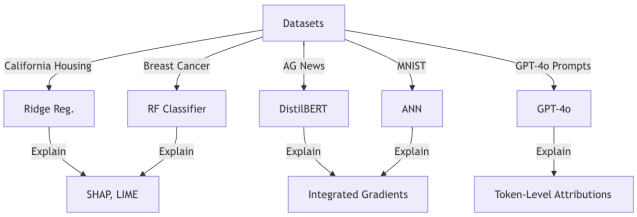

Figure 4: A visualization of the workflow linking datasets, models, and explainability methods.

3.1. Ridge Regression (Discriminative - Regression)

A ridge regression model was trained on the California Housing dataset:

$$\hat{y} = X\beta, \quad \text{with } \beta = \underset{\beta}{\operatorname{argmin}}\{\|y - X\beta\|^2 + \lambda\|\beta\|^2\}. \tag{1}$$

This model's straight forward methodology and linear structure make it an appropriate candidate for feature attribution methods like SHAP and LIME. One thing to note is that the dataset here is not ideal for a linear model. The visualization below demonstrates that the model non-linear nature makes it an ideal candidate for a non-linear model. We proceeded with this purposefully to better understand the impact this would have from a SHAP perspective.

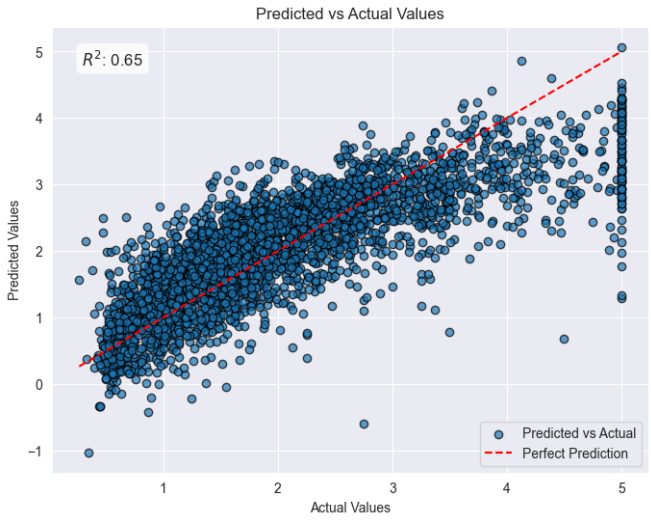

Figure 5: The correlation between the predicted and actual values of our linear model using the California Housing Dataset.

3.2. Random Forest Classifier (Discriminative - Classification)

For the Breast Cancer dataset, a Random Forest Classifier using SKLearn was implemented to tackle the binary classification problem of distinguishing between malignant and benign tumors. Random Forests operate as an ensemble learning method by constructing multiple decision trees during training and aggregating their predictions to produce a final output. The probability of the class $y = c$ is calculated as:

$$P(y = c) = \frac{1}{T} \sum_{t=1}^{T} P_t(y = c), \tag{2}$$

where T represents the total number of trees, and $P_t(y = c)$ is the class probability predicted by each individual tree.

This model is robust to overfitting and noise due to the averaging of multiple decision trees, which can be controlled via the Scikit-learn API. Random Forests inherently provide feature importance metrics making it possible to understand which input features most influence predictions. For example, in this study, diagnostic metrics like mean radius, texture, and symmetry were identified as critical for classifying tumors.

To enhance interpretability further, explainability techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) were integrated. SHAP provided a global overview of feature contributions for us, quantifying their impact on model predictions across the dataset. LIME complemented this by offering instance-specific explanations, highlighting which features drove predictions for individual samples.

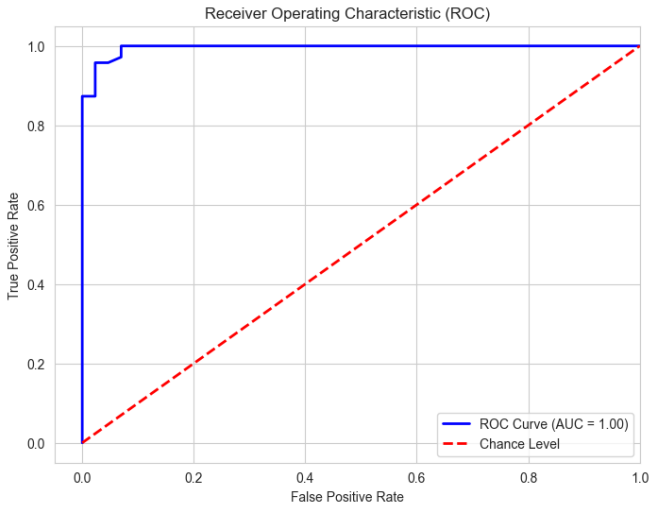

Figure 6: The ROC-AUC of the classification model's results.

This approach not only ensured we achieved a high accuracy and robustness of the classification task but also provided critical insights into the model's decision-making process, making it more interpretable and trustworthy in the context of sensitive medical diagnoses.

3.3. DistilBERT (Generative AI - Open Source)

DistilBERT, a distilled and optimized version of the original BERT (Bidirectional Encoder Representations from Transformers), was fine-tuned on the AG News dataset to be used for text classification tasks. This dataset consists of news articles classified into four categories: World, Sports, Business, and Sci/Tech. The DistilBERT model retains the core architecture of BERT but is 40% smaller, faster to train, and equally powerful in downstream tasks, making it a better choice for academic environments.

The model operates by transforming input text into contextual embeddings through attention mechanisms. These mechanisms allow the model to weigh the importance of different tokens relative to each other. This enables it to capture nuanced semantic

relationships within the text. For example, in a sentence about a business merger the model can focus more on words like "acquisition" or "stocks" to determine that the article belongs to the "Business" category.

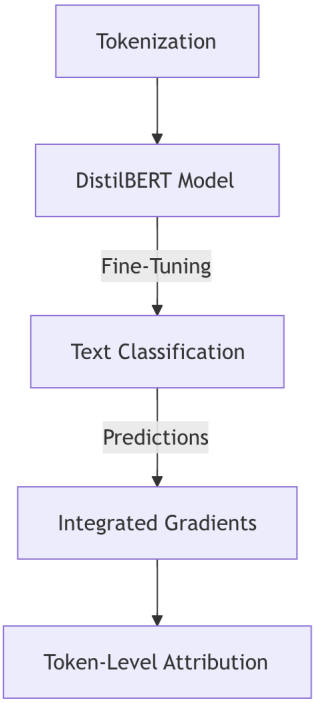

Figure 7: Workflow for the tokenization of, tuning, and prediction of data and subsequent analysis via integrated gradients.

3.4. GPT-4o (Generative AI - Closed Source)

A closed-source large language model (GPT-4o) was prompted for sentiment classification tasks. While proprietary restrictions limit direct internal analysis of weights, prompt engineering and word-level attributions through masking allowed us to infer which words drive the model's sentiment decisions. Although we lack direct architectural insight these methods provided partial transparency.

3.5. Neural Network (Deep Learning)

A feed-forward neural network was trained on the MNIST dataset to classify handwritten digits. Consisting of fully connected layers and ReLU activations, this model's complexity is representative of deep learning architectures. Integrated Gradients were used to visualize pixel-level contributions to predictions, bridging the gap between model complexity and interpretability.

4. Explainability

We used a variety of explainability techniques, each suited to different model types and data modalities:

- **SHAP:** Assigns feature attributions based on Shapley values, ideal for the regression and classification models which can reveal global and local feature importance.
- **LIME:** Locally approximates a black-box model with an interpretable one, which provides instance-specific explanations for classification tasks.
- **Integrated Gradients:** Attributes predictions to input features by integrating gradients along a path from a baseline. Applied to both deep learning (MNIST) and generative AI (DistilBERT, GPT-4o) scenarios.
- **Word-Level Attribution:** Masking words systematically and modifying the input prompts to determine the impact this has on the models ability to predict.

By pairing these techniques with each model, we determined which methods are most effective in uncovering the reasoning behind predictions providing transparency and trust.

5. Results

Explainability aided our understanding of each model's behavior in many different ways. By applying different interpretability tools, we gained insights into what drives model decisions, helping to validate assumptions, build trust, and inform potential improvements.

- **Ridge Regression (SHAP, LIME):** For the Ridge Regression model on the California Housing dataset, both SHAP and LIME confirmed linear assumptions about how features influence housing prices. When it comes to SHAP, by assigning Shapley values to each feature, SHAP offered a global and local perspective to determine feature impact on a decision. This allowed us to see

which features (average number of rooms, median income, etc...) consistently influenced predicted house values. Such global insights are valuable for policy decisions or urban planning since they highlight stable relationships across the dataset.

Figure 8: Visualizing the results from SHAP/LIME showing the underlying reasoning of a model's behavior.

Focusing on individual predictions, LIME created simplified, local surrogate models to explain why the model predicted certain prices for a specific home. For example, if a particular house's predicted price was high due to a combination of low population density and high average income, LIME's local linear approximations can make it clearer to stakeholders which factors are pivotal for that particular instance.

Together, these methods reinforced interpretability by confirming linear patterns and identifying the dominant features, making it easier to trust and act upon the model's predictions.

SHAP Provids a global understanding of which diagnostic features—like mean radius, mean texture, or cell compactness—had the greatest impact on classifying tumors as malignant or benign. By generating SHAP summary plots, we could see a consistent pattern of feature importance, confirming hypotheses or uncovering unexpected influences in a model.

On the other hand, LIME Offers local insights on individual patient cases. For instance, for a specific patient's tumor diagnosis, LIME showed which features tipped the model towards a malignant prediction in the model. This granular view helps clinicians understand the reasoning behind a single outcome, helping in human-AI collaboration for patient care.

With these interpretations, medical professionals can better trust the model's recommendations, as they are not just accurate but also explainable.

- **DistilBERT (Open Source Generative AI, Integrated Gradients):** DistilBERT's text classification on the AG News dataset benefited significantly from Integrated Gradients. By attributing prediction scores back to input tokens, we identified which words or common (and uncommon) phrases drove the classification decision. For instance, for a news headline the model might heavily rely on words like "government," "election," or "stocks" to assign categories such as "World" or "Business" news. Observing these token-level importances validated that the model's reasoning aligns with human thought and linguistic understanding.

  This alignment increases our confidence in the model and in deploying the model for content moderation or information retrieval, knowing it cares about relevant language cues.

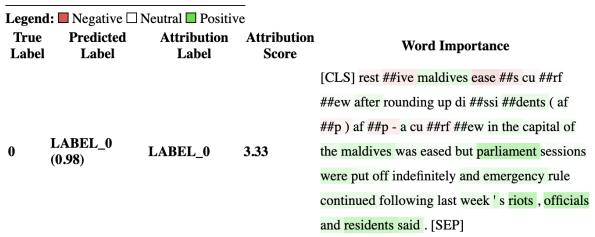

Figure 9: A visualization depiction of token importance of the AG Dataset showing the positive, negative, and neutral impact each token had on the result.

- **GPT-4o (Closed Source Generative AI, Token-Level Attributions and Prompt Engineering):** Despite GPT-4o's proprietary nature, we still gained valuable insights through indirect explainability methods such as masking.

  By examining how altering certain words in the prompt influenced the output sentiment (positive, negative, or neutral), we determined the model's sensitivity to particular linguistic indicators. For example, removing adjectives like "amazing" from a product review prompt could shift the classification from positive to neutral, confirming that emo-

tional descriptors strongly affect GPT-4o's sentiment judgments.

Although we lacked direct access to architectural transparency, this partial interpretability reassured us that the model's conclusions reflect meaningful linguistic patterns rather than arbitrary computational methods.

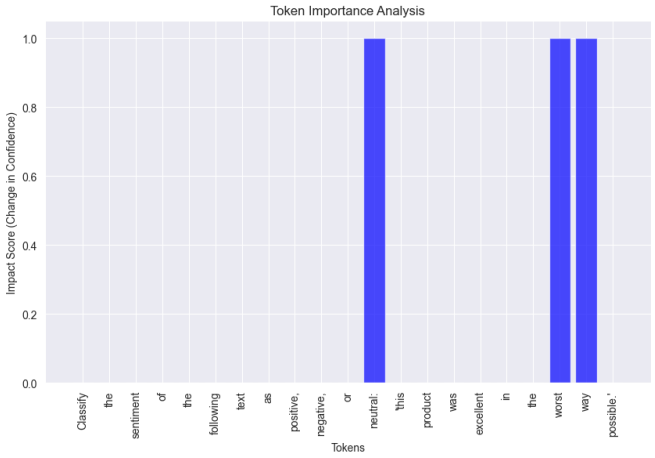

Figure 10: Visual depiction of token importance and which tokens had the most impact when removed or masked.

- **Neural Network (Deep Learning, Integrated Gradients):** In the MNIST digit classification model, Integrated Gradients provided a pixel-level attribution of predictions in classification. Highlighting key strokes and curves that define each digit shape, this method confirmed that the network focused on relevant visual patterns. For example a "7", the horizontal stroke and the angled line might be most influential. When introducing noise, we observed changes in attribution distributions, revealing the model's sensitivity and confirming that there are certain pixels that are indeed crucial for correct classification.

These insights validate that the network's reasoning is not random. Instead, it aligns with human thought of digit outlines, and increasing trust in its predictions.

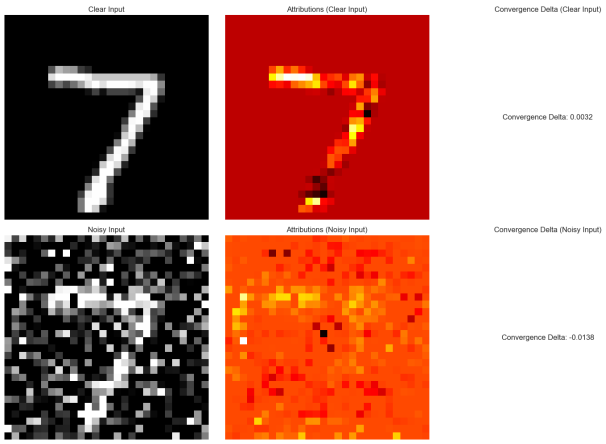

Figure 11: A visualization depiction of the explainability of the deep learning model on the MNIST dataset.

Comparing Methods Across Models

Our findings illustrate that different explainability methods excel in different contexts. Although we did not fully investigate the combination of every dataset and model type with every explainability method, we did identify a few patterns.

- *SHAP* and *LIME* are particularly effective with structured, tabular data through regression and classification tasks. SHAP's global and local consistency paired up with LIME's instance-based insights allowed for both high-level and fine-grained understanding.

- *Integrated Gradients* shined the most when explaining complex models and unstructured data, such as images (MNIST) and text (DistilBERT). Its gradient-based attributions helped decode and understand the hidden patterns within high-dimensional input spaces.

- Even closed-source models like GPT-4o benefited from token-level explainability through prompt engineering. While we lacked the ability to directly measure internal states, changes in output when key tokens were masked or altered gave us indirect insights and clues about decision-making processes.

Overall, each explainability technique contributed in a unique way to unpacking and understanding the model predictions, and reinforcing the notion that no single method is universally superior. Instead, the use of multiple techniques can provide a robust

interpretability framework suitable for diverse AI applications.

6. Conclusion

As AI systems continue to grow in use within regulated and high-stakes environments, ensuring their explainability remains highly important, especially for regulators like the FDA. This study surveyed a combination of datasets, models and interpretability techniques, to provide comprehensive insights into model decisions, enhancing trust, compliance, and utility.

Future directions include experimenting with hybrid explainability frameworks that combine global and local methods, and investigating fairness and bias through interpretability, and extending explanations to more complex datasets and tasks. In addition, another future direction would be to expand the scope of the project to additional datasets, models, and explainability techniques to better understand a wider variety of content.

References

[1] Lundberg, S. M. (2017) A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems.

[2] Ribeiro, M. T., Singh, S. (2016). Why Should I Trust You?: Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD.

[3] Sundararajan, M., Taly, A.(2017). Axiomatic Attribution for Deep Networks. International Conference on Machine Learning (ICML).

[4] Kokhlikyan, N., et al. (2020). Captum: A Model Interpretability Library for PyTorch.

[5] Vig, J. (2019). A Multiscale Visualization of Attention in the Transformer Model.

[6] Mothilal, R. K., Sharma, A.(2020). Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*20)

[7] Geada, R., et al. (2021). TrustyAI Explainability Toolkit.

[8] Lin, Z. Q., et al. (2019). Do Explanations Reflect Decisions? A Machine-centric Strategy to Quantify the Performance of Explainability Algorithms.

[9] Salih, A. M., et al. (2023). A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. Artificial Intelligence and Systems

[10] Vimbi, V., Shaffi, N.(2024). Interpreting Artificial Intelligence Models: A Systematic Review on the Application of LIME and SHAP in Alzheimer's Disease Detection. Brain Informatics

[11] Selvaraju, R. R., Cogswell, M., Das, A., et al. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[12] Doshi-Velez, F., & Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning.

[13] Alvarez-Melis, D. (2018). Towards Robust Interpretability with Self-Explaining Neural Networks. Advances in Neural Information Processing Systems (NeurIPS)

[14] OpenAI. (2024). ChatGPT. Retrieved from https://openai.com/chatgpt