# Towards Robust Interpretability with 3D Rotational Perturbations in Self-Explaining Neural Networks

Khoi Nguyen
Harvard College
khoinguyen@college.harvard.edu

Saketh Mynampati
Harvard College
sbmynampati@college.harvard.edu

February 13, 2025

**Abstract**

Self-Explaining Neural Networks (SENN) are proposed to be interpretable and robust classifiers that operate linearly with respect to higher-order concepts beyond raw pixels, such as strokes and orientations. However, the original framework falls short in relating these concepts to human interpretability and fails to clarify how perturbations influence predictions in the context of concept usage, particularly near decision boundaries. Traditional perturbations like Gaussian noise are insufficient for probing SENN's linear behavior, as they do not align with these concept-level representations or reveal why the model fails to predict accurately under such transformations.

In this work, we analyze the model's decision boundaries using 3D rotational perturbations, which provide interpretable transformations and clearer links between input modifications, concept activations, and predictions. By examining concept activations across rotated MNIST digits, we gain insights into SENN's distributed representations, where patterns of concept activations work in tandem to identify digits. Additionally, rotational perturbations allow us to explore the model's behavior near decision boundaries, highlighting the significance of concept scores in driving predictions.

Our results show that rotating digits around different axes uncovers stable decision-making for some digits and interpretable concept shifts for others. For example, we find that certain concept scores reliably spike at specific angles where a rotated digit visually transitions from one class to another. Moreover, while SENN demonstrates robustness to small rotations in certain axes, other rotations cause performance to degrade quickly. These findings underscore the utility of rotation-based perturbations for understanding SENN's robustness, interpretability, and the role of concepts in its decision-making process.

## 1 Introduction

Self-Explaining Neural Networks (SENN) [**senn**] are designed to enhance interpretability by modeling decisions linearly with respect to a set of higher-order concepts, such as strokes and orientations, rather than raw pixel-level features. Here, *concept-level representations* refer to intermediate features that SENN learns to associate with visually meaningful patterns. Instead of relying solely on pixel intensities, the model identifies abstract concepts—like certain types of curves, loops, or distinct stroke arrangements—that frequently occur in handwritten digits. By doing so, SENN can explain its predictions using terms more closely aligned with human perception. Figure 1 provides an example of learned concept prototypes, illustrating the kinds of patterns SENN deems fundamental to digit recognition.

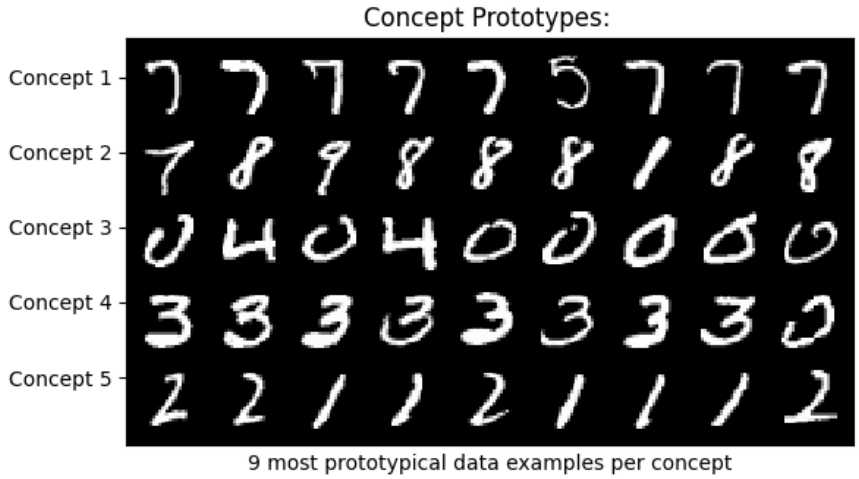


Figure 1: **Learned Concept Prototypes:** Example images representing the types of strokes and shapes associated with each concept. These visually meaningful patterns form the building blocks of SENN's concept-level representations.

### 1.1 Motivation

Traditional perturbation methods like Gaussian noise fail to effectively evaluate SENN's interpretability. Gaussian noise operates at the pixel level, which does not align with SENN's concept-level representations. Consequently, when concept scores diverge due to Gaussian perturbations, there is no meaningful interpretability insight. Such perturbations offer limited value in understanding how and why concept activations drive predictions, especially near decision boundaries. This makes the robustness measured by Alvarez-Melis et al. [**senn**] insufficient for accurately presenting a local explanation of the decision boundary across a desired explanation space (e.g., the set of digit-like strokes, loosely approximated by rotationally perturbing existing digits in the MNIST example).

To address these challenges, we propose using 3D rotational perturbations. Unlike Gaussian noise, these rotations directly modify higher-order features such as stroke orientation and shape, providing interpretable transformations that align with SENN's conceptual framework. By applying rotations to MNIST digits, we analyze changes in concept activations and gain insights into when and why the model's decision boundaries are met.

For example, rotating a digit "3" about an axis might cause it to visually resemble an "8" at certain angles. By observing concept scores during this transition, we can identify which concepts shift to justify the model's changing prediction. This analysis highlights the interplay between physical transformations, concept scores, and model predictions, and tests the SENN's robustness [**ml**].

The importance of these findings is illustrated in Figure 3, where we show that regularized concept scores remain interpretable under transformations, unlike those in unregularized settings.

### 1.2 Impact on Decision Boundaries and Robustness

Rotational perturbations enable a deeper understanding of SENN's robustness and interpretability near decision boundaries. Decision boundaries are regions in the input space where the model transitions between predictions, often corresponding to dramatic shifts in concept scores. By examining concept activations throughout a rotation, we discover how subtle changes in orientation can cause

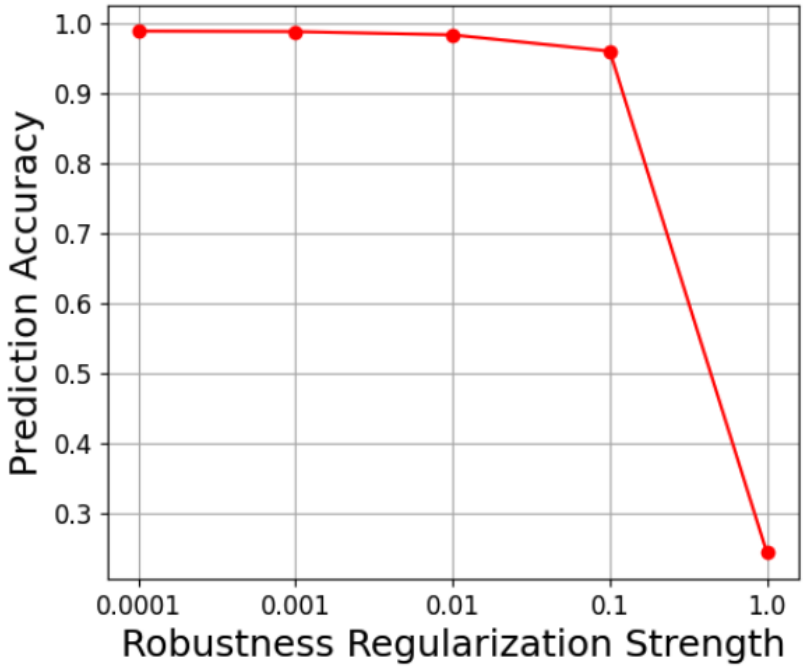


Figure 2: **Effect of Gaussian Noise:** Random pixel-level Gaussian noise offers limited insight into how concept activations relate to interpretability or decision boundaries, as it fails to produce meaningful changes in the learned concept space.

the model to favor different concepts and ultimately switch predicted classes.

This work demonstrates the utility of rotation-based perturbations in probing the robustness and interpretability of SENN. By aligning input transformations with conceptual representations, rotational perturbations provide clearer insights into how concept activations drive predictions and clarify model behavior at decision boundaries.

## 2 Methodology

### 2.1 Dataset Preparation

We use the MNIST dataset, which consists of 60,000 training images and 10,000 testing images of handwritten digits, each of size $28 \times 28$. MNIST is a well-studied, benchmark dataset often used in interpretability research due to its simplicity, high quality, and the ease with which humans can recognize and compare digit shapes. These characteristics make MNIST an ideal platform for probing how a model's concept space responds to controlled, interpretable perturbations.

To ensure stable training and facilitate more uniform concept activations, we normalize each image such that the overall dataset distribution has a mean of 0 and a standard deviation of 1. Specifically, pixel intensities are shifted and scaled so that:

$$X_{\text{norm}} = \frac{X - \mu_X}{\sigma_X}$$

where $\mu_X$ and $\sigma_X$ are the mean and standard deviation of the entire training set, respectively. We further create a validation set by taking 10% of the training data:

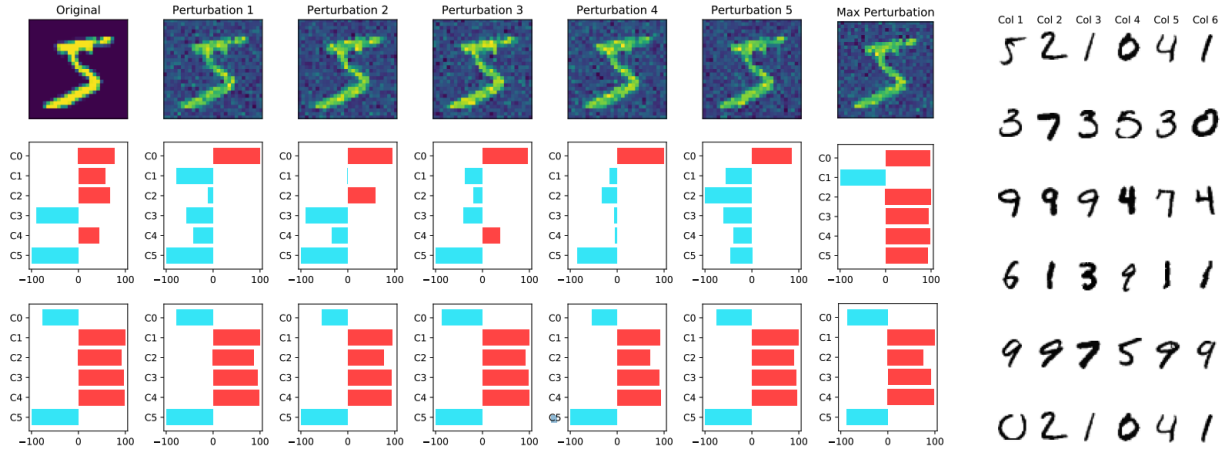


Figure 3: **Left:** The effect of gradient regularization on explanation stability. The unregularized version (second row) produces highly variable, sometimes contradictory explanations for slight perturbations of the same input. The regularized version ($\lambda = 2 \times 10^{-4}$) provides substantially more robust explanations. **Right:** Prototypes for the six learned concepts (rows).

- **Training Set:** 54,000 images
- **Validation Set:** 6,000 images
- **Testing Set:** 10,000 images

### 2.2 Training the SENN Model

We train a Self-Explaining Neural Network (SENN) on the normalized MNIST dataset with gradient regularization to ensure stable and robust concept-based explanations. The model outputs both a predicted class and a set of concept scores that correspond to interpretable features, such as the presence of specific strokes or orientations in the digit.

To maintain interpretability and prevent out-of-distribution concept activations, we rescale concept relevance to lie within $[-1, 1]$ and concept scores within $[-2, 2]$. These ranges were chosen empirically based on preliminary experiments indicating that narrower ranges (e.g., $[-1, 1]$ for concept scores) sometimes limited the expressiveness of concepts, while wider ranges (e.g., $[-5, 5]$) led to outliers and unstable values (for example, a 1 degree perturbation suddenly made activation scores go from units to hundreds range). The chosen bounds ensure that concept activations remain bounded and meaningful, even under perturbations, striking a balance between stability and flexibility in how concepts represent variations in the input.

### 2.3 3D Rotational Perturbations

To analyze interpretability and robustness, we apply 3D rotational perturbations to each digit in the testing set. We rotate the images along the $x$, $y$, and $z$ axes in a range of $-80°$ to $80°$ at increments of $2°$. We avoid going beyond $90°$ to prevent extreme transformations that would produce highly unrealistic digit shapes, ensuring that the perturbed images remain within a plausible range of appearances. Evaluating at $2°$ intervals strikes a balance between computational efficiency and the granularity needed to observe meaningful changes in concept activations and predictions.

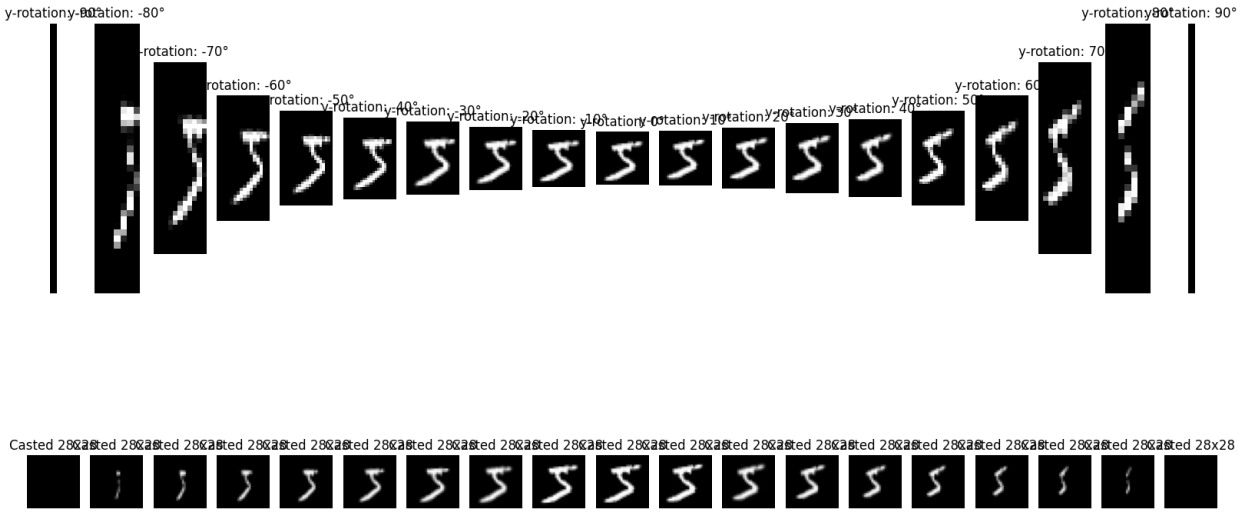


Figure 4: **3D Rotation Example:** A '5' rotated about the y-axis at various angles. Such rotations transform the pixel arrangement in interpretable ways that SENN's concepts can capture.

### 2.4 Rotation, Resize, and Recast Process

The rotational process involves:

1. **Image Rotation:** Each digit is rotated using a transformation matrix considering the specified rotation angles along the $x$, $y$, and $z$ axes.
2. **Resizing:** If the rotated image exceeds the $28 \times 28$ canvas, it is proportionally scaled down to fit within the canvas.
3. **Recasting onto Canvas:** The resized image is placed centrally onto a $28 \times 28$ canvas, maintaining consistency with the original dataset format and ensuring that the SENN model can process the perturbed images without additional modifications.

### 2.5 Experimental Setup and Evaluation

We evaluate SENN under these rotation-based perturbations as follows:

- Apply rotations from $-80°$ to $80°$ in increments of $2°$ along each axis to all test images.
- Record SENN's predictions and concept scores at each rotation angle.
- Measure how predictions and concept activations change as the input rotates, focusing on three main metrics:
  - **Prediction Stability:** How consistently the model predicts the same class over a range of rotation angles.
  - **Concept Activation Stability:** Whether concept activations remain interpretable and stable as the input is rotated.
  - **Decision Boundary Clarity:** How clearly concept activations reveal the moment a prediction switches between classes.

By focusing on these metrics and leveraging MNIST's simplicity and familiarity, we aim to understand how physical, interpretable transformations influence concept-level representations and decision boundaries within SENN.

# 3 Results

## 3.1 Single-Digit Rotation Analysis

We first examine individual digits rotated along different axes to understand how SENN's predictions and concept distributions shift under controlled, interpretable transformations.

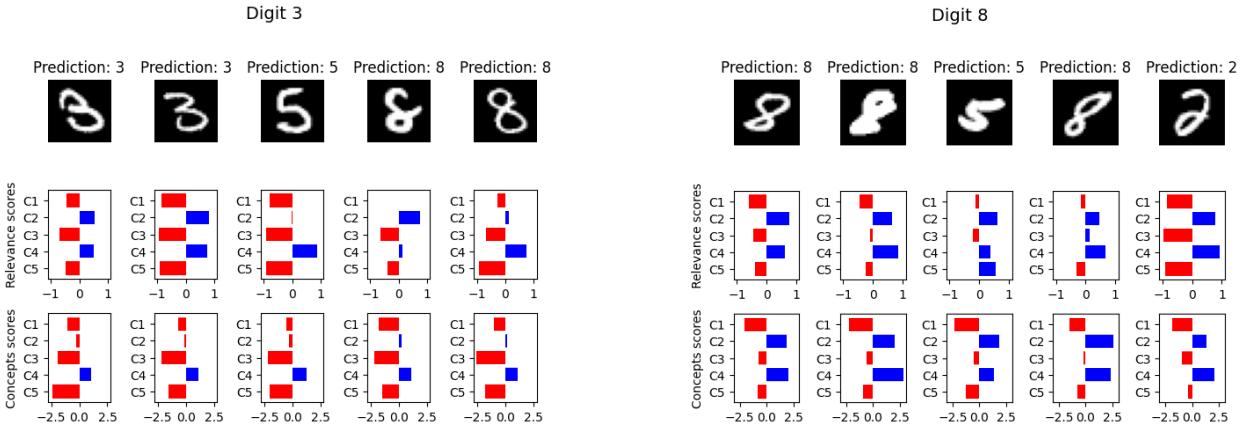


Figure 5: **Concept Distributions and Nearest Neighbors:** Using a k-nearest neighbors approach in concept space, we identify digits with similar concept distributions. **Left:** A digit predicted as '3' has nearest neighbors that include '5' and '8', indicating overlap in their learned concept features. For '3', we find that concept scores C3 and C5 are strongly negative, C2 is near zero, and C4 is somewhat positive. **Right:** A digit predicted as '8' has neighbors including '5' and '2', suggesting that '8' also shares conceptual similarities with these digits. For '8', C1 is strongly negative, C3 and C5 are weakly negative, and C2 and C4 are positive.

This overlap in concept distributions explains why certain digits are easily confused. Although a "3" may not resemble an "8" to a human, their concept embeddings share enough similarity that a slight rotation can shift a "3" toward an "8"-like configuration.

For a rotated "3," we observe that its concept scores gradually change in a predictable manner. Initially, the configuration matches a typical "3." As we rotate the digit about the y-axis, C2 becomes positive near 32 degrees, pushing the concept distribution toward that of an "8." This causes the model's prediction to switch from "3" to "8," providing a clear, measurable link between specific concept changes and decision boundary crossings.

## 3.2 Aggregated Results for Rotation Robustness

To assess overall robustness, we measure prediction accuracy across the test set as a function of rotation angle for $x$, $y$, and $z$ axes. As shown in Figure 7, the model maintains above 90% accuracy for rotations of up to approximately $\pm 25°$ about the $y$-axis. Beyond this range, accuracy declines, but the pattern of decline differs depending on the axis and the direction of the rotation:

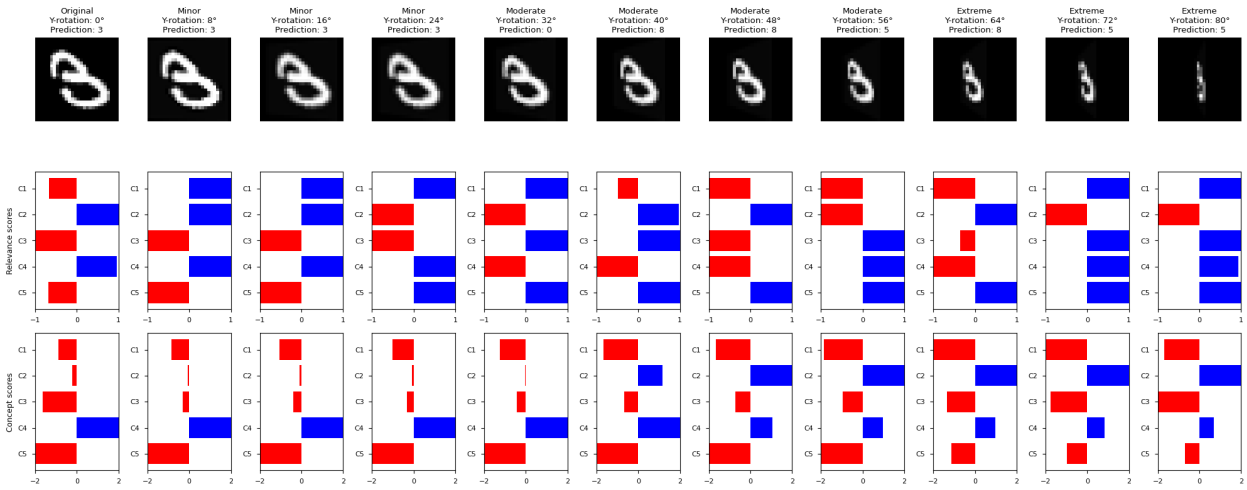


Figure 6: **Rotating a '3' about the y-axis:** As the '3' is rotated, its concept activation pattern evolves. Around 32 degrees of rotation, concept C2 begins to rise above zero, aligning the overall concept signature more closely with that of an '8'. This shift in C2's value correlates directly with the transition from predicting '3' to predicting '8', demonstrating that changes in concept scores trace the path across the decision boundary.

**X-axis Rotations:** Accuracy remains above 90% up to about $\pm 25°$. For positive rotations (tilting forward), accuracy steadily declines until around $+75°$, after which it stabilizes at a lower plateau. On the negative side (tilting backward), accuracy also decreases but reaches a stable lower accuracy around $-60°$. Thus, while both directions eventually plateau, they do so at different angles and performance levels.

**Y-axis Rotations:** The $y$-axis shows similar initial stability up to $\pm 25°$, but its subsequent behavior differs from the $x$-axis. On the positive side, accuracy decreases gradually and plateaus near $+75°$. In contrast, for negative rotations, accuracy continues to erode up to $-80°$ without reaching a clear plateau. This suggests that while the model can eventually settle into a stable accuracy regime at positive extremes, it struggles to find stability at the negative end within the tested range.

**Z-axis Rotations:** Rotations about the $z$-axis (in-plane rotations) are the most challenging. Even small deviations from zero cause a pronounced accuracy drop, reflecting a heightened sensitivity to changes in the digit's in-plane orientation. At very large positive angles (beyond $+70°$), the model's accuracy appears to level off at a consistently low value, indicating that, despite the severity of distortion, the model settles into a stable—albeit incorrect—interpretation.

# 4 Discussion

## 4.1 Connection Between Concepts and Predictions

Our rotation-based perturbations establish a direct relationship between input transformations, concept score evolution, and prediction changes. Unlike random pixel-level noise, rotations produce interpretable alterations that highlight how continuous movements in concept space lead to crossing decision boundaries. For instance, we observed how a "3" transitions into an "8" as C2 rises,

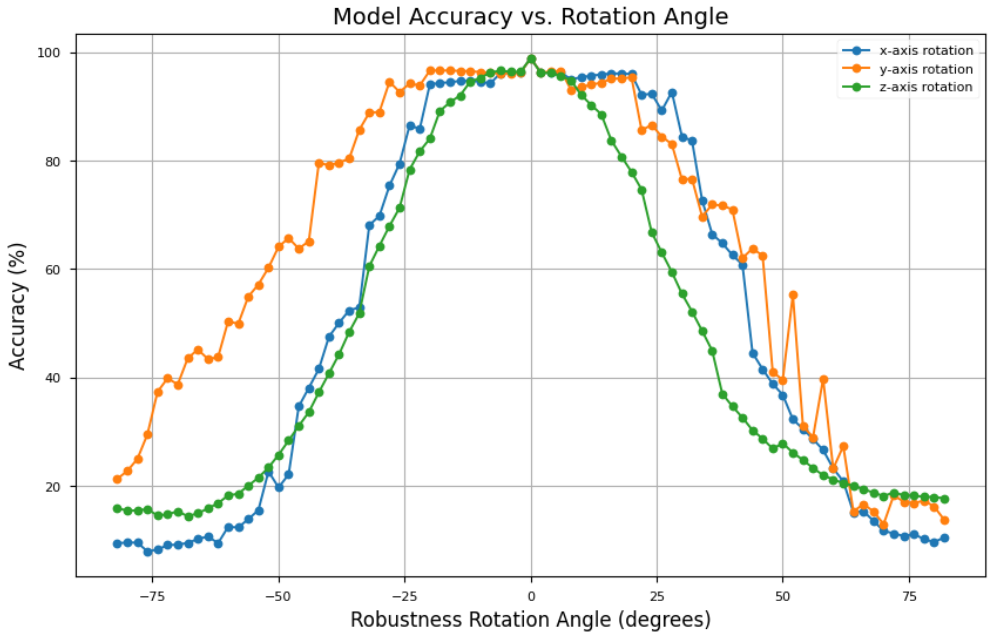


Figure 7: **Overall Prediction Accuracy as a Function of Rotation Angle:** The model exhibits initial robustness to small rotations, particularly along the $y$-axis. As the rotation angles increase, accuracy drops at varying rates depending on both the axis and direction of rotation.

illustrating that specific concept shifts map directly to changes in the predicted class. By observing these patterns, we gain insight into which concepts matter most at critical points and how they steer the model's decisions.

### 4.2 Interpreting Axis-Specific Robustness Patterns

SENN's rotational robustness varies by axis and direction. Moderate tilts around the $y$-axis are handled relatively well, whereas $x$-axis rotations stabilize at different extremes for positive and negative angles, revealing asymmetry in the learned representations. In contrast, even small $z$-axis (in-plane) rotations rapidly degrade accuracy, and while the model eventually reaches a stable interpretation at large angles, it is consistently incorrect. In sum, SENN's concept space is sensitive to both the axis and magnitude of rotation.

These detailed observations reveal that while SENN demonstrates initial resilience to small rotations, more extreme transformations produce a complex landscape of accuracy decay and stabilization. Each axis of rotation presents its own profile of vulnerability and eventual equilibrium, implying that SENN's concept space may handle certain geometric distortions with relative ease while struggling with others. The identification of stability regions at large angles—albeit at reduced accuracy—indicates that after enough distortion, the model settles into a stable but incorrect interpretation of the input. Together, these insights provide a more comprehensive understanding of rotational robustness and expose axes-specific patterns that can guide future improvements in concept regularization and model training.

### 4.3 Limitations

While concept scores offered valuable insights, relevance scores did not yield clear or consistent patterns. We expected relevance scores to highlight which concepts mattered most at various rotation angles, but no stable relationships emerged. Additionally, although we identified interpretable concept shifts for some single-digit rotations, other digits exhibited more erratic behavior. Attempts to quantify these fluctuations (e.g., measuring simple deltas in a given concept score like C2) proved inadequate, as they failed to capture when sudden, drastic changes occurred. Lastly, our focus on rotations alone leaves open questions about how other perturbations (e.g., shearing, scaling) might affect concept distributions and predictions.

### 4.4 Future Work

Several avenues can enhance our understanding of SENN's behavior:

- **Additional Perturbations:** Examining a broader range of transformations may reveal distinct patterns of concept stability and decision boundaries beyond rotations.
- **Deeper Class-Level Analysis:** Studying how each digit class responds to perturbations could uncover class-specific vulnerabilities and more nuanced concept configurations.
- **Improved Concept Metrics:** Developing more robust measures of concept change—beyond simple deltas—could better capture subtle or sudden shifts in activation patterns.
- **Clarifying Relevance Scores:** Identifying conditions under which relevance scores become meaningful may help integrate them into the interpretability toolkit.

By refining evaluation metrics, exploring a wider range of transformations, and investigating class-level responses, we aim to build more robust and interpretable SENNs and gain deeper insights into the interplay of concepts, geometry, and model decisions.

## References

[1] Alvarez-Melis, D., & Jaakkola, T. S. (2018). Towards robust interpretability with self-explaining neural networks. Advances in Neural Information Processing Systems (NeurIPS), 31, 7786-7795.

[2] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.