B-cos Explainable AI Analysis

# Iris Dataset Classification

Generated on: 2025-10-22 17:40:21

# Executive Summary

This report presents a comprehensive analysis of B-cos (B-cosine) networks for explainable AI on the Iris dataset. B-cos networks provide inherent interpretability through cosine similarity-based computations, making them ideal for applications where understanding model decisions is crucial.

## Key Performance Metrics

|  |  |  |
| --- | --- | --- |
| Metric | B-cos Model | Standard Model |
| Test Accuracy | 0.9333 | 0.9000 |
| Average Confidence | 0.9225 | 0.9567 |
| Confidence Std Dev | 0.1210 | 0.1033 |
| Average Sparsity | 9.20 | 0.00 |

# Dataset Information

The Iris dataset is a classic machine learning dataset containing 150 samples of iris flowers with 4 features (sepal length, sepal width, petal length, petal width) and 3 classes (setosa, versicolor, virginica). This dataset is ideal for demonstrating explainable AI techniques due to its clear feature meanings and biological interpretability.

# Model Architecture

Both B-cos and standard neural networks used identical architectures for fair comparison:

* • Input layer: 4 features
* • Hidden layer 1: 16 neurons
* • Hidden layer 2: 8 neurons
* • Output layer: 3 classes
* • Dropout: 0.1 for regularization
* • Total parameters: 243

# Performance Analysis

The B-cos model achieved a test accuracy of 0.9333 compared to 0.9000 for the standard model. This demonstrates that B-cos networks can maintain competitive performance while providing built-in interpretability.

## Class-wise Feature Importance

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Setosa | Versicolor | Virginica |
| sepal length (cm) | -1.9555 | 0.3694 | 1.4444 |
| sepal width (cm) | 1.1598 | -0.1726 | -0.9344 |
| petal length (cm) | 2.3325 | -0.5096 | -1.5103 |
| petal width (cm) | 2.2622 | -0.5051 | -1.4242 |

# Key Findings and Insights

## 1. PERFORMANCE COMPARISON:

* • Both models achieved similar accuracy (~93.3%)
* • B-cos model shows comparable performance to standard neural networks
* • Training convergence is similar for both approaches

## 2. INTERPRETABILITY ADVANTAGES:

* • B-cos networks provide built-in explainability through cosine similarity
* • Feature contributions are directly interpretable without post-hoc methods
* • Class-wise feature importance reveals meaningful patterns
* • Decision confidence analysis shows model reliability

## 3. TECHNICAL INSIGHTS:

* • B-cos layers normalize weights to unit vectors, enabling cosine similarity computation
* • Feature contributions can be extracted at any layer for multi-level explanations
* • The approach maintains computational efficiency similar to standard networks
* • Cosine similarity provides intuitive geometric interpretation

# Conclusion

This analysis demonstrates that B-cos networks successfully combine high performance with inherent interpretability on the Iris dataset. The B-cos model achieved 93.33% accuracy compared to 90.00% for the standard model, while providing meaningful insights into feature contributions and class-wise importance patterns.

The built-in explainability of B-cos networks makes them particularly valuable for applications where understanding model decisions is crucial, such as medical diagnosis, financial risk assessment, and legal decision support systems.