# B-cos Explainable AI Analysis

# Iris Dataset Classification

A Comprehensive Analysis of B-cosine Networks for Interpretable Machine Learning

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.

## *Data Split*

• Training set: 90 samples (60%)

• Validation set: 30 samples (20%)

• Test set: 30 samples (20%)

• Features were standardized using StandardScaler

# 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

## *B-cos Implementation*

The B-cos model uses custom linear layers that normalize weights to unit vectors and compute cosine similarity between inputs and weights. This provides inherent interpretability through geometric relationships in the feature space.

# Training Results

Both models were trained using Adam optimizer with learning rate scheduling and early stopping. The B-cos model achieved comparable performance to the standard neural network, demonstrating that interpretability can be achieved without sacrificing accuracy.

## *Final Training Metrics*

• B-cos Model: 96.67% training accuracy, 93.33% validation accuracy

• Standard Model: 97.78% training accuracy, 93.33% validation accuracy

• Both models converged within 100 epochs with early stopping

# 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.

## *Detailed Classification Results*

B-cos Model Classification Report:

• Setosa: 100% precision, 100% recall, 100% F1-score

• Versicolor: 90% precision, 90% recall, 90% F1-score

• Virginica: 90% precision, 90% recall, 90% F1-score

• Overall: 93.33% accuracy

Standard Model Classification Report:

• Setosa: 100% precision, 100% recall, 100% F1-score

• Versicolor: 89% precision, 80% recall, 84% F1-score

• Virginica: 82% precision, 90% recall, 86% F1-score

• Overall: 90.00% accuracy

# Explainability Analysis

The key advantage of B-cos networks is their inherent interpretability. Unlike standard neural networks that require post-hoc explanation methods, B-cos networks provide direct insights into feature contributions through cosine similarity computations.

## *Class-wise Feature Importance*

Analysis of feature contributions reveals meaningful biological patterns:

| 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 Insights:

• Setosa: Petal length and width are most important (positive contributions)

• Versicolor: Moderate importance across all features

• Virginica: Sepal length is most important, petal features are negative

## Interpretability Metrics

Quantitative analysis of interpretability reveals the advantages of B-cos networks:

| Metric | B-cos Model | Standard Model | Interpretation |
|---|---|---|---|
| Average Confidence | 0.9225 | 0.9567 | Both models show high confidence |
| Confidence Std Dev | 0.1210 | 0.1033 | B-cos shows more variation |
| Average Sparsity | 9.20 | 0.00 | B-cos uses ~9 important features |

## 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

### *4. WHEN TO USE B-COS NETWORKS:*

• When interpretability is crucial (medical, financial, legal applications)

• When you need to understand feature importance

• When stakeholders require model explanations

• When working with tabular data where features have clear meaning

• When you want built-in explainability without additional complexity

### *5. LIMITATIONS AND CONSIDERATIONS:*

• May require more careful hyperparameter tuning

• Cosine similarity assumption might not suit all data types

• Limited to linear transformations in each layer

• May need domain-specific adaptations for complex data

## Future Work and Recommendations

Based on this analysis, several directions for future research and practical applications emerge:

### *Research Directions:*

• Extend to more complex architectures (CNNs, RNNs)

• Apply to larger, more complex datasets

• Investigate hybrid approaches combining B-cos with standard layers

• Develop specialized B-cos variants for different data modalities

### *Practical Recommendations:*

• Use B-cos networks when explainability is a primary requirement

• Combine with standard networks for hybrid interpretable systems

• Validate explanations with domain experts

• Consider computational overhead vs. interpretability trade-offs

## 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.

Future work should focus on extending these techniques to more complex datasets and architectures while maintaining the interpretability advantages demonstrated in this study.