

Association for Information Systems

## AIS Electronic Library (AISeL)

---

NEAIS 2024 Proceedings

New England Chapter of Association for  
Information Systems

---

10-23-2024

# Developing Reliable Gradient Explanations for Artificial Intelligence: Addressing Consistency in Local Interpretability

Nolan M. Talaei

*University of Massachusetts Lowell, nolan\_talaei@uml.edu*

Asil Oztekin

*University of Massachusetts Lowell, asil\_oztekin@uml.edu*

Hongwei Zhu

*University of Massachusetts Lowell, hongwei\_zhu@uml.edu*

Luvai Motiwalla

*University of Massachusetts Lowell, luvai\_motiwalla@uml.edu*

Follow this and additional works at: <https://aisel.aisnet.org/neais2024>

---

### Recommended Citation

Talaei, Nolan M.; Oztekin, Asil; Zhu, Hongwei; and Motiwalla, Luvai, "Developing Reliable Gradient Explanations for Artificial Intelligence: Addressing Consistency in Local Interpretability" (2024). *NEAIS 2024 Proceedings*. 14.

<https://aisel.aisnet.org/neais2024/14>

This material is brought to you by the New England Chapter of Association for Information Systems at AIS Electronic Library (AISeL). It has been accepted for inclusion in NEAIS 2024 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact [elibrary@aisnet.org](mailto:elibrary@aisnet.org).

# Developing Reliable Gradient Explanations for Artificial Intelligence: Addressing Consistency in Local Interpretability

*Research-in-Progress*

Nolan M. Talaei  
University of Massachusetts Lowell  
nolan\_talaei@uml.edu

Asil Oztekin  
University of Massachusetts Lowell  
asil\_oztekin@uml.edu

Hongwei Zhu  
University of Massachusetts Lowell  
hongwei\_zhu@uml.edu

Luvai Motiwalla  
University of Massachusetts Lowell  
luvai\_motiwalla@uml.edu

## **ABSTRACT (REQUIRED)**

Interpreting machine learning models remains challenging, particularly in high-stakes applications where trust and transparency are vital. We introduce Reliable Gradient Explanations (RGE), a method designed to enhance the stability and consistency of gradient-based feature importance explanations. RGE combines first-order gradient information with second-order Hessian elements to refine feature importance based on output curvature, reducing instability in traditional methods. Preliminary results indicate that RGE improves explanation accuracy and stability across different model architectures. Ongoing research aims to refine RGE, evaluate its performance on diverse datasets, and compare it with established interpretability techniques, ultimately promoting more transparent and reliable AI-driven decisions.

## **Keywords (Required)**

Responsible Explainable AI (XAI), Stable Gradient-Based Interpretability, Feature Importance