Interpreting Neural Receivers: Mechanistic and Data Attribution Approaches

Marko Tuononen*
University of Eastern Finland, P.O. Box 111, 80101 Joensuu, Finland
Corresponding author: marko.tuononen@gmail.com

Abstract

This research studies interpretability methods for deep neural network–based radio receivers in future wireless communication systems. The work focuses on representation-level analysis and data attribution techniques that reveal how models encode physical-layer channel conditions and how training data influence receiver behavior. These insights enable validation, adaptation, and more trustworthy deployment of data-driven communication systems under realistic channel conditions, while providing deeper understanding of model behavior.

1 Introduction

Machine learning is expected to play an increasingly central role in future wireless communication systems. In particular, deep neural network–based radio receivers have been proposed as alternatives or complements to conventional signal-processing pipelines in the physical layer of next-generation networks (e.g., 6G [Farhadi *et al.*, 2025]). These models learn complex nonlinear signal transformations directly from data to recover transmitted information from noisy wireless signals, and have demonstrated strong performance in challenging propagation environments [Honkala *et al.*, 2021] and alternative receiver paradigms such as pilotless communication [Korpi *et al.*, 2023].

Despite these advantages, their adoption raises challenges related to interpretability. Deep neural networks often operate as black-box models whose internal decision processes are difficult to analyze. While interpretable machine learning has advanced significantly [Kim and Doshi-Velez, 2023], most methods have been developed for domains such as vision and language. Physical-layer machine learning differs substantially: inputs are non-semantic signals, outputs are high-dimensional, and deployment constraints are strict.

These challenges motivate interpretability approaches tailored to neural receivers. This research investigates how representation-level analysis and data attribution can reveal model behavior, relate learned representations to underlying physical channel properties, and support practical engineering tasks such as monitoring and adaptation.

*The author gratefully acknowledges the support of the Nokia Foundation and Nokia Corporation.

2 Research Problem

Interpretability is inherently context-dependent [Kim and Doshi-Velez, 2023], and methods effective in one domain may fail in another. Neural receivers present several distinctive challenges. First, inputs consist of complex-valued signals without direct semantic meaning, limiting the usefulness of standard input attribution methods. Second, models produce large vectors of outputs rather than single class predictions, making class-wise explanations impractical. Third, strict deployment constraints limit methods that modify model architectures or add computational overhead.

This research addresses these challenges by focusing on post-hoc, model-specific interpretability methods that analyze internal representations and training data influence. The work considers neural receiver architectures derived from DeepRx [Honkala *et al.*, 2021], a setting in which interpretability has received limited prior attention. DeepRx is a convolutional neural receiver for orthogonal frequency division multiplexing systems used in modern communication standards such as 5G NR and emerging 6G. The model takes received noisy signal samples as input and outputs estimates of the transmitted bits along with their associated confidence.

3 Research Contributions

The research studies interpretability of neural receivers through four complementary directions: mechanistic interpretability, training data attribution, out-of-distribution detection, and targeted model adaptation.

Mechanistic Interpretability. This work investigates how internal representations of neural receivers encode physically meaningful channel parameters. A post-hoc explainer model is trained in a supervised manner to predict channel characteristics from internal representations of a trained receiver [Tuononen *et al.*, 2025]. The predictive performance of the explainer is used as a proxy for the information content of individual layers and convolutional channels. Experiments show that signal-to-noise-ratio (SNR) information is not uniformly distributed across the network; instead, informativeness varies significantly across convolutional channels. These results demonstrate that neural receivers learn structured internal representations aligned with physical channel properties, and suggest opportunities for architecture analysis, pruning, and model simplification.

A related study analyzes the geometry of internal representations under varying channel conditions [Tuononen *et al.*, 2026a]. The results show that activations lie on a smooth low-dimensional manifold strongly aligned with SNR, providing insight into how neural receivers organize signal information and highlighting the role of continuous structure in representation learning, as opposed to discrete semantic classes.

Out-of-Distribution Detection. Building on these mechanistic insights, this work investigates how internal representations can be used to detect distribution shifts. A layerwise out-of-distribution detection framework is developed using channelwise aggregation of internal representations [Tuononen *et al.*, 2026a]. In-distribution behavior is modeled using statistics computed from training data, and deviations from this structure are used to detect anomalous inputs. Experimental results show that representation-based detectors can reliably identify certain types of distribution shifts, such as extreme multipath conditions, while other shifts such as high mobility remain more challenging. The analysis further shows that earlier network layers often provide stronger signals for distribution shift detection than deeper layers. These findings demonstrate how representation-level interpretability can inform the design of monitoring mechanisms for neural receivers.

Data Attribution and Model Adaptation. A complementary perspective is obtained through data attribution methods that analyze how individual training samples influence model behavior [Tuononen *et al.*, 2026b]. In this work, influence functions [Koh and Liang, 2017] are used to operationalize this perspective by estimating how removing specific training samples affects model predictions. Experiments show that influence-guided selection of fine-tuning data can accelerate adaptation to new channel conditions compared to random selection. Performance is evaluated as a reduction in the BER gap between the neural receiver and a genie-aided linear minimum mean square error (LMMSE) benchmark. These results demonstrate that data attribution can provide actionable guidance for efficient and targeted model adaptation.

Methodological Contributions. This work addresses numerical instability in k-nearest neighbor–based normalized mutual information estimation in high-dimensional settings by proposing a logarithmic reformulation [Tuononen and Hautamäki, 2025]. The approach improves numerical stability and enables more reliable information-theoretic analysis of internal representations, supporting mechanistic interpretability by quantifying dependence between representations and channel parameters.

4 Evaluation Methodology

The proposed methods are evaluated using DeepRx-based neural receivers trained on simulated wireless channel data and tested under both in-distribution and out-of-distribution conditions. Performance is measured using communication metrics such as bit-error rate (BER), alongside interpretability metrics derived from representation and attribution analysis. Ongoing work extends evaluation to measurement-based datasets from real wireless environments, focusing on deployment-oriented scenarios and site-specific adaptation.

5 Impact and Future Work

The increasing integration of machine learning into communication infrastructure raises important questions of reliability, transparency, and accountability in safety- and performance-critical systems. Interpretability methods address these challenges by enabling understanding of model behavior, detection and diagnosis of failures, and adaptation under changing conditions. By improving transparency and supporting informed decision-making, such methods contribute to the responsible deployment of neural receivers and to a more equitable and trustworthy use of this technology.

Future work will investigate the generality of learned representation structures in neural receivers, particularly their role in detecting and characterizing distribution shifts, and develop improved data attribution methods for model adaptation. This includes advancing information-theoretic analysis of neural representations and exploring underlying computational mechanisms. Further work will examine how interpretability can be integrated into system-level validation, monitoring, and adaptation workflows for future wireless communication systems.

More broadly, interpretability may serve not only as a diagnostic tool but also as a mechanism for discovery, providing insights that can inform the design of future communication systems and inspire new signal processing approaches.

References

[Farhadi *et al.*, 2025] Hamed Farhadi et al. 6G AI-driven Air Interface. *IEEE Commun. Mag.*, 63(10):118–125, 2025.

[Honkala *et al.*, 2021] Mikko Honkala, Dani Korpi, and Janne M. J. Huttunen. DeepRx: Fully Convolutional Deep Learning Receiver. *IEEE TWC*, 20(6):3925–3940, 2021.

[Kim and Doshi-Velez, 2023] Been Kim and Finale Doshi-Velez. Interpretability. In Kevin P. Murphy, editor, *Probabilistic Machine Learning: Advanced Topics*, chapter 33, pages 1073–1102. MIT Press, 2023.

[Koh and Liang, 2017] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *ICML*, 2017.

[Korpi *et al.*, 2023] Dani Korpi, Mikko Honkala, and Janne M.J. Huttunen. Deep Learning-Based Pilotless Spatial Multiplexing. In *ACSSC*, pages 1025–1029, 2023.

[Tuononen and Hautamäki, 2025] Marko Tuononen and Ville Hautamäki. Improving Numerical Stability of Normalized Mutual Information Estimator on High Dimensions. *IEEE SPL*, 32:2783–2787, 2025.

[Tuononen *et al.*, 2025] Marko Tuononen et al. Interpreting Deep Neural Network–Based Receiver Under Varying Signal-To-Noise Ratios. In *IEEE ICASSP*, 2025.

[Tuononen *et al.*, 2026a] Marko Tuononen et al. Out-of-Distribution Detection via Channelwise Feature Aggregation in Neural Network–Based Receivers. Under review; submitted to IJCAI, 2026.

[Tuononen *et al.*, 2026b] Marko Tuononen et al. Targeted Fine-Tuning of DNN-Based Receivers via Influence Functions. In *IEEE ICASSP*, 2026.