

Learning Dynamics and Interpretability in Quantum-Inspired Tensor Network Classifiers Across Biomedical, Neuroimaging, and Remote Sensing Applications

Quantum-inspired machine learning based on tensor networks has emerged as a powerful framework for addressing high-dimensional classification problems while providing physically interpretable representations of learned data structures. In particular, Matrix Product States (MPS), originally developed as variational ansätze for quantum many-body systems, offer a scalable approach to model complex correlations and higher-order interactions that are often inaccessible to conventional machine learning methods. Beyond predictive performance, the quantum formalism naturally provides a rich set of observables, including entanglement measures, local magnetizations, correlation functions, and reduced density matrices, that can be exploited to investigate both learning dynamics and model interpretability.

In this talk, we present a unified analysis of MPS-based classifiers across multiple application domains, including biomedical omics, structural neuroimaging, hyperspectral remote sensing, and wildfire susceptibility mapping. Using binary and multiclass classification problems of increasing complexity, we investigate how quantum-inspired representations capture relevant feature interactions and how their internal quantum-information structure evolves during training.

A central focus of our study is the phenomenon of grokking, whereby a model initially memorizes training data and only after prolonged optimization exhibits an abrupt improvement in generalization performance. Through extensive numerical experiments, we demonstrate that grokking can be interpreted as a dynamical transition in the underlying tensor-network representation. In binary classification tasks the onset of generalization is consistently accompanied by sharp changes in entanglement entropy, indicating a reorganization of the internal correlation structure of the learned quantum state. Simultaneously, observables such as local qubit magnetization and connected correlation functions reveal the emergence of ordered patterns associated with relevant predictive features. In contrast, tasks characterized by persistent overfitting, such as certain hyperspectral satellite-image experiments, maintain highly extended entanglement structures and fail to exhibit the characteristic transition associated with successful generalization.

To further characterize these learning dynamics, we complement standard performance measures with information-theoretic quantities. Transfer entropy is employed to identify directional dependencies among class-specific quantum masks, while O-information is used to distinguish synergistic from redundant higher-order interactions among class outputs. We find that grokking transitions are associated with the emergence of redundancy-dominated information structures, whereas overfitted regimes remain characterized by persistent synergistic interactions. These observations suggest that higher-order information dynamics provide valuable insights into the mechanisms governing generalization in tensor-network learning architectures, particularly in multiclass settings.

Beyond learning dynamics, we exploit the tensor-network representation as a source of interpretable feature interactions. In neuroimaging applications, trained MPS models are mapped into weighted graphs whose edges are defined through quantum connected correlations between gray-matter features extracted from structural MRI data. The resulting network representation enables simultaneous analysis of global spectral properties and local topological importance.

Applied to schizophrenia-versus-control and bipolar-versus-control classification tasks, the framework identifies stable neuroanatomical hubs, including Heschl's gyrus, insular cortex, and frontal regions, that consistently emerge across multiple train-test resamplings and centrality measures. Compared with Shapley-value-based explanations, graph centralities derived from quantum correlations exhibit lower variability, highlighting their robustness as interpretable neurobiological signatures. Furthermore, the feature network associated with bipolar disorder appears largely embedded within that of schizophrenia, providing a network-based characterization of the hierarchical organization of neuroanatomical alterations reported in the psychiatric literature.

A similar interpretability strategy is adopted for hepatocellular carcinoma gene-expression data. Here, local magnetization and correlation measurements extracted from the trained MPS identify biologically meaningful gene communities associated with disease discrimination. Subsequent enrichment analyses confirm the relevance of the detected subcommunities, demonstrating that tensor-network observables can be leveraged to reveal functionally coherent molecular signatures alongside predictive modeling.

Finally, we investigate classification of wildfire susceptibility in the Puglia region using AlphaEarth geospatial embeddings combined with MPS classifiers. The framework successfully addresses both binary and multiclass susceptibility prediction while maintaining interpretability through analysis of the underlying quantum state. By introducing level-resolved mixedness diagnostics based on reduced density matrices, we show that the classifier naturally encodes a hierarchy of class distinguishability. In particular, non-adjacent susceptibility classes become more separable than neighboring categories, revealing geometric relationships in the learned representation that are not directly observable from classification metrics alone. Moreover, the binary wildfire task exhibits a pronounced grokking transition, further supporting the connection between generalization and entanglement-driven reorganization of internal representations.

Taken together, these results establish quantum-inspired tensor networks as a unified framework that combines competitive classification performance with unprecedented access to the internal structure of the learning process. Across diverse scientific domains, entanglement measures, information-theoretic observables, correlation networks, and reduced-density-matrix diagnostics reveal common principles governing representation learning, interpretability, and generalization. Our findings suggest that the study of learning dynamics through the lens of quantum information may provide a principled route toward understanding and explaining complex machine learning systems in both binary and multiclass settings.

References

- [1] D. Pomarico et al. Grokking as an entanglement transition in tensor network machine learning. arXiv:2503.10483, 2025 (submitted to Quantum Machine Intelligence)
- [2] D. Pomarico et al. Transfer entropy and O-information to detect grokking in tensor network multi-class classification problems. *Technologies*, 13(10):438, 2025.
- [3] D. Pomarico et al. A tensor network framework for interpretable graph analysis of brain networks. Submitted to arXiv and PRX Intelligence, 2026.
- [4] D. Pomarico et al. Tensor Network Machine Learning for Wildfire Susceptibility Mapping: from Grokking Dynamics to Quantum Mixedness of Class Representations. In submission to arXiv and Scientific Reports, 2026.