Bayesian Neural Networks and Conformal Prediction for Out-of-Distribution Coverage in Multimodal Tabular Classification

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does combining Bayesian neural networks with conformal prediction affect out-of-distribution coverage in multimodal tabular classification compared to deep ensembles. 16 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: On the Out-of-Distribution Coverage of Combining Split Conformal Prediction and Bayesian Deep Learning. Research question: How does combining Bayesian neural networks with conformal prediction affect out-of-distribution coverage in multimodal tabular classification compared to deep ensembles?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.1/10.

3 Results

13 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 4.1/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Neural networks trained via stochastic gradient descent have been found to often be uncalibrated by being overly confide	×	0.11
Deep ensembles can be viewed as an approximation to the Bayesian model average (Wilson and Izmailov, 2020).	×	0.07
Bayesian neural networks for classification have been shown to be generally underconfident by overestimating aleatoric u	×	0.09
The credible set coverage on the calibration dataset for SGD at 0.05 error is 88%.	×	0.06
The credible set coverage on the calibration dataset for ENS at 0.05 error is 92%.	×	0.06
The credible set coverage on the calibration dataset for MFV at 0.05 error is 95%.	×	0.06
The credible set coverage on the calibration dataset for SGHMC at 0.05 error is 97%.	×	0.06
The credible set coverage on the calibration dataset for LAPLACE at 0.05 error is 99%.	×	0.06
The average set size on the calibration dataset for SGD at 0.05 error is 1.38.	×	0.02
The average set size on the calibration dataset for ENS at 0.05 error is 2.12.	×	0.02
The average set size on the calibration dataset for MFV at 0.05 error is 2.19.	×	0.02
The average set size on the calibration dataset for SGHMC at 0.05 error is 1.76.	×	0.02
The average set size on the calibration dataset for LAPLACE at 0.05 error is 1.72.	×	0.02
The observation model relating weights w and inputs x to labels y, p(y|x, w), is well-specified.	×	0.02
The prior over weights p(w) is well-specified.	×	0.02
Convolutional neural networks and zero-mean Gaussians have been shown to produce good inductive biases for image classif	×	0.05

References

- http://arxiv.org/abs/2107.07511v6

- http://arxiv.org/abs/2311.12688v2
- http://arxiv.org/abs/2512.17153v2