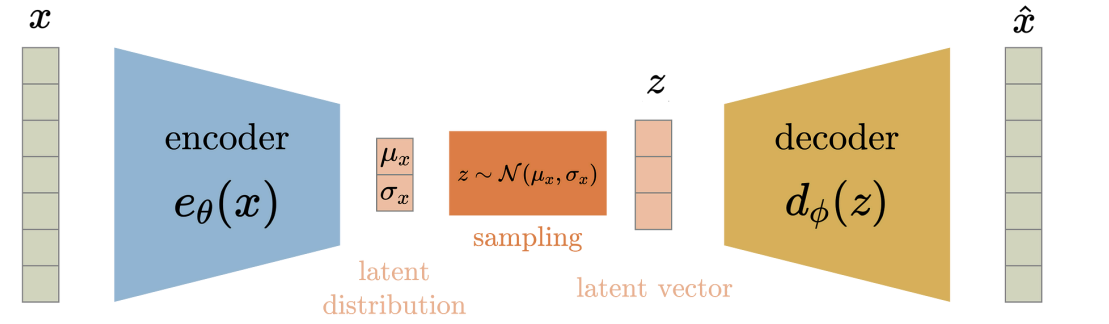
Report on Evidential Learning for Anomaly Detection

Introduction

In this report we go over the tests that have been done on the application of evidential learning to anomaly detection by leveraging Dirichlet Autoencoders (DAE). Traditional autoencoders (AE) and Variational Autoencoders (VAE) have been used in anomaly detection tasks. However they are unable to generate uncertainty estimates that are important for physics applications. In this report we go over the various models that were tested and how uncertainty quantification added to the model's predictions. We also introduce the Dirichlet Autoencoder (DAE) as an extension of the VAE that replaces the standard Gaussian prior with a Dirichlet distribution to better capture categorical data and improve explainability.

Variational Autoencoder (VAE)

The VAE is a generative model that incorporates a probabilistic latent space. The objective function of a VAE has two main components: reconstruction loss and Kullback-Leibler (KL) divergence loss. The goal of this loss is to make sure that the output is as close as possible to the input while the KL divergence loss regularizes the latent space by forcing a Gaussian prior. As a part of this work we benchmark the AE and VAE model on different aspects, some of them being latent space dimensions, layer depths, and activation functions. The figure below shows the general architecture of a VAE with the corresponding loss function.



$$\text{reconstruction loss} = \|x - \hat{x}\|_2 = \|x - d_\phi(z)\|_2 = \|x - d_\phi(\mu_x + \sigma_x \epsilon)\|_2$$

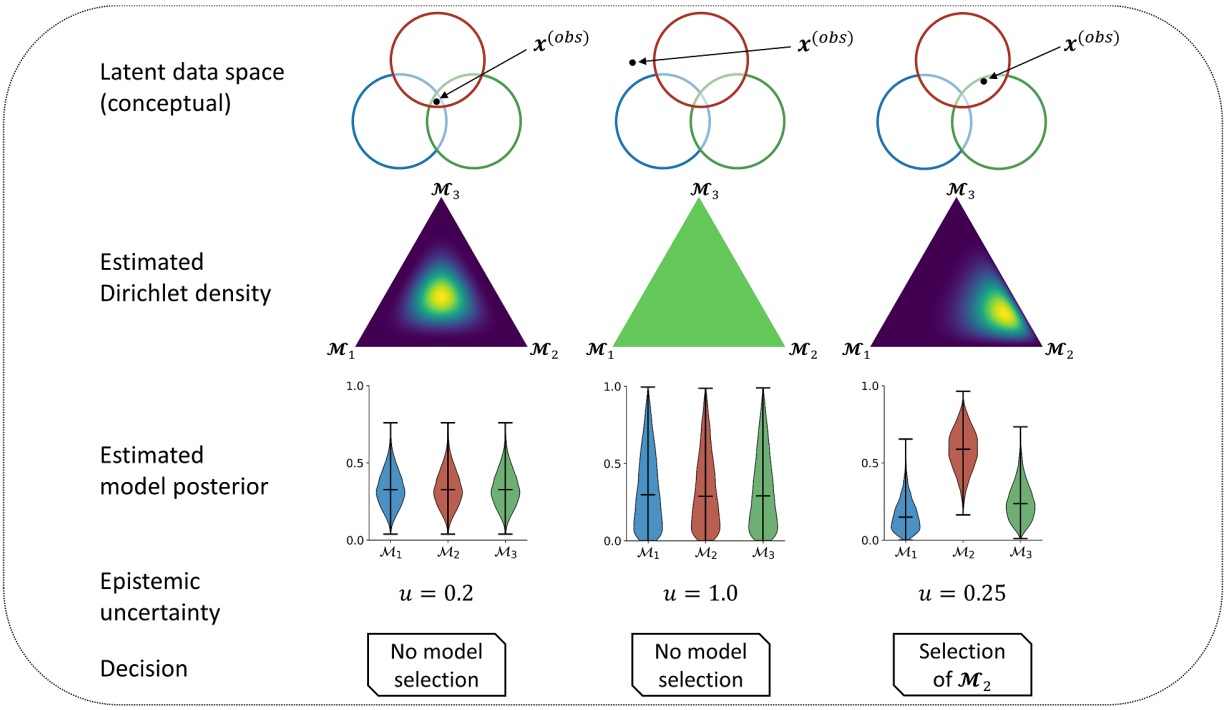
$$\mu_x, \sigma_x = e_\theta(x), \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$similarity\ loss = KL\ Divergence = D_{KL}(\mathcal{N}(\mu_x, \sigma_x) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I}))$$

$$loss = reconstruction\ loss + similarity\ loss$$

Methodology for Evidential Learning

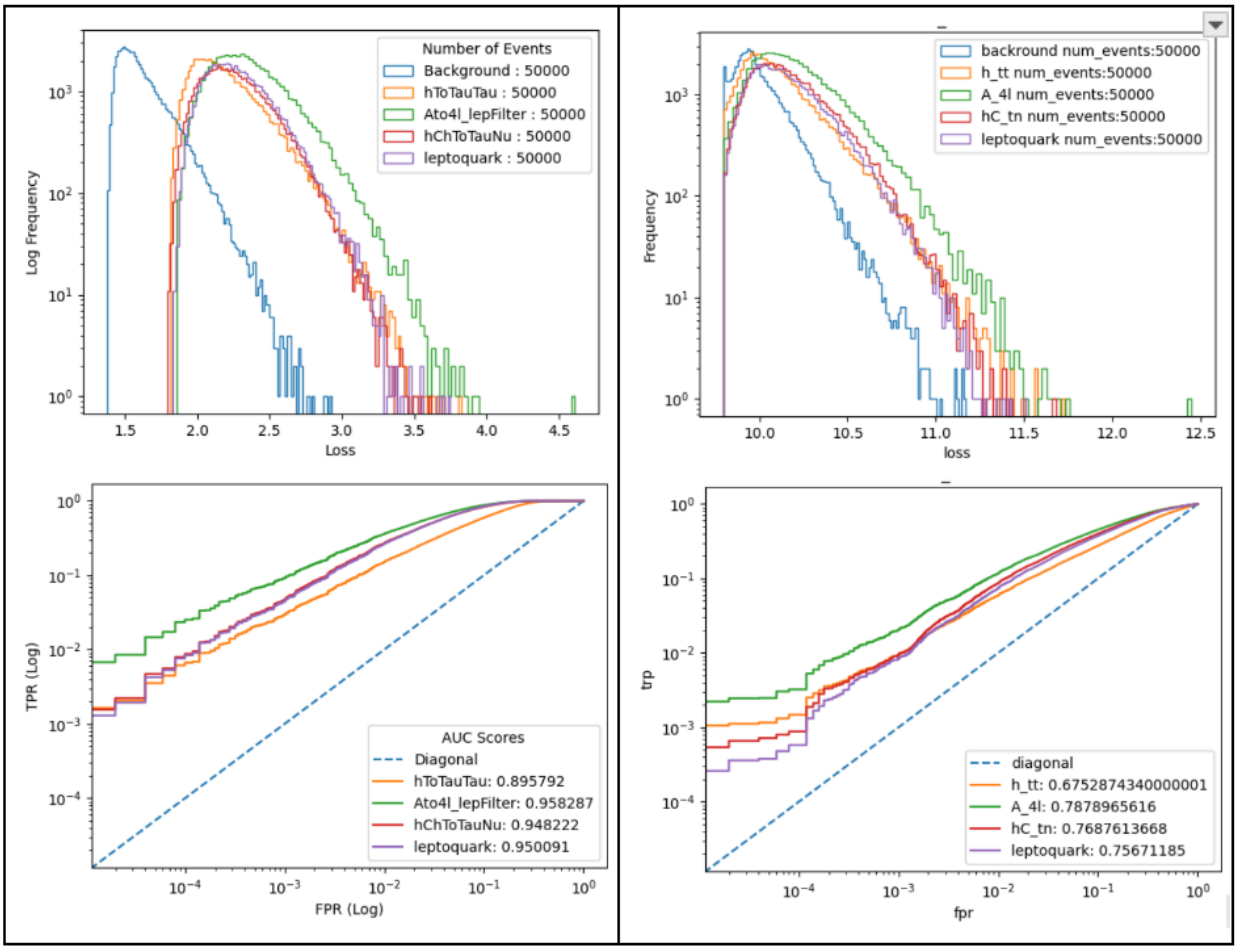
In Evidential Deep Learning (EDL) we add an additional layer of reasoning by replacing the Gaussian prior in VAEs with a Dirichlet distribution. This allows the model to estimate the probability of an event belonging to a given class while also quantifying the uncertainty associated with that prediction. The class for our use case is the background class. The image below shows how the evidential parameters help generate the uncertainty estimations.



The method for this experiment is as follows:

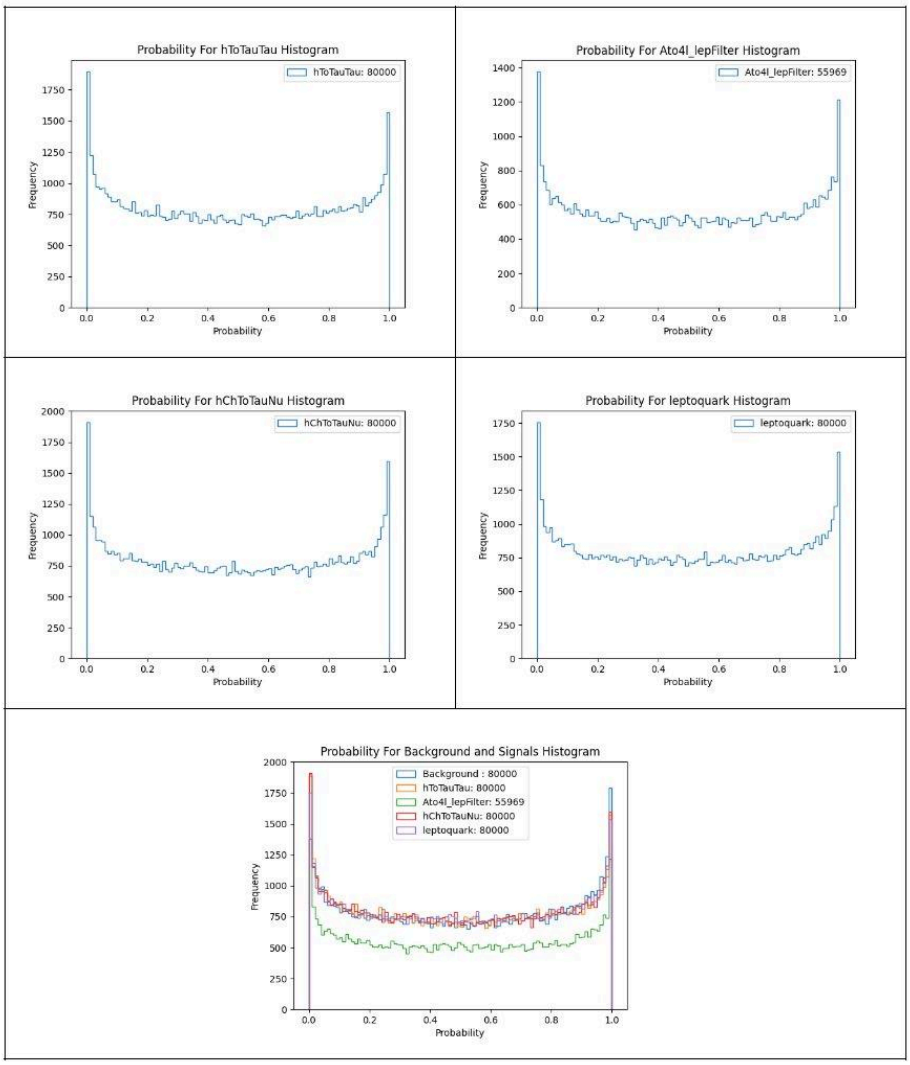
- Implementing a Dirichlet Variational Autoencoder (DVAE) where the latent space follows a Dirichlet prior.
- Training the DVAE on an anomaly detection dataset, with classes representing background and non background events.
- Benchmarking the DVAE against traditional AE and VAE models to evaluate performance in terms of reconstruction loss, uncertainty estimation, and classification accuracy.
- Testing different thresholding techniques to improve anomaly classification.

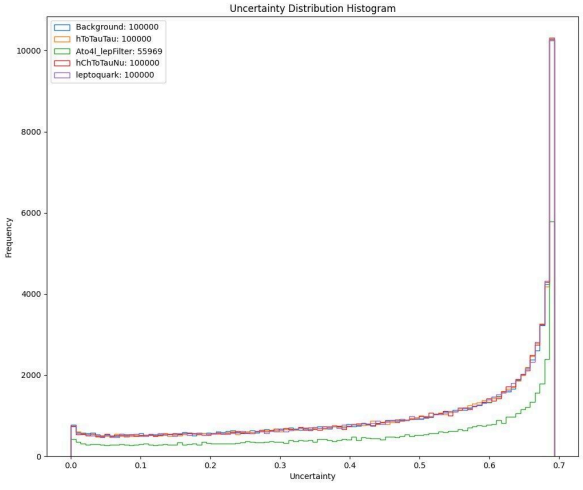
Model Results



The DVAE initially showed higher loss values compared to the VAE but improved significantly with tuning. With a latent space of 2 dimensions and a KL strength parameter of 0.01, the loss was significantly reduced. The DVAE does not perform as well as the VAE model and requires other strategies for improving the model.

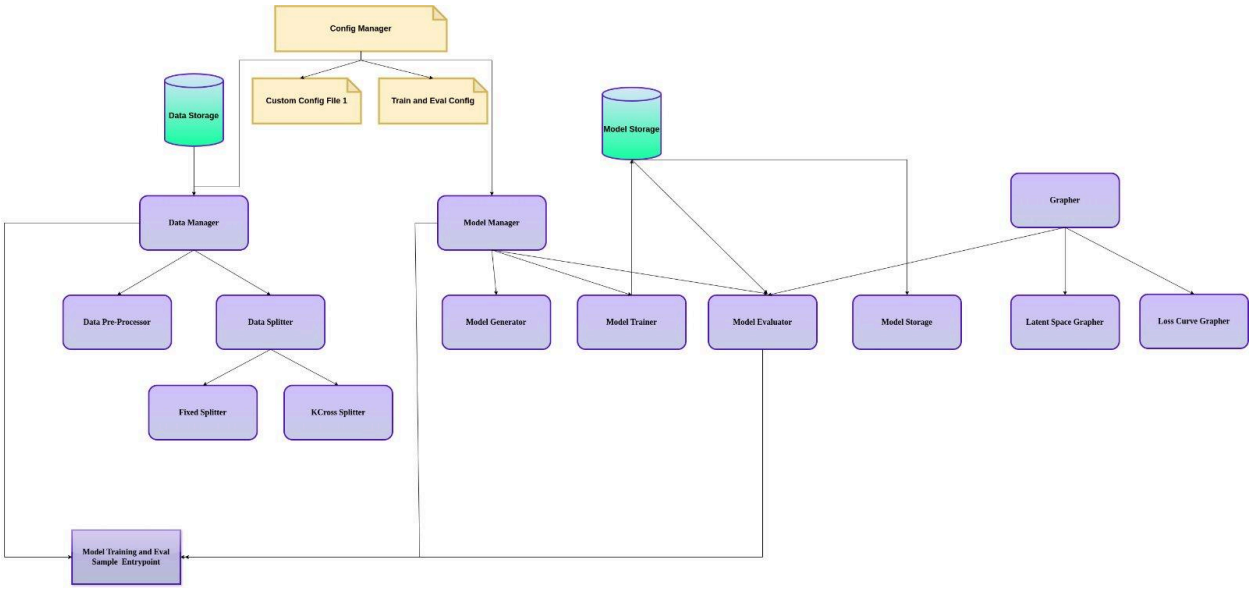
Note: As a part of this report we are only including results for the DVAE model. Additional benchmarking results are included in the overleaf document where we go into detail for every benchmarking step.


Probability For hToTauTau Histogram
hToTauTau: 80000
Frequency
Probability
Probability For Ato4l_lepFilter Histogram
Ato4l_lepFilter: 55969
Frequency
Probability
Probability For hChToTauNu Histogram
hChToTauNu: 80000
Frequency
Probability
Probability For leptoquark Histogram
leptoquark: 80000
Frequency
Probability
Probability For Background and Signals Histogram
Background : 80000
hToTauTau: 80000
Ato4l_lepFilter: 55969
hChToTauNu: 80000
leptoquark: 80000
Frequency
Probability


Uncertainty Distribution Histogram
Background: 100000
hToTauTau: 100000
Ato4l_lepFilter: 55969
hChToTauNu: 100000
leptoquark: 100000
Frequency
Uncertainty

The DVAE showed slightly better uncertainty estimates but not as expected. This could be attributed to low training data that was used to train the overall model due to compute limitations. By fine-tuning the thresholding mechanism we were able to slightly improve classification accuracy but this does not offset the performance due to the low training data. Note: This is an abridged version of the detailed analysis.

System Design



Originally all the analysis and model building were done in Jupyter notebooks which makes it a bit difficult to iterate over quickly. In order to work with multiple configurations quickly we developed a framework which is capable of training various models and performing different types of evaluations. This can also work with versioning and can be extended easily. We have also added features for tracking training and evaluation metrics using TensorBoard. This can be then extended by anyone to train a variety of models. The appendix has additional information about the architecture.

We designed a modular framework was designed with the following components:

Configuration Management:
JSON-based configuration files allow for easy modification of model parameters.

Data Management:
Handles preprocessing, splitting into training, validation, and test sets, and storing results.

Model Training and Evaluation
Implements benchmarking scripts, loss visualization, and model performance tracking via TensorBoard.

Layer Design
Encoder:
Input layer, multiple hidden layers with ReLU activation, and a latent Dirichlet parameter layer.

Decoder
Symmetric structure to reconstruct input data from the latent representation.

Uncertainty Estimation Module
Computes Dirichlet parameters to quantify confidence in predictions.
Note: Detailed Design decisions are also present in the overleaf document.

Conclusion

This study shows that Dirichlet Autoencoders can show promising results for anomaly detection with certain enhancements. While the DVAE performed slightly worse than VAE in raw reconstruction loss, its ability to capture uncertainty makes it a better choice in physics applications. For future work, one can look at improving the loss function, adding additional semantic information and testing on larger datasets to further improve performance.

References

[1] Dor Bank, Noam Koenigstein and Raja Giryes. “Autoencoders”. In: CoRR abs/2003.05991 (2020). URL: https://arxiv.org/abs/2003.05991.

[2] Diederik P. Kingma and Max Welling. “An Introduction to Variational Autoencoders”. In: CoRRabs/1906.02691 (2019). URL: http://arxiv.org/abs/1906.02691.

[3] Murat Sensoy, Lance Kaplan and Melih Kandemir. Evidential Deep Learning to Quantify Classification Uncertainty. 2018. URL: https://arxiv.org/abs/1806.
01768.

[4] Hai Siong Tan, Kuancheng Wang and Rafe McBeth. Evidential Physics Informed Neural networks. 2025. URL: https://arxiv.org/pdf/2501.15908

[5] Mengyuan Chen, Junyu Gao, and Changsheng Xu, Fellow, IEEE. Revisiting Essential and Nonessential Settings of Evidential Deep Learning. 2024. URL: https://arxiv.org/pdf/2410.00393

[6] Taeseong Yoon, Heeyoung Kim. Uncertainty Estimation by Density Aware Evidential Deep Learning. 2024. URL: https://arxiv.org/pdf/2409.08754

[7] Kai Ye, Tiejin Chen, Hua Wei, Liang Zhan. Uncertainty Regularized Evidential Regression. 2024. URL: https://arxiv.org/pdf/2401.01484

[8] Bertrand Charpentier, Ransalu Senanayake, Mykel Kochenderfer, Stephan Günnemann. Disentangling Epistemic and Aleatoric Uncertainty in Reinforcement Learning. 2022. URL: https://arxiv.org/pdf/2206.01558