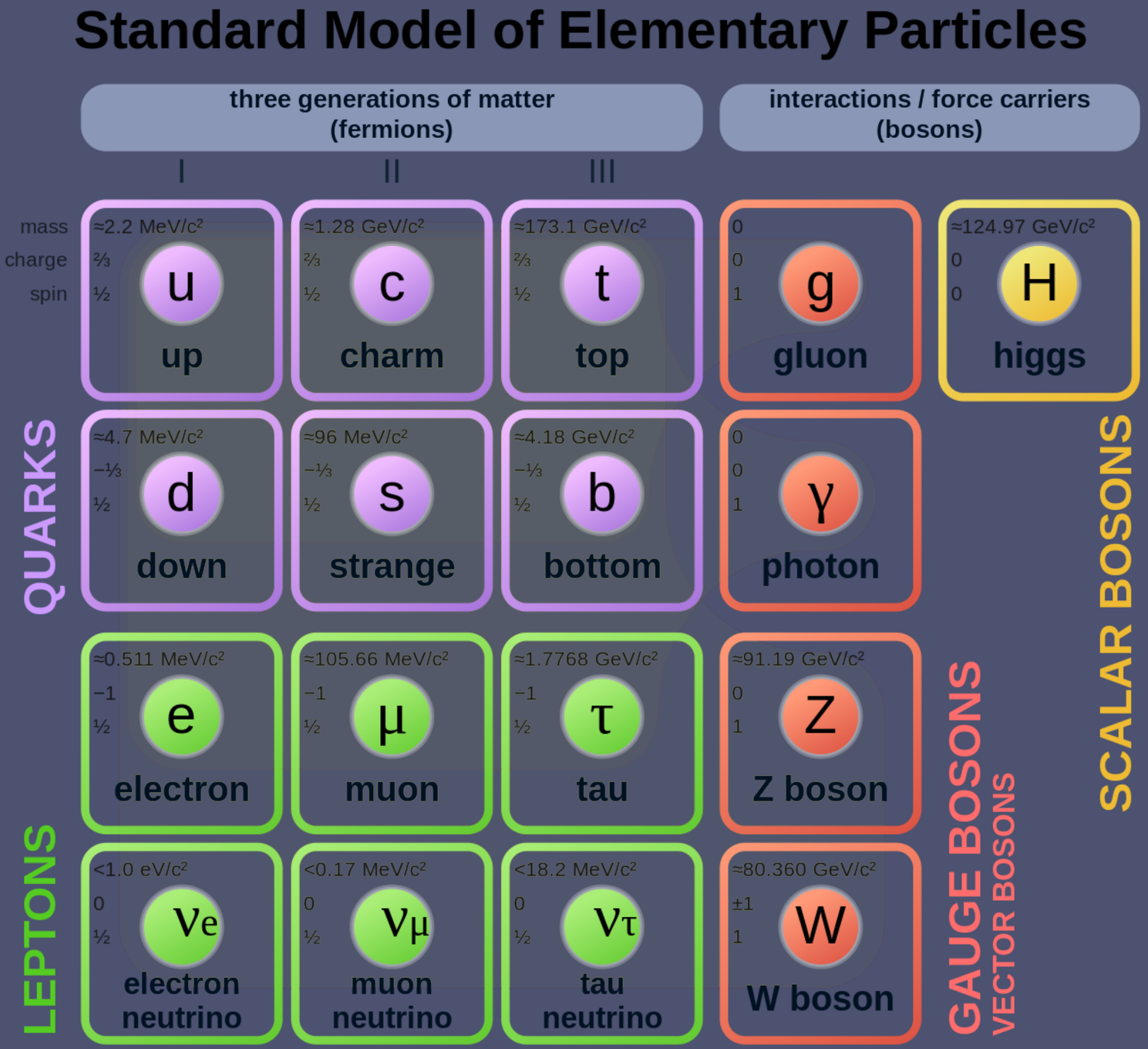
ANOMALY DETECTION AS A VALUABLE TOOL FOR UNCOVERING UNEXPECTED PHENOMENA AT COLLIDERS

LIP

Simão Silva Cardoso

Supervisors: Nuno Castro, Rute Pedro e Miguel Peixoto

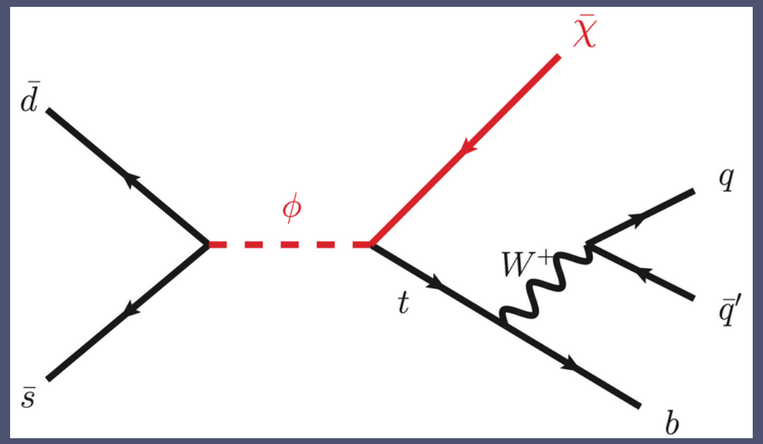


MOTIVATION

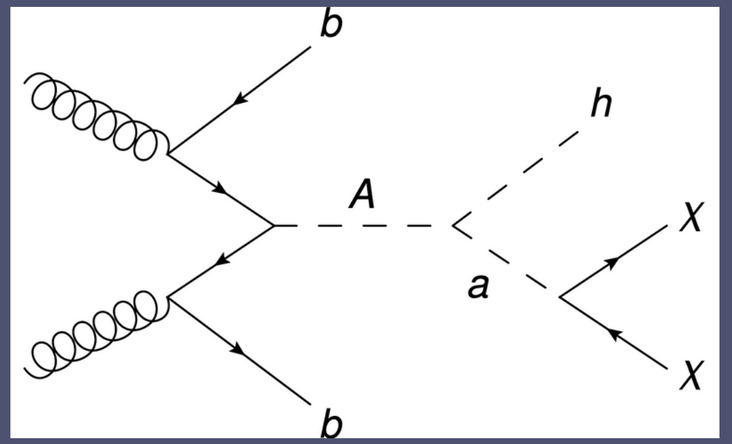
The Standard Model is the most successful theory of particle physics to date, but there are many open questions. Research on this topic can be assisted by using anomaly detection methods on colliders' data, finding possible signals that might hint at new physics.

SIMULATED DATASETS

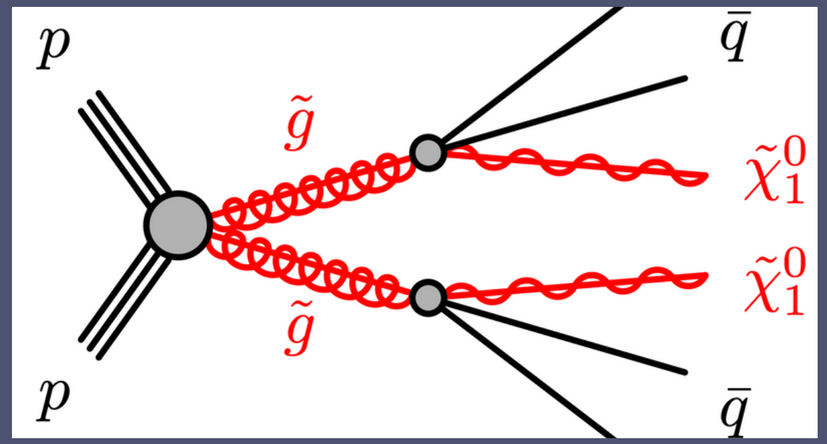
Monte Carlo events were simulated for all processes in the background.
The signals used are examples.



Resonant Dark Matter particles (S1)



Dark Matter production by two Higgs doublets (S2)



Gluino pair (S3)

Heavy Vectorial Triplet (S4)

Methods used

31 features used - some examples:
MET, 4-momentum jets and large jets, HT (scalar momentum sum),...

Supervised:

- Deep Neural Networks

Semi-supervised:

- Isolation Forest (w/o contamination)
- Autoencoder
- Variational Autoencoder (w/o contamination)

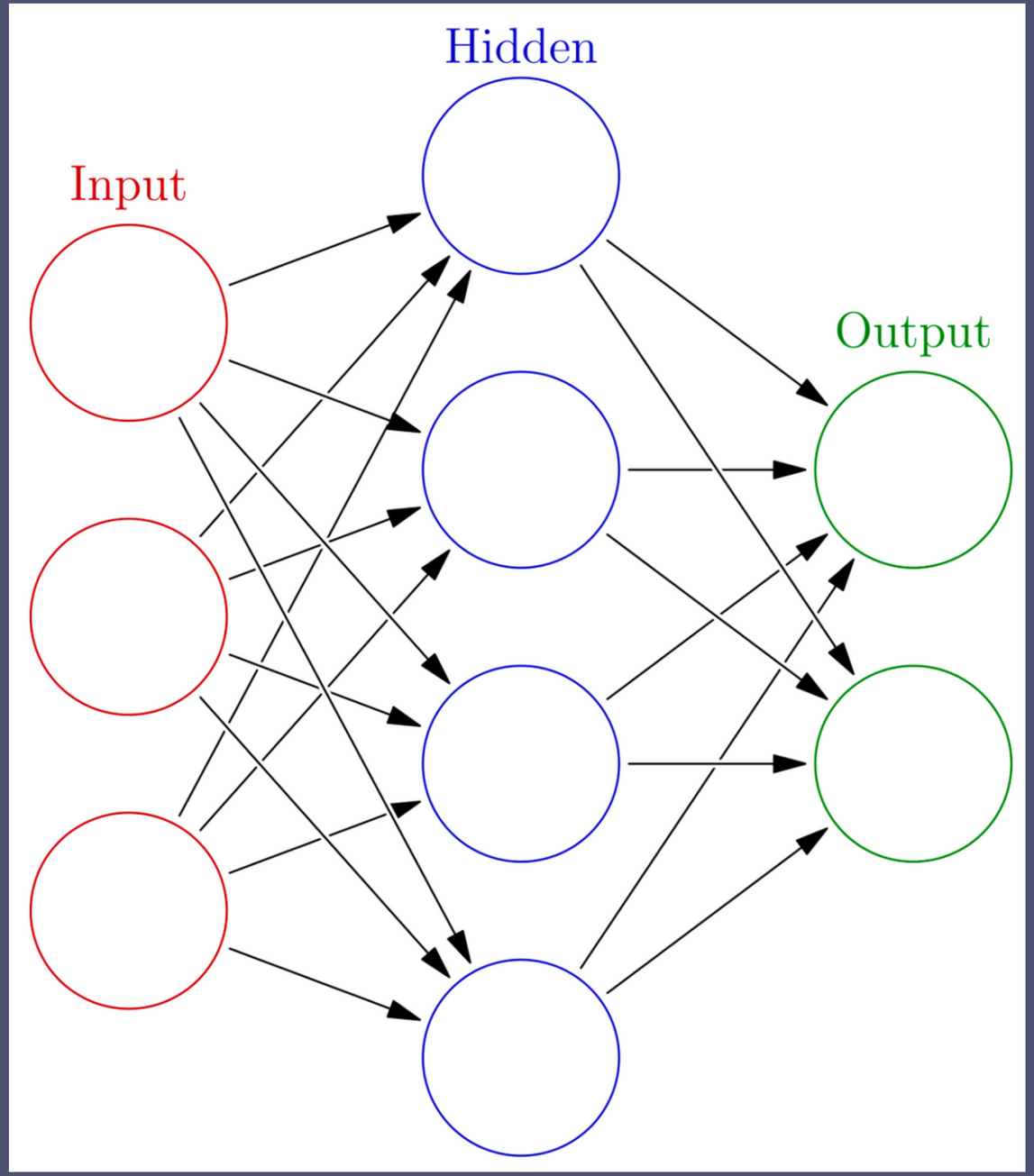
SUPERVISED LEARNING - NEURAL NETWORKS

It’s an algorithm whose architecture comprises layers with many neurons, leading to an output. Labels are used for training.

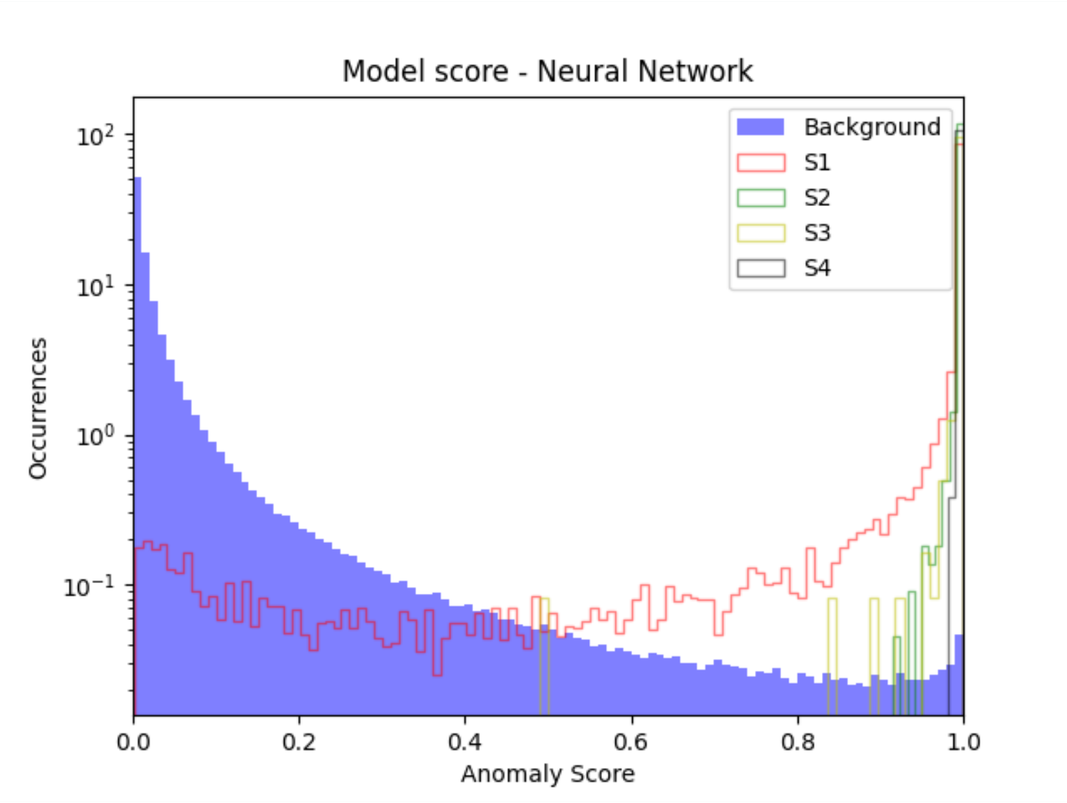
$$\mathbf{f}_l(\mathbf{z}) = \mathbf{g}_l(\mathbf{W}_l\mathbf{z} + \mathbf{b}_l)$$

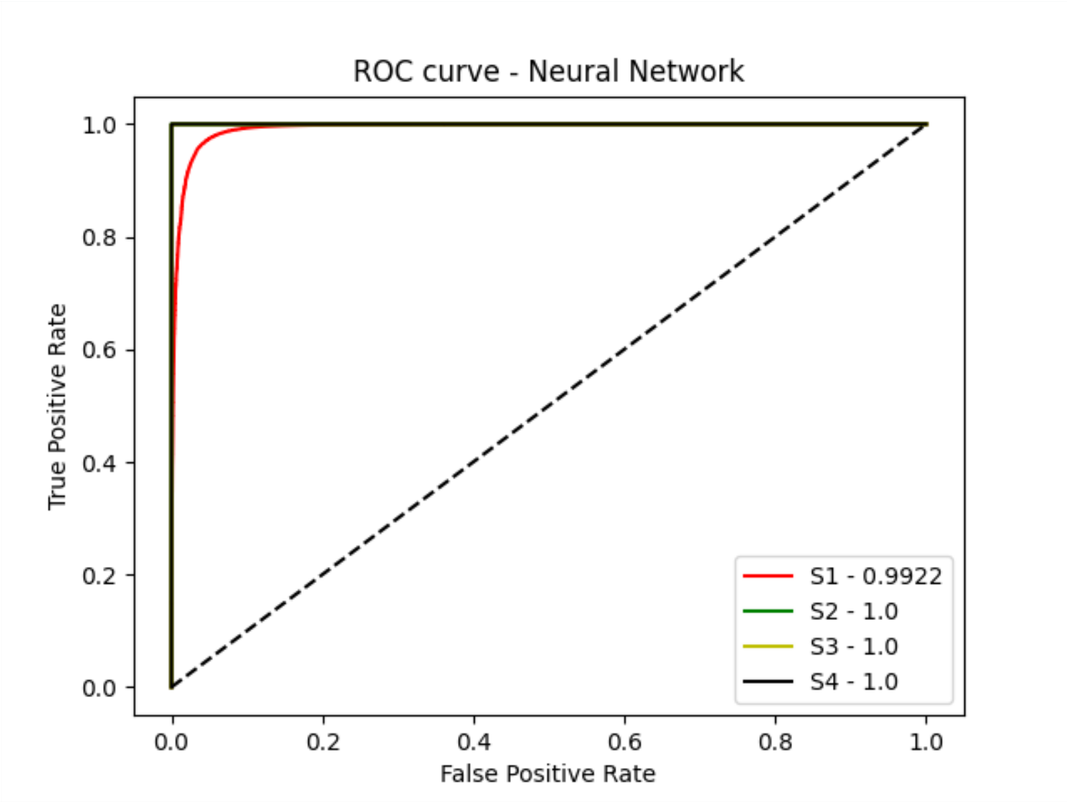
$$\min_{\mathbf{W},\mathbf{b}} \frac{1}{N} \sum_{i}^{N} [y_i \log_2[\mathrm{NN}(\mathbf{x}_i, \mathbf{W}, \mathbf{b})] +$$

$$(1 - y_i) \log_2[1 - \mathrm{NN}(\mathbf{x}_i, \mathbf{W}, \mathbf{b})]]$$



NEURAL NETWORKS - RESULTS



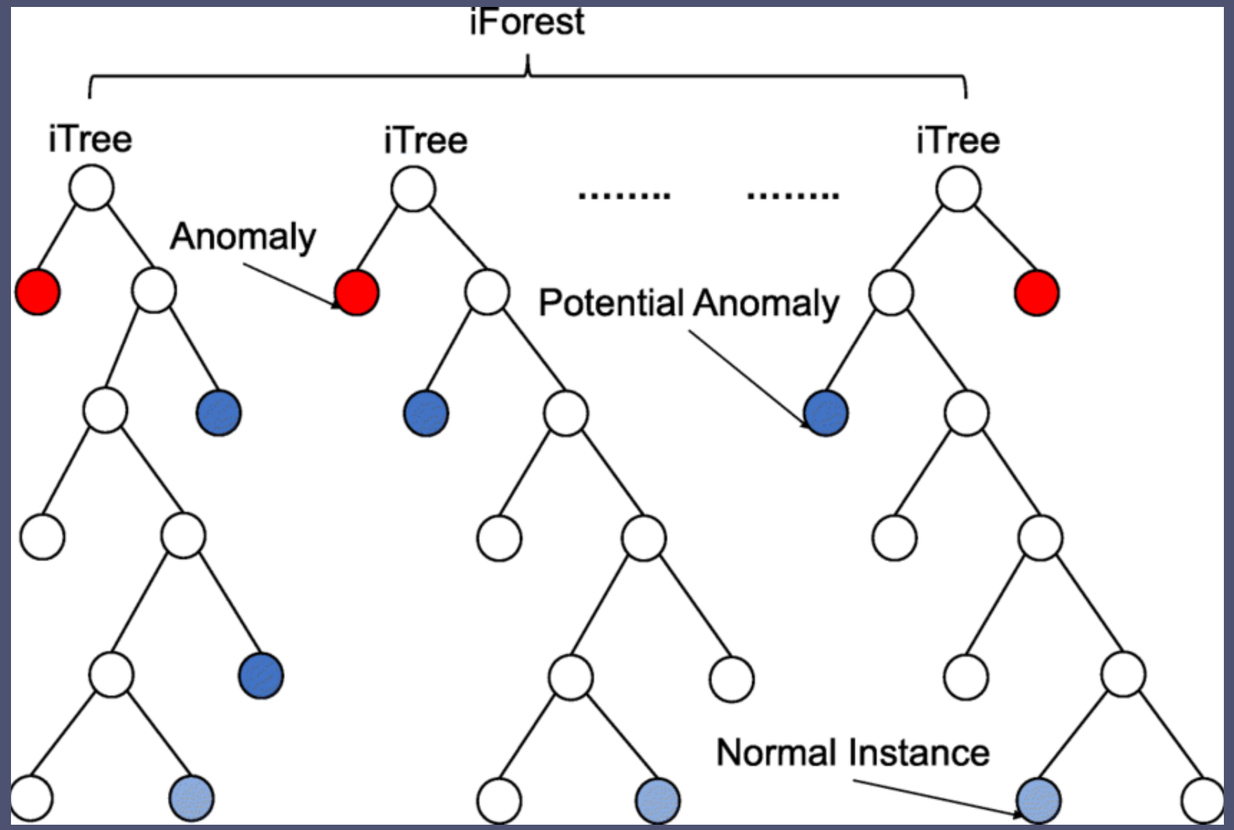


WHY PREFERING SEMI-SUPERVISED TO SUPERVISED METHODS?

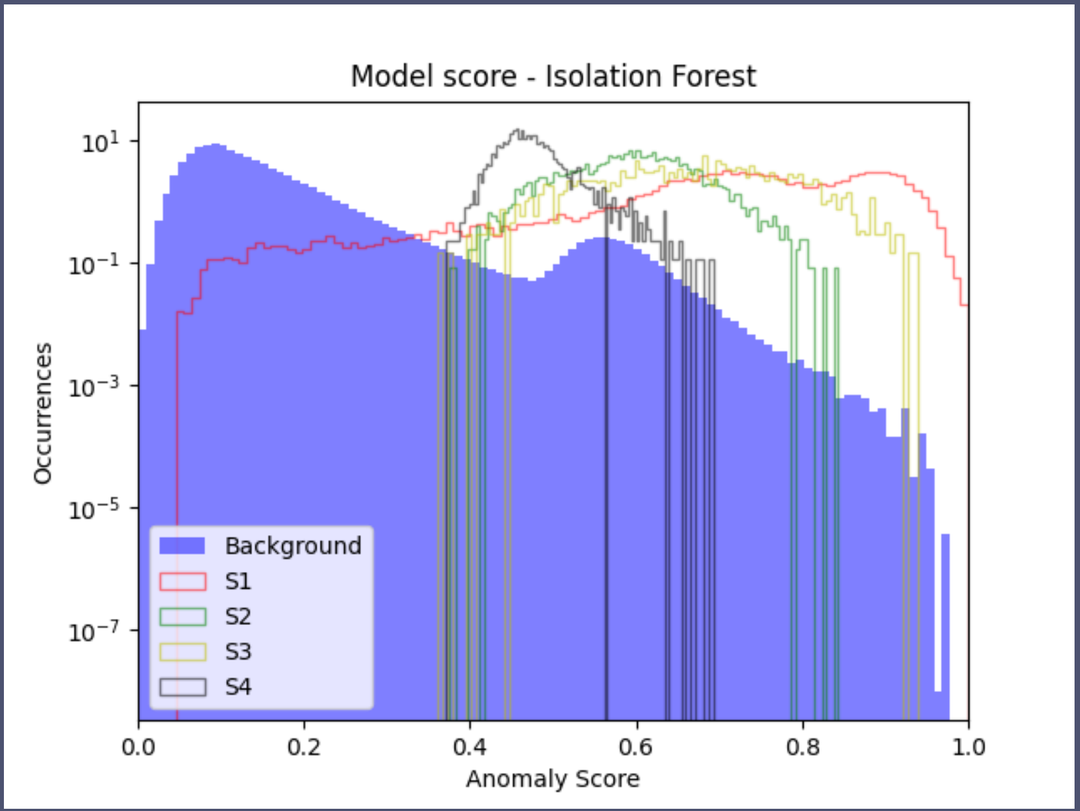
- Neural Network's performance might degrade when applied to other signals.
- Real datasets aren’t labelled.
- Semi-supervised methods are independent and offer great performance.

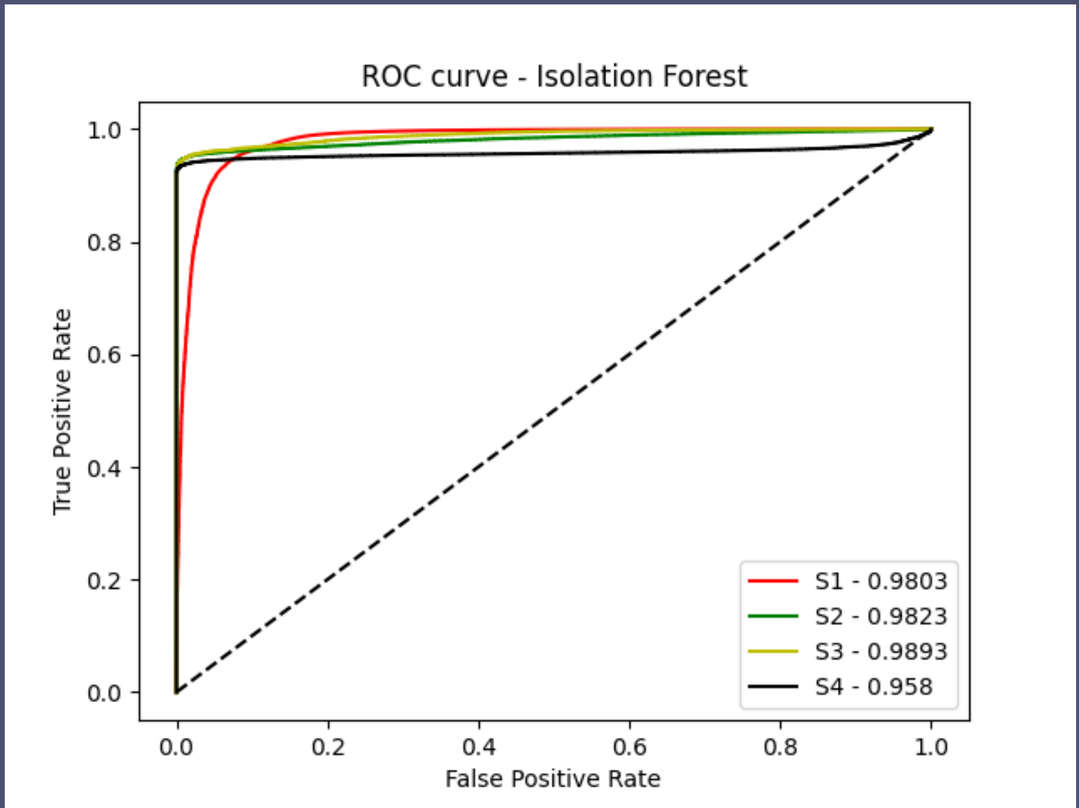
SEMI-SUPERVISED LEARNING - ISOLATION FOREST

Isolation Forest creates tree structures that can represent recursive partitioning during learning. The path length of the trees averaged over a forest of such random trees is used to measure the anomaly score.

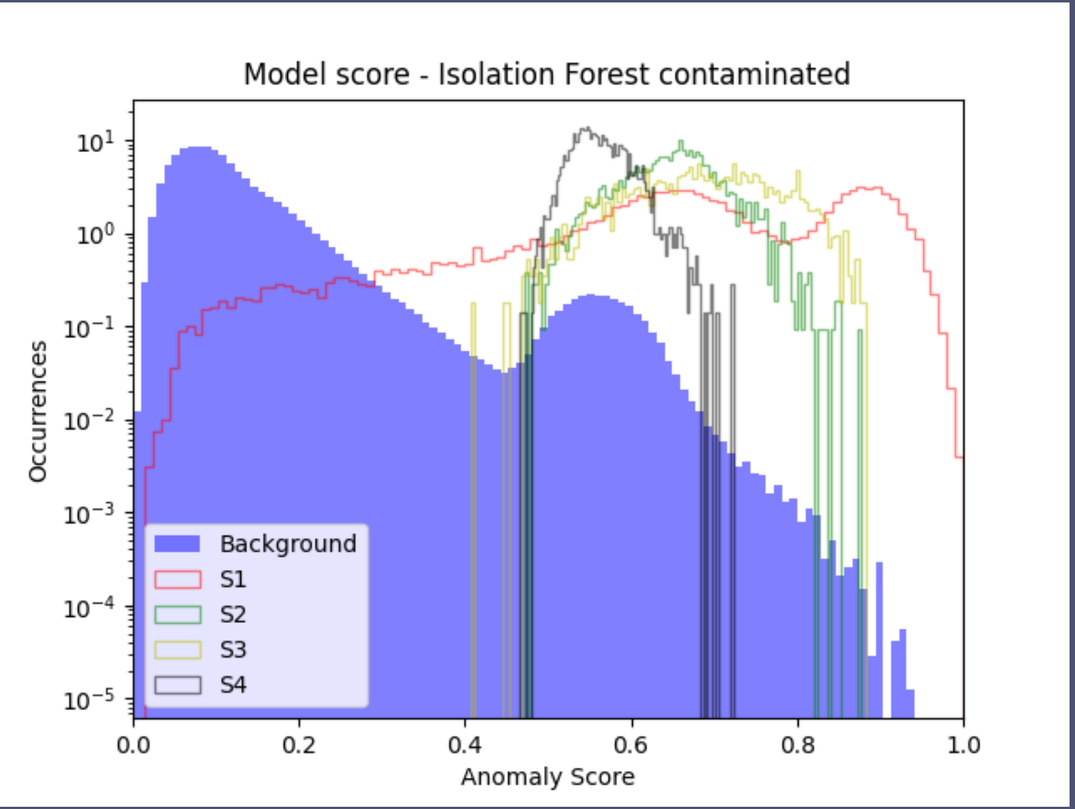


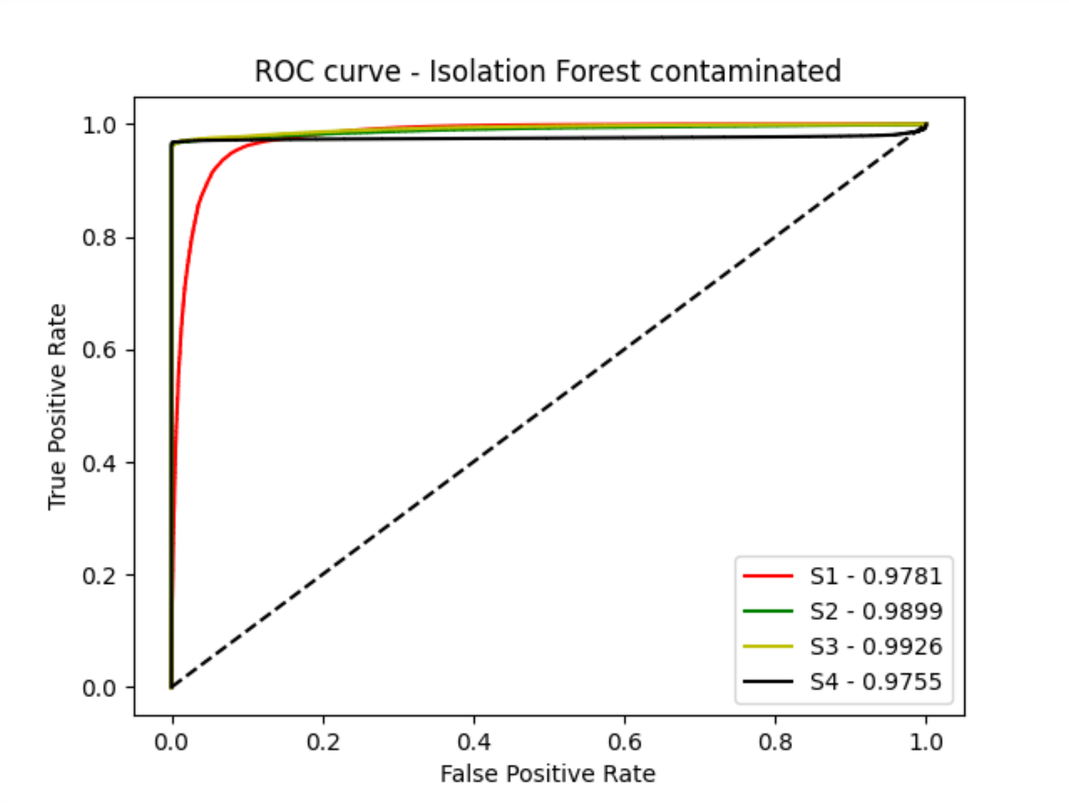
ISOLATION FOREST - RESULTS





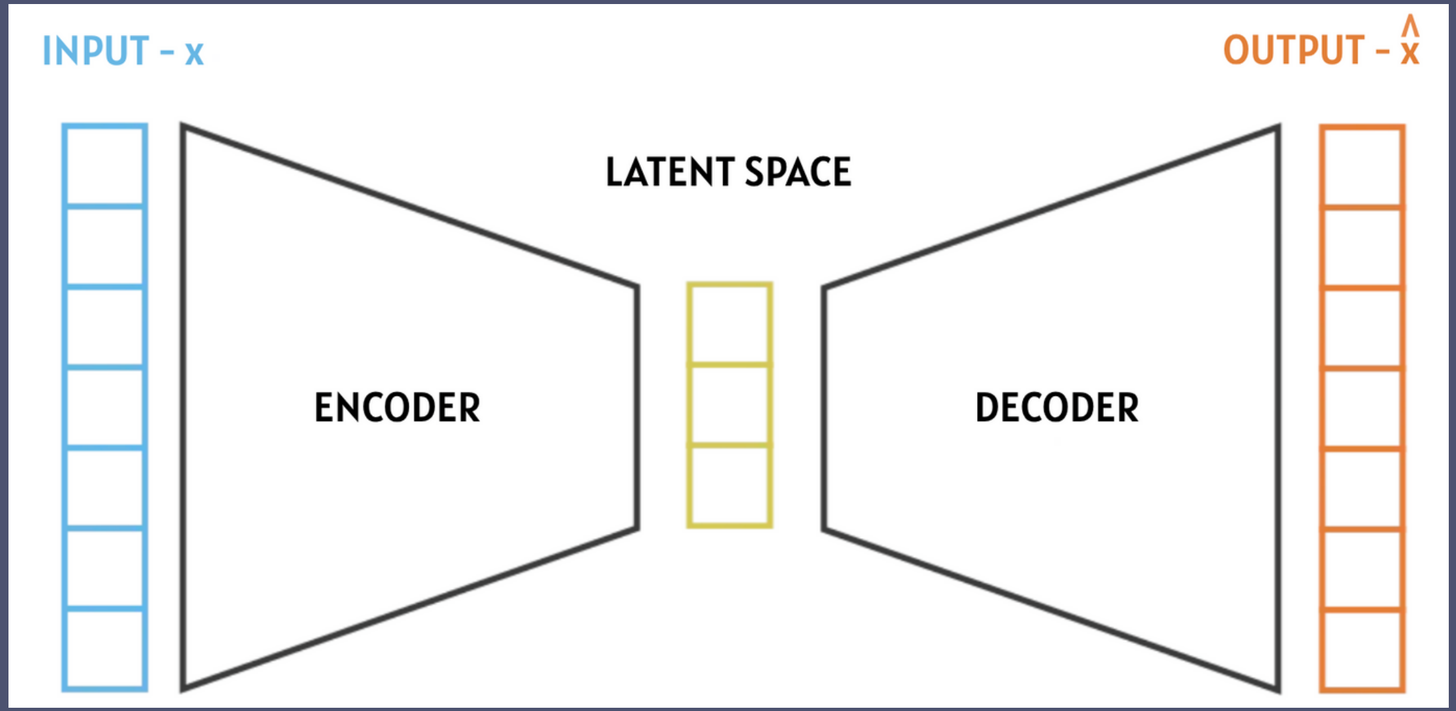
ISOLATION FOREST - RESULTS (CONTAMINATED)





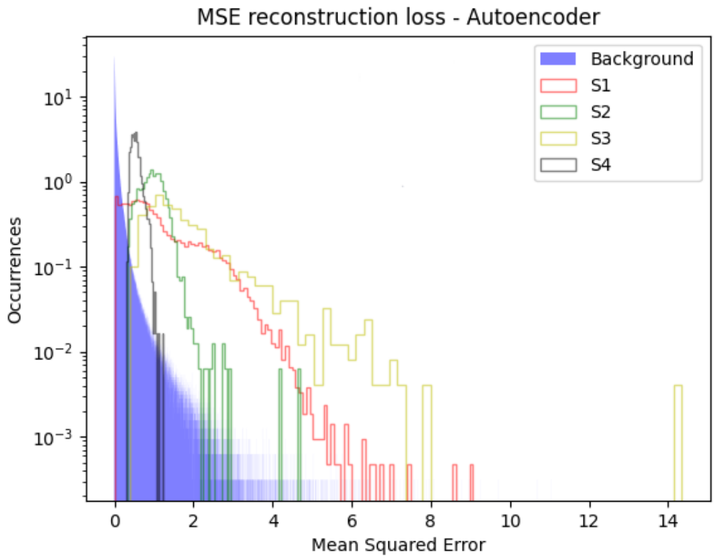
SEMI-SUPERVISED LEARNING - AUTOENCODER

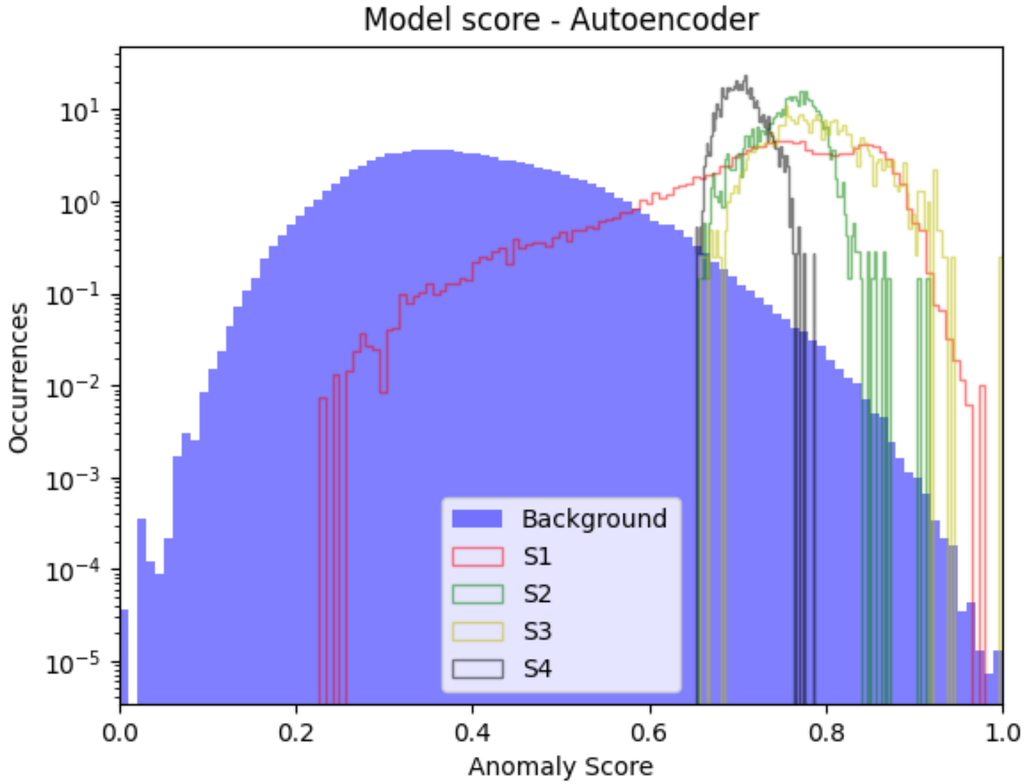
The Autoencoder is a deep architecture that learns to compress (encode) and then decompress (decode) data through a bottleneck called the latent space.

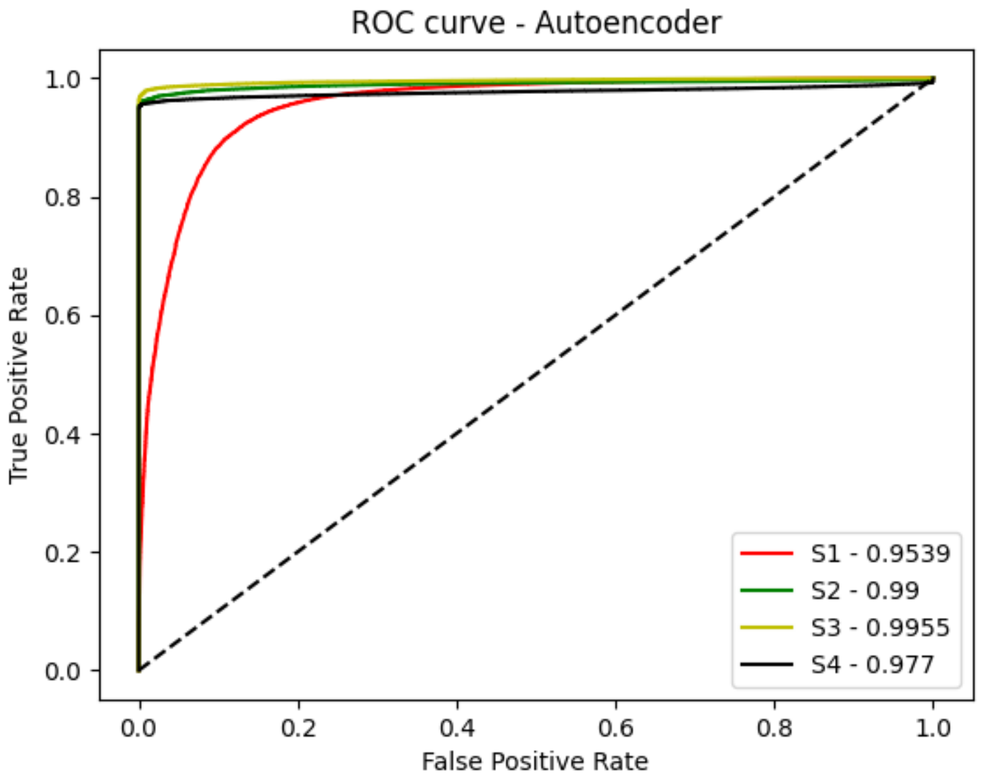


$$\text{loss} = \min_{\mathbf{W},\mathbf{b}} \frac{1}{N} \sum_{i}^{N} \|\text{AE}(\mathbf{x}_i, \mathbf{W}, \mathbf{b}) - \mathbf{x}_i\|^2$$

AUTOENCODER - RESULTS

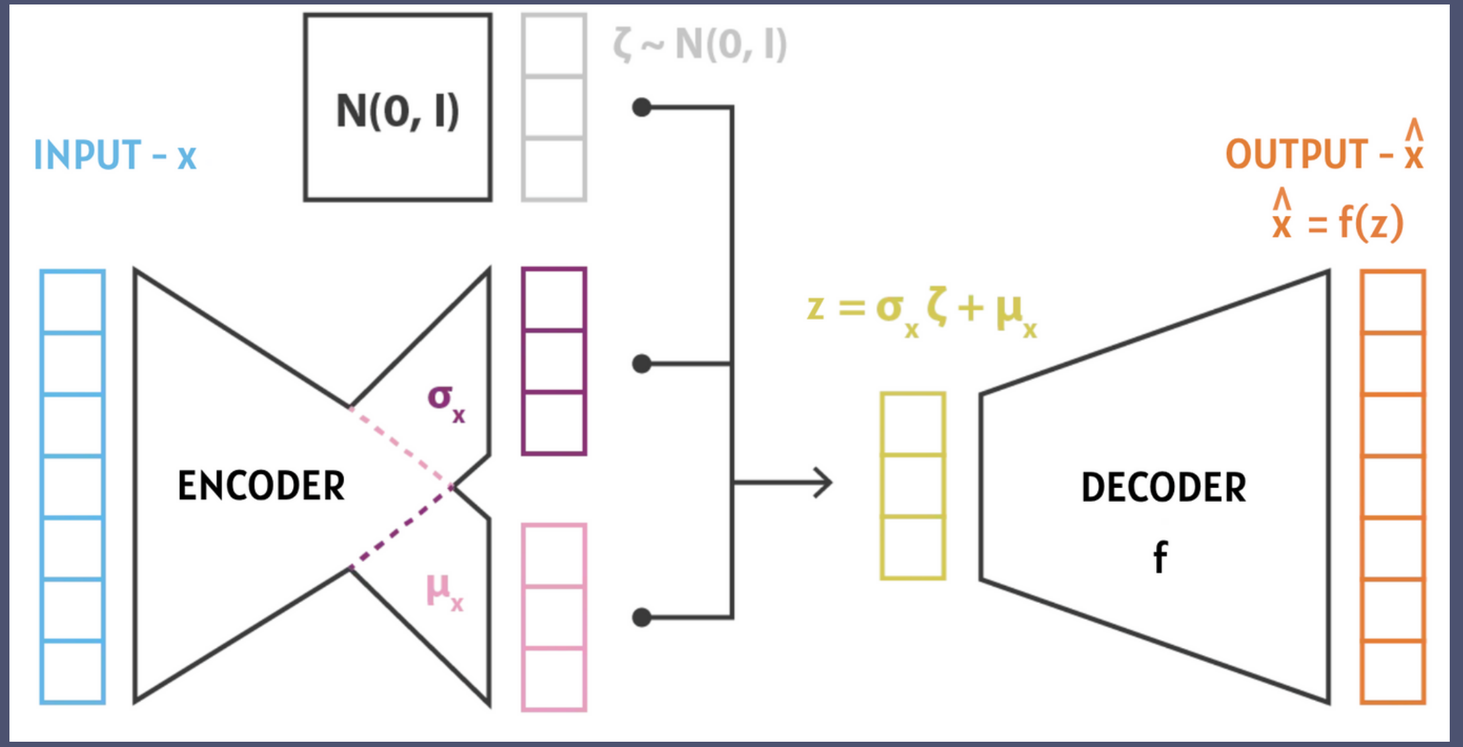






SEMI-SUPERVISED LEARNING - VARIATIONAL AUTOENCODER

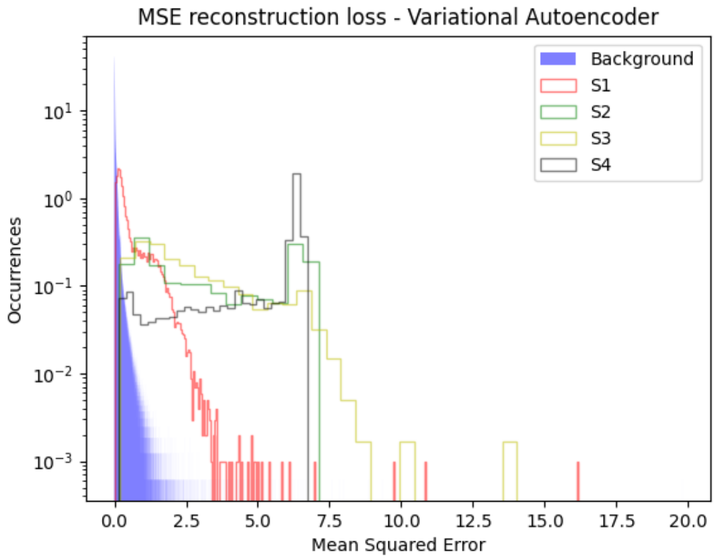
The Variational Autoencoder (VAE) is an AE whose training is regularised to avoid overfitting and ensure that the latent space has suitable properties that enable the generative process.

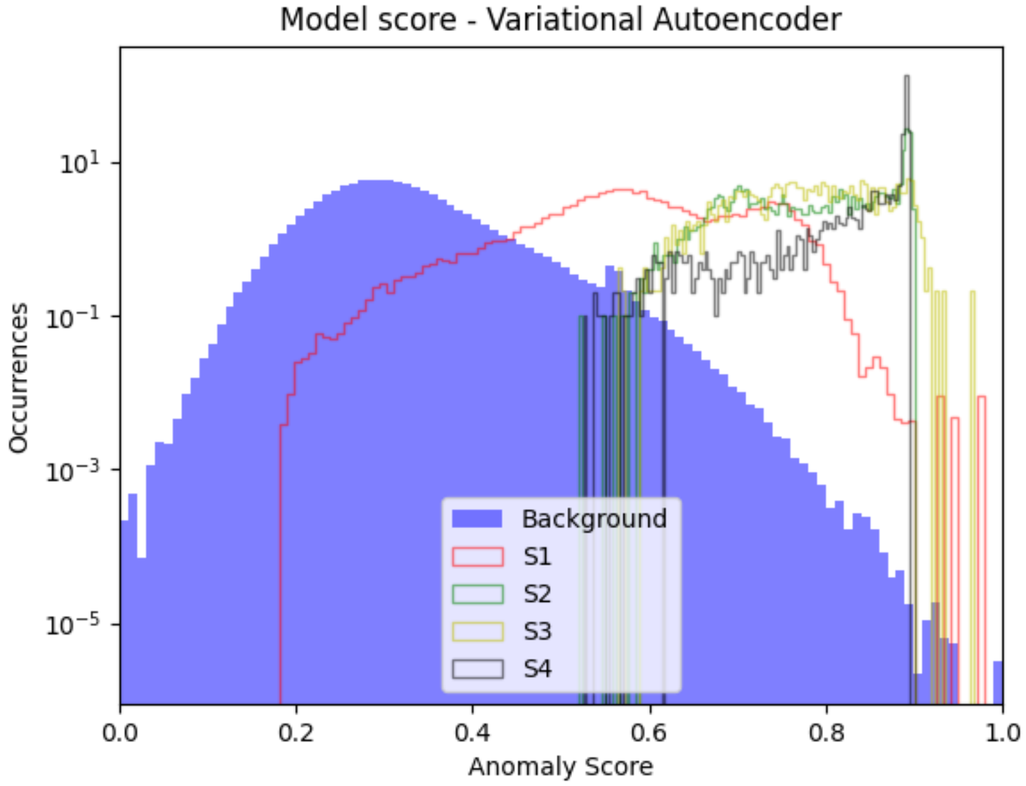


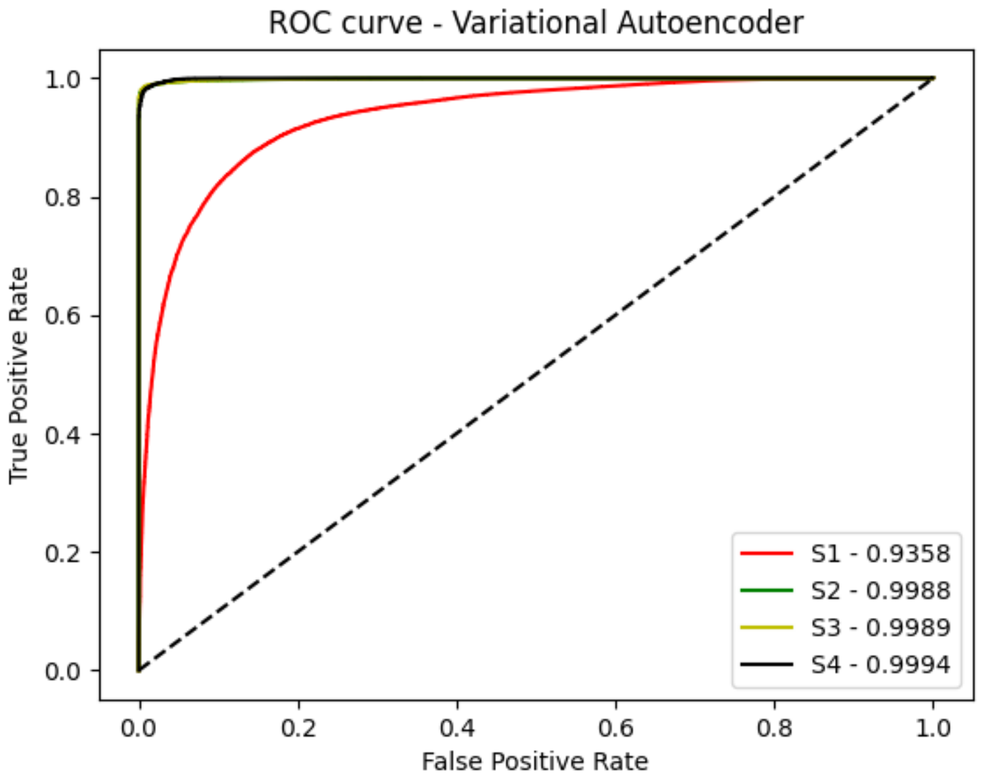
$$z = \mu_x + \sigma_x \odot \zeta$$

$$\text{loss} = \text{MSE}(\mathbf{x}, \hat{\mathbf{x}}) + \beta \text{KL}[N(\mu_x, \sigma_x), N(0, 1)]$$

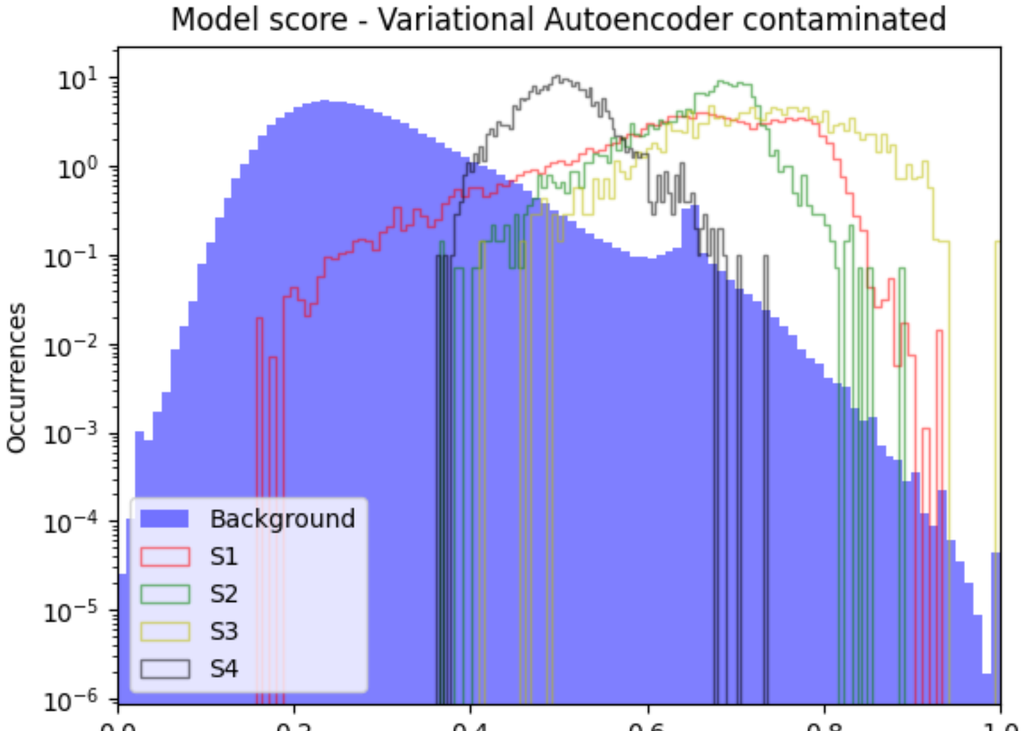
VARIATIONAL AUTOENCODER - RESULTS

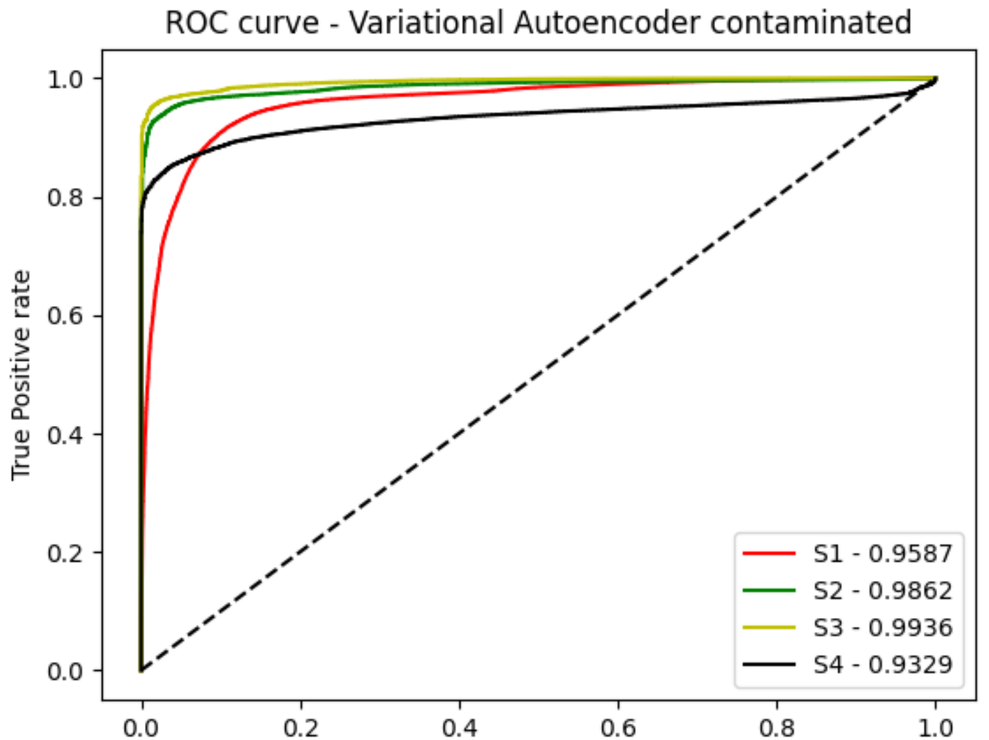




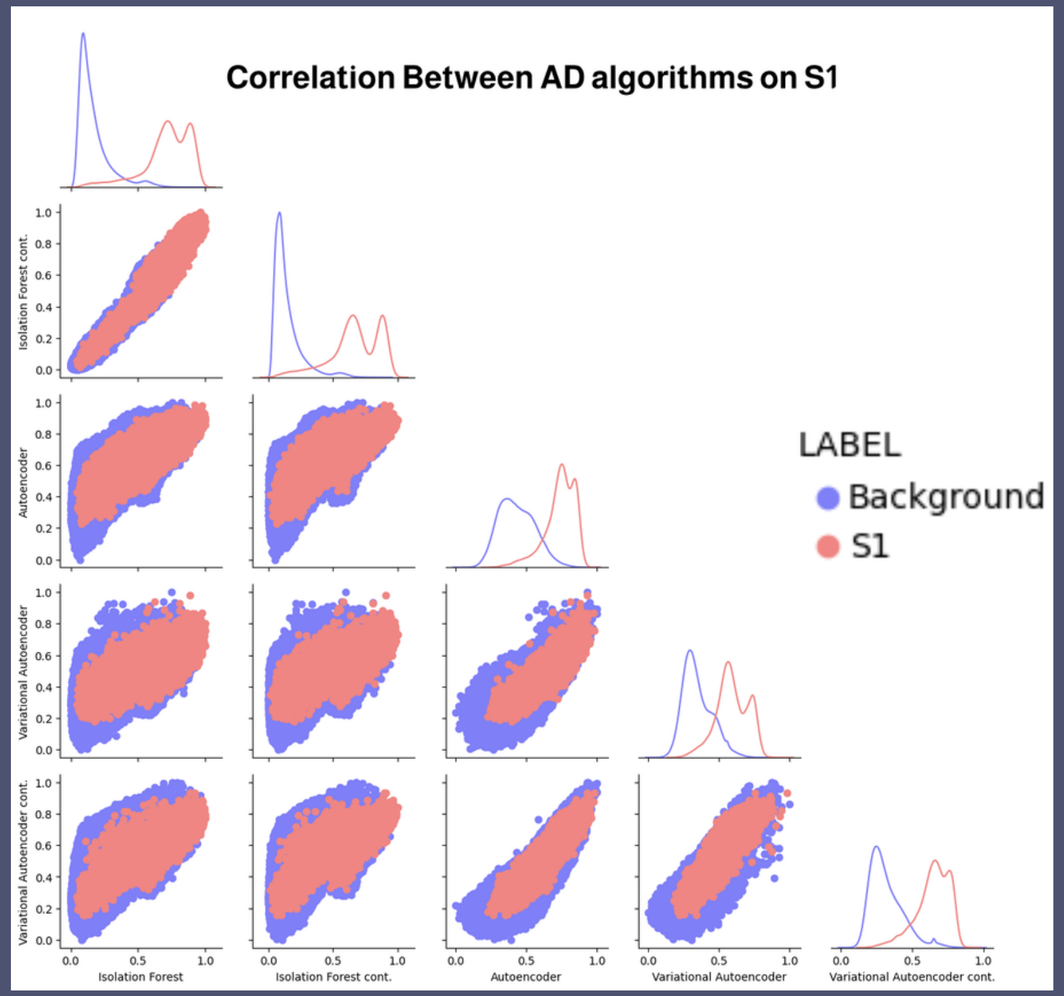


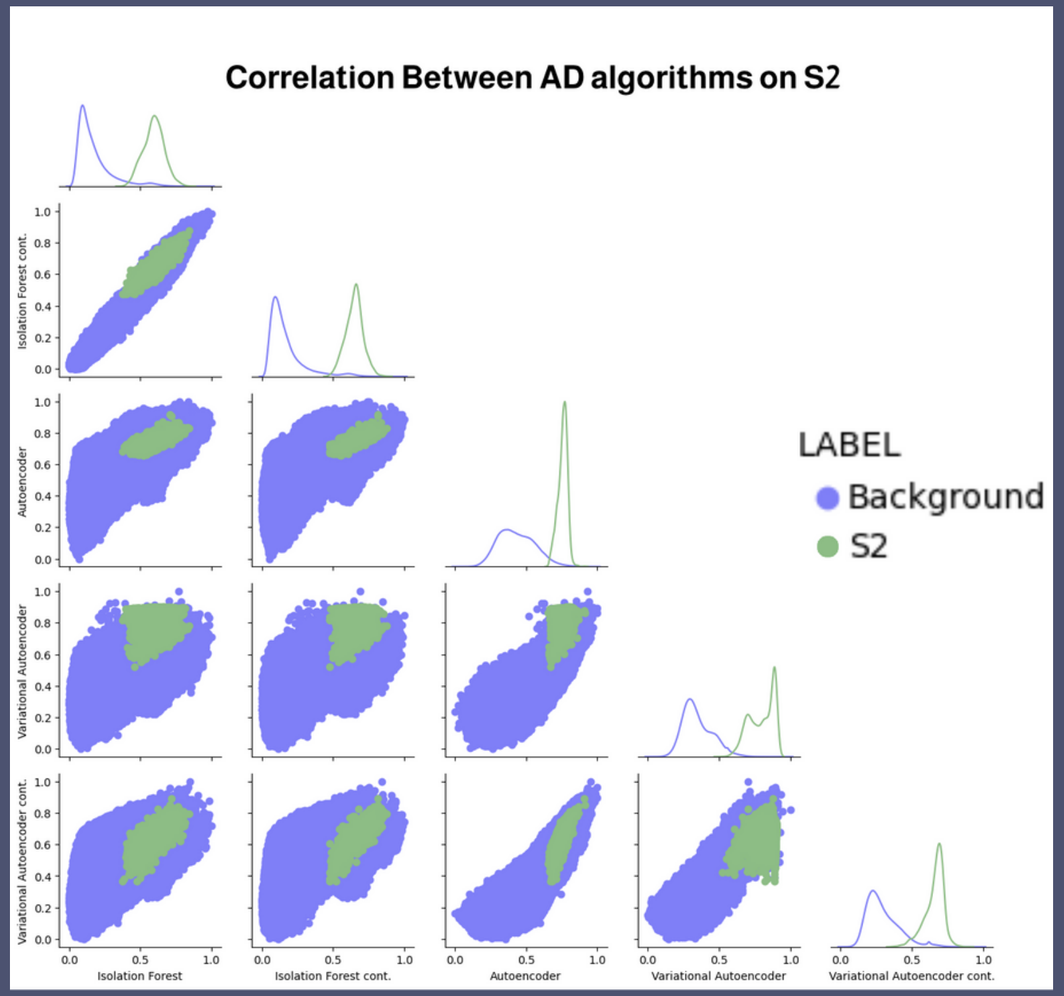
VARIATIONAL AUTOENCODER - RESULTS (CONTAMINATED)



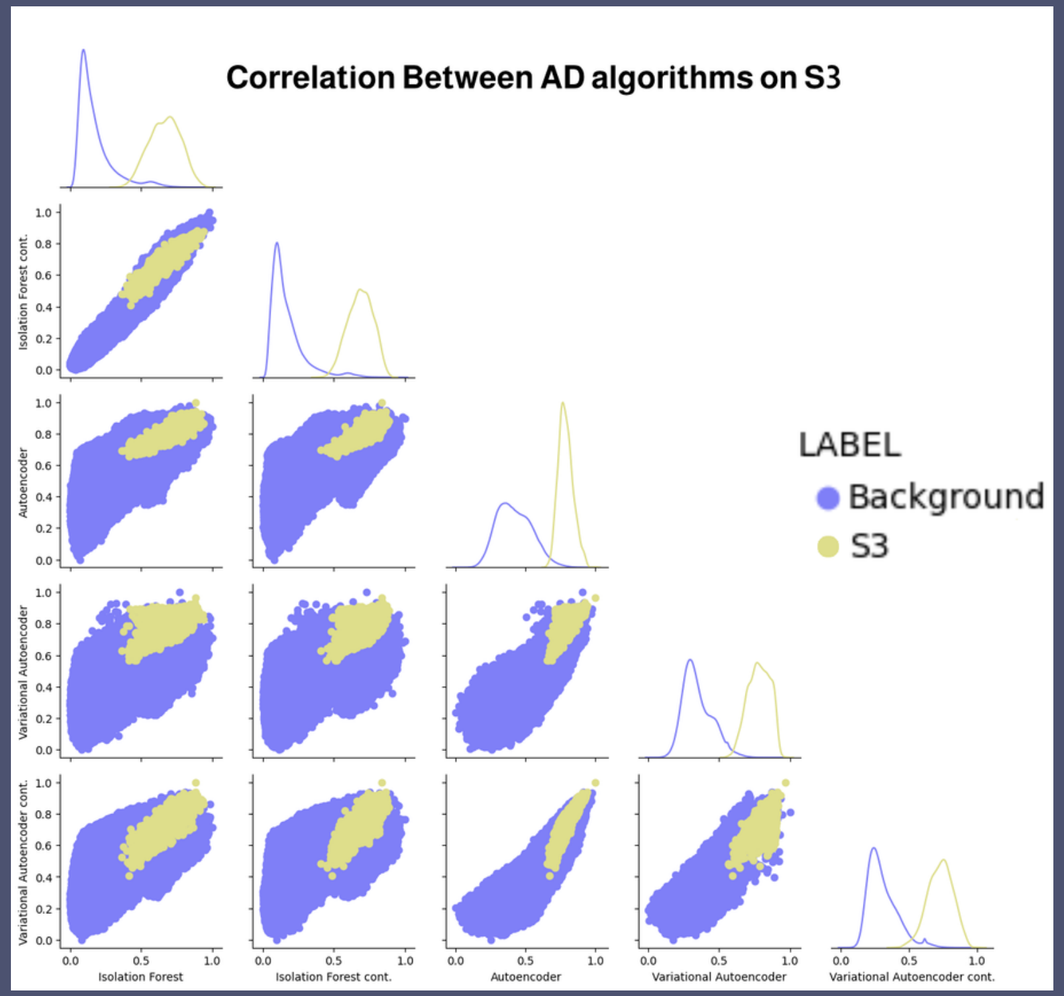


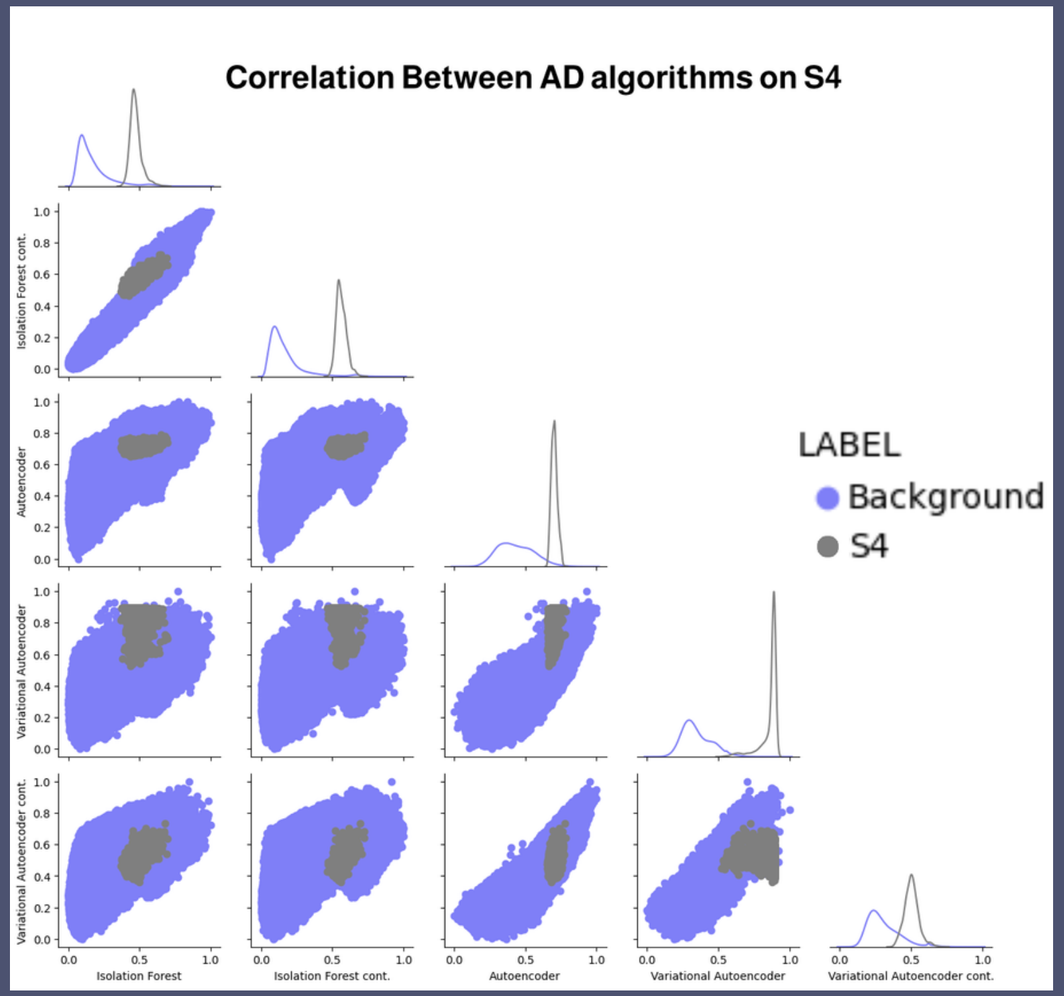
CORRELATION BETWEEN METHODS



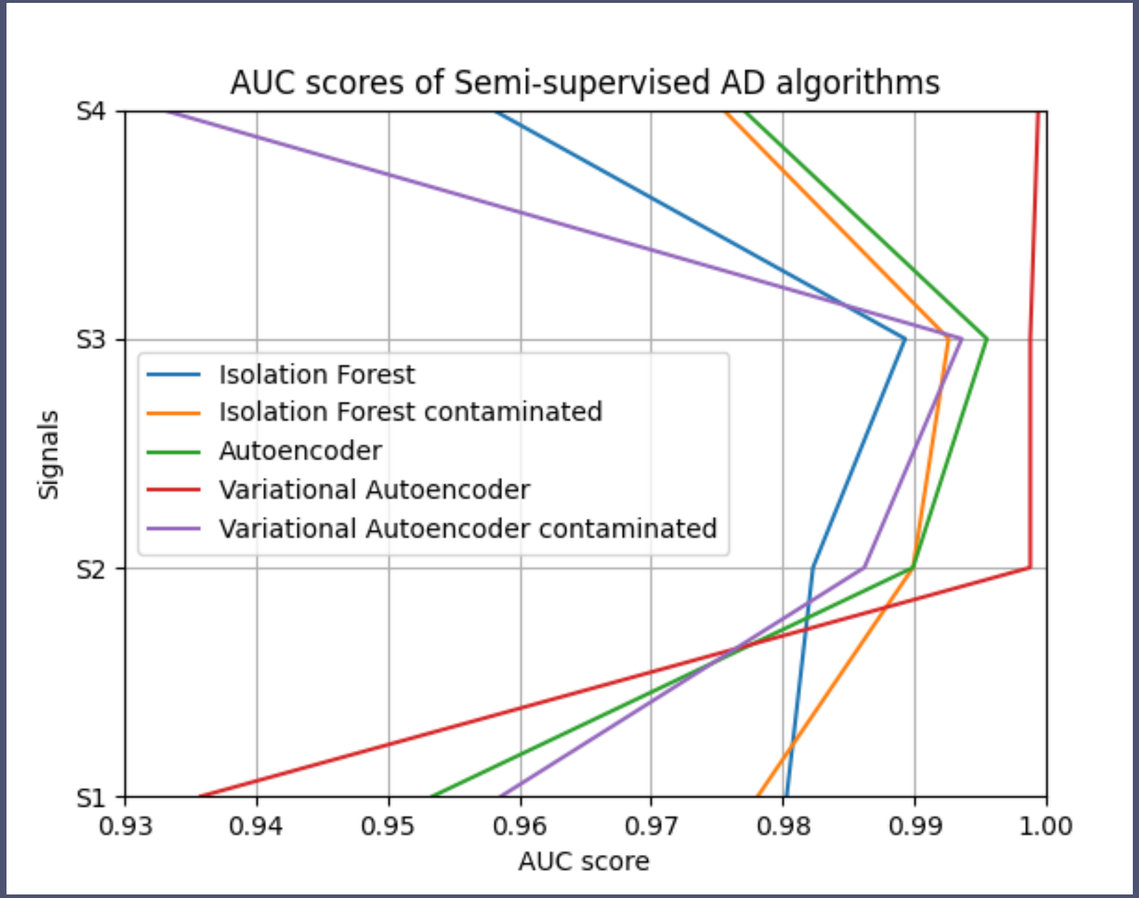


CORRELATION BETWEEN METHODS





AUC SCORES OF THE METHODS



CONCLUSION

Distinct algorithms for anomaly detection are highly effective in isolating diverse types of BSM physics. Furthermore, these algorithms can complement each other in unsupervised searches for new physics, making them potential tools in particle physics research.