ML4Crypto Classification: Model Comparison Report

Overview

This report summarizes the results obtained from various machine learning models on the ML4Crypto 2024 dataset, with a focus on binary classification of bitstreams.

Dataset Description

- Features: ID, Bitstream (1024-bit length), Label (0 or 1 for random generator type).
- Preprocessing: Bitstreams are converted into sequences, reshaped for models like LSTM, and split into training and testing datasets.

Model Comparisons

Model	**Accuracy**	**Cross Validation Accuracy**	**Advanced model accuracy**
Logistic Regression	**0.54**	**0.4875**	
SVM	**0.535**	**0.5215**	
Random Forest	**0.52**	**0.5045**	
MLP Neural Network	**0.5325**	**0.5089**	**0.5225**
CNN	**0.4975**	**0.4981**	
Deep CNN	**0.49**		**0.5025 (ResNet inspired CNN)**
LSTM	**0.4925**	**0.5095**	**0.495**

Analysis and Observations

- Logistic Regression: simple one but performs well, though cross validation accuracy is low.
- SVM: Best one with cross validation accuracy.
- MLP: performs well and good choice for experimentation yet cross validation accuracy is just above the average.
- CNN: though famous for pattern recognition, it did not perform well with the specific data set.
- DNN: With deeper architecture accuracy is less than shallow neural networks. Strange yet kind of expected as CNN performs below average.
- LSTM: Tried as some literature suggests, however, it did not get good accuracy.

Conclusion

This classification task is challenging and anything beyond 50% may be a non-trivial result as that might be considered as a flaw of the underlying random number generator.
Based on cross-validation and comparative results, SVM achieved the highest performance. Also, it should be noted that classical ML models with simple structures perform better than complicated models with more hidden layers. Future work may be done with feature extraction of random bitstreams. Finally, this challenge is a blind classification with no knowledge of underlying random number generators, which makes it quite difficult to choose and tune a particular ML model and mostly relies on trial and error methodology. If prior knowledge of the RNG is available then security analysis of the RNG will be more sophisticated than these brute force attack scenarios.