Can machine learning help us shed light on the dark side of the universe?

Divij Agarwal[1], Vasileios Kalaitzidis[1,2]

1. College of Science, Purdue University; 2. School of Physics & Astronomy, University of St Andrews

Introduction

- XENONnT detector is a dual-phase Xenon based instrument designed for the direct detection of Dark Matter using nuclear recoil events.
- Since the nuclear recoil for neutron scattering and WIMP scattering is the same, we can tune our detector to study neutrons.
- Calibration of the detector is done using a radioactive source of Yttrium-Beryllium which emits neutrons at 152keV which are slowed down by water and reach the detector at lower energies.
- The data from each event consists of a small peak (S1) produced by a photon (emitted at the time of the nuclear recoil) and is followed by a larger peak (S2) caused by an electron produced by the previous recoil event.
- However, these peaks can also be separately produced and incorrectly paired due to background events in the detector.
- **Our aim is to classify events as background or signal**
- Having the ability to reliably detect signals in the XENONnT detector is **crucial for dark matter research**

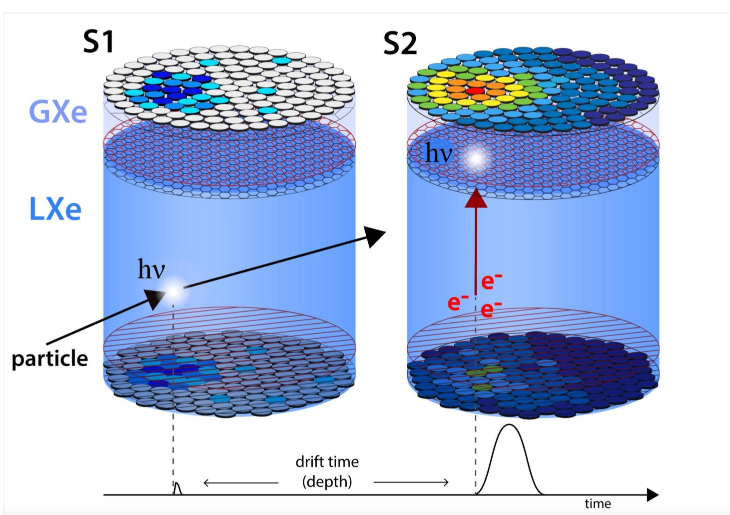


Figure 1: Schematic of dual-phase TPC

Citation: Di Gangi, P. (2021). The XENON Project. Universe, 7(8), 313

Selecting initial features

Initial features were selected using two criteria:

- Features that we expect to be different for signal and background based on the underlying physics of our experiment.
- Features which are strongly correlated with the occurrence of a signal. Alt S1 and Alt S2 are two such are features.

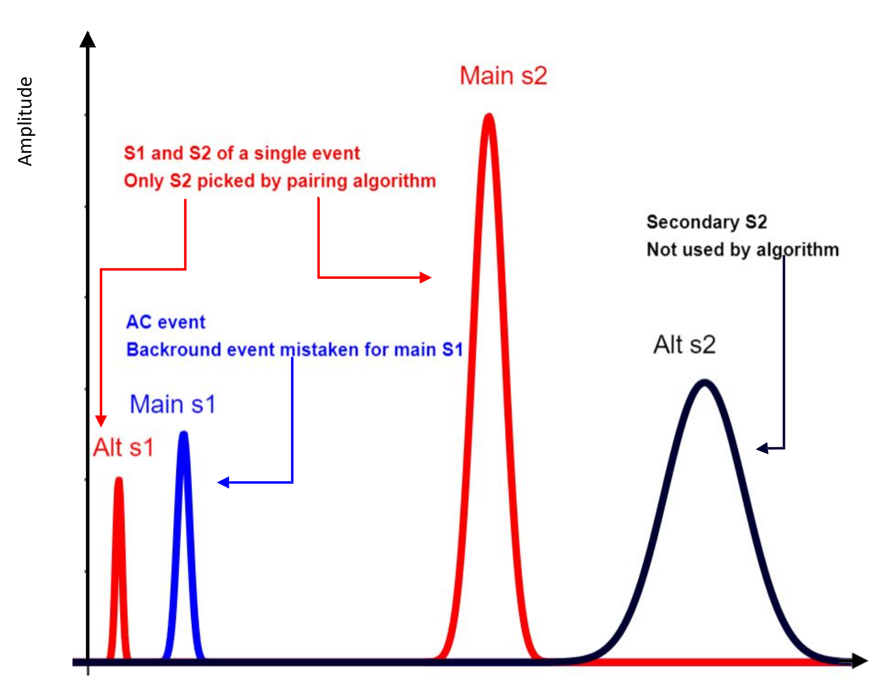


Figure 2: Alt s1 and alt s2 are features that are correlated with the Main s1 and s2 event
The pairing algorithm can incorrectly pair a background s1 and a correct s2 signal

Testing machine learning models

- XGBoost (Extreme boosting gradient tree)
- Neural networks (Convolutional neural network, recurrent neural network)
- Logistic Regression

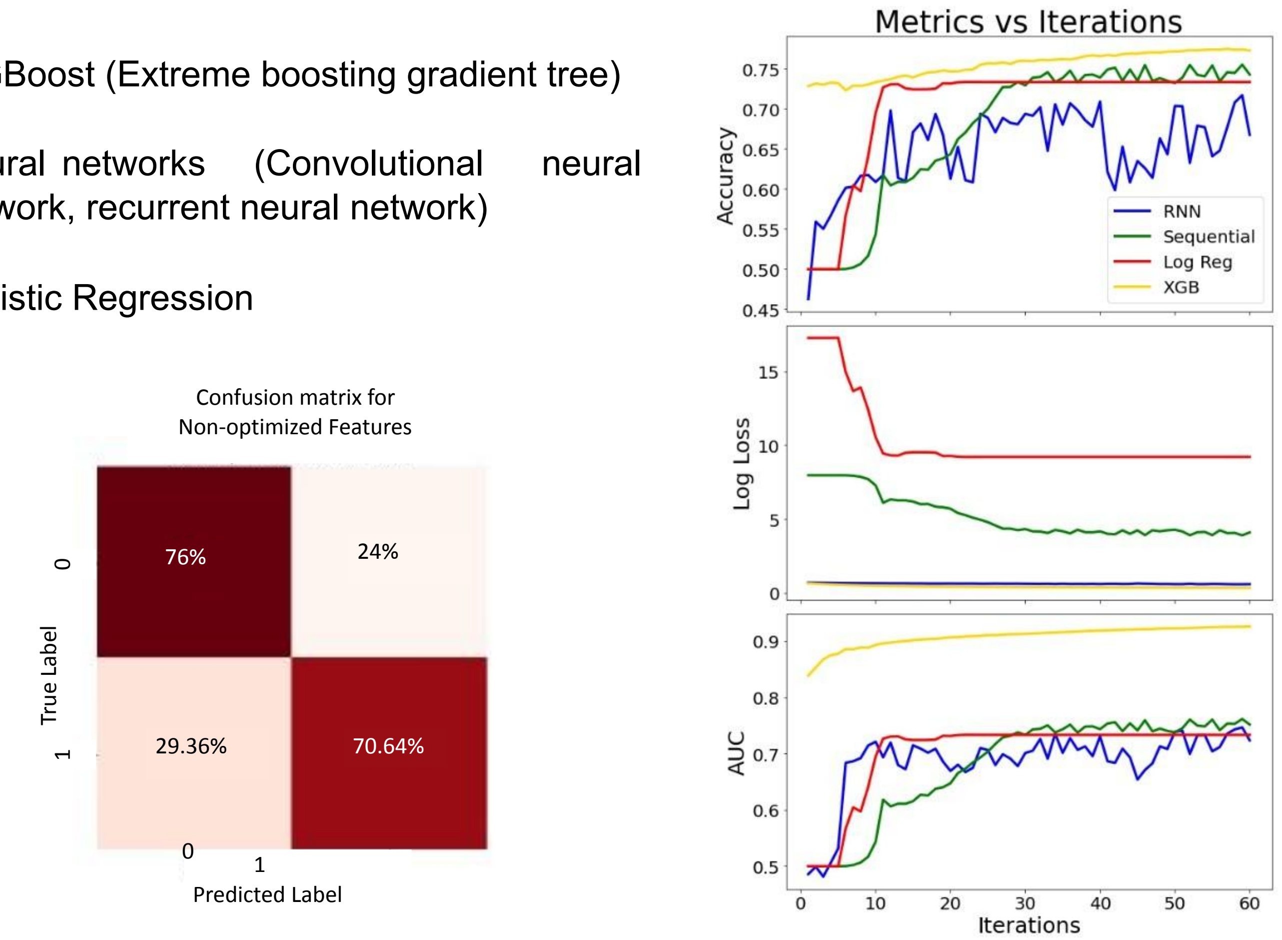


Figure 3: Performance evaluation of machine learning models for binary classification of signal vs background.

Selecting optimal features

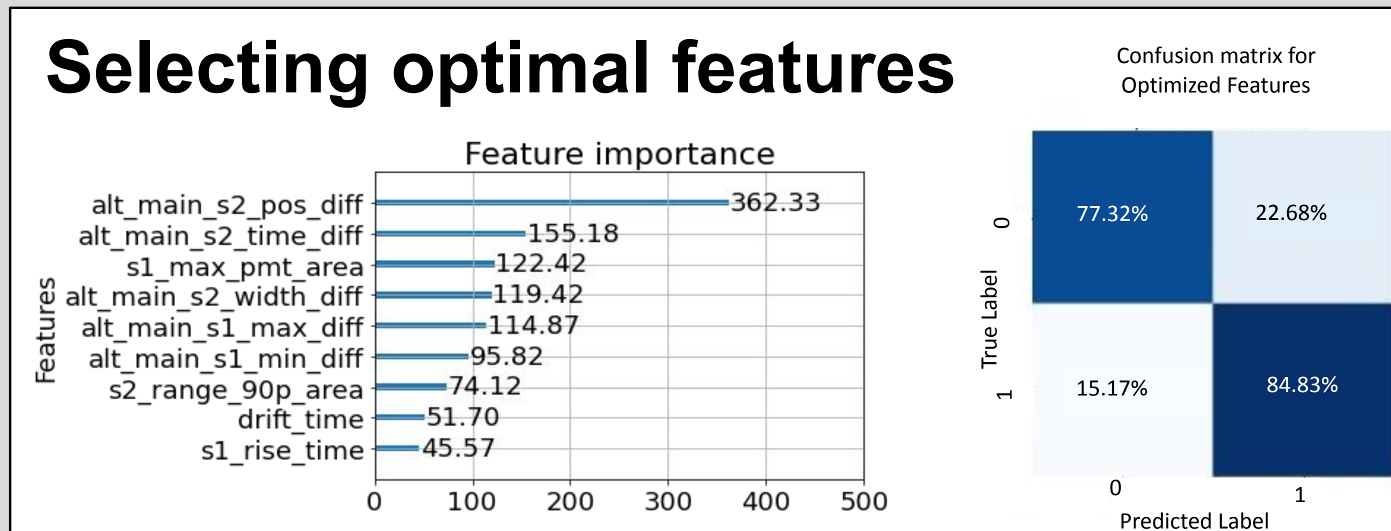
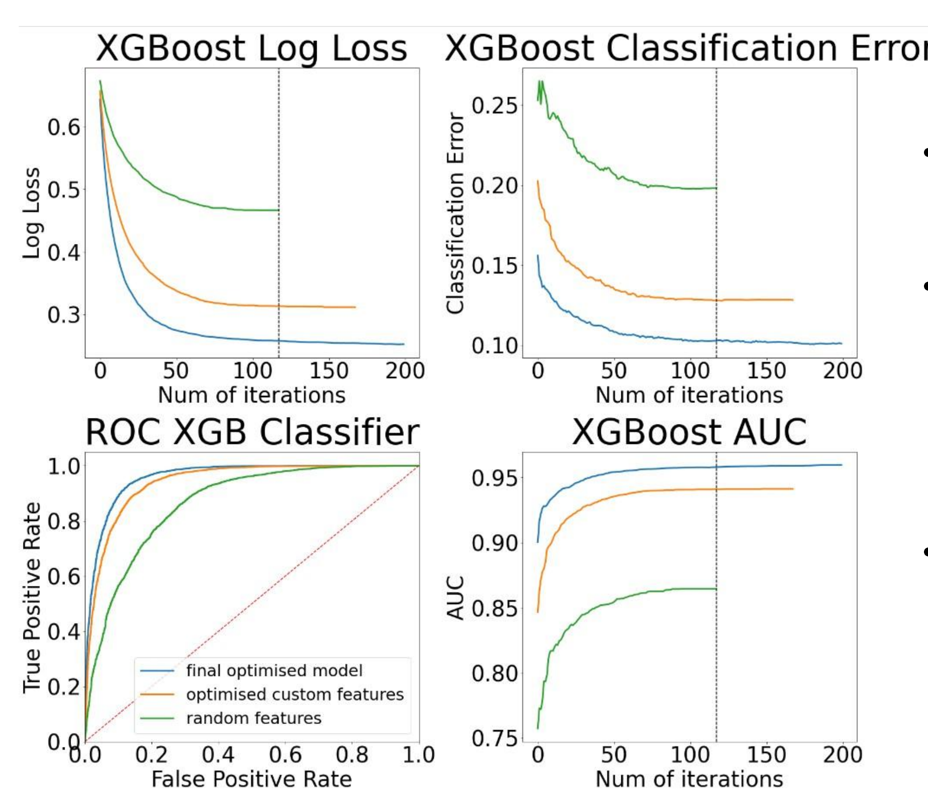


Figure 4: The most significant features for classifying signal and background events & confusion matrix

Hyperparameters optimization



- Algorithm maximized testing accuracy
- Penalized > 2.5% difference in training testing accuracy to avoid overfitting / underfitting
- grid search optimized hyperparameters for score improvement.

Figure 5: Difference in performance of XGBoost on various metrics with a) random features, b) optimized custom features and c) final optimizations and features.

Results

Accuracy: A tuned XGBoost was able to achieve approximately **90%** accuracy in classifying the signal and background in the test and validation datasets. The difference between the models' accuracy in training and testing datasets was less than 2.5 %.

On the tuned data the accuracies of all the models were:

	XGBoost	Log Reg	RNN	Seq
Accuracy	**90%**	81%	72%	78%

Accuracies are just a single metric. The NN models were unstable, however, they performed well on the log loss and AUC metrics. Log reg underperformed on the log loss metric having a value of 5 compared to the other models which score sub 1.

Customizing features and optimizing hyperparameters each boosted the model's accuracy by 2-3%.

To interpret our model we plot the predictions in our simulated data in different features parameter space.

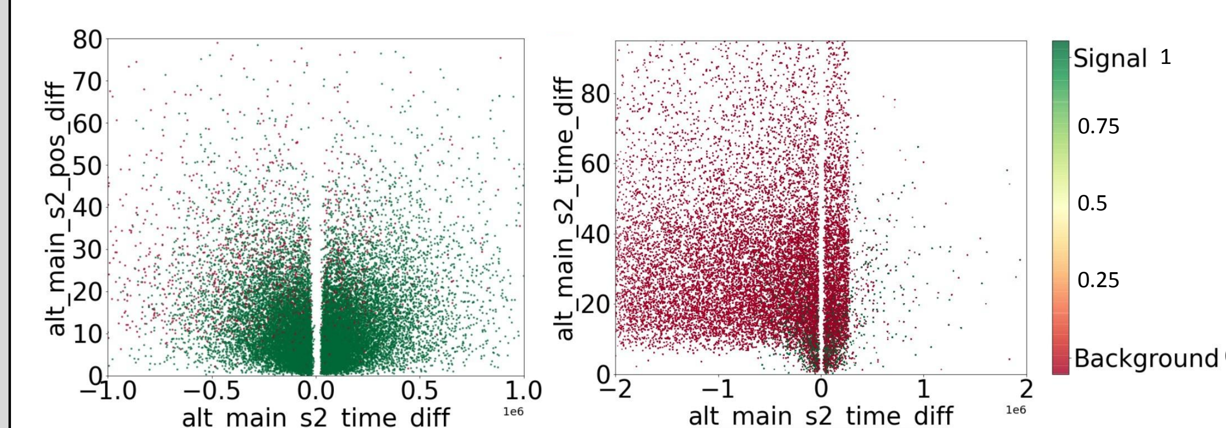


Figure 6: Left: Simulated signal data, Right: Background (AC) data
Green points represent data predicted as signal
Red points represent data predicted as background

Discussion & Conclusion

- XGBoost seems to be the best performing model for our task. However, there is still room for improvement.
- **Limitations:**
  - The simulation data is not perfect. The amount of data may not be enough to accurately train some ML models.
  - Furthermore, the dataset containing background events is data driven and is likely to contain some signals as well.
  - Thus, reaching higher accuracies may be unfeasible with the current data.
- **Future steps:**
  - Improving the quality of our data
  - Manually creating more data using methods like Gaussian noise

Enhancing signal vs. noise classification accuracy is a **critical step forward for more precise dark matter detection.**

Acknowledgement: We would like to thank our mentors Husheng Guan, Riya Singh and Sayan Ghosh, alongside with Prof Andreas Jung, for their constant support and feedback throughout our project.