Pulse-type classification for the Large Underground Xenon dark matter search

Kelly Stifter

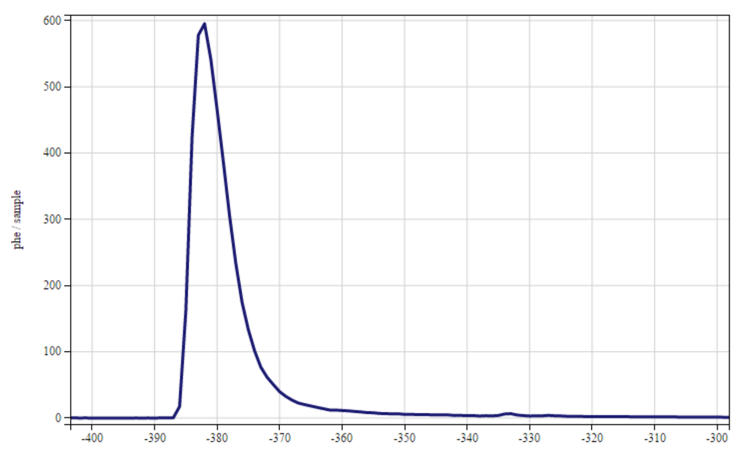
Problem Description

Goal: Develop a classifier to identify different pulse-types in data from the Large Underground Xenon (LUX) experiment.

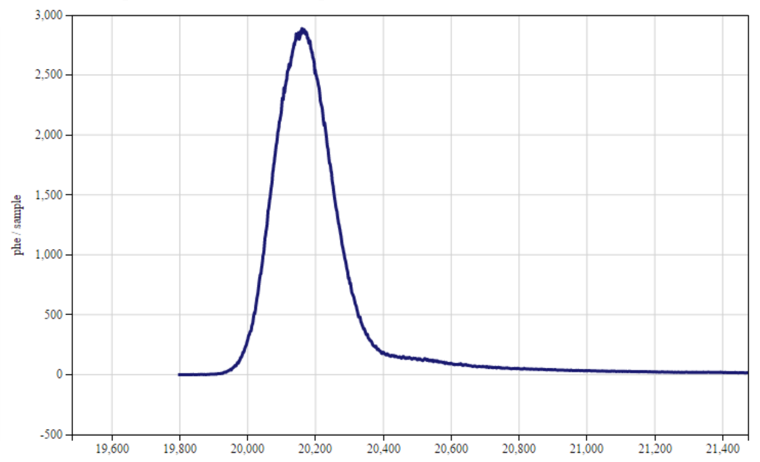
Motivation: Dark matter searches often rely on the ability to separate two main event types. The presence of specific pulse types helps identify these distinct event types. Since dark matter interactions are thought to be very rare, events and pulses must be predicted with a high level of accuracy.

Main pulse types:

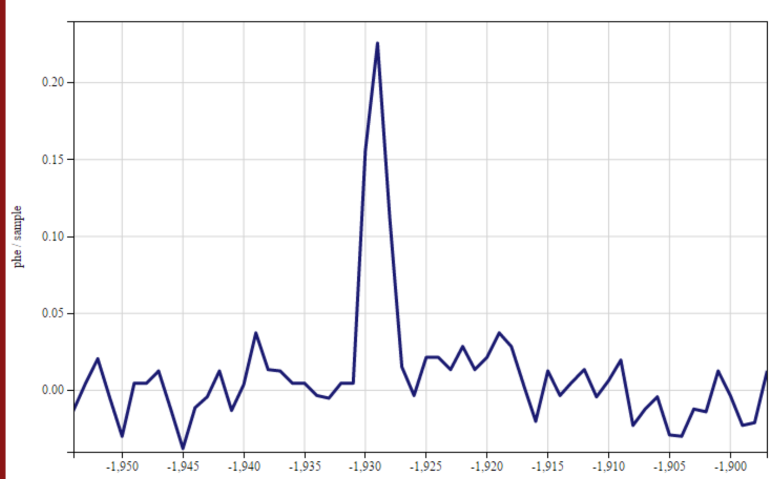
S1 (class 1)



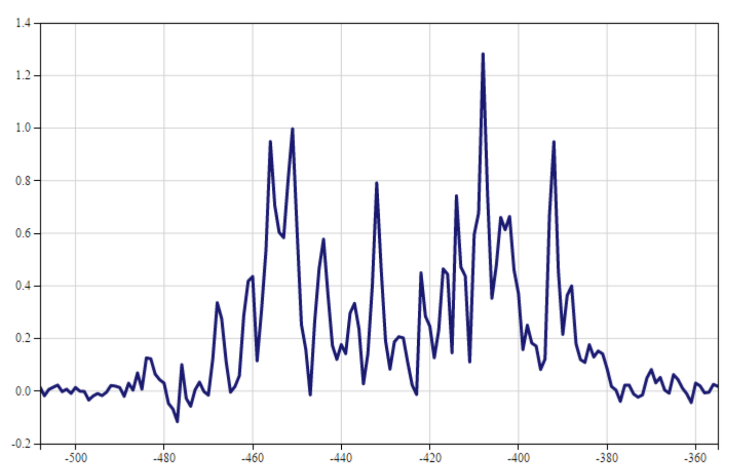
S2 (class 2)



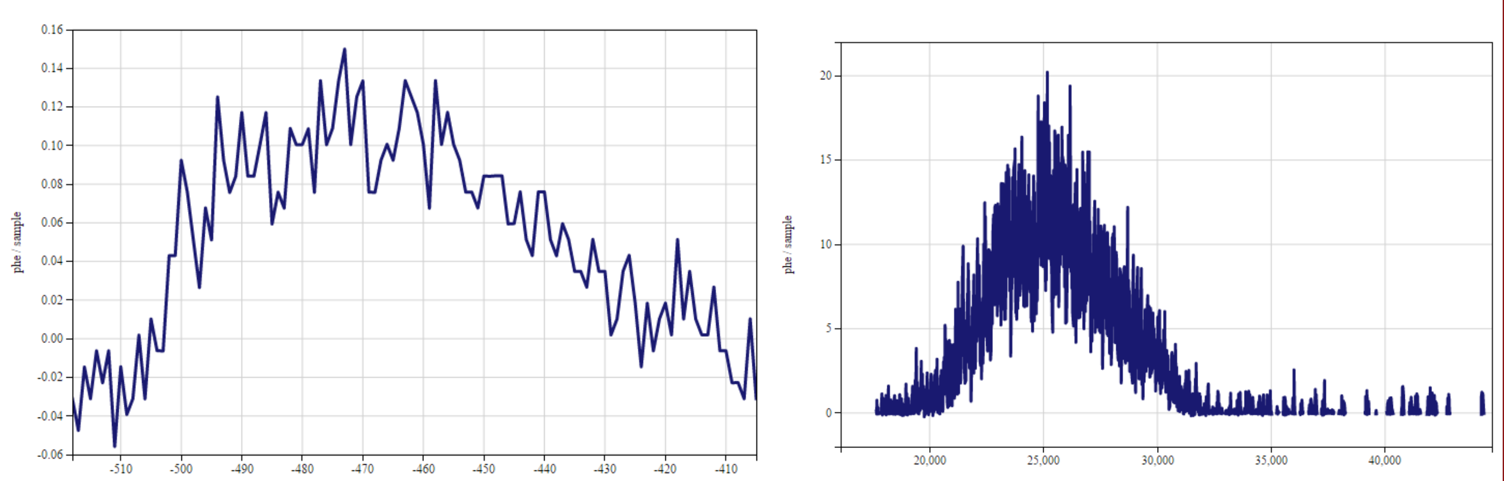
Single photon (class 3)



Single electron (class 4)



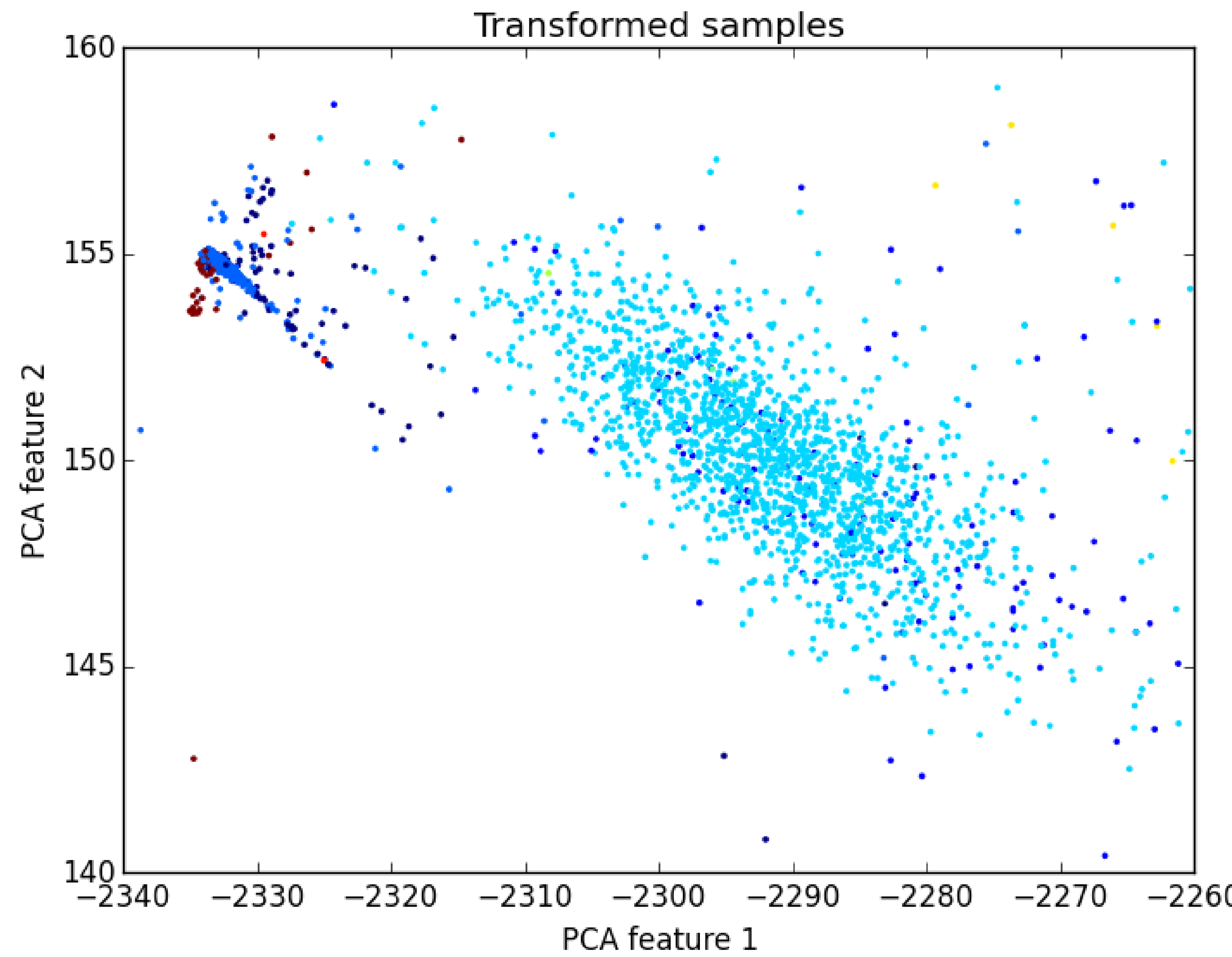
Noise/other



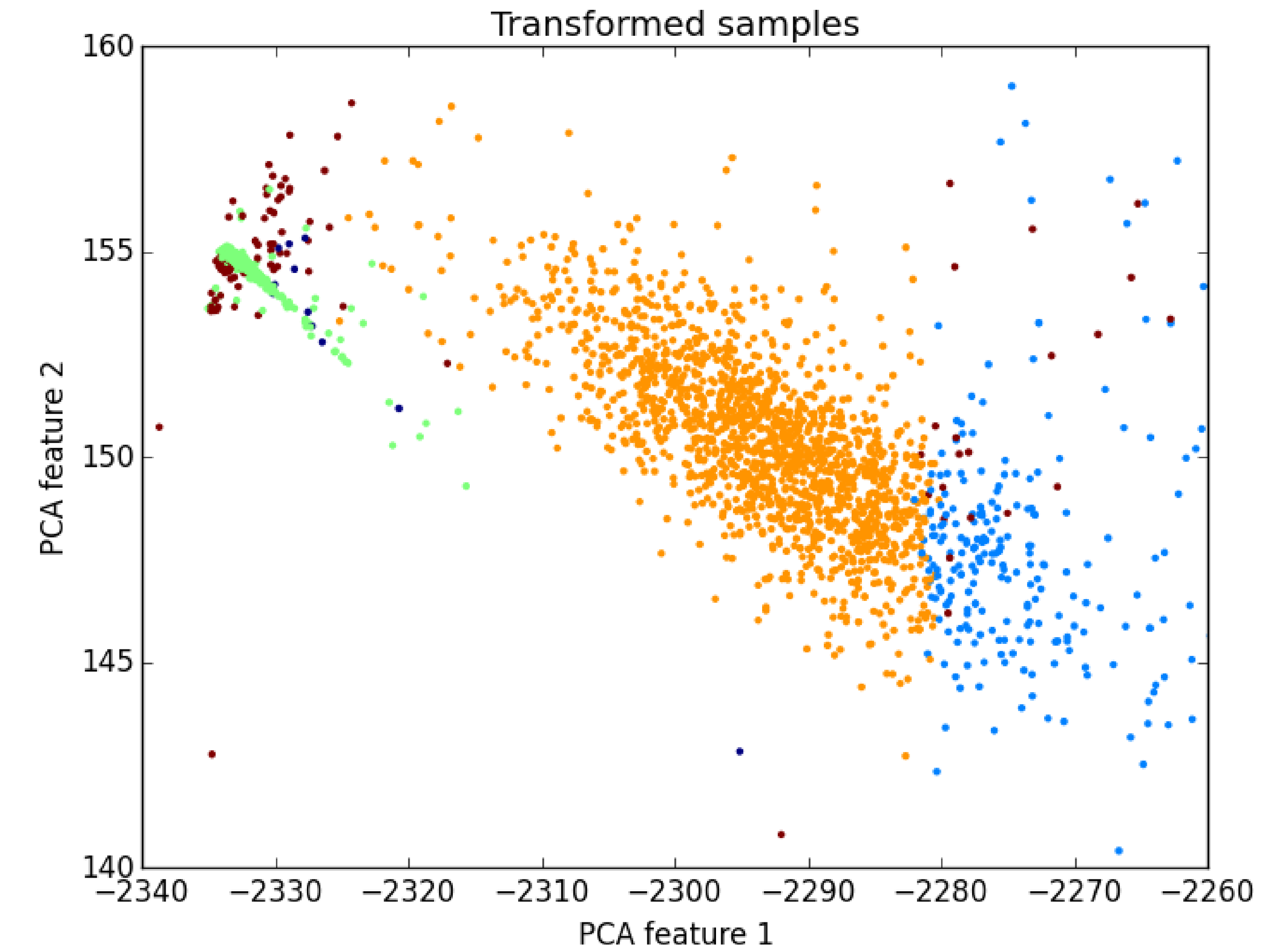
Features: 50 features, such as pulse width, area, etc., available for each pulse waveform, 42 evaluated through feature selection, ~10 selected for each pulse-type classifier.

Data Set

Hand scan: Classification of ~4,000 pulses, as determined by manual examination by LUX scientists. Total of 11 classes (5 existing, 5 new, 'I don't know').



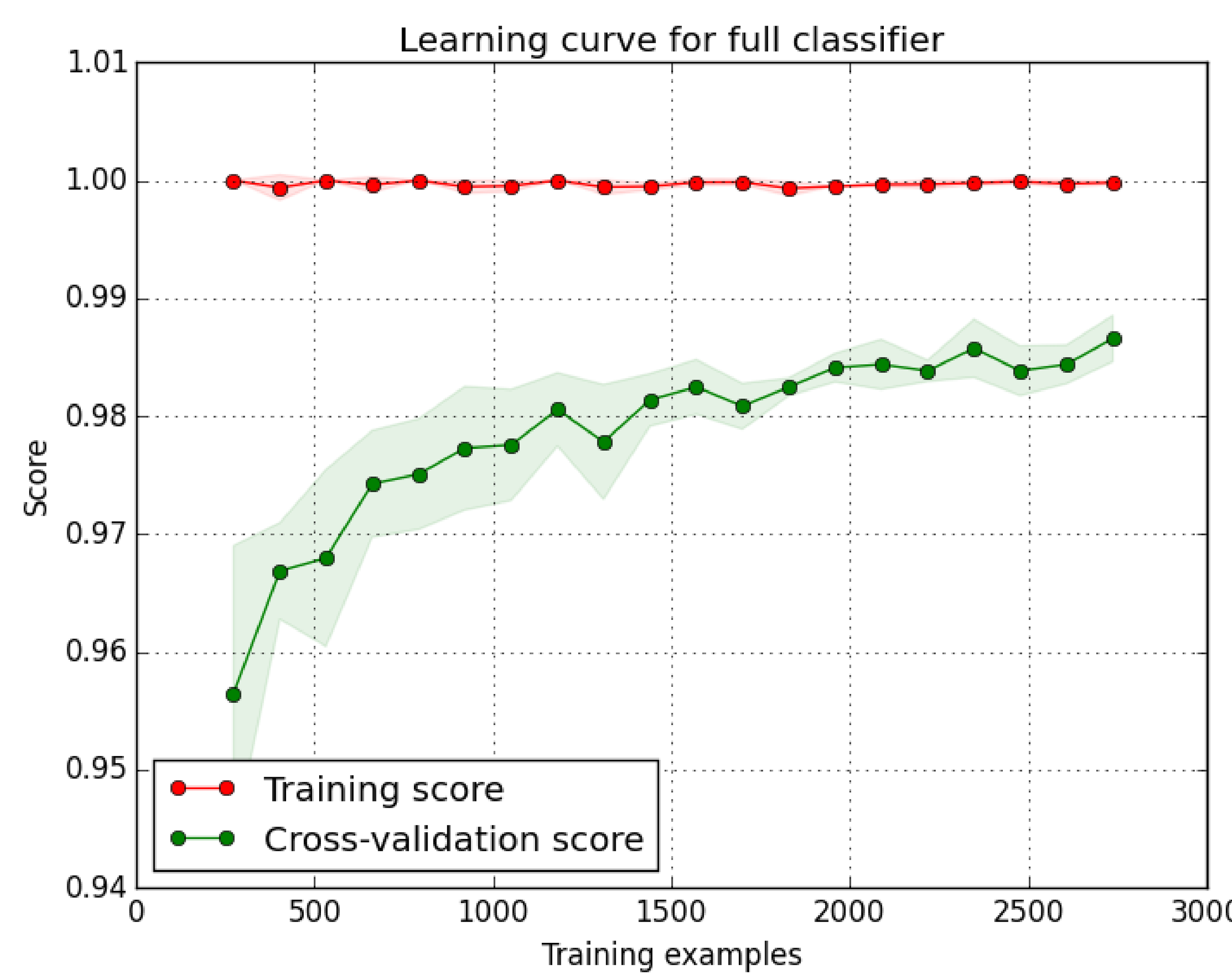
LUX classification: Classification done by the LUX analysis, based on simulation data. Total of 5 classes. Used in training classifier.



Final Classifier

Structure: Train four 'expert' random forest classifiers for classes 1-4. Anything not placed by the experts is assigned to class 5. If sample is assigned to both 1 and 3, override and assign it to 1. If sample is assigned to both 2 and 4, override and assign it to 2. If sample is assigned to other pair, any set of three, or all four classes, override and assign it to 5.

Performance: 4-fold cross-validation score of ***(98.5 ± 0.2)%***



Typical confusion matrix:

$$\begin{bmatrix} 51 & 0 & 0 & 0 & 5 \\ 0 & 110 & 0 & 0 & 0 \\ 0 & 0 & 218 & 0 & 4 \\ 0 & 0 & 0 & 312 & 2 \\ 0 & 3 & 2 & 0 & 24 \end{bmatrix}$$

Above: Confusion matrix indicates expert classifier 1 is weak link.

Left: Learning curve indicates more data would improve accuracy.

Conclusions

- Hand scan not reliable, ~80% accurate
- More features needed to successfully implement additional classes
- Random forests generally perform better than SVMs with this dataset
- Highest-scoring features are often unexpected, provide insights for physicists
- Providing probabilities from each classifier will contribute to event classification

Future work:

- Apply regularization, ensure classifier is not over-fitting
- Obtain more reliable hand scan data (planned for near future)
- Build pulse classifier based on waveform
- Expand and build full event classifier

Acknowledgements

Many thanks to James Morad for providing the hand-scan dataset and for his help with all things LUX-related, and to Mitchell McIntire for his machine learning expertise and helpful insights.