Machine Learning for Data Interpretability and Visualization Solutions

Katie Fotion[1], Justin Burau[2], Abdullah Rashed Ahmed[3], and Ryan Coffee[3]

[1]San Francisco State University.
[2]University of Washington.
[3]Linac Coherent Light Source, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA
+Contact: coffee@slac.stanford.edu

Introduction

With the approaching launch of LCLS-II, there is an urgent need for rapid data processing without loss of interpretability. Due to the scarcity of beam time and the black box nature of machine learning (ML), the application of ML to LCLS data is considered by some to be unreliable.

This project serves as an example of machine learning being carefully and effectively applied to the time tool to solve a common problem surrounding interpretability of data. The motivation is to prepare for the upcoming transition to the 1 MHz LCLS-II repetition rate, while enhancing overall understanding of data.

Keywords: machine learning, random forest, variable importance, time tool, data visualization

Research

Background

Predict the instant x-rays collide with the sample on LCLS-II

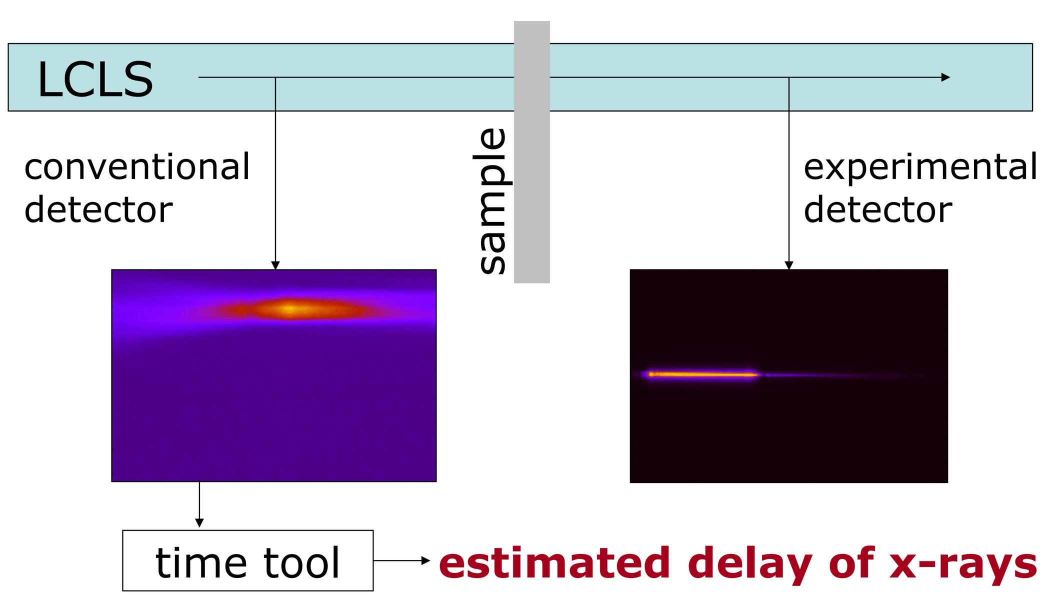

Figure 1. Diagram of current time tool setup with resolution image outputs.

Since the repetition rate of LCLS-II prohibits processing data from the conventional detector, we want to simulate the relative delay prediction given a signal from the experimental detector using ML, remotely training the model prior to an experiment, and then deploying on FPGA near the detector to enable rapid prediction speeds.

Figure 2. Flowchart of initial process.

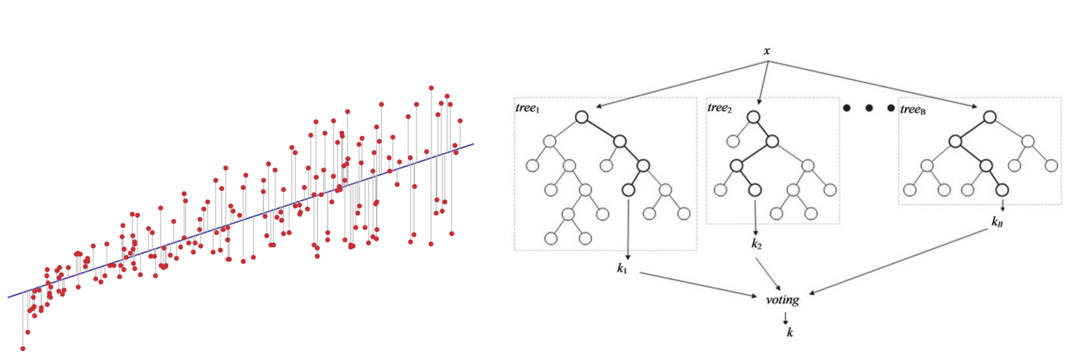

Figure 3. Linear regression[1] (left), random forest[2] (right). [1]https://i.stack.imgur.com/GXj8T.png [2]http://file.scirp.org/Html/6-9101686/f799e10c-50bd-48ec-9344-49d767083be5.jpg.

Model Selection

And the winner is: Random Forest, with a mean R^2 value of **0.9643** across runs of the same material.

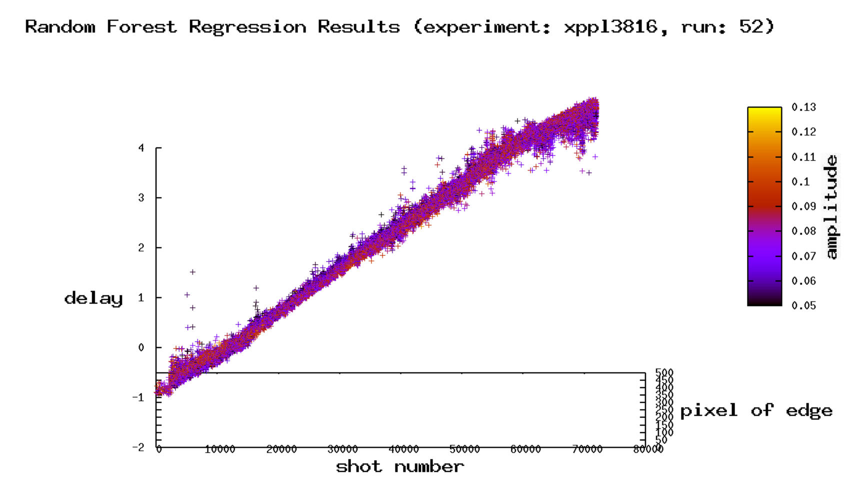

Figure 4. Shot number vs. pixel of detected edge vs. predicted delay, demonstrating results consistent with physics.

Feature Engineering

Determine the most effective form of input data to optimize speed without loss of performance and interpretability

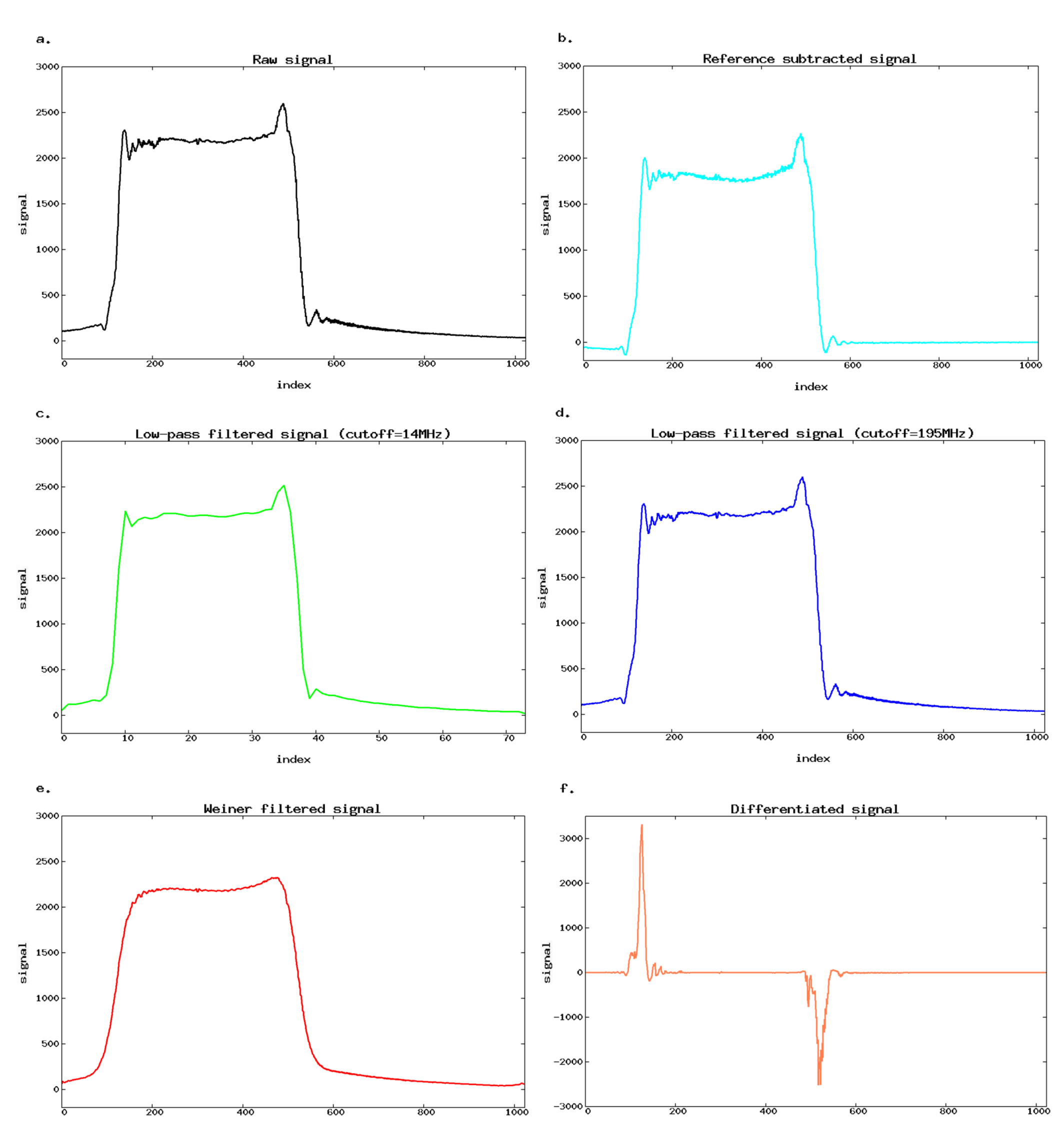

Figure 5. Comparison of a. raw signal, b. reference subtracted signal, c. low-pass filtered signal with frequency cutoff of 14MHz, d. low-pass filtered signal with frequency cutoff of 195MHz, e. Weiner filtered signal, and f. differentiated signal.

Determining the most informative features not only improves the model, but also teaches us about the nature of the data itself.

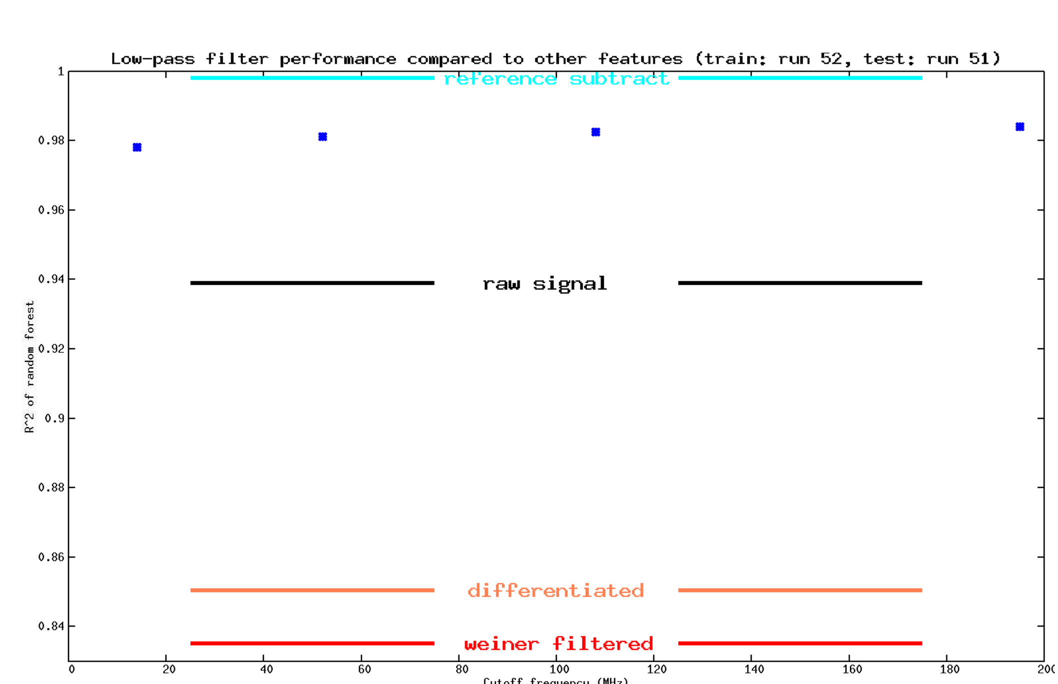

Figure 6. Feature engineering results demonstrated that analog features produced optimal, attainable results.

Results Visualization

Intuitively display model's reasoning in order to build user trust and prevent prolonged errors

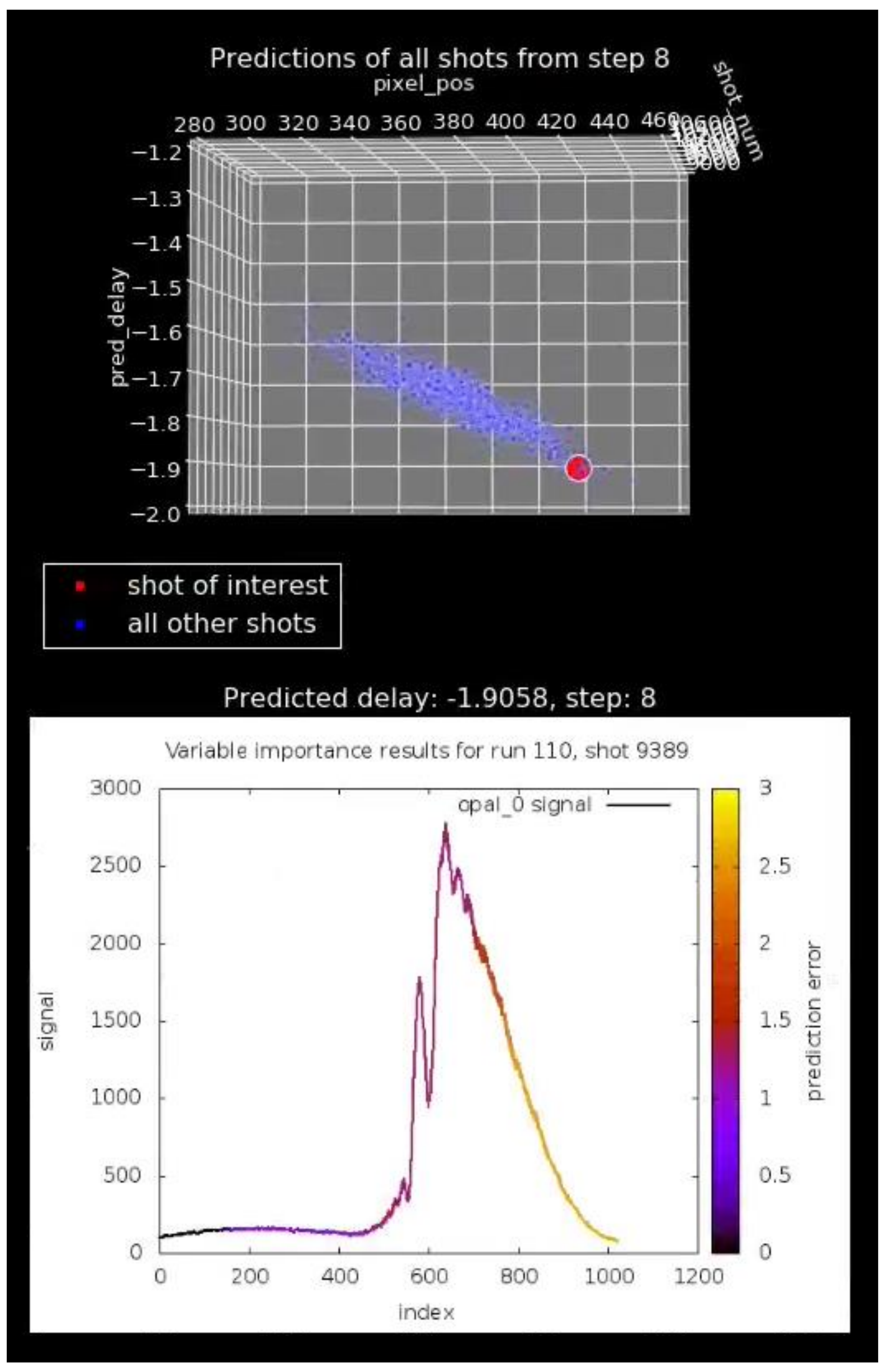

Figure 7. Variable importance visualization tool, which allows users to monitor the model with the option to accept or reject predictions.

Conclusion

This work exhibits that machine learning models can produce comparable results to algorithmic approaches, while improving compute speed and enhancing interpretability. This work will be continued during Fall 2017 to further both the time tool efforts and the machine learning initiative at SLAC as a whole.

Future work:
•Adapt variable importance measure to take variable interaction into account
•Further extend visualization capabilities to 3D and include parallel coordinate representation of machine parameters
•Apply similar strategies to various other problems at SLAC

Acknowledgements

Use of the Linac Coherent Light Source (LCLS), SLAC National Accelerator Laboratory, is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515.

Date: 08/09/2017