Improving Training Labeling for Machine Learning Algorithms

Using KULL Hydrodynamic Simulation Data (BubbleShock)

Joanna Held (University of Iowa), Cory Lanker (LLNL), Ming Jiang (LLNL)

DATA SCIENCE SUMMER INSTITUTE

Lawrence Livermore National Laboratory

ABSTRACT

Predicting conditions which lead to simulation failure generated using the Arbitrary Lagrangian-Eulerian (ALE) method allows users to take measures to avoid failure. In order to correctly predict these situations using machine learning algorithms, it is necessary to develop training labels which capture relevant information for classification.

PROBLEM DEVELOPMENT

Our aim is predicting the mesh elements that are at risk of crashing code for hydrodynamic simulations.

- Our data structure: 11,560 zones over 1000's of cycles.
  - **Zones** are geometric sections of the mesh. (Fig. 3)
  - **Cycles** capture snapshot-like progression of simulation but are not linearly related to time. (Fig. 3)
  - An **observation**, or zone and cycle combination, consists of the information in Fig. 1.
- Mesh flows with simulated material motion and zone geometry may reach impossible values (e.g. Maximum Angle > 180º) leading to code failure (Fig. 2).
- Prediction of zones that will cause code failure allows for compensatory measures (e.g. mesh relaxation).
  - Code failure prediction algorithms are trained on zone geometries from 100's of simulations using many different environments.
  - Algorithms require training labels for enormous amount of data.

Proper labeling of training data is necessary for creating successful prediction models.

- Label = 1: a zone that is a failure risk at a certain cycle (an observation).
- Label = 0: a zone not posing a failure risk at a certain cycle.

Metric	Metric
Area	Scaled Jacobian
Aspect Ratio	Shape
Condition	Shape and Size
Distortion	Shear
Jacobian	Shear and Size
Maximum Angle	Skew
Minimum Angle	Stretch
Oddy	Taper

Fig. 1: Information included in an observation for each zone and cycle.

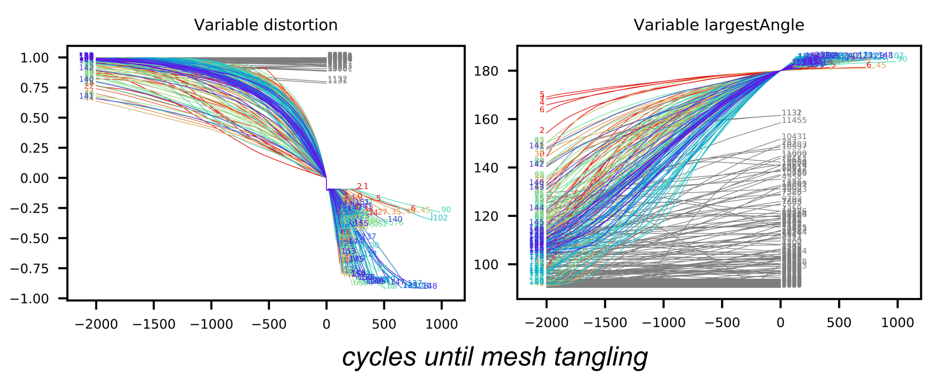
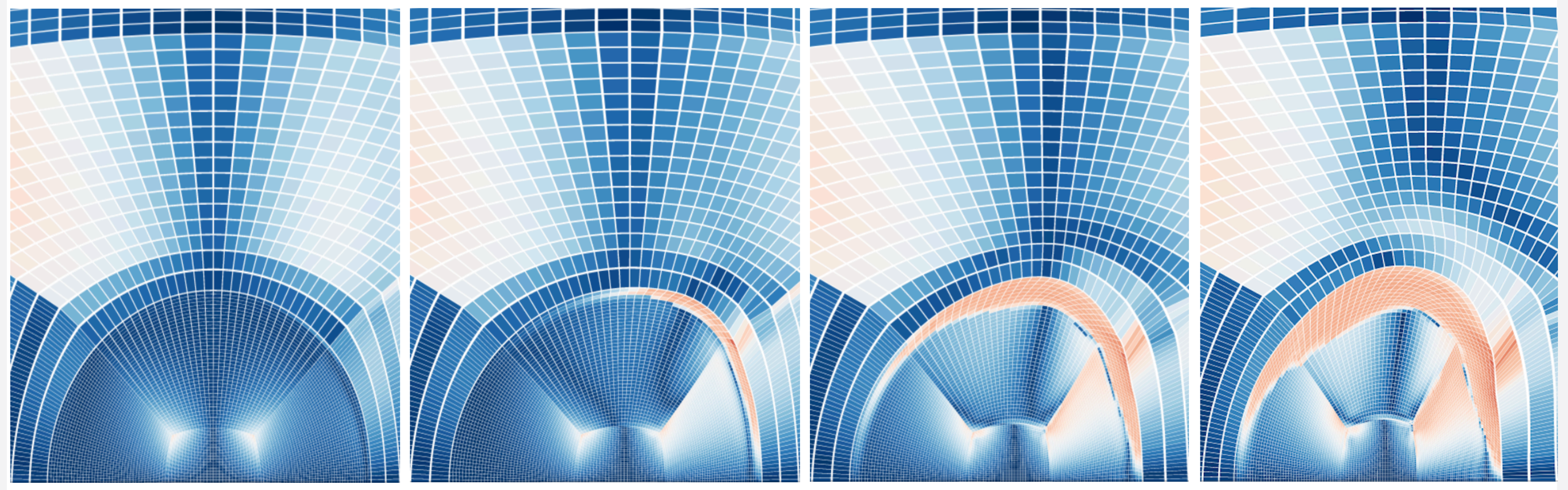


Fig. 2: Behavior of variable metrics for failed zones (in color) over cycles as they approach failure.



Cycle 4000
t = 44.10 μs

Cycle 4500
t = 48.40 μs

Cycle 5000
t = 51.25 μs

Cycle 5500
t = 52.80 μs

Fig. 3: BubbleShock central region at different points in time (cycle number). Colormap shows fluctuation of largest angle variable.

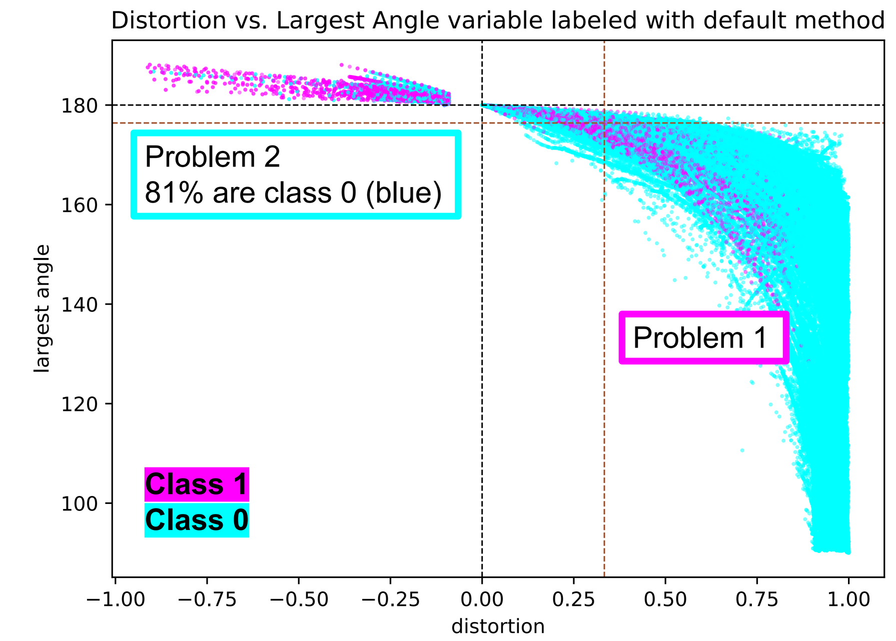
This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 LLNL-POST-782600.

A SIMPLE LABELING METHOD

Currently, observation labels are based solely on whether a zone causes code failure (1 = cause failure, 0 = did not cause failure), and this simplicity creates mislabeling.

- Only observations within n = 650 cycles until failure are assigned label = 1.

Figure 4 (right):
- Class 1 observations further from the failure point resemble class 0 and are mislabeled. Class 0 observations after point of failure should be class 1.
- The left group should be all class 1 (pink), and the right should be all class 0 (blue).



Problem 1. Cycles of failed zones far from failure should not be labeled 1.

- Labeling early causes the simulation to relax the mesh unnecessarily.
- Relaxation of mesh causes loss of physical accuracy of simulation.

Problem 2. There are label = 0 that are code failure risks.

- Labeling too late means the simulation did not relax the mesh soon enough to avoid mesh tangling.
- Mesh tangling can halt simulation progress.

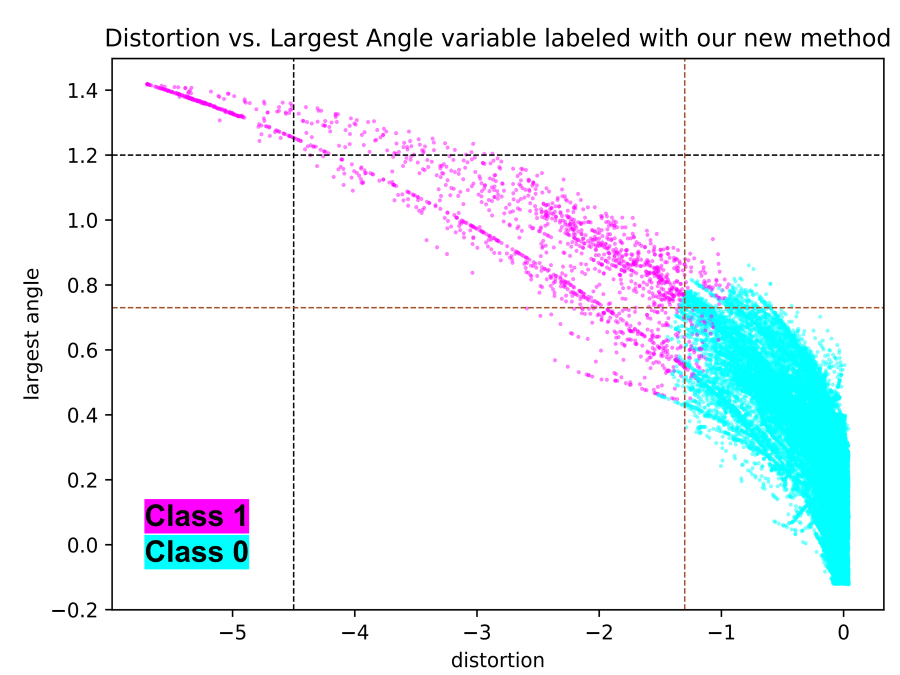
OUR IMPROVED LABELING METHOD

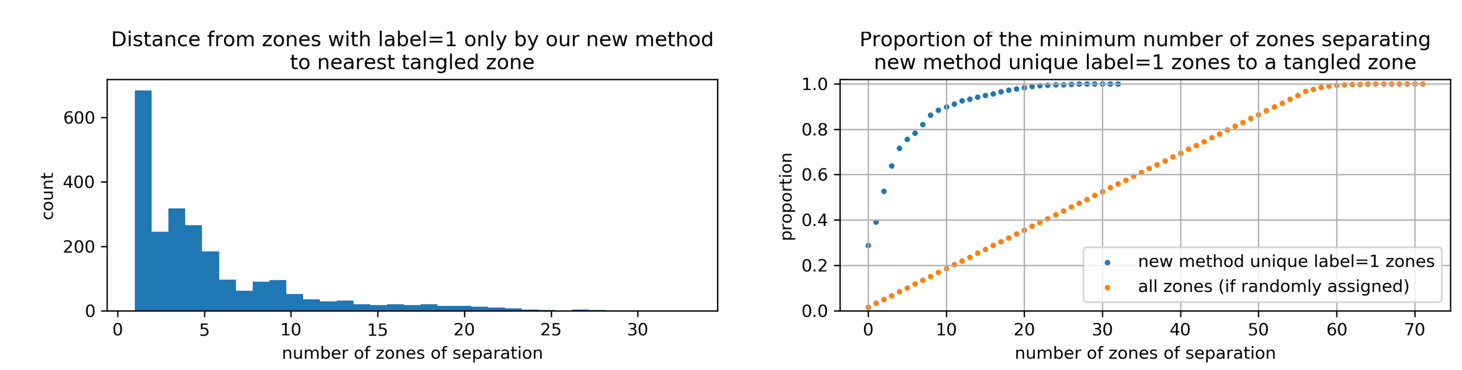
Instead of applying label = 1 to only zones that crash the code, base labeling on a distance metric of transformed variables.

- Linearize the 16 features with likelihood of mesh tangling.
- Scale variables according to influence on mesh tangling.
- Assign label = 1 to observations with this distance metric > threshold.

Figure 5 (right):
- Clearer separation of classes.
- Minimizes both problems 1 and 2 as explained in Fig. 4.
- Black and brown bars show transformed scale.

Figure 6 (below):
- Distance of observations labeled 1 by our new method from tangled observations.
- Our new method labels observations close to known fail risks as also a potential fail risk.





PREDICTION RESULTS

Obtained with Random Forest Classification

- Training data: observations sampled from speed=120, density=0.18 simulations.
- Testing data: observations from speed=160, density=0.22 simulations.
- Classification based on 400 fully-grown trees using 0.50 threshold.

A classifier using data with new labeling method (bottom row) improved class discrimination (left column) and prediction capabilities (right column)

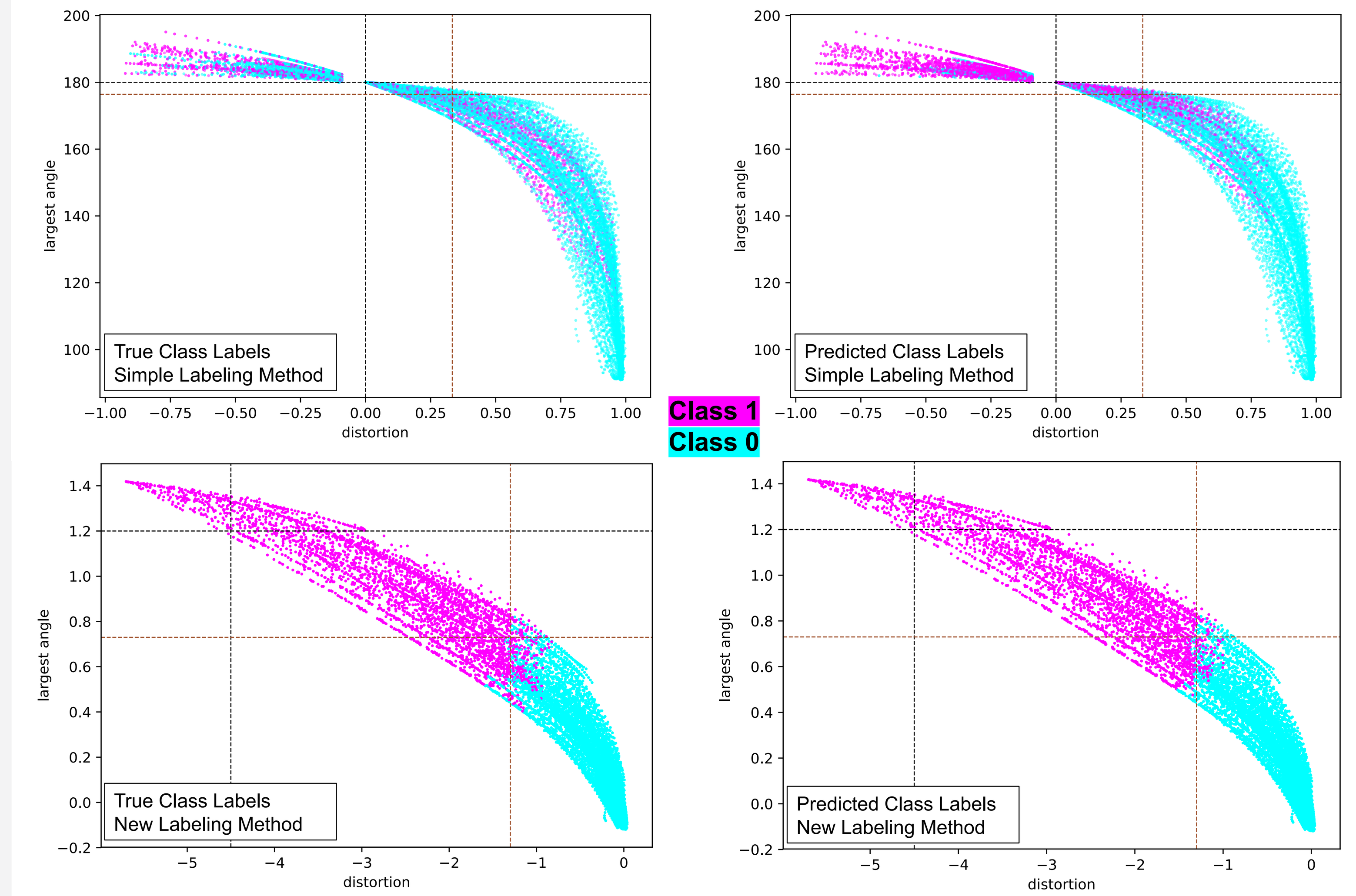
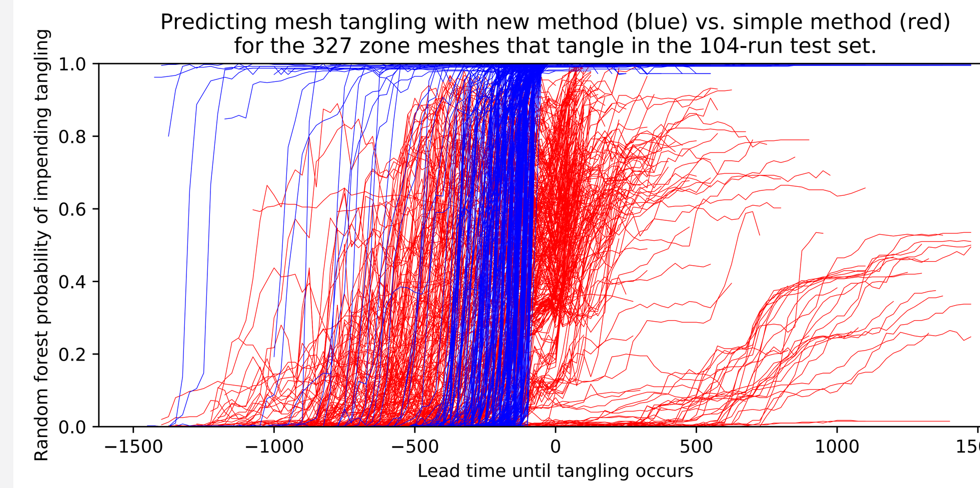


Fig. 7 (above): Left column displays true labels for test set by the default (top) and new (bottom) method. Right column shows predicted labels by each method respectively.



	simple	new
Predicted with at least 25 cycle lead time	51%	98%
Predicted at time of tangle event	0.5%	2%
Predicted after event	23%	0
Never predicted	25.5%	0

Fig. 8 (above): Predicting zones which are known to have tangled by the simple and new method. The table shows proportion of predictions at various benchmarks for both simple and new methods.

CONCLUSIONS

- Training labels that capture more information about classes improve performance in machine learning classification predictions.
- Future work involves exploring different transformation distance metrics.
- Additional computation is a potential limitation. However, once a framework is built, the new labeling takes little computing.
- Training labels that do not clearly represent a class introduce a bias that mars the ability for an algorithm to properly predict future observations.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.