Competence Measure Enhanced Ensemble Learning Voting Schemes

Francesca McFadden, freale1@umbc.edu

DATAWorks 2025, Alexandria, VA

Contributed Session 5C: Advancing T&E of Emerging and Prevalent Technologies / Improving Quality of T&E

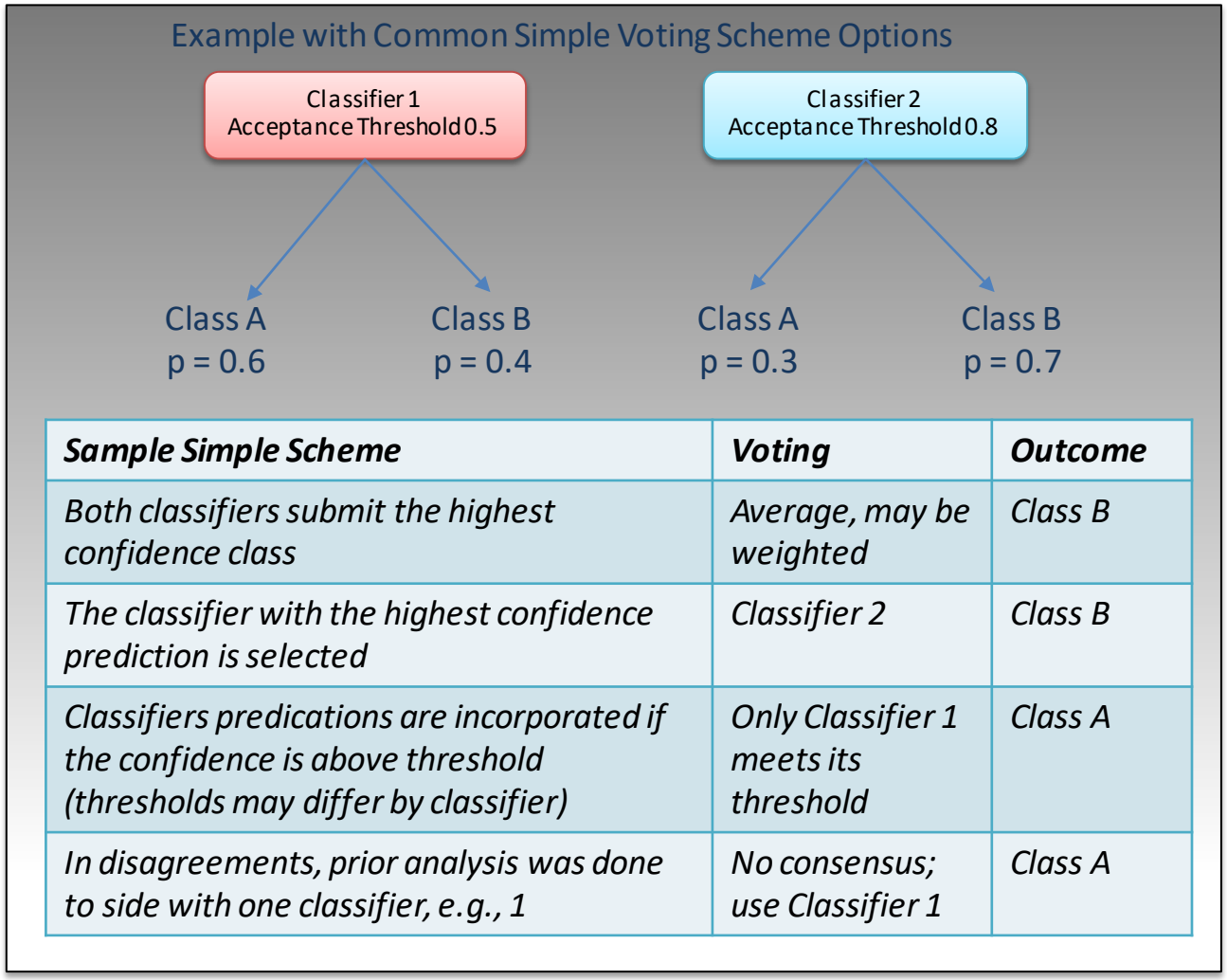
24 April 2025

Ensemble Learning Methods

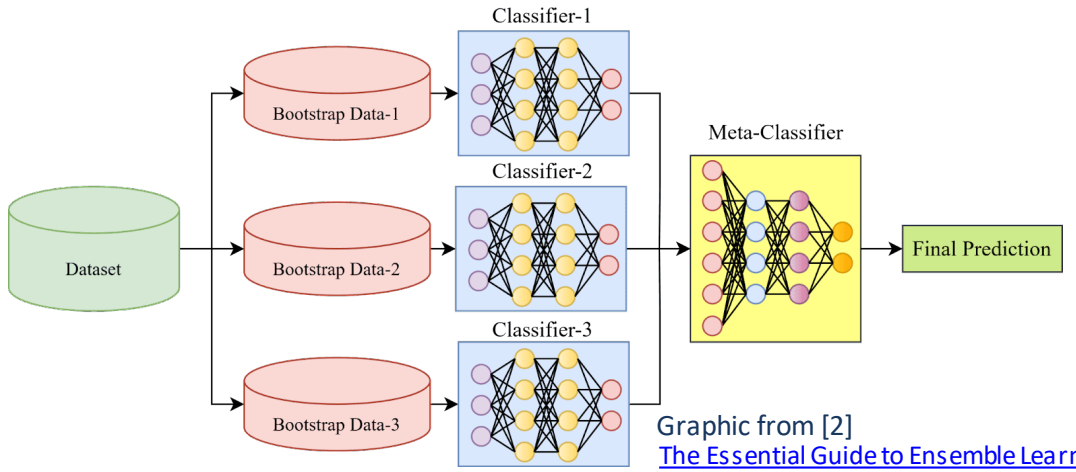
- Ensemble learning methods use the predictions of multiple classifier models.
  - A well-formed ensemble should be formed from classifiers with various assumptions, e.g., differing underlying training data, feature space selection, and therefore decision boundaries.
- A voting scheme is used to weigh the decisions of the individual classifier models to determine how they may be combined, fused, or selected among to predict class.
  - Voting schemes often consider individual reported classifier confidence in predictions.
- Complementary features, class representation, and training data distribution across the classifiers are to an advantage, but are not being fully exploited with existing schema.
- Network approaches attempting to learn the complementary traits of classifiers may result in loss of explainability to end users.

We show an approach for enhancing ensemble learning performance through integration of model competence measures in a simple voting scheme, exploiting the complementary traits of classifiers while preserving explainability to end users.

Voting Schemes



Sample Simple Scheme	***Voting***	***Outcome***
Both classifiers submit the highest confidence class	*Average, may be weighted*	*Class B*
The classifier with the highest confidence prediction is selected	*Classifier 2*	*Class B*
Classifiers predications are incorporated if the confidence is above threshold (thresholds may differ by classifier)	*Only Classifier 1 meets its threshold*	*Class A*
In disagreements, prior analysis was done to side with one classifier, e.g., 1	*No consensus; use Classifier 1*	*Class A*



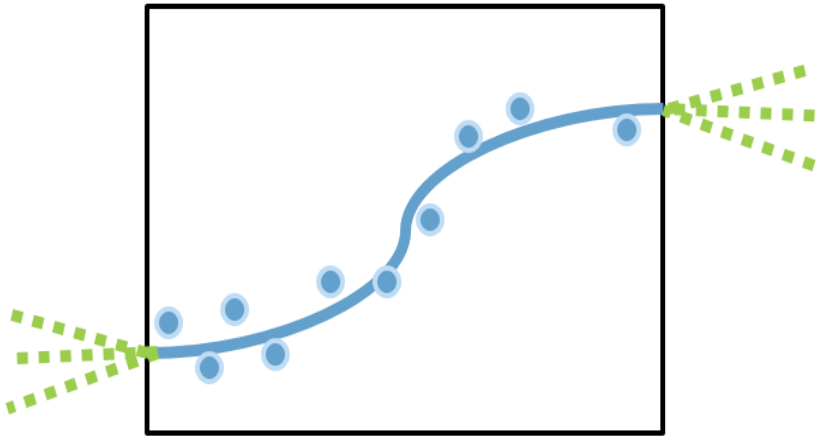
Graphic from [2]
The Essential Guide to Ensemble Learning

- More advanced approaches apply a training framework or network wrapping the individual classifiers to attempt learning where they are complementary [3]:
  - Form a polynomial decision boundary from the ensemble
  - Bagging parallel ensemble
  - Bootstrapping sequential ensemble
  - Stacked classifiers
  - Weighting in gating network and Fuzzy Ensembles
- The more advanced techniques may be able to learn complementary traits of the classifiers, but lose transparency in how the decisions are weighed to end users.

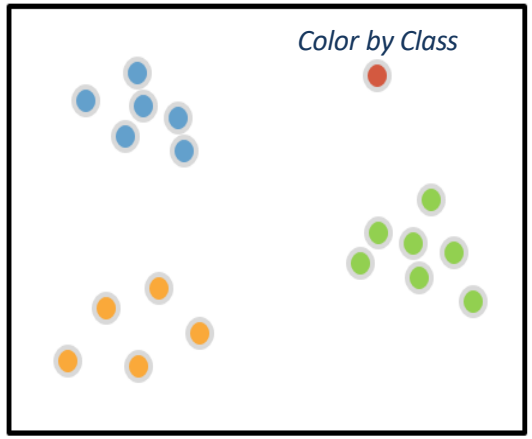
[2] Kondu, Rohit, The Essential Guide to Ensemble Learning, V7 Labs, 11 Jan 2024.
[3] Polikar, Robi, "Ensemble based systems in decision making," in *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21-45, Third Quarter 2006.

Why Estimate Model Competence

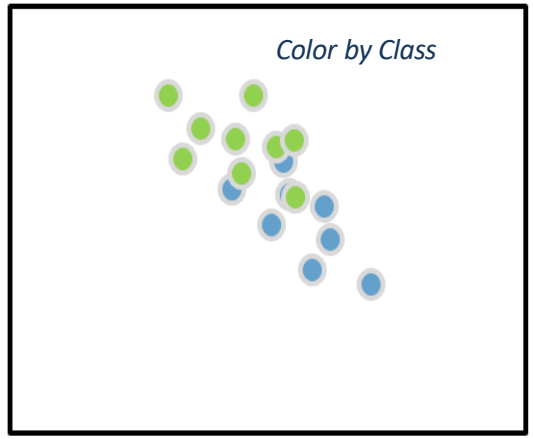
Distributional Differences



Class Representation



Feature Selection



Historically model confidence is used to estimate the effectiveness of a machine learning model's prediction. Model confidence is incorporated in existing voting schemes to weigh consensus of model predictions.

However, model confidence alone does not provide an indication where prediction of true class may be impacted by lack of representation in model training or possible class predictions.

Competence Measure Background

The Accurate layerwise interpretable competence estimation (ALICE) score [1] has distributional, model, and data uncertainty factors. The scores are compared to a threshold and the model is deemed competent for values above it. Both a correctness threshold and a risk threshold must be set based on the original definition, often requiring expert judgement.

$$p\left(\varepsilon\left(f(x), \hat{f}(x)\right) < \delta \mid x\right) \approx p(D \mid x) \sum p\left(\varepsilon\left(f(x), c_j\right) < \delta \mid c_j, x\right)\ p\left(c_j \mid x, D\right)$$

$f(x)$ true & $\hat{f}(x)$ predicted class of input x with $\delta > 0$ user set threshold

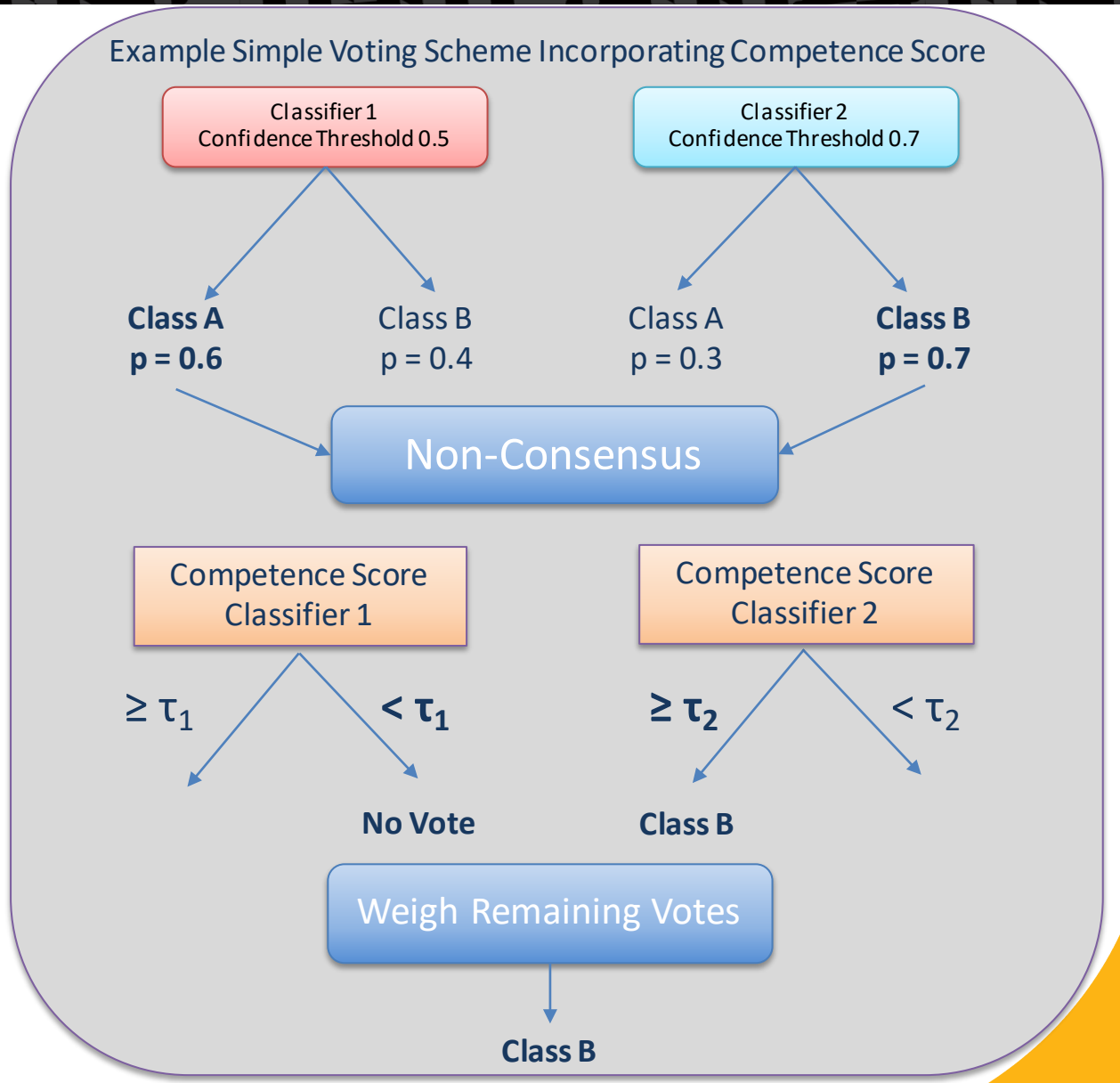
D is the set of all training data points, c_j is the one hot label per class

The ALICE score for an input x is an indicator of whether the model will be competent to predict the true class label of an input x. An in-distribution factor is incorporated. Consequently the score accounts for additional components that confidence does not. We will employ this method to estimate model competence in this presentation.

Reference: [1] V. Rajendran & W. LeVine. Accurate layerwise interpretable competence estimation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alch´e-Buc, E. Fox, R. Garnett, editors, Advances in Neural Information Processing Systems, vol 32, pgs 13981–13991. Curran Associates, Inc., 2019.

Approach

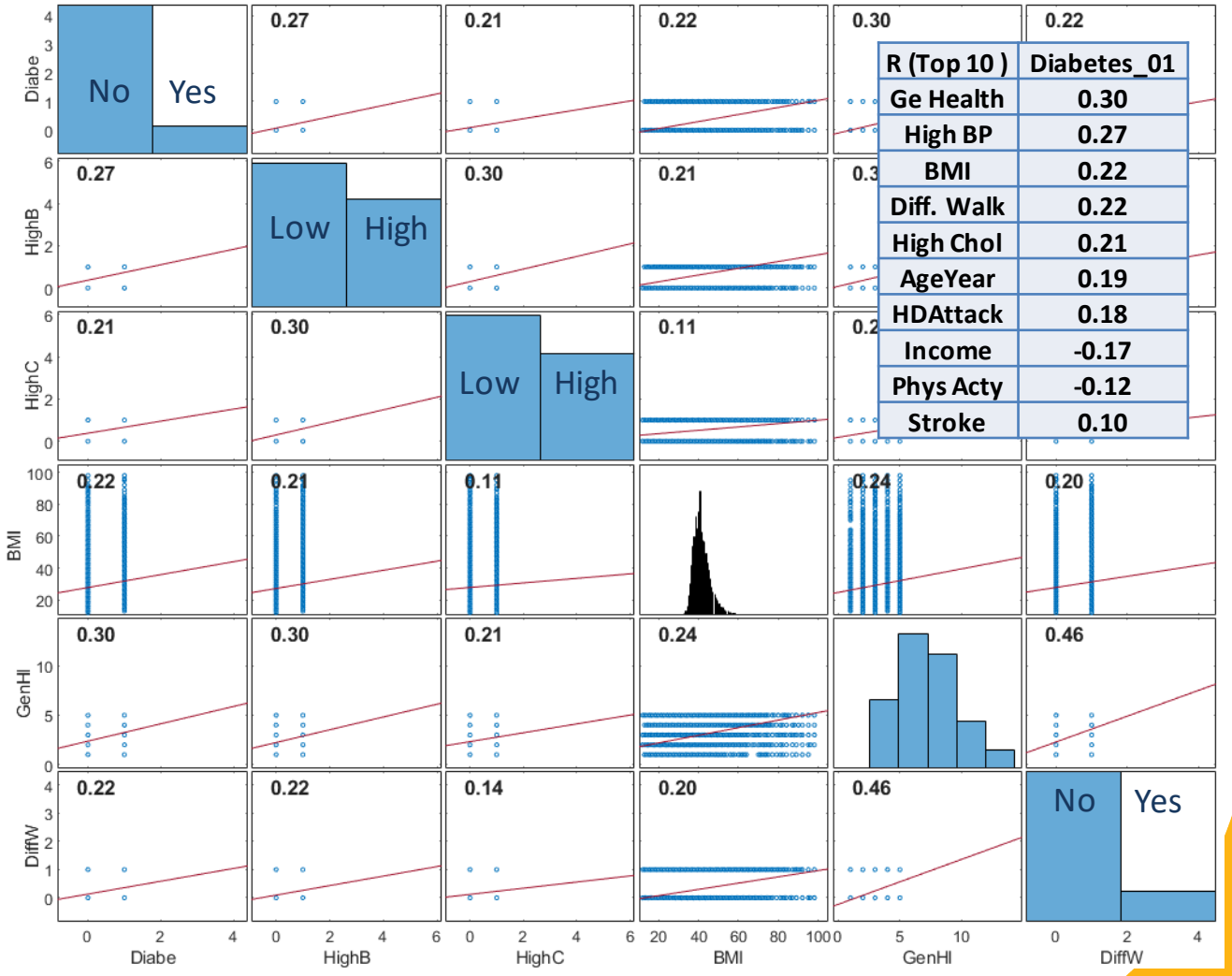
- The purpose of the described concept is to enhance current voting scheme approaches by integrating individual model competence measures
  - Ensures input data are appropriate to the prediction space of the individual classifiers
  - This approach appends confidence-based schemes with ensuring that inputs are consistent with the training data of the individual models.
- When there is non-consensus, consideration of the individual classifiers in the voting for the specified input will be based on achieving a threshold model competence measure.
  - If non-consensus remains after this filtering step, traditional single best source selection or averaging may be applied.
- These simple threshold filtering and averaging techniques maintain transparency in which classifier predictions are used and when filtering occurs to end users.



Diabetes Health Indicators Dataset

- Behavioral Risk Factor Surveillance System (BRFSS) 2015 survey [4]
  - Annual Centers for Disease Control and Prevention (CDC) survey Americans from all 50 states and 3 US territories on health-related risk factors, chronic conditions, and behaviors
  - Cleaned data set from Kaggle [5] was employed in the workflow
- 253,680 interviews with indication
  - no diabetes and/or only gestational (during pregnancy) diabetes (0)
  - prediabetes and/or diabetes (1)
- The data includes 21 features including a mixture of feature types with quantitative and qualitative responses,
  - binary, e.g., smoker or not,
  - integer, e.g., body mass index (BMI),
  - categorical scale, e.g., a general health score from 1-5; excellent to poor values

Top 5 Correlation Matrix to diabetes indicator –
high blood pressure, high cholesterol, BMI, general health, difficulty walking

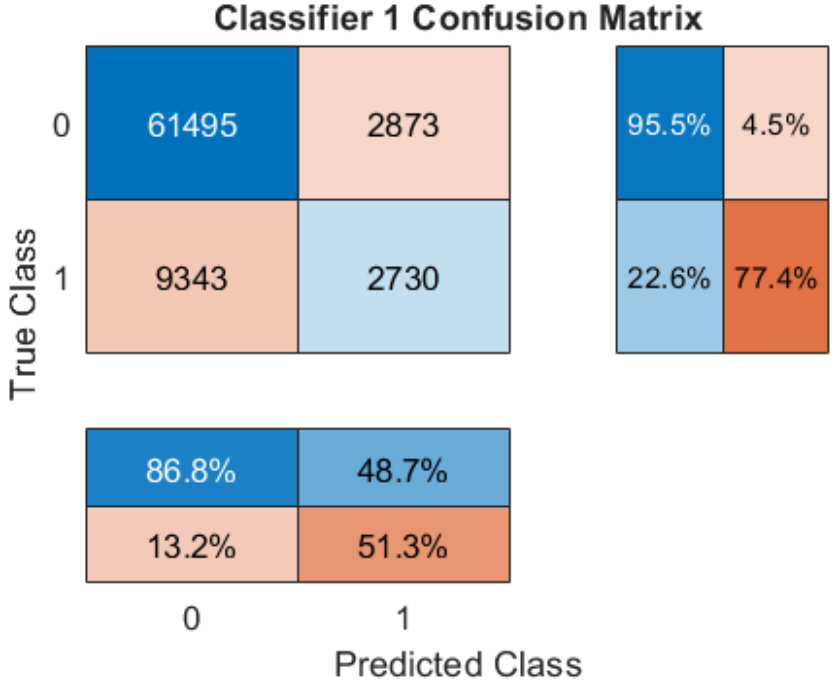


R (Top 10)	Diabetes_01
Ge Health	0.30
High BP	0.27
BMI	0.22
Diff. Walk	0.22
High Chol	0.21
AgeYear	0.19
HDAttack	0.18
Income	-0.17
Phys Acty	-0.12
Stroke	0.10

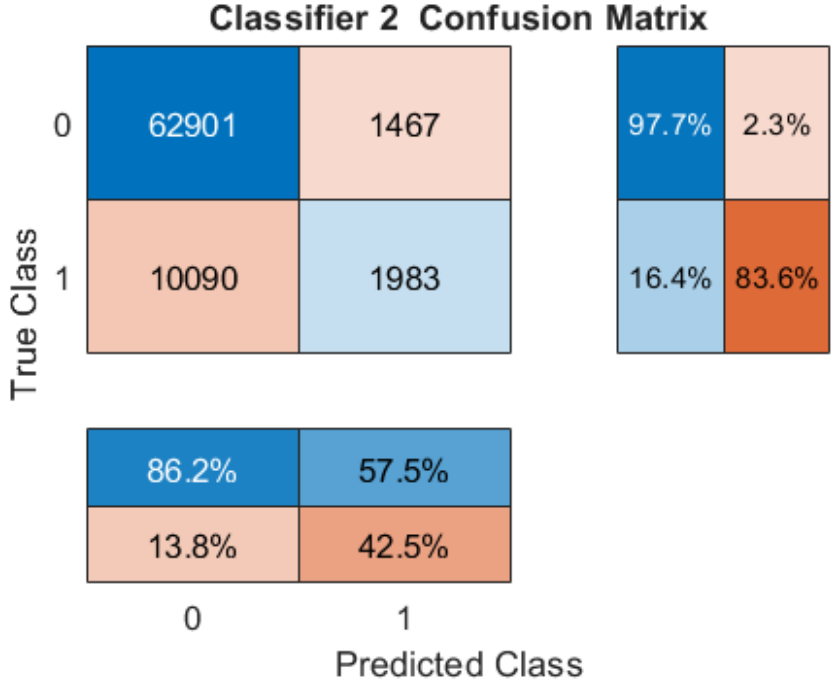
[4] Centers for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System Survey Questionnaire. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2015.
[5] Teboul, Alex. Diabetes Health Indicators Dataset, Kaggle, 2022.

Set of Classifiers

- Classifier 1 Random Forest (100 Trees)
  - Random 60% of data used as a training set
  - Top 10 features correlated to diabetes used



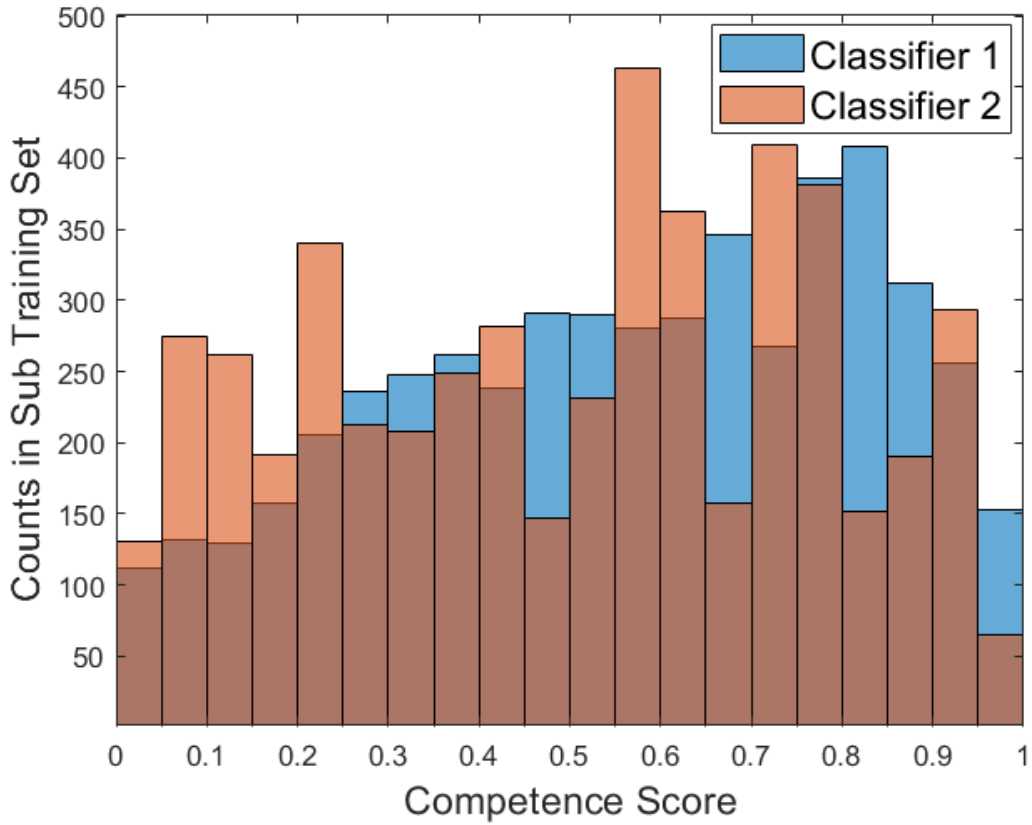
- Classifier 2 Random Forest (100 Trees)
  - Random 60% of the data used as a training set
  - Top 5 features correlated to diabetes used

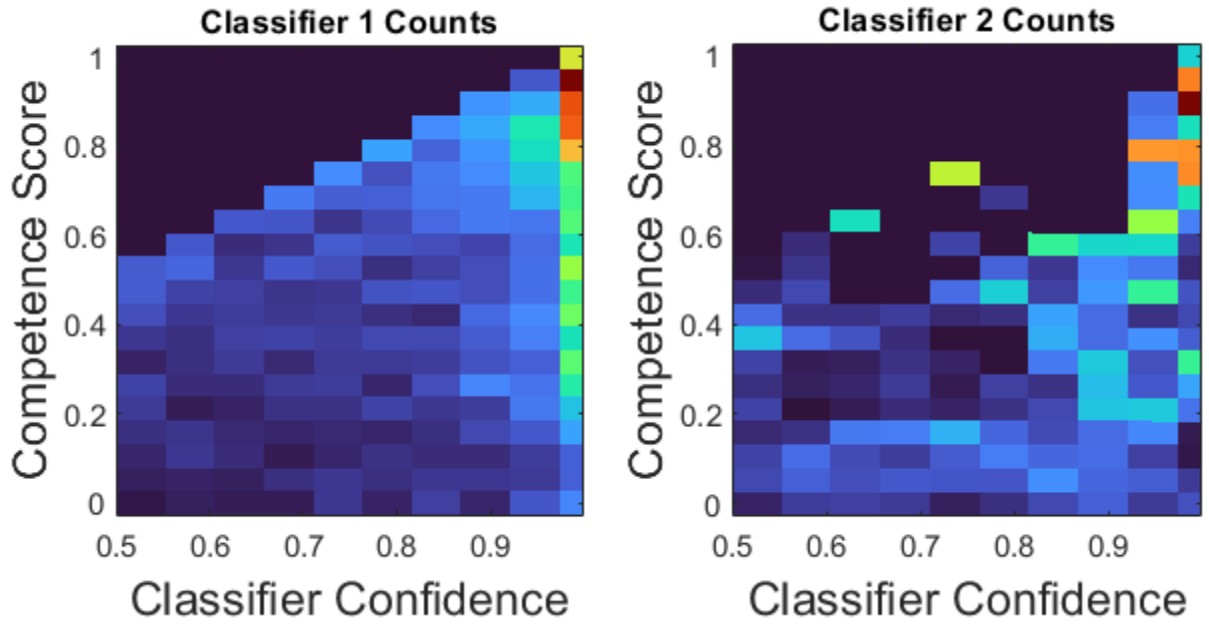


These two random forest classifiers using different features and training data subsets were produced. In the confusion matrices, 30% of the data was preserved as a test set.

Competence Scores of Classifiers

A subset of the training data (5000 points) was then used to analyze the competence score distributions to determine thresholds which may be used in the ensemble.



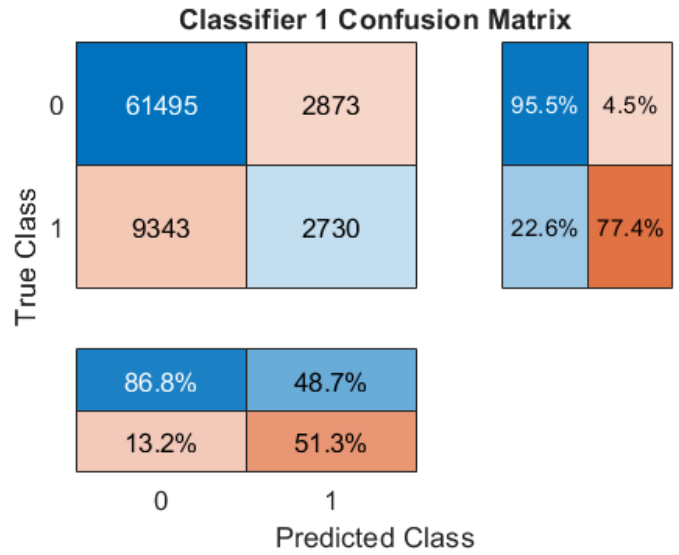


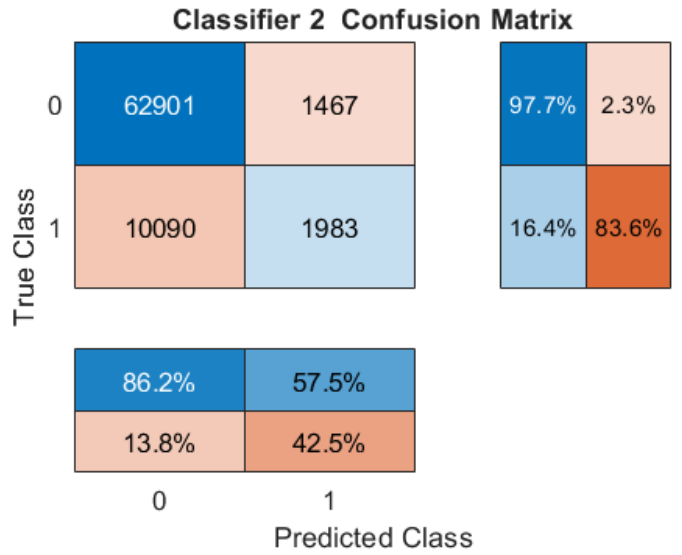
- Competence score thresholds require experts to set and thresholds may be low for real world data sets. Due to distribution of count values, chose competence threshold of 0.5 for Classifier 1 and 0.2 for Classifier 2. Scores below threshold will not be included when non-consensus occurs.

Regions of Consensus

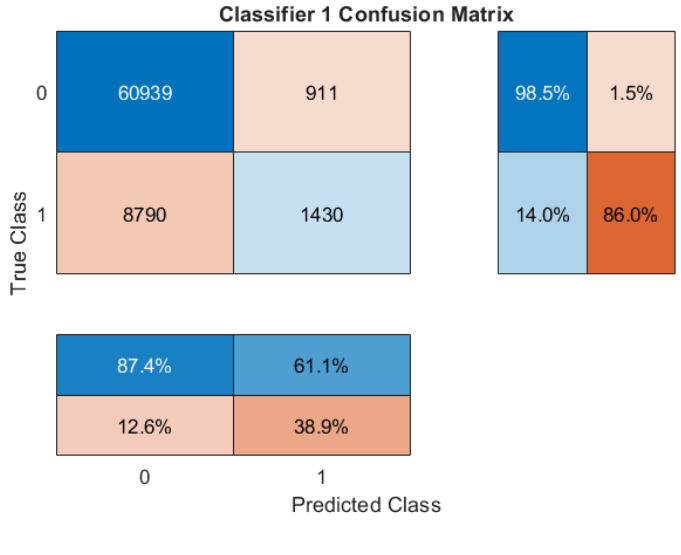
- Consensus was obtained for 94.3% (72070 of 76411) of the test set cases
- As expected, when there is consensus the performance is improved

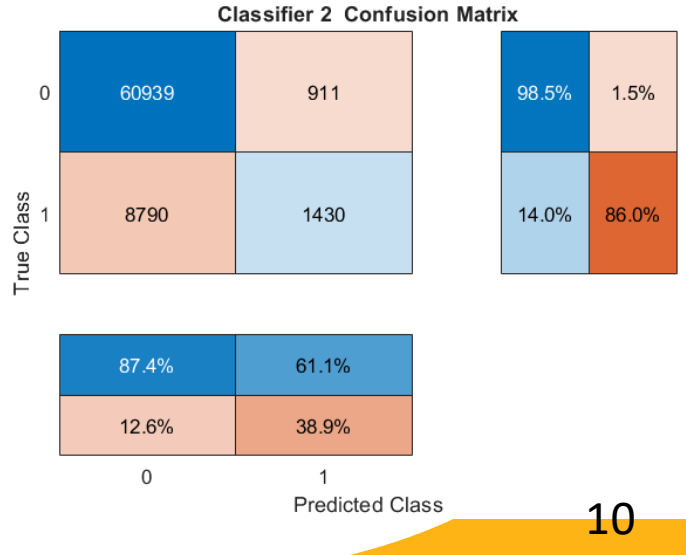
All Test Set Results



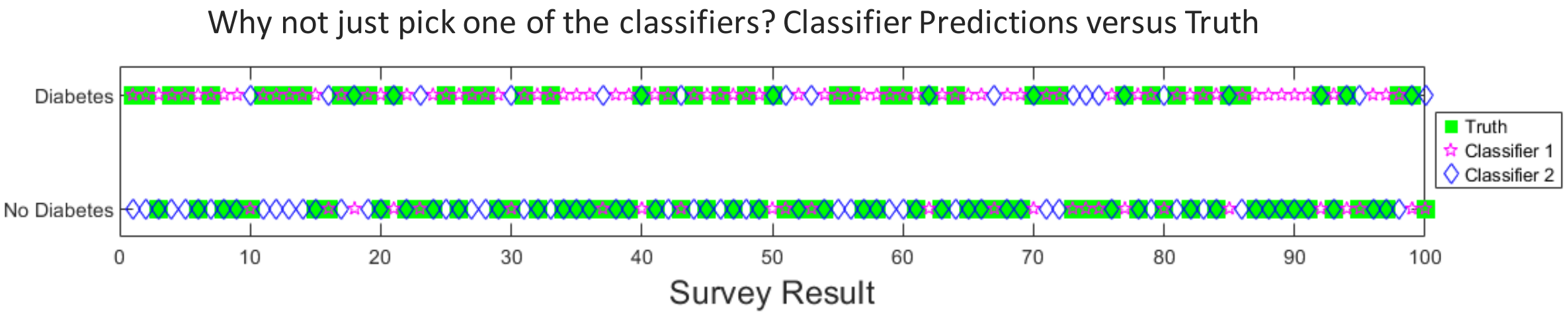


Consensus Cases Only





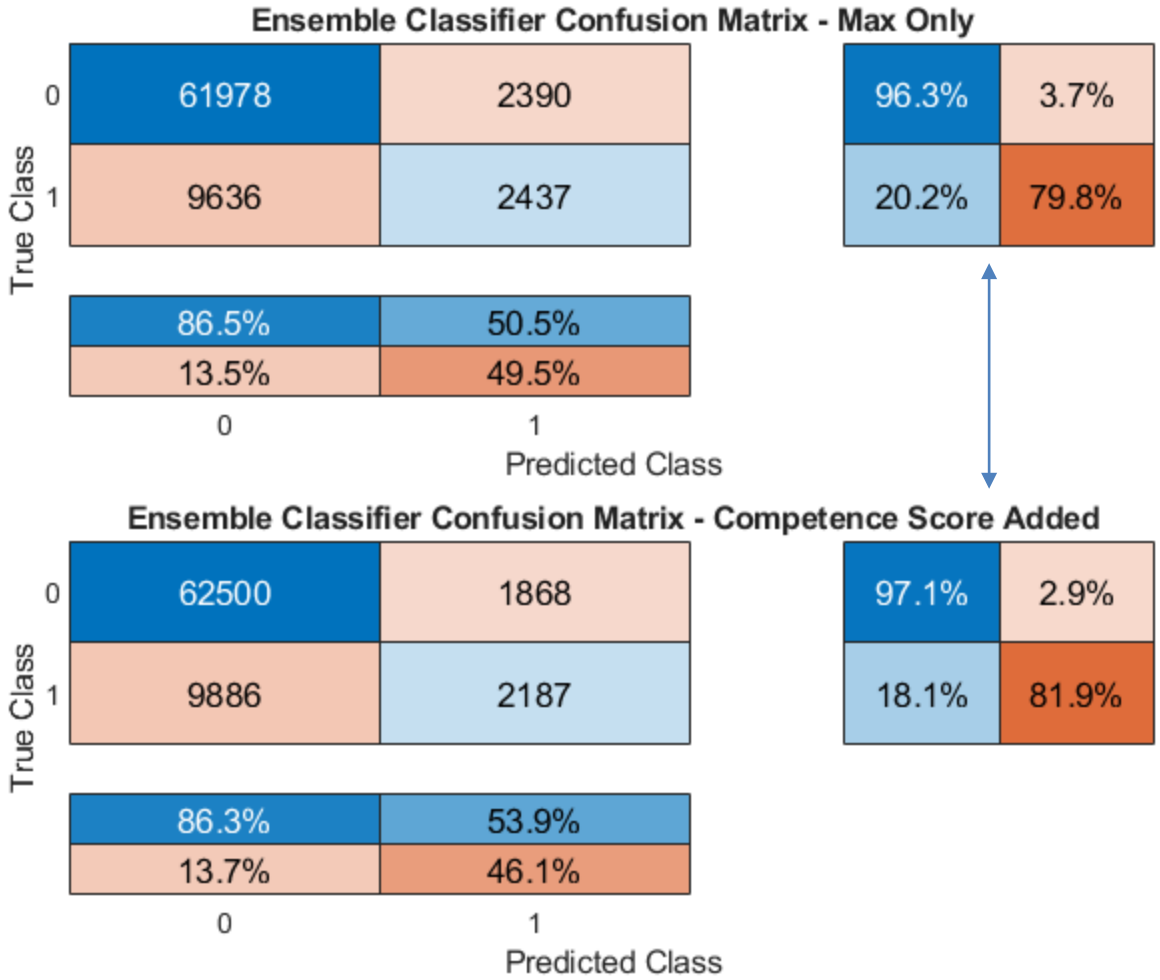
Incorporation of Non – Consensus Results



- Opportunity exists to find areas where Classifier 1 and Classifier 2 are individually accurate
- We attempt to use our new approach to select appropriately when the classifiers differ

Results

- Incorporating the competence score performed slightly better than the max posterior method for true positive rate, true negative rate, false negative rate, and false positive rate
- There are data sets where the differences between the ensemble learning will be pronounced, but we were still able to exploit some of the classifier differences in this example
  - This process will be attempted on several other data sets and classifiers to evaluate where it works best
- We were able to identify and log which classifier was used or selected for each point, leading to more transparency in selection for human machine teaming applications



Discussion, Applications, & Future Work

- Demonstrated an approach for incorporating competence score estimation into ensemble learning methods
  - While there was some promise in performance, will apply to additional data sets and classifiers to identify opportunity for further enhancement from the approach
- This approach enables dynamic integration
  - Model competence scores may be generated at the speed of decision [6]
- Approach is more explainable to end users than network learning ensemble techniques
  - From this approach recommender system visualizations may be formed to make ensemble learning with many classifiers more easily understood by end users

[6] McFadden, Francesca, "Applications of model competence estimation" [Conference Presentation], Society of Industrial and Applied Mathematics (SIAM) Mathematics of Data Science (MDS) Conference, Atlanta, GA, USA, 21-25 October 2024. https://meetings.siam.org/sess/dsp_programsess.cfm?SESSIONCODE=80798

References

[1] V. Rajendran & W. LeVine. Accurate layerwise interpretable competence estimation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alch´e-Buc, E. Fox, R. Garnett, editors, Advances in Neural Information Processing Systems, vol 32,pgs 13981–13991. Curran Associates, Inc., 2019.

[2] Kondu, Rohit, The Essential Guide to Ensemble Learning, V7 Labs, 11 Jan 2024.

[3] Polikar, Robi, "Ensemble based systems in decision making," in IEEE Circuits and Systems Magazine, vol. 6, no. 3, pp. 21-45, Third Quarter 2006, doi: 10.1109/MCAS.2006.1688199.

[4] Centers for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System Survey Questionnaire. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2015.

[5] Teboul, Alex. Diabetes Health Indicators Dataset, Kaggle, 2022.

[6] McFadden, Francesca, “Applications of model competence estimation” [Conference Presentation], Society of Industrial and Applied Mathematics (SIAM) Mathematics of Data Science (MDS) Conference, Atlanta, GA, USA, 21-25 October 2024. https://meetings.siam.org/sess/dsp_programsess.cfm?SESSIONCODE=80798

Matlab was the environment used to process data, create classifier models, and generate plots and confusion matrices for the results shown in this presentation