INFORMATICS PROFESSIONALS. LEADING THE WAY.

A QUEST for Model Assessment: Identifying Difficult Subgroups via Epistemic Uncertainty Quantification

Precision Medicine and Disease Subtyping - "Just You, Just Me"

S103

Katherine E. Brown[1,2], Steve Talbert[3], and Douglas Talbert[2]

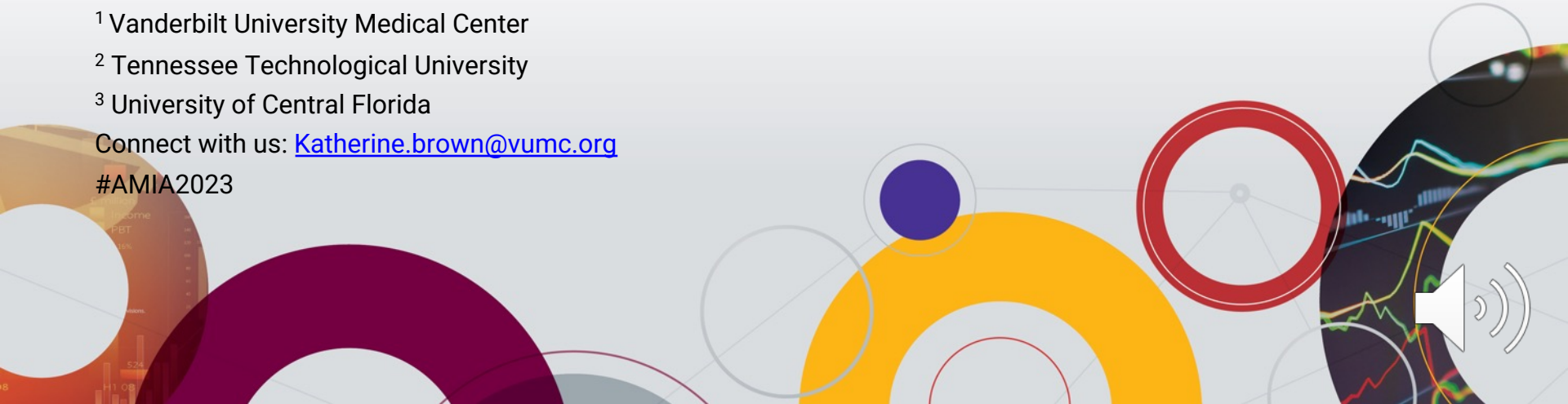
[1] Vanderbilt University Medical Center

[2] Tennessee Technological University

[3] University of Central Florida

Connect with us: Katherine.brown@vumc.org

#AMIA2023



Disclosure

I have no relevant relationships with commercial interests to disclose.

Learning Objectives

After participating in this session the learner should be better able to:

- Understand the benefits of utilizing uncertainty quantification in machine learning model development
- Learn how decision trees can be utilized to form rules that assign a discrete level of uncertainty to data points
- Evaluate subpopulations of data based on predicted uncertainty level

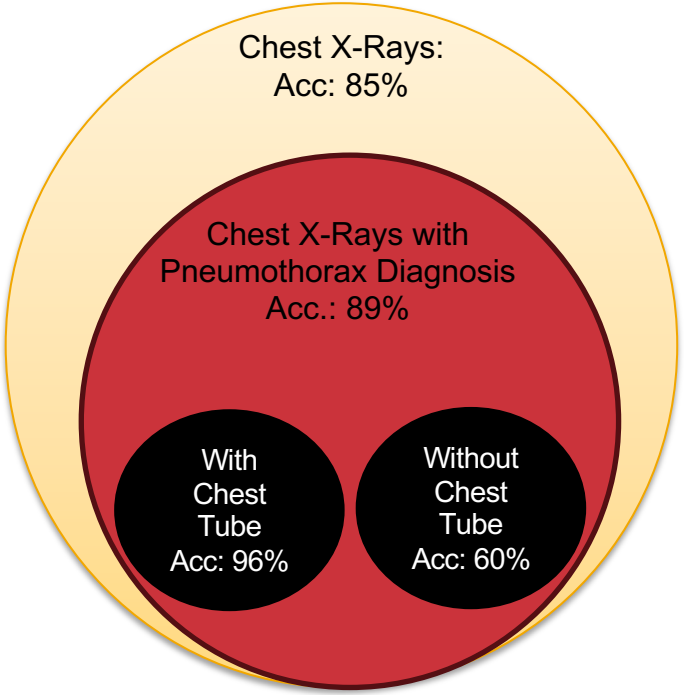
Hidden Stratification

Hidden stratification occurs when a subgroup of data exists with higher error (equiv. lower accuracy) than its super-group [1]

Prevalent in chest x-ray detection of pneumothoraxes

- Pneumothorax occurs when air is outside the lung but within the chest cavity
- When the result of a trauma, the condition is treated promptly in ER

Deep learning-based analysis of chest x-rays have been noted to have higher error on chest x-rays with pneumothoraxes without a chest tube than pneumothoraxes with a chest tube



[1] Oakden-Rayner, L., Dunnmon, J., Carneiro, G., & Ré, C. (2020). Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. Proceedings of the ACM conference on health, inference, and learning, 151–159.

Uncertainty Quantification

Uncertainty in ML: Measured as variation of model predictions under stochasticity [2,3]

High Variation implies High Uncertainty

Epistemic uncertainty: Uncertainty in a model due to insufficient data or improperly tuned model hyperparameters [4]

Epistemic uncertainty is effective at identifying unseen (or out-of-distribution) data [4]

[2] Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *Proceedings of The 33rd International Conference on Machine Learning*, 1050–1059.
[3] Malinin, A., Prokhorenkova, L., & Ustimenko, A. (2020). Uncertainty in gradient boosting via ensembles. *ArXiv Preprint ArXiv:2006.10562*.
[4] Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, *30*, 5574–5584.

Research Questions and Hypotheses

Motivation: Few techniques use uncertainty quantification to inform model assessment, and we have seen that uncertainty can serve as a powerful indicator of incorrectness.

Hypothesis: Uncertainty quantification can be used as a model performance analysis tool that can identify higher- and lower-accuracy patient subpopulations

Research Questions to Test Hypothesis:

- (RQ 1) Does an inherently interpretable model predict a discretized uncertainty label derived from the underlying uncertainty value with high fidelity?
- (RQ 2) Is there a statistically significant performance difference between subsets of data associated with different levels of predicted uncertainty?

Quantifying Uncertainty for Estimating Subgroup Types (QUEST)

Inputs

Supervised training and validation dataset (labeled with correct classification labels)

Classification model trained and validated upon above data

Uncertainty quantification technique

Discretization technique

Supervised, transparent rule induction algorithm

Output

Set of rules defining subgroups associated with different levels of uncertainty
(Each $r \in R$ is an if-then rule describing one subgroup).

Using QUEST for Model Assessment

Three Notes on QUEST

1. A large proportion of the training data is withheld from training and used as validation data, and QUEST is trained on uncertainty information for training and validation data
2. Classification label is replaced with a discretized uncertainty label
3. An inherently interpretable model is used to model uncertainty

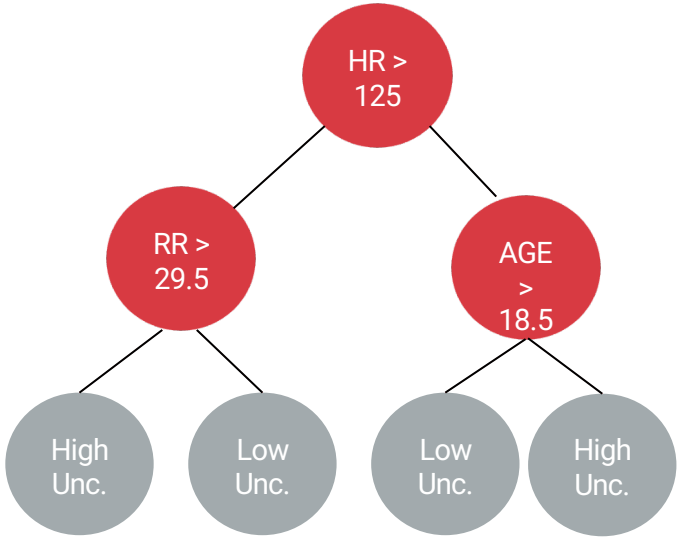
Model assessment via QUEST primarily requires utilizing the predicted uncertainty label as an indicator of potential incorrectness.

Decision Trees

Each leaf has the resulting uncertainty level

Each non-leaf node has a Boolean condition

Conjunction of conditions from root to leaf describes the patients in the subgroup



(HR > 125) and (Age > 18.5) => High Unc.
(HR > 125) and (Age <= 18.5) => Low Unc.
(HR <= 125) and (RR > 29.5) => Low Unc.
(HR <= 125) and (RR > 29.5) => High Unc.

Empirical Evaluation

Neural network classification model

10-fold CV

QUEST Implementations

- $QUEST(\alpha_2)$: LOW, HIGH
- $QUEST(\alpha_3)$: LOW, MEDIUM, HIGH

Measure average test set classification accuracy weighted by test set coverage

Welch's T-Test for samples of unequal size and variance

- LOW vs. HIGH
- LOW vs. MEDIUM (α_3 only)
- MEDIUM vs. HIGH (α_3 only)

KY Trauma Triage

Records originate from Level 1 Trauma Center

Predict if patient has Injury Severity Score > 15

32 total features

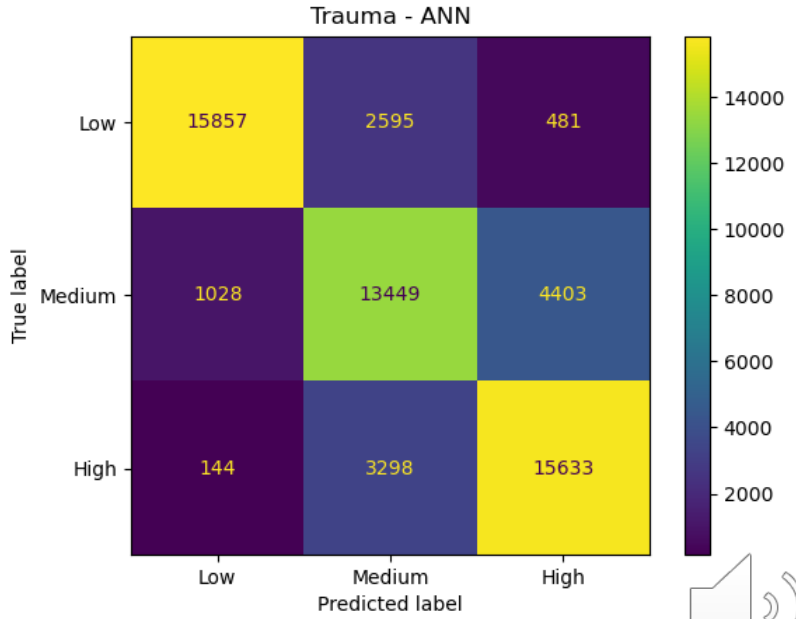
- Physiological parameters
- Anatomical criteria
- Mechanism of injury
- Age
- Multiple computed injury scores: AMPT, GCS

Accuracy	F1 Score	AUROC
0.846	0.733	0.909

QUEST Fidelity to Discretized Uncertainty

(RQ 1) Does an inherently interpretable model predict a discretized uncertainty label derived from the underlying uncertainty value with high fidelity?

Version	No. Rules	Fidelity
$QUEST(\alpha_2)$	34.0 (8.58)	0.85 (0.01)
$QUEST(\alpha_3)$	33.1 (11.84)	0.79 (0.02)



Identifying Performance Differences in Predicted Subgroups

(RQ 2) Is there a statistically significant performance difference between subsets of data associated with different levels of predicted uncertainty?

Version	Low Unc. Cls. Acc.	High Unc. Cls. Acc.	P-Value
$QUEST(\alpha_2)$	0.937	0.753	0.005

Version	Low Unc. Cls. Acc.	Med. Unc. Cls. Acc.	High Unc. Cls. Acc.	P-Value (Low vs Med.)	P-Value (Med. vs High)	P-Value (Low vs. High)
$QUEST(\alpha_3)$	0.988	0.838	0.730	0.030	0.085	0.009

Conclusions

Introduced QUEST – a system to identify low and high performing subgroups of predictions based on epistemic uncertainty

- Used pruned decision trees to classify data to different levels of discretized uncertainty

Answers to Research Questions

- Proposed method had high fidelity to discretized uncertainty
- Predicted subgroups of differing uncertainty levels are linked to varying levels of performance

Uncertainty quantification can be used as a model performance analysis tool that can identify higher- and lower-accuracy patient subpopulations

QUEST is a first step in an uncertainty-based solution to hidden stratification.

Thank you!

Katherine Brown	Steve Talbert	Douglas Talbert
katherine.brown@vumc.org X: @katiebrown_phd_ Web: katherinebrown539.github.io	steven.talbert@ucf.edu	dtalbert@tntech.edu