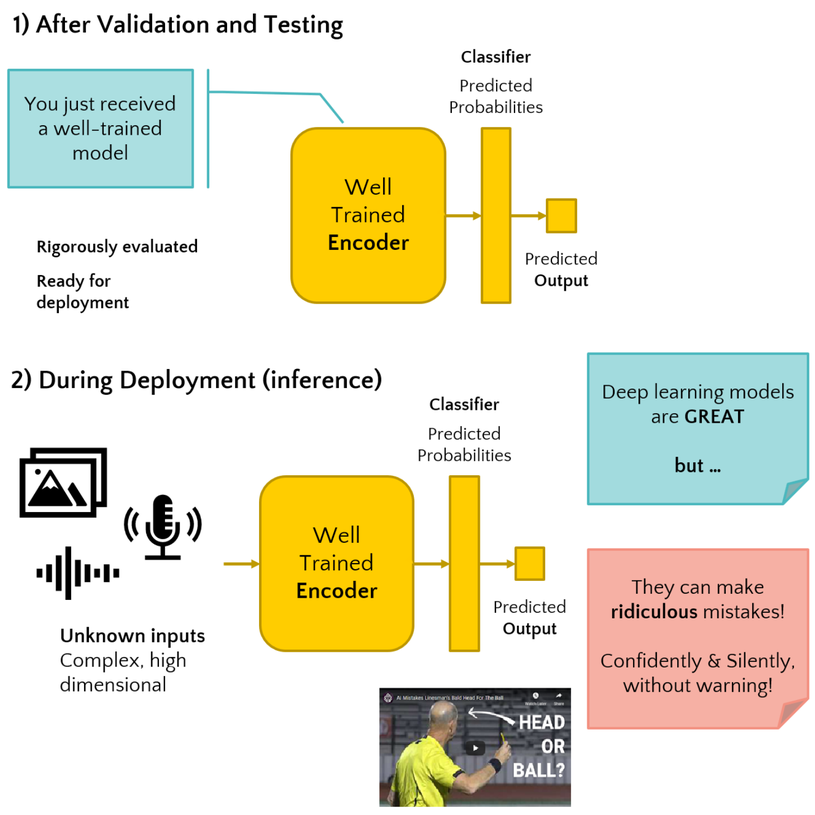
Trust Me Not: Trust Scoring for Continuous Model Monitoring

Nandita Bhaskhar, Daniel Rubin, Christopher Lee-Messer

Stanford University

Motivation: Continuous Model Monitoring



Problem Formulation

Goal: During inference, for *every* incoming input, determine:

- Is the trained model's prediction trustworthy?
- ***accept*** or ***reject*** model predictions?

Desiderata: Continuous Model Monitoring

Notion of Trust

Trust: (i) Nuanced, (ii) Difficult to quantify, (iii) Subjective, (iv) Context driven, (v) Boils down to belief

Various flavours of Trust

- Probability: Model confidence
- Uncertainty: Classical techniques
- Explainability: Glass Box Model
- Fairness: Human metrics for fairness
- Adversarial attacks: Robustness
- Generalizability: Diverse test sets

How do humans do it?:

- Surprisingly good at this
- Past learnt knowledge, lived experiences, human intuition
- e.g. encountering a language shift

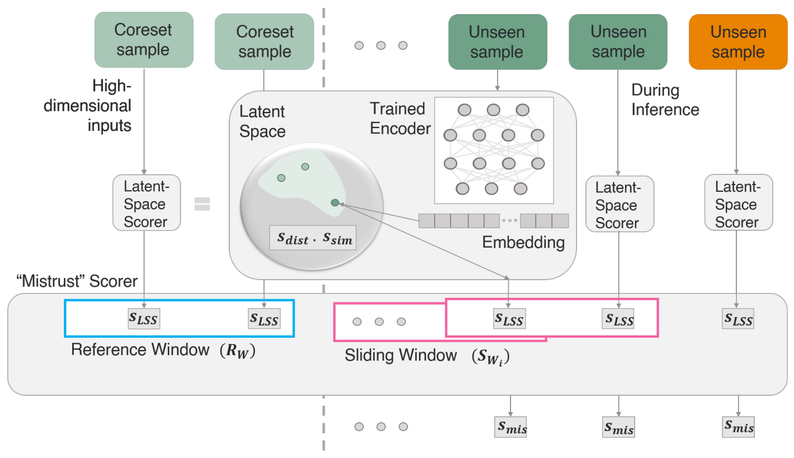
Limitations in Current Approaches

Current Approaches: Calibration, Bayesian Neural Networks, Variational methods, Ensembling, Monte Carlo Dropout, etc.

Limitations

- Difficult to train
- Separate / new architectures
- Modifications to training strategies
- Computationally expensive
- May need exposure to labelled outliers
- **Insensitive to semantic content!!**
- **Ultimately, do NOT fulfil desiderata**

TRUST-LAPSE: Our Approach



- **Project** complex, high dimensional inputs to the **Latent-Space**
- Compare latent-space embeddings with those of **coreset** using **different metrics** to get **Latent Space Score**
- Estimate correlations over **SEQUENCES** of these scores (set-based approach vs instance-based approach) to give **Sequential Mistrust Score**
- Final Decision: **Trust** / **Mistrust**

Latent-Space Scorer

Coreset (sampled from trainset)

$$\text{coreset} = \{h(x_i) \mid x_i \sim \mathcal{D}_{train}\}; \quad |\mathcal{D}_{train}| \geq |\text{coreset}|$$

Distance-based Metric

$$s_{\text{dist}}(x) = \min_{h(x_i) \in \text{coreset}} d(h(x_i), h(x))$$

Angle-based Similarity

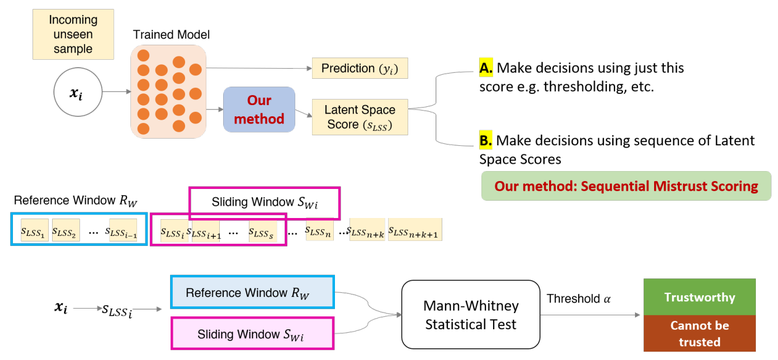
$$s_{\text{sim}}(x) = \max_{h(x_i) \in \text{coreset}} \text{sim}(h(x_i), h(x))$$

Latent Space Score

$$s_{\text{LSS}}(x) = s_{\text{dist}}(x) \cdot s_{\text{sim}}(x)$$

- **Distance-based Metric**: Mahalanobis Distance with class-wise separate covariance, no label smoothing
- **Angle-based Metric**: Cosine Similarity

Sequential Mistrust Scorer



Key Insights

- Well-trained encoder: encodes world into ***hierarchical, geometric, latent space***
- Metrics in latent-space: capture how ***similar*** are two inputs and how ***near*** or ***far*** are two inputs
- Different metrics capture different aspects of the latent-space. **Combining** them has value
- Track these over time as a **sequence** for continuous model monitoring

Results: Out of Distribution Detection

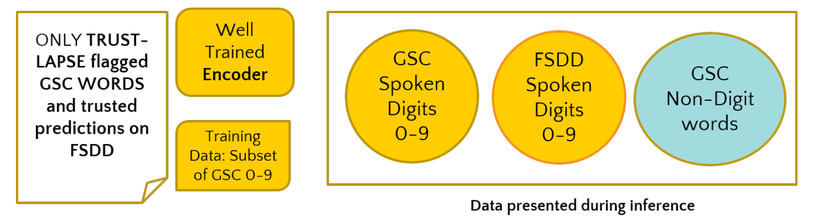
SOTA on vision, audio and challenging clinical EEG domains

SOTA — AUROC↑ / AUPR↑ / FPR80↓

	Audio	EEG Data	Vision
Task	Speech Classification	Seizure Detection	Image Classification
OOD Sets	Other spoken words	Other institutions	SVHN
MSP	0.626 / 0.527 / 0.515	0.358 / 0.421 / 0.754	0.760 / 0.770 / 0.358
Predictive Entropy	0.615 / 0.515 / 0.515	0.393 / 0.495 / 0.742	0.761 / 0.752 / 0.357
KL_U	0.553 / 0.475 / 0.579	0.390 / 0.472 / 0.719	0.775 / 0.786 / 0.347
ODIN	0.466 / 0.448 / 0.712	0.325 / 0.388 / 0.790	0.748 / 0.776 / 0.402
Vanilla Mahalanobis	0.680 / 0.636 / 0.520	0.633 / 0.651 / 0.525	0.738 / 0.782 / 0.477
Test-Time Dropout	0.649 / 0.619 / 0.523	0.647 / 0.619 / 0.583	0.716 / 0.725 / 0.494
TRUST-LAPSE (ours)	**0.739 / 0.704 / 0.439**	**0.771 / 0.701 / 0.335**	**0.814 / 0.827 / 0.311**

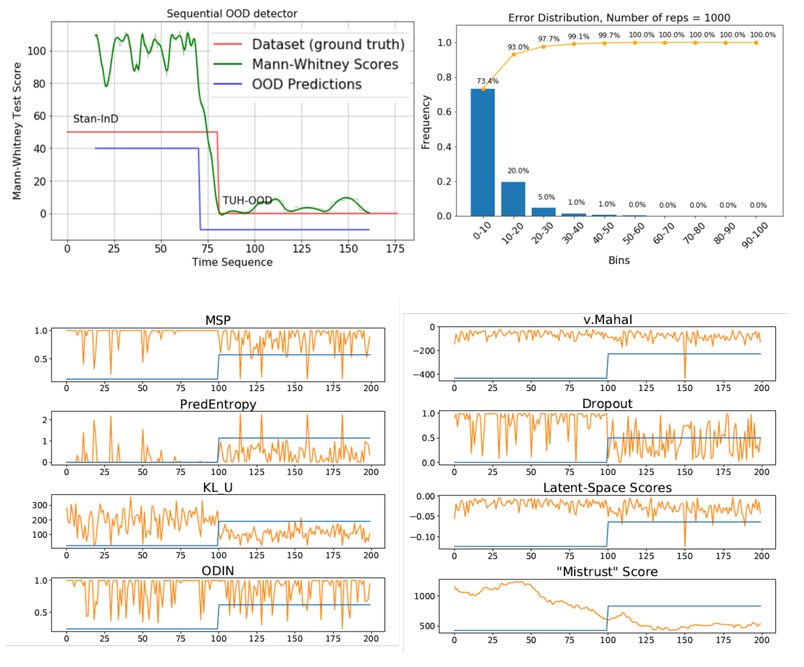
Results: Semantic Shifts

- Other methods: more prone to **dataset statistics**, NOT semantic content, unlike TRUST-LAPSE
- **Counterfactual experiment** with 2 spoken word datasets: GSC and FSDD



- TRUST-LAPSE detects semantic shifts on ALL domains (vision, audio, EEG) unlike other methods. Details in paper.

Results: Drift Detection



- EEG: >73% of 1000 datastreams (of length 10,000) have <10% error and >93% have <20% error
- Audio: >85% of streams have <10% error and >97% have <20% error
- Vision: near 99% detection accuracy for >95% of streams

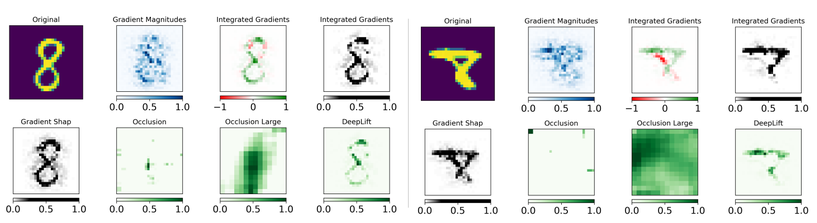
More Key Results

TRUST-LAPSE can detect **lack of generalization in trained models**

TRUST-LAPSE performance depends on **encoder capacity**

Just **1-2%** of trainset in the **coreset** is sufficient for good TRUST-LAPSE performance

Ties to Explainability



- High (low) trust scores from TRUST-LAPSE correlate with correct (incorrect) predictions and good (poor) attributions
- **(left)** Digit 8, Prediction 8. Trust score: 0.95 (high). **(right)** Digit 8, Prediction 7. Trust score: 0.44 (low).

Conclusions & Contact Details

TRUST-LAPSE is a **simple** yet **powerful** and **flexible** framework that we can use for any model and any task for monitoring a model in deployment, essential in safety-critical domains like healthcare, self-driving, etc.

Questions? Hiring? Reach Out :)

Email: nanbhas@stanford.edu, **Website**: www.stanford.edu/∼nanbhas, **Twitter**: @BhaskharNandita, **LinkedIn**: https://www.linkedin.com/in/nanditabhaskhar/