筑波大学
University of Tsukuba

Characterizing Reliability of Three-version Traffic Sign Classifier System through Diversity Metrics

Qiang Wen and Fumio Machida

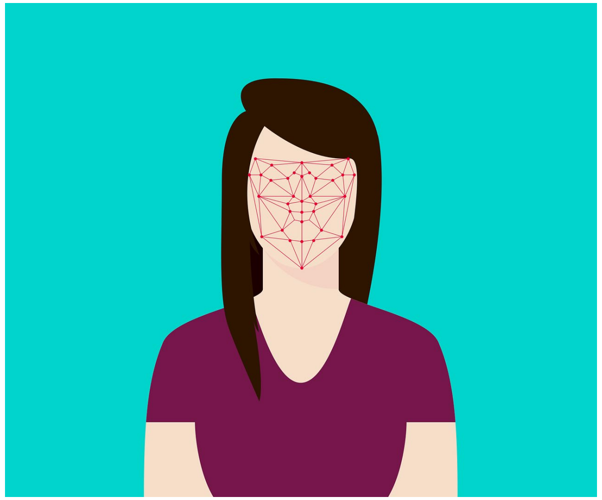
Laboratory for System Dependability

Department of Computer Science

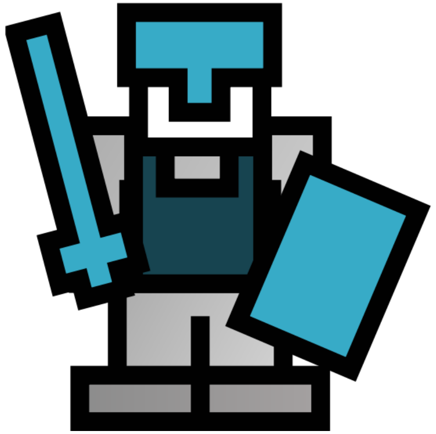
Outline

1. Introduction
2. Related Work
3. Reliability Model
4. Objective
5. Experiment Configuration and Results
6. Conclusion & Future Work

1. Introduction – Machine Learning Systems



- **Machine learning (ML) models have been used in many intelligent software systems.**
  - Face recognition
  - Medical diagnosis
  - Autonomous robots and vehicles



1. Introduction – Reliability Issues of ML Systems

- Outputs of ML models for real-world input data are not always correct

- Error outputs of ML models may induce undesirable consequences (e.g., traffic accidents in automated driving)

2. Related Work – Reliability Issues

■ **Approaches to ML system reliability improvement**

- Data validations [1]
  - Detect real-world error-inducing corner cases at runtime
  - Require a white box model for deep neural networks
- Safety monitors [2]
  - Detect out-of-distribution data at runtime
  - Need to be trained together with the ML model in advance

✓ Redundant architecture [3-4]

  - Achieve improved reliability by a simple redundancy scheme with diversity

2. Related Work – On Reliable ML Systems

■N-version ML system approach

- Multiple ML models [5]
- Diversified input data [7]

■Issue of parameter estimation

- ➢ Estimation of diversity parameters
- ➢ The impacts of estimated diversity parameters on system reliability

2. Related Work – Diversity Measures

■Diversity Metrics

- Mutual error rate [6]
- Coverage of errors [7]
- Gini coefficient and the Shannon equitability index [8]

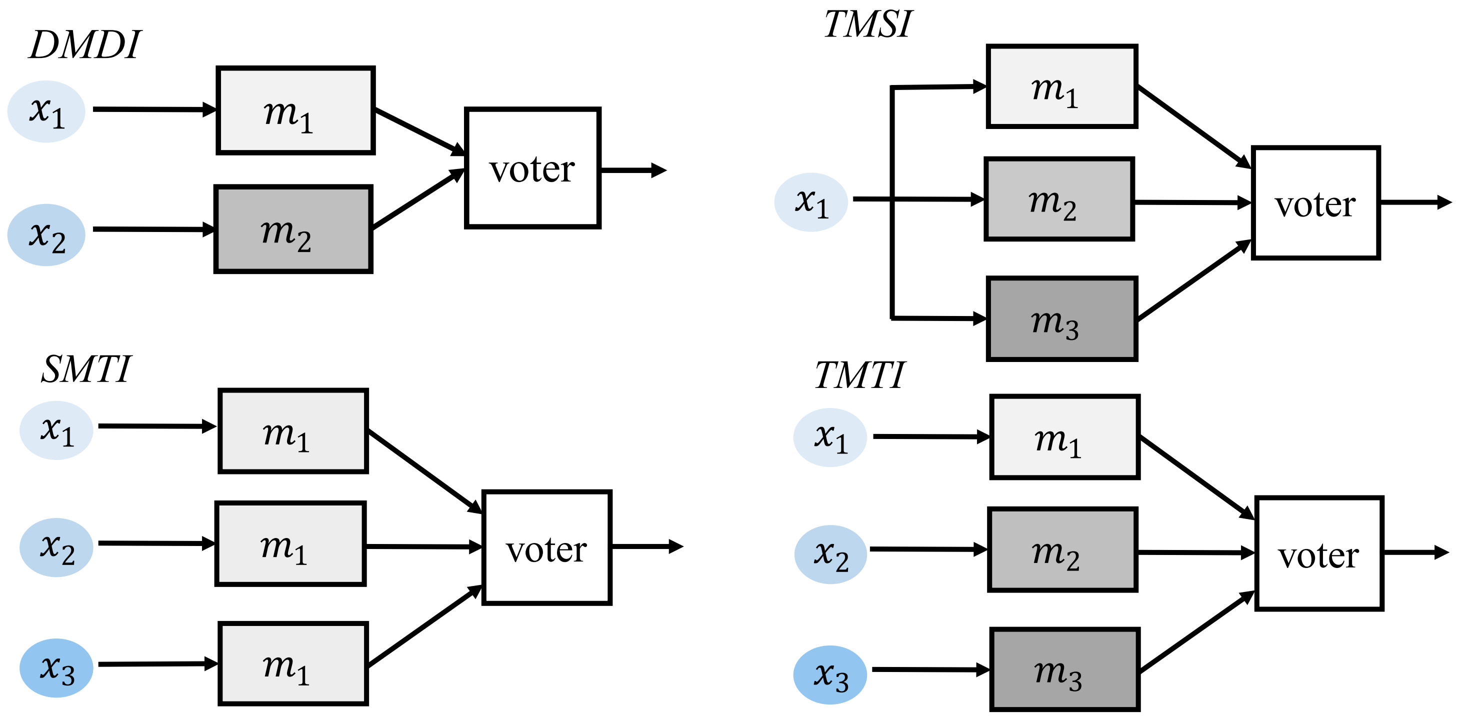
➢ The metrics are not applicable for diversity in different input data sources.

➢ The joint impact of model diversity and input diversity on system reliability is not discussed.

3. Reliability Model – N-version ML Architectures

■ Two-version and three-version ML architectures

- **Double model with double input system (DMDI)**
- **Triple model with single input system (TMSI)**
- **Single model with triple input system (SMTI)**
- **Triple model with triple input system (TMTI)**



3. Reliability Model – Conventional Reliability Model

■ A conventional reliability model for a three-version system

$$R = 1 - [3\alpha f(1-\alpha) + \alpha^2 f] = 1 - \alpha f(3 - 2\alpha)$$

- ***R*:** Reliability of a three-version N-version Programming model
- ***α*:** A dependent failure parameter
- ***f*:** The failure probability of each version

■ Shortcomings

- The ratio of the dependence is **homogeneous** which may not be true in reality.
- The dependent failure parameter is **not enough** to represent the dependence of input data.

3. Reliability Model – Diversity Metrics [4]

- $\boldsymbol{\alpha_{i,j}}$ **:** Model diversity- Intersection of errors $\alpha_{i,j} \in [0,1]$
  - $\boldsymbol{E_i}$ **,** $\boldsymbol{E_j}$ **:** The input sets that make ML models m_i and m_j output error
  - A smaller intersection value is better-ML models are unlikely to reach a mutual error

$$\alpha_{i,j} = \frac{|E_i \cap E_j|}{min\{|E_i, E_j|\}}$$

- $\boldsymbol{\beta_{\mathrm{i,s}|t}}$ **:** Input diversity- Conjunction of errors $\beta_{\mathrm{i,s}|t} \in [0,1]$
  - $\boldsymbol{x_s}$ **,** $\boldsymbol{x_t}$**:** Input data to ML models from different data sources (i.e., $s \neq t$)
  - A smaller conjunction value is better-the probability of a mutual error becomes small

$$\beta_{\mathrm{i,s}|t} = Pr[x_s \in E_i \mid x_t \in E_i]$$

3. Reliability Model – Reliabilities [4][9]

- **Reliability of DMDI:**

$$\boldsymbol{R_{2,2}}(\boldsymbol{m_1}, \boldsymbol{m_2}; \boldsymbol{x_1}, \boldsymbol{x_2}) = 1 - \left[\beta_{1,2|1} \cdot \alpha_{1,2} + (1 - \beta_{1,2|1}) \cdot \frac{p_2 - \alpha_{1,2} \cdot p_1}{(1 - p_1)}\right] \cdot p_1$$

- **Reliability of TMSI:**

$$\boldsymbol{R_{3,1}}(\boldsymbol{m_1}, \boldsymbol{m_2}, \boldsymbol{m_3}; \boldsymbol{x_1})$$
$$= 1 - (\alpha_{1,2} \cdot p_1 + \alpha_{1,3} \cdot p_1 + \alpha_{2,3} \cdot p_2 - 2\alpha_{1,2} \cdot \alpha_{1,3} \cdot p_1)$$

- **Reliability of SMTI:**

$$\boldsymbol{R_{1,3}}(\boldsymbol{m_1}; \boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}) = 1 - (\beta_{1,2|1} p_1 + \beta_{1,3|1} p_1 + \beta_{1,3|2} p_2' - 2\beta_{1,2|1} \beta_{1,3|1} p_1)$$

- **Reliability of TMTI:**

$$\boldsymbol{R_{3,3}}(\boldsymbol{m_1}, \boldsymbol{m_2}, \boldsymbol{m_3}; \boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3})$$
$$= 1 - [p_{2,2}(m_1, m_2; x_1, x_2) + p_{2,2}(m_1, m_3; x_1, x_3) + p_{2,2}(m_2, m_3; x_2, x_3) -$$
$$2p_{2,2}(m_1, m_2; x_1, x_2) \cdot p_{2,2}(m_1, m_3; x_1, x_3)/p_1]$$

3. Reliability Model – Variants of Reliability Models

■ **Five variants in the evaluation of TMSI reliability**

$$\boldsymbol{R_{3,1}(m_1, m_2, m_3; x_1)}$$
$$= 1 - (\alpha_{1,2} \cdot p_1 + \alpha_{1,3} \cdot p_1 + \alpha_{2,3} \cdot p_2 - 2\boldsymbol{\alpha_{1,2} \cdot \alpha_{1,3} \cdot p_1})$$

$$\begin{cases} \boldsymbol{t_1 = \alpha_{1,2} \cdot \alpha_{1,3} \cdot p_1} \\ \boldsymbol{t_2 = \alpha_{1,2} \cdot \alpha_{2,3} \cdot p_1} \\ \boldsymbol{t_3 = \alpha_{1,3} \cdot \alpha_{2,3} \cdot p_1} \\ \boldsymbol{t_4 = \dfrac{t_1 + t_2 + t_3}{3}} \\ \boldsymbol{t_5 = \sqrt[3]{t_1 t_2 t_3}} \end{cases}$$

4. Objective

Objective

- Theoretical investigation of the reliability of N-version ML systems with model diversity and input diversity.
- Lack of discussion on the effectiveness of **diversity metrics** for reliability prediction.

Empirical Experiment

- Conduct experiments on **traffic sign recognition tasks** using deep neural networks
- Evaluate the reliability of **three-version traffic sign classifier architectures**
- Compare observed reliability with predicted reliability based on estimated diversity parameter values.

5. Experiment Configuration

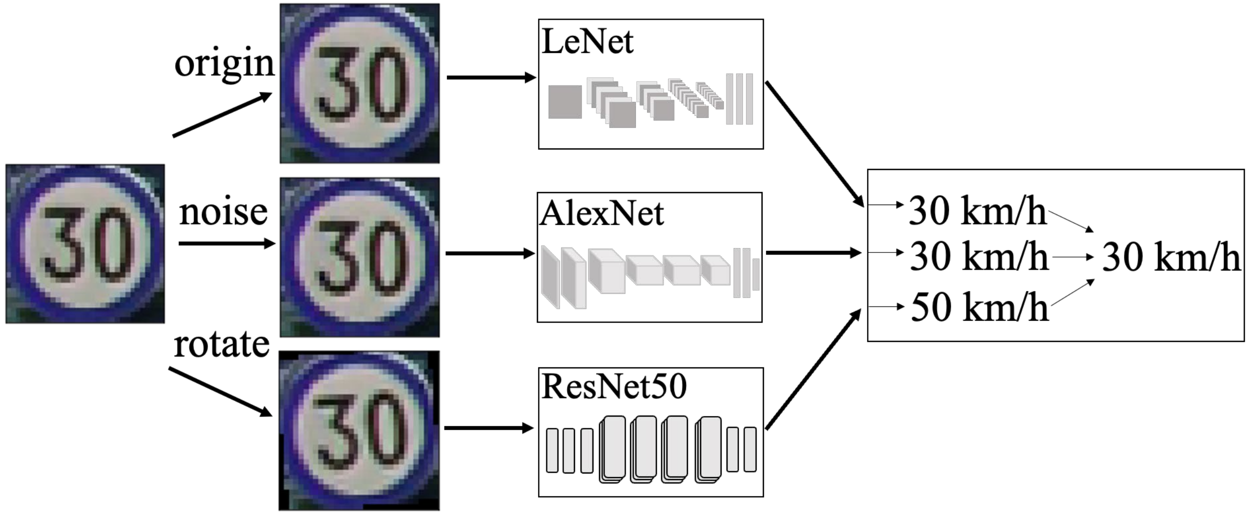
■ *Model Diversity*

- LeNet
- AlexNet
- ResNet50

■ *Input Diversity*

- Original data
- Noise-added data
- Rotated data

(rotate 5 degrees counterclockwise)



A three-version system by TMTI architecture

5. Experiment Configuration

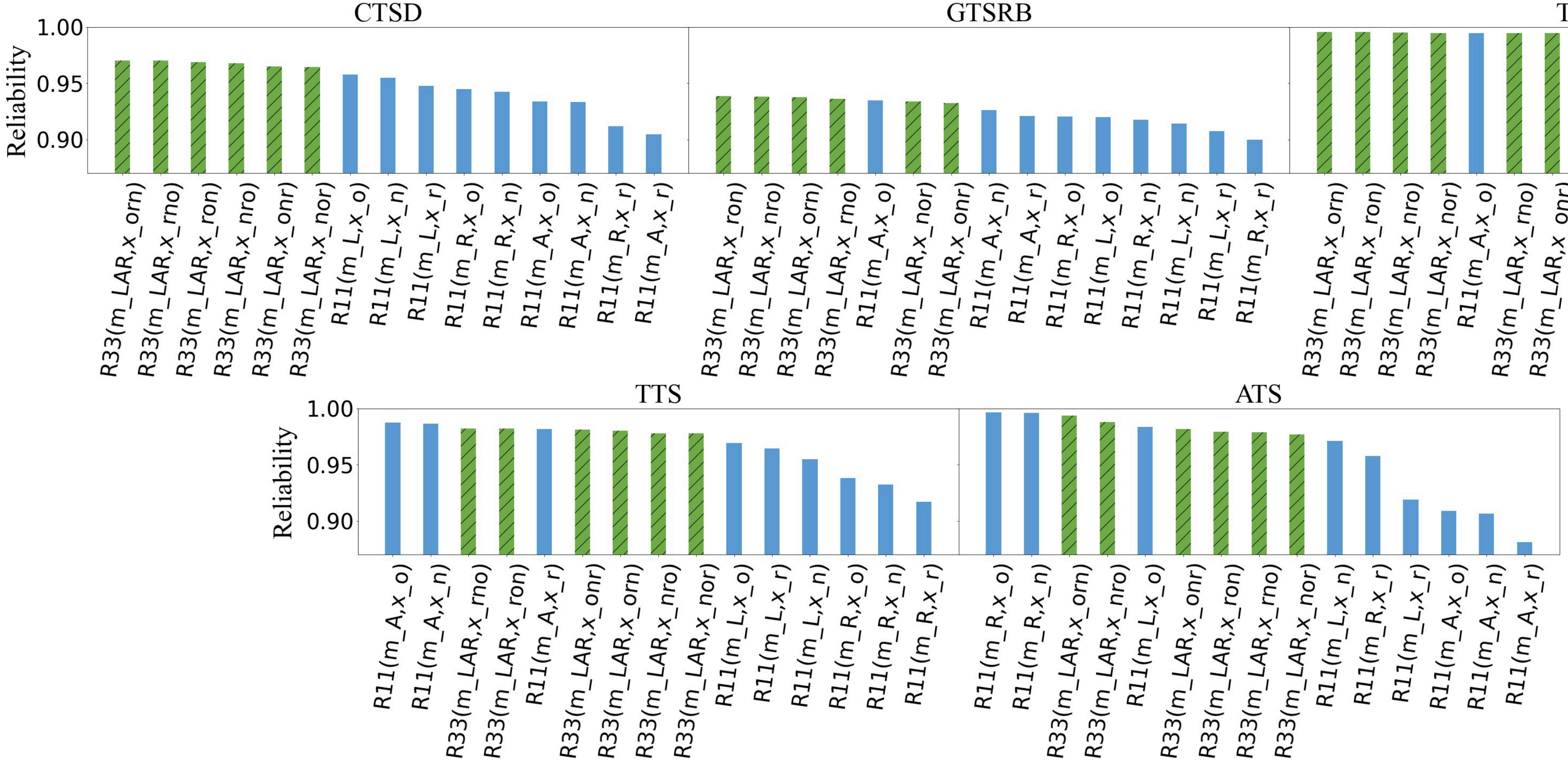
■ *Datasets*

Five different traffic sign datasets

- Chinese Traffic Sign Dataset (CTSD)
- German Traffic Sign Recognition Benchmark (GTSRB)
- Traffic Sign Classification Dataset (TSCD)
- Turkey Traffic Sign (TTS)
- Arabic Traffic Signs (ATS)

5. Experiment Results – Research Question 1

◆**Does the implementation of a three-version system architecture effectively enhance reliability?**

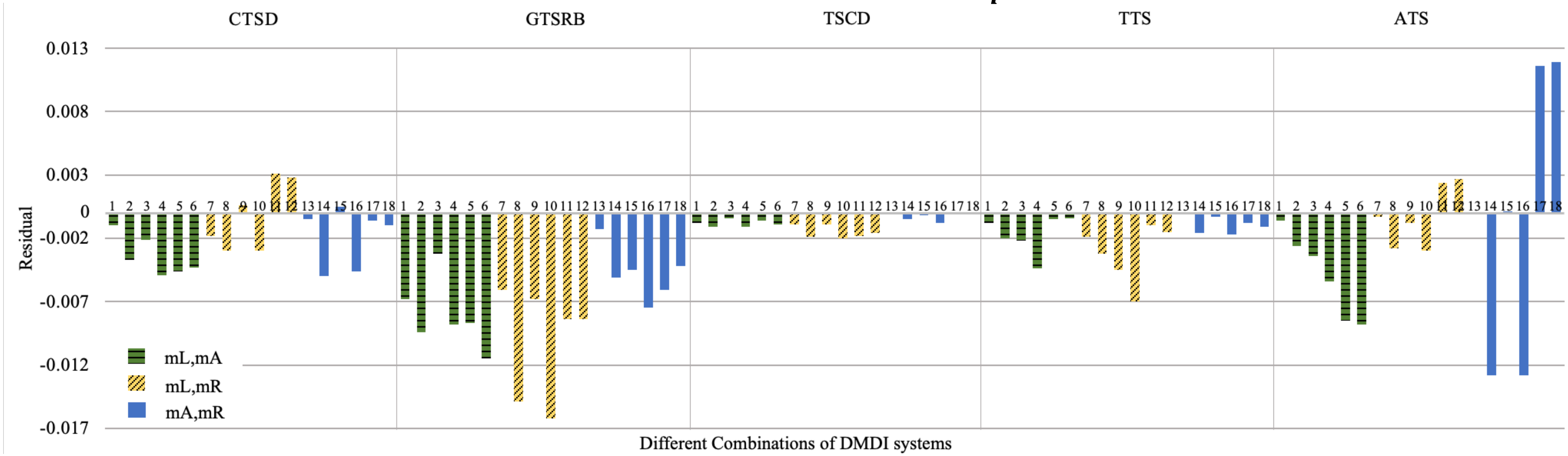


Observation 1. Three-version ML system architectures, especially the TMTI architecture, have the potential to efficiently improve system reliability compared to single models.

5. Experiment Results – Research Question 2

◆**How can the reliability models using diversity parameters estimate well the reliability of traffic sign classifier architectures?**

- e *(Prediction residual)* $= R_{observed} - R_{predicted}$



DMDI residual between observed results and model results

Observation 2. The prediction residuals are mostly less than 0.017 across five data sets in most architectures except the SMTI architecture.

5. Experiment Results – Research Question 3

◆**How does the last term of the three-version reliability model impact on the reliability prediction?**

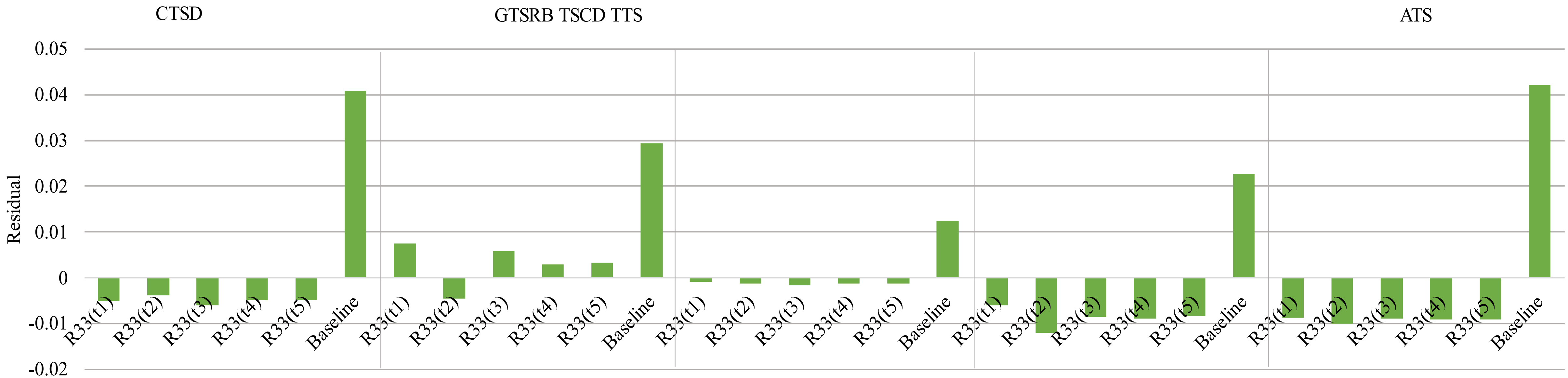
- **Five variants in the evaluation of TMTI reliability**

$$\boldsymbol{R_{3,3}(m_1, m_2, m_3; x_1, x_2, x_3)}$$
$$= 1 - [p_{2,2}(m_1, m_2; x_1, x_2) + p_{2,2}(m_1, m_3; x_1, x_3) + p_{2,2}(m_2, m_3; x_2, x_3) -$$
$$2\boldsymbol{p_{2,2}(m_1, m_2; x_1, x_2) \cdot p_{2,2}(m_1, m_3; x_1, x_3)/p_1}]$$

$$\begin{cases} t_1 = \dfrac{p_{2,2}(m_1, m_2; x_1, x_2) \cdot p_{2,2}(m_1, m_3; x_1, x_3)}{p_1} \\ t_2 = \dfrac{p_{2,2}(m_1, m_2; x_1, x_2) \cdot p_{2,2}(m_2, m_3; x_2, x_3)}{p_1} \\ t_3 = \dfrac{p_{2,2}(m_1, m_3; x_1, x_3) \cdot p_{2,2}(m_2, m_3; x_2, x_3)}{p_1} \\ t_4 = \dfrac{t_1 + t_2 + t_3}{3} \\ t_5 = \sqrt[3]{t_1 t_2 t_3} \end{cases}$$

5. Experiment Results – Research Question 3

■ **Residual between observed results and model results for TMTI**

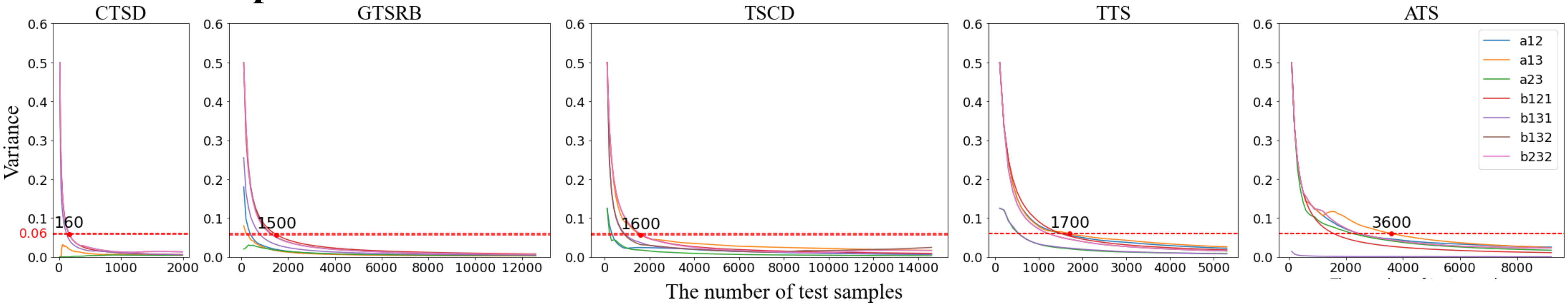


Observation 3. The residuals of five variants of TMSI, SMTI, and TMTI reliability predictions are equally effective. No variant shows evident superiority over the others.

5. Experiment Results – Research Question 4

◆How many samples are required to obtain good estimates of the diversity parameter values?

■ The trends of variances of estimated diversity parameters over the number of samples



Observation 4. For some data sets, we can obtain fairly good estimates of diversity parameters by a relatively small number of samples (less than a few thousand samples). In such cases, we may predict the reliability of three-version systems by measuring the diversities from early samples.

5. Experiment Results – Discussion

■ Suggestions for reliable ML system design

- Adopt a three-version architecture, specifically emphasizing TMTI, for improved system reliability.

- Apply reliability models to select the most reliable three-version architecture based on observed diversities.

- For the architecture comparison purpose, a relatively small number of samples may be satisfactory for obtaining reasonable estimates of diversity parameters.

5. Experiment Results – Discussion

■ Limitations

- Our observations are limited to traffic sign image recognition tasks.
- Decision schemes and voting rules for other tasks (e.g., object detection) require further investigation.
- Other system design factors, such as performance, resource consumption, energy, and cost need to be considered together with reliability.

6. Conclusion & Future Work

■ Conclusion

- We investigate the reliability of N-version ML systems and the associated diversity metrics estimated from the empirical data.
- We focus on traffic sign recognition tasks and conduct experiments on five different traffic sign datasets.
- We answer five research questions and give suggestions for reliable ML system design.

■ Future work

- Explore other ML tasks
- Consider the cost and performance of N-version ML systems

Reference

[1] W. Wu, H. Xu, S. Zhong, M. Lyu, and I. King, Deep validation: Toward detecting real-world corner cases for deep neural networks, In Proc. of the 49th IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), pp. 125-137, 2019.

[2] R. S. Ferreira, J. Arlat, J. Guiochet, and H. Waselynck, Benchmarking safety monitors for image classifiers with machine learning, In Proc. of IEEE Pacific Rim International Symposium on Dependable Computing (PRDC), pp. 7-16, 2021.

[3] F. Machida, On the diversity of machine learning models for system reliability, In Proc. of IEEE Pacific Rim International Symposium on Dependable Computing (PRDC), pp. 276-285, 2019.

[4] F. Machida, N-version machine learning models for safety critical systems, In Proc. of the DSN Workshop on Dependable and Secure Machine Learning, pp. 48-51, 2019.

[5] T. Zoppi, A. Ceccarelli, A. Bondavalli, Detecting Intrusions by Voting Diverse Machine Learners: Is It Really Worth?, IEEE Pacific Rim Int'l Symp. on Dependable Computing (PRDC), pp. 57-66, 2021.

[6] H. Xu, Z. Chen, W. Wu, Z. Jin, S. Kuo, M. R. Lyu, NV-DNN: towards fault-tolerant DNN systems with N-version programming, In Proc. of the 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W), pp. 44-47, 2019.

[7] M. Takahashi, F. Machida, and Q. Wen, How Data Diversification Benefits the Reliability of Three-Version Image Classification Systems, IEEE Pacific Rim Int'l Symp. on Dependable Computing (PRDC), pp. 34-42, 2022.

[8] A. Chan, N. Narayanan, A. Gujarati, K. Pattabiraman, S. Gopalakrishnan, Understanding the Resilience of Neural Network Ensembles against Faulty Training Data, In Proc. of 21st International Conference on Software Quality, Reliability and Security (QRS), pp. 1100-1111, 2021.

[9] Q. Wen, F. Machida, Reliability Models and Analysis for Triple-model with Triple-input Machine Learning Systems, In Proc. of the 5th IEEE Conference on Dependable and Secure Computing, pp. 1-8, 2022.

Thank you for your attention!