Robustness and Sensitivity of AI/ML systems - Two Sides of a Coin

An updated version of DNV mini conference on AI assurance 17.11.2021

Shen Yin

Department of Mechanical and Industrial Engineering

Norwegian University of Science and Technology (NTNU)

Content

1 Background & Motivation

2 Positive thinking: Robustness issue

3 Reverse thinking: Sensitivity issue

4 A balance issue

5 Remarks

1 Background & Motivation

- In the era of big data, Artificial Intelligence (AI) shows superior ability for information extraction. From an engineer/user point of view, AI is mostly discussed within Machine Learning (ML) framework.
- In the past few years, we focused on several practical demand driven research projects in manufacturing industry, human medical care, cyber security, etc.
- We use/modify existing ML methods to meet various practical application requirements

Motivation: Based on the previously mentioned experience, we found an interesting phenomenon: Most of the practical requirements we met can be categorized as robustness or sensitivity issues of ML related to data/data-uncertainty. This talk is dedicated to discuss robustness and sensitivity of ML from an engineer/user point view

Manufacturing industry

Human health care

Cyber security

Content

1 Background & Motivation

2 Positive thinking: Robustness issue

3 Reverse thinking: Sensitivity issue

4 A balance issue

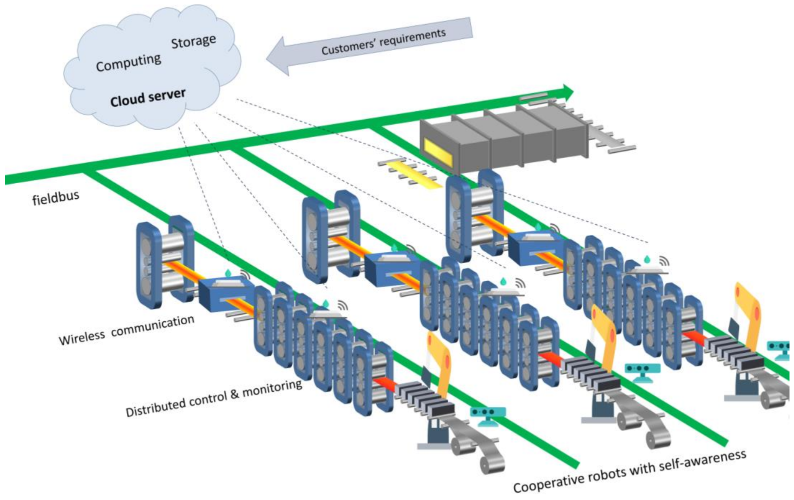
5 Remarks

2 Positive thinking: Robustness issue

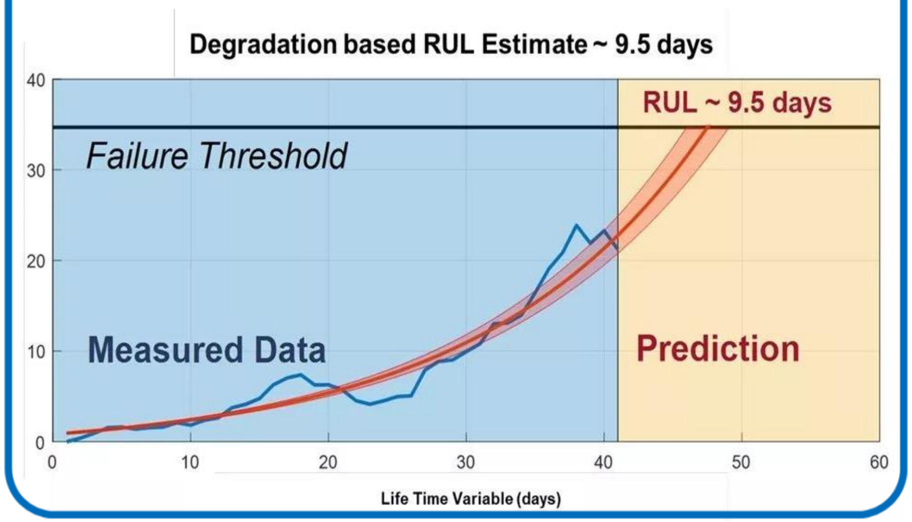
- A positive/straightforward thinking of a designed ML system should show enough robustness to the data uncertainty.
- From a user/engineer perspective, data uncertainty generally includes noise, disturbance, missing data, outlier, and sometimes abnormal behavior (which is not expected in the nominal system).
- A majority of the practical issues can be considered in order to enhance the robustness of the ML system, i.e., a well-designed ML approach should be robust to the data uncertainty.
- Two examples of robustness issue: Remaining useful life prediction and human health diagnosis.

2 Case Ⅰ: Remaining useful life (RUL) prediction

Critical component



Remaining useful life (RUL) prediction

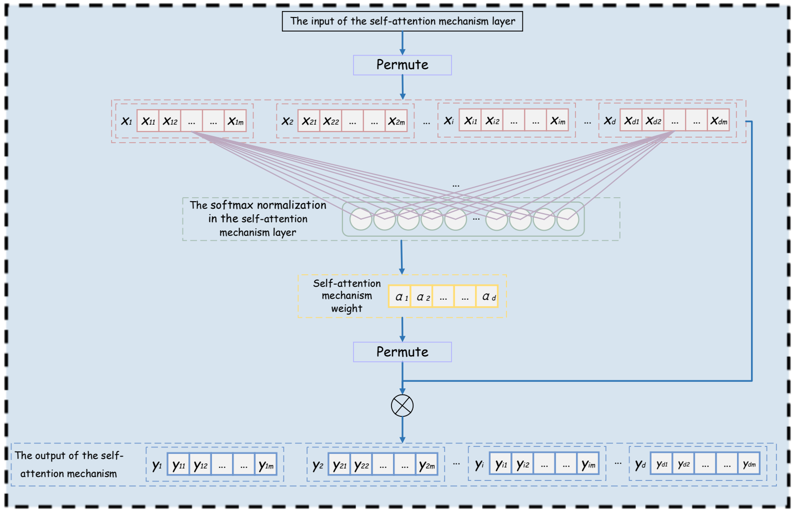


□ **Prediction of RUL is to prognosis future status of critical devices and perform maintenance actions at the appropriate time before system failure. It serves as one of the most important issues in Reliability, Availability, Maintenance and Safety (RAMS).**

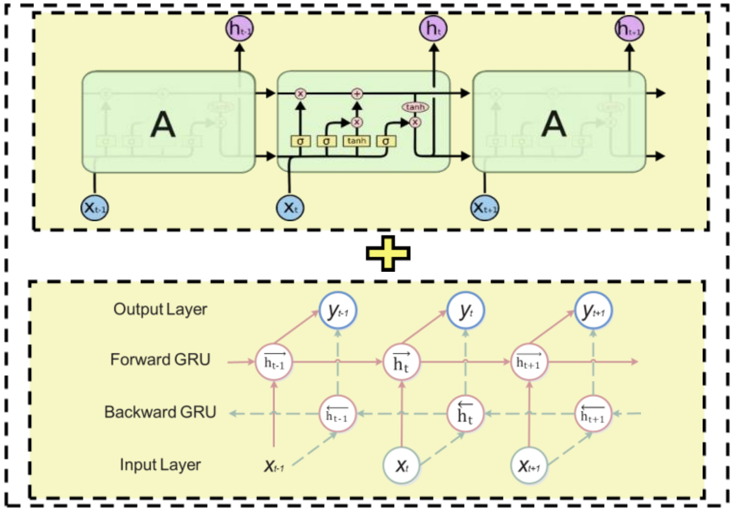
□ **Data uncertainty will bring difficulties for RUL prediction. Not only the missing/outliers/noises, different sampling time etc., but also different data lengths. E.g., huge amount of data collected from many life cycles of a key device. The data might have different lengths for different batches due to various working conditions.**

2 Case Ⅰ: Remaining useful life (RUL) prediction

Temporal self–attention mechanism layer (TSAM)



Bidirectional gated recurrent unit (BiGRU)

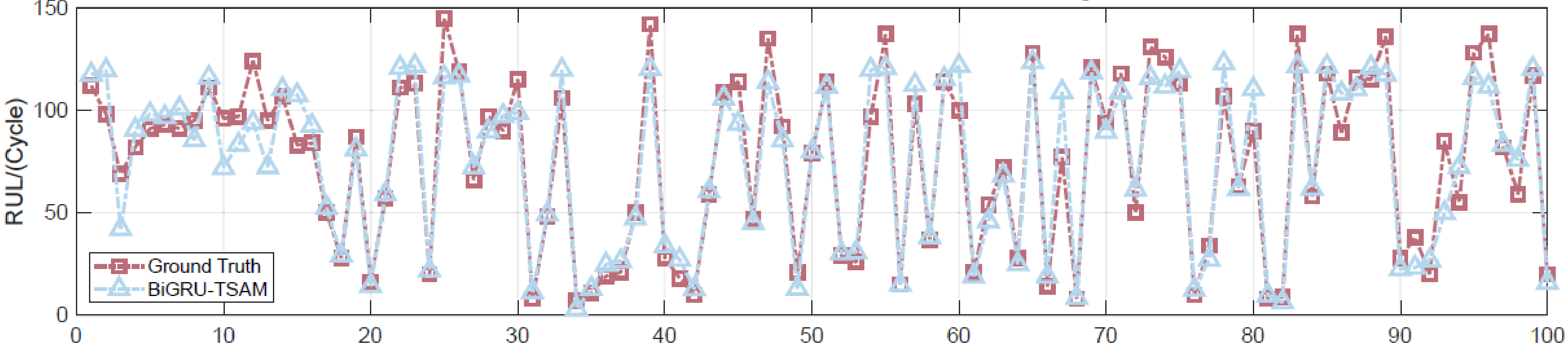


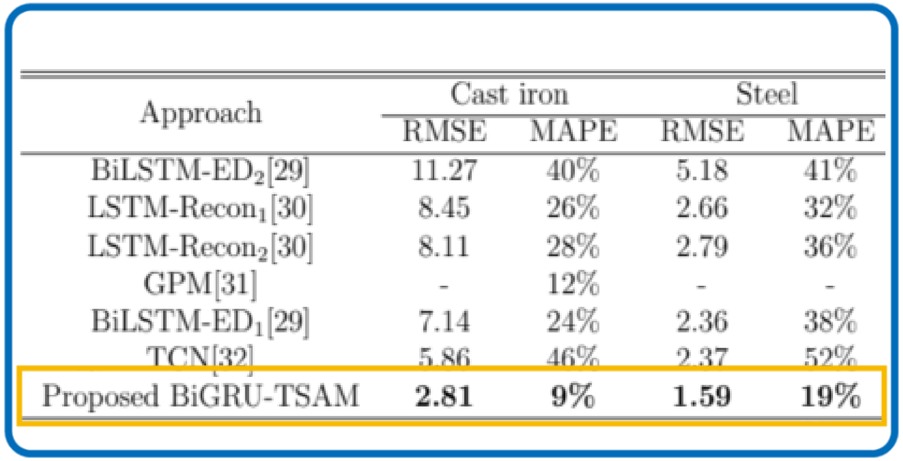
- Bidirectional Gated Recurrent Unit (BiGRU) and attention mechanism are considered to overcome the difficulties and enhance robustness to data uncertainty. BiGRU and attention mechanism are efficient to deal with data with different lengths and in addition extract more important information.

- A temporal self–attention mechanism layer (TSAM) is proposed in order to learn the significance/importance of the dataset (time slots). Then different weights are assigned in order to cope with the high data uncertainty.

2 Case Ⅰ: Remaining useful life (RUL) prediction

Prediction results of different critical components



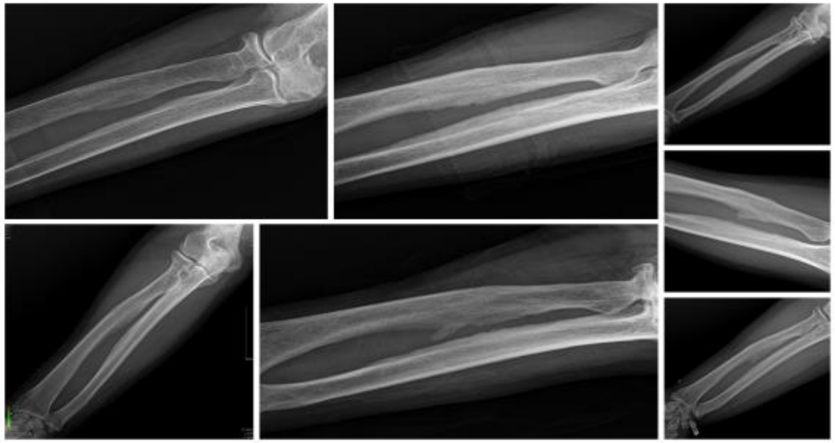


Approach	Cast iron		Steel	
	RMSE	MAPE	RMSE	MAPE
BiLSTM-ED$_2$[29]	11.27	40%	5.18	41%
LSTM-Recon$_1$[30]	8.45	26%	2.66	32%
LSTM-Recon$_2$[30]	8.11	28%	2.79	36%
GPM[31]	-	12%	-	-
BiLSTM-ED$_1$[29]	7.14	24%	2.36	38%
TCN[32]	5.86	46%	2.37	52%
Proposed BiGRU-TSAM	**2.81**	**9%**	**1.59**	**19%**

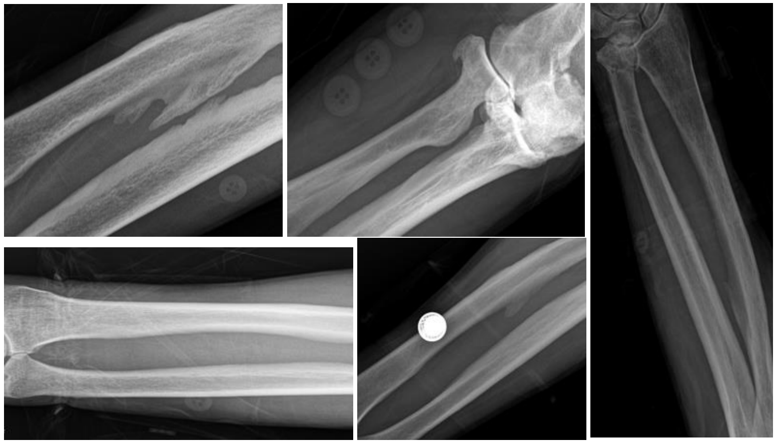
- **The proposed TSAM enhances the robustness related to data uncertainty caused by different data lengths and offer weightings.**
- **It shows that RMSE and MAPE are significantly improved compared with some standard/popular approaches.**

Prediction of remaining useful life based on bidirectional gated recurrent unit with temporal self-attention mechanism. **Reliability Engineering & System Safety** (Revised and Resubmitted)

2 Case Ⅱ: Human health diagnosis



Different position and posture

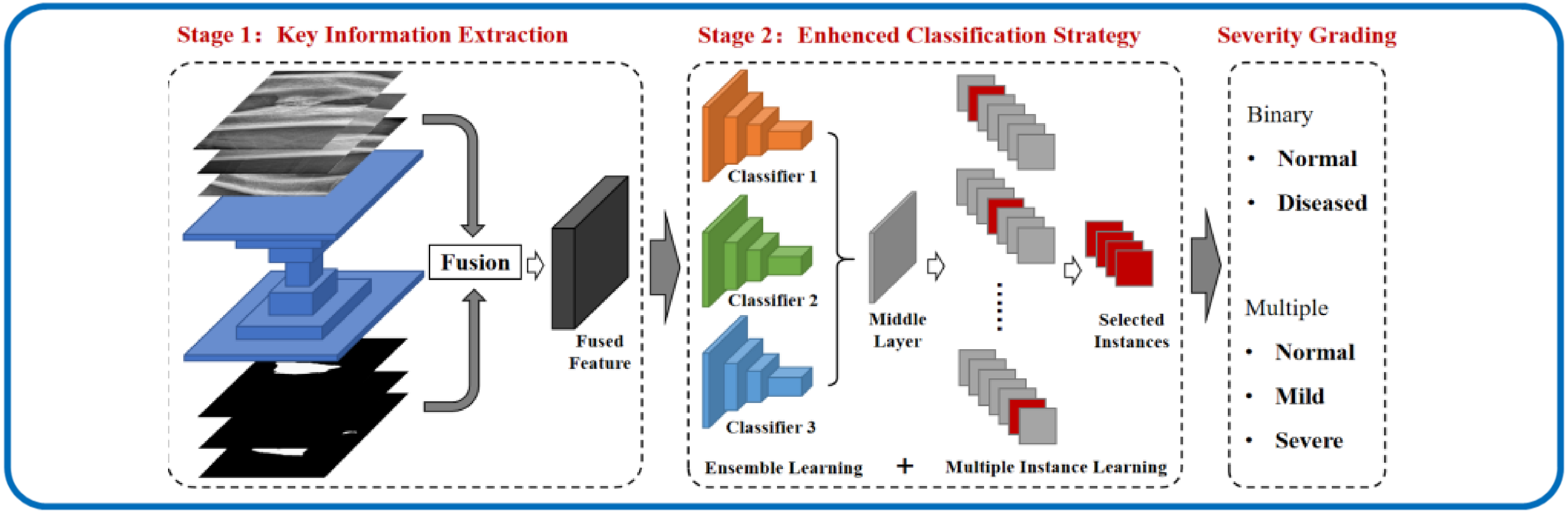


Tissue unrelated to the lesion

- Another application background to enhance robustness to AI/ML algorithm is on human health diagnosis. Especially, for initial screening of high incidence illness diagnosis for economically underdeveloped areas, auxiliary staff (not medical staff) can only use portable X-ray equipment.

- Limited by the environment and quality of portable X-ray/CT equipment, it is expected that the ML approaches shall show high robustness related to uncertainty/disturbances, including different position and posture of patients, tissue unrelated to the lesion (cloths, coins, buttons, etc.)

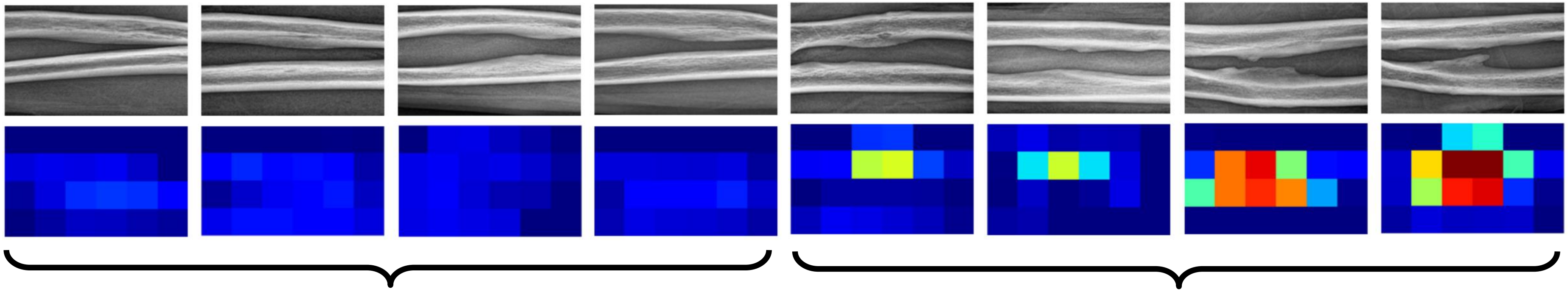
2 Case Ⅱ: Human health diagnosis



- Convolutional Neural Network (CNN) and multi-instance learning are considered to overcome the difficulties and enhance robustness to this kind of data uncertainty. CNN has a nice feature extraction ability for medical images. In addition, multi-instance learning is utilized for classification purposes (final diagnosis).

- An image segmentation network is proposed to get coarse lesion segmentation results and reduce the influence on the unrelated tissues. A multiple parallel neural network is designed as feature extraction process to improve the feature robustness.

2 Case Ⅱ: Human health diagnosis



Normal images

Abnormal images

- The portable equipment with the proposed ML algorithm have been applied in economically underdeveloped areas for initial screening of high incidence illness: Skeletal fluorosis diagnosis. Accuracy for diagnosis is 95.74%, accuracy for severity grading is 90.06%.

- The proposed ML approach could enhance the robustness related to the medical images with different position and posture, and for the images with tissue unrelated to the lesion. In addition, the proposed method can provide heatmaps directly showing the lesion areas.

Integrated learning approach based on fused segmentation information for skeletal fluorosis diagnosis and severity grading, **IEEE Transactions on Industrial Informatics**, 2021, 17(11): 7554-7563. https://doi.org/10.1109/TII.2021.3055397

2 Positive thinking: Robustness issue

Brief summary

- A majority of the practical problems might be summarized as improving the robustness of AL/ML to overcome data imperfections/uncertainties.
- In addition, for a part of practical demands, sensitivity issue also need to be considered. Unlike Robustness, sensitivity offers another way of thinking and can be regarded as the other side of the coin.

Content

1 Background & Motivation

2 Positive thinking: Robustness issue

3 Reverse thinking: Sensitivity issue

4 A balance issue

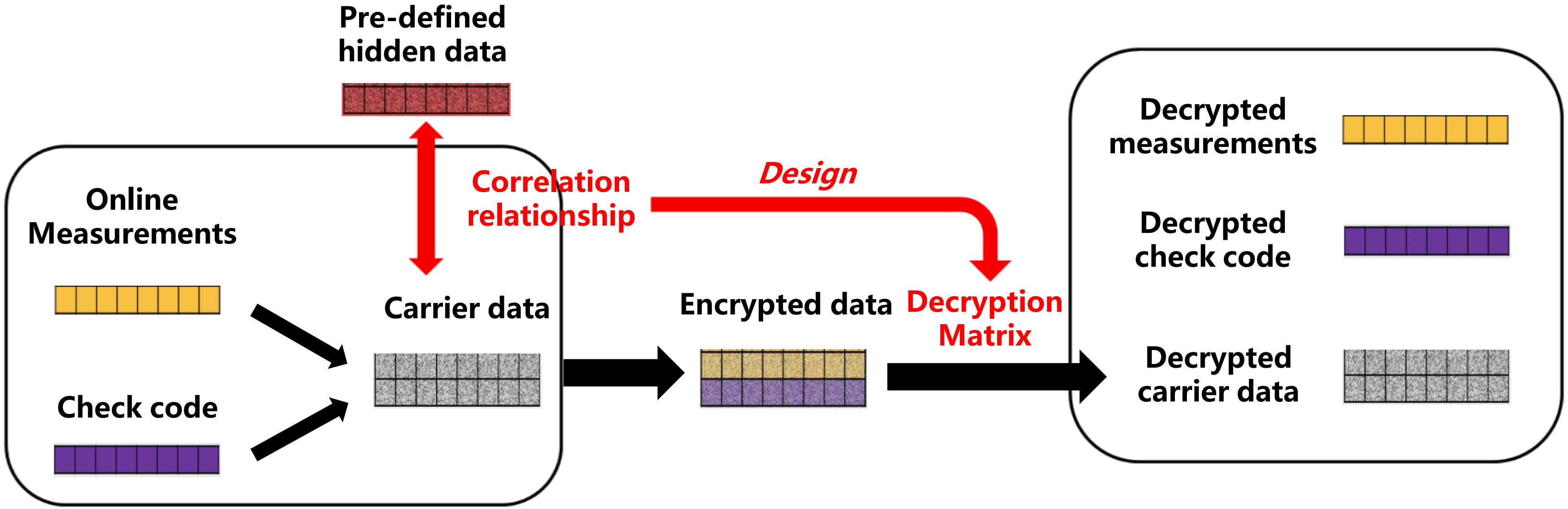
5 Remarks

3 Case Ⅲ: Attacks detection during data transmission

- The risks of data breaches and cyber-physical attacks have become evident in the era of Industry 4.0. Among malicious attacks, spoofing attacks are the most challenges to detect, and generally spoofing attacks will modify the data during transmission.

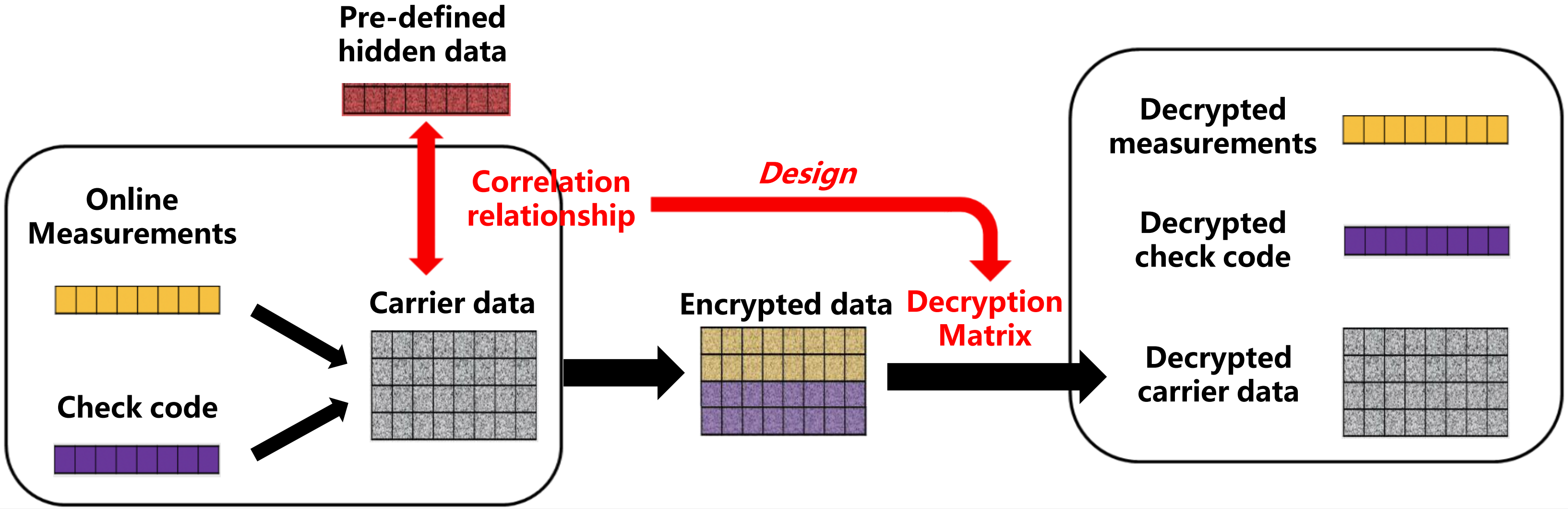
- In fact, if malicious attack can be regarded as a kind of data uncertainty (as an abnormal behavior), the problem of malicious attack detection can be transferred as enhancement of sensitivity related to attacks.

3 Case Ⅲ: Attacks detection during data transmission



- Attack detection V1: Instead of transmitting measurements directly, carrier data are added to the measurements to realize encrypted transmission.
- Carrier data are designed to be fully correlated to a set of predefined hidden data. Meanwhile, the measurements are unrelated to the hidden data. Based on these setup, we propose a multivariate statistical analysis (MVA)-based method to recover the measurements from the encrypted data. In addition, check code is used to judge whether the transmission process is attacked.
- However, we prove that there are many zero elements in the "Decryption Matrix" . As a result, it decreases the sensitivity of attack detection.
- A trick is used to avoid zero elements in the "Decryption Matrix" (use a pair of carrier data for encryption), and the sensitivity of attack detection is much improved.

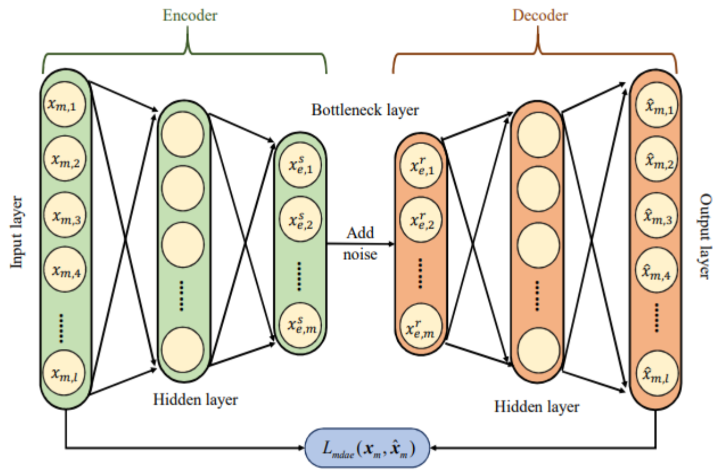
3 Case Ⅲ: Attacks detection during data transmission



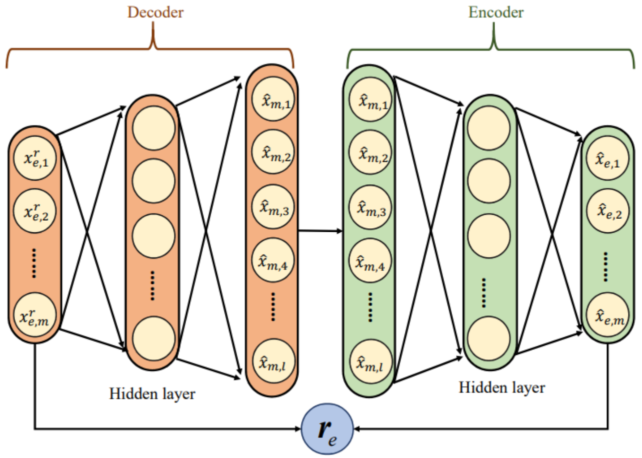
- Attack detection V1: Instead of transmitting measurements directly, carrier data are added to the measurements to realize encrypted transmission.
- Carrier data are designed to be fully correlated to a set of predefined hidden data. Meanwhile, the measurements are unrelated to the hidden data. Based on these setup, we propose a multivariate statistical analysis (MVA)-based method to recover the measurements from the encrypted data. In addition, check code is used to judge whether the transmission process is attacked.
- However, we prove that there are many zero elements in the "Decryption Matrix" . As a result, it decreases the sensitivity of attack detection.
- A trick is used to avoid zero elements in the "Decryption Matrix" (use a pair of carrier data for encryption), and the sensitivity of attack detection is much improved.

3 Case Ⅲ: Attacks detection during data transmission

Autoencoder training



Residual generation



- **Attack detection V2: Autoencoder based methods are considered to enhance the sensitivity to attack. Since the autoencoder is an efficient tool for data reconstruction.**

- **A residual generation autoencoder is proposed. In the training process, the normal data is used to build the autoencoder. In case of attacks, the residual of the middle layer features would be significantly increased. Sensitivity related to the attacks will be enhanced by the study of the residual.**

An integrated data-driven scheme for the defense of typical cyber-physical attacks. **Reliability Engineering & System Safety** (Accepted)

3 Reverse thinking: Sensitivity issue

Brief summary

- A part of the practical problems can be categorized as a sensitivity issue. To improve/enhance the sensitivity of the designed ML approach serves as an efficient tool for malicious attack detection, fault detection and classification. Multivariate statistical analysis (MVA)-based methods and autoencoder based approaches show superior performance compared based on the case studies.
- A balance issue is also worth a particular consideration: From selected objectives perspective OR an adversarial perspective.

Content

1 Background & Motivation

2 Positive thinking: Robustness issue

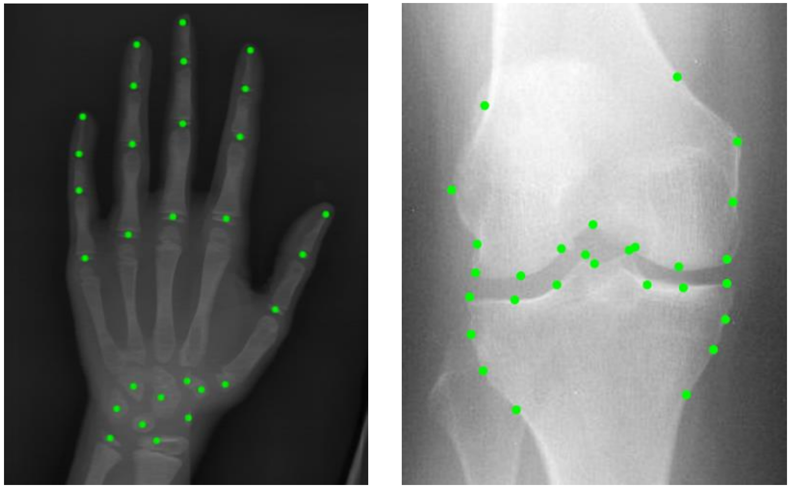
3 Reverse thinking: Sensitivity issue

4 A balance issue

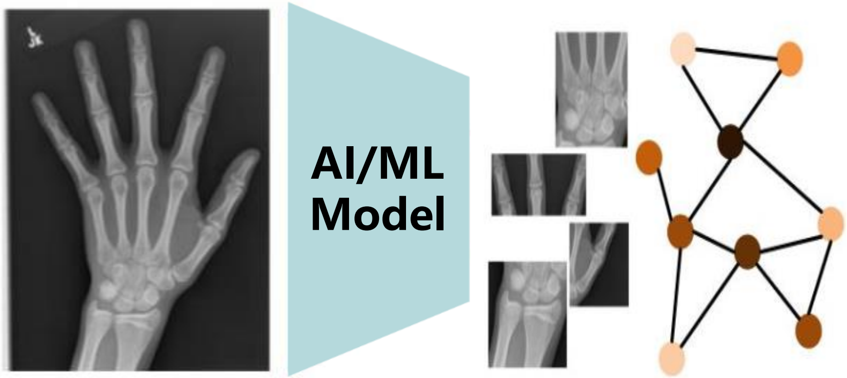
5 Remarks

4 Case IV: Key points in human health diagnosis

Key points



Sensitive to areas of interest through key points



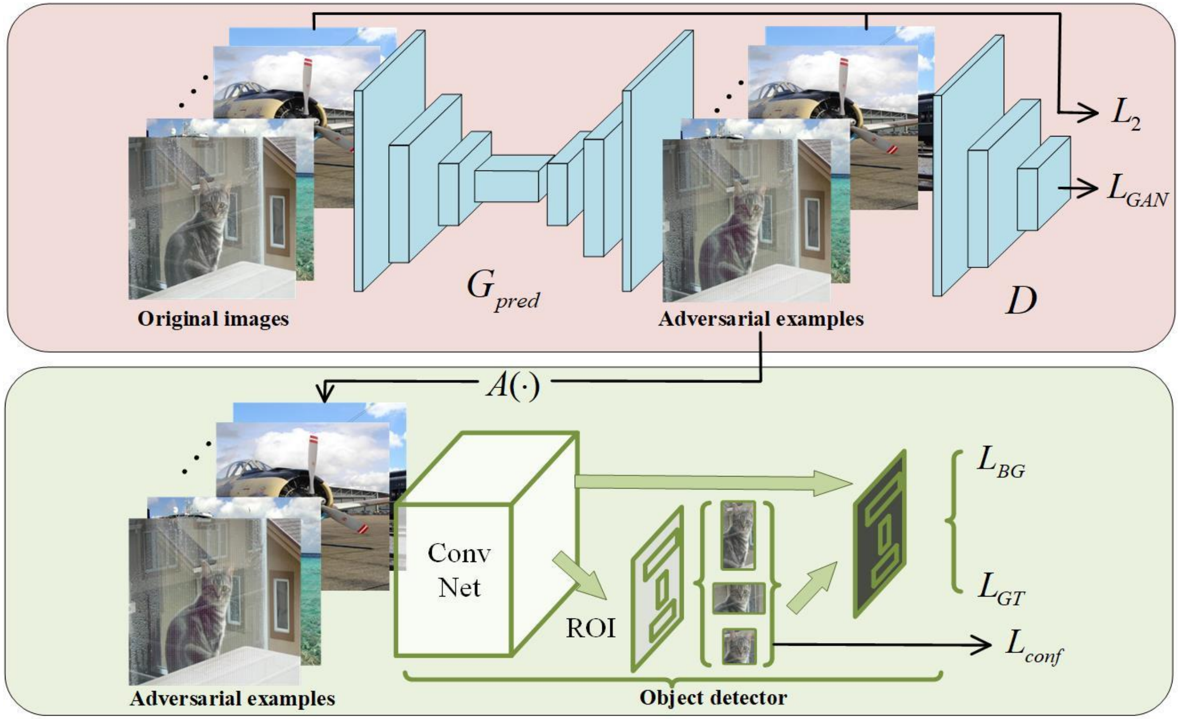
- Most of the image based human health diagnosis approaches focus on certain selected objectives/areas. It is expected that ML approach is sensitive to the objectives of interesting but robust to the others.
- Key points could be proposed for the balance purpose. The labeled key points on the image make ML approach only sensitive to special areas.
- In processing work: Key Points aided ML approach to cope with medical images for human health diagnosis. The diagnostic performance shows significant improvement compared with other popular approaches.

KRGCN: Key Region Aggregation Graph Convolutional Network for Bone Age Measurement. **IEEE Transactions on Instrumentation and Measurement** (Submitted)

4 Case V : Balance issue: Adversarial perspective

- □ What kind of data uncertainty could breakdown the robustness of the designed ML approaches?
- □ What kind of data uncertainty could breakdown the sensitivity of the designed ML approaches?
- □ Does such kind a of uncertainty exist? If so, from an adversarial perspective, it might be helpful to build a KPI to evaluate/validate the designed ML approach.
- □ We performed some interesting experiments and found that there exists certain artificial data uncertainty, which could breakdown the robustness of the popular ML approaches.

4 Case V : Balance issue from adversarial perspective



$$L_{GAN}(G,D) = \min_{G} \max_{D} E_I[logD(I)] + E_I[log(1 - D(G(I))]$$

Original purpose of GAN

$$L_{GT} = maxE_\theta[\frac{1}{k}\sum_{j=1}^{K} -logPr(C_i|F_j(\hat{I})]$$

$$L_{BG} = minE_\theta[\frac{1}{k}\sum_{j=1}^{K} -logPr(C_b|F_j(\hat{I})]$$

The proposed bidirectional adversarial attack loss function

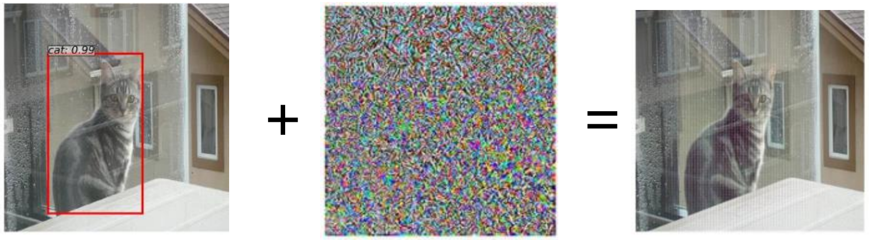
- □ **An adversarial generation method based on the generative adversarial network (GAN) is proposed.**
- □ **The original purpose of GAN is to generate image (will introduce minor data uncertainty) similar to the original image. The proposed bidirectional adversarial attack loss function makes the data uncertainty have a specific purpose: Breakdown the robustness of the designed ML.**
- □ **Bidirectional adversarial attack loss function: Maximize the background probability and minimize the ground truth probability simultaneously.**

4 Case V : Balance issue from adversarial perspective

Can be recognized

Data uncertainty

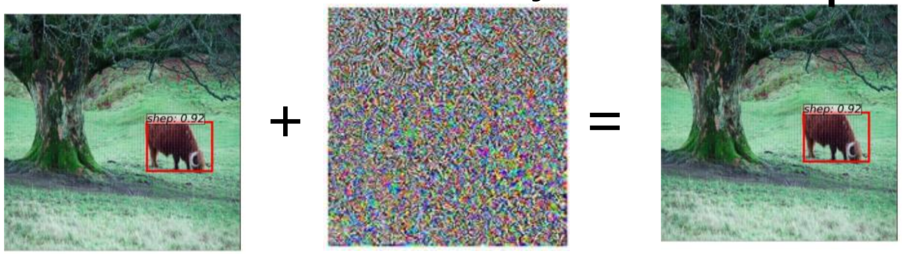
Can not be recognized

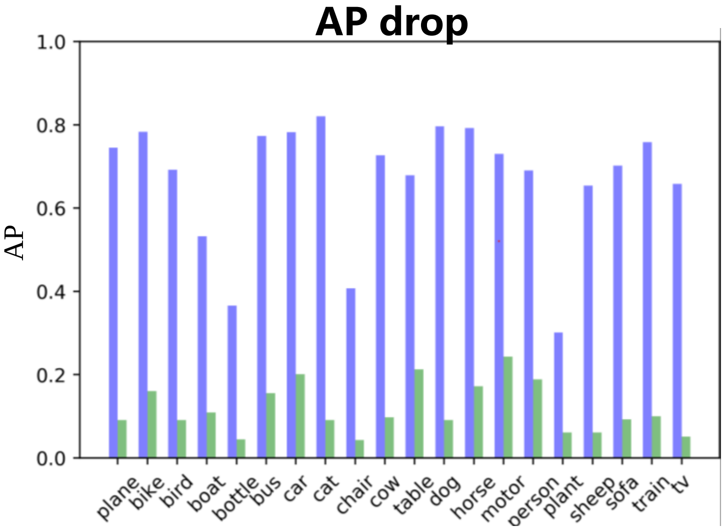


Be recognized as cow

Data uncertainty

Be recognized as sheep





Dataset : PASCALVOC

http://host.robots.ox.ac.uk/pascal/VOC/

AP : Average Precision

🟦 **: results of Faster RCNN**

🟩 **: results in case of data uncertainty**

The data uncertainty generated by the proposed approach will destroy Faster RCNN.

Playing against deep neural network-based object detectors: A novel bidirectional adversarial attack approach, **IEEE Transactions on Artificial Intelligence**, 2021. https://doi.org/10.1109/TAI.2021.3107807

4 A balance issue

Brief summary

- For human health diagnosis, key points are used to make ML sensitive to certain areas of interest but robust to others.
- From the adversarial perspective, there exists certain data uncertainty, which could destroy the robustness of the underlying ML system.
- It is still possible that there exists such a kind of data uncertainty, which could destroy sensitivity of the underlying ML systems.

Content

1. Background & Motivation
2. Positive thinking: Robustness issue
3. Reverse thinking: Sensitivity issue
4. A balance issue
5. Remarks

5 Remarks

Remarks:

- Robustness and sensitivity are two sides of a coin. From a user/engineer perspective, most of practical demands related to ML/AI approaches could be categorized.
- Challenge I: How to measure data uncertainty?
- Challenge II: How dose data uncertainty influence the underlying ML/AL analytically?
- Challenge III: What are KPIs to design an AI/ML/DT system? What are the most important KPIs for evaluation/verification? Any analytical and interpretable solution?

Thank you for comments!