인공지능 V&V 방법 및 방향 제시

SEUNG JUN LEE
Department of Nuclear Engineering, UNIST

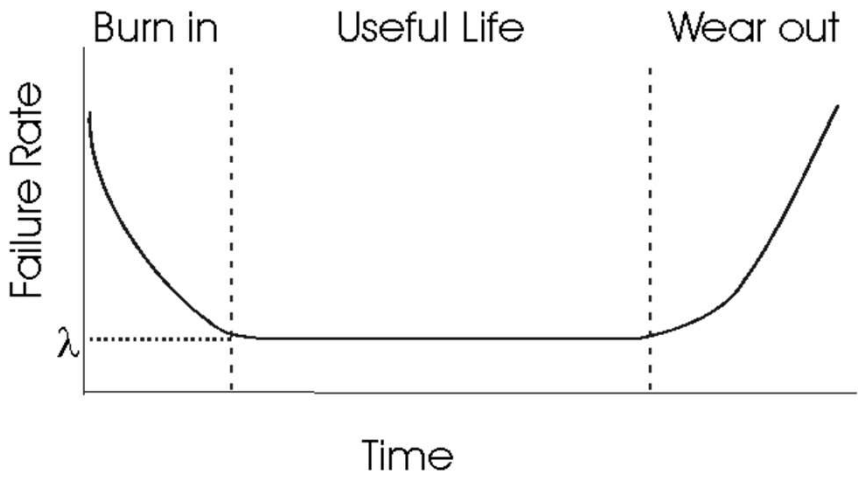
SW Reliability

Hardware	Software
Failures can be caused by deficiencies in design, production, use, and maintenance.	Failures are primarily due to design faults.
Failures can be due to wear or other energy-related phenomena.	There are no wear-out phenomena.
No two components are identical.	There is no variation.
Repairs can be made to make equipment more reliable, as in the case of preventive maintenance.	There is no preventive maintenance for software.
Reliability can depend on burn-in or wear-out phenomena.	Reliability is not so time-dependent.
Reliability may be related to environmental factors such as temperature, vibration, humidity, etc..	External environment does not affect reliability except insofar as it might impact program inputs.
Reliability can be improved by redundancy.	Reliability cannot be improved by redundancy if parallel code paths are identical.
Failures can occur in components of a system in a pattern that is, to some extent, predictable from the stresses on the components and other factors.	Failures are rarely predictable from analysis of code lines on a line-by-line basis.

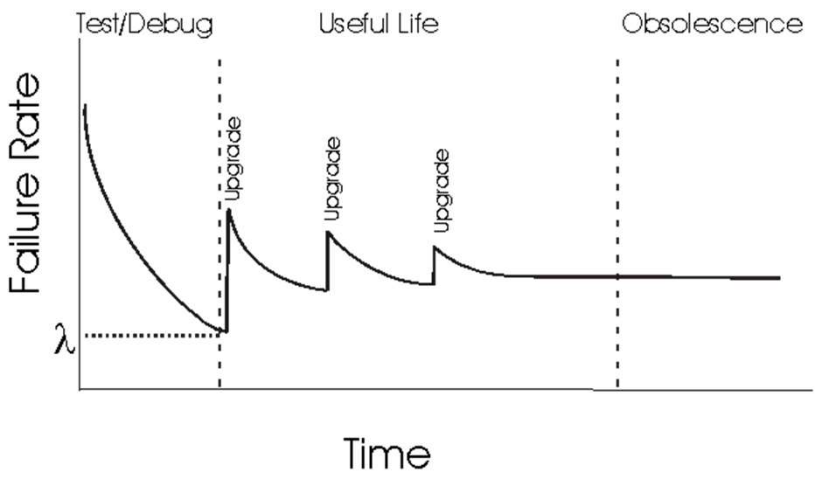
Specific differences between hardware and software reliability

SW Reliability

- SW faults
  - Software faults are design faults caused by human error
  - Hardware reliability vs Software reliability



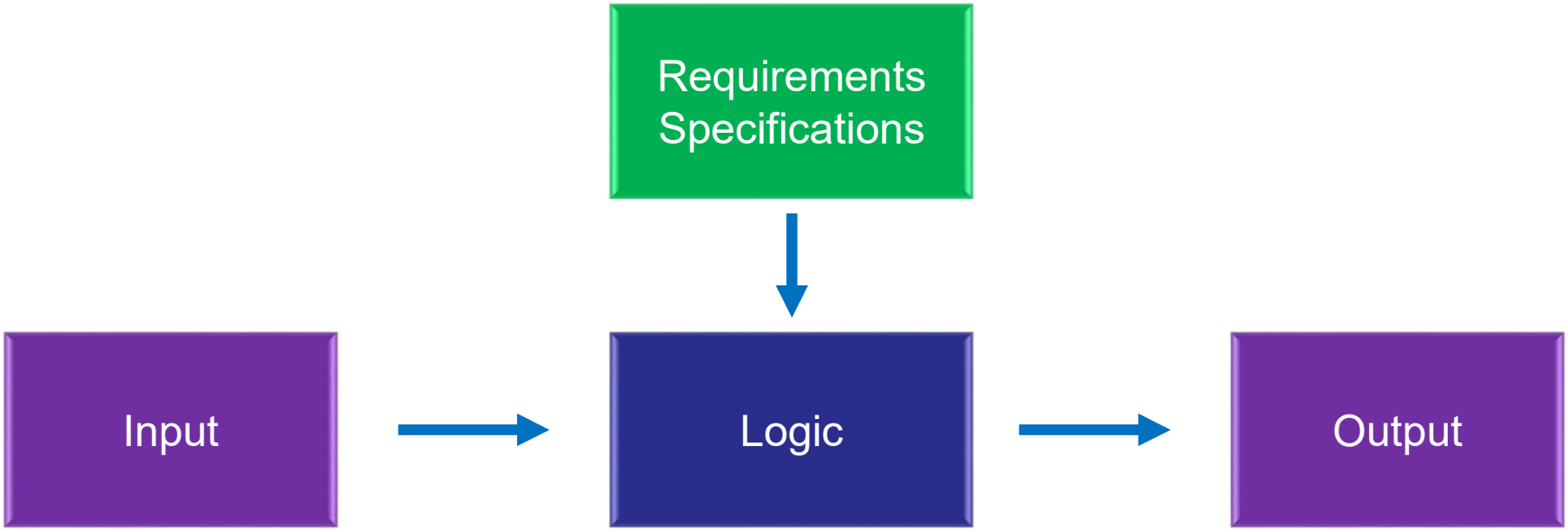
Hardware reliability



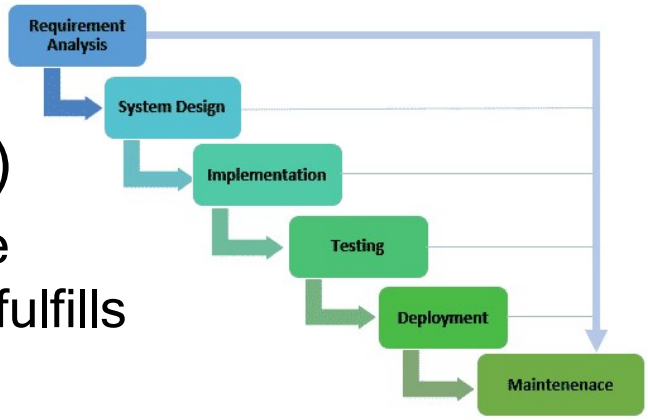
Software reliability

SW V&V

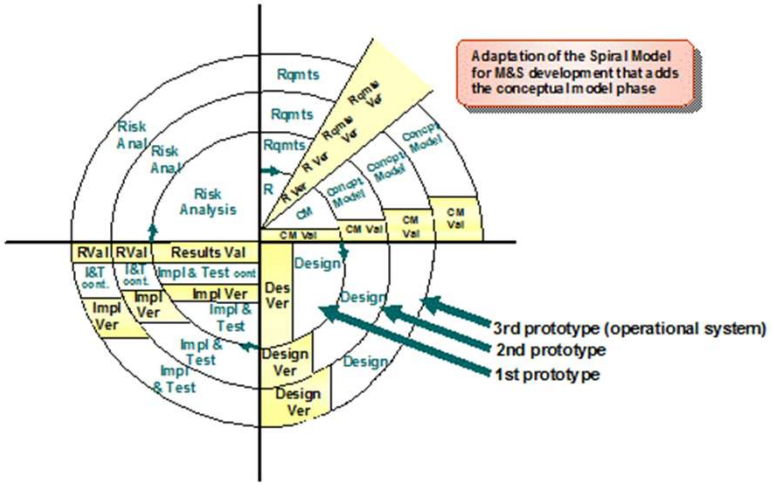
- SW always generates the same output for the given input.

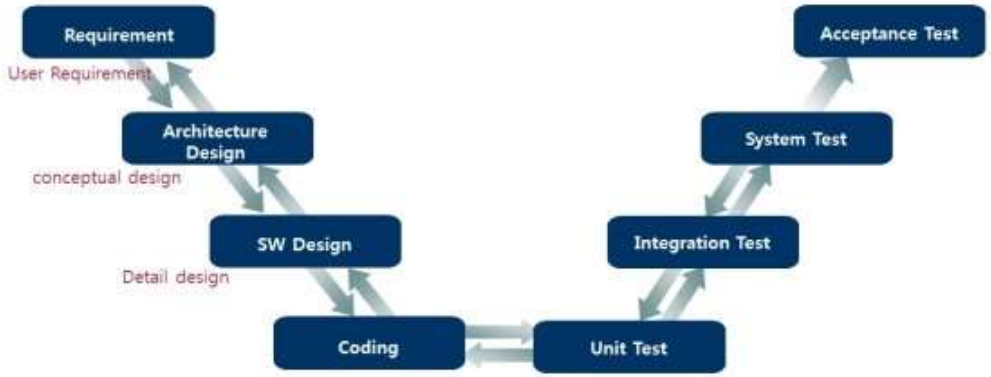


SW V&V

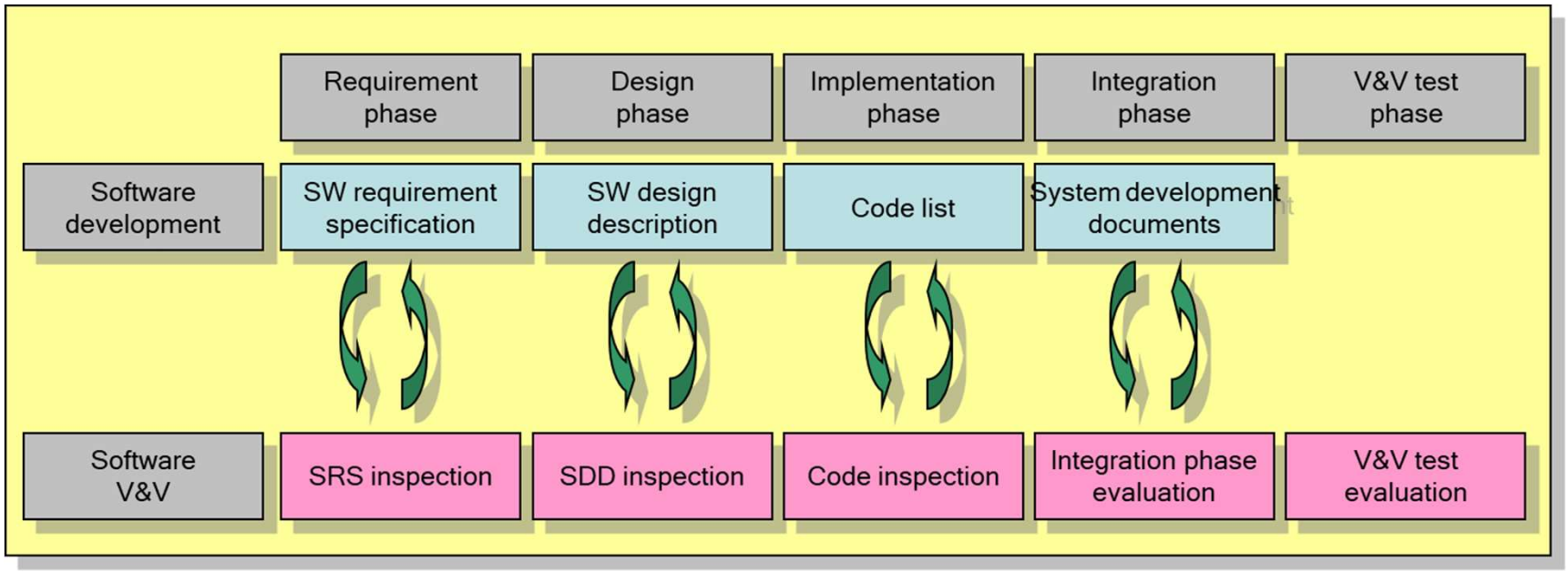


- SW verification and validation (V&V)
  - The process of checking that a software system meets specifications and that it fulfills its intended purpose.
  - Verification: Are we building the product right?
  - Validation: Are we building the right product?





Software Development Life Cycle



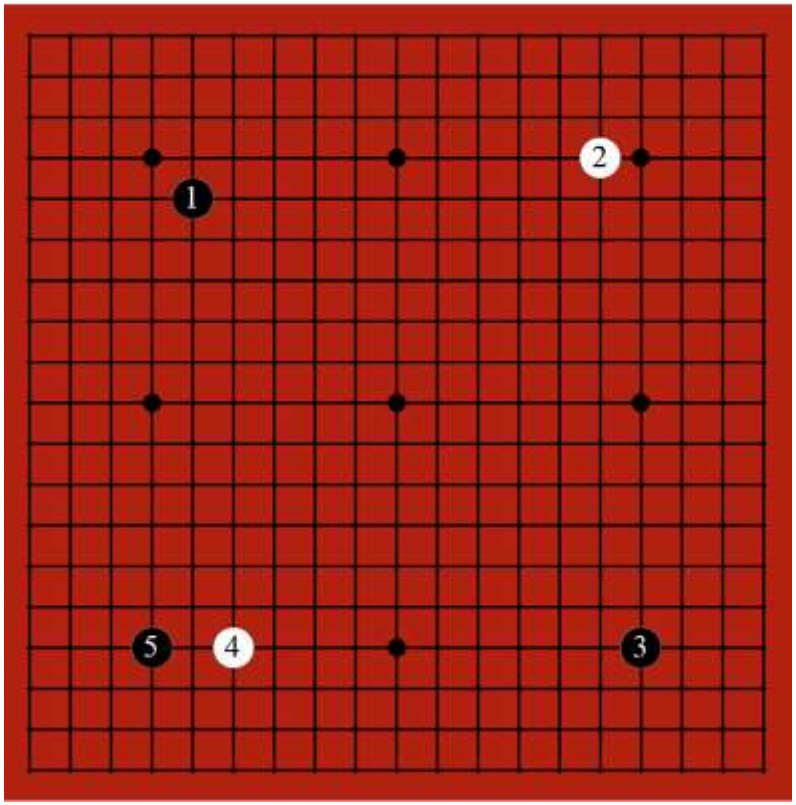
What is V&V for AI SW?

What is AI for?

- Not for simple problems
- Not for problems which can be solved with clear logics or equations

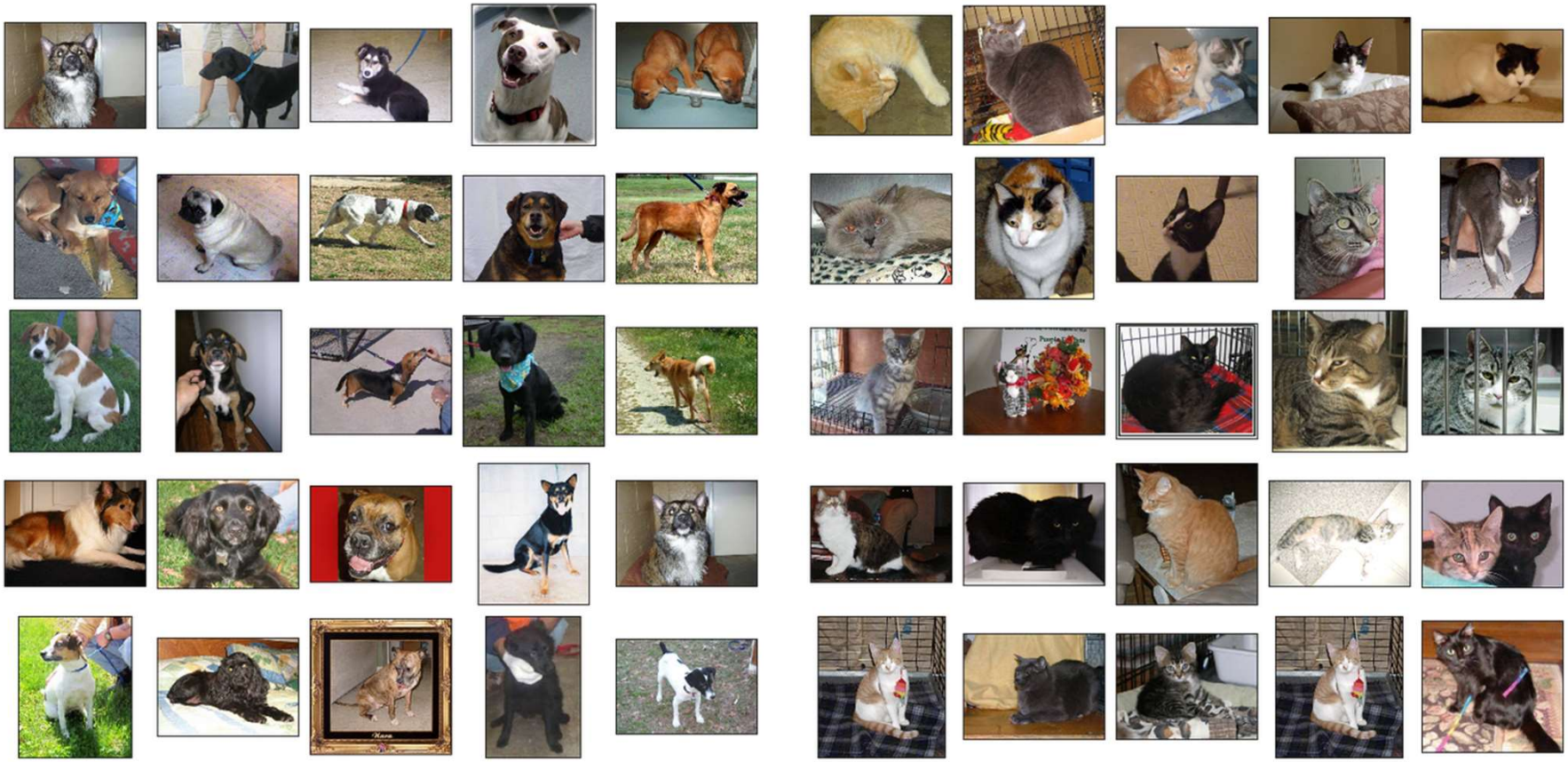
Go (바둑)

경우의 수 = 361×360×359×358×357×…



- 1.4379232588843891 E 768

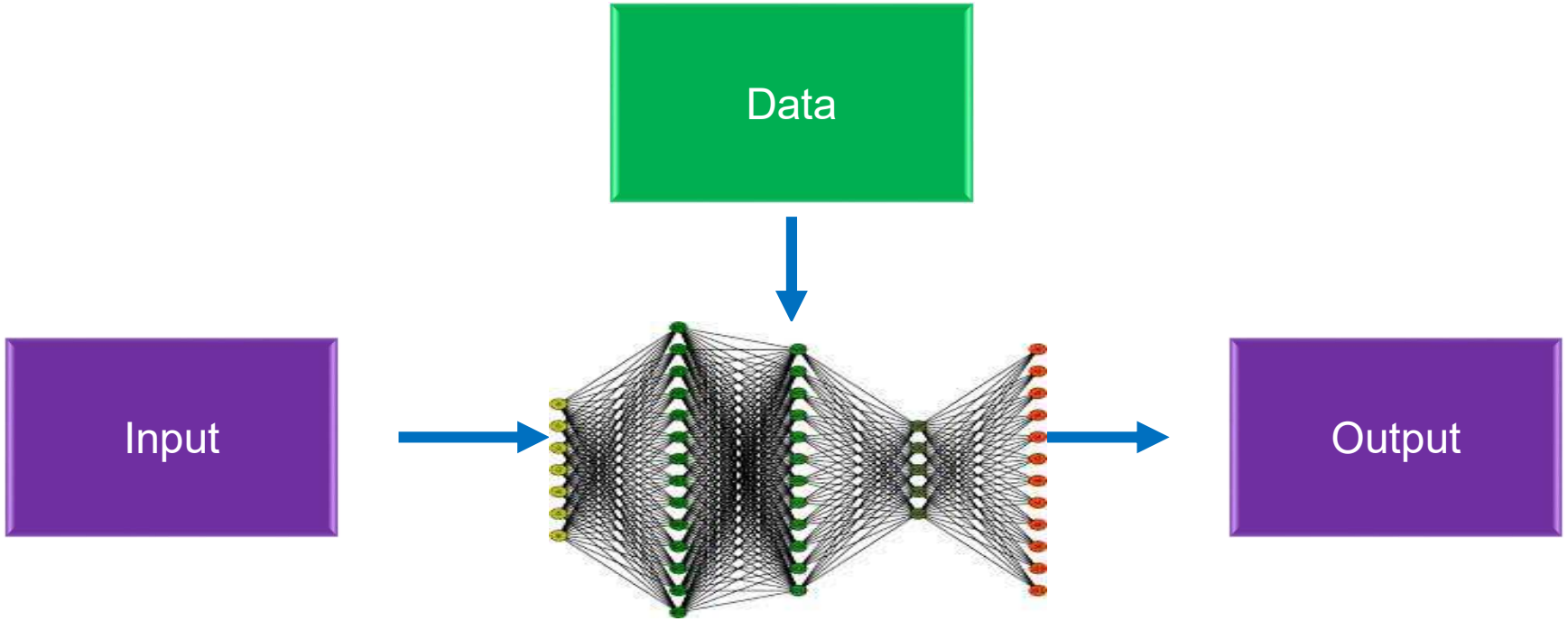
Dogs and Cats



- Cannot be solved with a rule-based approach

Artificial Intelligence

- To train the AI model using the given data
  - Requirements and specifications → Data
  - Logic → Trained model



SW V&V and AI V&V

- SW V&V is to check the integrity of each phase.
  - Requirements and specifications
  - Design
  - Implementation
  - Testing

- AI V&V
  - ??

AI V&V?

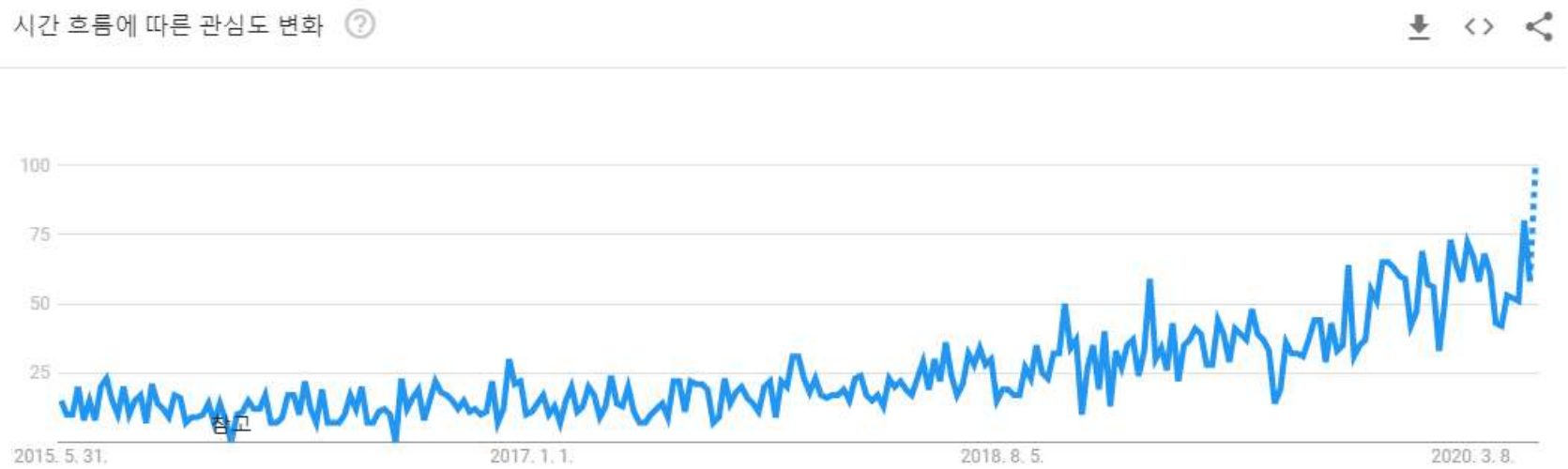
- V&V of AI or V&V of human
  - Artificial Intelligence is a model of the neural network in a human brain
- Cannot be done by logic verification
  - If the trained model can be verified with logics or equations, then it is not an appropriate problem for using AI
- Code verification is not meaningful.
- Validation through testing is possible

```
model.add(LSTM(100, input_shape=(self.config.input_time_step, int(len(self.config.input_para)))
model.add(Dropout(0.3))
model.add(LSTM(100, return_sequences=True))
model.add(Dropout(0.3))
model.add(LSTM(100, return_sequences=True))
model.add(Dropout(0.3))
model.add(LSTM(100, return_sequences=False))
```

XAI: Explainable AI

XAI

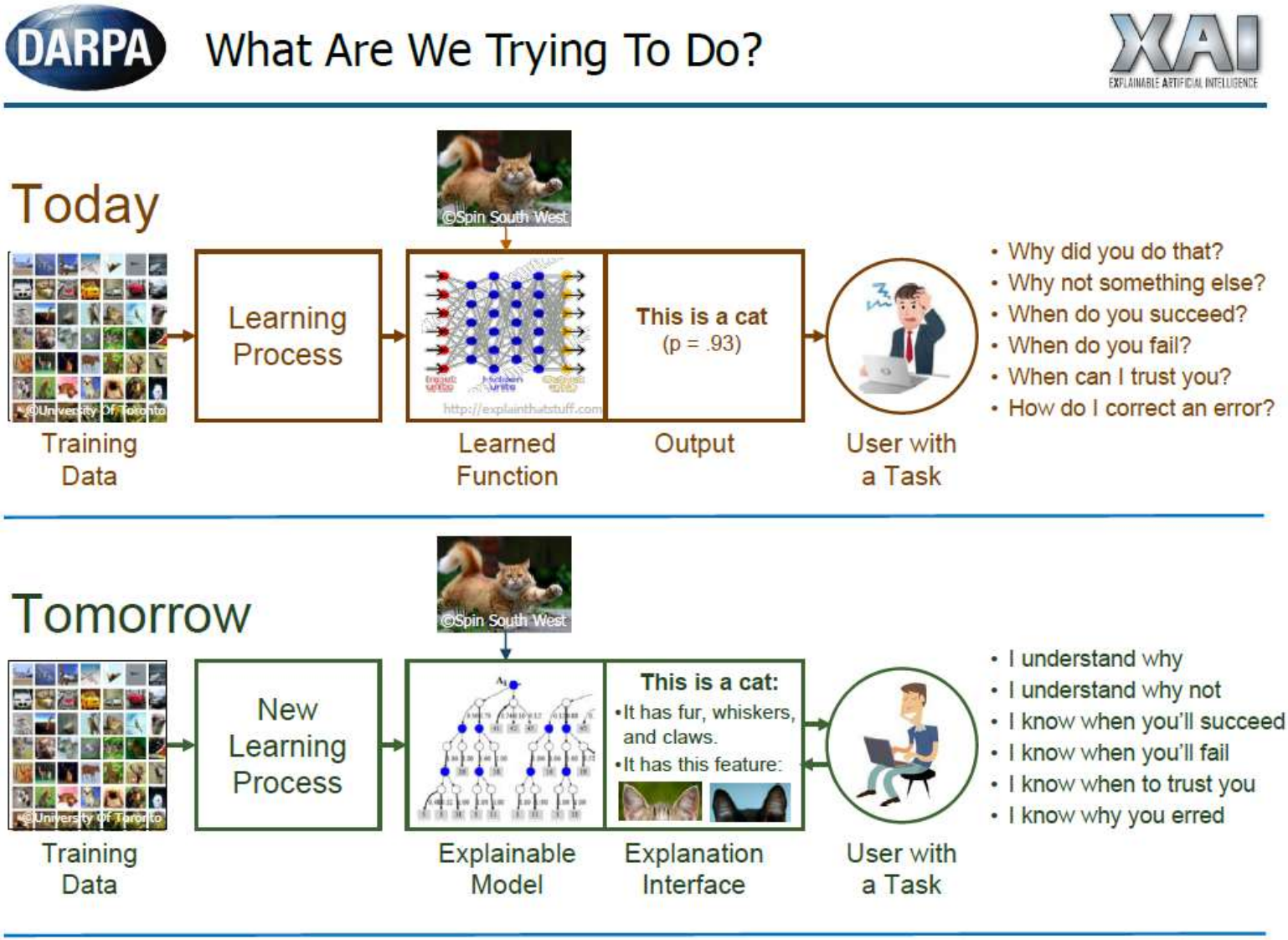
- AI applications are not able to explain their autonomous decisions and actions to human users.
- For certain AI applications, explanations may not be essential, however, explanations are essential for many critical applications including a NPP.
- Interest about explainable AI (XAI, 설명가능 인공지능) has increased.



주제 '설명가능 인공지능' 관심도 변화 (2015~)

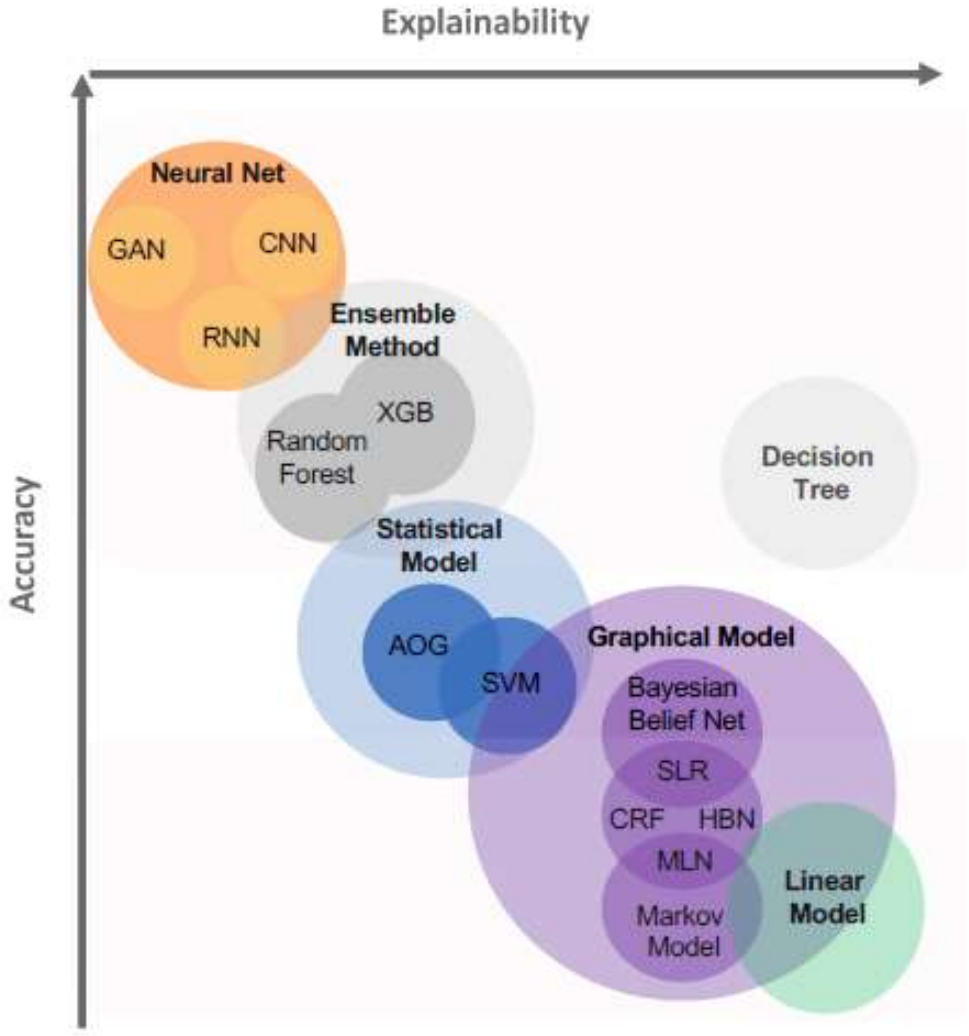
Ref. Google Trend - https://trends.google.co.kr/trends

XAI



Ref. Explainable Artificial Intelligence (XAI), David Funning, DARPA/I2O

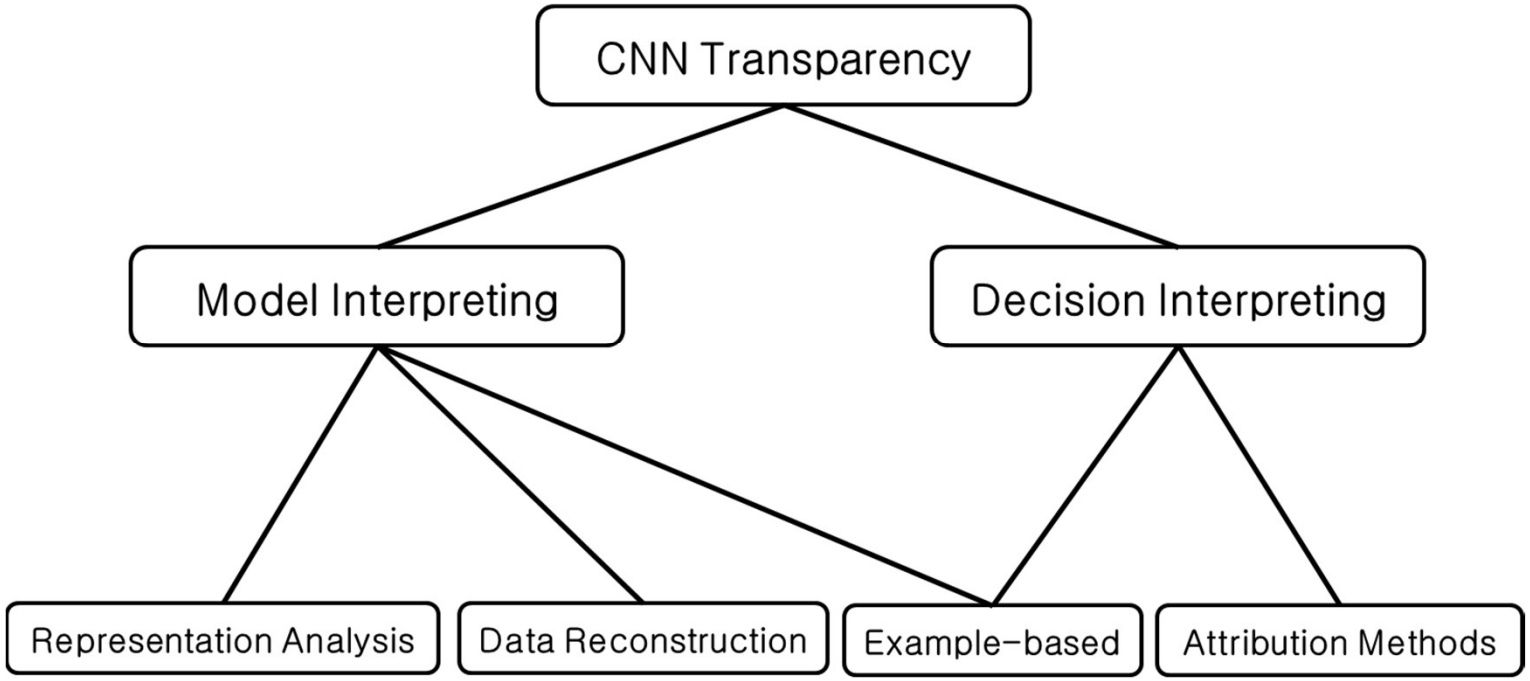
XAI



1) XAI—Explainable artificial intelligence, Gunning et al., Sci. Robot. 4, eaay7120 (2019)

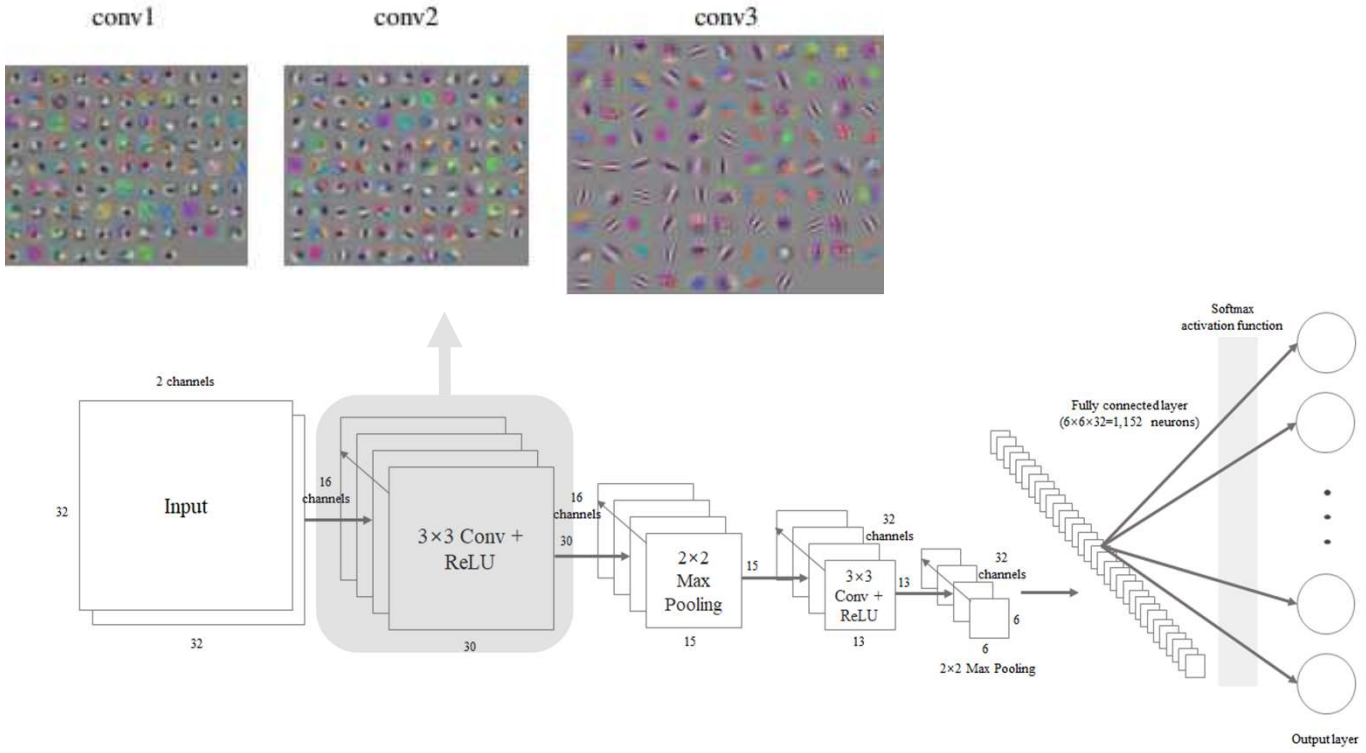
Introduction

- Transparency of deep learning neural network (DNN)
  - Explanation of model
    - What kind of input maximally activate a particular neuron?
  - Explanation of predictions
    - Why does the model arrive at this particular prediction?



CNN transparency

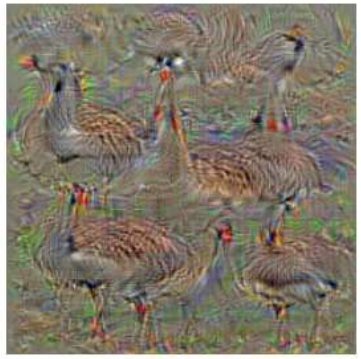
- Weight Visualization
  - Visualize adjusted weight of trained convolutional layer
  - Understand what CNN has learned
  - Explainable?



Ref. Striving for simplicity: the all convolutional net, Jost Tobias et al., University of Freiburg ,2015

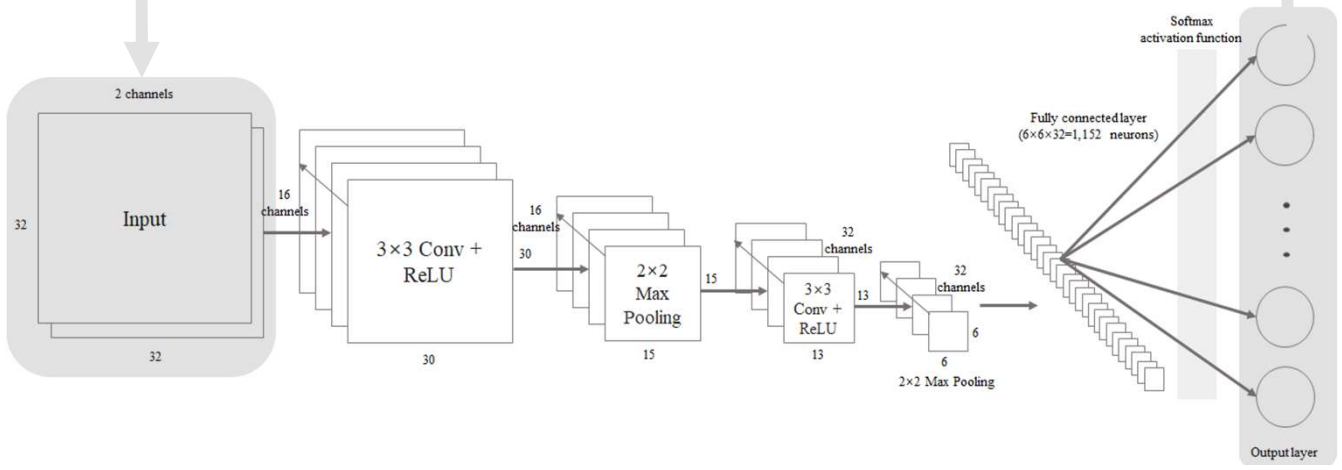
CNN transparency

- Activation Maximization
  - Generate a pattern (arbitrary input) such that maximize an activation of certain output class and minimize activations of other classes.
  - Goose? (Lower the quality of the interpretation for given classes)



$$\max_{x \in \mathcal{X}} p_\theta(\omega_c \mid x)$$

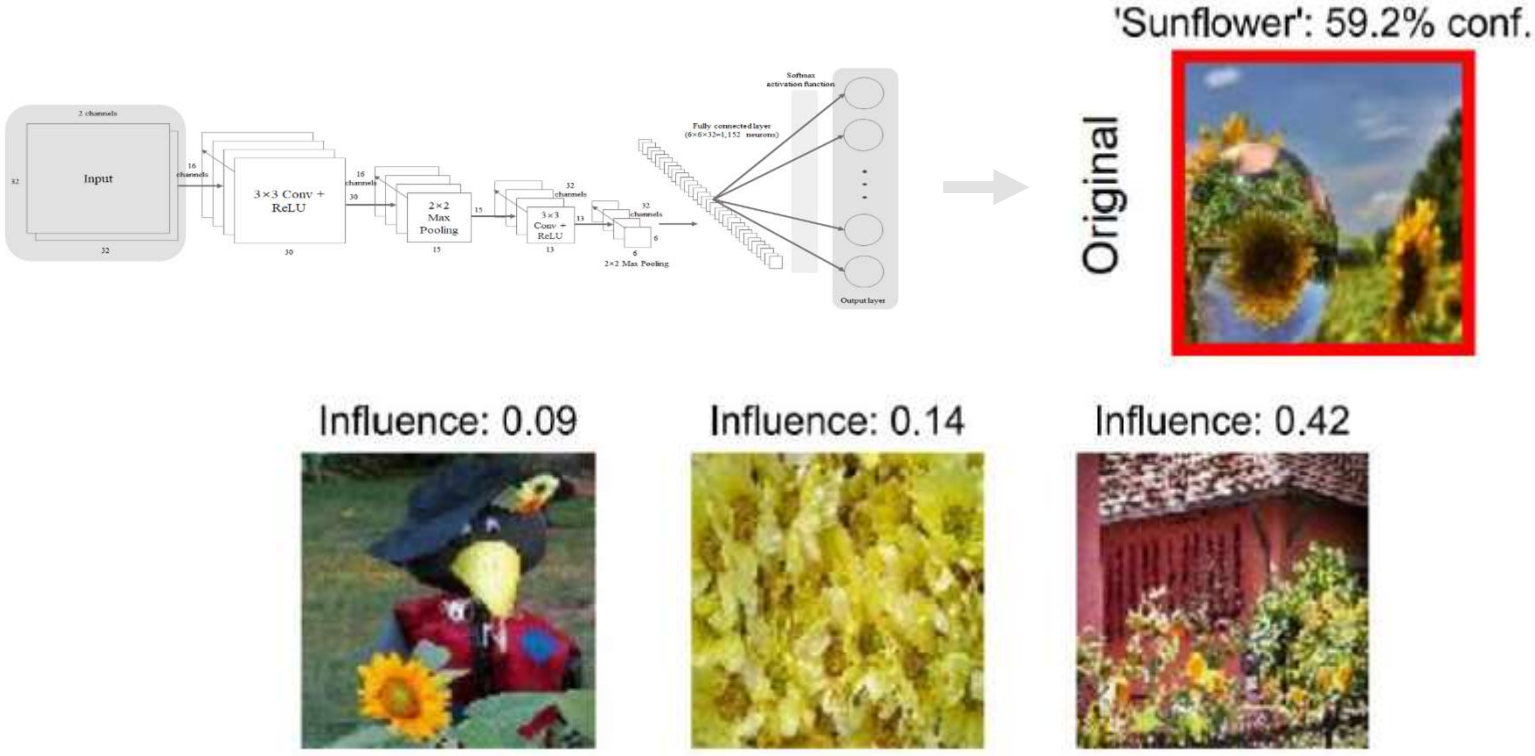
Dumbel = 0
Cup = 0
Fox = 0
Goose = 1
Lemon = 0



Ref. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, Jost Tobias et al., University of Freiburg ,2015

CNN transparency

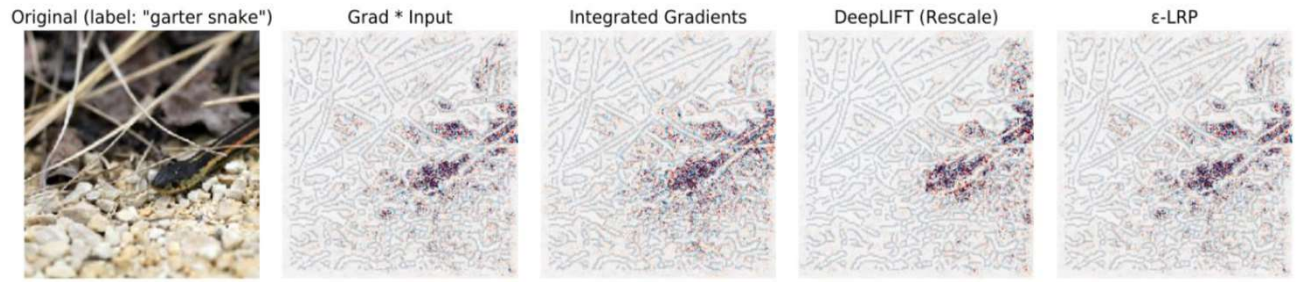
- Find influential training image for certain decision
  - Find training instance influenced the decision most.
  - Still does not specifically highlight which features were important.

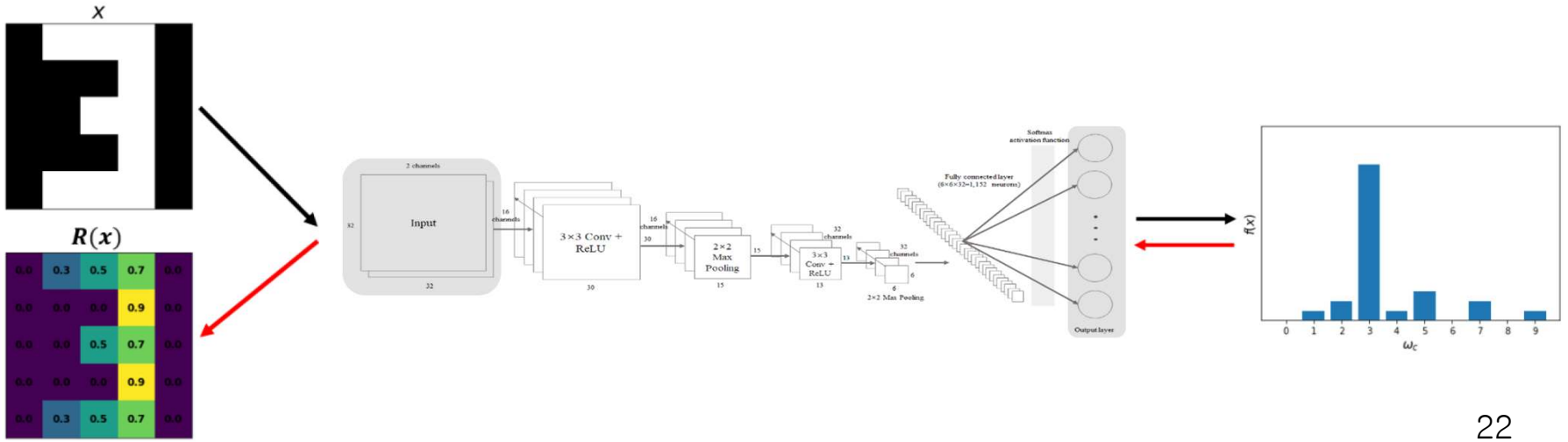


Ref. "Understanding Black box Predictions via Influence Functions ", https://arxiv.org/pdf/1703.04730.pdf

CNN transparency

- Attribution Method
  - Find the attribution of each pixels for certain decision
  - Generally visualized as heatmaps
  - Saliency map with backpropagation algorithm is the baseline method.

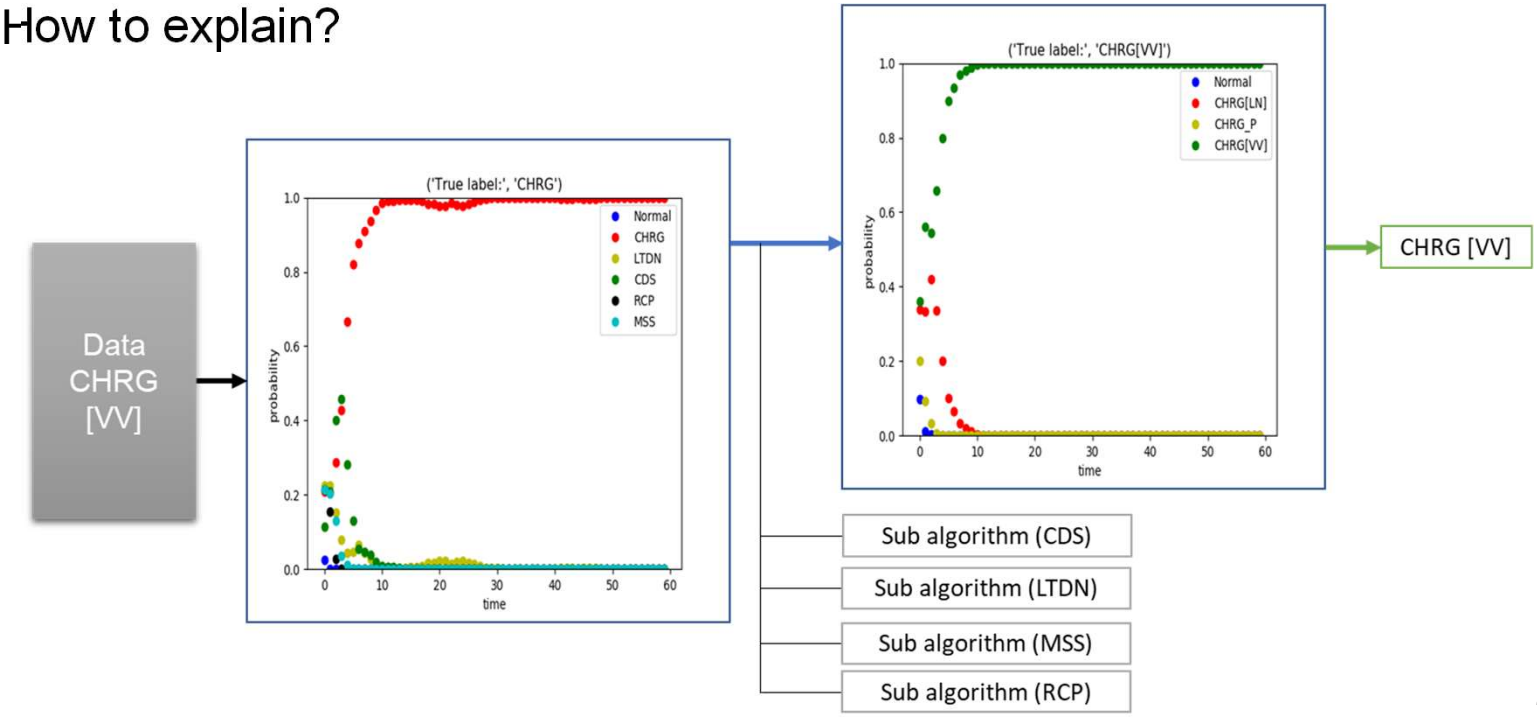




XAI for Abnormal Diagnosis Model

- Total 82 abnormal operating procedures (AOPs) in APR-1400
- Stage : Indicating specific cause of events → 224 Stages

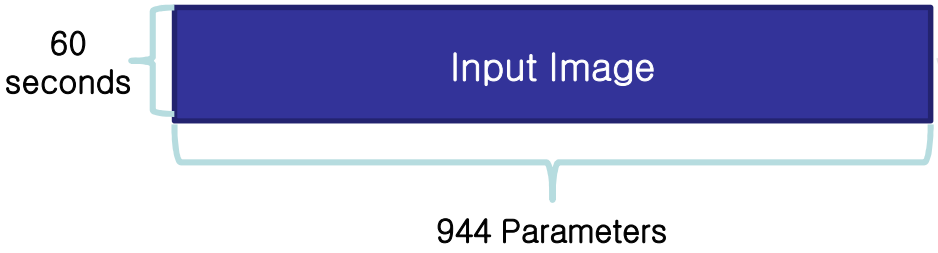
- Abnormal Diagnosis using AI
  - Good performance
  - How to explain?

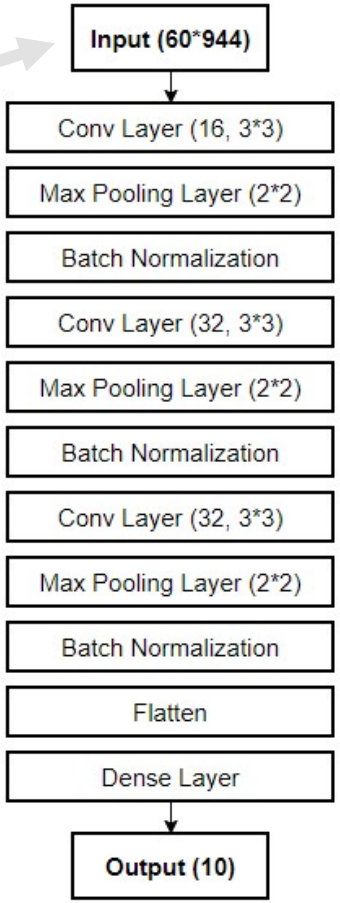


XAI for Abnormal Diagnosis Model

- CNN model for abnormal diagnosis
  - Trained for 1 normal and 9 abnormal states
  - Input shape : 60 seconds x 944 parameters
  - More than 99 % accuracy

Acc	Loss	Val_acc	Val_loss
1	0.00024	0.9933	0.0081

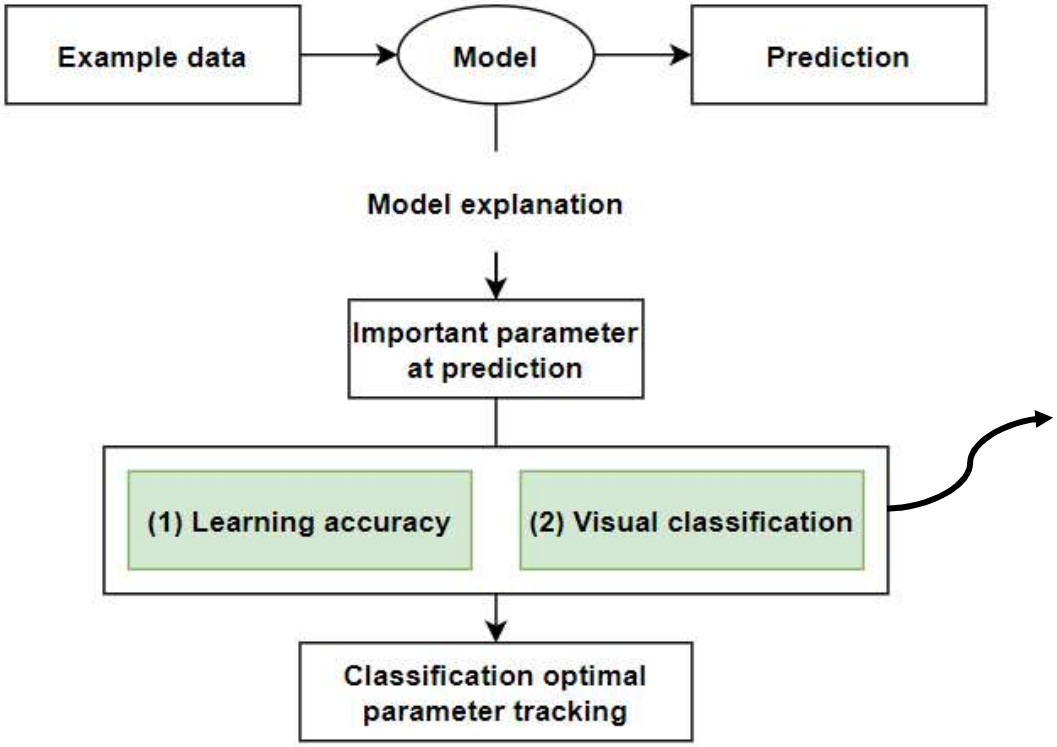




Label	Malfunction	Description
Normal	-	Initial condition #2 MOL 100%
SGTL	IMF mf_RCS01	Steam generator A tube leak
CHRG	IMF mf_CVC01	Charging line break upstream of FT-121
LTDN	IMF mf_CVC05	Letdown line leak inside containment
CDS	IMF mf_CON01	Loss of condenser vacuum
CWS	IMF mf_CWS01	Circulating water tube leak in LP condenser
RCP	IMF mf_CCW01	CCW service loop header leak to aux atm
MSS	IMF mf_MRS01	SG-1 steam line 1A break inside containment
LFH	IMF mf_CON11	Feedwater heater 4A tube break
HFH	IMF mf_MFW10	Feedwater header break

XAI for Abnormal Diagnosis Model

Experiment set-up



(1) Learning accuracy

-Newly train using a **total 10 parameters, <u>one of the most relevant parameters for each label</u>**

-Based on **learning curve and classification accuracy**, determine whether parameters are effective to prediction

(2) Visual classification

- Presentation with **flowchart** about visual classification
- It is judged to be distinguishable when trend difference of the relevant parameter for corresponding label is **<u>more than 20% about others</u>**
  (ex. Increase, decrease, maintain)

XAI for Abnormal Diagnosis Model

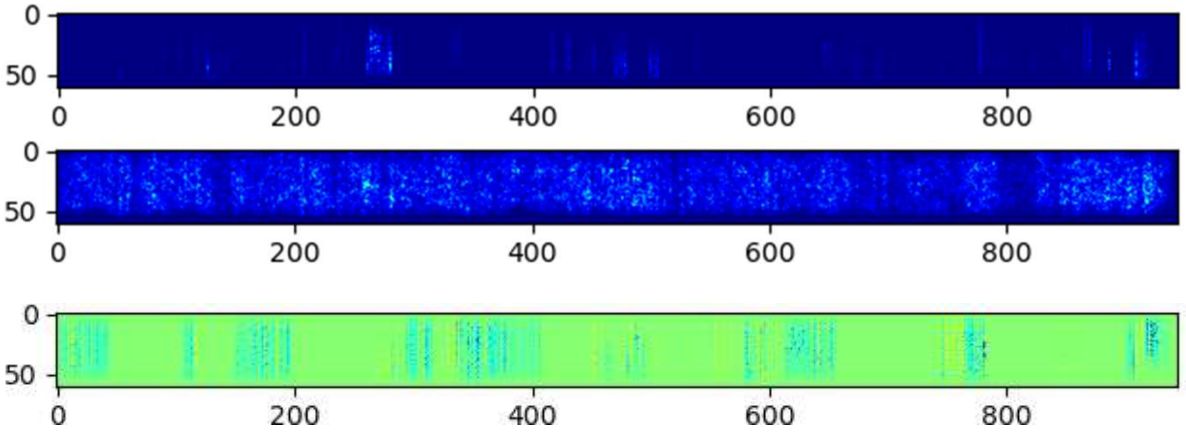
Heatmap of each AI explanation method

The closer to red, the more positive the label affects

Not contributed  High contributed

The parameter corresponding to the column in the red is an important thing for model prediction.

- ✓ **Deep SHAP**
- ✓ **Saliency map**
- ✓ **Guided Grad-CAM**

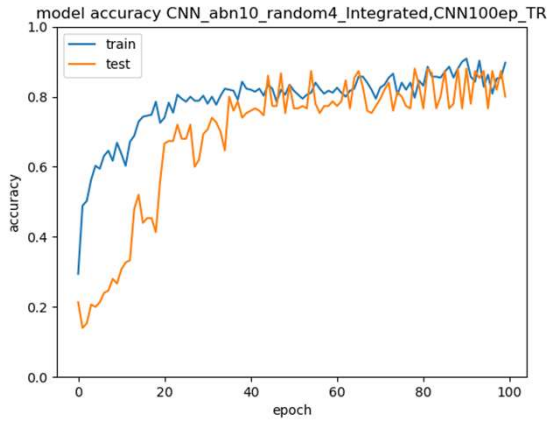


<Heatmap about CDS>

XAI for Abnormal Diagnosis Model

- 10 Randomly selected
  - Learning accuracy and Classification accuracy

	precision	recall	f1-score	support
Normal	0.36	1	0.53	17
CDS[VR]	1	1	1	18
CHRG[LN]	1	1	1	16
CWS[LN]	1	0.14	0.25	14
HFH[LN]	1	1	1	10
LFH[TB]	1	0.88	0.94	17
LTDN[LN]	1	1	1	15
MSS[LN]	1	1	1	11
RCP[LC]	0	0	0	16
SGTL[TL]	1	1	1	16

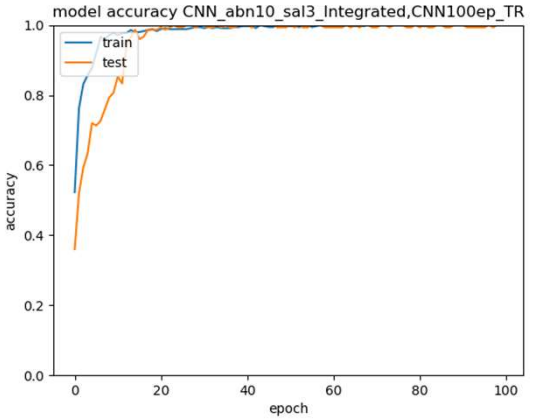


- **Poor learning accuracy**
- Classification **not possible**.

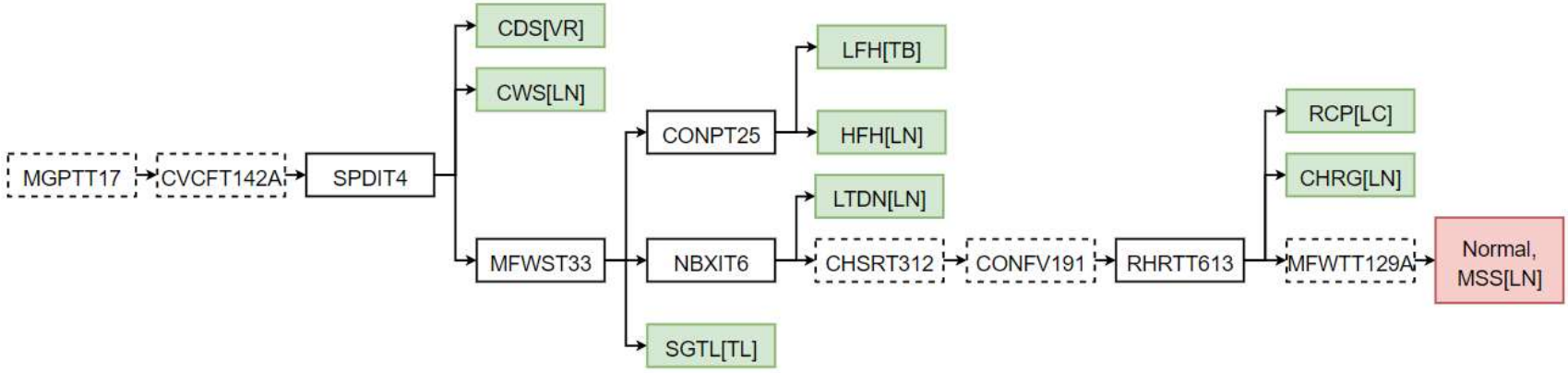
XAI for Abnormal Diagnosis Model

- Saliency map
  - Learning accuracy and Classification accuracy using only top parameters

	precision	recall	f1-score	support
Normal	1	1	1	17
CDS[VR]	1	1	1	18
CHRG[LN]	1	1	1	16
CWS[LN]	1	1	1	14
HFH[LN]	1	1	1	10
LFH[TB]	1	1	1	17
LTDN[LN]	1	1	1	15
MSS[LN]	1	1	1	11
RCP[LC]	1	1	1	16
SGTL[TL]	1	1	1	16



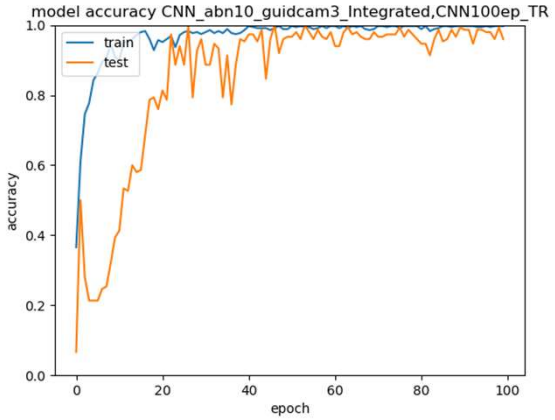
  - Flowchart about visual classification using only top parameters
    - (* Usless Useful Classified Unclassified)



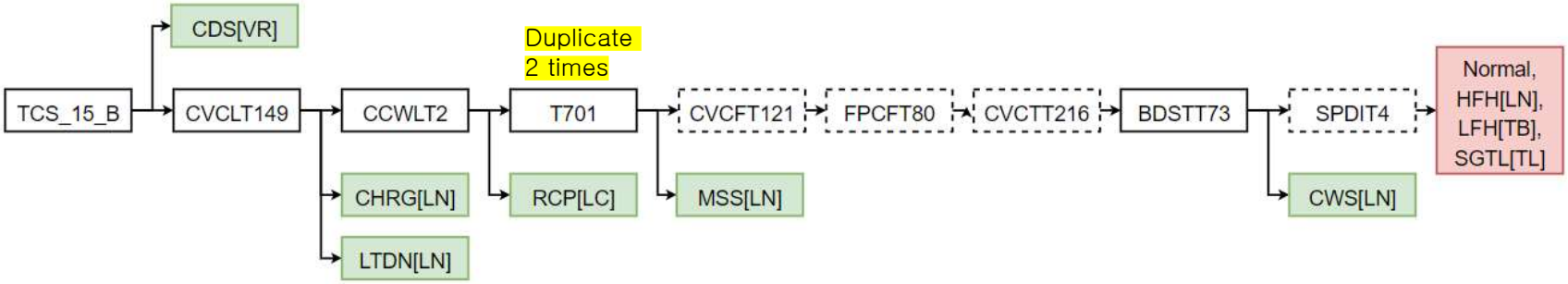
XAI for Abnormal Diagnosis Model

- Guided Grad CAM
  - Learning accuracy and Classification accuracy using only top parameters

	precision	recall	f1-score	support
Normal	0.77	1	0.87	17
CDS[VR]	1	1	1	18
CHRG[LN]	1	1	1	16
CWS[LN]	1	1	1	14
HFH[LN]	0.91	1	0.95	10
LFH[TB]	1	0.65	0.79	17
LTDN[LN]	1	1	1	15
MSS[LN]	1	1	1	11
RCP[LC]	1	1	1	16
SGTL[TL]	1	1	1	16



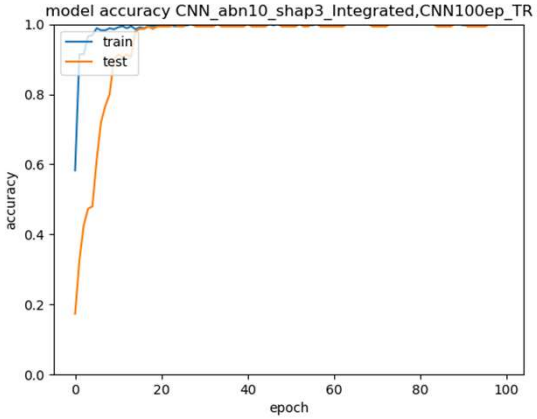
  - Flowchart about visual classification using only top parameters



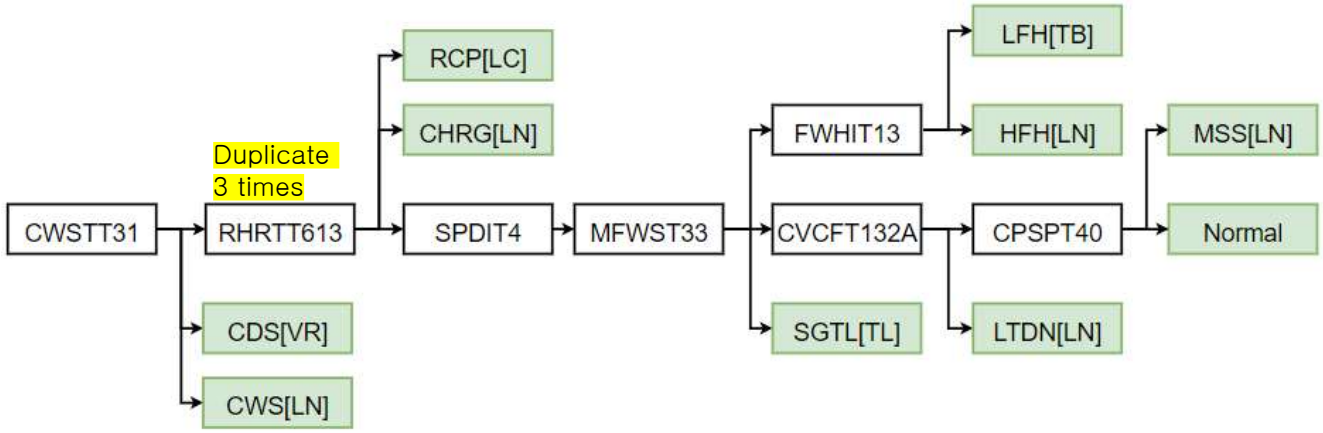
XAI for Abnormal Diagnosis Model

- Deep Explainer SHAP
  - Learning accuracy and Classification accuracy using only top parameters

	precision	recall	f1-score	support
Normal	1	1	1	17
CDS[VR]	1	1	1	18
CHRG[LN]	1	1	1	16
CWS[LN]	1	1	1	14
HFH[LN]	1	1	1	10
LFH[TB]	1	1	1	17
LTDN[LN]	1	1	1	15
MSS[LN]	1	1	1	11
RCP[LC]	1	1	1	16
SGTL[TL]	1	1	1	16



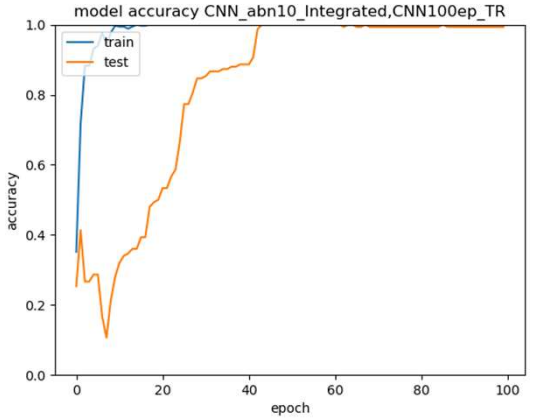
  - Flowchart about visual classification using only top parameters



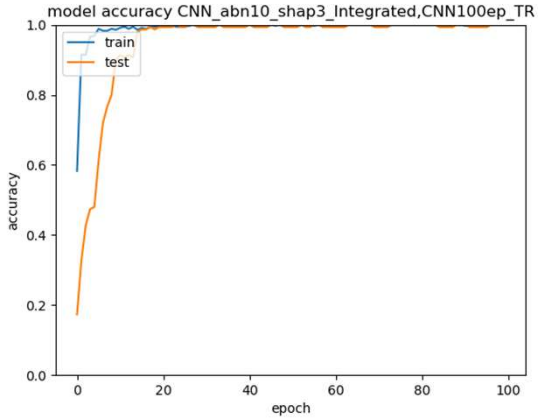
XAI for Abnormal Diagnosis Model

- Learning accuracy

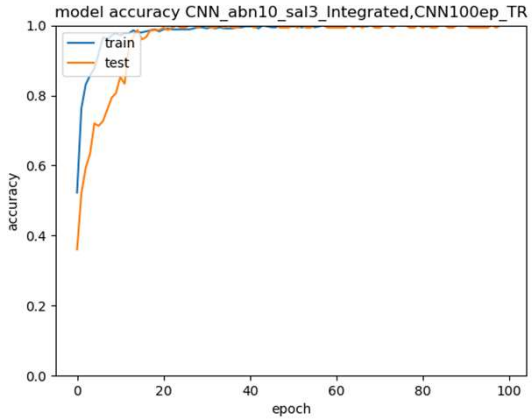
All Parameter



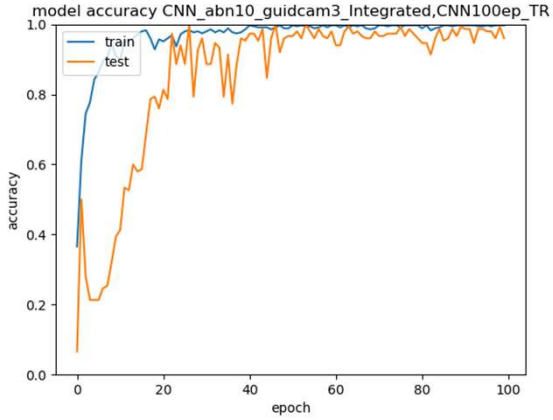
Deep Explainer SHAP



Saliency map



Guided-Grad-CAM



XAI for Abnormal Diagnosis Model

- Comparison with each method

Exclude duplicates

	Random	Saliency	Guided CAM	Deep SHAP
Param num.	10	10	9	7
Acc	0.8971	1	0.9971	1
Loss	0.3187	0.0038	0.0273	0.0014
Val_acc	0.8	1	0.96	1
Val_loss	0.3773	0.0053	0.1257	0.0018
Visual classification group num.		9	7	10

✓ **Saliency map** and **Deep explainer SHAP** method is better to explain AI model and to extract relevant parameters.

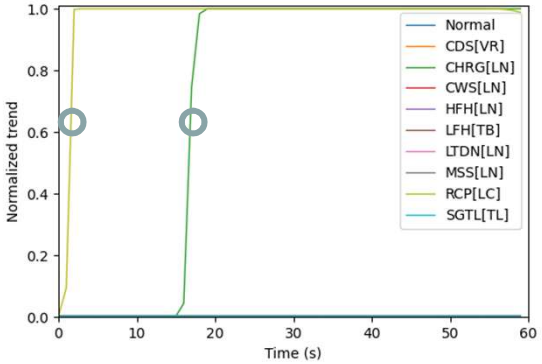
- Comprehensive high contributed parameter
  - **<u>6 highest relevant parameters that duplicate in each label</u>**

Parameter	Description
hmi_RHRTT613_VALUE	[RHR] RHR PMP B DISCH TEMP
hmi_SPDIT4_VALUE	[SPD] 13.8 KV TO SITE
hmi_MFWST33_VALUE	[MFW] MAIN FEEDWATER PUMP TURBINE A SPEED
hmi_FWHIT13_VALUE	[FWH] HTR DRN PMP B CURRENT
hmi_NBXIT6_VALUE	[NBX] LC FDR BKR NBX209 CURRENT
hmi_CPSPT40_VALUE	[CPS] CTMT-AUX BLD DIFF PRESS

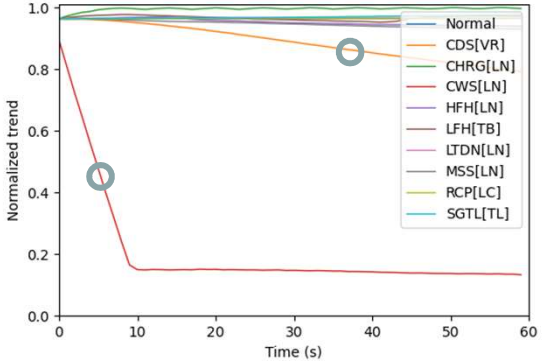
XAI result of Abnormal Diagnosis model

- Most important 6 parameter trends for each state

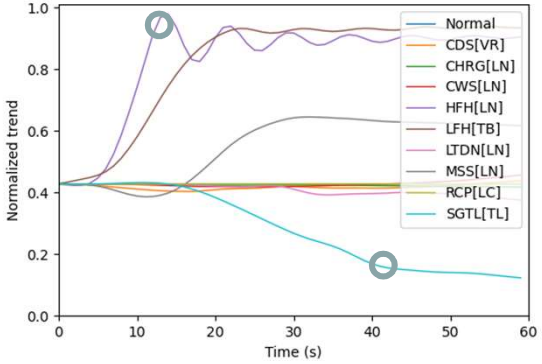
✓ CHRG[LN] / RCP[LC]



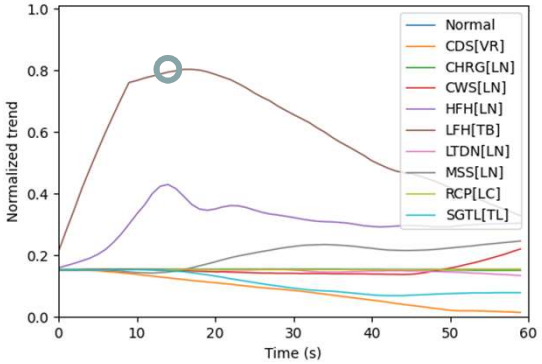
✓ CWS[LN]
(additional distinguishable CDS[VR])



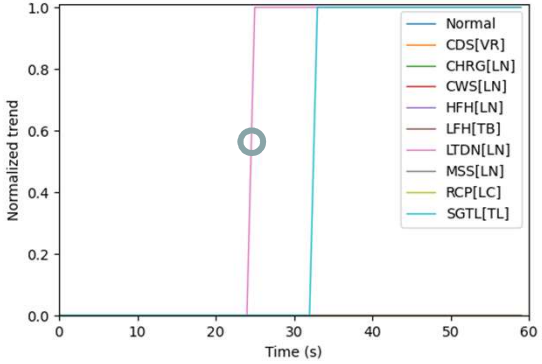
✓ HFH[LN] / SGTL[TL]



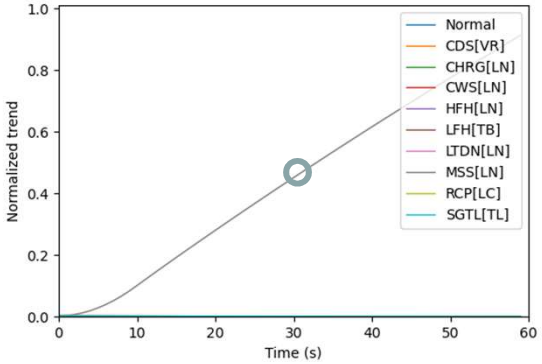
✓ LFL[TB]



✓ LTDN[LN]



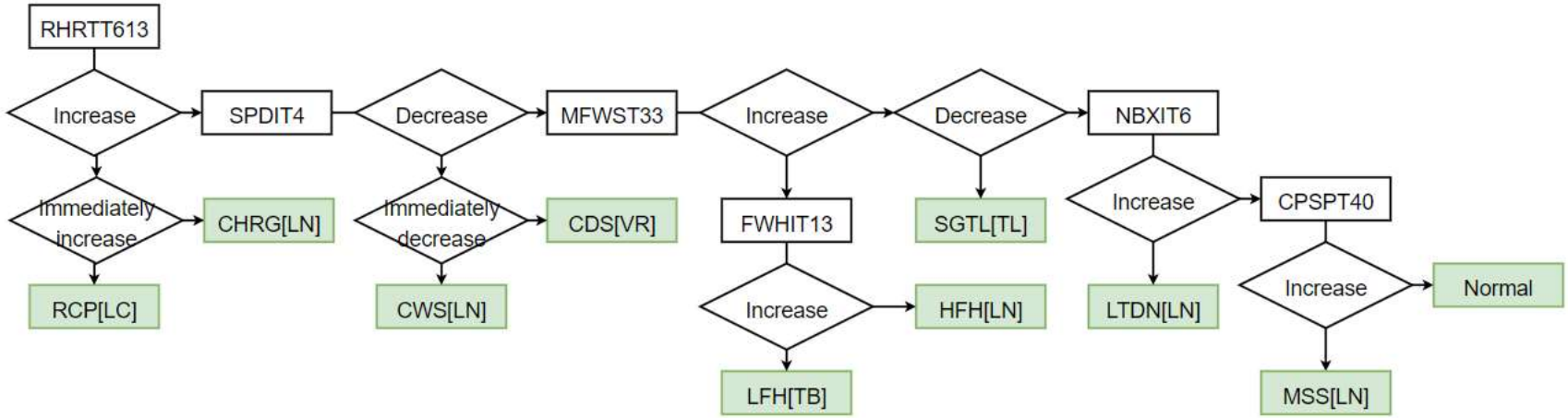
✓ MSS[LN]



XAI result of Abnormal Diagnosis model

- Flowchart for abnormal diagnosis

(↓: Y / →: N)



  - Classification of 10 states with only 6 parameters (994 -> 6)

Conclusions

- AI V&V has completely different aspects from SW V&V
  - Better accurate but less explainable
  - Hard to be verified and validated.
- Characteristics of applications
  - Usually applied to the problems which are hard to be modeled with logics and equations
  - How to validate Alpha go?

- Validation based on testing is the easiest way
  - Testing coverage is the most important.
  - Perfect validation is not possible.
- XAI could be one option, but different approach is required according to the application
  - Validation for abnormal diagnosis and autonomous operation