

**SURESOFT**  
Intelligence for Safe World

# VERIFAI-M

AI 모델의 신뢰성 검증 도구

# 여러분의 인공지능 모델은 안전합니까?



Trustworthy AI를 위해서는



다양한 환경에서  
모델의 안정성을  
검증해야 합니다



인공지능 모델의  
취약 구간을  
파악해야 합니다



인공지능 모델의  
동작과 실패 결과를  
확인해야 합니다.



인공지능 모델의  
성능과 변화를  
관리해야 합니다

## Trustworthy AI

- Trustworthy AI에 대한 국내/국외 요구 확대
- 안전하고 신뢰할 수 있는 인공지능을 위한 글로벌 규제 논의 증가

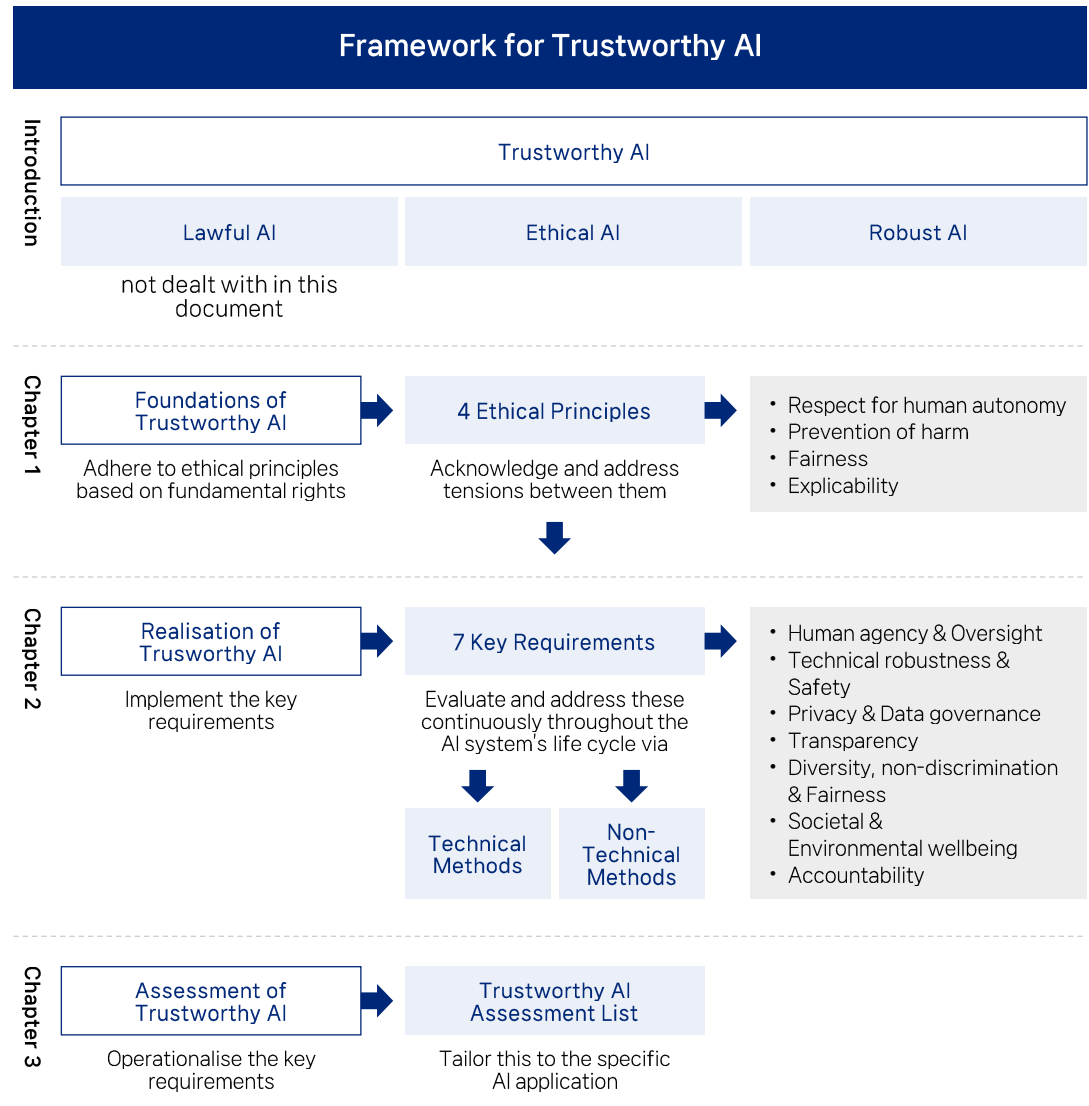


Figure 1 : The Guidelines as a framework for Trustworthy AI

## Trustworthy AI의 구현

- Trustworthy AI를 구현하기 위한 ISO/IEC 표준 제시

| 요구 사항                 | 세부 요구 사항                                                                                                                                                                                         | 구현 방법                                                                                                                                                                                          | 레퍼런스 위치                                                                                                                                                                                                                                                                                                                                                                                                                  |
|-----------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 인간의 행동<br>주체성과 감독     | <ul style="list-style-type: none"> <li>- Fundamental rights</li> <li>- Human agency</li> <li>- Human oversight</li> </ul>                                                                        | <ul style="list-style-type: none"> <li>- 의사결정 과정에서 사용자 개입 제어 설계</li> <li>- AI 사용 여부 명확화</li> <li>- 위험 수준에 따라 사용자 개입 기능 제공</li> </ul>                                                           | <ul style="list-style-type: none"> <li>- [EU AI Act] Art14, Art.27 , Art 50</li> <li>- [ISO/IEC 25059] 5.2, 5.7</li> </ul>                                                                                                                                                                                                                                                                                               |
| 기술적<br>강건성과 안전성       | <ul style="list-style-type: none"> <li>- Resilience to attack and security</li> <li>- Fallback plan and general safety</li> <li>- Accuracy</li> <li>- Reliability and Reproducibility</li> </ul> | <ul style="list-style-type: none"> <li>- 적대적 공격 테스트</li> <li>- 메타모픽 테스트</li> <li>- 강건성 테스트</li> <li>- 표준 지표 기반 성능 평가</li> <li>- 테스트 반복을 통해 결과 신뢰성 확보</li> <li>- 변화에 따른 성능을 지속적으로 추적</li> </ul> | <ul style="list-style-type: none"> <li>- [EU AI Act] Art.15(3), Art.15(4), Art.15(5)</li> <li>- [ISO/IEC TR 24028] 9.7, 9.10</li> <li>- [ISO/IEC TR 24029-1] Annex A</li> <li>- [ISO/IEC 24029-2] 5, 6</li> <li>- [ISO/IEC TS 4213] 5.3, 6, 7</li> <li>- [ISO/IEC 25059] 5.4, 5.5</li> <li>- [ISO/IEC TR 29119-11] 6.1.11, 6.2, 8.4, 9.2, A.5.3</li> <li>- [ISO/IEC TS 42119-2] p.5</li> <li>- [ISO 21448] 11</li> </ul> |
| 개인정보 보호 및<br>데이터 거버넌스 | <ul style="list-style-type: none"> <li>- Privacy and data protection</li> <li>- Quality and integrity of data</li> <li>- Access to data</li> </ul>                                               | <ul style="list-style-type: none"> <li>- 개인정보 관리 체계 구축</li> <li>- 데이터 품질 평가 및 관리</li> <li>- 데이터 독립성 및 누수 검증</li> <li>- 데이터 접근 권한 관리</li> </ul>                                                 | <ul style="list-style-type: none"> <li>- [EU AI Act] Art.10</li> <li>- [GDPR]</li> <li>- [ISO/IEC TR 29119-11] 6.1.3, 7.5, 7.6</li> <li>- [ISO/IEC 5259] Part 2 6, Part 3 6, Part 4 8</li> </ul>                                                                                                                                                                                                                         |
| 투명성                   | <ul style="list-style-type: none"> <li>- Traceability</li> <li>- Explainability</li> <li>- Communication</li> </ul>                                                                              | <ul style="list-style-type: none"> <li>- 모델 및 데이터 이력 추적</li> <li>- 설명 가능성 분석 적용</li> <li>- 설명 품질 평가</li> </ul>                                                                                 | <ul style="list-style-type: none"> <li>- [EU AI Act] Art.12, Art.13, Art.18, Art.50</li> <li>- [ISO/IEC TR 24028] 9.3, 9.3.3, 9.3.7</li> <li>- [ISO/IEC 25059] 5.6, 6.3</li> <li>- [ISO/IEC TR 29119-11] 7.4, A.7</li> </ul>                                                                                                                                                                                             |
| 다양성, 차별 금지<br>및 공정성   | <ul style="list-style-type: none"> <li>- Avoidance of unfair bias</li> <li>- Accessibility and universal design</li> <li>- Stakeholder Participation</li> </ul>                                  | <ul style="list-style-type: none"> <li>- 편향 탐지 및 공정성 평가</li> <li>- 데이터 대표성 검증</li> <li>- 편향 재학습 트리거 설정</li> </ul>                                                                              | <ul style="list-style-type: none"> <li>- [EU AI Act] Art.10(2)(f)</li> <li>- [ISO/IEC TR 24028] 9.5</li> <li>- [ISO/IEC TS 4213] 5.3.2</li> <li>- [ISO/IEC 25059] 6.2, Annex B</li> <li>- [ISO/IEC TR 29119-11] 5.1.5, 6.1.8</li> </ul>                                                                                                                                                                                  |
| 사회 및 환경 복지            | <ul style="list-style-type: none"> <li>- Sustainable and environmentally friendly AI</li> <li>- Social impact</li> <li>- Society and Democracy</li> </ul>                                        | <ul style="list-style-type: none"> <li>- 모델 경량화 및 최적화</li> <li>- 에너지 효율 및 지속 가능성 평가</li> <li>- 윤리 기준 기반 영향 평가</li> </ul>                                                                       | <ul style="list-style-type: none"> <li>- [EU AI Act] Annex III</li> <li>- [ISO/IEC 25059] 6.2</li> <li>- [ISO/IEC TR 29119-11] 5.3, 5.4</li> <li>- [ISO PAS 8800] Annex G.7.2~3</li> </ul>                                                                                                                                                                                                                               |
| 책임                    | <ul style="list-style-type: none"> <li>- Auditability</li> <li>- Minimisation and reporting of negative impacts</li> <li>- Trade-offs</li> <li>- Redress</li> </ul>                              | <ul style="list-style-type: none"> <li>- 검증 결과 이력 관리</li> <li>- 위험 관리 체계 구축</li> <li>- 규제 대응 문서화 및 적합성 평가</li> </ul>                                                                           | <ul style="list-style-type: none"> <li>- [EU AI Act] Art.9, Art.11, Art.18, Art.43, Art.47, Art.48, Art.72</li> <li>- [ISO/IEC 25059] Annex B</li> <li>- [ISO/IEC TR 29119-11] 10.3</li> <li>- [ISO/IEC TS 42119-2] p.5</li> <li>- [ISO PAS 8800] 8</li> </ul>                                                                                                                                                           |

## VERIFAI-M : AI 모델 신뢰성 검증 플랫폼

### Vision

우리는 VERIFAI-M을 통해 **인공지능 모델을 신뢰할 수 있도록 시험하고 평가하는 것을 지원**하고자 합니다.  
VERIFAI-M은 AI 소프트웨어를 위해 선도적이고 중요한 책임을 수행하여 안전하고 신뢰성 있는 AI 발전에 기여합니다.

#### Features 01

모델 및 접속 환경  
지원



- 다양한 Task 지원
- 로컬 및 원격 접근 환경 지원

#### Features 02

테스트 데이터  
증강



- 모델 종속적 데이터 증강
- 모델 독립적 데이터 증강

#### Features 03

모델 메트릭  
측정



- Task 별 메트릭 측정
- 테스트 케이스 별 메트릭 측정

#### Features 04

이력  
관리



- 검증 결과 관리
- 모델 및 데이터 추적 관리

**VERIFAI-M Is All You Need**  
Beyond AI Performance, We verify trustworthiness

## I 모델 및 접속 환경 지원

- 현재 7종의 AI 모델 Task 를 지원하며, 다양한 도메인에 대응할 수 있도록 지속적으로 확장하고 있습니다.
- 로컬 및 원격 환경에서 모델에 유연하게 접근하여, 다양한 실행 환경에서 검증 수행이 가능합니다.

### 지원 모델 Task



Image  
Classification

이미지 내 객체를  
분류하는 모델



Object  
Detection 2D

이미지 내 객체의  
위치와 클래스를  
탐지하는 모델



Segmentation 2D

이미지 내 픽셀 단위  
영역을 분할하는 모델



Tabular  
Classification

구조화된  
센서 데이터를  
분류하는 모델



Tabular  
Regression

구조화된  
센서 데이터를  
회귀 예측 하는 모델



Text  
Generation

자연어 텍스트를  
생성하는 모델



E2E Driving

자율주행을 위한  
End-to-End  
예측 모델

### 접속 환경



#### 로컬 환경

사용자 로컬 환경에서  
모델에 직접 접속 하여 검증 수행

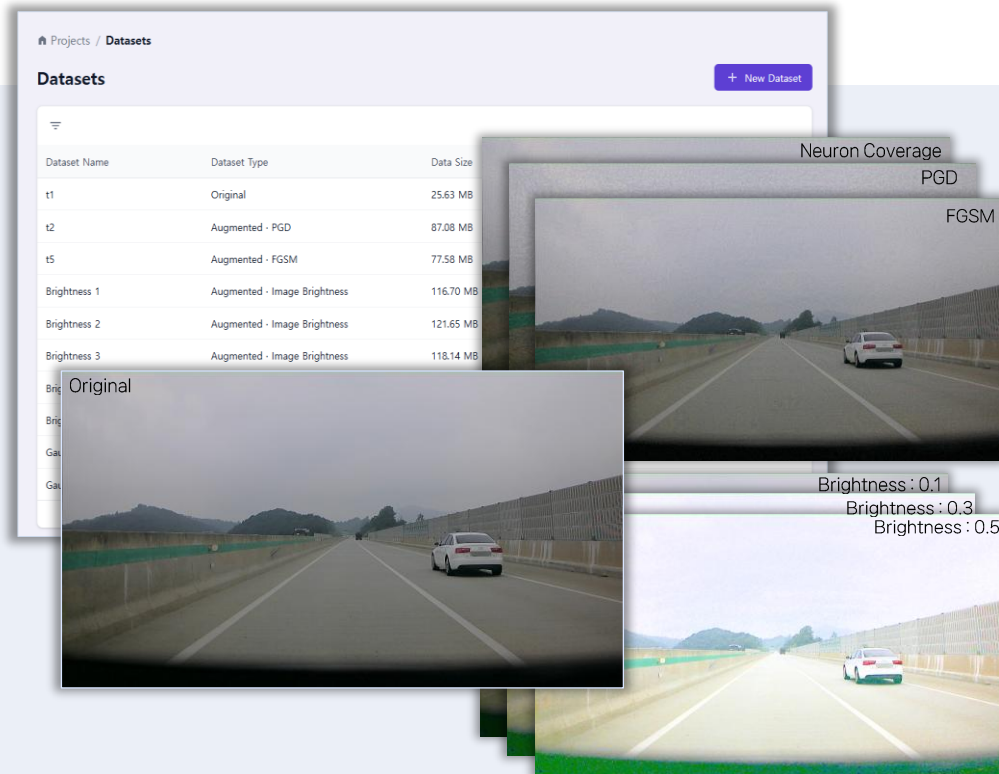


#### 원격 환경

원격 서버 환경에서  
모델에 유연하게 접속하여 검증 수행

## I 테스트 데이터 증강

- 사용자 보유 데이터셋을 체계적으로 관리할 수 있습니다.
- 사용자 데이터를 활용하여 다양한 변형 기법을 적용한 테스트 데이터를 증강할 수 있습니다.
- 다양한 조건에서 모델이 실패하는 지점을 사전에 탐지할 수 있습니다.



### 지원 하는 증강 방법

#### ▪ 모델 종속적 데이터 증강

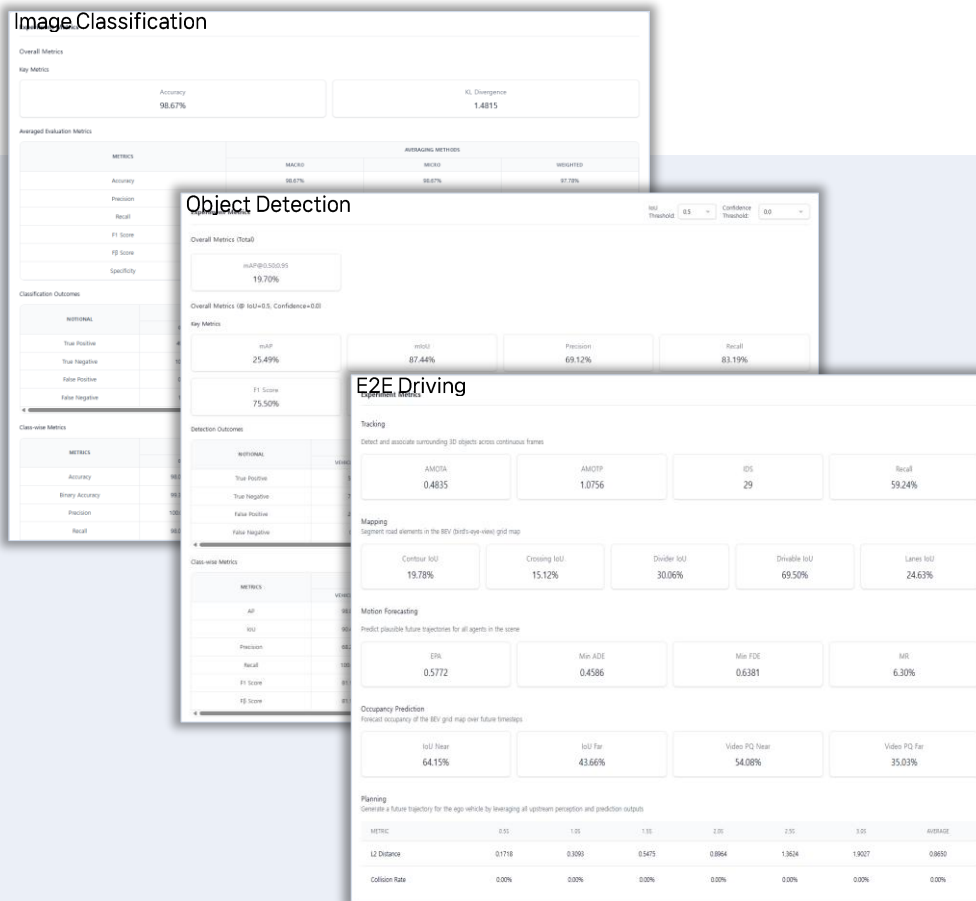
- PGD
- FGSM
- Neuron Coverage

#### ▪ 모델 독립적 데이터 증강

- Image Brightness
- Image Contrast
- Image Saturation
- Image Pixelate
- Image Motion Blur
- Image Gaussian Blur
- Image Gaussian Noise
- Image JPEG Compression

## | 모델 성능 지표 측정

- 결과에 대한 분석을 통해 모델을 평가할 수 있습니다.
- 모델 Task 별 특화된 지표와 표준 기반 평가를 통해 모델 성능을 다각도로 분석 할 수 있습니다.



### 모델 Task 별 지원 Metric

- 분류 모델
  - ISO/IEC 4213 기반 총 15종
- 객체 탐지 모델
  - mAP, mIoU 등 총 6종
- 회귀 모델
  - MAE, MSE 등 총 6종
- 생성 모델
  - BLEU Score, ROUGE Score 등 총 8종
- 자율주행 Planning 모델
  - AMOTA, EPA 등 총 17종



## 모델 성능 지표 측정

- 국제 표준 · 국가 인증 · 학계 벤치마크의 지표 체계를 통합 하였습니다.
- 최신 동향을 지속 반영하는 신뢰 기반 AI 모델 평가 솔루션을 제공합니다.



### 국제표준

- ✓ ISO/IEC 4213 (Classification) 등 국제표준 적용
- ✓ Model Task 별 최신 동향 지속 모니터링
- ✓ 신규 표준 재 개정 시 즉각 반영



### AI-MASTER

- ✓ Trustworthy AI 핵심 요구사항 적용 및 국내 관련 법령 기반 국제표준 시험 방법을 체계화한 문서
- ✓ 체계화된 맞춤형 AI 신뢰성 인증 기준을 평가 항목으로 적용



### Model Task 별 벤치마크

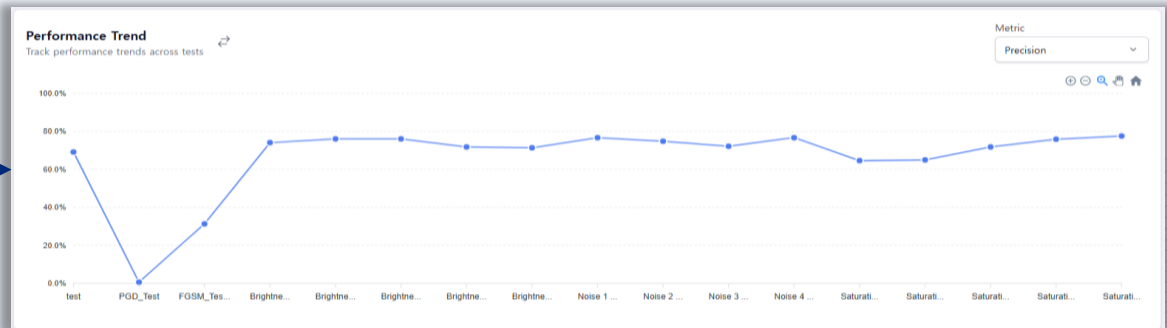
- ✓ 학계 및 산업계가 Task별로 합의·채택한 Task-specific Evaluation Metrics 적용
- ✓ Task 별 최신 벤치마크 Metric 지속 수집

국제표준, 국가 인증, 학계 벤치마크 기반의 측정 지표 제공

## I 검증 결과 이력 관리

- 모델 및 데이터 변경 이력을 기반으로 검증 결과를 지속적으로 추적합니다.
- 변화에 따른 성능 차이를 분석하여 검증 결과를 체계적으로 관리합니다.

### ① 성능 추세 분석



### ② 강건성 비교

**Robustness Summary**  
Compare metric robustness across augmented tests

| METRIC NAME                      | Original | Augmentation |          |           |              |              |              |              |              |           |           |           |           |           |  |   |
|----------------------------------|----------|--------------|----------|-----------|--------------|--------------|--------------|--------------|--------------|-----------|-----------|-----------|-----------|-----------|--|---|
|                                  |          | test         | PGD_Test | FGSM_Test | Noise 1 Test | Noise 2 Test | Noise 3 Test | Noise 4 Test | Noise 5 Test | Sat1 Test | Sat2 Test | Sat3 Test | Sat4 Test | Sat5 Test |  |   |
| mAP@0.5:0.95                     | ↑        | 0.19704      | 0.00045  | 0.08418   | 0.19373      | 0.17251      | 0.15573      | 0.14005      | 0.12683      | 0.19164   | 0.18550   | 0.19883   | 0.19005   | 0.15805   |  |   |
| IOU_MAPMacro                     | ↑        | 0.25482      | 0.00103  | 0.12524   | 0.25501      | 0.21749      | 0.19676      | 0.17824      | 0.16012      | 0.24083   | 0.23321   | 0.25592   | 0.24278   | 0.20614   |  |   |
| IOU_CONF_MeanIoUPerClassMicro    | ↑        | 0.87436      | 0.61202  | 0.83499   | 0.86999      | 0.88260      | 0.88083      | 0.87820      | 0.88610      | 0.87979   | 0.87832   | 0.87306   | 0.87781   | 0.87927   |  |   |
| IOU_CONF_PrecisionPerClassMicro  | ↑        | 0.89118      | 0.00637  | 0.31313   | 0.76667      | 0.74766      | 0.72115      | 0.76667      | 0.78481      | 0.64539   | 0.64964   | 0.71756   | 0.75882   | 0.77551   |  | + |
| IOU_CONF_RecallPerClassMicro     | ↑        | 0.83186      | 0.11504  | 0.54867   | 0.81416      | 0.70796      | 0.66372      | 0.61062      | 0.54867      | 0.80531   | 0.78761   | 0.83186   | 0.77676   | 0.67257   |  |   |
| IOU_CONF_F1ScorePerClassMicro    | ↑        | 0.75502      | 0.01207  | 0.39671   | 0.78970      | 0.72727      | 0.69124      | 0.67980      | 0.64583      | 0.71654   | 0.71200   | 0.77049   | 0.76856   | 0.72038   |  |   |
| IOU_CONF_FbetaScorePerClassMicro | ↑        | 0.75502      | 0.01207  | 0.39671   | 0.78970      | 0.72727      | 0.69124      | 0.67980      | 0.64583      | 0.71654   | 0.71200   | 0.77049   | 0.76856   | 0.72038   |  |   |
| Pass Rate                        | ↑        | 0.36000      | 0.00000  | 0.38000   | 0.40000      | 0.38000      | 0.34000      | 0.32000      | 0.34000      | 0.26000   | 0.26000   | 0.38000   | 0.38000   | 0.36000   |  |   |

**Projects / Tests**

**Tests**

Performance Trend

검증 결과 추적  
검증 이력을 지속 추적

| Test Name         | Model Name | Dataset Name     | Dataset Type                | Created At  | Status |
|-------------------|------------|------------------|-----------------------------|-------------|--------|
| test              | detection  | t1               | Original                    | 2026-04-... | Done   |
| PGD_Test          | detection  | t2               | Augmented - PGD             | 2026-04-... | Done   |
| FGSM_Test         | detection  | t5               | Augmented - FGSM            | 2026-04-... | Done   |
| Brightness 1 Test | detection  | Brightness 1     | Augmented - Image Bright... | 2026-04-... | Done   |
| Brightness 2 Test | detection  | Brightness 2     | Augmented - Image Bright... | 2026-04-... | Done   |
| Brightness 3 Test | detection  | Brightness 3     | Augmented - Image Bright... | 2026-04-... | Done   |
| Brightness 4 Test | detection  | Brightness 4     | Augmented - Image Bright... | 2026-04-... | Done   |
| Brightness 5 Test | detection  | Brightness 5     | Augmented - Image Bright... | 2026-04-... | Done   |
| Noise 1 Test      | detection  | Gaussian_noise 1 | Augmented - Image Gauss...  | 2026-04-... | Done   |

## I 강건성 시험

- 강건성 시험 : 노이즈, 변형이 가해진 입력 데이터에 대해, 모델의 예측 결과가 원본과 일관 되게 유지 되는지 검증 하는 시험



### 관련 연구

- Hendrycks & Dietterich, Benchmarking Neural Network Robustness to Common Corruptions and Perturbations
- Taori et al., Measuring Robustness to Natural Distribution Shifts in Image Classification
- N Carlini et al., Audio adversarial examples: Targeted attacks on speech-to-text
- GR Mode et al., Adversarial examples in deep learning for multivariate time series regression

## I 강건성 시험

- 원본 데이터에서 출발하여 변형·실행·분석으로 이어지는 강건성 시험 절차

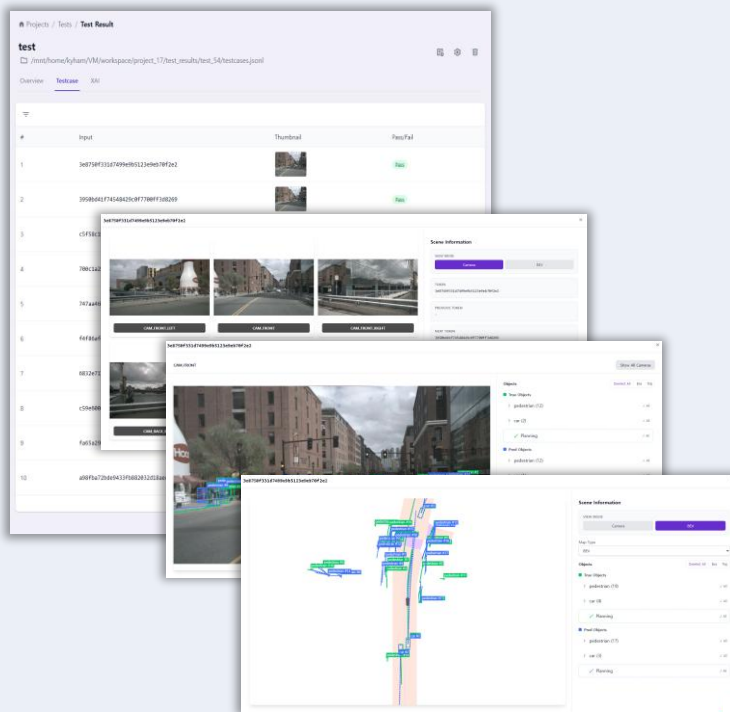


데이터 증강부터 결과 관리까지, 전체 검증 과정을 하나의 플랫폼에서 통합 수행

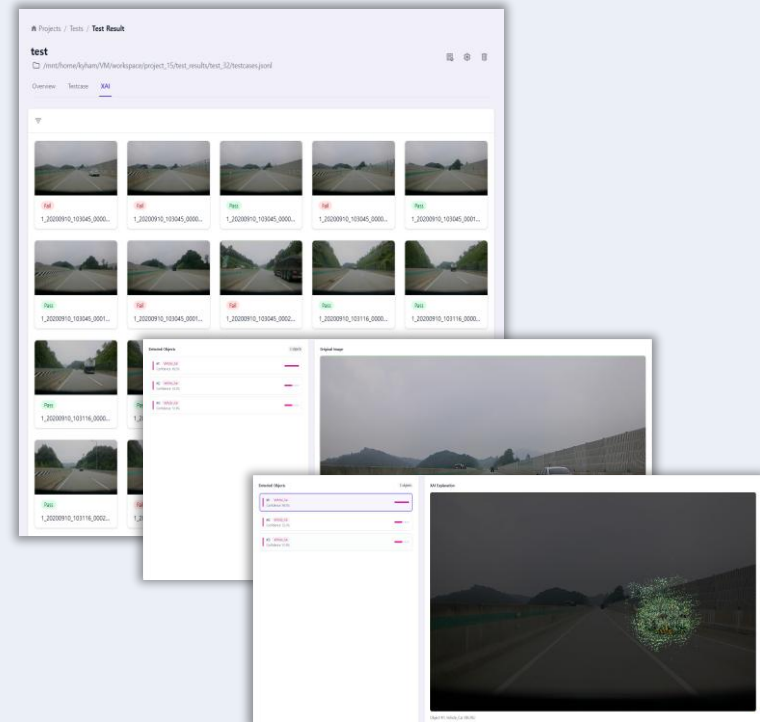
## I 설명 가능성 시험

- 테스트 케이스별 모델의 추론 결과와 Pass/Fail 유무를 확인 할 수 있습니다.
- 테스트 케이스별 모델의 주요 관심사를 확인 할 수 있습니다.,

### 테스트 케이스 별 모델의 추론 결과



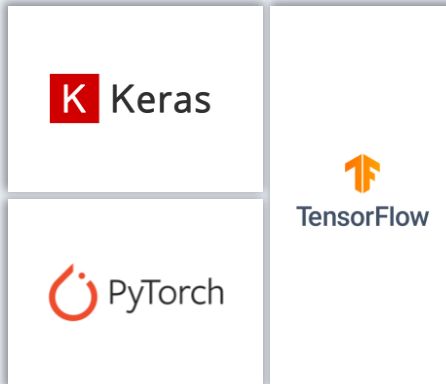
### 테스트 케이스 별 모델의 주요 관심사



## ■ VERIFAI-M 검증을 위한 입력

- 검증 대상 모델, 데이터 처리 스크립트 필요

### 모델



성능을 측정하고자 하는 대상 AI 모델

Pytorch, Keras, Tensorflow  
포맷의 모델 파일

### 데이터



모델 입력 데이터

모델의 성능을 정확하게 측정하기 위해  
다양한 상황의 데이터 필요

### 데이터 처리 스크립트



입력 받은 데이터를 모델에  
입력되는 형태로 구조화하기 위한 Python 코드

그 외 모델의 구동에 필요한 모델  
Custom 코드

## | 상세 환경

| 구분    |          | 세부 사항                                                                                                                                                                                                                                                                                  |
|-------|----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 지원 모델 |          | <ul style="list-style-type: none"> <li>Image Classification, Object Detection 2D, Segmentation 2D, Tabular Regression, Tabular Classification, Text Generation, E2E Driving</li> </ul>                                                                                                 |
| 개발 환경 | 운영체제     | <ul style="list-style-type: none"> <li>Windows 10 이상 (64bit 지원)</li> <li>Linux (Debian, RPM 계열)</li> </ul>                                                                                                                                                                             |
|       | 지원 프레임워크 |  PyTorch    Keras    TensorFlow |

## | 하드웨어 최소 장비 사양

| CPU             | GPU          | RAM      | 저장 공간    |
|-----------------|--------------|----------|----------|
| 8 core 3 GHz 이상 | 30 TFLOPS 이상 | 32 GB 이상 | 40 GB 이상 |





**Thank You**