Evidence-based Participatory Decision Making for Cancer Prevention through implementation research

Building AI Trust through Bias Identification

ONCODIR

Dr. Panos Tsakanikas, ICCS
ptsakanikas@cn.ntua.gr

ONCODIR Overview...

ONCODIR

Project Objectives

IDENTIFY main correlations, barriers and significant factors of CRC

ENSURE equal and affordable access to cancer prevention strategies for everyone between and within EU countries

PROVIDE innovative AI-powered personalised prevention approaches

ENHANCE the ongoing evidence-based CRC prevention programmes for precise CRC primary prevention

ESTABLISH risk-based stratification for citizens considering structural and behavioural intervention through participatory approach

DESIGN intelligent monitoring tools for policy makers through a participatory co-designing approach

ONCODIR is developing a platform based on artificial intelligence and privacy principles. It will provide recommendation services based on input from citizens, clinicians and policy-makers. We will consider factors such as lifestyle, nutrition and economics.

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

Bias in AI & APPO Bias Detection Framework

Types of Bias

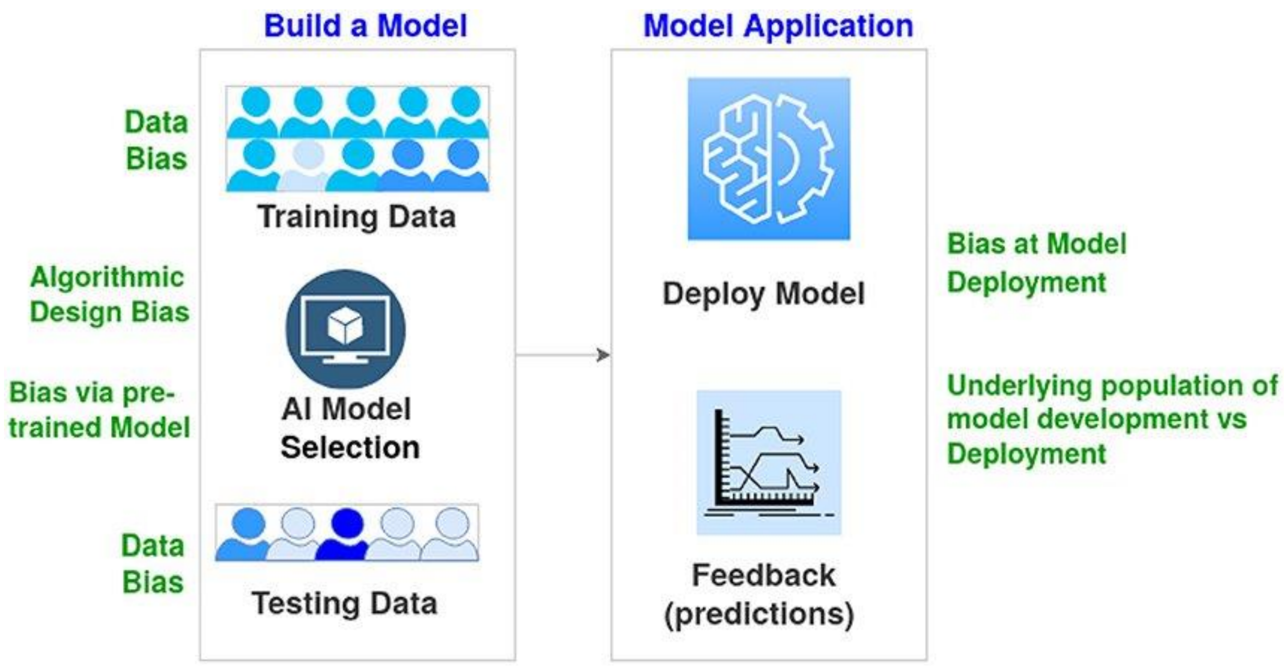
Systematic unfairness in AI models due to data, features, or algorithms.

Types of Bias in AI

- Data Bias: When training data is not representative of real-world scenarios.
- Algorithmic Bias: When ML models systematically favor one group over another.
- Label Bias: Labels used for training contain human prejudices.
- Hidden Bias: High-dimensional patterns in data create **unintended unfairness**.

Applications in ONCODIR

- **Predictive Diagnostics**: AI misclassifying conditions due to biased training data.
- **Recommendations**: Bias in recommendations.
- **Public Health Data**: Underprediction of disease risks in specific regions.



Bias in AI & APPO Bias Detection Framework

Challenges in Bias Detection

- Traditional metrics (KL, JS, TVD, KS) detect **bias in individual features**.
- High-dimensional data can hide **complex biases**.
- Advanced techniques like **anomaly detection** & **SHAP** are needed.

For Developers/Tech

General Aspects of Bias

- **Sources**: Data collection, sampling, and annotation.
- **Impacts**: Ethical concerns, legal implications, fairness issues, **enable human-in-the-loop**.
- **Mitigation**: Data balancing, fairness-aware algorithms, post-hoc adjustments.

For health domain

Insights in real life (related to Bias)

- **Better Patient Outcomes**: Biased AI can misdiagnose or recommend incorrect treatments. APPO helps avoid this.
- **Fairness for All**: Ensures that AI doesn't favor one group over another based on age, gender, region, or other factors.
- **Trustworthy AI**: Doctors and patients need confidence in AI decisions. APPO makes those decisions more transparent and fa
- **Regulatory Compliance**: Helps meet ethical, legal, and fairness standards in healthcare AI.

Bias in AI & APPO Bias Detection Framework

APPO Bias Detection Framework

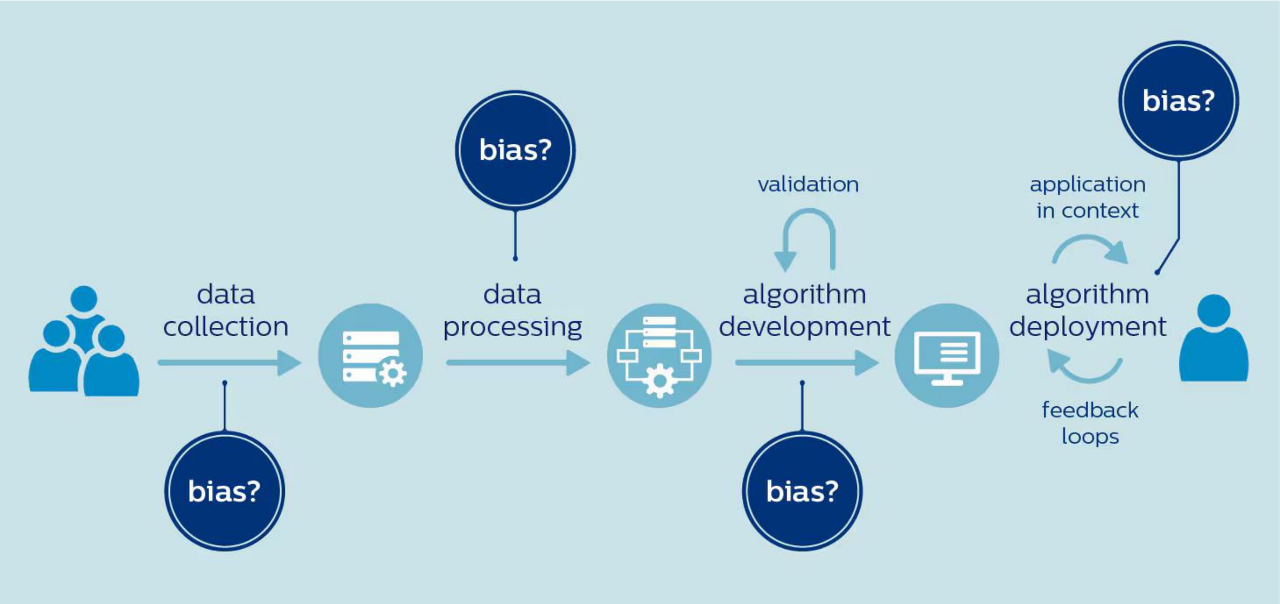
APPO Bias Detection Framework

Multi-phase detection approach using:

- Data quality assessment
- Traditional bias metrics (KL, JS, TVD, KS)
- High-dimensional bias detection (PCA, t-SNE, UMAP **+ clustering**)
- Community Detection: Isolation Forest, LOF, Graph-based clustering.
- SHAP-based explanations
- Autoencoder-based hidden bias detection.

Key Features & Innovations

- Combining standard & high-dimensional bias detection.
- Dynamic anomaly scoring.
- Automated thresholding & severity alerts.
- JSON reporting for bias auditing.



Bias in AI & APPO Bias Detection Framework

APPO workflow

Workflow diagram for the **Multi-Phase Bias Detection Framework**

Data Preprocessing & Profiling

- Cleans data, encodes categorical features, normalizes numerical data, and generates a profiling report.

Traditional Bias Metrics

- KL Divergence, JS Divergence, Lp-Norm, TVD, and KS statistics to detect bias in feature distributions.

High-Dimensional Bias Detection

- PCA, UMAP, or t-SNE for dimensionality reduction, and anomaly detection with LOF and Isolation Forest.

SHAP-Based Bias Detection

- Autoencoder and XGBoost to compute SHAP values, revealing features contributing to bias

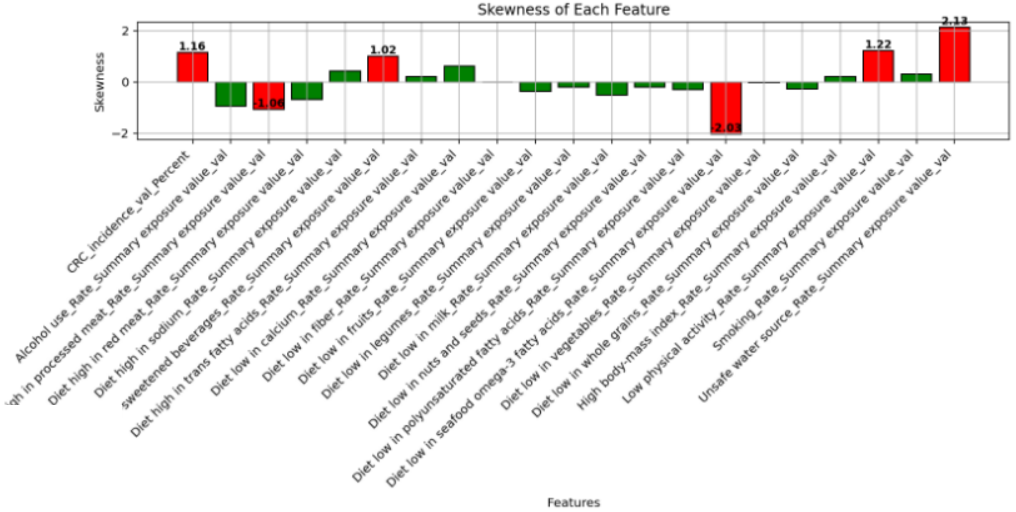
Community Bias Detection

- Applies clustering methods (K-Means, DBSCAN) to detect biased structures in community networks.

Bias Alerts & Report Generation

- Aggregates all insights, generates severity levels, and produces a final JSON bias report.

PDF report - screenshot



s bar plot. This plot shows the skewness values for each feature, highlighting the statistical distribution

2. Bias Alerts and Key Findings

This section highlights specific bias risks identified in the dataset, categorized by severity levels.

- Severe bias detected in 'CRC_incidence_val_Percent'. Consider re-evaluating data collection methods.

- Moderate bias detected in 'Alcohol use_Rate_Summary exposure value_val'. Investigate potential data skew or imbalance.

- Severe bias detected in 'Diet high in processed meat_Rate_Summary exposure value_val'. Consider re-evaluating data collection methods.

- Moderate bias detected in 'Diet high in red meat_Rate_Summary exposure value_val'. Investigate potential data skew or imbalance.

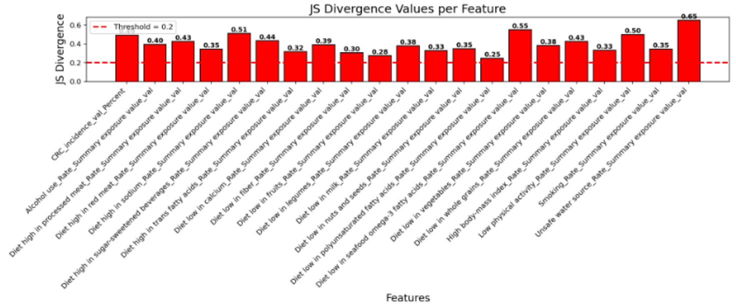


Figure: Js divergence bar plot

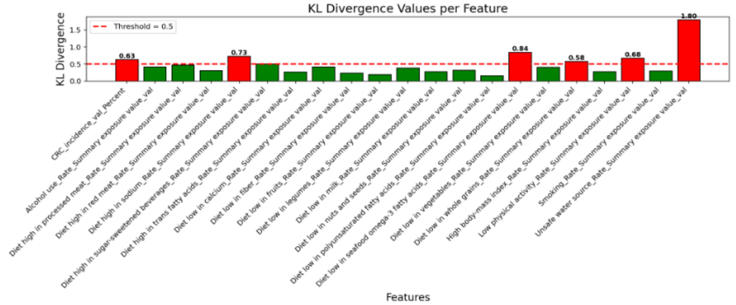
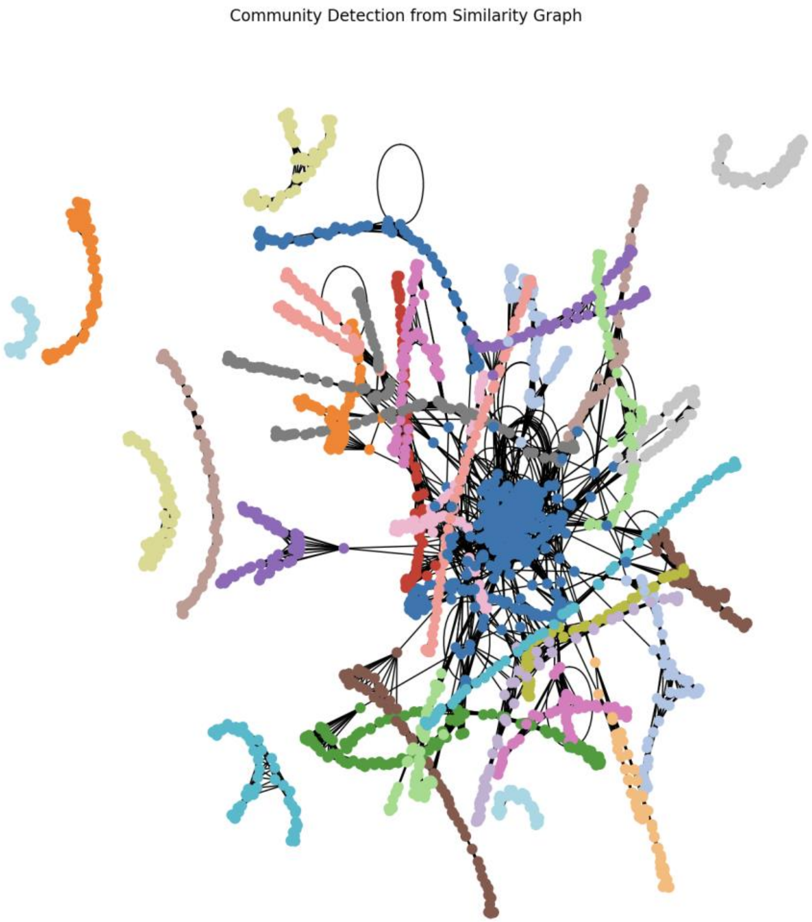


Figure: Kl divergence bar plot

PDF report – screenshot (2)



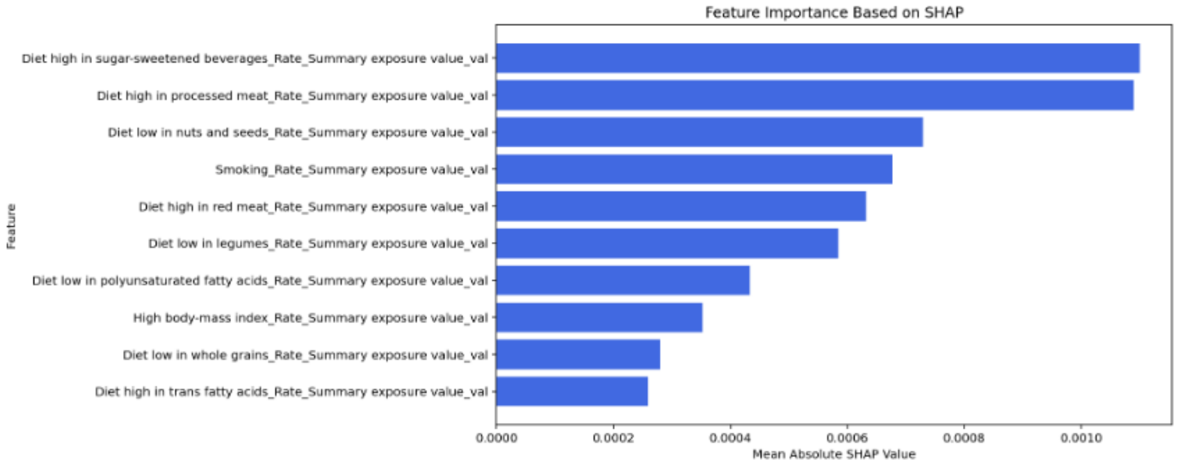
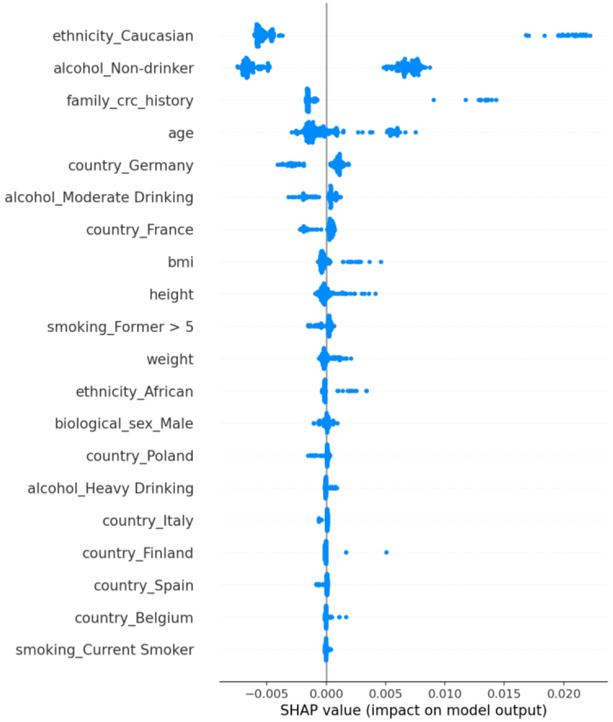


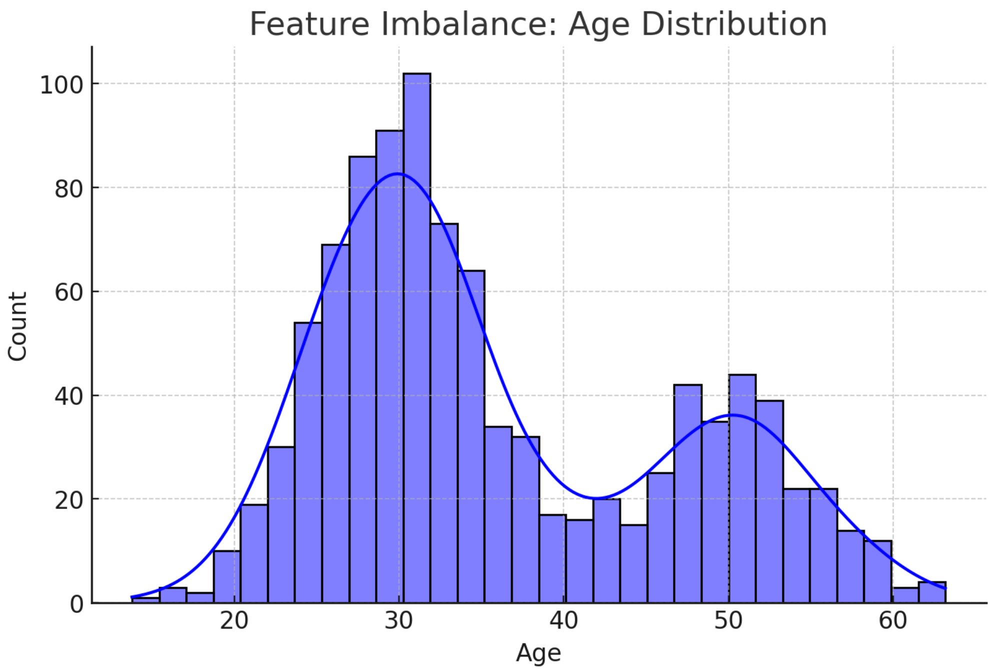
Figure: Shap feature importance. This figure illustrates anomaly scores, helping to identify unusual data points or bias patterns.



Bias in AI & APPO Bias Detection Framework

Data quality assessment:

- imbalanced data,
- missing data,
- skewness & kurtosis of data,
- coefficient of variation (CV),
- Etc…



Feature Imbalance: Age Distribution

The histogram depicts the distribution of the **Age** feature.

Interpretation: There is a clear overrepresentation of individuals around age 30, while those around age 50 are underrepresented.

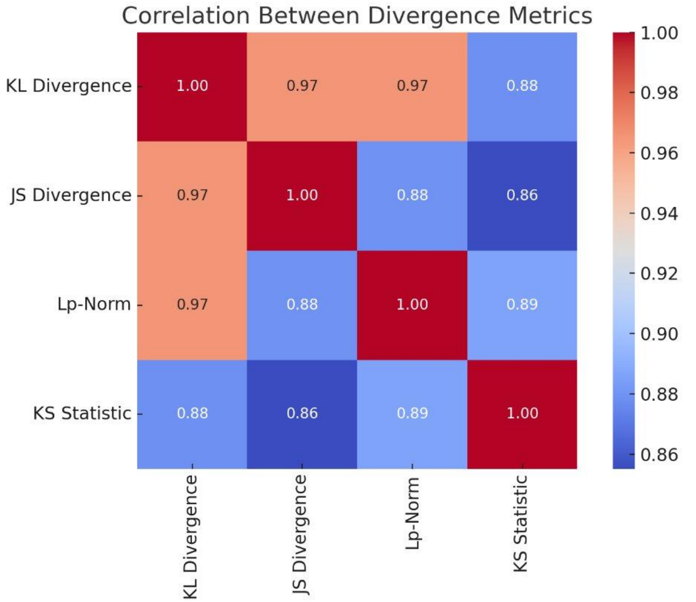
Bias Indicator: If age is a sensitive feature (e.g., for a disease probability of incidence), this imbalance may lead to biased model behavior.

Bias in AI & APPO Bias Detection Framework

Metrics Employed

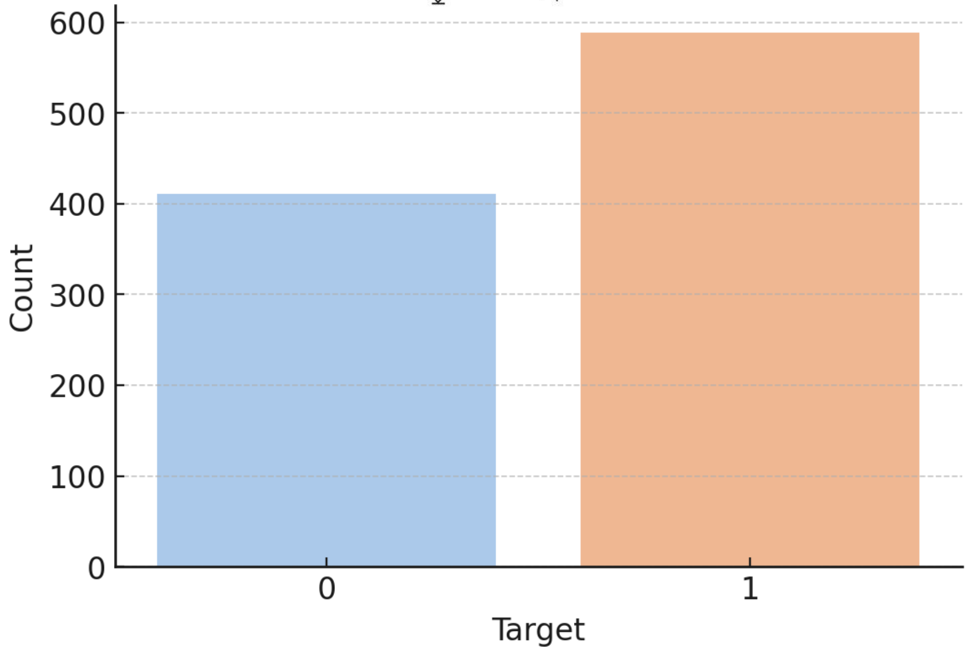
Traditional Bias Metrics - Distribution based!

- KL Divergence: Measures divergence between probability distributions.
- JS Divergence: A symmetric version of KL Divergence.
- **TVD**: Maximum difference between probability distributions.
- KS Test: Statistical test measuring distribution differences.
- Lp-Norm: measure the distance between two points in space



For example - Target Variable Disparity – Demonstrates how the target variable is unevenly distributed across different groups.

- A bar plot of the counts of each target class (0 and 1).
- One class (e.g., target=1) is significantly more frequent than the other, which indicates potential **class imbalance**.
- **Bias Indicator**: If the imbalance aligns with a sensitive feature (e.g., age, gender, or ethnicity), the model might **favor one group** over another.

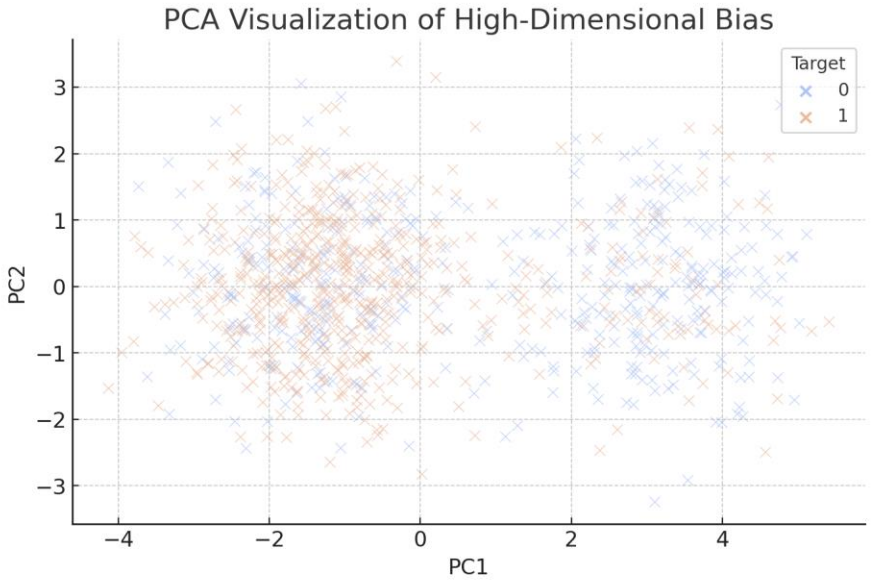


Bias in AI & APPO Bias Detection Framework

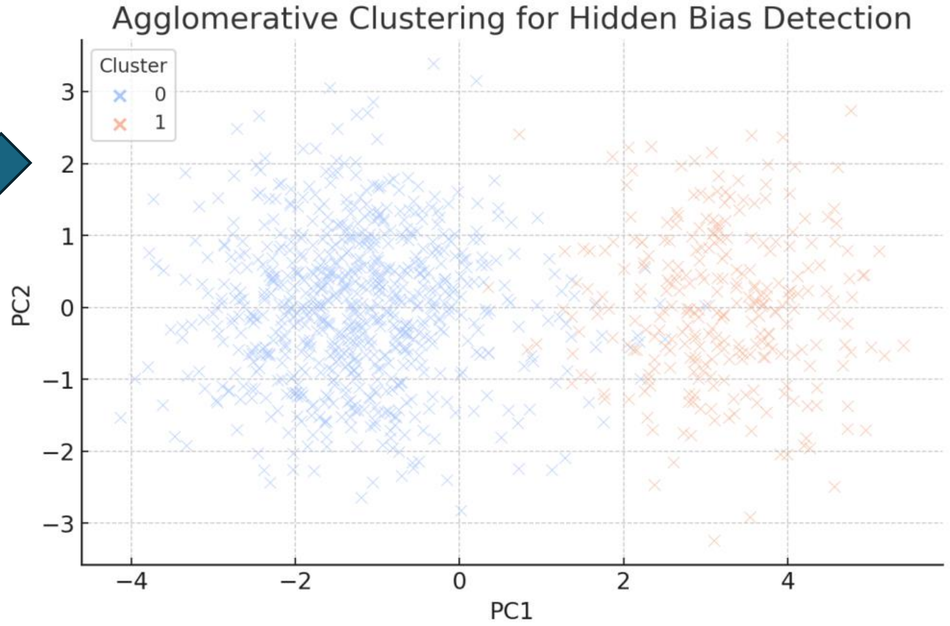
Metrics Employed

High-Dimensional Bias Detection

- PCA, t-SNE, UMAP for dimensionality reduction.
- K-Means, DBSCAN for clustering.
- Detecting hidden bias patterns.



Interpretation: The **separation** of the two classes in PCA space suggests that the model might rely heavily on certain patterns in the data.
Bias Indicator: If one cluster is predominantly from an overrepresented group (e.g., younger individuals), the model may favor that group, leading to biased outcomes.



Bias Indicator: If clusters strongly correlate with a sensitive attribute (e.g., age or gender), it indicates that the dataset naturally **reinforces certain separations**, potentially leading to biased predictions.

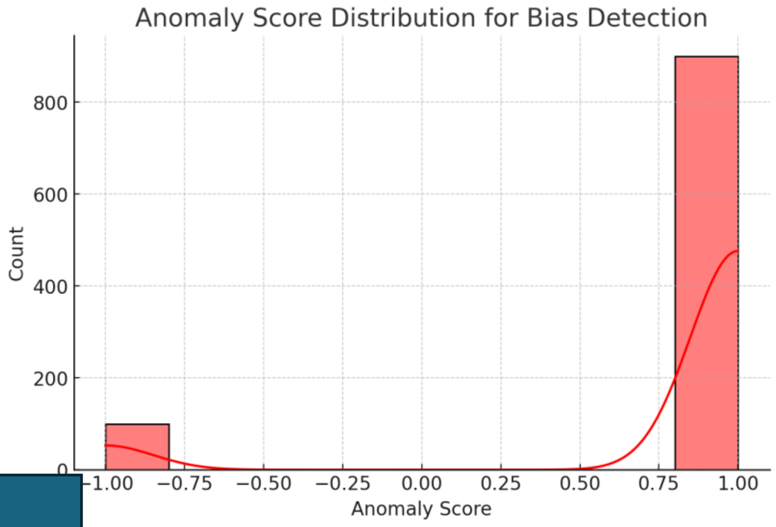
Bias in AI & APPO Bias Detection Framework

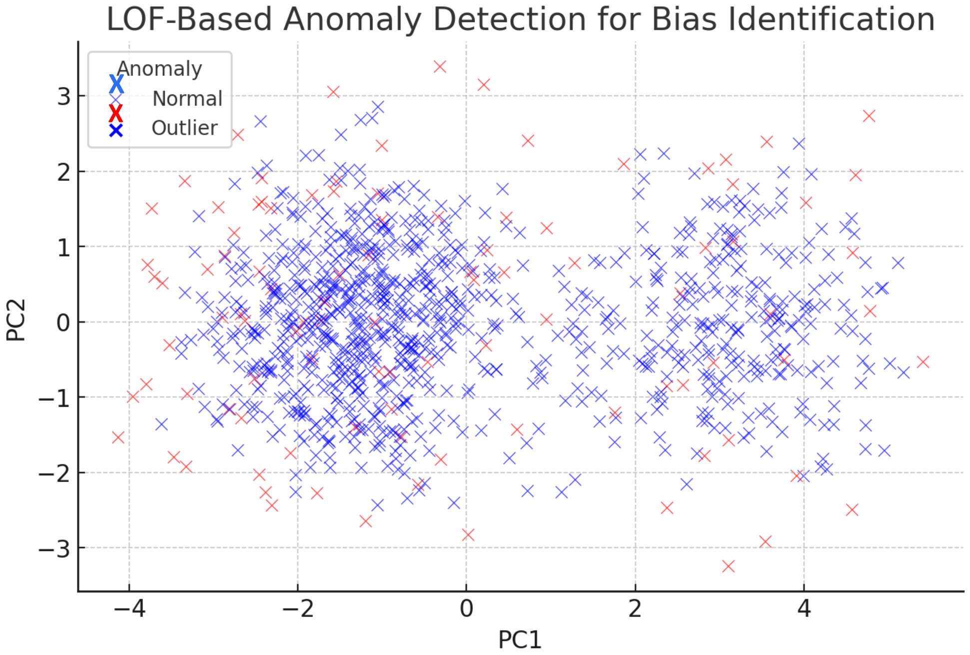
Metrics Employed

Anomaly Detection

- Isolation Forest & Local Outlier Factor (LOF) for anomaly scoring.
- Detecting unfair distributions in data.

Anomaly detection shows hidden biases, where certain groups appear as "outliers," meaning the model may perform poorly on them.





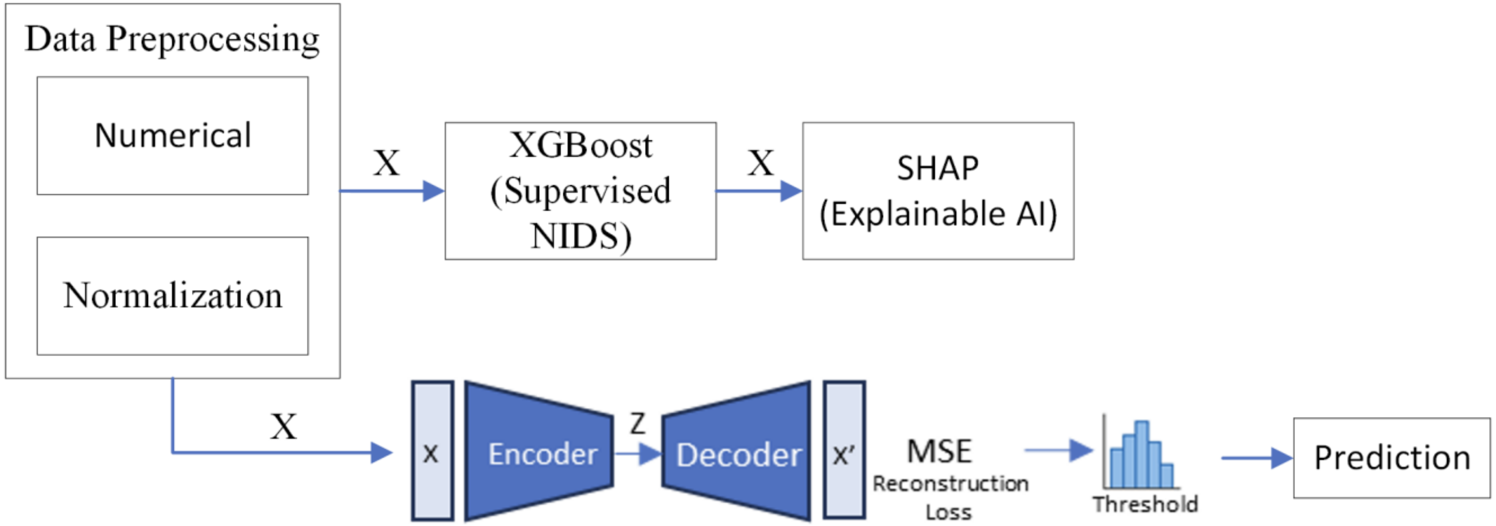
- The distribution of **anomaly scores** assigned by the **Isolation Forest** algorithm.
- The scores indicate how "unusual" a data point is compared to the rest of the dataset.
- **If a subgroup (e.g., older individuals) has higher anomaly scores, it suggests they are underrepresented**

Bias in AI & APPO Bias Detection Framework

SHAP-Based Bias Detection

- **Autoencoder** (AE) learns feature representations (trained to compress and reconstruct the dataset).
- The **reconstruction error** is used as an anomaly measure (higher errors suggest potential bias).
- **XGBoost surrogate model** predicts error levels.
- **SHAP Explainability Analysis**:
- Computes **SHAP values** for each feature.
- Flags features contributing most to **model errors**.
- Generates **SHAP feature importance plots**.

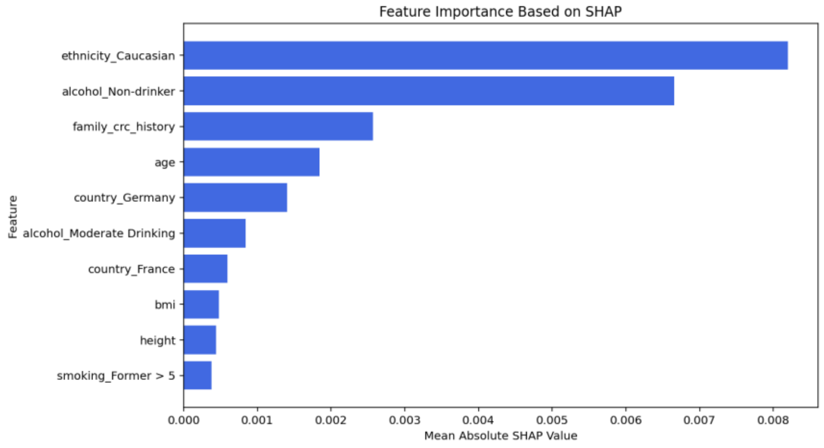
Output: SHAP-based feature attributions with bias severity levels.



Bias in AI & APPO Bias Detection Framework

SHAP Analysis for Model Bias

- SHAP-based feature importance.
- Permutation Importance for additional bias detection.
- Identifying the most biased features.



<u>Bias Risk Overview</u>:

•**High-Risk Bias:** Ethnicity and country-based features being strong influencers suggest potential biases.

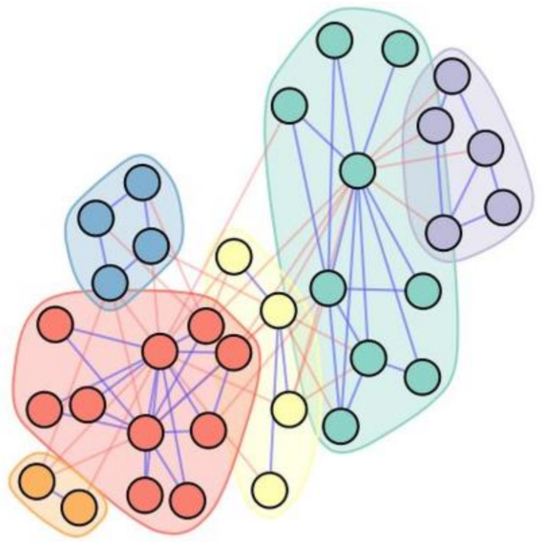
•**Moderate-Risk Bias:** Lifestyle factors like smoking and drinking may indirectly introduce bias.

•**Lower-Risk Bias:** Health-related factors (age, BMI) are expected influencers but still need fairness checks.

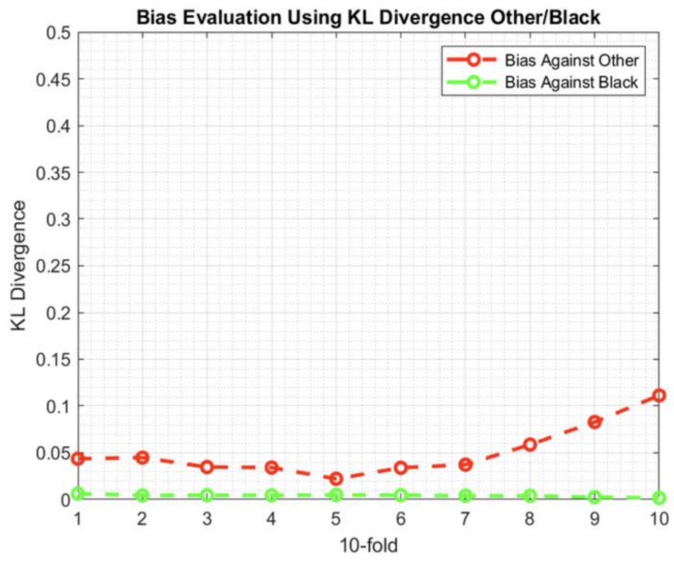
Bias in AI & APPO Bias Detection Framework

Community Bias Detection

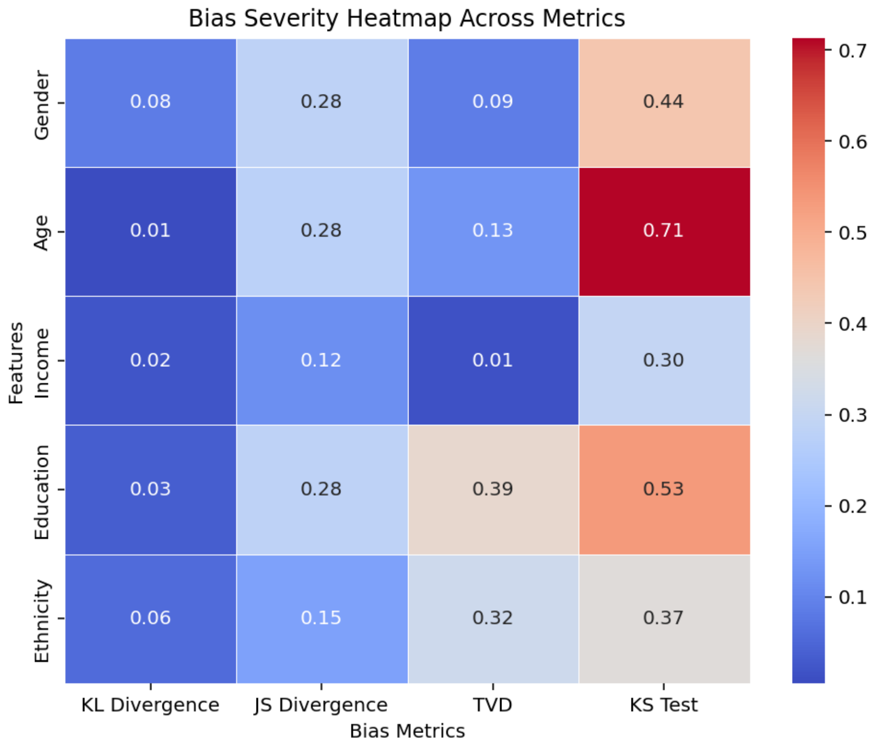
- Graph-based clustering to detect systemic bias.



- **Graph Representation:** Nodes represent individuals, and edges represent interactions or relationships between them.
- **Community Detection Results:** Different colors or shapes indicate the communities identified by the algorithm.
- **Uneven Community Sizes:** If the algorithm detects communities of vastly different sizes without justification, it may indicate a bias favoring larger or more connected groups.
- **Homogeneity Within Communities:** Overly homogeneous communities concerning attributes like race, gender, or age might suggest that the algorithm is grouping individuals based on these attributes, potentially reinforcing existing biases.
- **Isolation of Minority Nodes:** If nodes representing minority groups are isolated or grouped into separate communities without substantial reasoning, it could indicate bias in the detection process.



Bias in AI & APPO Bias Detection Framework



Automated Dynamic Bias Thresholding:

1. Mild Bias: 1 metric exceeded (Monitor)
2. Moderate Bias: 2 metrics exceeded (Investigate)
3. Severe Bias: 3+ metrics exceeded (Reassess)
4. Pervasive Bias: 5 metrics exceeded (Critical action needed)

Automated Bias Alerts

- System dynamically generates alerts when bias is detected.
- Thresholds adjust dynamically based on dataset distribution.
- Each exceeded threshold contributes to severity classification.
- Bias Mitigation Strategies

Feature	Severity	Alert Message
Age	Severe	"Bias detected, review data collection."
Gender	Moderate	"Potential bias, investigate distribution."

Bias Reporting & Auditing

JSON-Based Bias Reporting

- Ensures reproducibility.
- Provides structured, interpretable bias reports.

Insights for end users – Example of the usefulness and benefits from bias assessment.

Age Bias in Cancer Risk Prediction Model

E.g.: A cancer risk prediction model is deployed to support early diagnostics. It uses demographic and lifestyle data to estimate individual risk levels. However, a bias assessment revealed that the training dataset had an overrepresentation of individuals aged around **30** and an underrepresentation of those aged **50 and above**.

Bias Indicator:

- Age distribution histogram showed a skewed dataset.
- PCA and clustering analysis revealed that younger individuals formed distinct clusters influencing model outcomes.
- SHAP analysis showed **age** as a top bias-contributing feature.

Real-World Impact, on the basis of end users (patients & clinicians):

- Older individuals will receive **systematically lower** risk scores despite having **higher actual** risk.
- Potential lead to delayed or missed screenings for the elderly population.
- Trust in AI recommendations **are declined** when clinicians observe mismatches with real-world clinical assessments.

Benefits of Bias Assessment:

- **Early Detection:** The APPO framework flagged age imbalance as a potential bias source.
- **Fairness Improvement:** Data was rebalanced, and model retraining improved predictions across age groups.
- **Clinical Trust:** Improved alignment between AI predictions and clinical observations restored user confidence.
- **Regulatory Compliance:** Bias reporting supports **ethical and legal requirements** for **equitable AI in healthcare**.

Bias in AI & APPO Bias Detection Framework

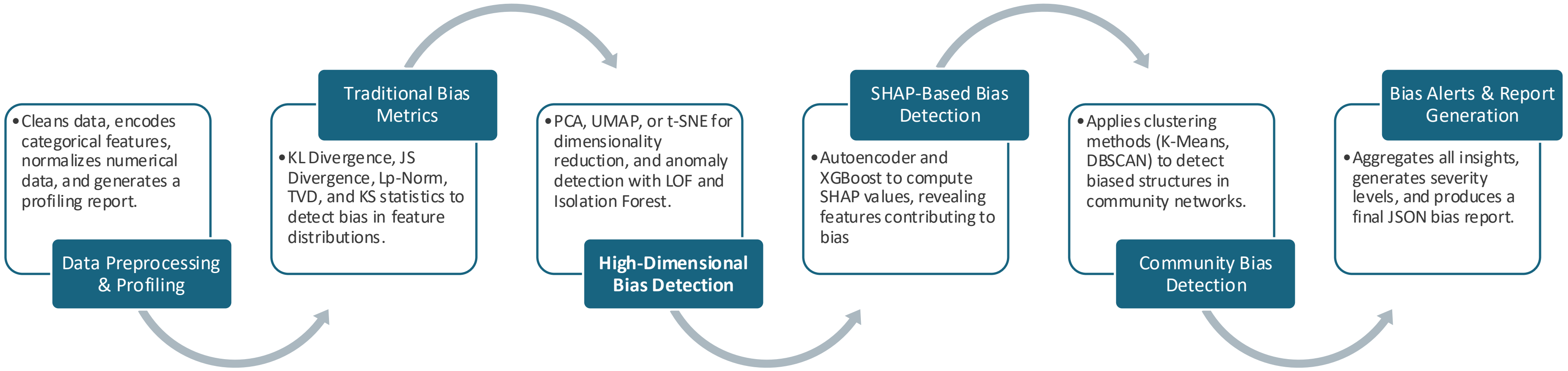
Summary & Future Improvements

Key Features & Innovations:

- **Multi-Phase Bias Detection**:
  - Data quality checks, anomaly scoring, clustering.
- Automated Bias Thresholding:
  - Severity levels from mild to pervasive bias.
- Bias Auditing & Reporting:
  - JSON-based structured reports for reproducibility.

Future Enhancements to be included:

- Automating bias detection in CI/CD pipelines.
- Advanced high-dimensional fairness auditing.



ONCODIR

https://www.oncodir.eu/

https://www.linkedin.com/company/oncodir/posts/?feedView=all

ONCODIR

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.