ISSN: **3107-6513**

Bias Drift Detection in AI Models

Shri Keerthanaa[#1], Thamizharasan[*2]
#Department of Computer Science, Rathinam College of Arts and Science (Autonomous),
Coimbatore, Tamilnadu, India
shrikeerthanaa48@gmail.com, *tamilofficialmailbox@gmail.com*

Abstract *- As Artificial Intelligence (AI) models are increasingly deployed in sensitive sectors like finance and recruitment, maintaining long-term fairness is a critical challenge. While models may be fair at the time of training, changes in real-world data distributions can lead to "bias drift," where predictions become discriminatory over time. This research proposes a system that continuously monitors deployed models using metrics such as Demographic Parity (DP) and Equal Opportunity (EO). The system employs a statistical thresholding approach to detect significant deviations and provides real-time alerts via an interactive Dash-based dashboard. Experimental results demonstrate the system's ability to successfully identify bias drift in an Adult Income model, providing a practical solution for responsible AI governance*

Keywords *- Machine Learning, Bias Drift, AI Fairness, Demographic Parity, Continuous Monitoring, Dash, Data Preprocessing.*

1. INTRODUCTION

AI and Machine Learning (ML) models are increasingly used to support decision-making in healthcare, finance, and recruitment. However, these models are prone to developing bias over time due to shifts in incoming data or environmental conditions, a phenomenon known as bias drift. Bias drift is a silent challenge that can degrade the fairness of AI systems after deployment, leading to ethical concerns and legal risks.

Traditional MLOps primarily focuses on technical performance metrics like accuracy and latency, often neglecting fairness monitoring in production. This project, "Bias Drift Detection in AI Models," aims to fill this gap by developing a system that continuously evaluates model predictions against fairness metrics.

1.1 Object and Scope

The objective of this research is to develop an automated framework for the real-time detection of bias drift. The scope involves training models on diverse datasets (Adult Income, Heart Disease, and Credit Risk), implementing a monitoring script to simulate real-time data processing, and visualizing ethical trends through a web-based dashboard.

2. LITERATURE REVIEW

Existing AI monitoring systems primarily focus on performance metrics and system reliability. Tools such as Prometheus, Grafana, and AWS SageMaker Model Monitor track metrics like accuracy, latency, and data drift. While these systems are effective in detecting technical issues, they do not adequately address fairness monitoring.

Fairness-focused tools such as IBM AI Fairness 360 and Microsoft Fair learn provide techniques for bias detection and mitigation. However, these tools are mostly used during the pre-deployment phase. They perform static analysis and do not support continuous monitoring of fairness in production environments.

Recent research has highlighted the importance of monitoring fairness dynamically. However, many existing solutions are complex, expensive, or difficult to integrate into real-world systems. There is a need for a lightweight,

ISSN: 3107-6513

scalable, and real-time solution for bias drift detection.

The proposed system addresses this gap by combining fairness evaluation, statistical drift detection, and real-time visualization into a single integrated framework.

3. METHODOLOGY

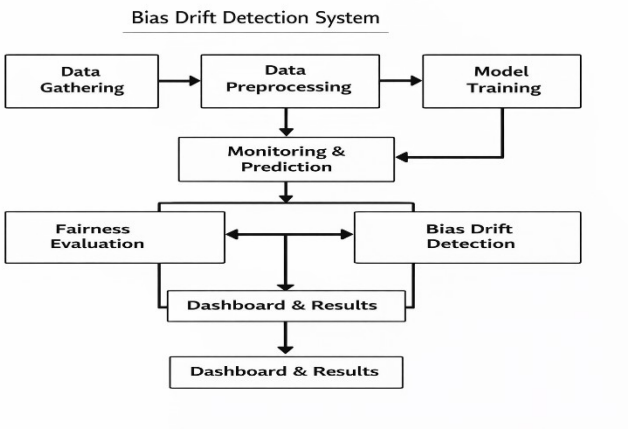

Fig 1. Data Work Flow

3.1. Data Gathering Module

This foundational stage involves collecting structured datasets such as Adult Income, Heart Disease, and Credit Risk from public repositories. The module ensures that sensitive attributes like age and gender are present, as these are critical for subsequent fairness evaluations.

3.2. Data Preprocessing Module

Raw data is cleaned to ensure consistency. Key tasks include

- Median Imputation: Handling missing values to maintain data integrity.
- Encoding: Converting categorical variables into numeric formats for compatibility with ML models.
- Feature Selection: Dropping irrelevant columns (like IDs) to prevent the introduction of unintentional bias.

3.3. Model Training Module

Using scikit-learn, various classification models are trained and optimized for predictive accuracy. Once validated, these models are serialized into .pkl files via joblib, allowing them to be easily loaded into the live monitoring script.

3.4. Machine Learning Classification Algorithms

These algorithms are used to train the models that generate predictions for your datasets.

- **Random Forest Classifier**: This is the primary algorithm demonstrated in your sample code for the **Adult Income** model. It works by building multiple decision trees and averaging their predictions to improve accuracy and handle complex data patterns.
- **Scikit-Learn Suite**: The project is designed to support various algorithms available in the scikit-learn library. While Random Forest is featured, the system architecture allows for the integration of other classifiers for the **heart disease** and **Credit Risk** datasets

3.5. Monitoring & Prediction Module

A dedicated script (monitor.py) simulates real-time deployment by processing incoming data in

ISSN: 3107-6513

batches. It generates predictions using the pre-trained models and passes the results to the evaluation engine.

3.6. Fairness Evaluation Module

The Fairness Evaluation Module serves as the system's ethical engine, transforming model predictions into quantifiable metrics to assess bias. This module specifically monitors how different protected groups are treated to ensure that the AI does not perpetuate or amplify existing societal inequalities.

3.6.1 Demographic Parity (DP) Calculation

Demographic Parity ensures that the likelihood of a positive outcome is equal across different demographic groups, regardless of their true labels.

- **Logic**: It measures the difference in the rate of positive predictions (e.g., "Income > 50K" or "Low Credit Risk") between a protected group (like Female) and a reference group (like Male).
- **Goal**: To ensure the model's selection rate is independent of sensitive attributes.

3.6.2 Equal Opportunity (EO) Calculation

Equal Opportunity is a more refined metric that focuses specifically on the "True Positive Rate" (TPR) to ensure that qualified individuals from all groups are treated equally.

- **Logic**: It ensures that individuals who should actually receive a positive outcome (the "qualified" ones) have an equal probability of being correctly identified by the model, regardless of their group.
- **Application**: This is essential for ensuring fairness in "qualified" populations, such as identifying equally healthy patients in your heart disease model or equally solvent applicants in the Credit Risk model.

3.7. Bias Drift Detection Module

This module acts as the "alarm" for the system. It uses statistical thresholding to compare current fairness metrics against historical averages. If a metric deviates by more than two standard deviations, it flags a "Bias Drift" event. This was successfully demonstrated when the Adult Income Model DP metric shifted to 0.1958, triggering a drift alert.

3.8. Visualization & Dashboard Module

The frontend, built with Dash and Plotly, converts complex backend data into intuitive visuals. Users can monitor trend graphs, gauge charts, and real-time alert cards to track model behaviour and ensure accountability.

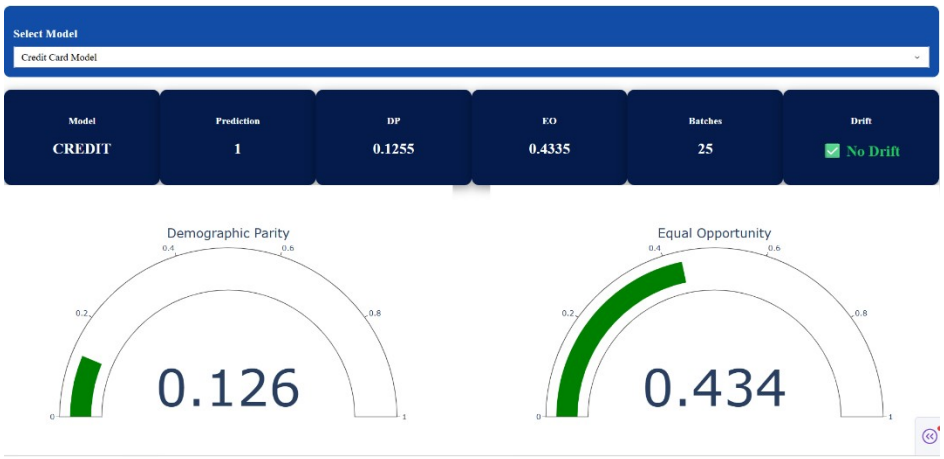

Fig 2. Monitoring Dashboard

ISSN: **3107-6513**

4. RESULTS

The system was tested using simulated real-time data batches. The monitoring script recorded the following results across 25 batches:

Model Name	DP (Initial)	DP (Final)	Drift Status
Adult Income	0.1775	0.1958	Drift Detected
Heart Disease	0.5543	0.6058	No Drift
Credit Risk	0.1255	0.1255	No Drift

The results confirm that static pre-deployment audits are insufficient for long-term fairness. By treating fairness metrics as live signals, this system effectively bridges the gap between technical performance and ethical accountability.

5. CONCLUSION

The Bias Drift Detection System successfully addresses the challenge of maintaining fairness in deployed AI models. By shifting the focus from static pre-deployment audits to continuous, real-time monitoring, the system ensures that ethical degradation is identified and addressed promptly. The integration of fairness metrics with an interactive dashboard provides stakeholders with the transparency needed to manage responsible AI systems effectively. Future work will explore automated retraining mechanisms and the integration of explainable AI (XAI) tools to diagnose the root causes of detected drift.

6. ACKNOWLEDGEMENT

This article / project is the outcome of research work carried out in the **Department of Computer Science under the DBT Star College Scheme**. The authors are grateful to the Department of Biotechnology (DBT), Ministry of Science and Technology, Govt. of India, New Delhi, and the **Department of Computer Science** for the support.

7. REFERENCES

Core Fairness & Bias Drift References

1. Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning.
   - *Relevance*: Provides the formal mathematical foundations for the Equal Opportunity (EO) metric used in your system.
2. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness Through Awareness.
   - *Relevance*: One of the seminal papers defining Demographic Parity (DP) and the ethical necessity of treating different demographic groups equitably.
3. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A Survey on Concept Drift Adaptation.
   - *Relevance*: Offers a comprehensive overview of how data distributions shift over time, which is the underlying cause of the bias drift your project detects.
4. Bellamy, R. K., et al. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias.
   - *Relevance*: References the IBM toolkit mentioned in your existing system analysis, highlighting the shift from static auditing to continuous monitoring.

ISSN: **3107-6513**

5. Bird, S., et al. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI.

- *Relevance*: Supports your methodology for assessing ethical implications using Python-based frameworks.

Statistical & Technical Tool References

6. Massey, F. J. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit.

- *Relevance*: Provides the basis for the ks_2samp algorithm used in your detect_drift module to identify significant changes in data batches.

7. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python.

- *Relevance*: The primary library used for your Random Forest implementation and model evaluation.

8. Sievert, C. (2020). Interactive Data Visualization with Plotly, Dash, and R.

- *Relevance*: Documents the capabilities of the Dash and Plotly frameworks used to build your interactive real-time dashboard.