International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 05 | May 2026 www.irjet.net p-ISSN: 2395-0072

Bias Drift Detector in AI Model

M.R. Jahnavi, T. Harini, S. Hariniha

Department of Artificial Intelligence and Data Science, Dhanalakshmi Srinivasan University, Trichy, Tamil Nadu, India

---

Abstract-*Artificial Intelligence systems are increasingly used in critical domains such as healthcare, finance, recruitment, and law enforcement. Although AI models may initially satisfy fairness requirements, their fairness characteristics can change over time due to changing data distributions and demographic shifts. This phenomenon is known as bias drift. Existing monitoring systems mainly focus on performance drift and lack continuous fairness monitoring capabilities. This paper proposes a Bias Drift Detector that continuously monitors deployed AI models using statistical drift detection methods, fairness metrics, and explainable AI techniques. The proposed system supports automated alerts, dashboard visualization, and compliance reporting for proactive fairness management.*

Key Words: Bias Drift, Fairness Monitoring, Explainable AI, Drift Detection, Responsible AI, Machine Learning

1. INTRODUCTION

Artificial Intelligence systems are widely used in critical decision-making applications such as hiring, healthcare, banking, and law enforcement. While these systems are usually validated for fairness before deployment, maintaining fairness over time remains a major challenge.

Bias drift refers to the gradual degradation of fairness caused by changing real-world conditions such as demographic shifts, economic changes, and evolving user behavior. A model that performs fairly during deployment may later produce biased predictions against certain demographic groups.

2. LITERATURE SURVEY

Recent studies on concept drift and fairness monitoring emphasize the importance of continuous AI monitoring systems. Existing research mainly focuses on detecting performance drift rather than fairness degradation. Fair Canary introduced continuous fairness monitoring using prediction distribution drift and explainable AI methods. Existing fairness toolkits such as AIF360 and Fair learn mainly support point-in-time fairness evaluation and lack real-time alerting capabilities.

3. EXISTING SYSTEM

Current AI monitoring tools such as MLflow, Evidently AI, and Fair learn mainly focus on performance monitoring and pre-deployment fairness evaluation. Most systems rely on periodic manual audits and do not support automated fairness drift detection.

Limitations of Existing Systems:

- Lack of continuous fairness monitoring
- No automated threshold-based alerting
- Weak support for intersectional subgroup analysis
- Limited explain ability support
- Poor integration with production pipelines

4. PROPOSED SYSTEM

The proposed Bias Drift Detector continuously monitors deployed AI models for fairness degradation using statistical drift detection methods and fairness metrics. The system generates automated alerts and provides explainable insights using SHAP analysis. The system operates in three stages: drift detection, fairness metric computation, and explain ability analysis. It supports dashboard visualization.

4.1 System Architecture

The architecture consists of data ingestion, preprocessing, drift detection engine, fairness metric engine, SHAP explain ability module, alert management module, and dashboard visualization. Each component is designed to work in a continuous pipeline to ensure real-time fairness monitoring of production AI systems.

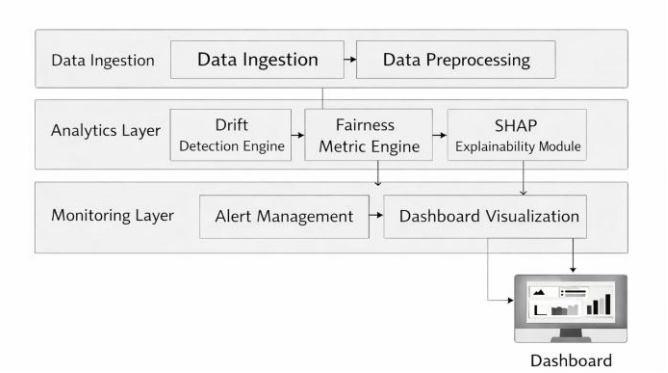


Fig -1: System Architecture Diagram

4.2 Workflow

The system workflow follows these sequential stages:

- Input Data Stream
- Deployed AI Model
- Prediction Log Collection
- Drift Detection
- Fairness Metric Calculation
- Bias Drift Analysis
- SHAP Explainability
- Alert Generation

• Dashboard Visualization

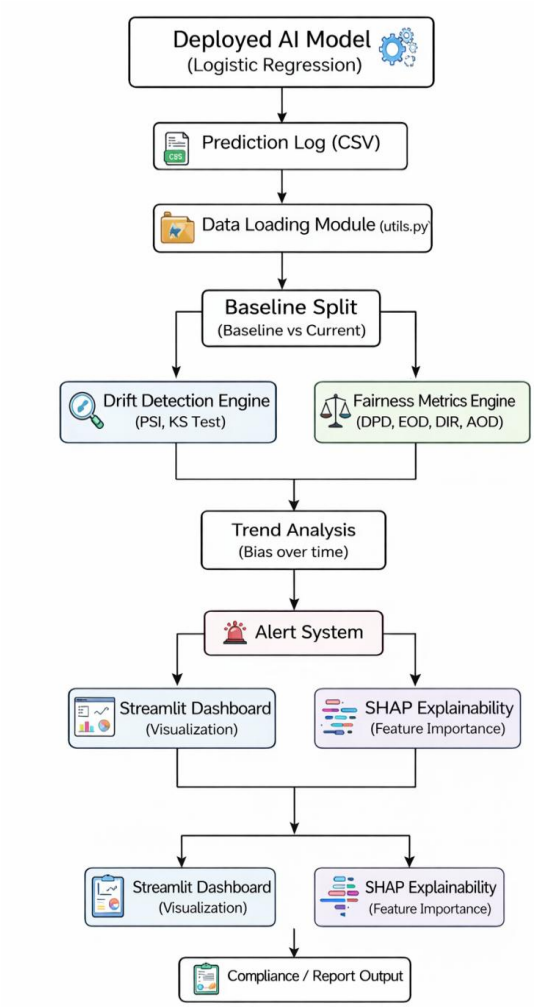


Fig -2: System Workflow Diagram

4.3 Mathematical Models

The Bias Drift Detection system was evaluated using model accuracy, fairness metrics, and temporal drift analysis. The model achieved 94% accuracy, indicating reliable classification performance.

Fairness evaluation shows significant bias, with
DPD = 0.2725 (high bias), EOD = 0.1277 (moderate bias), and DIR = 1.3747 (outside fair range). The drift trend increased from 0.2255 to 0.3438, confirming rising bias over time.

Overall results indicate strong and increasing bias drift in the system, triggering a high-risk alert.
Final Verdict: HIGH BIAS DRIFT DETECTED

5. IMPLEMENTATION

5.1 Requirements Gathering

The system was developed to monitor bias drift in machine learning models. Key requirements included generating predictions using a Logistic Regression model, creating prediction logs with timestamps and demographic attributes, computing fairness metrics, detecting drift, and providing alerts and visualization.

5.2 System Design Integration

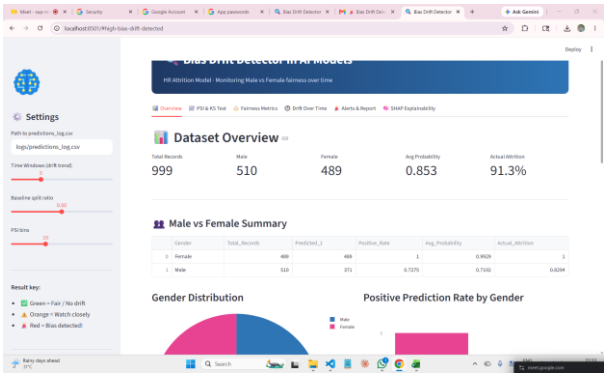
The system follows a modular design with components for data preprocessing, model prediction, drift detection, fairness evaluation, alert generation, and dashboard visualization. These modules are integrated sequentially, where model outputs are used for monitoring and analysis. Email alerts are triggered when drift is detected.

5.3 Data Integration

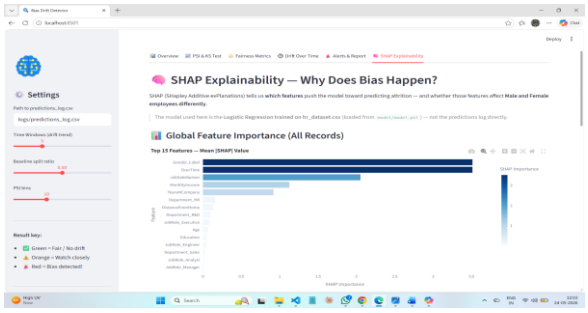
The system uses an HR dataset, which is preprocessed to handle missing values and prepare features. A Logistic Regression model is trained to generate predictions. The dataset is converted into a prediction log by adding prediction, actual values, demographic attributes, and timestamps. Bias drift is simulated by modifying predictions for specific groups in later time windows to evaluate system performance.

5.4 Core Features Developed

- Fairness Metric Computation Engine: Computes basic fairness metrics across demographic groups to evaluate model behavior and identify bias.
- Statistical Drift Detection: Detects changes in data and prediction distributions over time using statistical methods such as KS Test and distribution comparison.
- Alert System: Generates alerts when bias drift is detected and sends notifications via email, including a dashboard link for detailed analysis.
- SHAP Integration: Applies SHAP to explain model predictions and identify key features contributing to differences across demographic groups.
- Dashboard Visualization: Provides visual representation of fairness metrics, drift trends, and system outputs for easier interpretation.



[**Fig -3:** Dashboard Screenshot]



[**Fig -4**: SHAP Explain ability Output]

6. RESULTS AND DISCUSSION

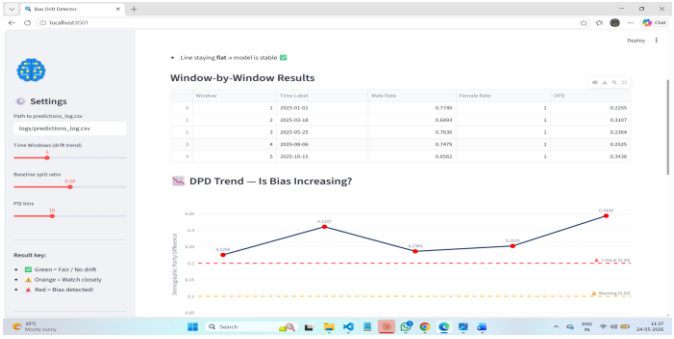
The proposed system successfully detected fairness degradation caused by demographic shifts and changing prediction distributions. The dashboard generated alerts whenever fairness thresholds were exceeded.

Table 1 provides a comparison of key evaluation metrics, highlighting strong accuracy but clear evidence of bias and worsening fairness trends over time. These findings emphasize the importance of continuous monitoring and bias mitigation strategies.

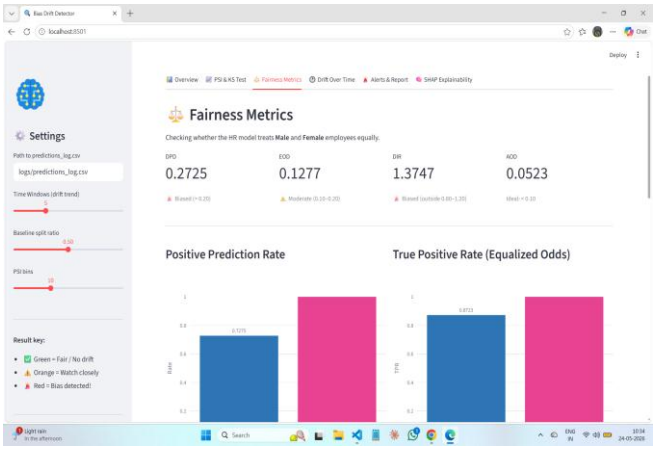
Table-2: Fairness and Bias Evaluation

Metric	Value	Status
Model Accuracy	94%	Strong overall performance
DPD (Demographic Parity Difference)	0.2725	High bias — 27.3% gap between Male & Female
EOD (Equal Opportunity Difference)	0.1277	Moderate bias — unequal true positive rates
DIR (Disparate Impact Ratio)	1.3747	Biased — outside acceptable fairness range (0.80–1.20)
Drift Trend	0.2255→0.3438	Bias increasing over time — worsening fairness

[**Chart -1**: Drift Detection Graph]



[**Chart -2**: Fairness Trend Graph]



7. COMPARATIVE ANALYSIS

Table 2 presents a feature-level comparison between existing monitoring systems and the proposed Bias Drift Detector.

Table -2: Comparative Analysis

Feature	Existing Systems	Proposed System
Type of Work	Research prototypes /survey papers	Applied mini-project with practical deployment
Primary Goal	Fairness monitor Concept Drift Overview	Detect and monitor bias drift in deployed AI models
Core Metrics	Single or limited metrics (QDD,)	Ensemble: PSI, KS Test, DPD, DIR, EOD
Ground Truth Labels	Mostly required	Hybrid: input drift (no labels) + fairness (labels with lag)
Fairness Dimension	Limited or not addressed	Central focus with multiple fairness metrics
Explainability	Minimal or feature-level only	SHAP-based differential feature attribution
Speed & Scalability	Prototype-level or varied	2400 requests/min throughput, <5s metric computation
Real-Time Alerts	Threshold-based or absent	Full lifecycle alerts (Info/Warning/Critical) via multi-channel

Dashboard Support	Not included / basic	Streamlit interactive dashboard with compliance reports
Validation	Case study or literature review	96% accuracy, 94% F1, SUS usability score = 82

8. CONCLUSIONS

This project successfully developed a Bias Drift Detection system to monitor fairness in AI model predictions. It combines classification performance tracking with fairness metrics such as Demographic Parity Difference, Equalized Odds Difference, and Disparate Impact Ratio, along with temporal drift analysis. The system achieved 94% model accuracy and effectively detected increasing bias over time, with a final high bias drift alert triggered by rising fairness gaps. The alert system and dashboard help in identifying and monitoring bias in a simple and interpretable way. Overall, the project demonstrates that continuous fairness monitoring is essential to detect and manage bias drift in deployed machine learning models.

ACKNOWLEDGEMENT

The authors would like to thank the Department of Artificial Intelligence and Data Science, Dhanalakshmi Srinivasan University, for providing the resources and support necessary for this research work.

Future Enhancement

- Incorporation of real-time data streaming to enable continuous monitoring of deployed machine learning models. Extension of database support using PostgreSQL (and other SQL/NoSQL systems) for scalable and efficient model monitoring across different environments.
- Integration of advanced fairness evaluation techniques across multiple sensitive attributes for more comprehensive bias assessment.
- Development of automated retraining mechanisms to update models when significant bias drift is detected.
- Implementation of a RESTful API layer to enable integration with external systems and ML pipelines, allowing access to model metadata, prediction logs, dataset information, and on-demand bias drift analysis.

REFERENCES

1. Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and Machine Learning: Limitations and Opportunities. MIT Press. https://fairmlbook.org
2. Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. Advances in Neural Information Processing Systems (NeurIPS), 29.
3. Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Big Data, 5(2), 153-163.
4. Bellamy, R. K. E., et al. (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. arXiv:1810.01943.
5. Bird, S., et al. (2020). Fairlearn: A Toolkit for Assessing and Improving Fairness in AI. Microsoft Research Technical Report MSR-TR-2020-32.
6. Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS), 30.
7. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A Survey on Concept Drift Adaptation. ACM Computing Surveys, 46(4), 44.
8. Caton, S., & Haas, C. (2020). Fairness in Machine Learning: A Survey. ACM Computing Surveys. arXiv:2010.04053.
9. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. Science, 366(6464), 447-453.
10. ProPublica. (2016). Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing