FailNetX: A Temporal Stress-Testing Framework for Machine Learning Systems

Vedavi V[1], Dr.M.C.S Geetha[2]
[1]*UG Scholar, Department of Data Science Kumaraguru College of Liberal Arts and Science Coimbatore, India*
[2]*Assistant Professor, Department of Computer Applications, Kumaraguru College of Technology Coimbatore, India*

Abstract: Machine learning models often perform well during training but lose reliability after deployment due to changes in data distributions. This paper presents AI Resilience Lab, a web-based platform designed to proactively evaluate model behaviour under realistic data drift conditions before production deployment. The system integrates a drift simulation engine, a SHAP-based explainability module for identifying feature-level causes of degradation, and a unified risk assessment framework that quantifies model reliability. Implemented using React and FastAPI, the platform supports time-based simulations across varying drift intensities. Experimental results on benchmark datasets demonstrate effective performance degradation analysis, accurate root cause identification, and early warning of potential model failure. The proposed approach addresses key limitations of existing MLOps practices by enabling proactive model robustness evaluation instead of reactive monitoring.

Keywords: Machine Learning, Data Drift, Model Robustness, Explainable AI, SHAP, Risk Assessment, MLOps, Model Monitoring.

I.INTRODUCTION

Machine learning models are increasingly deployed in real-world applications such as healthcare, finance, and decision-support systems. While these models often perform well during training and validation, their performance frequently degrades after deployment due to changes in data distributions over time. This phenomenon, commonly referred to as data drift, includes covariate shift in input features, concept drift in feature-target relationships, and degradation in data quality caused by noise or missing values.

Most existing machine learning workflows rely on post-deployment monitoring techniques that detect issues only after model performance has declined. Such reactive approaches increase operational risk and provide limited insight into the underlying causes of failure. In addition, current tools for drift detection, explainability, and risk assessment are often fragmented, requiring manual interpretation and intervention.

To address these challenges, this paper proposes the AI Resilience Lab, an integrated platform that enables proactive evaluation of model robustness through controlled drift simulation, explainable analysis, and unified risk scoring. By simulating future data scenarios and analysing model behaviour before deployment, the system helps practitioners anticipate performance degradation, identify vulnerable features, and make informed decisions regarding retraining and deployment readiness.

II. LITERATURE REVIEW

2.1 Data Drift and Model Performance Degradation

Machine learning models deployed in real-world environments are highly sensitive to changes in data distributions over time. This issue, commonly referred to as data drift, has been widely studied due to its direct impact on prediction accuracy and system reliability. Gama et al. [1] provided one of the earliest comprehensive studies on concept drift, classifying it into sudden, gradual, and recurring forms. Their work primarily focused on post-deployment drift detection and adaptive learning strategies.

Rabanser et al. [2] evaluated several statistical techniques for identifying dataset shift, including the Kolmogorov–Smirnov test and Maximum Mean Discrepancy. While these methods were effective in detecting distributional changes, the study highlighted their limitations in anticipating future drift scenarios. Similarly, Lu et al. [3] reviewed concept drift handling techniques and emphasised the importance of continuous monitoring. However, most existing approaches rely on detecting drift only after it has occurred, which limits their usefulness for proactive decision-making.

These studies demonstrate that although drift detection techniques are well established, they

remain largely reactive and provide limited support for testing model robustness before deployment.

2.2 Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) has gained increasing importance, particularly in applications where transparency and trust are critical. Lundberg and Lee [4] introduced SHAP, a unified framework based on Shapley values that assigns consistent feature importance scores for model predictions. SHAP has since become a widely adopted explanation technique due to its solid theoretical foundation and compatibility with various model types.

Ribeiro et al. [5] proposed LIME, which focuses on explaining individual predictions using local approximations. While LIME provides valuable insights at the instance level, its explanations may vary across runs and lack global consistency. Molnar [6] presented a detailed overview of interpretability techniques, discussing trade-offs between model accuracy and explainability.

Although XAI methods effectively explain individual predictions, most existing techniques do not address how feature importance evolves or how explanations can be linked to performance degradation caused by data drift. This limits their ability to support root cause analysis in dynamic production environments.

2.3 Machine Learning Operations (MLOps)

As machine learning systems have become more prevalent in production, MLOps practices have emerged to manage deployment, monitoring, and maintenance challenges. Breck et al. [7] discussed data validation and monitoring strategies used in large-scale ML systems, highlighting issues such as training-serving skew. Sculley et al. [8] introduced the concept of technical debt in machine learning pipelines and emphasised the importance of continuous monitoring to prevent long-term degradation.

Paleyes et al. [9] surveyed real-world ML deployment case studies and identified gaps in proactive evaluation and long-term reliability assessment. While existing MLOps tools provide monitoring dashboards and alerting mechanisms, they largely focus on detecting problems after deployment. Polyzotis and Roy [10] emphasised data lifecycle management but did not address resilience testing under future drift conditions.

Overall, current MLOps solutions lack integrated tools that combine drift simulation, explainability, and risk assessment within a single framework.

2.4 Identified Research Gaps

Based on the literature, several gaps are evident. First, most existing methods focus on detecting drift after deployment rather than evaluating model behaviour under simulated future conditions. Second, explainability techniques are rarely applied to understand performance degradation over time. Third, monitoring, explanation, and risk analysis are often implemented as separate systems, increasing operational complexity. Finally, there is a lack of unified metrics that combine performance degradation, drift severity, and confidence loss into a single interpretable risk indicator.

This research addresses these gaps by proposing an integrated platform that enables proactive evaluation of model robustness through drift simulation, explainability, and unified risk assessment.

III. EXISTING SYSTEM

Most organisations monitor deployed machine learning models using dashboards that track metrics such as accuracy, precision, and recall, with alerts triggered when thresholds are exceeded. While these tools provide visibility into past performance, they are largely reactive, often identifying issues only after the model has already caused operational or business impact. Statistical monitoring techniques based on distribution tests are also used, but they become difficult to manage for high-dimensional data and frequently produce false alarms that require manual tuning.
A/B testing is another common practice for comparing model versions, but it involves exposing real users to experimental models, which is risky in critical applications and offers only limited, comparison-based insights. Overall, existing approaches suffer from a lack of proactive evaluation, limited explainability, fragmented tooling, and the absence of a unified risk metric that can be easily understood by both technical teams and decision-makers, making timely and informed model maintenance challenging.

IV. PROPOSED SYSTEM

The proposed system, AI Resilience Lab, is designed to evaluate the robustness of machine learning models before deployment by simulating realistic data drift scenarios. Instead of relying on reactive post-deployment monitoring, the system enables

proactive testing of model behaviour under changing data conditions, helping identify potential performance risks in advance.AI Resilience Lab follows a layered architecture consisting of a user interface, a processing layer, and a data storage layer. Users upload datasets and configure drift settings through a web interface. The system automatically preprocesses data, trains a baseline model, simulates different types of data drift over time, and evaluates model performance. An explainability module identifies features responsible for performance changes, while a risk assessment module combines multiple indicators into a unified risk score with actionable recommendations. This integrated workflow supports early decision-making and improves model reliability in real-world deployments.

User Interface (React)

↓

REST API (FastAPI)

↓

Data Preprocessing

↓

Drift Simulation Engine

↓

Model Evaluation

↓

Explanation Engine (SHAP)

↓

Risk Scoring & Recommendations

V.METHODOLOGY

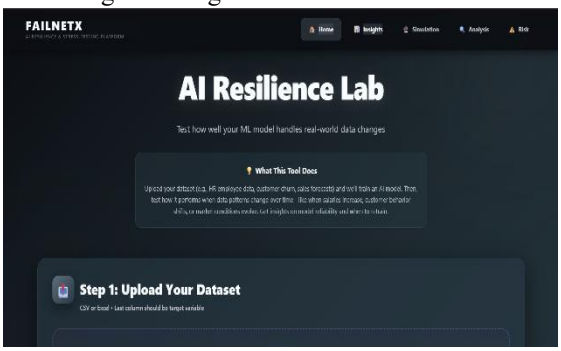
This study follows an experimental methodology to evaluate the effectiveness of the AI Resilience Lab framework. The methodology includes system implementation, controlled data drift simulation, explainability analysis, and risk assessment using benchmark datasets under different drift conditions. Three publicly available datasets were selected to represent diverse real-world scenarios: a telecom Customer Churn dataset, a Credit Card Fraud dataset with significant class imbalance, and a Medical Diagnosis dataset for multi-class classification. Standard preprocessing steps such as handling missing values, feature encoding, and data normalisation were applied before training the models.

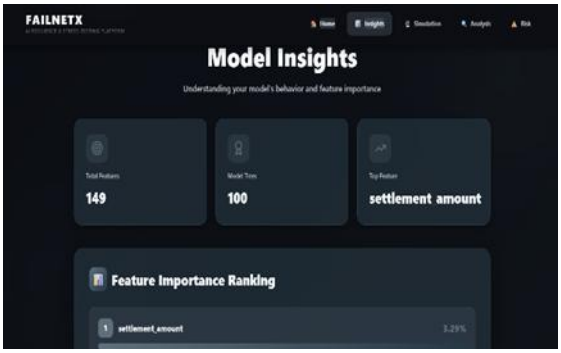
Random Forest classifiers were used as baseline models due to their stable performance and suitability for SHAP-based explanations. Drift simulations were carried out across multiple time steps with varying intensity levels to model covariate drift, concept drift, and combined drift scenarios. At each time step, model performance metrics, feature importance changes, and risk scores were recorded to analyse degradation trends and early warning behaviour.

VI. RESULTS AND ANALYSIS

6.1 Baseline Model Performance

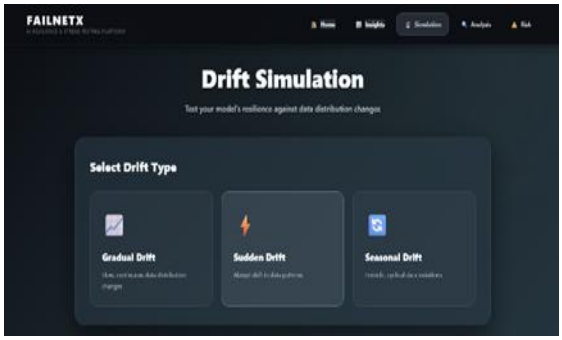
Initial evaluation showed that all baseline models achieved strong performance before drift was introduced. The Customer Churn dataset achieved high classification accuracy, the Credit Card Fraud dataset maintained strong results despite class imbalance, and the Medical Diagnosis dataset showed reliable multi-class prediction performance. These results established a stable reference point for evaluating degradation under drift.





6.2 Effect of Data Drift on Performance

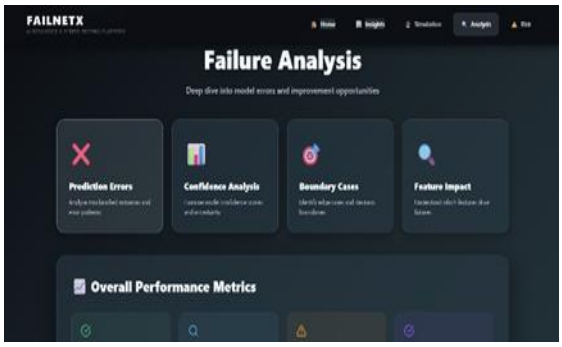
As drift intensity increased over time, a gradual decline in model performance was observed across all datasets. Combined drift scenarios resulted in the most significant degradation, indicating that simultaneous changes in data distribution and feature relationships have a compounding effect on model reliability. These findings align with real-world observations reported in prior studies.



6.3 Explainability and Root Cause Analysis

The SHAP-based explanation module successfully identified changes in feature importance across time steps. Several features exhibited noticeable shifts in contribution as drift progressed, indicating their role in performance degradation.

The consistency between SHAP explanations and permutation-based importance measures confirms the reliability of the explanation engine for root cause identification.



6.4 Risk Assessment and Early Warning

The unified risk scoring framework effectively classified model reliability into predefined risk levels. Risk scores increased progressively with drift intensity, enabling early identification of potential failure conditions. In multiple cases, the system detected rising risk levels several time steps before significant performance loss occurred, demonstrating its usefulness as an early warning mechanism.

6.5 Comparative Discussion

Compared to conventional MLOps tools that rely on post-deployment monitoring, the proposed system enables proactive testing of model robustness before deployment. By integrating drift simulation, explainability, and risk assessment into a single workflow, AI Resilience Lab reduces manual analysis effort and supports informed decision-making.

VII.CONCLUSION

This paper presented AI Resilience Lab, a unified framework for proactively evaluating the robustness of machine learning models under realistic data drift conditions. The system integrates drift simulation, SHAP-based explainability, and a weighted risk assessment mechanism within a single web-based platform. Through controlled experiments on benchmark datasets, the platform demonstrated its ability to simulate gradual performance degradation, identify feature-level causes of model behaviour changes, and provide early warning signals before critical failure occurs.

By shifting the focus from reactive post-deployment monitoring to pre-deployment resilience evaluation, the proposed approach helps practitioners better understand long-term model behaviour and take informed actions such as retraining or mitigation in advance. Overall, AI Resilience Lab contributes a practical and systematic solution for improving the reliability and trustworthiness of deployed machine learning systems.

VIII.SCOPE FOR FUTURE ENHANCEMENT

Future work will focus on improving drift realism by learning drift patterns from historical production data and exploring adversarial drift generation techniques. The platform can be extended to support additional model types, including neural networks, regression tasks, and multi-modal data. Incorporating alternative explainability methods such as counterfactual analysis and causal inference may further enhance root cause identification. Integration with real-time production pipelines and automated remediation strategies represents another promising direction for improving long-term model reliability.

REFERENCES

[1] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, “A survey on concept drift adaptation,” *ACM Computing Surveys*, vol. 46, no. 4, pp. 1–37, 2014.

[2] S. Rabanser, S. Günnemann, and Z. C. Lipton, "Failing loudly: An empirical study of methods for detecting dataset shift," in *Advances in Neural Information Processing Systems*, 2019.
[3] J. Lu, A. Liu, F. Dong, G. Zhang, J. Gama, and G. Gu, "Learning under concept drift: A review," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 12, pp. 2346–2363, 2018.
[4] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017.
[5] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. ACM SIGKDD*, pp. 1135–1144, 2016.
[6] C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/
[7] E. Breck, N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, "Data validation for machine learning," in *Proc. SysML Conference*, 2019.
[8] D. Sculley et al., "Hidden technical debt in machine learning systems," in *Advances in Neural Information Processing Systems*, 2015.
[9] A. Paleyes, R. G. Urma, and N. D. Lawrence, "Challenges in deploying machine learning: A survey of case studies," *ACM Computing Surveys*, vol. 55, no. 6, pp. 1–29, 2022.
[10] G. I. Webb, R. Hyde, H. Cao, H. L. Nguyen, and F. Petitjean, "Characterizing concept drift," *Data Mining and Knowledge Discovery*, vol. 30, no. 4, pp. 964–994, 2016.