Project Design
Phase-I Solution
Architecture

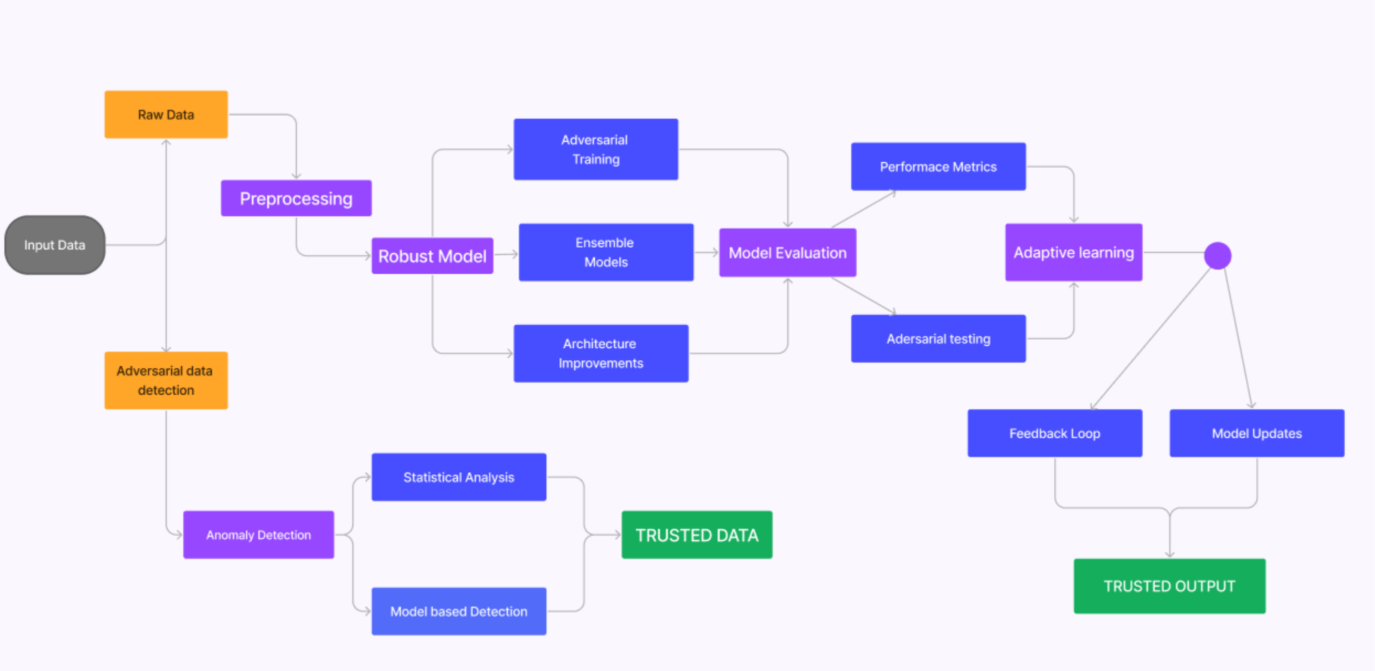
Date	24 October 2023
Team ID	Team-593456
Project Name	Project - Adversarial attacks and Defenses
Maximum Marks	4 Marks

Solution Architecture:

The solution architecture below offers a detailed framework for bolstering the defense of AI systems against adversarial attacks while simultaneously upholding the integrity of their outputs. It encompasses several essential phases, starting with data processing. Raw input data is subjected to preprocessing, which includes techniques like normalization and feature extraction. In addition, there's a critical adversarial data detection step that sifts out potentially malicious inputs, ensuring that only trusted data is processed further.

The core of the architecture lies in the development of a robust AI model. This is achieved through a multi-faceted approach, which includes adversarial training, where adversarial examples are incorporated during model training to enhance its resistance to attacks. Ensemble models are employed to combine the strength of multiple models, further fortifying the system's overall resilience. Architectural improvements are made to the model to enhance its ability to withstand adversarial attacks.

To validate the model's effectiveness and robustness, the architecture incorporates thorough evaluation and testing. Performance metrics are used to assess the model's accuracy and its capability to mitigate false positives. Real-world readiness is assured through adversarial testing, where various attack scenarios are simulated, and the model's performance in resisting such attacks is evaluated. The architecture is designed to be adaptive, incorporating a continuous feedback loop for monitoring the model's performance and making necessary updates to adapt to evolving threats. It concludes with the delivery of trusted and secure AI outputs, fortifying AI systems against adversarial intrusions while preserving their reliability in a dynamic and evolving threat landscape.


Raw Data
Input Data
Preprocessing
Robust Model
Adversarial Training
Ensemble Models
Architecture Improvements
Model Evaluation
Performace Metrics
Adersarial testing
Adaptive learning
Feedback Loop
Model Updates
TRUSTED OUTPUT
Adversarial data detection
Anomaly Detection
Statistical Analysis
Model based Detection
TRUSTED DATA