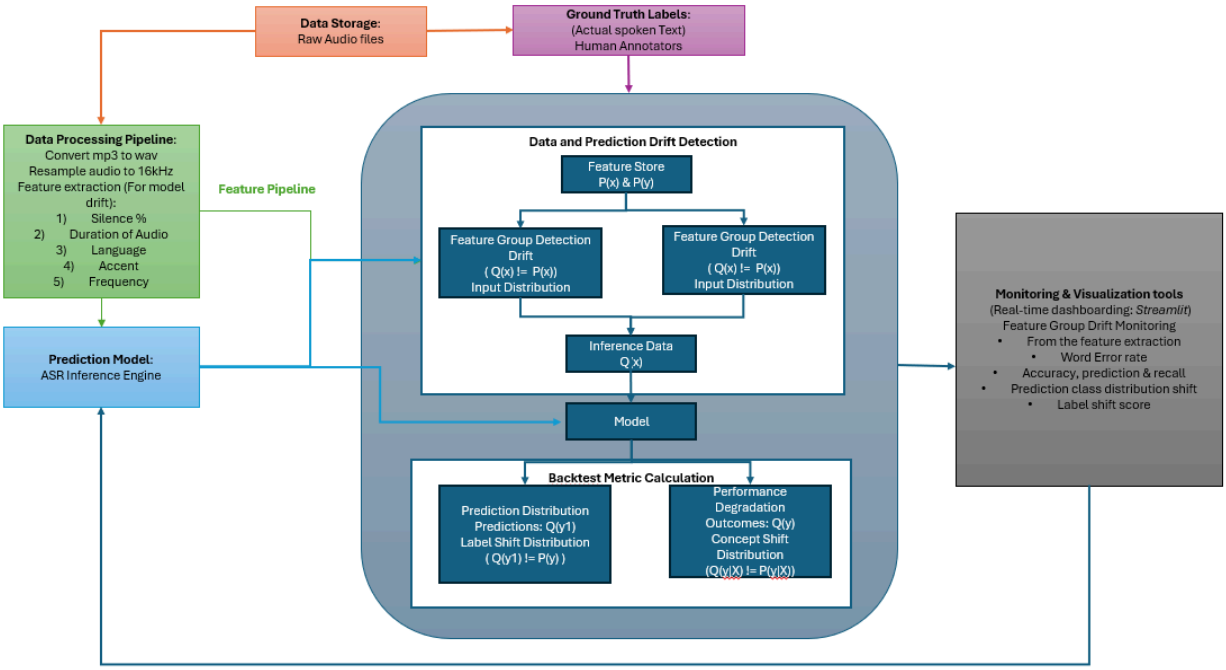

Fig 1: Proposed Model Pipeline

The Model monitoring pipeline provides an end-to-end solution for tracking and evaluating the ASR model's performance across time while managing model drift in production deployment. In order to track model drift we will follow 8 main steps.

Firstly, the data storage will be a centralised repository which contains raw audio files. These files act as an input for the ASR model and also are used as data for monitoring of the model and retraining based on the feedback loop. Consistent access to historical data is assured. Then we obtain ground truth labels likely manually prescribed by human annotators for greater accuracy.Prior to feeding the model with input data, the audio is converted from MP3 to WAV format. In the same step we perform feature extraction which is crucial for drift monitoring. Example features can include: silence percentage, duration, detected language, accent, and frequency. The preprocessed audio data is passed into the ASR model where the transcription predictions are generated. These predictions are logged for later drift analysis and performance monitoring.

Next, we perform data and prediction drift detection. This is essential for monitoring how data and model behavior change over time. The first step in drift detection is to compare the distribution of training data (P(X)) with the distribution of incoming feature data (Q(X)). A significant difference suggests feature group drift, which could have an impact on the model's performance. This drift can be measured with the use of instruments like Kolmogorov-Smirnov tests. The system simultaneously tracks prediction distributions (Q(y1)) and contrasts them with the training data's label distribution (P(y)). This procedure, called label shift detection, finds

changes in the kinds of predictions the model is producing. For example, even though the input data seems consistent, the prediction distribution may alter if users abruptly submit different kinds of audio or language content.

The system then assesses the model's performance in this step by contrasting its predictions (Q(y1)) with ground truth results (Q(y)) that are supplied by human annotators. Word Error Rate (WER), which detects insertions, deletions, and substitutions, is the main metric used to assess transcription accuracy. A rising WER could be a sign of concept drift, which is the change in the relationship between inputs and expected outputs, possibly brought on by background noise, language, or accent differences. Additional metrics, such as accuracy, precision, and recall, if applicable, are also used to monitor performance degradation. This thorough assessment shows whether retraining is required and whether the model's behavior has deviated from its training goals.

This will be coupled with insights derived in a real-time dashboard that can be built on softwares such as streamlit which can plot distributions and time series insights to study feature drift and WER trends. Thresholds can be set such that an alert is sent when these thresholds are breached to take necessary action. This can include a retaining process through a feedback loop where the new labelled data is incorporated to update the model.