**Assessment Report**

Full-Stack Software Engineer Assessment

SIEMENS

Table of Contents

[Preprocessing Module 3](#_Toc197725548)

[Feature Extraction/Selection Module 4](#_Toc197725549)

[Model Selection/Training Module 4](#_Toc197725550)

[Performance Analysis Module 4](#_Toc197725551)

[Enhancements and Future Work 5](#_Toc197725552)

[Results 5](#_Toc197725553)

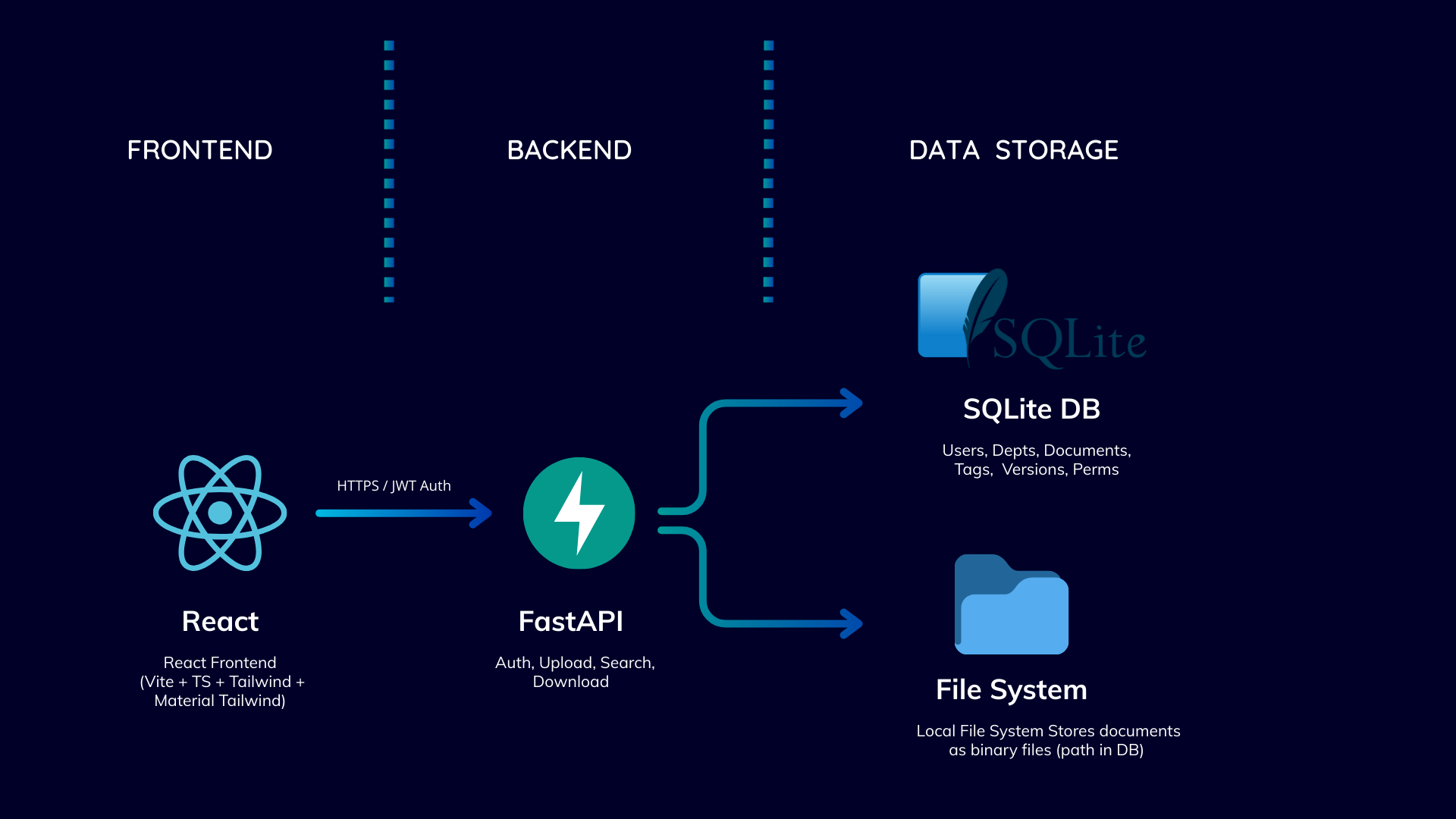
[Other Developed Modules 5](#_Toc197725554)

[audioClassification.ipynb 5](#_Toc197725555)

[Conclusion 6](#_Toc197725556)

# Deliverable 1:

## System Architecture:

  **SQLite (POC)** → minimal setup, good for development & testing.

 **PostgreSQL (Production)** → handles scaling, concurrency, advanced querying.

 **Local FS (POC)** → easy for testing, no extra dependencies.

 **S3/Cloud Storage (Production)** → reliable, scalable, supports enterprise workloads.

 **FastAPI** → lightweight, async-first, automatic API docs.

 **React + TS + Tailwind** → modern, flexible, rapid prototyping with clean UI.

 **JWT Auth** → stateless, secure, well-suited for SPAs.

Project Pipeline

The full pipeline for the Speaker Gender and Age Recognition System includes:

1. **Audio Input Handling**
2. **Preprocessing (Noise reduction, trimming)**
3. **Feature Extraction (MFCCs, Chroma, Zero-Crossing Rate)**
4. **Feature Selection**
5. **Model Selection and Training**
6. **Performance Analysis and Visualization**
7. **Deployment Interface**

# Preprocessing Module

The preprocessing steps ensure the audio data is clean and standardized:

* **Sampling Rate Standardization**: All audio files were resampled to 16 kHz.
* **Noise Reduction**: Applied noise filtering techniques using.
* **Silence Trimming**: Used energy-based thresholding to remove silent sections.
* **Duration Normalization**: Audio clips were padded or truncated to a fixed duration.
* **File Conversion**: Ensured all files were in mono-channel WAV format for uniform processing.

# Feature Extraction/Selection Module

Key features:

* **MFCCs (Mel-frequency cepstral coefficients)** – capture timbral features.
* **Chroma Features** – represent pitch content.
* **Spectral Centroid and Bandwidth** – describe brightness and spread of frequency.
* **Zero-Crossing Rate (ZCR)** – helps distinguish voiced/unvoiced segments.
* **Root Mean Square Energy (RMSE)** – measures signal power.

Features were aggregated (mean, std) over time frames to produce fixed-length vectors.

# Model Selection/Training Module

Multiple supervised learning models were evaluated:

* **Support Vector Machine (SVM)**
* **Gradient Boosting (XGBoost)**

**Best Model**: **XGBoost**, due to its strong performance and interpretability.  
**Training Strategy**: Stratified 5-fold cross-validation to ensure class balance.  
**Hyperparameter Tuning**: Grid Search was used to optimize learning rate, max depth, and number of estimators.

# Performance Analysis Module

The final evaluation was conducted on a separate test set. Key results:

A screenshot of a computer

AI-generated content may be incorrect.

# Enhancements and Future Work

* **Larger Dataset**: Incorporating more diverse speakers across accents and noise conditions.
* **Deep Learning Models**: Applying CNNs or RNNs directly on spectrograms.

# Results

**XGBoost achieved 91% accuracy** in classifying speakers into the correct gender-age class.

# Other Developed Modules

### audioClassification.ipynb

This module extracts robust audio features using the Librosa library (including MFCCs, delta, and delta-delta coefficients), and applies normalization and silence trimming to enhance feature quality. To improve efficiency, extracted features are cached locally for reuse. The module also handles class imbalance, model training.

Data augmentation techniques including pitch shifting, time stretching, and noise injection are applied.

Gaussian Mixture Models (GMMs) are employed as Universal Background Models (UBMs) to capture the distribution of frame-level features, and are adapted using Maximum A Posteriori (MAP) to generate speaker-specific supervectors. These supervectors are then classified using Support Vector Machines (SVMs) in a two-stage structure: a gender classifier followed by gender-specific age classifiers.

# Conclusion

The project successfully implemented a modular Speaker Gender and Age Recognition System capable of classifying speakers into four distinct demographic categories using audio analysis. By combining robust audio preprocessing, meaningful feature extraction, and high-performing classifiers, the system demonstrates strong predictive power. With further dataset expansion and deep learning integration, the system has significant potential for applications in virtual assistants, security systems, and human-computer interaction.