Project Title: auralDine

Student Name(s): V Sai Sumedha, Yuvika Sai Simhadri, K Sri Karuna Reddy

Roll Number(s): 23WH1A05F0, 23WH1A05G7, 23WH1A05H6

Department & Institution: BVRIT Hyderabad College of Engineering for Women

Date of Submission: 30/11/25

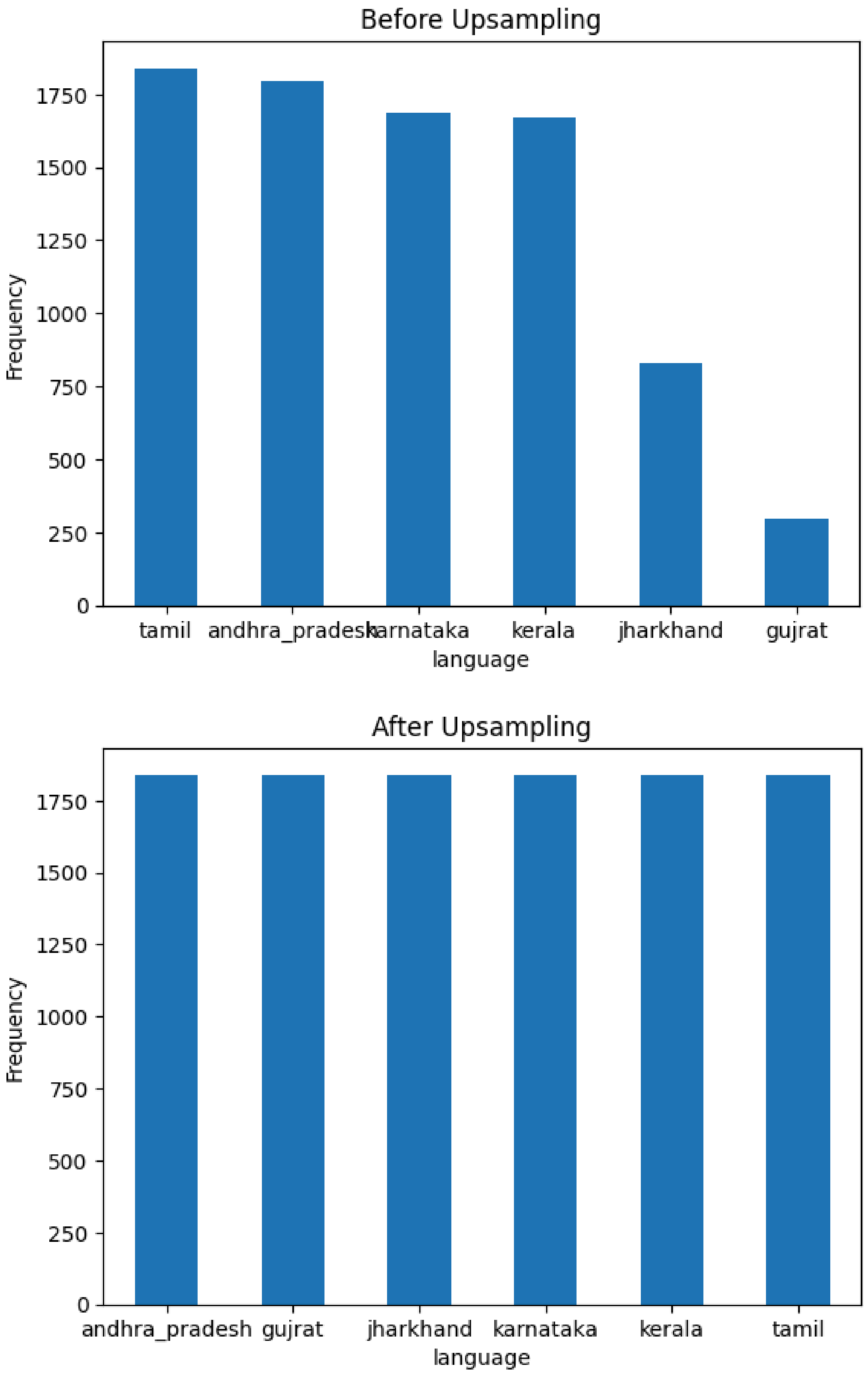
1. Project Scope and Deliverables

1.1 Project Scope

AuralDine is an AI speech analytics system that detects a speaker's Indian regional accent from English speech and recommends regional cuisines that reflect their cultural roots. It combines MFCC features, HuBERT embeddings, and deep learning classifiers so that diners can discover "the taste of home" based on how they speak.

1.2 Deliverables

- MFCC-based feature extraction and upsampling pipeline (balanced to 1840 clips per accent).
- HuBERT Base embedding extraction (13 representational layers, 768-D per layer).
- Accent classification models:
  - MFCC + Random Forest baseline.
  - HuBERT layer-wise SVMs.
  - HuBERT + CNN, Transformer, and BiLSTM classifiers.
- Hyperparameter tuning and validation for deep models.
- Age generalization study (adult-trained models tested on child speech).
- Word-level vs sentence-level accent detection analysis.
- Final trained HuBERT + CNN-BN model (.pt).
- Confusion matrices, ROC curves, classification reports, and t-SNE visualisations.
- AuralDine application that maps predicted accent → region → cuisine recommendations.



Class balance before and after upsampling for MFCC features.

2. Model Development

2.1 Feature Extraction

- MFCC Features:

- Audio is resampled to a consistent sampling rate and converted to 13-coefficient MFCCs with delta and delta-delta coefficients, giving 39-D feature vectors.
- Features are averaged over time to form one fixed-length vector per clip and used as input to classical classifiers (KNN, SVM, Decision Tree, Random Forest, Logistic Regression, Naive Bayes).

- HuBERT Embeddings:

- HuBERT Base is loaded to produce frame-level 768-D embeddings from a convolutional feature encoder followed by 12 transformer layers.
- For each audio clip, embeddings from each layer are mean-pooled over time, yielding 13 separate 768-D vectors (layers 0–12) used for layer-wise SVMs, deep CNN/Transformer models, word vs sentence analysis, and age generalization experiments.

2.2 Model Architectures

MFCC baselines (classical ML)

- Multiple classifiers were trained on MFCCs; test accuracies were:
    - KNN: ≈ 99.00%
    - Decision Tree: ≈ 94.23%
    - SVM: ≈ 83.70%
    - Random Forest: ≈ 99.46%
    - Logistic Regression: ≈ 79.71% (with convergence warnings)
    - Gaussian Naive Bayes: ≈ 75.48%

Random Forest was chosen as the MFCC baseline with 99.46% test accuracy on 6-way accent classification.

HuBERT + classical ML

- SVMs with RBF kernels were trained on 768-D HuBERT embeddings. Early layers (2-3) achieved ≈ 99% accuracy, while higher layers focused more on content and showed slightly reduced accent sensitivity.

HuBERT + deep models

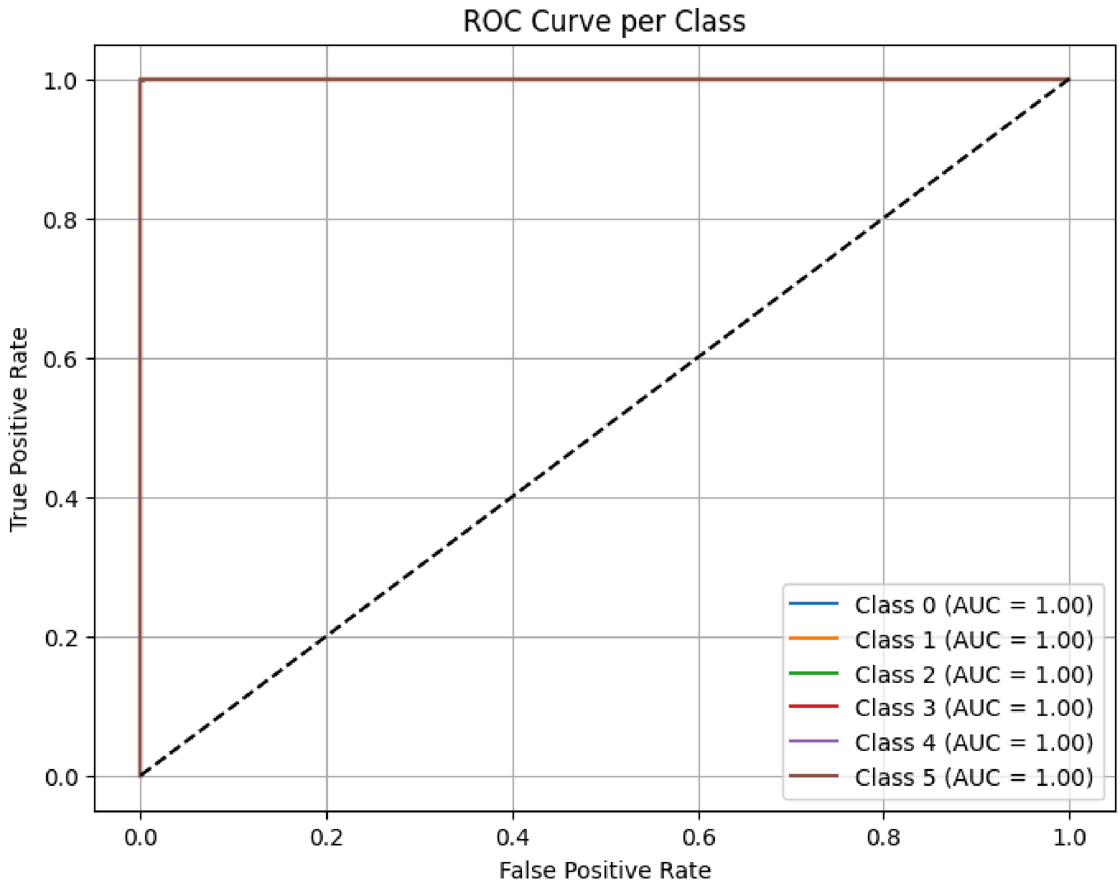
- CNN variants reached ≈ 99.9% accuracy by Epoch 5.
- Transformer models achieved ≈ 99.7–99.8% accuracy.
- BiLSTMs remained near 26–31% accuracy and were discarded.

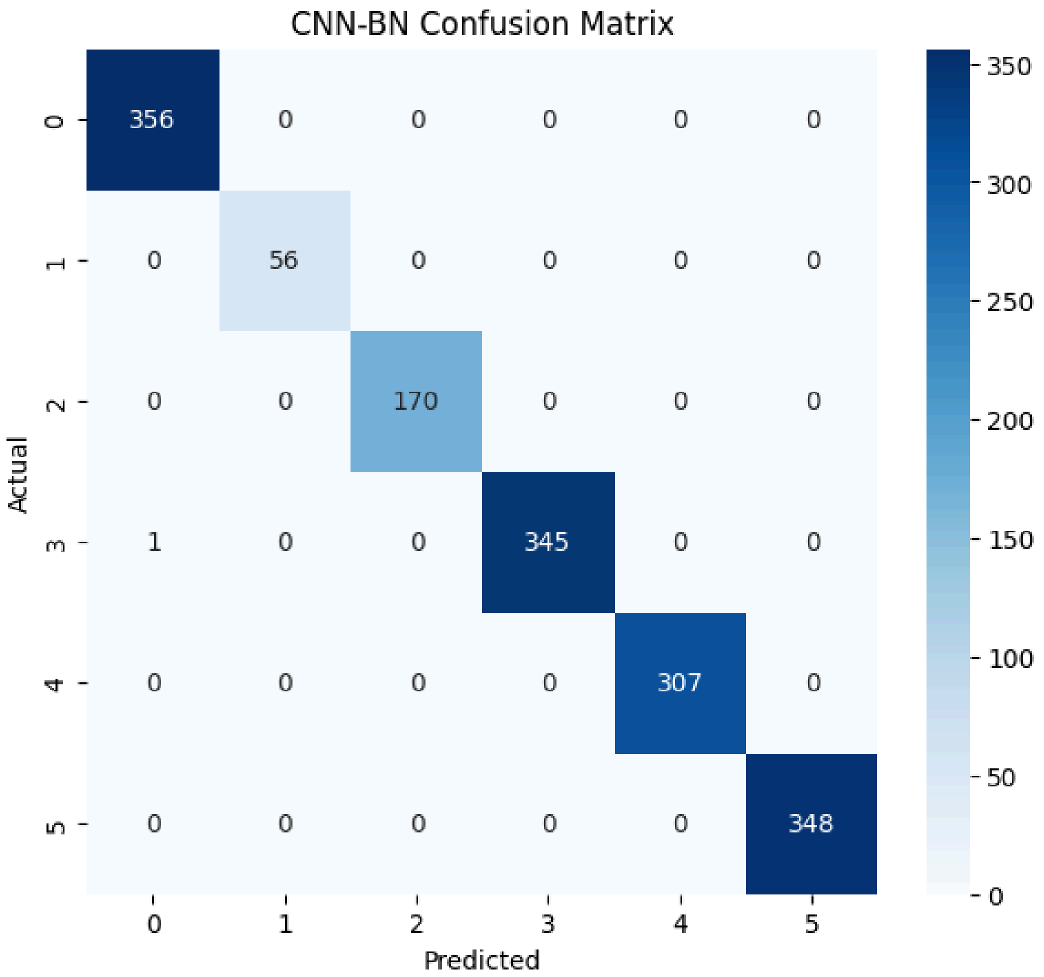
2.3 Training and Optimization

- Dataset cleaning removed rare labels; remaining six accents were upsampled to 1840 clips each to correct imbalance.
- Data split: 70% training, 15% validation, 15% test.
- Hyperparameters were tuned using validation accuracy/F1:
    - Random Forest: number of trees and depth.
    - CNN/Transformer: channels, kernel size, depth, dropout.

Final validation (1583 adult samples):

- TransformerModel_Small:
    - Validation accuracy ≈ 99.56%, macro F1 ≈ 0.996.
- CNN1D_BN:
    - Validation accuracy ≈ 99.94%, macro and weighted F1 ≈ 0.999, with a perfectly diagonal confusion matrix and per-class AUC ≈ 1.0





Validation ROC curves and confusion matrix for HuBERT + CNN-BN.

CNN1D_BN was therefore selected as the final production classifier.

3. Generalization Across Age Groups

The HuBERT CNN-BN model, trained on adult speech with clear accent cues, showed accent confusion when tested on child speech, predicting mostly Tamil (55 files), Andhra (14), Karnataka (4), and Jharkhand (2). Confidence analysis revealed:

- HuBERT CNN-BN on children:
  - Avg confidence ≈ 0.66, min ≈ 0.32, max ≈ 0.98.
- MFCC + Random Forest on the same children:
  - Avg confidence ≈ 0.36, min ≈ 0.24, max ≈ 0.58.

.This confirms HuBERT generalizes better across age but both models struggle with child speech due to higher pitch, unstable phoneme production, and incomplete accent formation, highlighting domain shift issues.

4. Word-Level vs Sentence-Level Accent Detection

Word-level experiments (1869 clips) used short vocabulary words, while sentence-level experiments (1023 clips) used longer sentences with richer phonetic context..

Performance:

- Word-level (CNN-BN):
  - Accuracy ≈ 99.63%, macro F1 ≈ 0.996, weighted F1 ≈ 0.996.
- Sentence-level (CNN-BN):
  - Accuracy 100%, macro and weighted F1 = 1.0 (perfect confusion matrix).

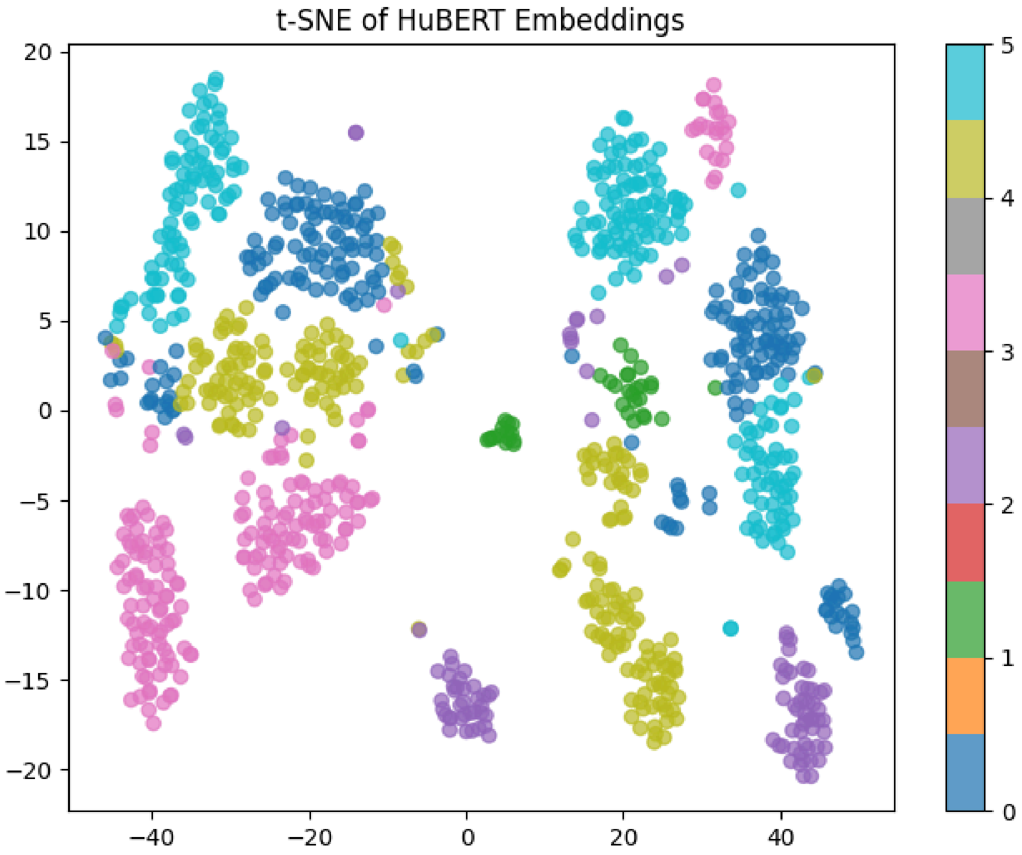
Table 1. Word- vs Sentence-Level Accent Detection (HuBERT + CNN-BN)

Comparison Criteria	**Word-Level**	**Sentence-Level**
Accuracy	99.6%	100%
Robustness	Medium – less context; slightly more errors	High – strong, stable cues across phones
Interpretability	High – easier to inspect phoneme-level errors	Medium – multiple phonemes entangled

5. HuBERT Layer-wise Analysis

Embeddings from all 13 layers (0–12) were extracted, mean-pooled, and used to train separate SVMs to identify which layer best captured accent information. Lower transformer layers (around layers 2–3) produced the highest SVM accuracies (~99%), mid-layers shifted towards lexical/semantic information, and higher layers became less sensitive to accent and more focused on speech content.

A t-SNE projection of HuBERT embeddings shows six clearly separated clusters, illustrating how the learned representation organizes accents in latent space.



Cluster structure of HuBERT embeddings for six Indian accents.

6. Results and Discussion

Experiment	Model	Feature	Accuracy / Metric	Key Observation
Baseline	Random Forest	MFCC	99.46% test accuracy	Strong baseline with handcrafted features; poorer confidence and age generalization than HuBERT.
HuBERT layer-wise	SVM (best lower layer)	HuBERT layer 2–3	≈99% accuracy	Early HuBERT layers contain strong accent cues; performance slightly below tuned CNN.
Adult validation	CNN-BN	HuBERT sequence	99.94% val accuracy, F1 ≈ 0.999	Near-perfect adult performance; diagonal confusion matrix and AUC ≈ 1.0 for all classes.
Transformer val.	Transformer Small	HuBERT sequence	99.56% val accuracy, F1 ≈ 0.996	Very strong but marginally weaker than CNN-BN.
Word-level	CNN-BN	HuBERT sequence	99.6% accuracy, F1 ≈ 0.996	High performance with limited context; good for analyzing specific word pronunciations.
Sentence-level	CNN-BN	HuBERT sequence	100% accuracy, F1 = 1.0	Most robust setting; longer context yields perfectly separable accents.
Adult → Child	CNN-BN vs RF	HuBERT vs MFCC	Confidence: 0.66 (CNN-BN), 0.36 (RF)	Both models show degraded certainty on child speech; HuBERT remains better calibrated than MFCCs.

Overall, HuBERT + CNN-BN clearly outperforms MFCC-based methods and classical SVMs, especially when evaluated on sentence-level speech, while age generalization experiments expose the need for domain adaptation before real-world deployment.

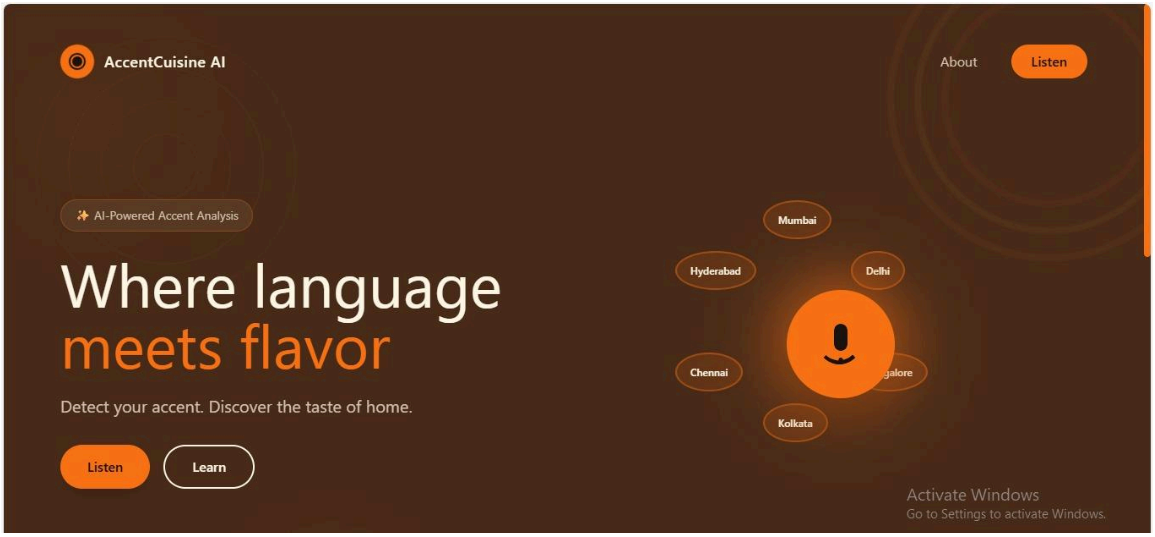
7. Application Development

The final AuralDine application integrates the HuBERT + CNN-BN classifier into a simple front-end:

1. Capture English speech from microphone or uploaded file.
2. Extract HuBERT embeddings using the Base model.
3. Run the embeddings through the trained CNN-BN model (.pt / .joblib) to predict one of six accents (e.g., tamil, andhra_pradesh, karnataka, kerala, jharkhand, gujarat).
4. Map detected accent → region → cuisine using a rule-based mapping, for example:

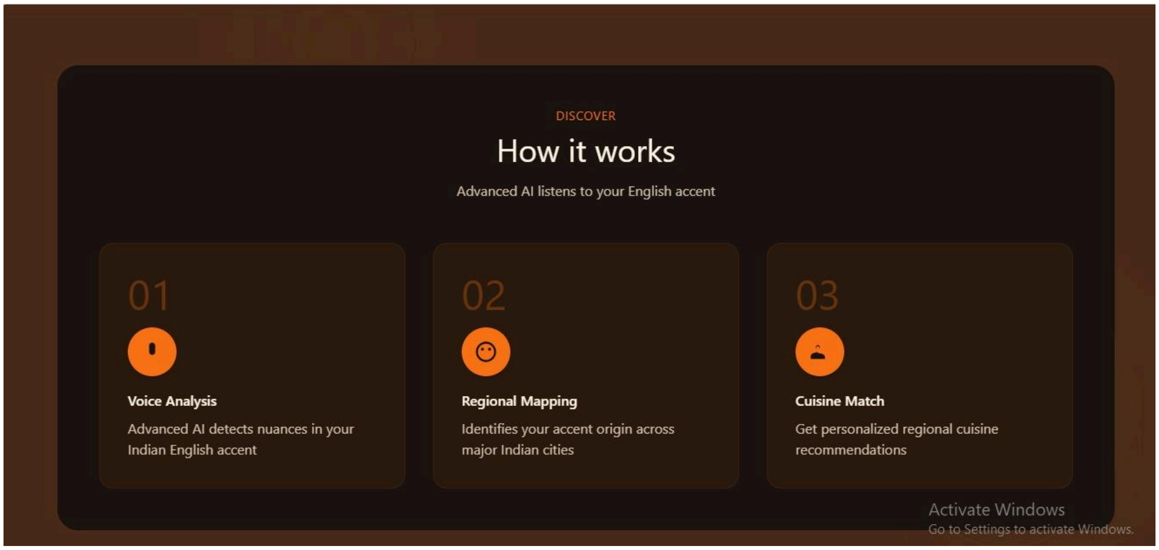
Detected Accent	Inferred Region	Recommended Dishes
Malayalam	Kerala	Appam, Puttu, Avial
Tamil	Tamil Nadu	Dosa, Pongal, Chettinad Chicken
Telugu	Andhra/Telangana	Pesarattu, Gutti Vankaya, Hyderabadi Biryani
Kannada	Karnataka	Bisi Bele Bath, Neer Dosa, Ragi Mudde
Hindi (North)	North India	Roti, Chole Bhature, Rajma Chawal

5. Display recommended dishes in a web UI (e.g., Streamlit or React) along with the predicted accent and region.



Where language meets flavor

Detect your accent. Discover the taste of home.



DISCOVER

How it works

Advanced AI listens to your English accent

01

Voice Analysis

Advanced AI detects nuances in your Indian English accent

02

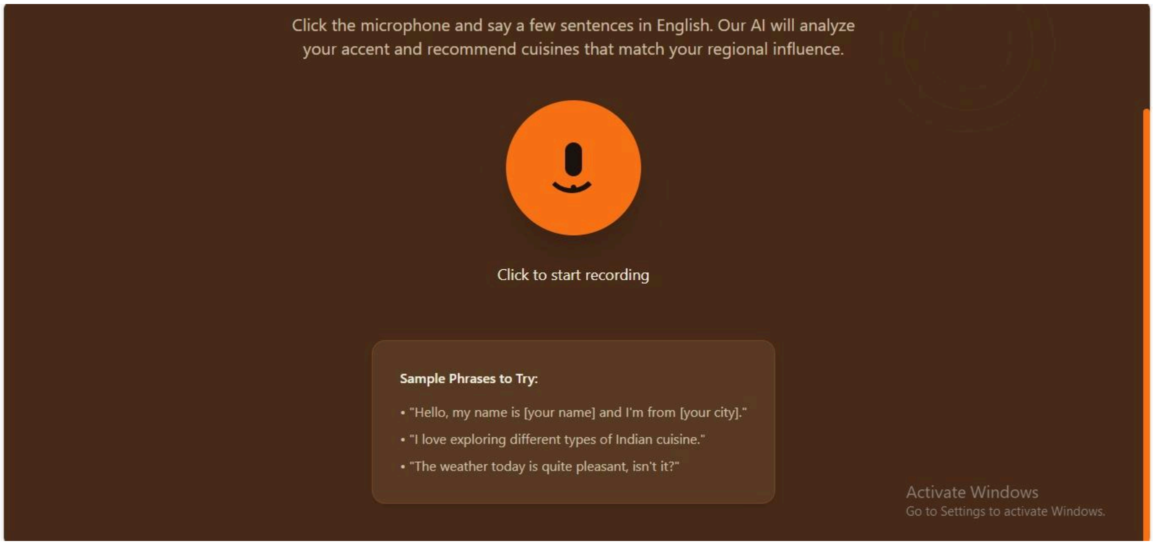
Regional Mapping

Identifies your accent origin across major Indian cities

03

Cuisine Match

Get personalized regional cuisine recommendations

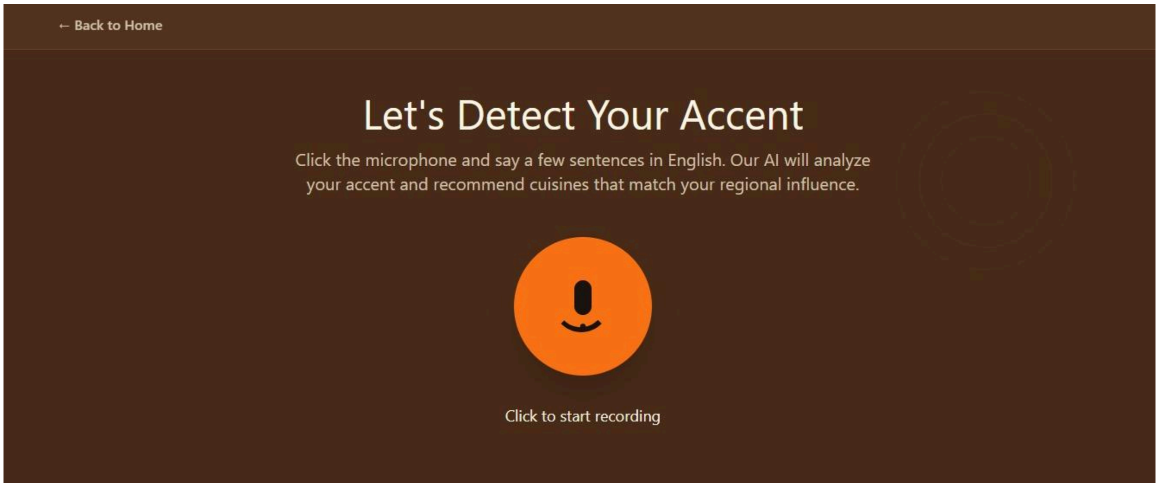


Click the microphone and say a few sentences in English. Our AI will analyze your accent and recommend cuisines that match your regional influence.

Click to start recording

Sample Phrases to Try:

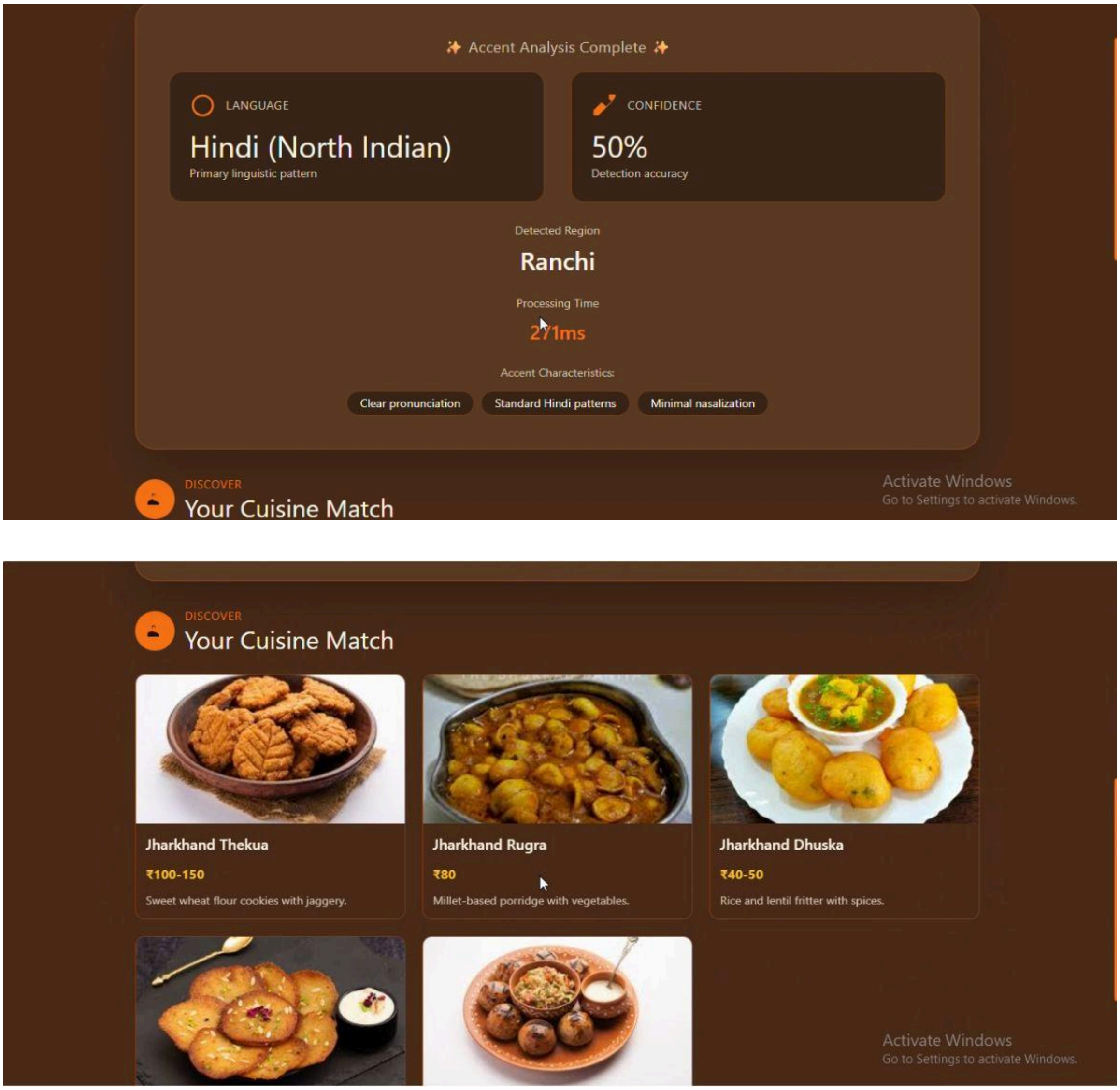
- "Hello, my name is [your name] and I'm from [your city]."
- "I love exploring different types of Indian cuisine."
- "The weather today is quite pleasant, isn't it?"



Let's Detect Your Accent

Click the microphone and say a few sentences in English. Our AI will analyze your accent and recommend cuisines that match your regional influence.

Click to start recording



✨ Accent Analysis Complete ✨

LANGUAGE

Hindi (North Indian)

Primary linguistic pattern

CONFIDENCE

50%

Detection accuracy

Detected Region

Ranchi

Processing Time

271ms

Accent Characteristics:

Clear pronunciation | Standard Hindi patterns | Minimal nasalization

DISCOVER

Your Cuisine Match

Activate Windows

Go to Settings to activate Windows.

DISCOVER

Your Cuisine Match

Jharkhand Thekua

₹100-150

Sweet wheat flour cookies with jaggery.

Jharkhand Rugra

₹80

Millet-based porridge with vegetables.

Jharkhand Dhuska

₹40-50

Rice and lentil fritter with spices.

Activate Windows

Go to Settings to activate Windows.

AuralDine prototype interface showing detected accent and recommended dishes.

8. Tools and Frameworks Used

- Python for all experiments and application logic.
- Librosa for audio I/O, resampling, and MFCC extraction.
- PyTorch and Transformers for HuBERT model loading and CNN/Transformer implementation.
- scikit-learn for SVMs, Random Forest, metrics, and evaluation utilities.
- NumPy and Pandas for data preprocessing and upsampling.
- Matplotlib and Seaborn for visualizations (class balance plots, confusion matrices, ROC, t-SNE).
- joblib / PyTorch serialization for model saving.
- Streamlit / Flask for the prototype front-end.

9. Conclusion

AuralDine demonstrates that self-supervised HuBERT embeddings combined with a tuned CNN-BN classifier can achieve near-perfect Indian accent recognition for adult English speech and power a culture-aware cuisine recommendation system.

HuBERT-based models substantially outperform MFCC baselines, especially in sentence-level settings, while additional studies on age generalization and word-vs-sentence behaviour provide deeper insight into robustness and sensitivity of different representations.

10. Future Work

- Collect more diverse and real-world speech samples
- Improve child speech generalization via domain adaptation
- Incorporate prosody features (pitch, intonation)
- Build an end-to-end neural model
- Deploy a mobile version of the AuralDine system
- Expand cuisine database to more regions

11. References / GitHub

GitHub Repository: click here