COMSYS HACKATHON-4

Track 2: Audio Identification Quest

Vajratiya Vajrobol
Pushkar Baranwal
IIC-team
Institute of Informatics and Communication, Delhi University

Contents :

- Dataset details
- Data Analysis
- Data pre-processing
- Model development
- Conclusion and Future works

DATASET DETAILS

- **DATASET SOURCES:**

AUDIO CLIPS FROM VARIOUS PERSONALITIES ON YOUTUBE
AUDIO CLIPS COLLECTED VIA CROWDSOURCING

- **AUDIO FILE FORMAT:**

ALL AUDIO FILES ARE IN .WAV FORMAT.

- **DATASET SPLITS:**

TRAINING SET: 1453 RECORDS
TESTING SET: 637 RECORDS

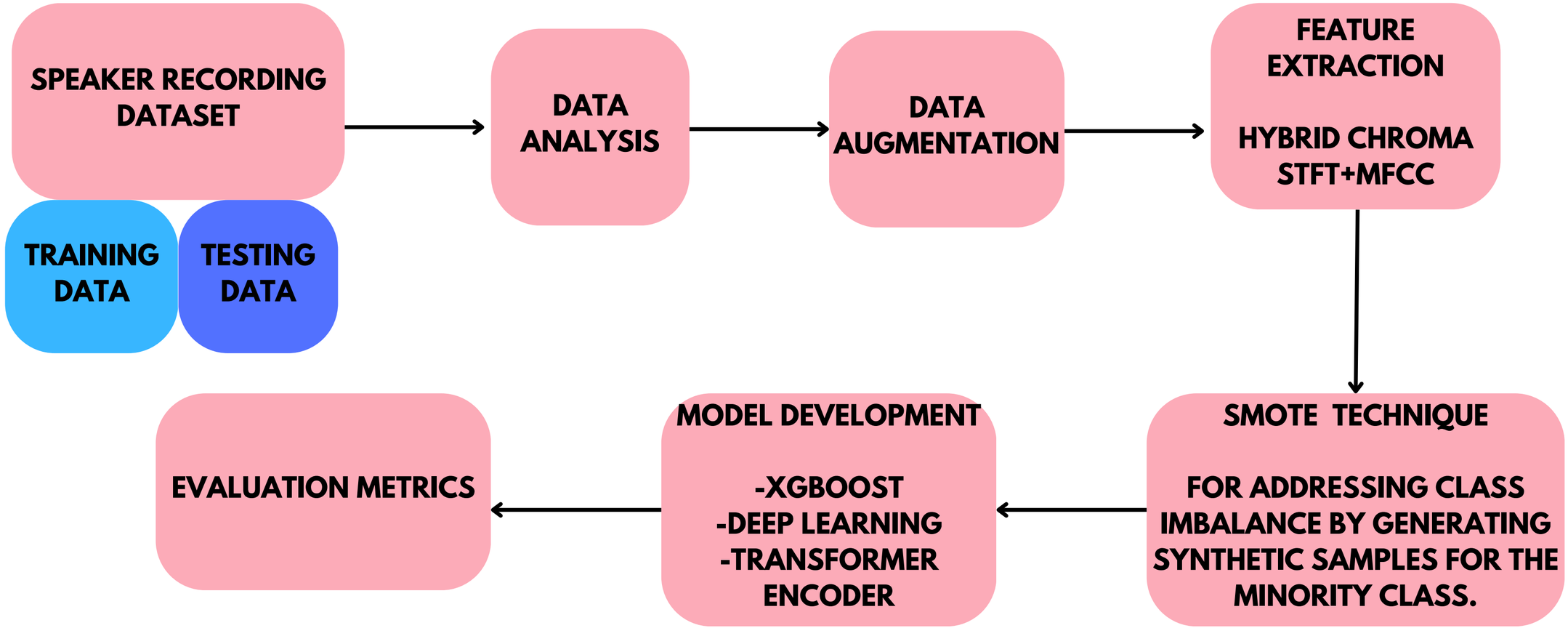
- **LABEL DEFINITIONS:**

AGE GROUP CLASSIFICATION:
LABEL 0: AGE 15 AND UNDER
LABEL 1: AGE BETWEEN 16 AND 40 (INCLUSIVE)
LABEL 2: AGE 41 AND ABOVE

- **GENDER CLASSIFICATION:**

MALE
FEMALE

THE PROCESS OF SPEAKER IDENTIFICATION



DATA ANALYSIS

- **FOCUSING ON TRAINING DATA**

```
df['age_group'].unique()
```

```
array([30, 26, 23, 19, 25, 42, 38, 41, 22, 20, 70, 27, 32, 74, 50, 65, 47,
       36, 66, 21, 55, 29, 24, 16, 43, 51, 31, 60, 37, 64, 53, 18, 59, 48,
       52, 28, 56, 35, 63, 76, 58, 39, 44, 77, 33, 68, 57, 73, 81, 75, 40,
       80, 45])
```

FIGURE 1 : THE AGE GROUP LABEL OF TRANING DATASET

THERE ARE A PROBLEM WITH THE DATA AS THERE IS NO DATA IN AGE GROUP 0 (AGE 15 AND UNDER)

age_group	
1	1113
2	340

FIGURE 2 : THE DISTRIBUTION OF AGE GROUPS

THE DATASET IS IMBALANCED REGARDING AGE GROUPS

DATA ANALYSIS

Gender	
0	1034
1	414
feamle	5

FIGURE 3 : THE DISTRIBUTION OF GENDER LABEL

-THE DATASET IS IMBALANCED REGARDING GENDER LABEL

-THE SPELLING OF LABEL IS DIFFERENT AND WRONG SPELLING.

WE HAVE ENCODED FEMALE AS 1 AND MALE AS 0

DATA AUGMENTATION TECHNIQUES

- NOISE: ADDS RANDOM NOISE TO INCREASE THE ROBUSTNESS OF THE MODEL BY SIMULATING REAL-WORLD VARIATIONS.
- STRETCH: ALTERS THE SPEED OF THE AUDIO TO HELP THE MODEL GENERALIZE TO DIFFERENT SPEAKING RATES.

- SHIFT: SHIFTS THE AUDIO TO SIMULATE DIFFERENT STARTING POINTS, IMPROVING TEMPORAL INVARIANCE.

- PITCH: MODIFIES THE PITCH TO INCREASE DIVERSITY IN TRAINING DATA, HANDLING VARIATIONS IN VOICE TONES.

WHY USE DATA AUGMENTATION: IT ENHANCES MODEL ROBUSTNESS, IMPROVES GENERALIZATION TO UNSEEN DATA, AND PREVENTS OVERFITTING BY PROVIDING MORE DIVERSE TRAINING EXAMPLES.

FEATURE EXTRACTION

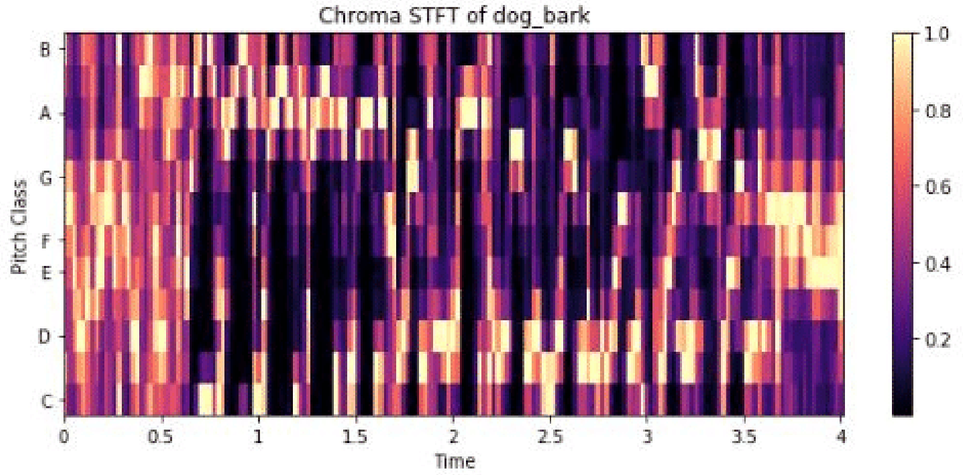


FIGURE 4. CHROMA STFT

WE USED MFCC + CHROMA STFT FOR FEATURE EXTRACTION

- **MFCCS HELP CAPTURE KEY CHARACTERISTICS OF SPEECH, MAKING THEM ESSENTIAL FOR ACCURATE SPEECH PROCESSING TASKS.**

- **CHROMA VALUE OF AN AUDIO BASICALLY REPRESENT THEINTENSITY OF THE TWELVE DISTINCTIVE PITCH CLASSES THAT ARE USEDTO STUDY MUSIC. THEY CAN BE EMPLOYED IN THE DIFFERENTIATIONOF THE PITCH CLASS PROFILES BETWEEN AUDIO SIGNALS**

COMBINING BOTH FEATURES CREATES A RICHER REPRESENTATION OF AUDIO DATA, IMPROVING CLASSIFICATION AND DETECTION TASKS ACROSS DIFFERENT AUDIO TYPES.

MODEL DEVELOPMENT AND VALIDATION SET RESULT

	precision	recall	f1-score	support
0	0.91	0.93	0.92	207
1	0.82	0.77	0.80	84
accuracy			0.89	291
macro avg	0.87	0.85	0.86	291
weighted avg	0.89	0.89	0.89	291

FIGURE 5 : XGBOOST

Validation Classification Report:

	precision	recall	f1-score	support
0	0.93	0.93	0.93	259
1	0.83	0.82	0.82	105
accuracy			0.90	364
macro avg	0.88	0.87	0.88	364
weighted avg	0.90	0.90	0.90	364

FIGURE 6 :DEEP LEARNING (MLP)

	precision	recall	f1-score	support
0	0.96	0.72	0.83	207
1	0.58	0.93	0.71	84
accuracy			0.78	291
macro avg	0.77	0.83	0.77	291
weighted avg	0.85	0.78	0.79	291

FIGURE 7 : TRANSFORMER ENCODER

IN THE TRAINING PERIOD, THE F1-SCORE MACRO AVERAGE REVEALED THAT MLP OUTPERFORMED XGBOOST (0.88 VS. 0.86), WHILE THE TRANSFORMER ENCODER LAGGED AT 0.77. HOWEVER, UPON EVALUATING THE TEST SET, XGBOOST DEMONSTRATED SUPERIOR PERFORMANCE COMPARED TO THE OTHER MODELS. THIS DISCREPANCY SUGGESTS THAT XGBOOST'S ROBUSTNESS AND GENERALIZATION CAPABILITIES ALLOWED IT TO PERFORM BETTER ON UNSEEN DATA, LIKELY DUE TO ITS EFFECTIVE HANDLING OF FEATURE INTERACTIONS AND ABILITY TO MITIGATE OVERFITTING. THUS, WHILE MLP SHOWED STRONG TRAINING PERFORMANCE, XGBOOST PROVED TO BE MORE RELIABLE IN REAL-WORLD APPLICATIONS.

TESTING RESULTS

THE BEST SCORE FROM CHROMA STFT +MFCC WITH XGBOOST MODEL BASED
-PUBLIC SCORE 0.5173
-PRIVATE SCORE 0.4725

CONCLUSION AND FUTURE WORKS

IN THIS STUDY, WE EXPLORED THE TASK OF SPEAKER IDENTIFICATION USING A DATASET THAT EXHIBITED CLASS IMBALANCE AMONG DIFFERENT SPEAKERS. BY EMPLOYING VARIOUS AUDIO FEATURE EXTRACTION TECHNIQUES SUCH AS MEL-FREQUENCY CEPSTRAL COEFFICIENTS (MFCC) AND CHROMA SHORT-TIME FOURIER TRANSFORM (STFT), WE DEVELOPED A ROBUST MODEL CAPABLE OF RECOGNIZING SPEAKERS BASED ON THEIR UNIQUE VOCAL CHARACTERISTICS. DESPITE THE CHALLENGES POSED BY THE IMBALANCED DATA DISTRIBUTION, OUR APPROACH DEMONSTRATED GOOD RESULTS IN IDENTIFYING SPEAKERS ACCURATELY, HIGHLIGHTING THE IMPORTANCE OF SELECTING APPROPRIATE FEATURES AND IMPLEMENTING EFFECTIVE DATA AUGMENTATION STRATEGIES TO MITIGATE THE EFFECTS OF IMBALANCE.

REFERENCES

HTTPS://WWW.RESEARCHGATE.NET/PUBLICATION/346659500_URBAN_SOUND_CLASSIFICATION_USING_CONVOLUTIONAL_NEURAL_NETWORK_AND_LONG_SHORT_TERM_MEMORY_BASED_ON_MULTIPLE_FEATURES/FIGURES?LO=1

HTTPS://WWW.GOOGLE.COM/SEARCH?Q=MFCC+GEEKFORGEEK&RLZ=1C1CHBF_ENIN979IN979&OQ=MFCC+GEEKFORGEEK&GS_LCRP=EGZJAHJVBWUYBGGAEEUYOTIJCAEQIRGKGKAB0GEINDU5OWOWAJEOAGIWAGE&SOURCEID=CHROME&IE=UTF-8