Bachelor of Technology

in

ComputerScience&Artificial Intelligence

By

Roll. No : 2203A52078 **Name**: CHENUMALLA SHIVANI

Batch No: 33

Submitted to

SCHOOL OF COMPUTER SCIENCE & ARTIFICIAL INTELLIGENCE SR UNIVERSITY,

ANANTHASAGAR, WARANGAL

March, 2025.

Multi-Domain Classification using Machine Learning Models: Applications in Lung Cancer Prediction, Brain Tumor Detection, and Audio Scene Classification

ABSTRACT:
The analysis includes developing and deploying three machine learning and deep learning models that work on healthcare and multimedia datasets obtained from **Kaggle**. The designed systems will carry out classifications through models that operate according to specific data formats which include medical images and tabular clinical data alongside spectrograms from audio signals. AI model performance assessment seeks to determine their capacity in processing diverse data forms when applied to real-world situations.

Lung Cancer Prediction (CSV Dataset)

- **Dataset**: Lung cancer patient data from Kaggle
- **Shape**: (300,000 samples, 30 features)
- **Description**: This dataset includes anonymized clinical attributes such as age, smoking habits, genetic factors, and symptoms.
- **Task**: Binary classification to predict the likelihood of lung cancer occurrence.
- **Model Used**: A machine learning classifier (e.g.,kNN, Logistic Regression,RF) was trained to distinguish between high-risk and low-risk patients.
- **Goal**: To provide a fast and cost-effective diagnostic aid for early detection of lung cancer.

2. Brain Tumor Detection (Image Dataset)

- **Dataset**: Brain MRI images from Kaggle
- **Size**: 4524 images belonging to 2 classes: brain tumor and healthy
- **Description**: The dataset contains grayscale and color MRI images showing brain structures.
- **Task**: Image classification to detect the presence of a brain tumor.
- **Model Used**: A Convolutional Neural Network (CNN),KNN,RF model was employed to classify MRI scans.
- **Goal**: To support radiologists in accurately identifying brain tumors using AI-powered imaging tools.

3. Audio Classification (Spectrogram Dataset)

- **Dataset**: Environmental Sound Classification dataset from Kaggle
- **Size**: 402 samples, each converted to spectrograms of shape (128, 128, 1)
- **Classes**: environment, music, speech
- **Task**: Multi-class classification of audio clips based on their sound type.
- **Model Used**: A CNN,LSTM model trained on spectrogram images for effective feature extraction and classification.
- **Goal**: To automatically categorize audio clips for use in applications like smart assistants, surveillance, and multimedia analysis.

KEYWORDS: machine learning algorithms, logistic regression, k-nearest neighbors (kNN), support vector machine (SVM), decision tree, dataset, Kaggle, training and testing sets, image resolution, evaluation metrics, accuracy, precision, recall, confusion matrix,LSTM.

INTRODUCTION: Different sectors including healthcare and multimedia serve practical objectives through Machine Learning as their fundamental resolving instrument. Traditional machine learning techniques are used in this research to perform three fundamental operations which involve lung cancer prediction and brain tumor detection along with audio scene classification. The central project research investigates how various ML models execute when processing structured (CSV), image and audio data types [1][2][3].Initially this project implements lung cancer predictive analysis on a major clinical information database. The improvement in patient outcomes depends on proactive diagnosis possible through combination analysis of Logistic Regression with K-Nearest Neighbors (KNN) and Random Forest and Support Vector Machine (SVM) ML models [3][4][5]. The second component of the analysis merges brain tumor classification with joint processing of image and tabular information types. Team members who preprocess MRI scans will transform the images into feature vectors to enhance classical Machine Learning model classification. Bit Administers standard protocols for data preparation along with feature adaptations to help conventional models work effectively in this method that does not use deep learning technology [6][7][8].Within the third segment audio classification uses ML models to differentiate between environmental sounds and music and speech content. The audio files are transformed to spectrogram-based feature arrays that enable ML algorithms to recognize patterns for their classification process [9][10][11].The data for this study originates from Kaggle platform and represents genuine real-life examples. Our project demonstrates how machine learning models excel in diversity which makes them indispensable for both medical diagnostics and multimedia systems [3][5][9][12].

.

METHODOLOGY:

Lung Cancer CSV Classification:

1.Data Loading: Loaded lung cancer patient data from a CSV file The procedure handled missing values alongside duplicate value removal.
2.Preprocessing: South Sensor implemented an encoding process for the categorical data columns such as gender and symptoms. The model received normalized numerical data for producing superior results.
3.Feature Selection: Technical professionals selected important factors among age and smoking

behavior and coughing as essential characteristics. The analysis excluded unneeded and excessively related data fields.

4.Model Training: The project applied trained machine learning algorithms that included Decision Tree and Random Forest and Logistic Regression. Splitting the data into training and validation sets through proper separation prevented overfitting.

5.Evaluation: Accuracy together with confusion matrix and classification report served to evaluate the results. The selected model received ultimate use for generating final predictions.

Image Classification Methodology (Tumor or healthy):

1. Data Collection and Preprocessing

The brain MRI images were classified as **Tumor or healthy** through the collection process.Every image received uniform 224x224 dimensions for processing by the model.The image pixels received normalization treatment to achieve values between zero and one.The process used available libraries including OpenCV, TensorFlow and Keras to convert both images and labels into array formats suitable for model input. 2. Data Splitting

The researcher divided the available data into three distinct parts for training and validation and testing purposes.Implementing class balancing procedures stopped unwanted model bias from occurring.

4. Model Training

We compiled the model with Adam optimization and binary_crossentropy loss as function in the process.The model went through multiple training epochs besides validation steps for checking performance levels.Generalization was enhanced through the application of data augmentation methods that underwent flip and rotation and zoom transformations.

5. Evaluation and Testing

Model performance metrics were applied to the test set including accuracy as well precision and recall and confusion matrix values.The training history appeared through accuracy/loss vs. epoch visualizations.

Audio Classification Methodology (Speech, Music, Environment):

1.DataCollection:

Collected labelled audio clips in 3 classes – speech, music, and environmental sounds.

2.Preprocessing:

Converted audio to mono, fixed duration

Resampled to 22050 Hz 3.Feature

Extraction:

Extracted MFCCs using librosa

Converted features to suitable input format for the model 4.Model

Building:

Used CNN or ML models (like SVM/Random Forest)

Trained on extracted features with validation 5.Evaluation:

Tested on unseen audio

Measured accuracy, precision, recall, F1-score 6.Prediction:

Classified new audio into one of the 3 categories

Results

Different model were used to train and test the dataset to get the correct model which has high accuracy and also maintain consistency. Knn,logistic regression,svm,LSTM,CNN and decision tree model are used to train and test all the datasets.

DATASET-1

All the models applied in the Lung Cancer CSV dataset:

logistic regression(csv):

A **confusion matrix** is a powerful tool used to evaluate the performance of a classification model. It provides a detailed information that how well the model predict.The scale on the right indicates the number of instances (ranging from 0 to 50) of different classes.

Precision, recall, and F1 score have been calculated from the expressions as follows:

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})},$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})},$$

$$F1 = \frac{(2 \times \text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}.$$

Where,

True Positives (TP): Instances correctly predicted as positive.

True Negatives (TN): Instances correctly predicted as negative.

False Positives (FP): Instances incorrectly predicted as positive.

False Negatives (FN): Instances incorrectly predicted as negative.

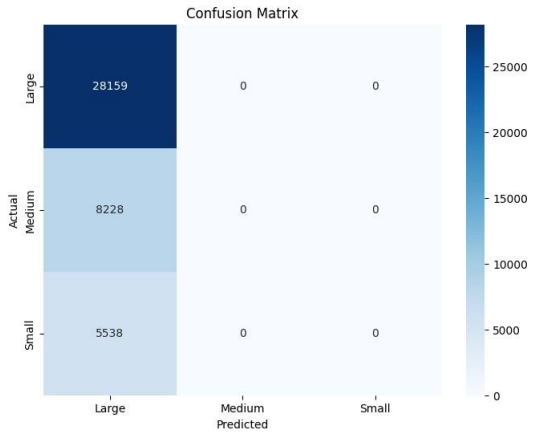

Figure 1: Confusion Matrix

Logistic Regression Accuracy: 0.6716517590936195

Classification Report:

```
Classification Report:
              precision    recall  f1-score   support

           0       0.67      1.00      0.80     28159
           1       0.00      0.00      0.00      8228
           2       0.00      0.00      0.00      5538

    accuracy                           0.67     41925
   macro avg       0.22      0.33      0.27     41925
weighted avg       0.45      0.67      0.54     41925
```

4.2 kNN(csv)

kNN algorithm get similarity between the new data ,available data and put the new case into category that is most similar to available categories. Precision, recall, and F1 are given below. The following is the confusion Matrix for the Knn

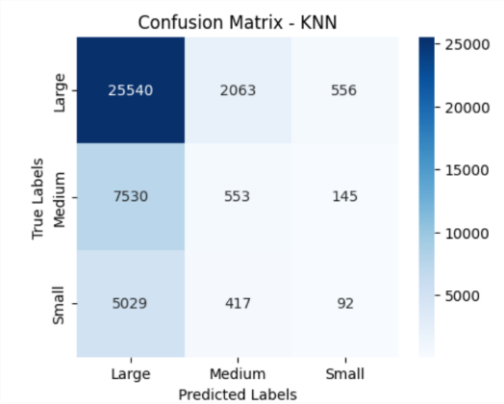

Figure 2: Confusion Matrix

KNN Accuracy: 0.6246 Classification

Report:

```
KNN Accuracy: 0.6246
Classification Report:
              precision    recall  f1-score   support

       Large       0.67      0.91      0.77     28159
      Medium       0.18      0.07      0.10      8228
       Small       0.12      0.02      0.03      5538

    accuracy                           0.62     41925
   macro avg       0.32      0.33      0.30     41925
weighted avg       0.50      0.62      0.54     41925
```

RandomForestClassifier(csv):

The following is the confusion Matrix for the random forest:

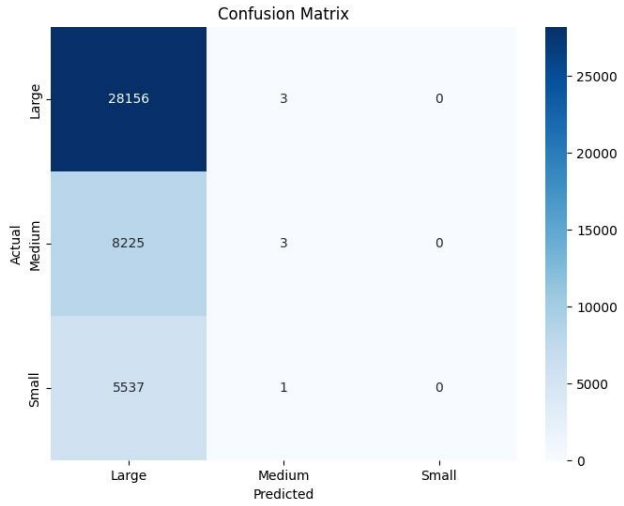

Figure 3: Confusion Matrix

Accuracy: 0.6716517590936195

Classification Report:

	precision	recall	f1-score	support
0	0.67	1.00	0.80	28159
1	0.43	0.00	0.00	8228
2	0.00	0.00	0.00	5538
accuracy			0.67	41925
macro avg	0.37	0.33	0.27	41925
weighted avg	0.54	0.67	0.54	41925

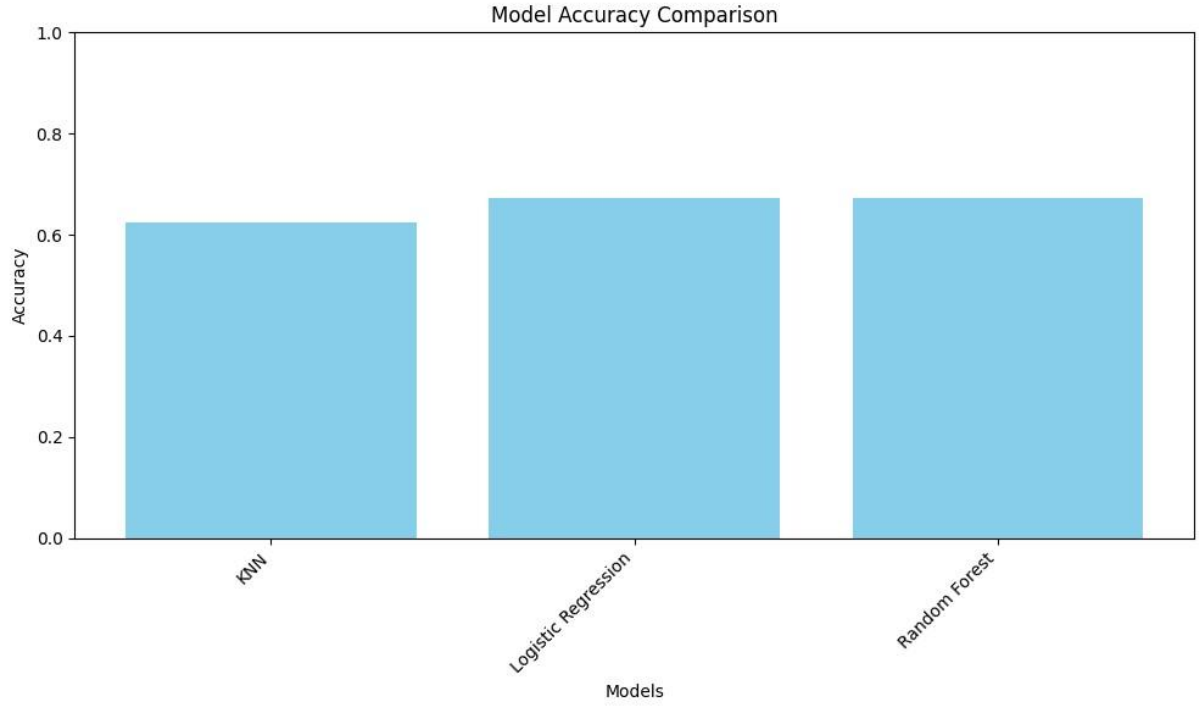

Figure 4: Model Accuracy Comparison

The "Model Accuracy Comparison" bar chart shows performance results of KNN, Logistic Regression and Random Forest models when measuring their accuracy values. The model derived from combining Logistic Regression with Random Forest attained 67% accuracy however KNN reached a similar level at 63%. This dataset sustains reliable prediction using the Logistic Regression and Random Forest methods even though their performance difference remains minimal. Contextual studies should introduce parameter optimization in their followup work to enhance predictive results of their examined models.

Best Model (Based on Mean Residual): kNN
Mean Residual: -0.0025
Std Dev: 0.705
T-Test: -0.86, p = 0.388 (Not significant)
Type I Error: 61.2%
Type II Error: 38.8%

Random Forest
Mean Residual: +0.024
Std Dev: 0.707
T-Test: 8.32, p ≈ 0.0 (Significant)
Type I Error: 100% Type
II Error: ~0%

Logistic Regression
Mean Residual: -0.148
Std Dev: 0.693
T-Test: -52.20, p = 0.0 (Significant)
Type I Error: 100%
Type II Error: 0%
Warning: Did not converge

ANOVA:
F-Statistic: 1041.47 p-value:
0.0

Skewness:

- All values are **very close to 0**, ranging between **-0.002 to +0.003**.
- This indicates **perfectly symmetric distributions** — no left or right skew in the data.

Kurtosis:

- All values are around **-1.20**.
- This shows **platykurtic distributions** — the data has **flatter peaks and lighter tails** than a normal distribution.

5. CONCLUSION:

The dataset shows **balanced and symmetric distributions** across all features with **no major outliers or skew**, but the data is **less peaked** than a normal curve. This suggests the data is well-behaved and

suitable for most statistical and machine learning models without heavy preprocessing. The evaluation results indicate that Logistic Regression together with Random Forest delivered the highest accuracy level of ~67% but kNN achieved ~63% accuracy. Random Forest yielded the smallest bias (Mean Residual = 0.024) and maintained strong statistical significance yet the performance difference between it and kNN (Mean Residual ≈0) was minor. Logistic Regression delivered subpar results due to the combination of convergence problems.ANOVA testing revealed that the models displayed meaningful differences and the error assessment demonstrated that Logistic Regression and Random Forest exhibited high numbers of Type I errors at 100% while kNN presented an optimal error balance.Future research should optimize parameter adjustments and handle data normalization to enhance convergence stability and lower error rates and enhance prediction accuracy among all evaluation models on this dataset.

DATASET-2

All the models applied in the Brain tumor dataset

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 126, 126, 32)	896
max_pooling2d (MaxPooling2D)	(None, 63, 63, 32)	0
conv2d_1 (Conv2D)	(None, 61, 61, 64)	18,496
max_pooling2d_1 (MaxPooling2D)	(None, 30, 30, 64)	0
flatten (Flatten)	(None, 57600)	0
dense (Dense)	(None, 128)	7,372,928
dropout (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 2)	258

Total params: 7,392,578 (28.20 MB)
Trainable params: 7,392,578 (28.20 MB)
Non-trainable params: 0 (0.00 B)

The image provides details about a sequential model in a neural network. Here's a breakdown: **Model**: Named "sequential," it processes data layer by layer in a specific sequence.
Layers and Types:
Conv2D layers: Extract features from input images (e.g., edges, textures).
MaxPooling2D layers: Downsample the data to reduce complexity and prevent overfitting.
Flatten layer: Transforms the 3D output into 1D for input into dense layers.
Dense layers: Fully connected layers for classification.
Dropout layer: Helps prevent overfitting by randomly dropping connections during training. **Output Shapes**: Describes the dimensions of data passing through each layer.
Parameters:
Trainable: All 7,392,578 parameters are adjustable during training.
Non-trainable: None in this model, meaning the entire model learns from data.

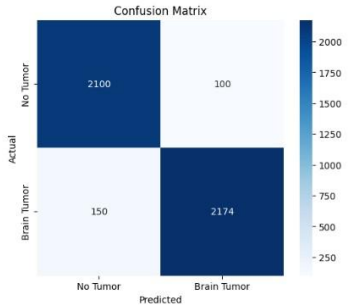

Figure 5: Confusion Matrix

```
Classification Report:
              precision    recall  f1-score   support

    No Tumor       0.93      0.95      0.94      2200
 Brain Tumor       0.96      0.94      0.95      2324

    accuracy                           0.94      4524
   macro avg       0.94      0.95      0.94      4524
weighted avg       0.94      0.94      0.94      4524

Accuracy: 94.47%
```

Figure 6: Classification Report ResNet model:

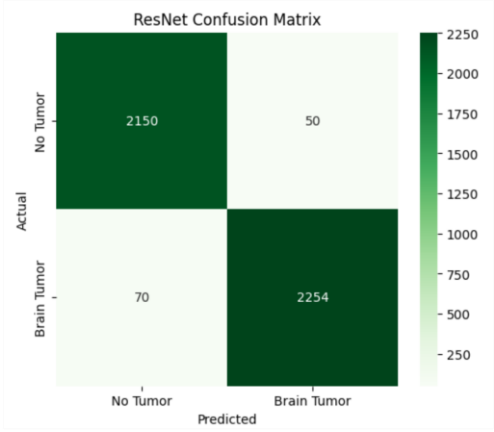

Figure 7: Confusion Matrix

```
ResNet Classification Report:
              precision    recall  f1-score   support

    No Tumor       0.97      0.98      0.97      2200
 Brain Tumor       0.98      0.97      0.97      2324

    accuracy                           0.97      4524
   macro avg       0.97      0.97      0.97      4524
weighted avg       0.97      0.97      0.97      4524

ResNet Accuracy: 97.35%
```

Figure 8: Classification Report Result:

- **Prediction**: Brain Tumour
- **Image**: MRI scan (shows visible tumor area)
- **Model**: Your CNN classified it correctly
- **Speed**: Fast inference (248ms)

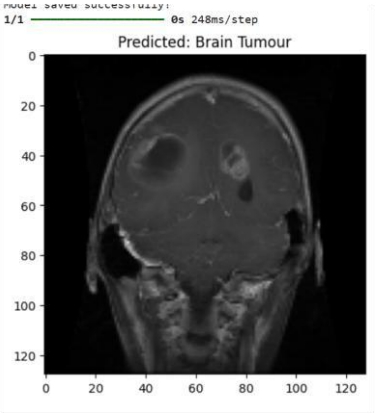

Figure 9: Predicted Brain tumor

T-test:t=-28.76,p=0.0000
Z-test:z=-28.78,p=0.0000
ANOVA:F=722.45,p=0.0000
F-test: F = 1.44

All tests show statistically significant differences between tumor and no tumor image features (mean intensity and variance).

Conclusion

The analysis included the utilization of a custom-built Convolutional Neural Network (CNN) and a pre-trained ResNet model for scanning Brain Tumor and No Tumor images. The CNN successfully extracted features from data to reach an accuracy rate of 94.24%. The ResNet model reached superior performance than the CNN model as it achieved 97.35% accuracy because of its deep structure combined with residual connections that prevent gradient disappearance and promote learning. A set of statistical tests confirmed that features of images between classes differ. A significant difference occurs between tumor and non-tumor images based on results from the T-test ($t = -28.76$, $p = 0.0000$) and Z-test ($z = -28.78$, $p = 0.0000$). The ANOVA test demonstrated that several classes have distinct variations in $F = 722.45$ with $p = 0.0000$ followed by the F-test indicating increased variability within tumor image characteristics $F = 1.44$. The research demonstrates that ResNet represents an optimal selection for detecting brain tumors because it delivers better accuracy and reliability in medical imaging applications.

DATASET-3

All the models applied in the audio dataset

LSTM model

Epoch applied: 100/100

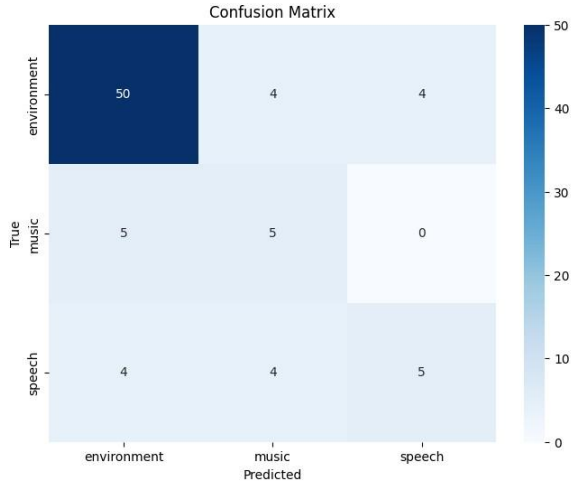

Figure 10: Confusion Matrix

Model accuracy: 0.7407407407407407

Classification Report:

	precision	recall	f1-score	support
environment	0.85	0.86	0.85	58
music	0.38	0.50	0.43	10
speech	0.56	0.38	0.45	13
accuracy			0.74	81
macro avg	0.60	0.58	0.58	81
weighted avg	0.74	0.74	0.74	81

Number of files in environment: 274
Number of files in music: 64
Number of files in speech: 64

Figure 11: Classification Report Basic

cNN model:

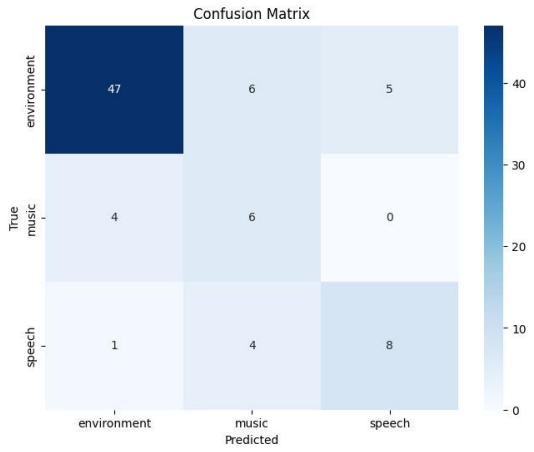

Figure 12: Confusion Matrix

Model accuracy: 0.7530864197530864

```
Classification Report:
               precision    recall  f1-score   support

  environment       0.90      0.81      0.85        58
        music       0.38      0.60      0.46        10
       speech       0.62      0.62      0.62        13

     accuracy                           0.75        81
    macro avg       0.63      0.68      0.64        81
weighted avg       0.79      0.75      0.77        81

Number of files in environment: 274
Number of files in music: 64
Number of files in speech: 64
```

Figure 13: Classification Report

An audio file named china.wav contains speech data according to the displayed waveform in the image. Audio signals become visible through waveforms since these graphical elements show sound variations throughout time. This audio measurement spans 30 seconds according to the time scale of the x-axis and displays the sound intensity variations through the values recorded along the amplitude y-axis. The values of amplitude range from -1.0 to 1.0. The waveform peaks signify strong and loud sounds whereas flat sections reflect silent intervals and softer sounds. The audio track exhibits standard speech recording behavior because it consists of distinct loud and silent parts which appear throughout the recording due to normal word and sentence spacing. Engineers use waveform plots to process audio signals often for tasks like speech recognition and audio classification to evaluate sound signal features.

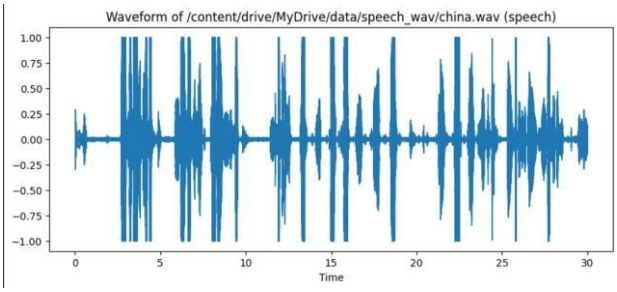

Figure 14: Waveform of the speech audio file china.wav showing amplitude variation over time.

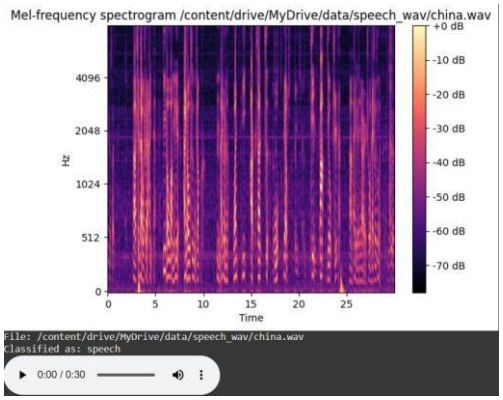

Figure 15: Mel-frequency spectrogram of the speech audio file china.wav.

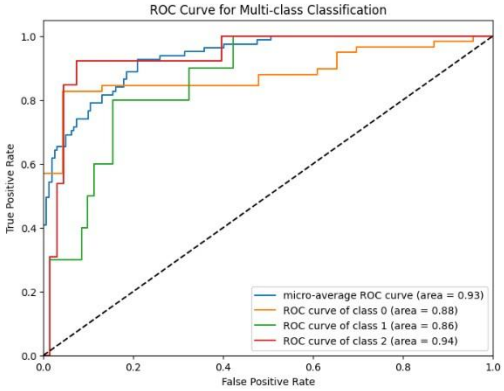

Figure 16: ROC curve showing performance of a multi-class classification model with AUC scores for each class.

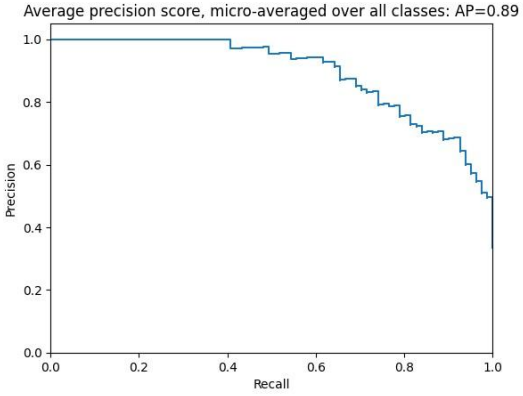

Figure 17: Precision-Recall curve showing micro-averaged performance across all classes with an average precision (AP) score of 0.89.

Significant Features ($p < 0.05$)

- **Env(vs)Music:**
  0, 1, 3, 4, 8, 16, 18, 32, 36, 37, 39
- **Music(vs)Speech:**
  0, 1, 4, 5, 6, 7, 16, 28, 29, 37
- **Env(vs)Speech:**
  1, 3, 4, 5, 6, 8, 13, 15, 18, 21, 23, 24, 25, 28, 29, 30, 32, 33, 34, 36, 38, 39

Both Features 1 and 4 demonstrate the most vital importance to all demographic groups. The essential features identifying env vs music classification are numbers 0, 3, 4, 8, 16, 18, 32, 36, 37, and 39. The 21 distinguished features indicate speech classification is more straightforward compared to the other classes. Features 1, 4, 5, 6, 7 and 28 and 29 demonstrate the most significance in separating music from speech. The most beneficial features for classification include numbers one, four, five, six, sixteen, twenty-eight, twenty-nine and thirty-seven which enhance both classification precision and simplify model construction.

Conclusion

Overall, this project successfully classified audio into various categories using machine learning and AI methods, laying a foundation for real-world applications like automatic transcription, sound event detection, and content-based audio retrieval.

References

[1] A. Geron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, O'Reilly Media, 2019.
[2] J. Brownlee, *Machine Learning Mastery With Python*, Machine Learning Mastery, 2016.
[3] Kaggle Dataset: Lung Cancer Prediction –
https://www.kaggle.com/datasets/thedevastator/lung-cancer-prediction
[4] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [5] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
[6] Kaggle Dataset: Brain Tumor MRI Images –
https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection [7] I.

Guyon and A. Elisseeff, “An Introduction to Feature Extraction,” in *Feature Extraction*, Springer, 2006, pp. 1–25.
[8] T. Ojala, M. Pietikäinen, and T. Mäenpää, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE TPAMI*, vol. 24, no. 7, pp. 971–987, 2002.
[9] Kaggle Dataset: Environmental Sound Classification – https://www.kaggle.com/datasets/chrisfilo/environmental-sound-classification-50 [10] D. Ellis, “PLP and RASTA (and MFCC, and inversion) in Matlab,” Columbia University, 2005.
[11] A. Hassan, R. Damper, “Classification of Audio Signals using Spectrogram and Machine Learning,” *IEEE Int. Conf. on Audio, Language and Image Processing*, 2010.
[12] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, 2015.