Music Genre Classification With Deep Learning

SCIE30001 student presentation

Fiona Zhang

TABLE OF CONTENTS

PLAYLIST

Playlist (35min)

1. Abstract & Problem Statement
2. Dataset & Data Augmentation
3. Input Representations
4. Baseline Models
5. Proposed Model
6. Results & Conclusion
7. Extension: deployment

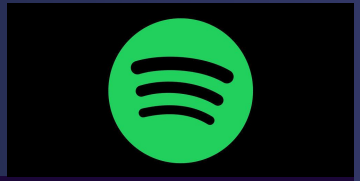
Fiona's Presentation

2:54

3:49

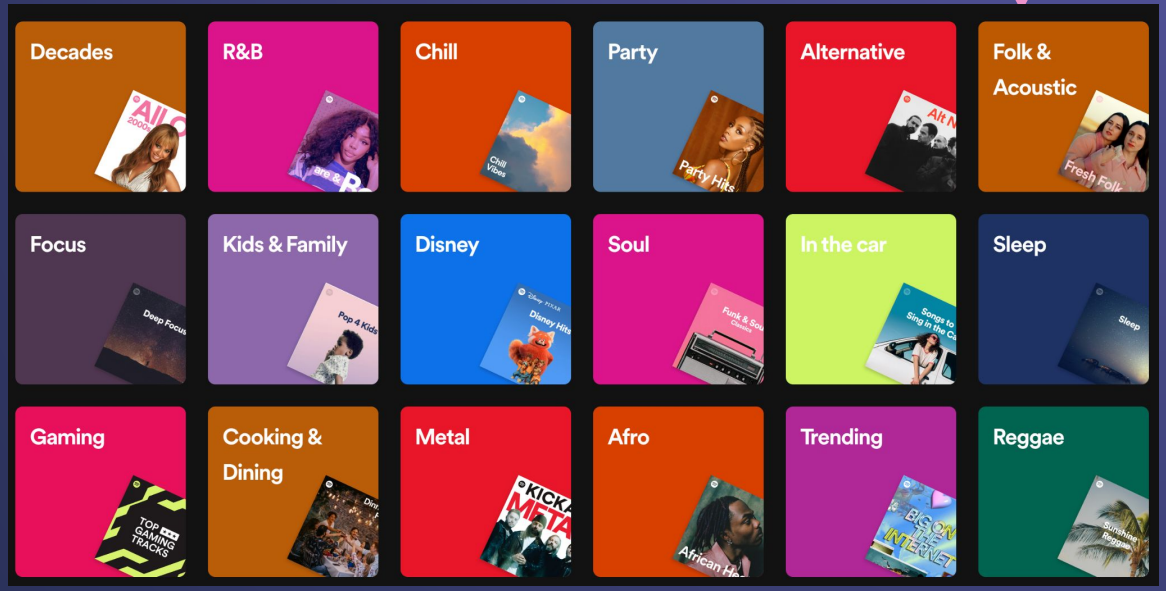
01

Abstract & Problem Statement



```
{
    "genres": ["acoustic", "afrobeat", "alt-rock", "alternative", "ambient", "anime", "black-metal", "bluegrass", "blues", "bossanova", "brazil", "breakbeat", "british", "cantopop", "chicago-house", "children", "chill", "classical", "club", "comedy", "country", "dance", "dancehall", "death-metal", "deep-house", "detroit-techno", "disco", "disney", "drum-and-bass", "dub", "dubstep", "edm", "electro", "electronic", "emo", "folk", "forro", "french", "funk", "garage", "german", "gospel", "goth", "grindcore", "groove", "grunge", "guitar", "happy", "hard-rock", "hardcore", "hardstyle", "heavy-metal", "hip-hop", "holidays", "honky-tonk", "house", "idm", "indian", "indie", "indie-pop", "industrial", "iranian", "j-dance", "j-idol", "j-pop", "j-rock", "jazz", "k-pop", "kids", "latin", "latino", "malay", "mandopop", "metal", "metal-misc", "metalcore", "minimal-techno", "movies", "mpb", "new-age", "new-release", "opera", "pagode", "party", "philippines-opm", "piano", "pop", "pop-film", "post-dubstep", "power-pop", "progressive-house", "psych-rock", "punk", "punk-rock", "r-n-b", "rainy-day", "reggae", "reggaeton", "road-trip", "rock", "rock-n-roll", "rockabilly", "romance", "sad", "salsa", "samba", "sertanejo", "show-tunes", "singer-songwriter", "ska", "sleep", "songwriter", "soul", "soundtracks", "spanish", "study", "summer", "swedish", "synth-pop", "tango", "techno", "trance", "trip-hop", "turkish", "work-out", "world-music"]
}
```

featured genres



These are all **editorial playlists**

R&B around the world

Show all

Are & Be

The pulse of R&B music today. Cover: Janelle...

Mood Ring

This is R&B in Canada, elevate your aura with...

R&B UK

100% British - 100% R&B. Cover: No...

Ginger Me

A collection R&B and Afropop cuts from We...

TrenChill K-R&B

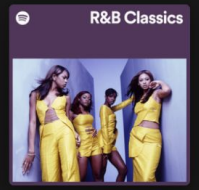
Enjoy trendy & chill Korean R&B music....

Latin R&B

The home of Latin R&B. La casa del R&B Latin...

Popular R&B Playlists

Show all



R&B Classics

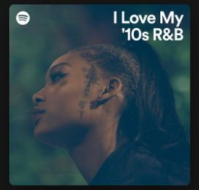
Throw it back with these classic R&B jam...

Hip Hop + R&B FM

Hip Hop and R&B to soundtrack your life.

I Love My '00s R&B

The most essential R&B songs from the 2000s...



I Love My '10s R&B

The most essential R&B songs from the 2010s....

I Love My '90s R&B

Celebrating 25 years of "Never Say Never."...

Energy Booster: R...

Need to get your energy level up?

Why Automatic Music Genre Classification (AMGC)

(*Music Business Worldwide*)

82+ Million songs

Total songs on Spotify

60,000+

New songs everyday on Spotify

Automatic procedure to deal with large amounts of songs are essential

→ Automatic Music Genre Classification (AMGC)

Automatic Music Genre Classification (AMGC)

Aim: train a machine learning model to automatically categorize a piece of music into one of many genres

Benchmark: human performance accuracy 70% [11]

Fundamental step that is essential to many applications:

- Music recommendation systems
- Playlist generation
- Music library organization

Fionaa

Top tracks this month

Only visible to you

#	Title	Album		
1	Stargaze Nvr/Mnd, NÜ	Stargaze	♥	1:54
2	a place to call home E NÜ, Nvr/Mnd	a place to call home	♥	1:30
3	Weather Any Storm Cody Francis	Weather Any Storm	♥	2:40
4	Time Machine E WILLOW	WILLOW	♥	2:23
5	Summertime In Paris E Jaden, WILLOW	ERYS (Deluxe)	♥	4:30
6	afraid of change NÜ, Kayli Marie	afraid of change	♥	2:44
7	Pauline À La Plage Jasing Rye	Pauline À La Plage	♥	5:13
8	NÜ - so close to giving up NÜ, vict molina	NÜ - so close to giving up	♥	2:29
9	小島 Tang Siu Hau	小島	♥	3:27
10	I Tried to Get to You NÜ, dyslm, yaeow	I Tried to Get to You	♥	2:10


```
5     "followers": {
6       "href": null,
7       "total": 2305125
8     },
9     "genres": ["afrofuturism", "pop", "post-teen
pop", "pov: indie"],
10    "href":
"https://api.spotify.com/v1/artists/3rWZHrfrsPBxVy
692yAIxF",
11    "id": "3rWZHrfrsPBxVy692yAIxF",
```

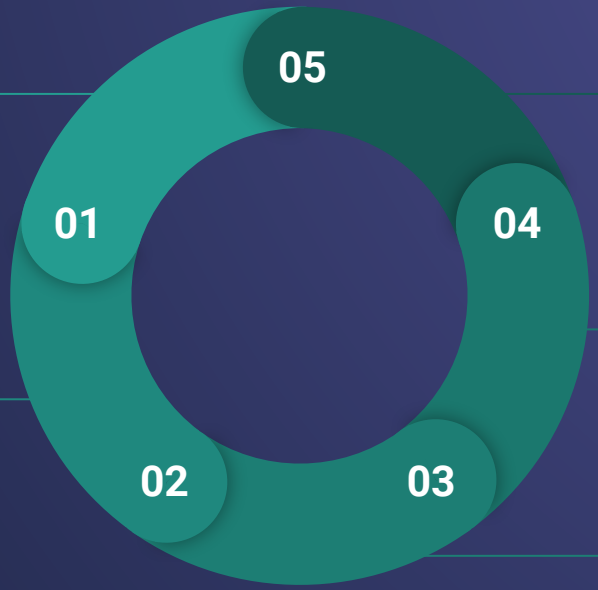
Automatic Music Genre Classification (AMGC)

Music?

- Temporal Dynamics

Rhythm

Beat, Meter



Timing & Expression
rubato (flexible timing), ritardando (slowing down), accelerando (speeding up)

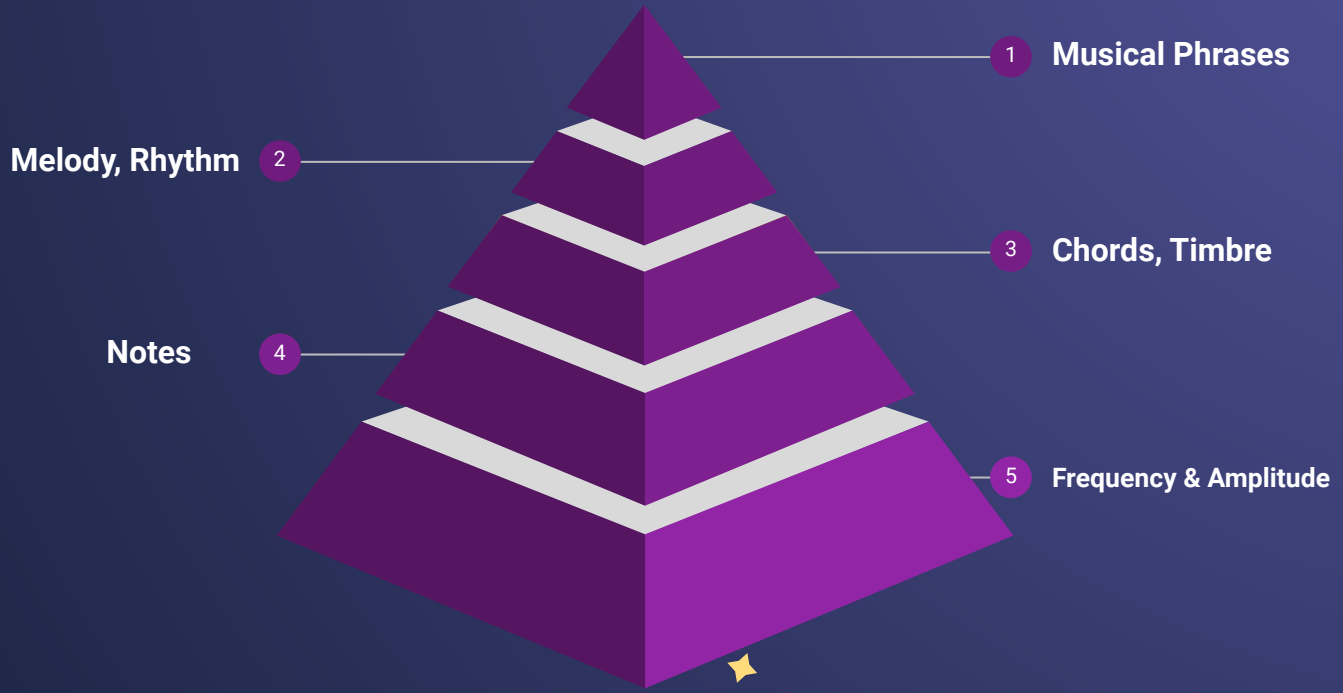
Form/Structure
verse-chorus-verse
sonata

Tempo
beats per minute (BPM)

Automatic Music Genre Classification (AMGC)

Music?

- Hierarchical structure



Automatic Music Genre Classification (AMGC)

Methods	Accuracy
Machine Learning	
KNN	0.54[2], 0.61[4], 0.64[5]
SVM	0.60[2], 0.78[5]
Random Forest	0.81[7]
Neural Network	
CNN	0.72[12], 0.82[2]
State-of-the-art	
Sparse representations	0.91[8]
YOLOv4	0.945[9]
BBNN	0.939[10]

Automatic Music Genre Classification (AMGC)

***Limitations*:**

Traditional machine learning method:

- Focused on hand-crafted feature extraction → fail to capture the complexity and nuances within music data, heavily relies on the quality of the manually engineered features

Traditional CNN:

- Primarily designed for image data
- Effective in capturing spatial hierarchies in data
- Do not inherently model the sequence of data with temporal dynamic features [10]

Motivations for this project

1. Use Deep Learning on AMGC task instead of Traditional Machine Learning; do not rely on any hand-crafted features

2. Use of visual input to leverage the power of pre-trained models and techniques developed for image classification tasks

3. Propose an architecture that can capture both the complex hierarchical and temporal characteristics of music

Thesis

Propose a novel method based on visual Mel-spectrum for AMGC task that can capture both spatial and temporal information

Inspired by the success of ResNet-18 in image classification tasks, propose its use for extracting high-level features from the spectrograms

Incorporate a GRU layer to treat the time-sequence data (from left to right of the spectrogram), capturing the temporal dynamics inherent in the music data

02

Dataset & Data Augmentation

GTZAN Dataset

Item	Data	Details
Genre	10	[blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, rock]
Length	∼30 s	Minimum length: 29.93s; Maximum length: 30.65s; Mean length: 30.02s
Sample rate	22,050 Hz	
Song format	WAV	
Total	1000	100 files/genre

Data Augmentation

Limitation of GTZAN dataset: only 1,000 data, all with a (slightly) different length

→ For each clip, randomly sampled a contiguous 3-second[1] window at 5 non-overlapping random locations, thus augmenting data to 5000 clips of 3 seconds each

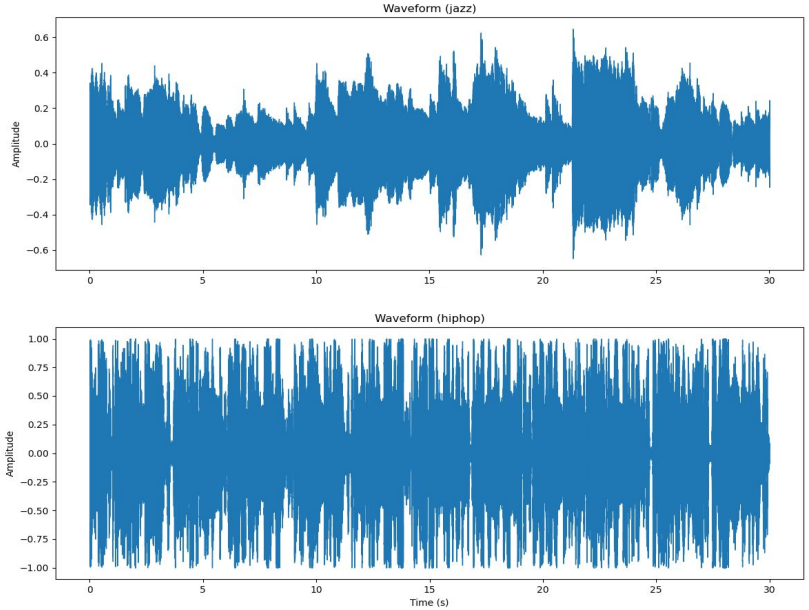
Augmented Data:

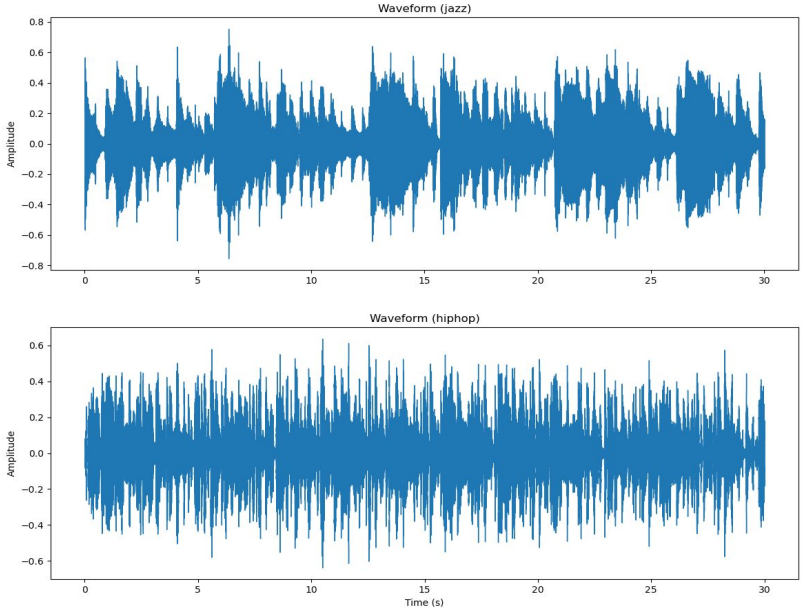
- 10 genres
- 500 clips/genre
- 3s/clip
- Raw audio shape: (5000, 6,6150)

03

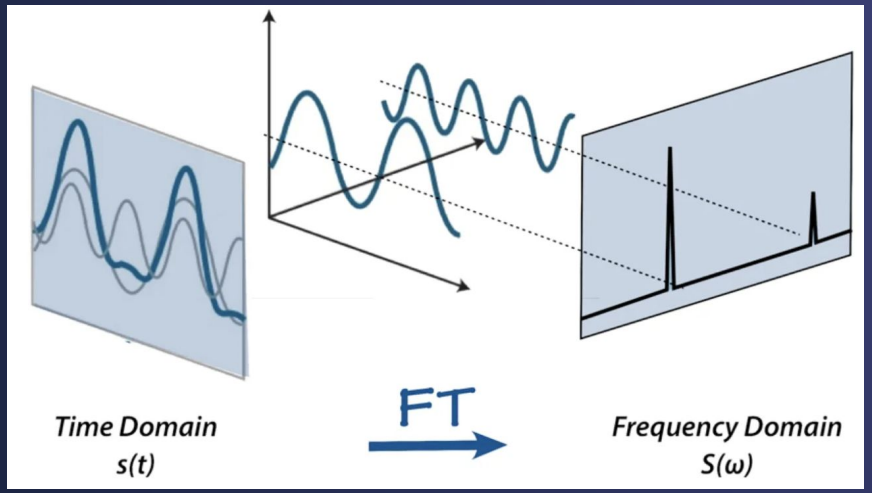
Understanding the Input

Raw data

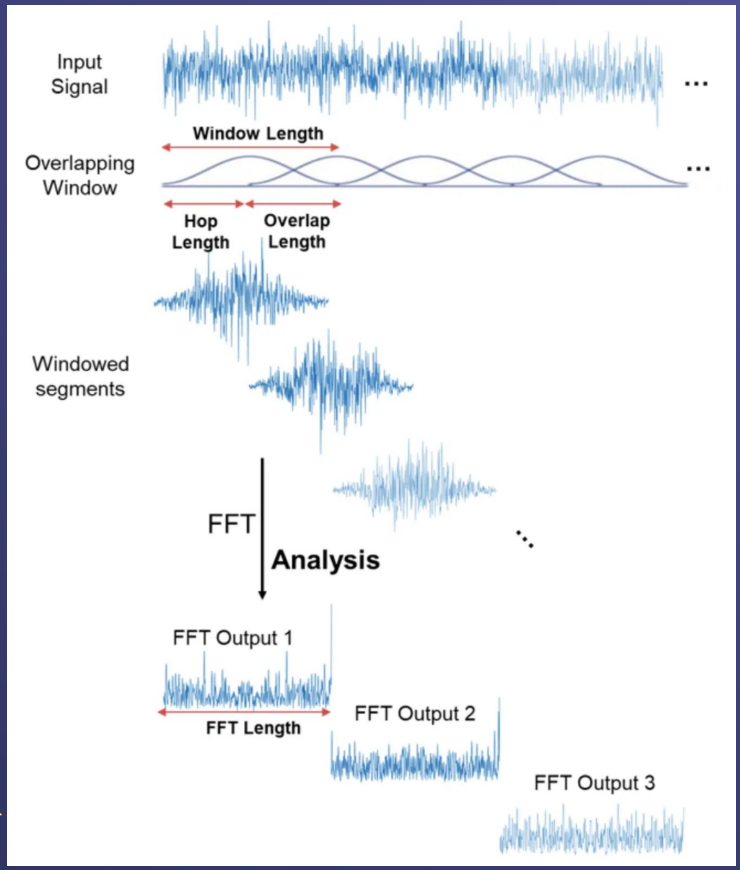

Waveform (jazz)
Amplitude
Waveform (hiphop)
Amplitude
Time (s)


Waveform (jazz)
Amplitude
Waveform (hiphop)
Amplitude
Time (s)

Short-Time Fourier Transforms (STFT)

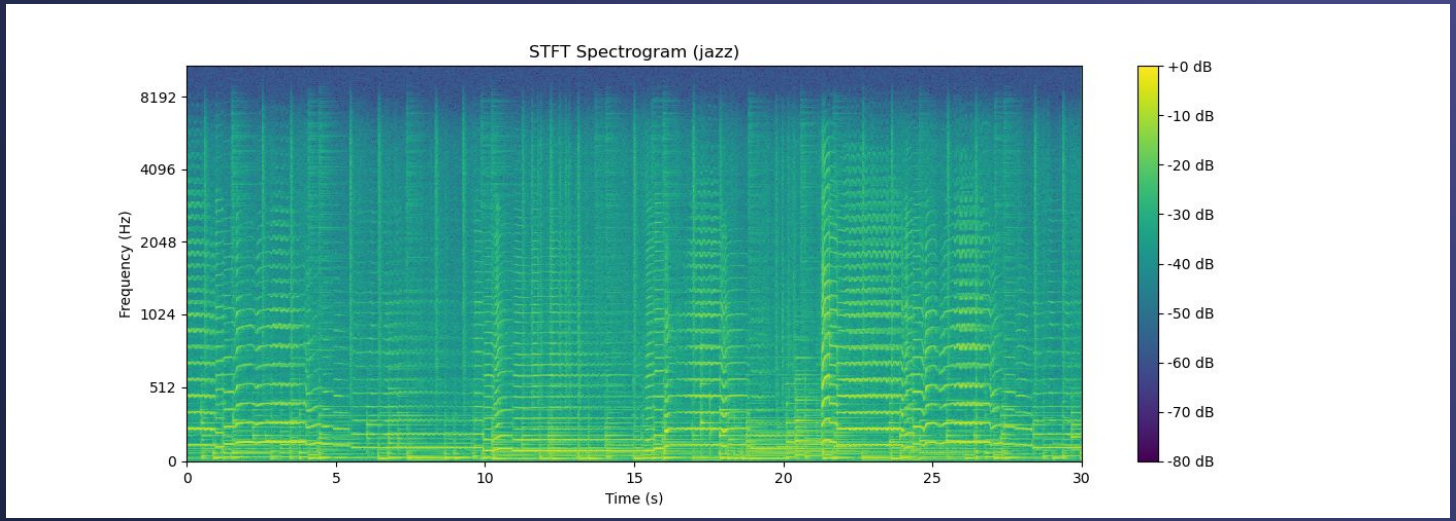


[14]



Short-Time Fourier Transforms (STFT)

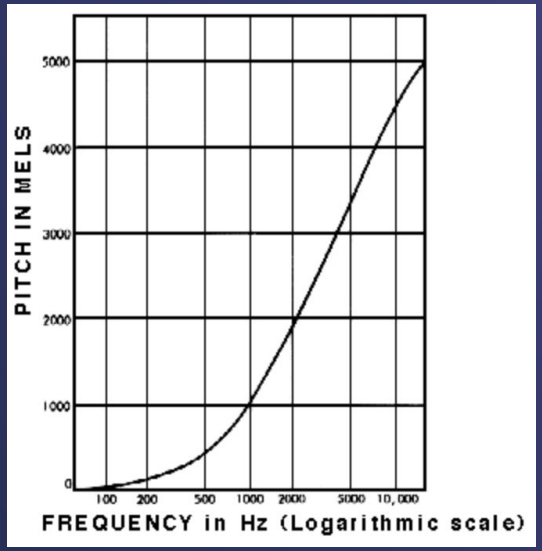
```python
# Compute the STFT of the audio
stft = librosa.stft(audio)
# Convert the complex-valued STFT to magnitude for visualization
magnitude, _ = librosa.magphase(stft)
magnitude_db = librosa.power_to_db(magnitude, ref=np.max)
# Plot the spectrogram
plt.figure(figsize=(14, 5))
librosa.display.specshow(magnitude_db, x_axis='time', y_axis='mel', sr=sample_rate, cmap='viridis')
plt.title('STFT Spectrogram (blues)')
plt.xlabel('Time (s)')
plt.ylabel('Frequency (Hz)')
plt.colorbar(format='%+2.0f dB')
```



Mel Spectrogram

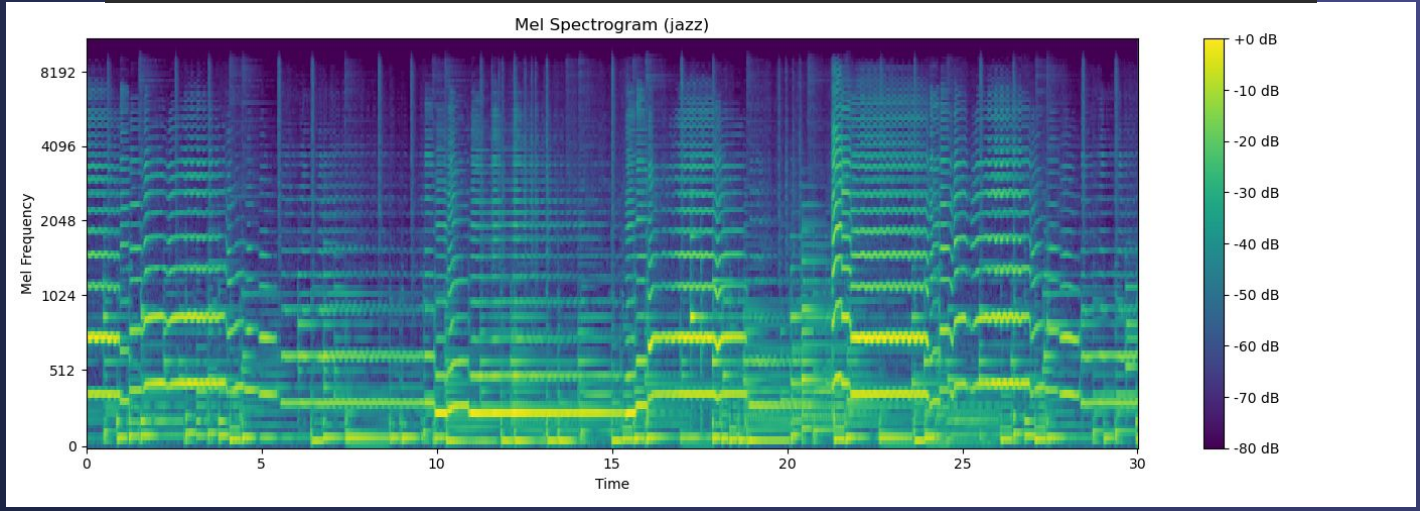
Studies have shown that humans do not perceive frequencies on a linear scale.

In 1937, Stevens, Volkmann, and Newmann proposed a unit of pitch [13]



Mel Spectrogram

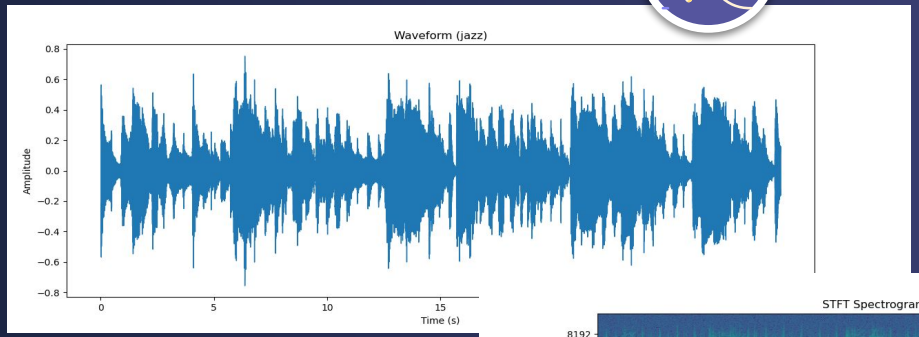
```python
# Compute the Mel spectrogram
mel_spectrogram = librosa.feature.melspectrogram(y=audio, sr=sample_rate, n_mels=128)
# Convert to dB scale
mel_spectrogram_db = librosa.power_to_db(mel_spectrogram, ref=np.max)
# Visualize the Mel spectrogram
plt.figure(figsize=(14, 5))
librosa.display.specshow(mel_spectrogram_db, x_axis='time', y_axis='mel', sr=sample_rate, cmap='viridis')
plt.colorbar(format='%+2.0f dB')
plt.title('Mel Spectrogram (blues)')
plt.xlabel('Time')
plt.ylabel('Mel Frequency')
```



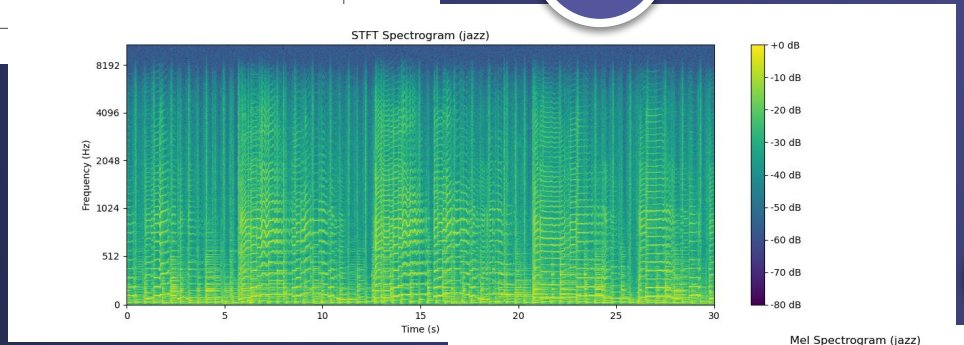
Parameters

Parameter	Value[2][3]
Audio Length (seconds)	3
Sampling Rate (hertz)	22,050
Window Length (frames)	2,048
Overlap Length (frames)	512
FFT Length (frames)	2,048
Num Bands (filters)	128

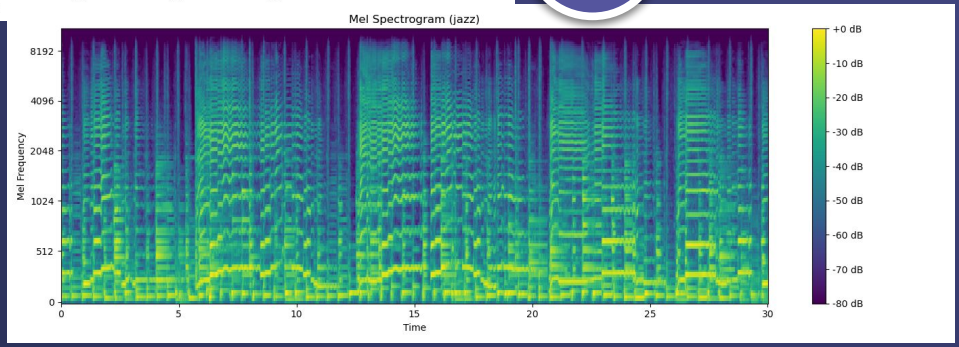
Raw waveform



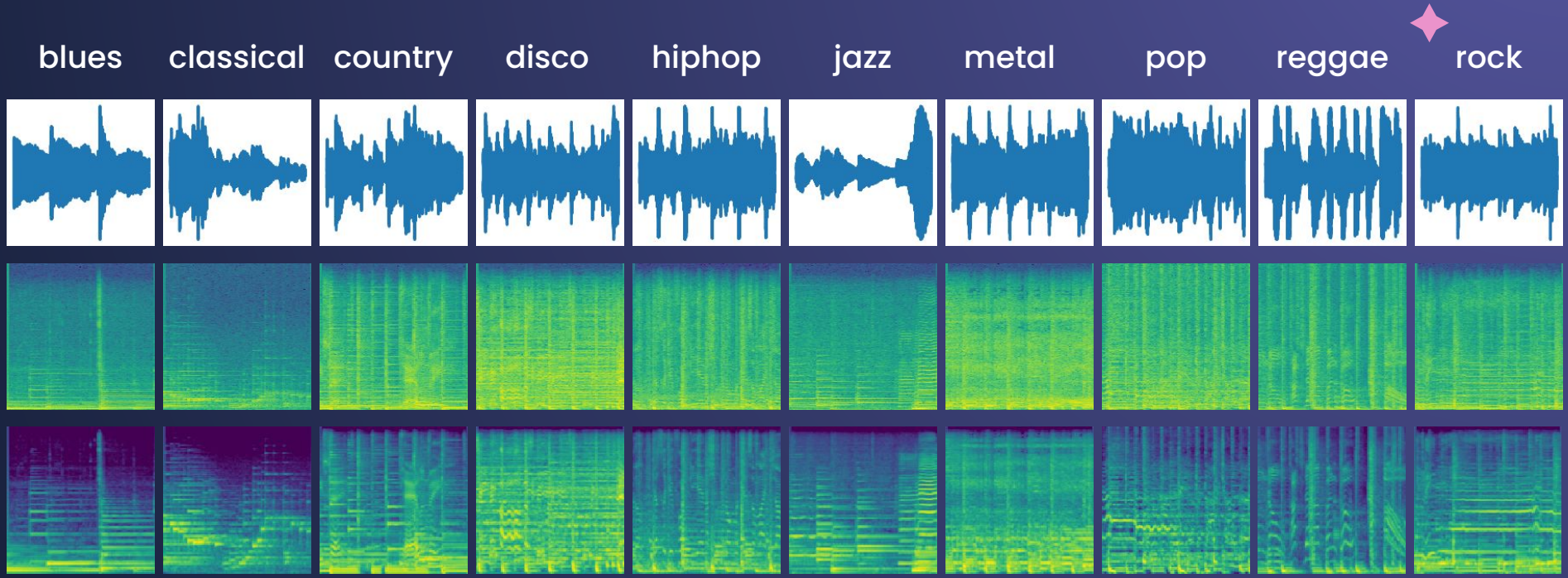
STFT spectrogram



Mel spectrogram



Input


blues
classical
country
disco
hiphop
jazz
metal
pop
reggae
rock

Time Series Data vs Spectrograms

Time series data

- Pros:
  - Raw Representation
  - Simplicity
  - More Explicit Temporal Dynamics
- Cons:
  - High Dimensionality
  - Susceptibility to Noise
  - Lack of Frequency Information

Spectrograms

- Pros:
  - Rich Information (captures frequency)
  - Robustness
  - Mel spectrogram: Account for Human Auditory Perception
- Cons:
  - Computational Expense

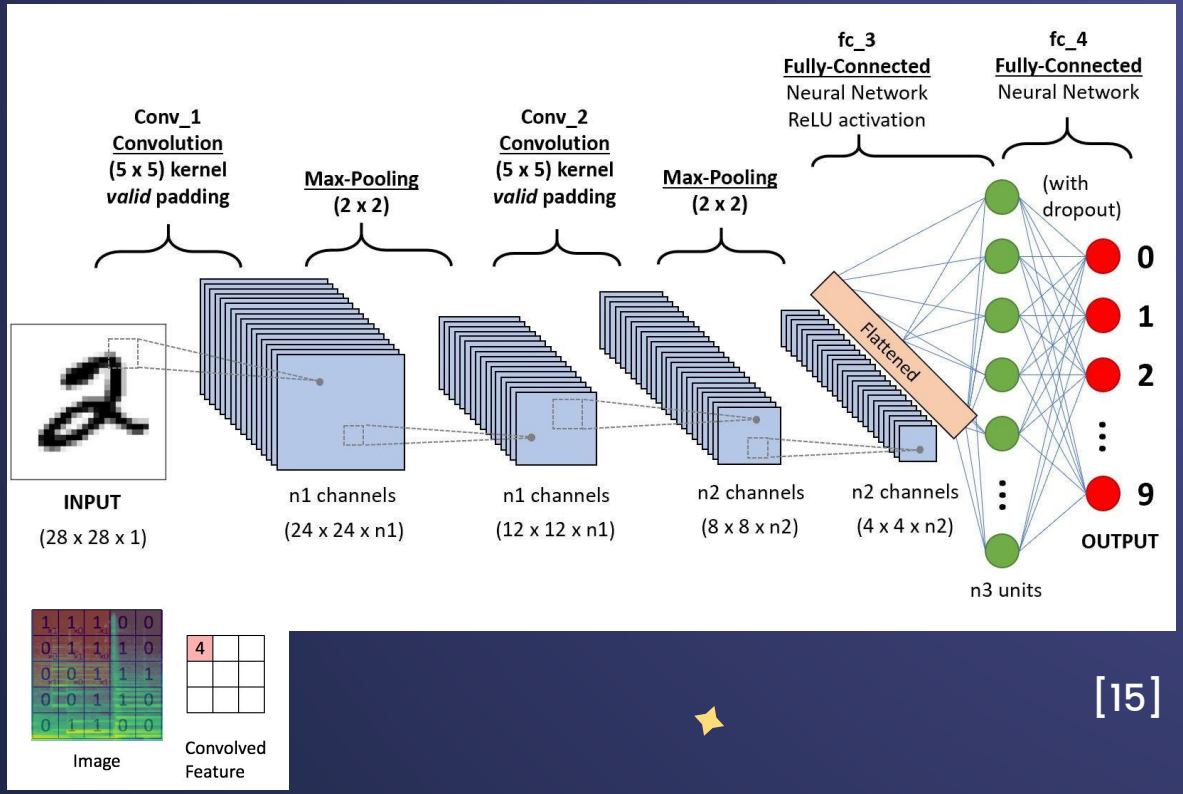
04

Baseline Models Using Machine Learning

Methods	Raw data	STFT Spectrogram	Mel Spectrogram
KNN:			
n = 10	0.25	0.67	0.66
n = 15	0.26	0.64	0.63
n = 20	0.27	0.60	0.60
SVM	0.12	0.69	0.65

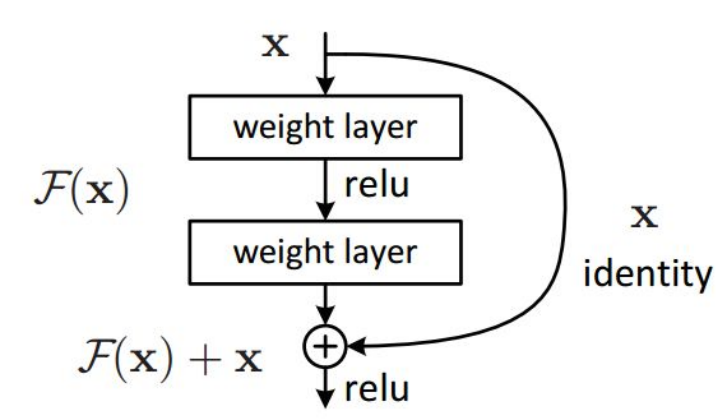
Proposed Model 05

Convolutional Neural Networks (CNN) based



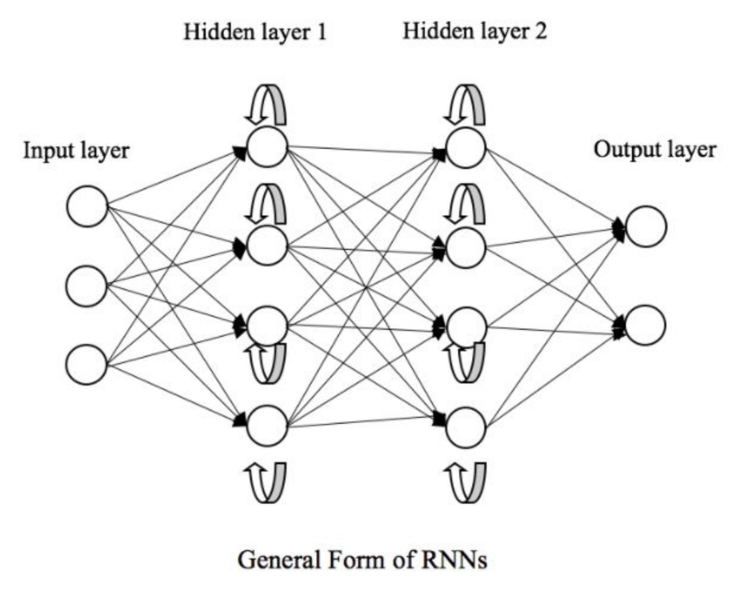
[15]

Residual Network (ResNet)



ResNet18: a version of the architecture that is 18 layers deep

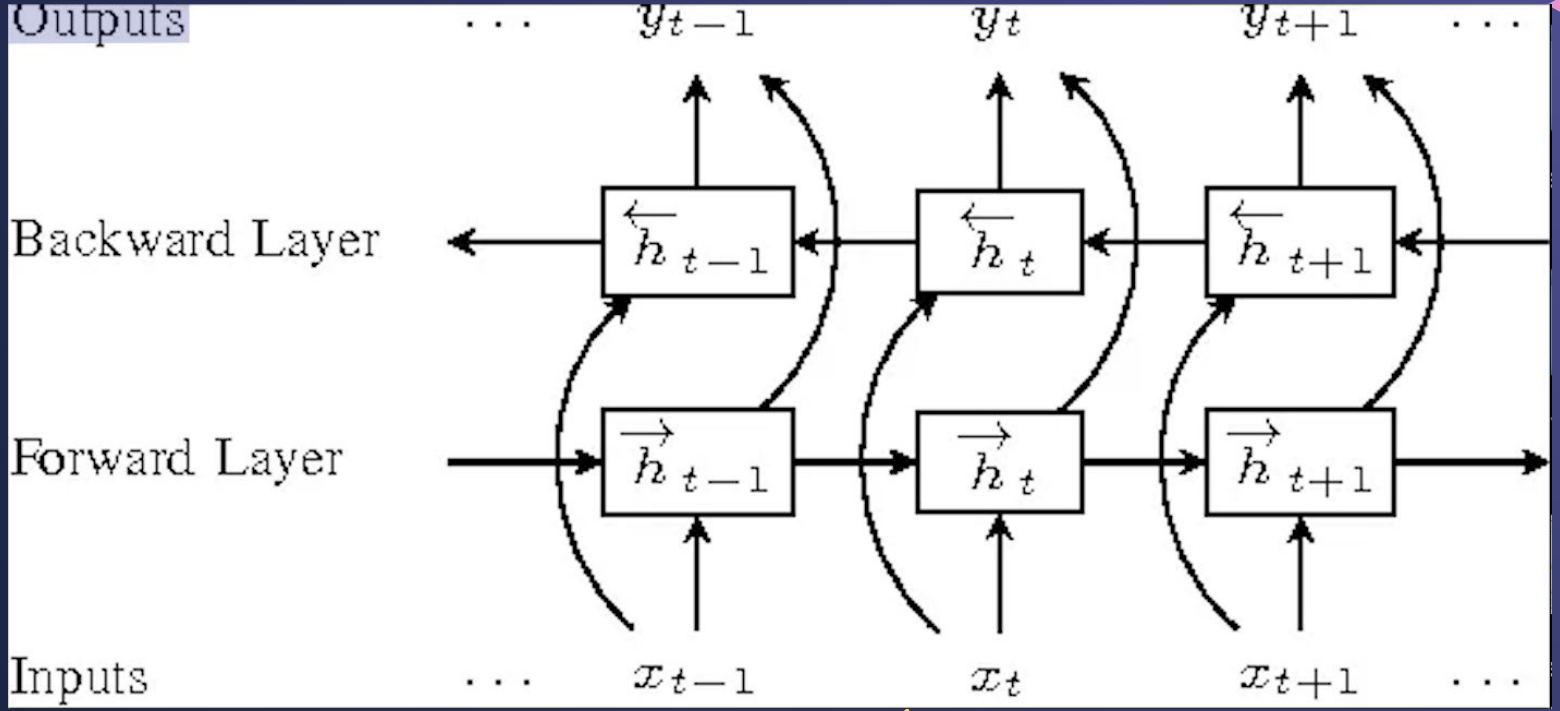
Recurrent Neural Networks (RNNs)



RNNs can use their internal state (memory) to process sequences of inputs - effective for tasks where temporal dynamics are important

Standard recurrent networks (RNNs) only take sequence of previous context but ignore the backwards dependencies which are also important for feature learning

Gated Recurrent Unit (GRU)

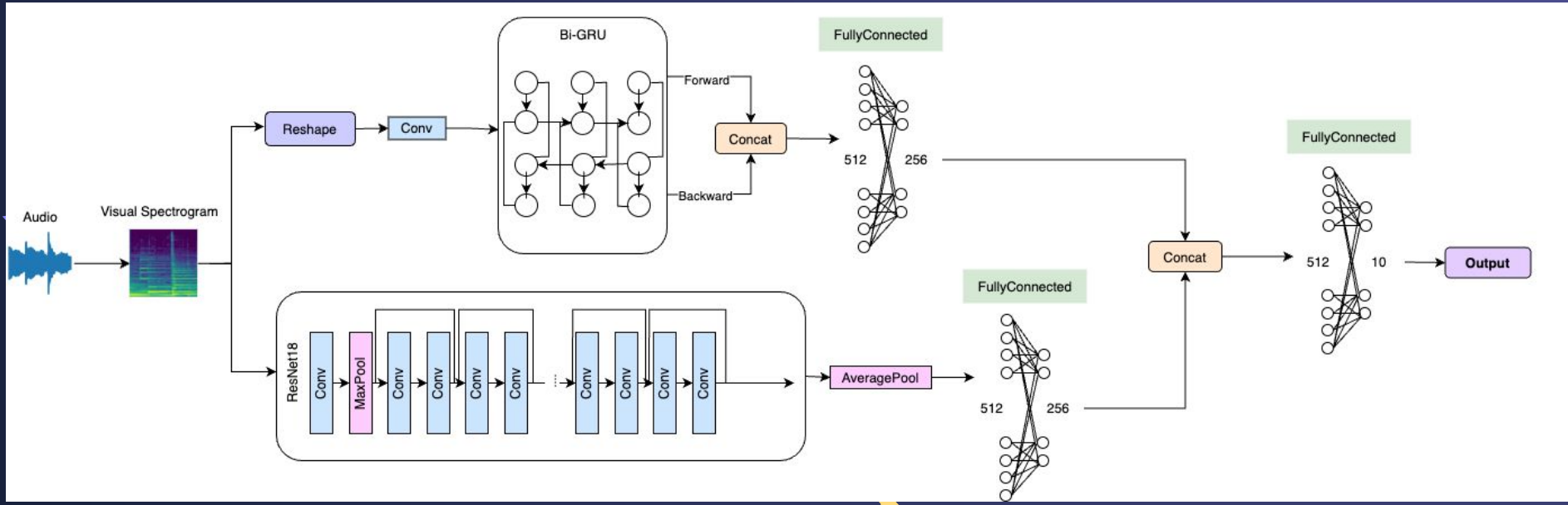


Proposed Idea

Considerations:

1. Utilising ResNet as feature extractor to capture spatial information
2. Utilising GRU to capture temporal features which are lost in CNNs

Proposed Architecture: Parallel Bi-GRU/ResNet Hybrid



06

Results & Discussion

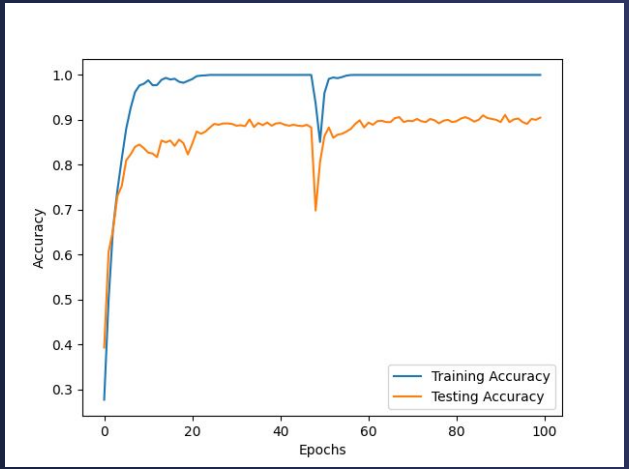
Comparison of Deep Learning Models

Models	STFT Spectrogram				Mel Spectrogram			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
CNN	0.61	0.61	0.61	0.61	0.66	0.66	0.66	0.66
Bi-GRU	0.71	0.71	0.71	0.71	0.78	0.78	0.78	0.78
ResNet18	0.70	0.70	0.70	0.70	0.87	0.87	0.87	0.87
Res-BiGRU	0.86	0.86	0.86	0.86	0.90	0.90	0.90	0.90

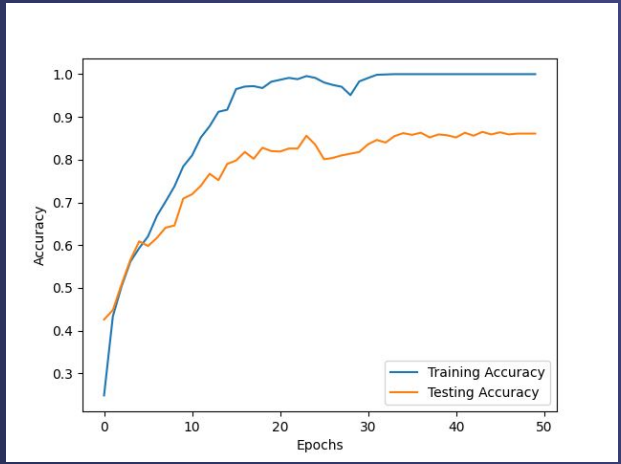
$$precision = \frac{TP}{TP + FP} \qquad F1 - score = \frac{2 * (Recall * Precision)}{Recall + Precision}$$

$$RecaII = \frac{TP}{TP + FN} \qquad Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

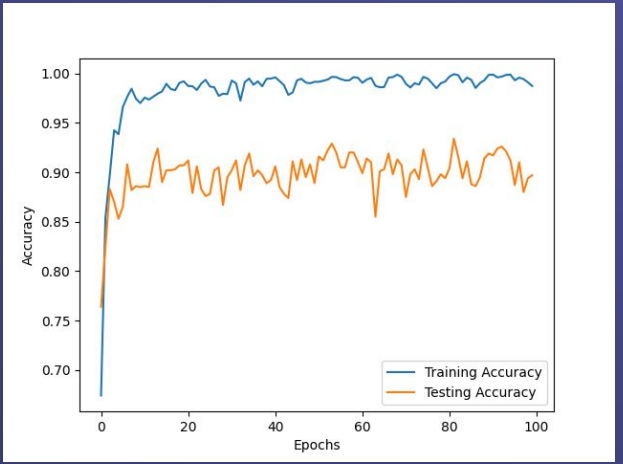
Accuracy Curve



ResNet18

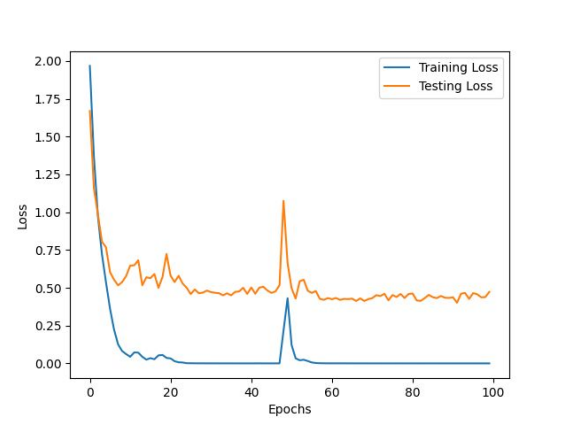


Bi-GRU

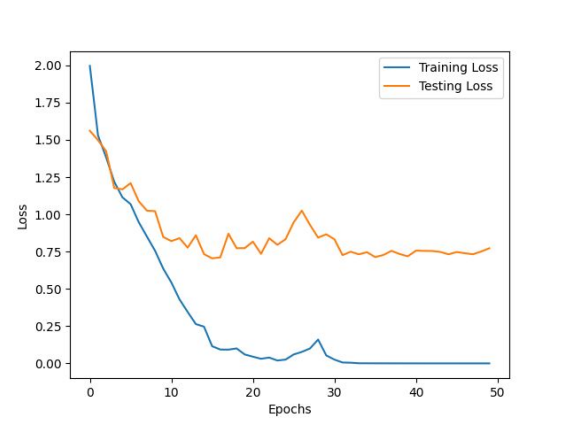


Hybrid

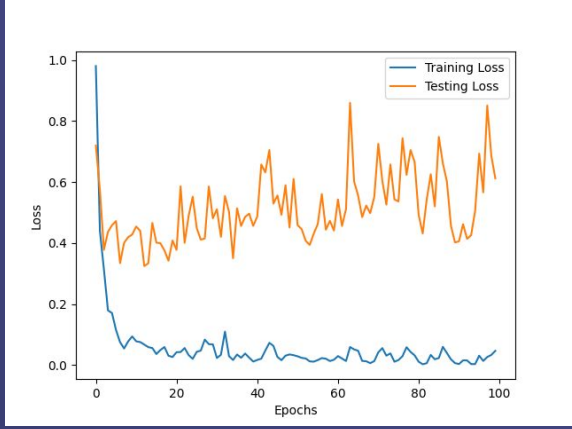
Loss Curve



ResNet18

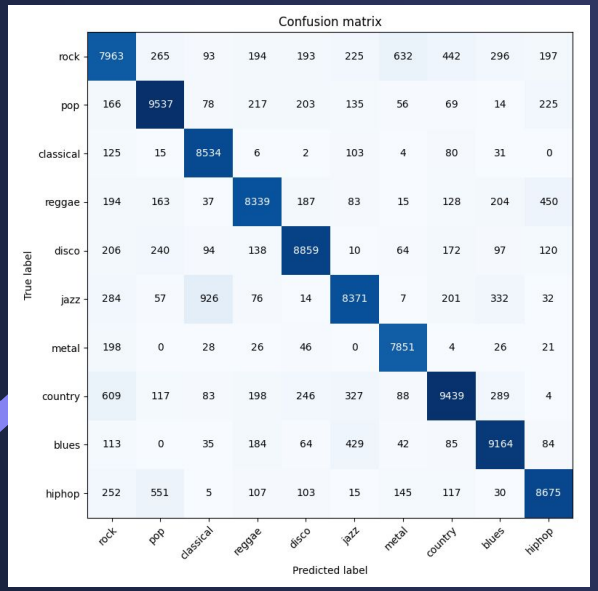


Bi-GRU

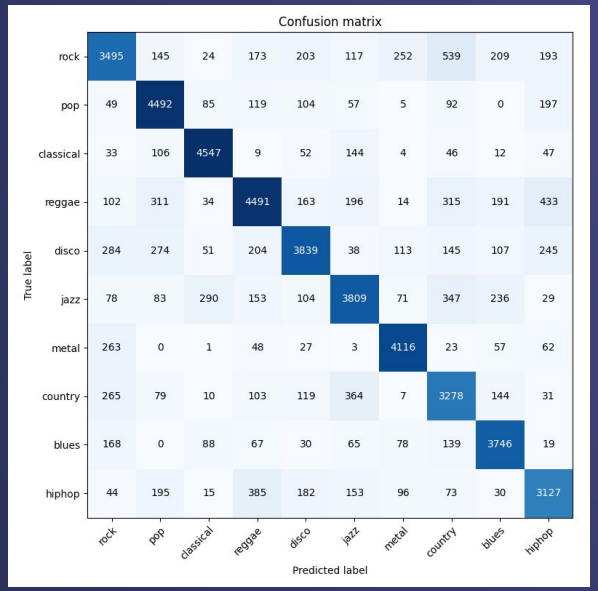


Hybrid

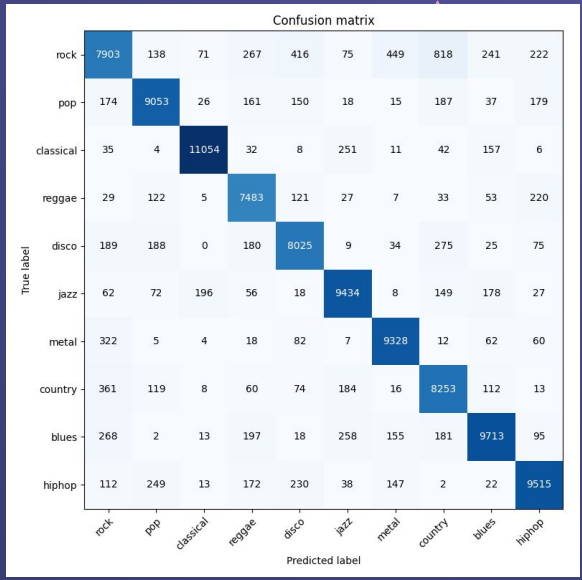
Confusion Matrix



ResNet18



Bi-GRU



Hybrid

Insights & Discussion

1: Hybrid Approach

Combining different types of models to tackle complex tasks

2: Transfer learning of image information on audio data

Transferring methodology and techniques from image analysis to audio analysis

3: Mel Spectrograms vs STFT Spectrograms

Human-centered approach in model design

1: Transfer learning of image on audio

Converting audio files into visual mel spectrograms → use image-based model (ResNet18) to audio data

Transferring methodology and techniques from image analysis to audio analysis

2: Mel Spectrograms vs STFT Spectrograms

Mel spectrograms: align more closely with human auditory system, indeed yielding better results

Human-centered approach in model design

3: Hybrid Approach

Integrates ResNet18 and GRU models to simultaneously extract hierarchical and temporal features from music data

Combining different types of models to tackle complex tasks

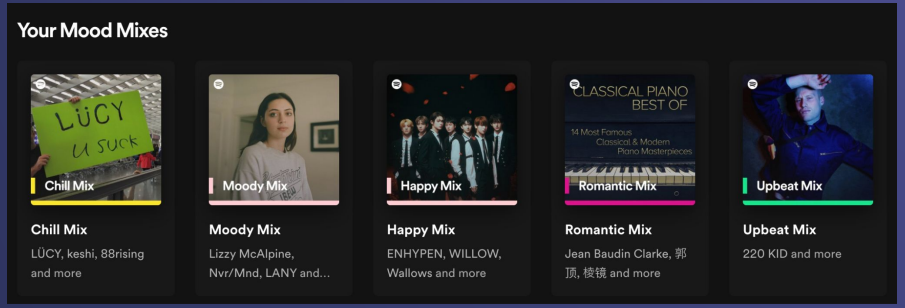
Conclusion

1. The combination of Residual Networks (ResNet) and Gated Recurrent Units (GRU) can create an effective hybrid deep learning model for processing audio data.

2. Images can serve as an alternative input representation for deep learning models on the Automatic Music Genre Classification (AMGC) task.

3. When considering spectrograms as image inputs, Mel Spectrograms are generally more suitable compared to Short-Time Fourier Transform (STFT) Spectrograms.

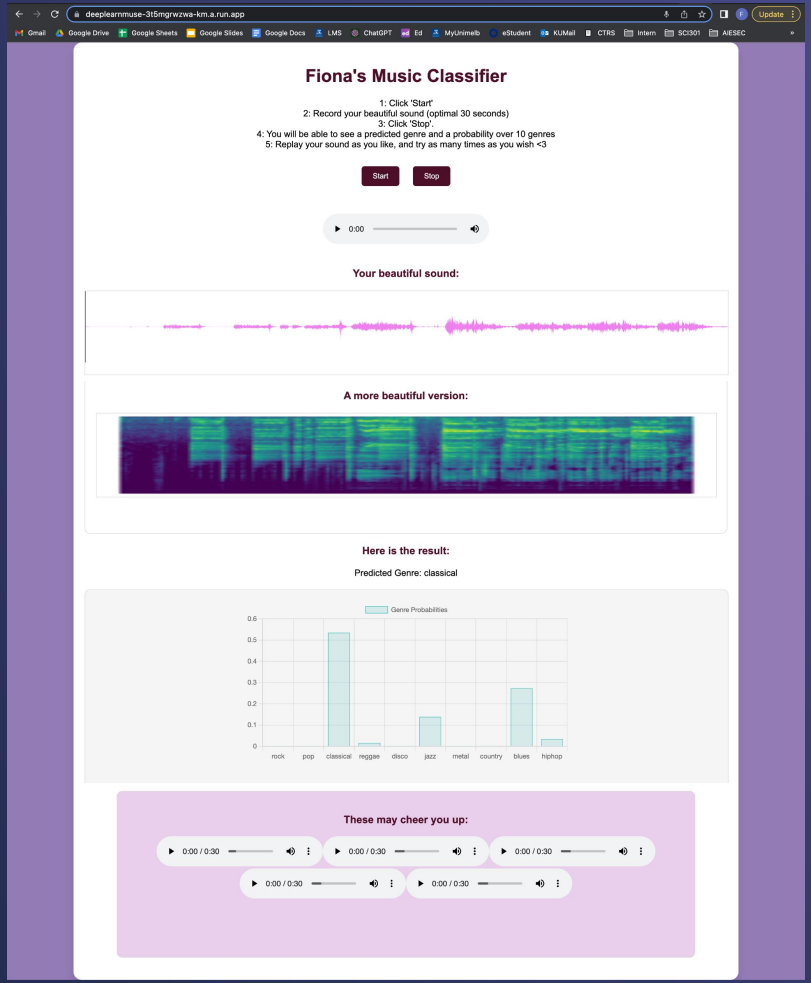
Future Work



1. **Multitask Learning**
2. **Generative Model**
3. **Explainability and Interpretability of model**

Extension: Deployment 07

https://deeplearnmuse-3t5mgrwzwa-km.a.run.app/



Powered by Google Engine

Acknowledgement

I want to extend my deepest gratitude to my two incredible supervisors who have been instrumental in the completion of this project.

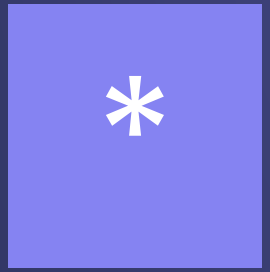
To Lars, thank you for agreeing to supervise my project and for your patience throughout this process. Your guidance has been instrumental in keeping me on the right track. Your insights and expertise have been invaluable in my understanding of this complex task. Your passion has greatly enhanced my learning experience, I am truly thankful.

To Nestor, your profound background knowledge and ideas have been a source of inspiration for me. Your expertise in deep learning have enriched my understanding, and always put me on the right track. I'm genuinely grateful for the way you continually inspire me to learn more, and thank you for always helping me when I have troubles!

I believe this experience will be a precious resource for my future academic endeavors. Thank you!!!

Reference

[1] D. Perrot and R. Gjerdigen, “Scanning the dial: An exploration of factors in identification of musical style,” in Proc. Soc. Music Perception Cognition, 1999, p. 88, (abstract).
[2] Derek A. Huang, Arianna A. Serafini, Eli J. Pugh. Music Genre Classification. http://cs229.stanford.edu/proj2018/report/21.pdf.
[3] Cheng, Y.-H.; Kuo, C.-N. Machine Learning for Music Genre Classification Using Visual Mel Spectrum. Mathematics 2022, 10, 4427. https://doi.org/10.3390/math10234427.
[4]George Tzanetakis and Perry Cook. Musical genre classification of audio signals. IEEE Transactions on speech and audio processing,
10(5):293–302, 2002.
[5] https://ijcert.org/ems/ijcert_papers/V4I206.pdf
[7]https://medium.com/hackerdawn/music-genre-classification-using-random-forest-219fc2446666#:~:text=If%20Music%20is%20a%20Place,Road%2C%20Classical%20is%20a%20Temple.
[8]Y. Panagakis, C. Kotropoulos, and G. R. Arce. Music genre classification via sparse representations of auditory temporal modulations. In 2009 17th European Signal Processing Conference, pages 1–5, Aug 2009.
[9]Cheng, Y.-H.; Kuo, C.-N. Machine Learning for Music Genre Classification Using Visual Mel Spectrum. Mathematics 2022, 10, 4427. https://doi.org/10.3390/math10234427.
[10]Liu, C., Feng, L., Liu, G. et al. Bottom-up broadcast neural network for music genre classification. Multimed Tools Appl 80, 7313–7331 (2021). https://doi.org/10.1007/s11042-020-09643-6.
[11] D. Perrot and R. Gjerdigen, “Scanning the dial: An exploration of factors in identification of musical style,” in Proc. Soc. Music Perception Cognition, 1999, p. 88, (abstract).
[12]S. Sugianto and S. Suyanto, "Voting-Based Music Genre Classification Using Melspectogram and Convolutional Neural Network," 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2019, pp. 330-333, doi: 10.1109/ISRITI48646.2019.9034644.
[13] S. S. Stevens, J. Volkmann, E. B. Newman; A Scale for the Measurement of the Psychological Magnitude Pitch. J Acoust Soc Am 1 January 1937; 8 (3): 185–190. https://doi.org/10.1121/1.1915893
[14] https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53
[15] https://saturncloud.io/blog/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way/



Questions & Answers & Feedback