ANNAMALAI

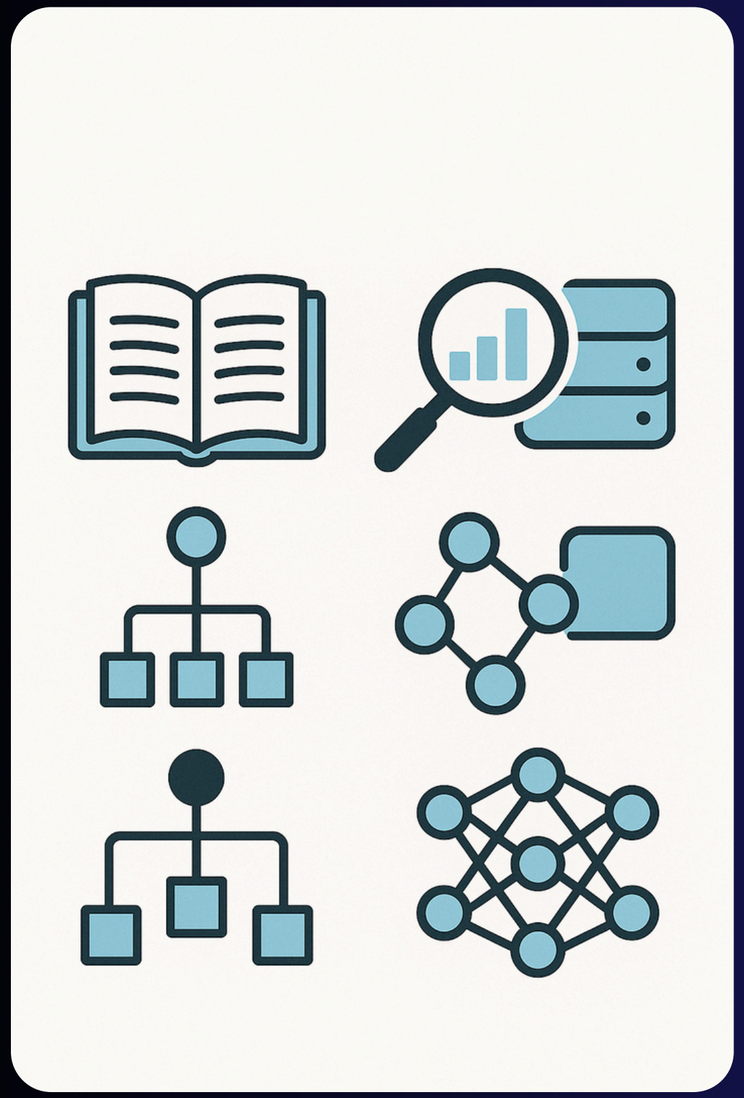
SPEECH EMOTION RECOGNITION (SER) APPLICATION

Audio-based Neural Network for Affective Multiclass Analysis and Labelling using Artificial Intelligence [ANNAMALAI]

24 April 2025

Ankita, Sheng Xiang, Joel, Fuo En, Anthony

ANNAMALAI



AGENDA

1. Background
2. Literature Review
3. Datasets Used
4. Methodology
5. Traditional ML Algorithms
6. Feed-forward Neural networks (FFNNs)
7. Convolutional Neural networks (CNNs)
8. transformer models
9. demo

ANNAMALAI

BACKGROUND

Understanding Speech is COMPLEX due to variability in:

- Tone
- Context
- Accent

Misinterpretation in fields like law or mental health support can lead to SERIOUS consequnces

CHALLENGING but CRUCIAL problem to solve

ANNAMALAI

SPEECH EMOTION RECOGNITION (SER)

Identifies Emotional Cues

AI detects tone, pitch, and intensity to recognize core emotions like happiness, sadness, anger and fear

01

Handles Speech Variability

Adapts to different accents, tones and languages, improving accuracy across diverse speakers

02

Reduces Misinterpretation

Provides objective emotional analysis, minimizing human bias and preventing misjudgments

03

Enhance Emotional Awareness

Helps individuals and professionals better understand emotional states for improved interactions

04

Supports Decision Making

Enables informed responses with better emotional insights of stakeholders

05

ANNAMALAI

AI FOR GOOD

Good Health and Well-Being (SDG 3)

Early detection of emotional distress. Enables timely mental health support and interventions through emotion-aware insights.

Peace, Justice, and Strong Institutions (SDG 16)

Fairer evaluations in investigative settings based on emotional cues in speech. Supports credibility assessment and reduces bias.

ANNAMALAI

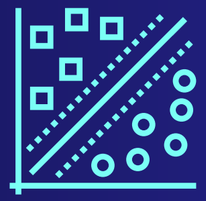
LITERATURE REVIEW

ANNAMALAI

TRADITIONAL APPROACH

Acoustic speech features were extracted from raw speech signals:

1. Spectral features, like MFCCs (Mel Frequency Cepstral Coefficients)
2. Prosodic features, pitch, energy, and speech intensity



Classifiers used are:

1. Support Vector Machines (SVMs)
    - Known to have good generalizability in noisy SER tasks
2. Hidden Markov Models (HMMs)
3. Decision Trees

ANNAMALAI

CURRENT APPROACH

Deep learning approaches are increasingly being used due to:

Enhanced Recognition Rates

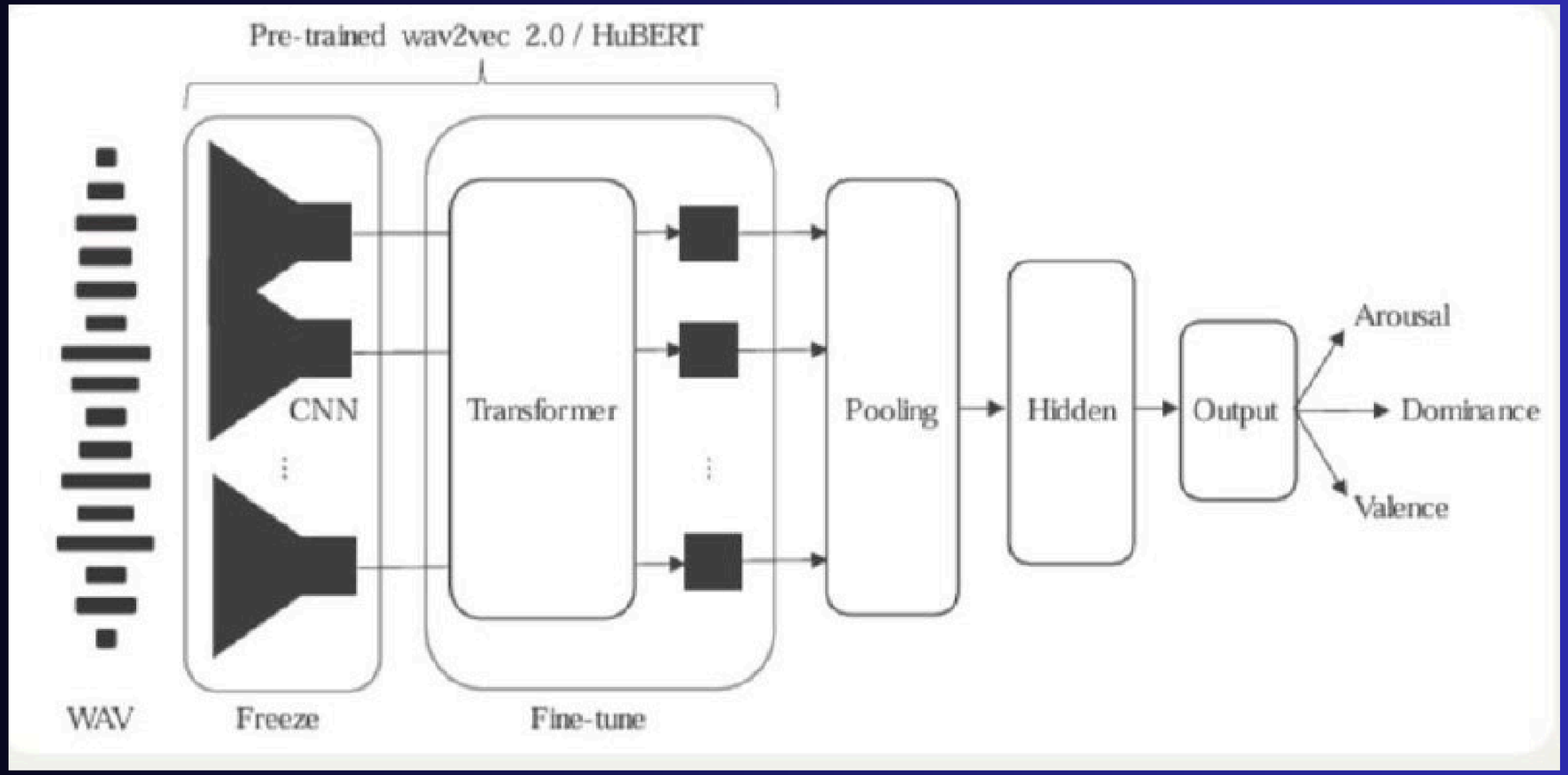
Robustness

End-to-end architectures used to learn features directly from speech signals:

1. Deep Neural Networks (DNNs)
2. Convolutional Neural Networks (CNNs)
3. Recurrent Neural Networks (RNNs)
4. Long Short-Term Memory (LSTMs)

ANNAMALAI

CASE STUDY



Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., & Schuller, B. W. (2023). Dawn of the transformer era in speech emotion recognition: Closing the valence gap. arXiv preprint arXiv:2203.07378. https://arxiv.org/abs/2203.07378

ANNAMALAI

NOISY SER

ANNAMALAI

NOISY SER

Preprocessing

- Noise reduction reduces unwanted background noise from a speech signal
- Voice Activity Detection (VAD) that isolates speech portions from non-speech ones (i.e., silence, noise)
- Signal normalisation

Noise Robust Feature Extraction

- Common noise-robust features include MFCCs and W-WPCCs.
- Prosodic features (e.g., rhythm, pitch, stress) capture emotion cues.
- Mel spectrograms are widely used with CNNs for deep feature learning.

Data Augmentation

- Artificial noise like AGWN, pink noise, and environmental sounds is added to clean speech data.
- Enhances SER robustness across varied real-world conditions.

ANNAMALAI

DATASETS USED

Combined Dataset – 74,705 Samples

- CREMA-D (7,442 Samples)
- SAVEE (480 Samples)
- MELD (9,989 Samples)
- TESS (2,800 Samples)
- MLEnd (32,654 Samples)
- ESD (17,500 Samples)
- RAVDESS (1,440 Samples)
- JL Corpus (2,400 Samples)

ANNAMALAI

DATASETS USED

Combined Dataset – **74,705** Samples

↓ Removal of Corrupt audio (3)

Dataset (w/o Corrupt Audio): **74,702** Samples

↓ Removal of Outliers: Audio <1s or >4s (9,948)

Dataset (w/o Outliers): **64,754** Samples

↓ Removal of Rare Classes* (1,580)

Final Dataset: **63,174** Samples

*Rare Classes: Anxious, Apologetic, Assertive, Encouraging, Concerned, Excited & Calm

ANNAMALAI

DATASETS USED

Final Dataset: **63,174** Samples

Stratified Sampling: Ensuring Proportional Representation

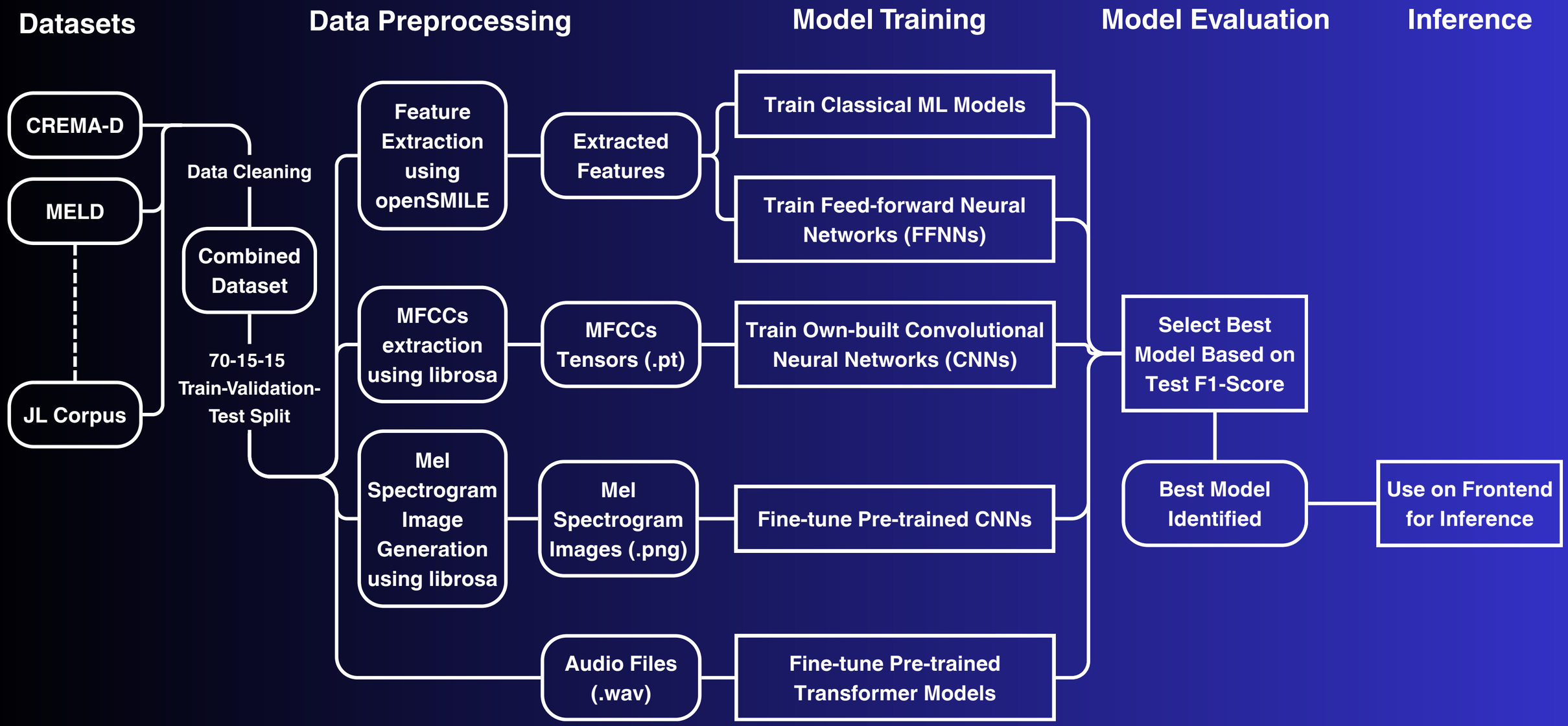
70%
Training Set
Data used to fit the model's parameters

15%
Validation Set
Held-out data used for tuning hyper-parameters and selecting the best model

15%
Testing Set
Unseen data used for the final, unbiased evaluation of model performance.

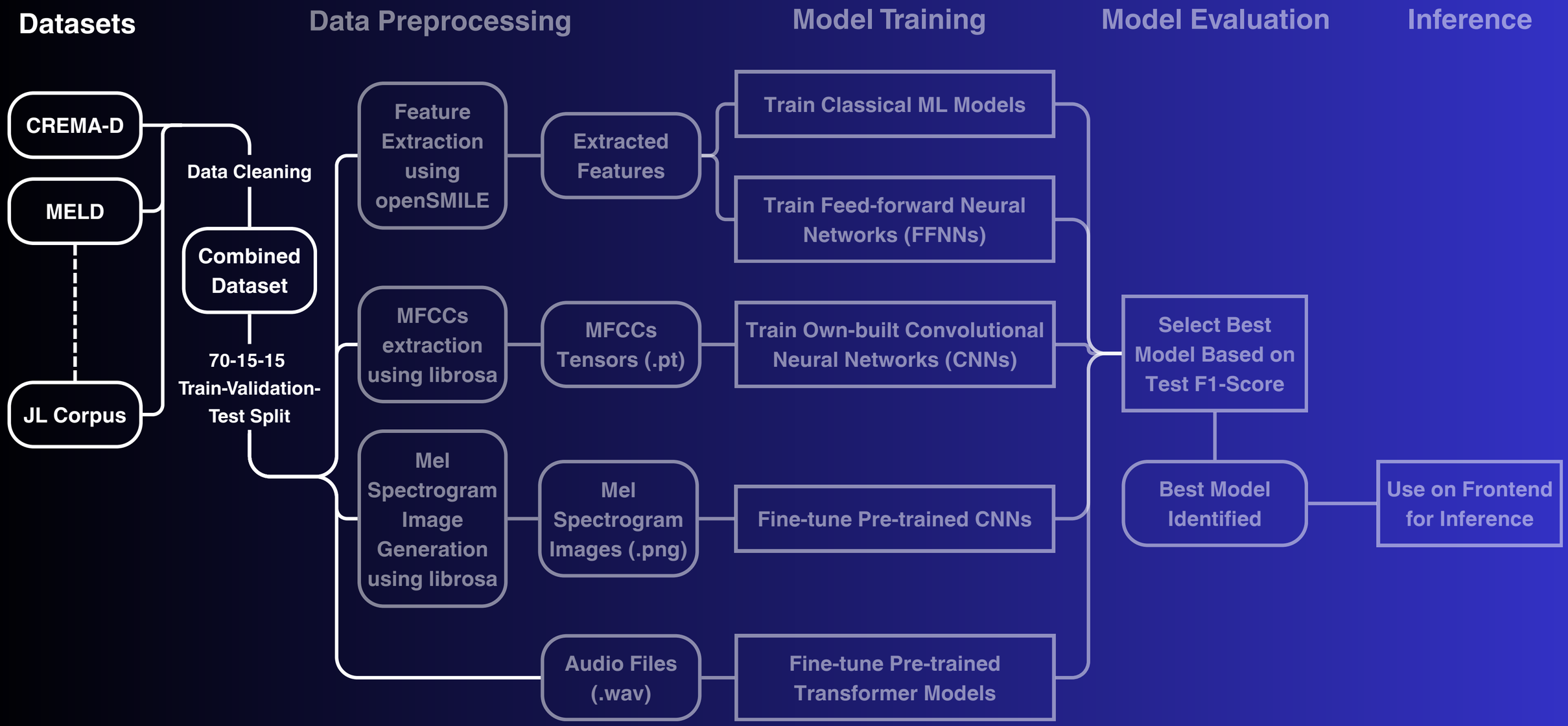
ANNAMALAI

METHODOLOGY



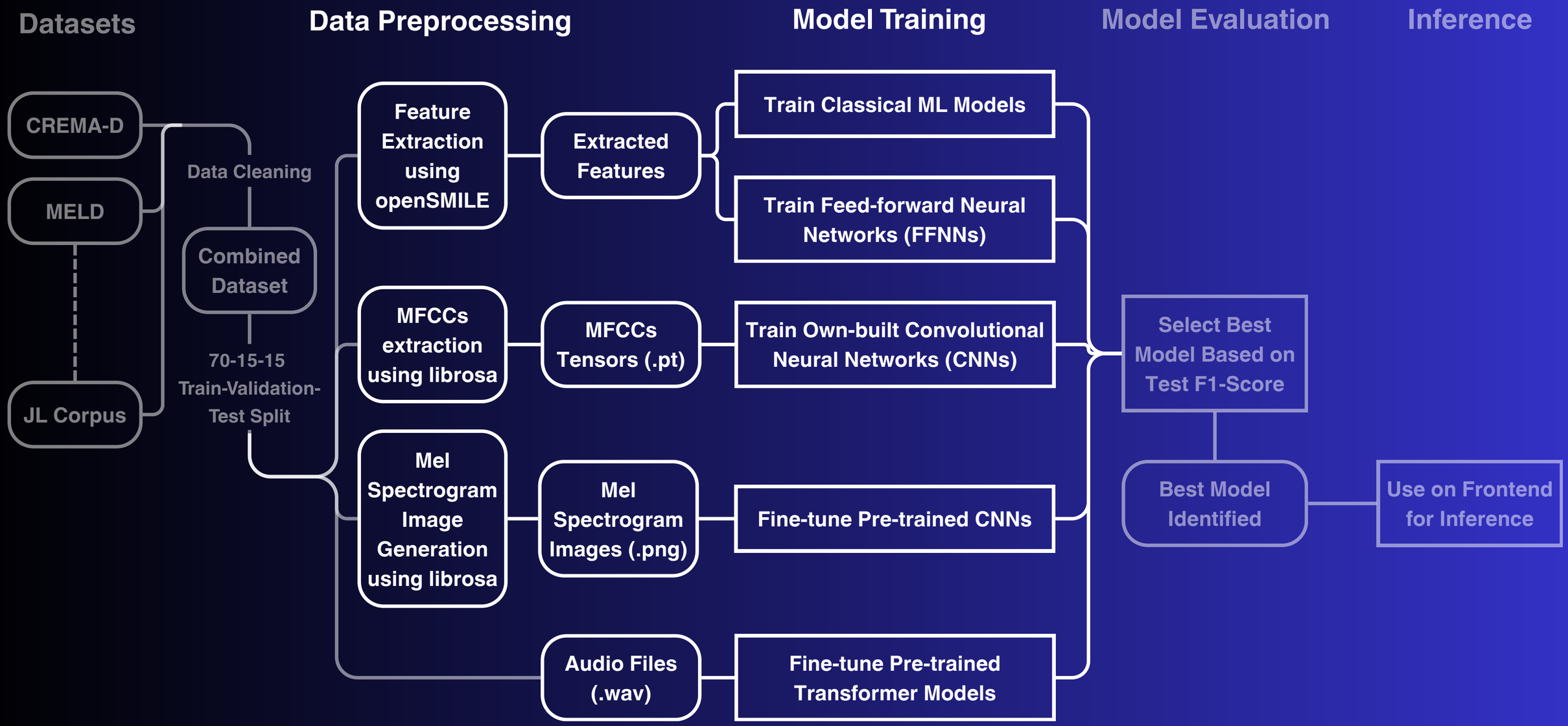
ANNAMALAI

METHODOLOGY



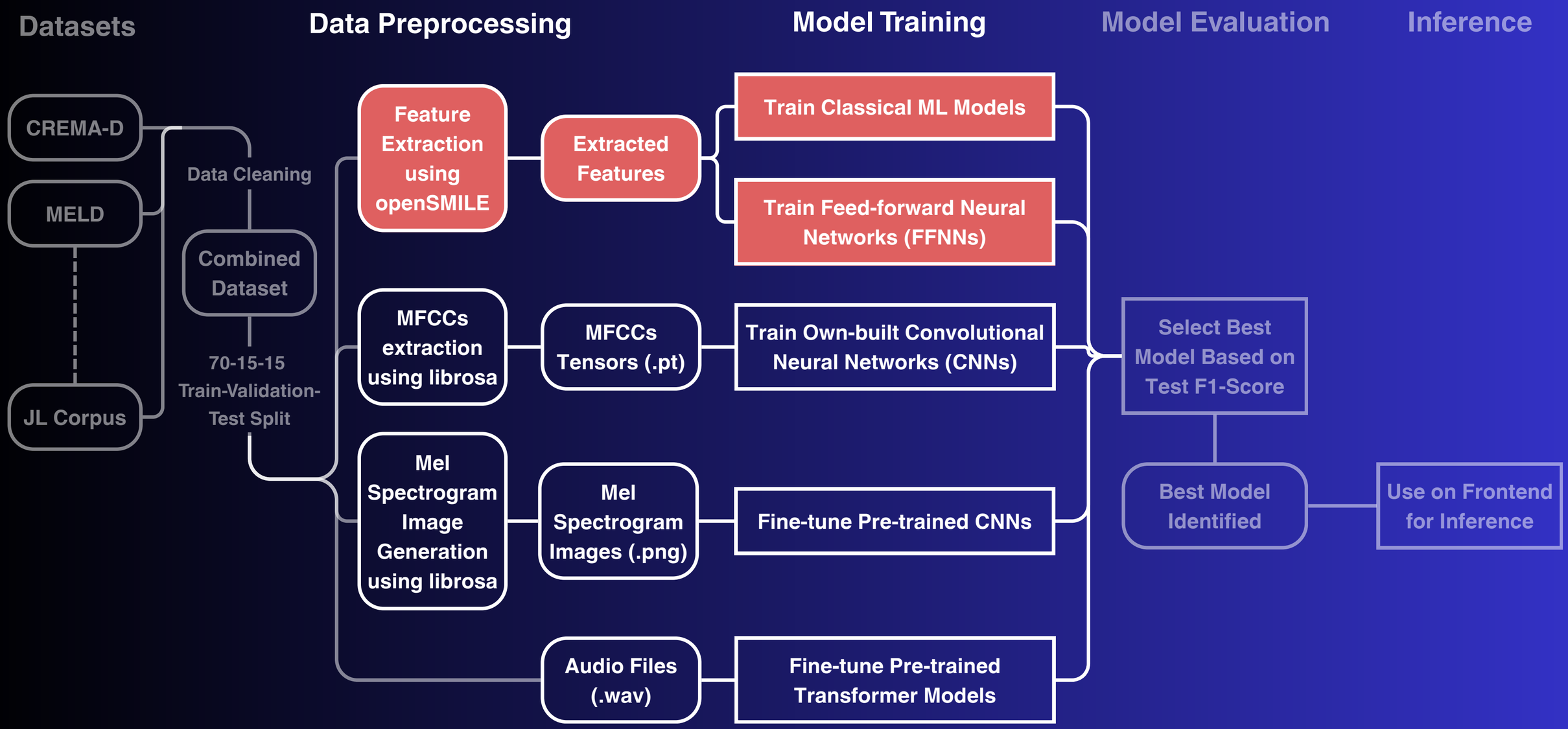
ANNAMALAI

METHODOLOGY



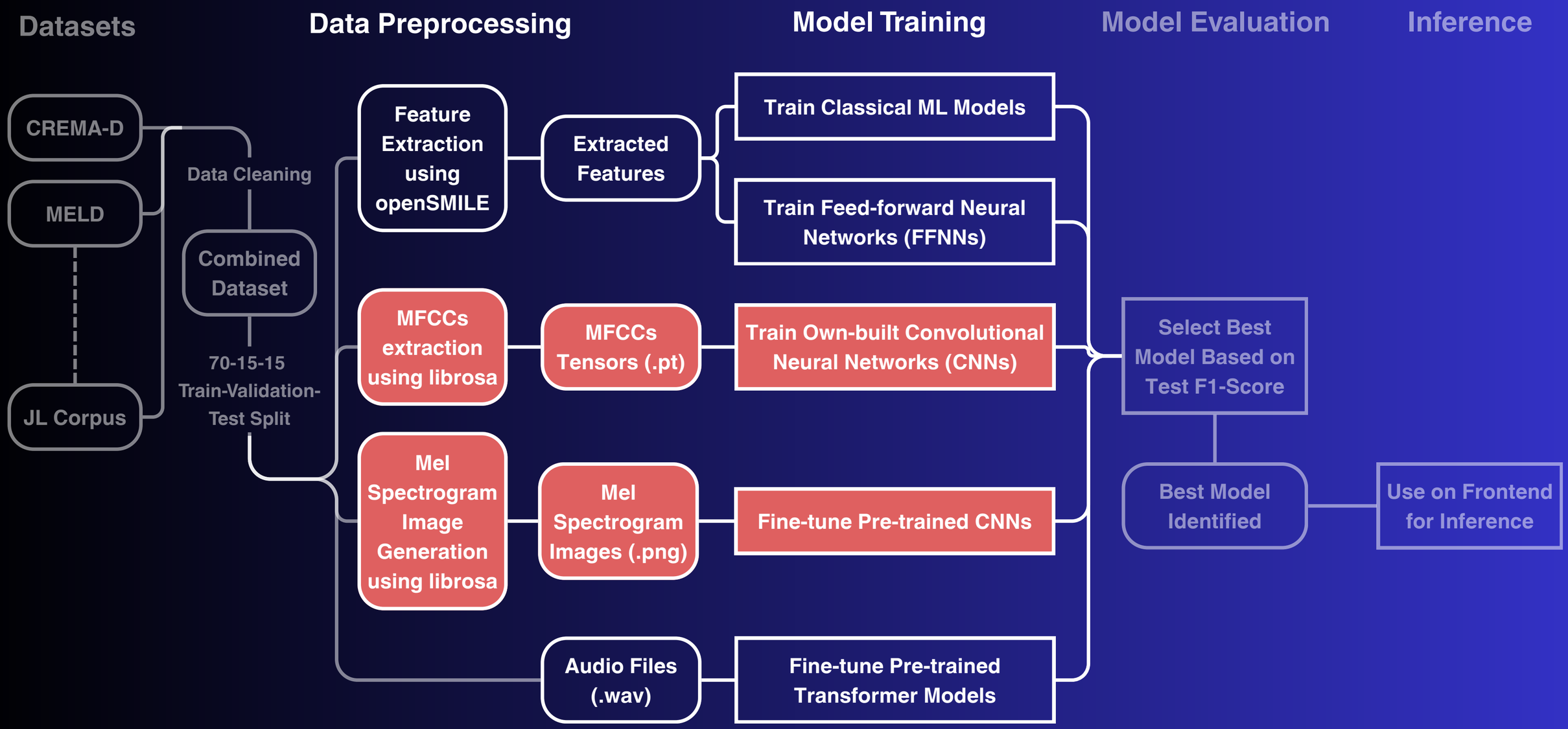
ANNAMALAI

METHODOLOGY



ANNAMALAI

METHODOLOGY



ANNAMALAI

METHODOLOGY

Datasets

CREMA-D

MELD

JL Corpus

Data Preprocessing

Data Cleaning

Combined Dataset

70-15-15 Train-Validation-Test Split

Feature Extraction using openSMILE

MFCCs extraction using librosa

Mel Spectrogram Image Generation using librosa

Audio Files (.wav)

Model Training

Extracted Features

Train Classical ML Models

Train Feed-forward Neural Networks (FFNNs)

MFCCs Tensors (.pt)

Train Own-built Convolutional Neural Networks (CNNs)

Mel Spectrogram Images (.png)

Fine-tune Pre-trained CNNs

Fine-tune Pre-trained Transformer Models

Model Evaluation

Select Best Model Based on Test F1-Score

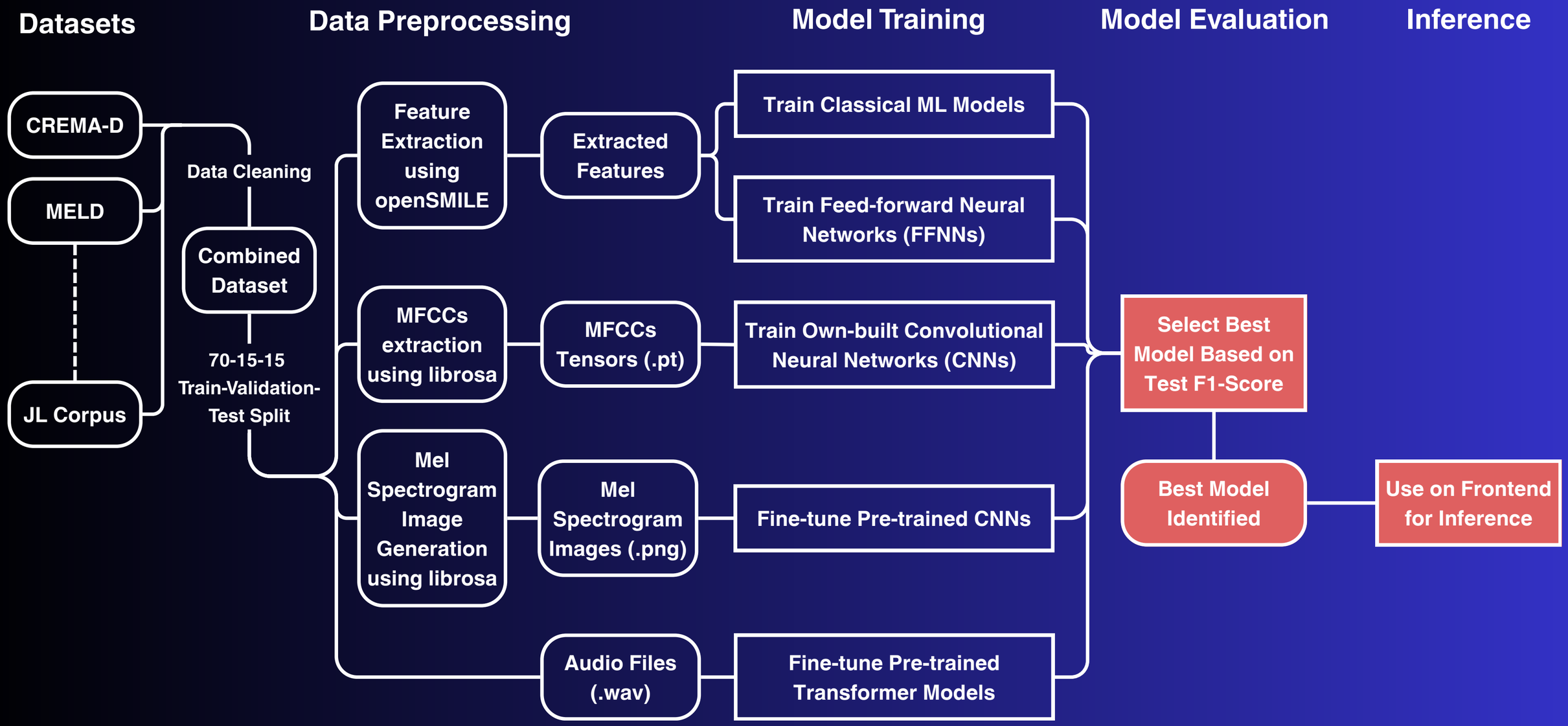
Best Model Identified

Inference

Use on Frontend for Inference

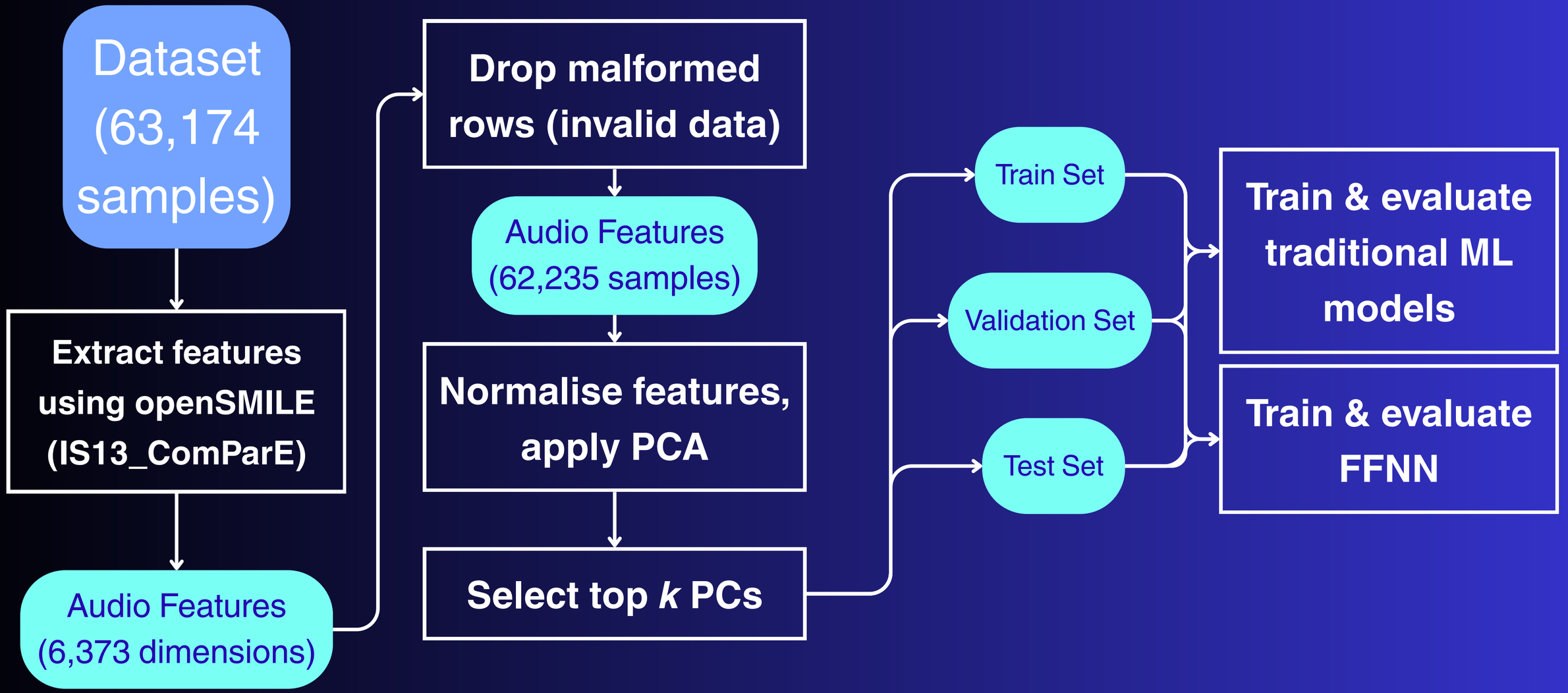
ANNAMALAI

METHODOLOGY



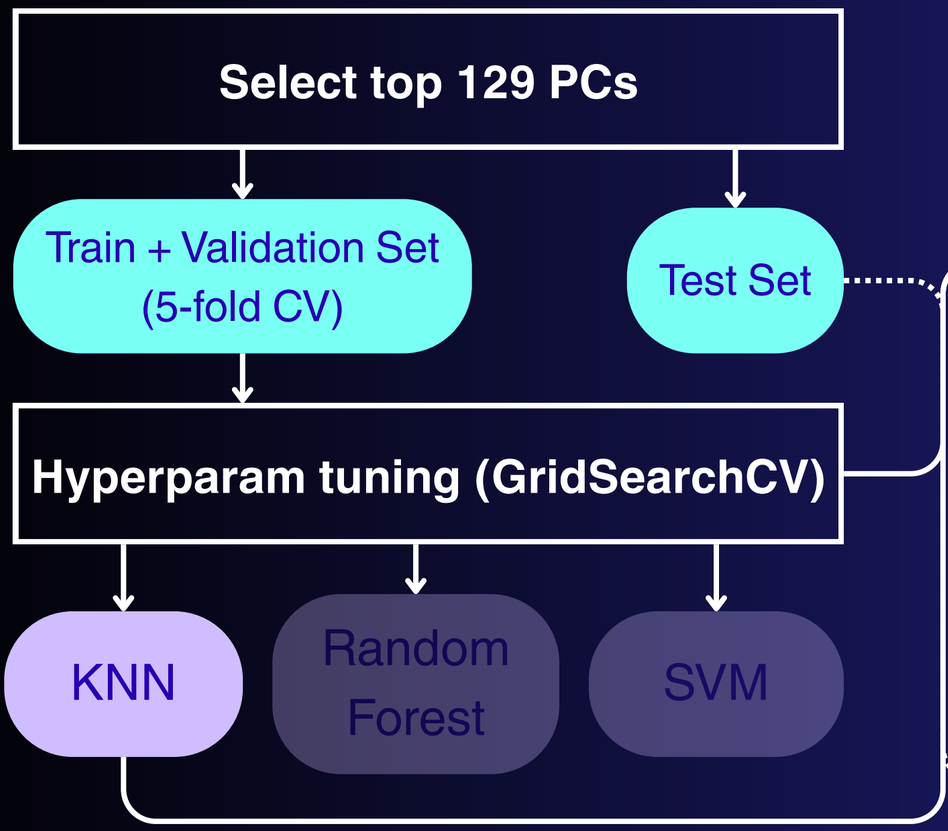
ANNAMALAI

TRADITIONAL ML & FFNN: OPENSMILE FEATURE EXTRACTION



ANNAMALAI

TRADITIONAL ML ALGORITHMS: KNN

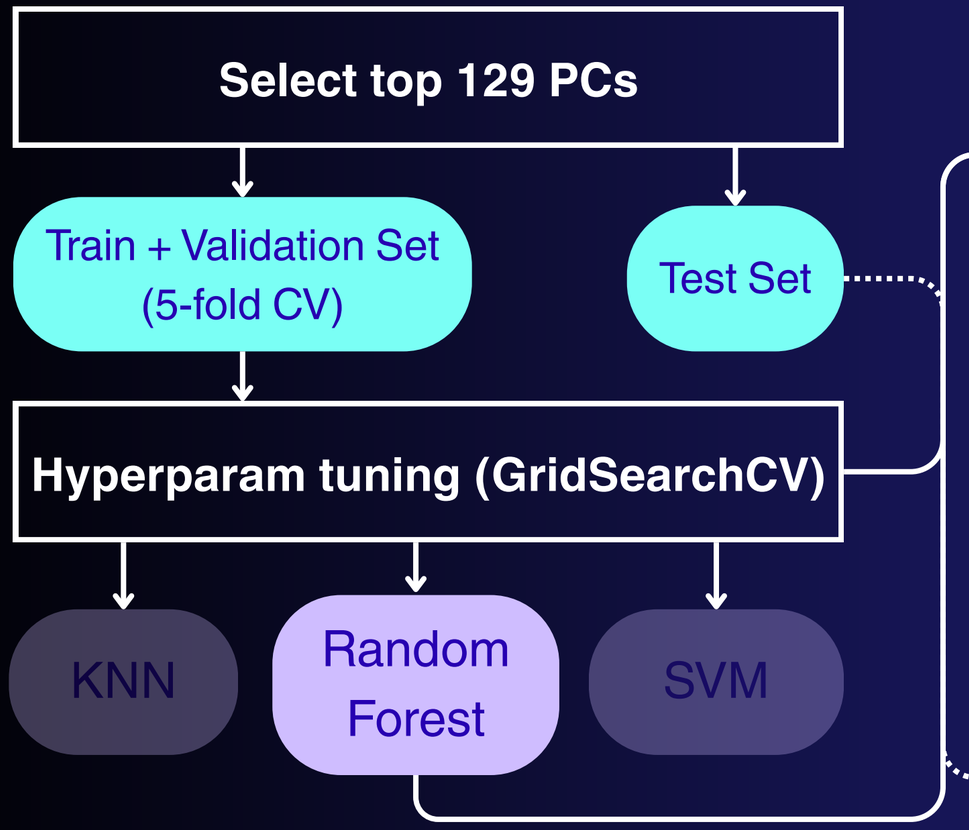


	Tried Values	Best Value
n_neighbors	1-15	15
weights	uniform, distance	distance
metric	euclidean, manhattan	Manhattan

Test Accuracy	**Test Precision**	**Test Recall**	**Test F1-score**
0.6466	0.6503	0.6466	0.6429

ANNAMALAI

TRADITIONAL ML ALGORITHMS: RANDOM FOREST

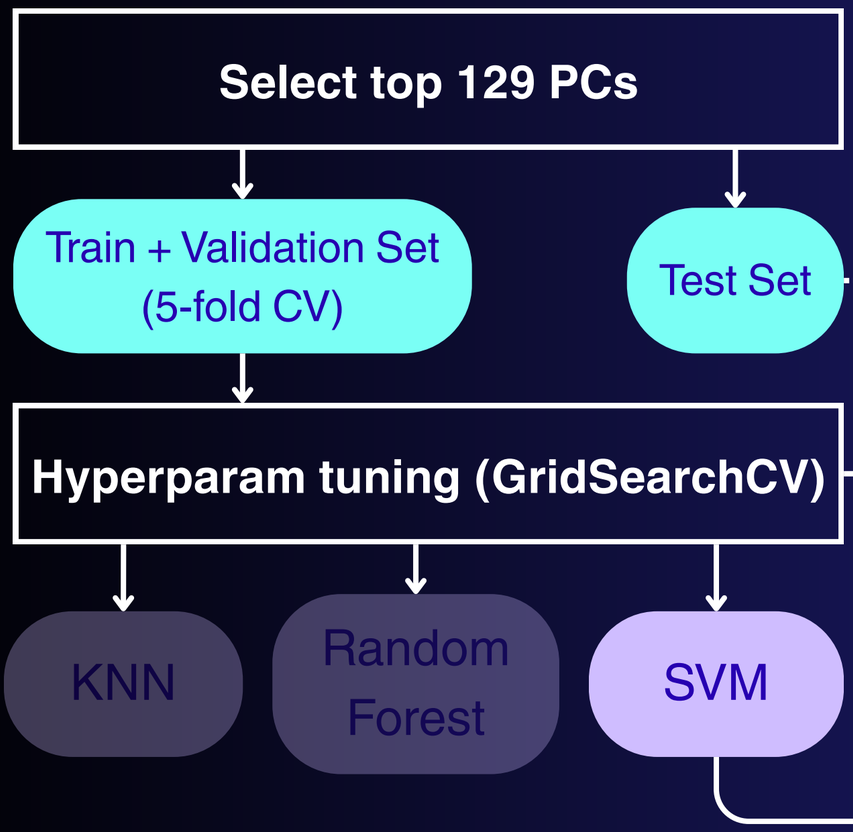


	Tried Values	Best Value
n_estimators	100, 200	200
max_depth	None, 10, 20	None
min_samples_split	2, 5	2
max_features	sqrt, log2	sqrt

Test Accuracy	**Test Precision**	**Test Recall**	**Test F1-score**
0.6443	0.6491	0.6443	0.6402

ANNAMALAI

TRADITIONAL ML ALGORITHMS: SVM

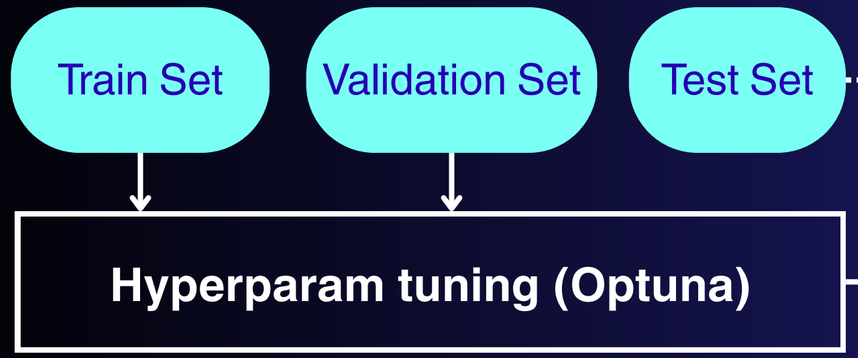


	Tried Values	Best Value
C	0.1, 1, 10	10
kernel	linear, rbf	rbf
gamma	scale, auto, 0.01, 0.1	scale

Test Accuracy	Test Precision	Test Recall	Test F1-score
0.7166	0.7175	0.7166	0.7156

ANNAMALAI

FFNNS

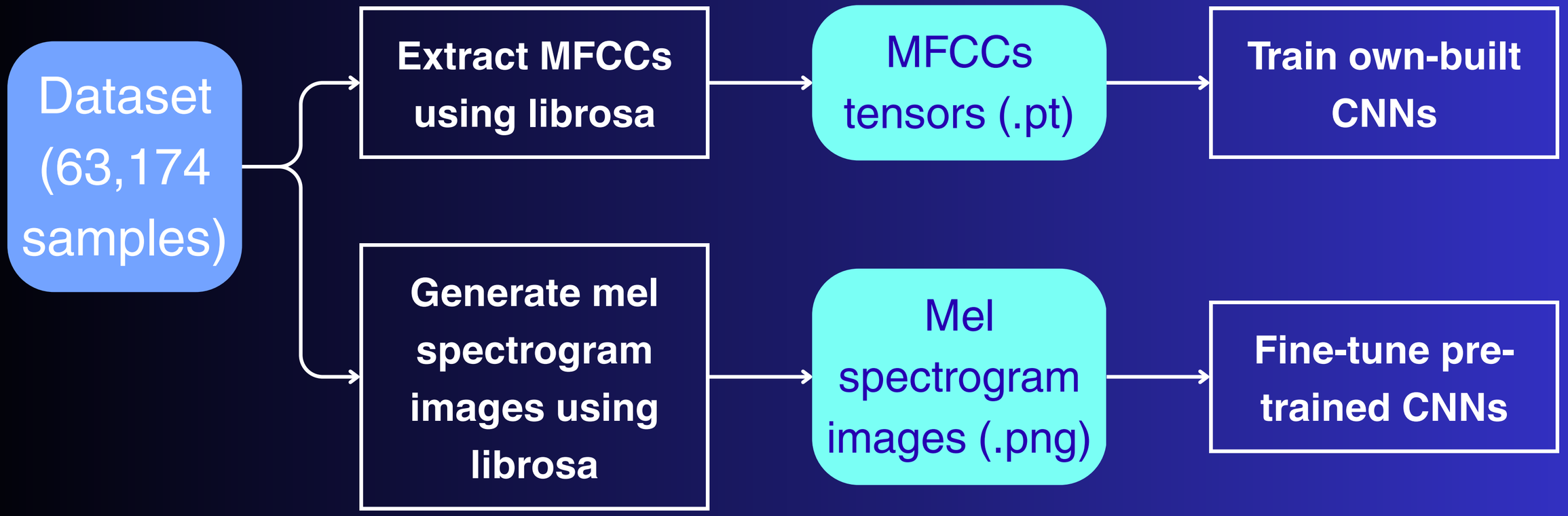


Test Accuracy	Test Precision	Test Recall	Test F1-score
0.7228	0.7223	0.7228	0.7218

	Tried Values	Best Value
Number of principal components	[100, 1000], in steps of 100	600
Hidden dimension size	2^*x*, where *x* is an integer between [4, 9]	512
Dropout	[0.1, 0.5], in steps of 0.1	0.2
Number of layers	1, 2 or 3	1
Learning rate	0.00001, 0.0001 or 0.01	0.0001

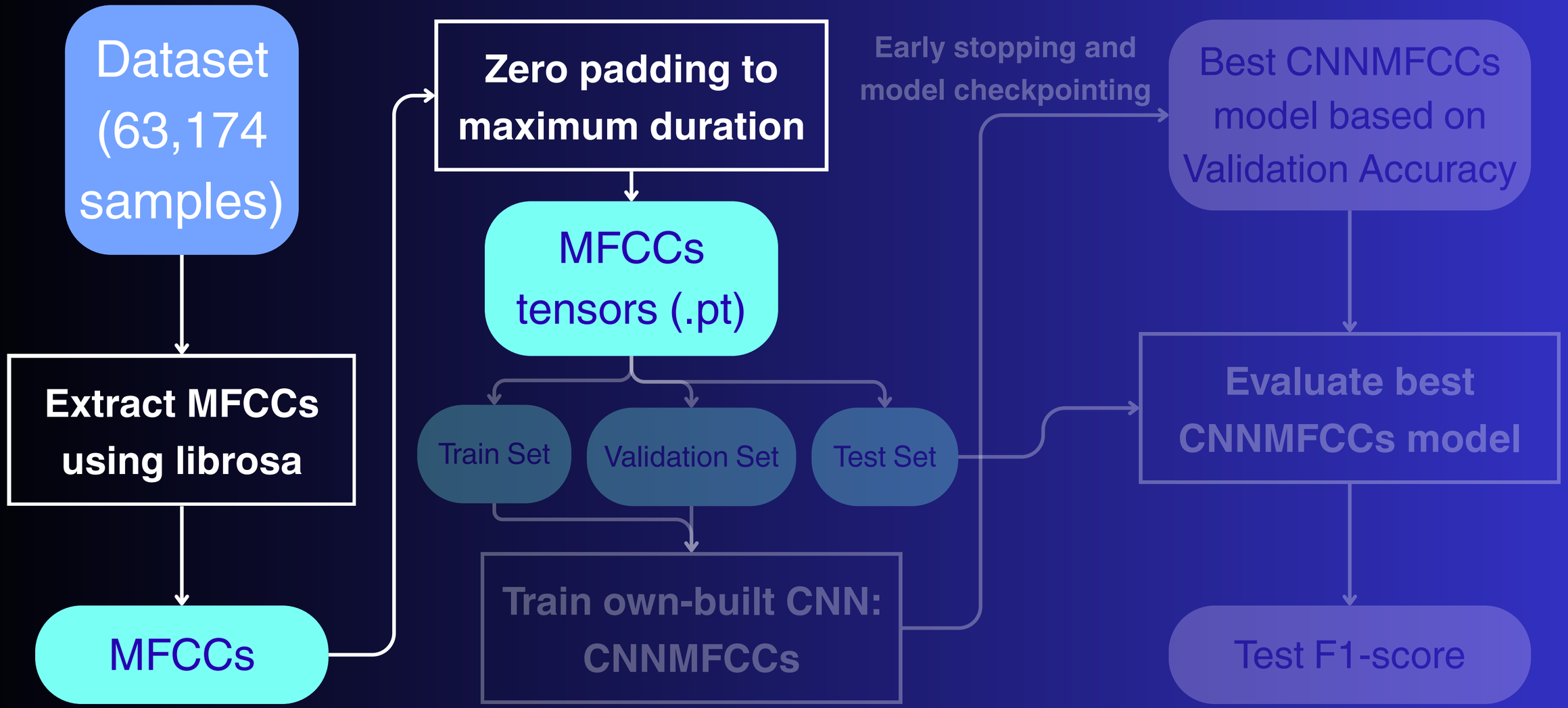
ANNAMALAI

CNNS



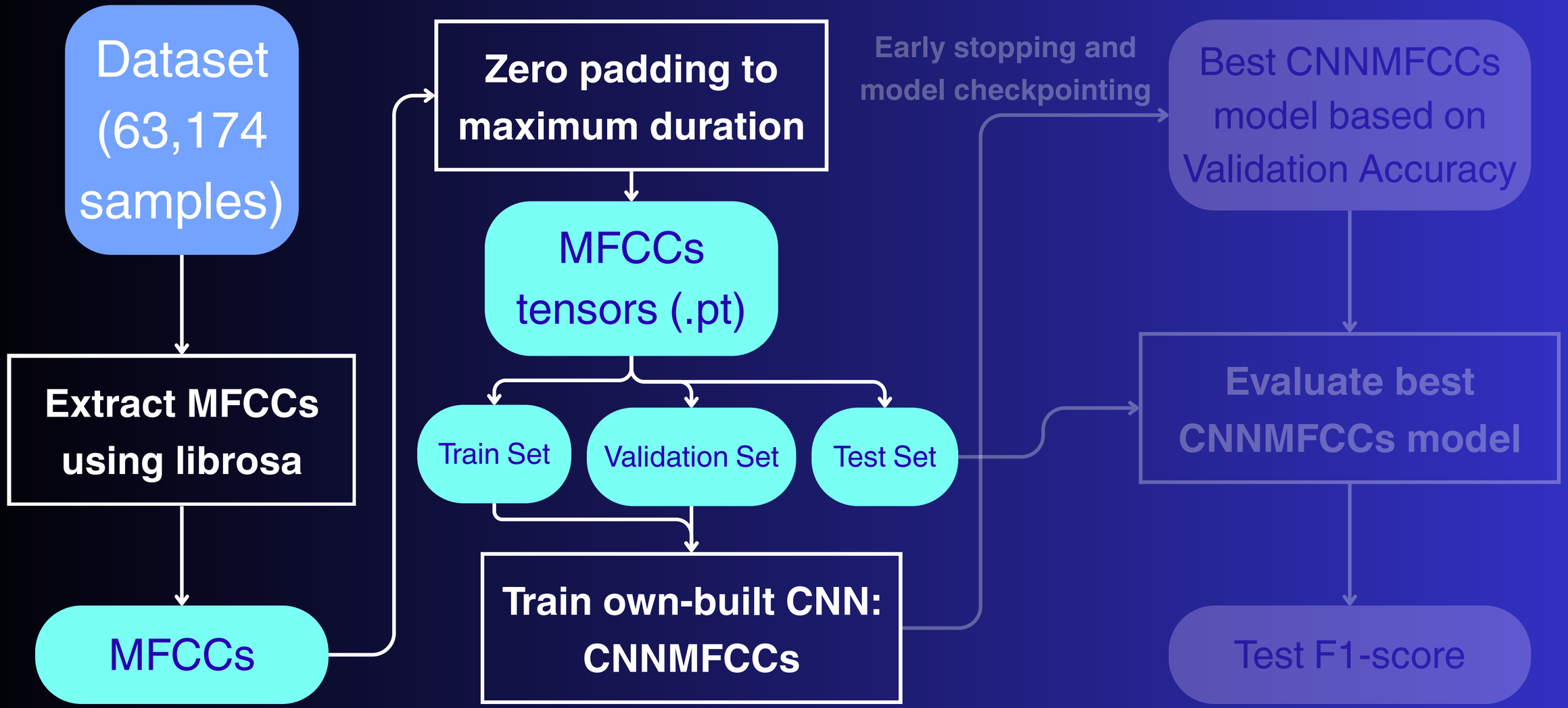
ANNAMALAI

CNNS: OWN-BUILT CNNS



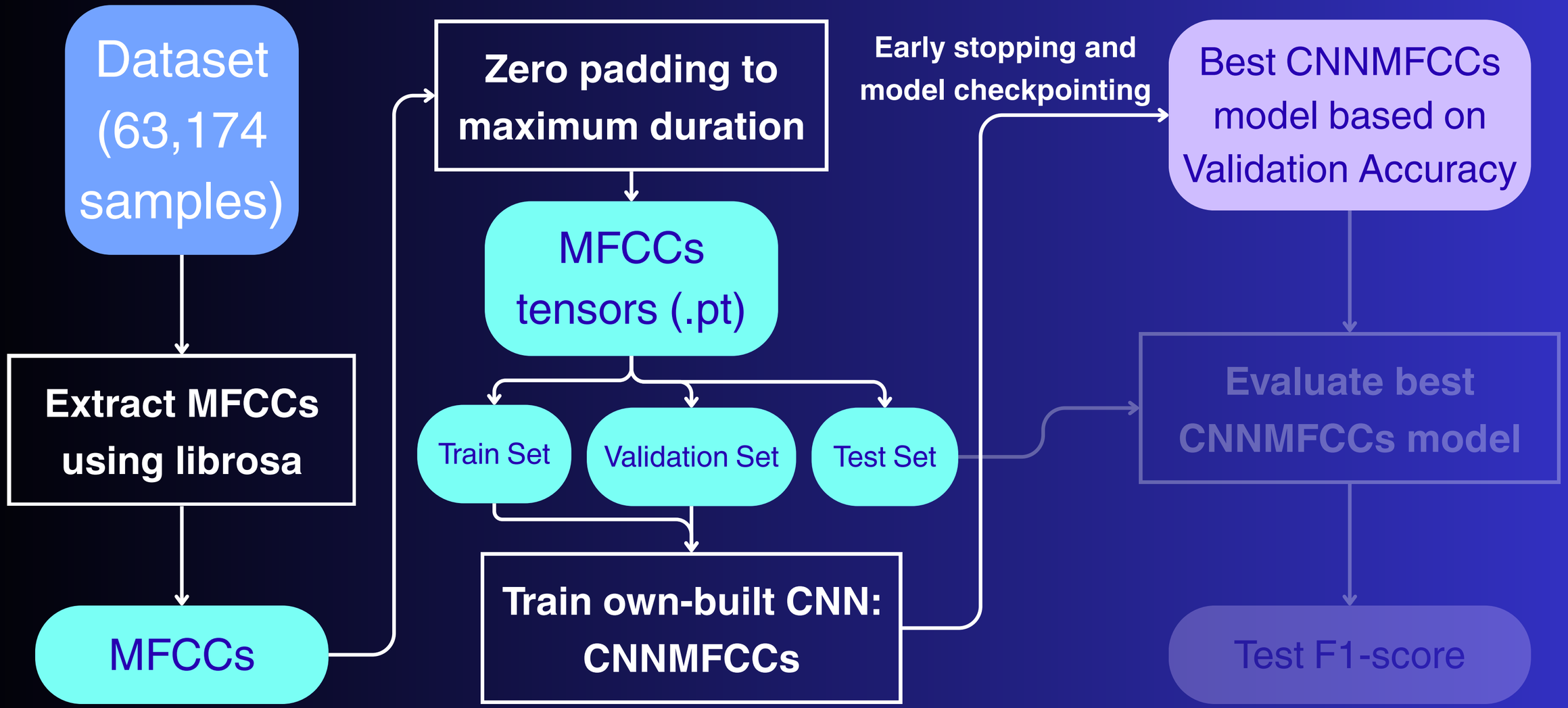
ANNAMALAI

CNNS: OWN-BUILT CNNS



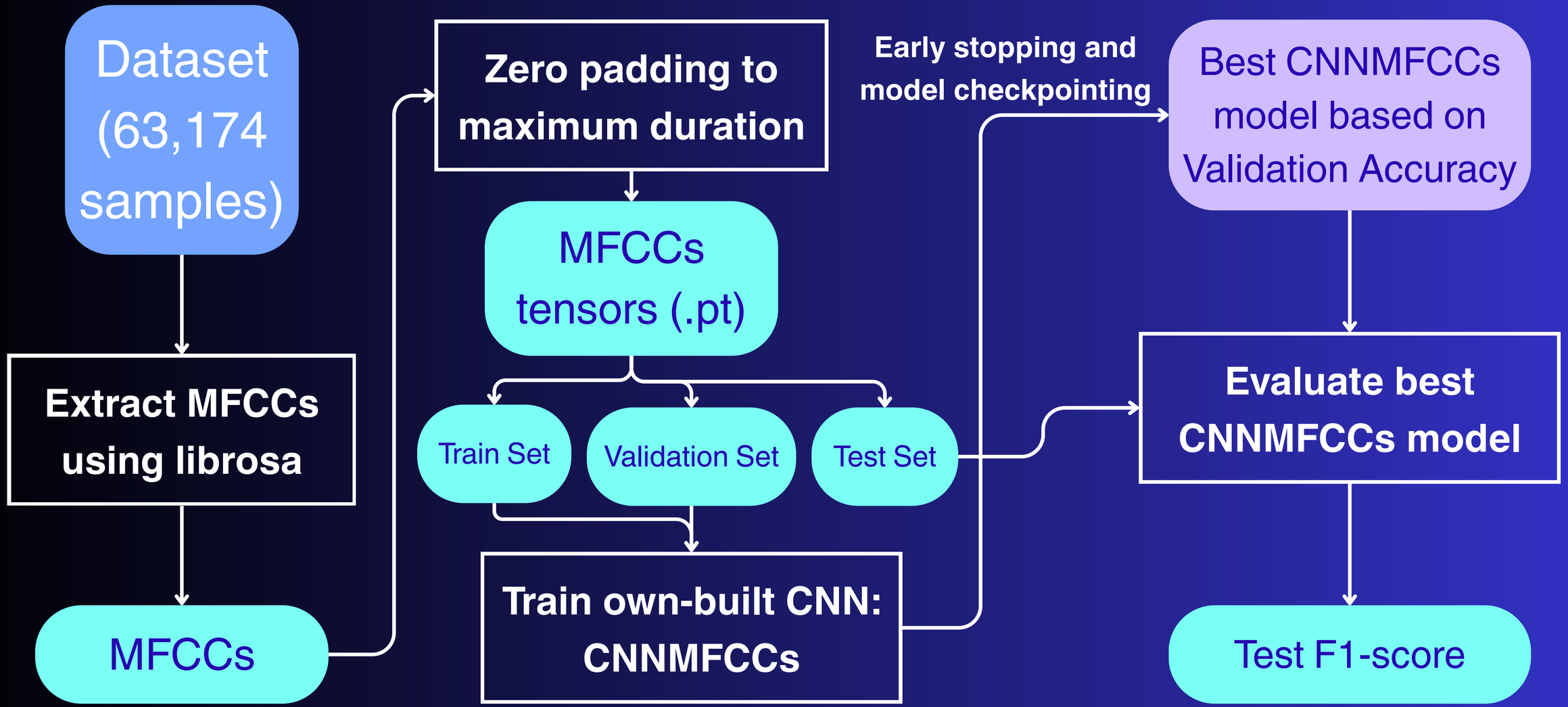
ANNAMALAI

CNNS: OWN-BUILT CNNS



ANNAMALAI

CNNS: OWN-BUILT CNNS



ANNAMALAI

CNNS: OWN-BUILT CNNS

CNNMFCCs

CNNMFCCs2

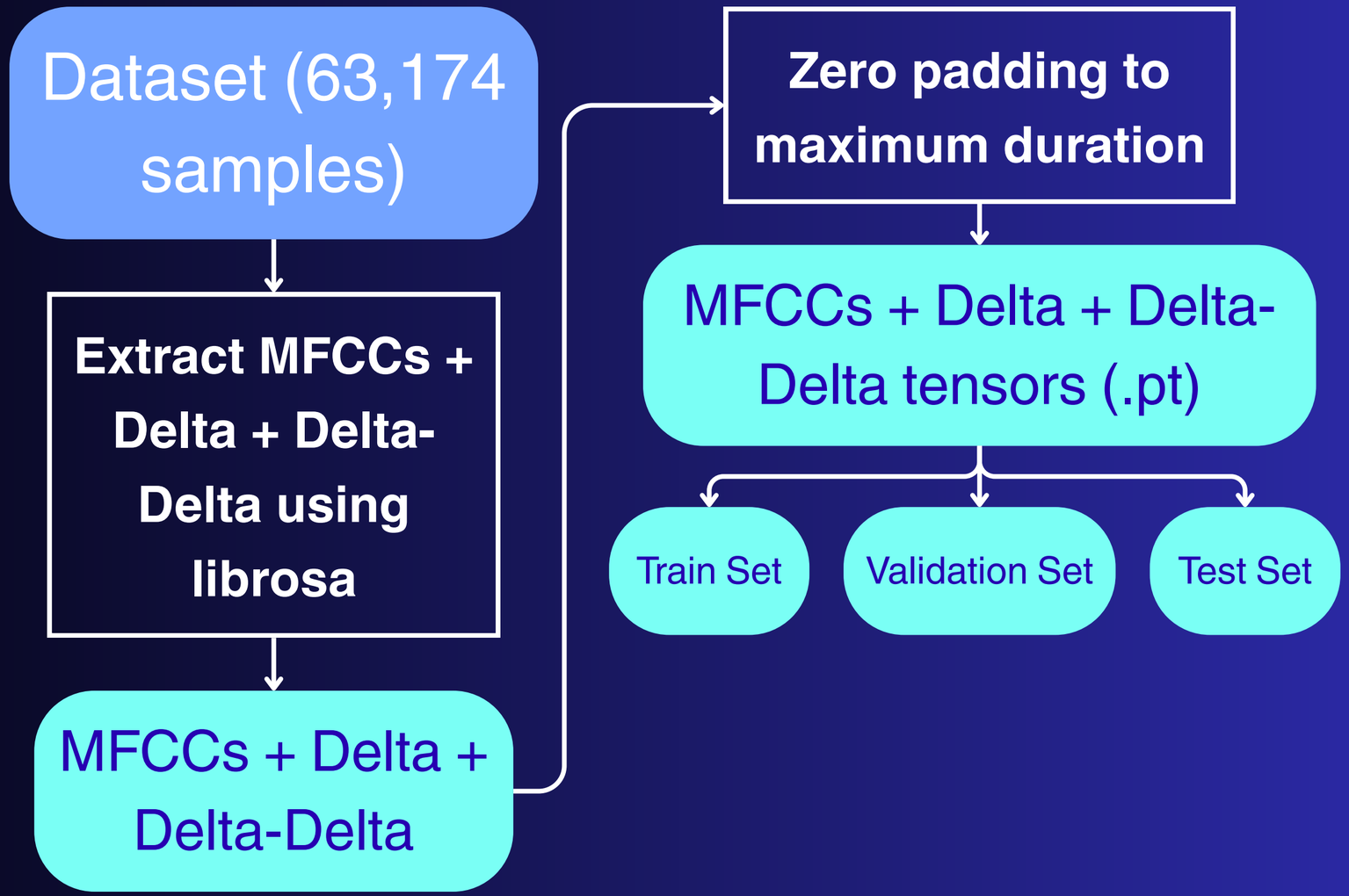
With enhancements like batch normalization, higher dropout

CNNMFCCs3

With different activation function (SiLU instead of ReLU)

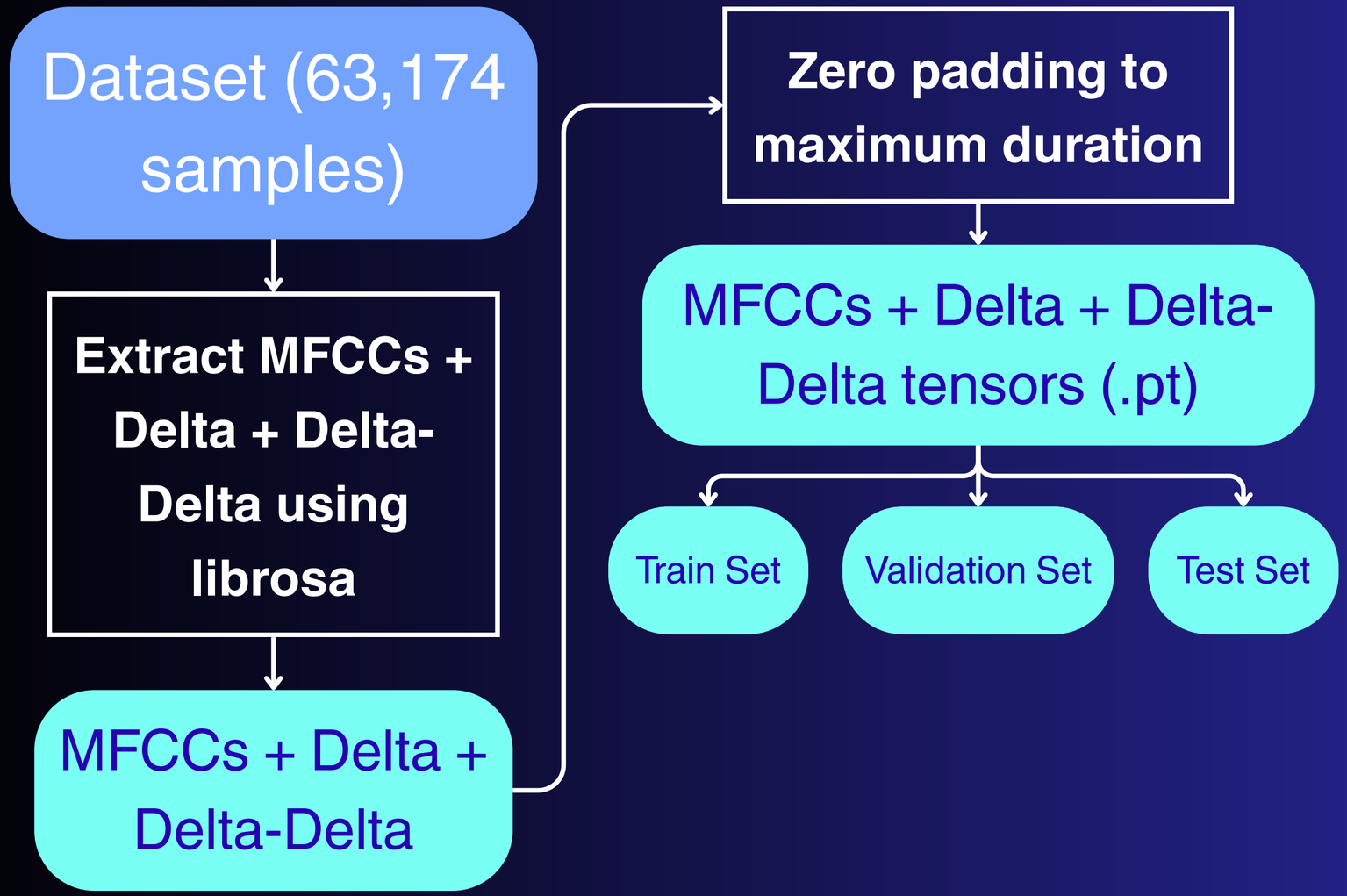
ANNAMALAI

CNNS: OWN-BUILT CNNS



ANNAMALAI

CNNS: OWN-BUILT CNNS



CNNMFCCsDelta

CNNMFCCsDelta2

With enhancements like batch normalization, higher dropout

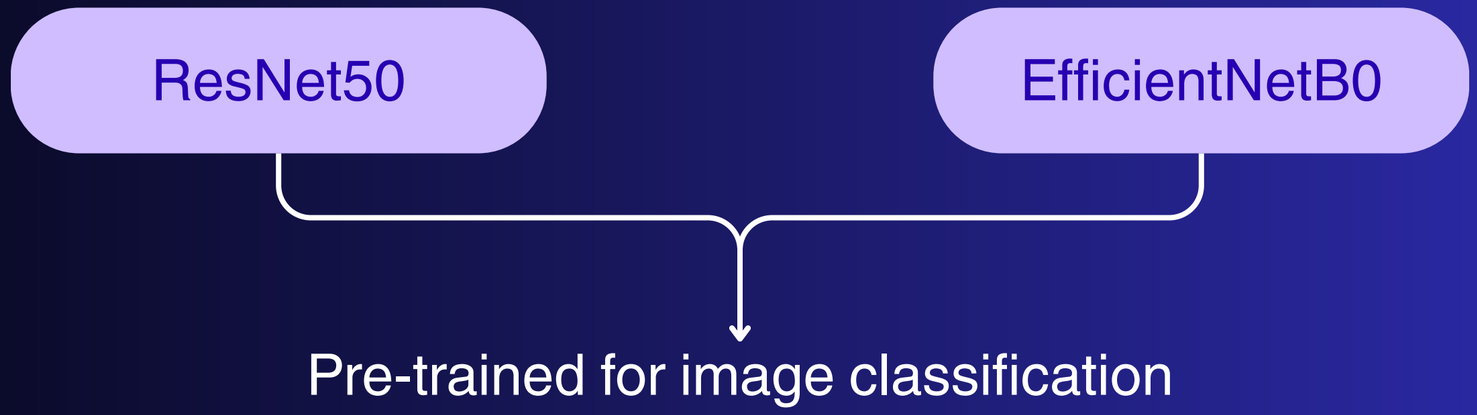
ANNAMALAI

CNNS: OWN-BUILT CNNS

Data Preprocessing	CNN Architecture	Test F1-score
MFCCs, zero padding	CNNMFCCs	0.6223
	CNNMFCCs2	0.7352
	CNNMFCCs3	0.7325
MFCCs + Delta + Delta-Delta, zero padding	CNNMFCCsDelta	0.7154
	CNNMFCCsDelta2	**0.7402**

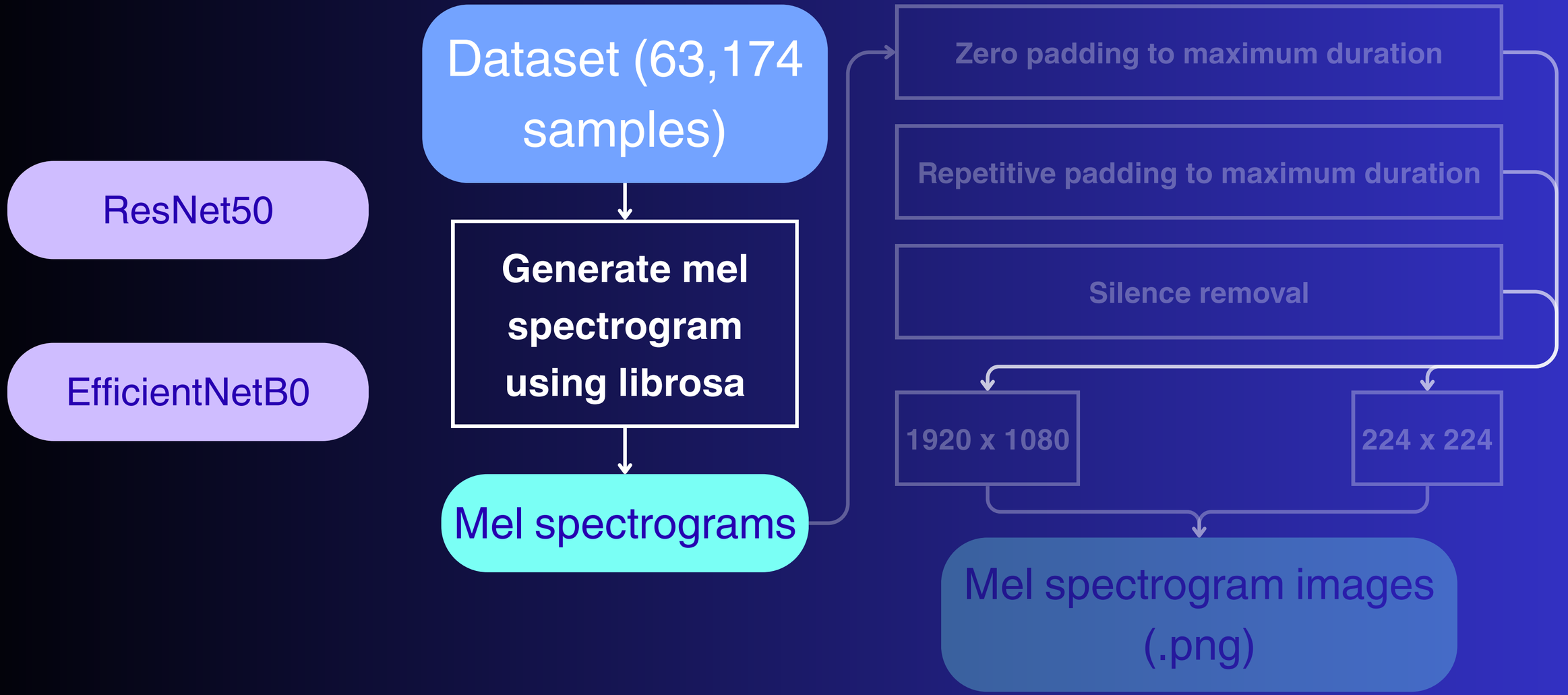
ANNAMALAI

CNNS: PRE-TRAINED CNNS



ANNAMALAI

CNNS: PRE-TRAINED CNNS



ANNAMALAI

CNNS: PRE-TRAINED CNNS

ResNet50

EfficientNetB0

Dataset (63,174 samples)

Generate mel spectrogram using librosa

Mel spectrograms

Zero padding to maximum duration

Repetitive padding to maximum duration

Silence removal

1920 x 1080

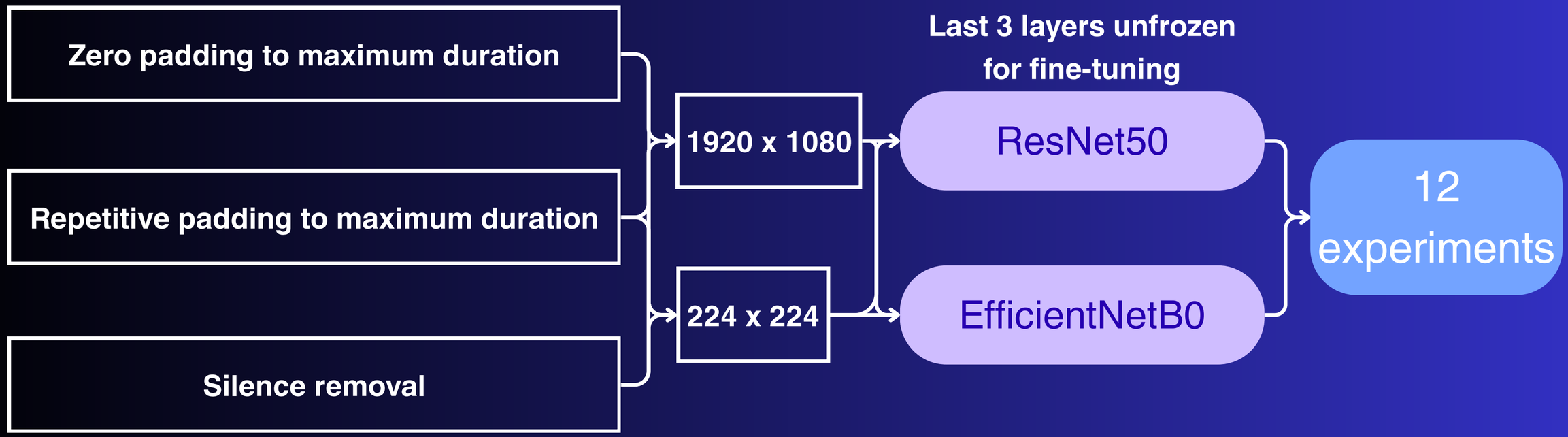
224 x 224

Mel spectrogram images (.png)

ANNAMALAI

CNNS: PRE-TRAINED CNNS

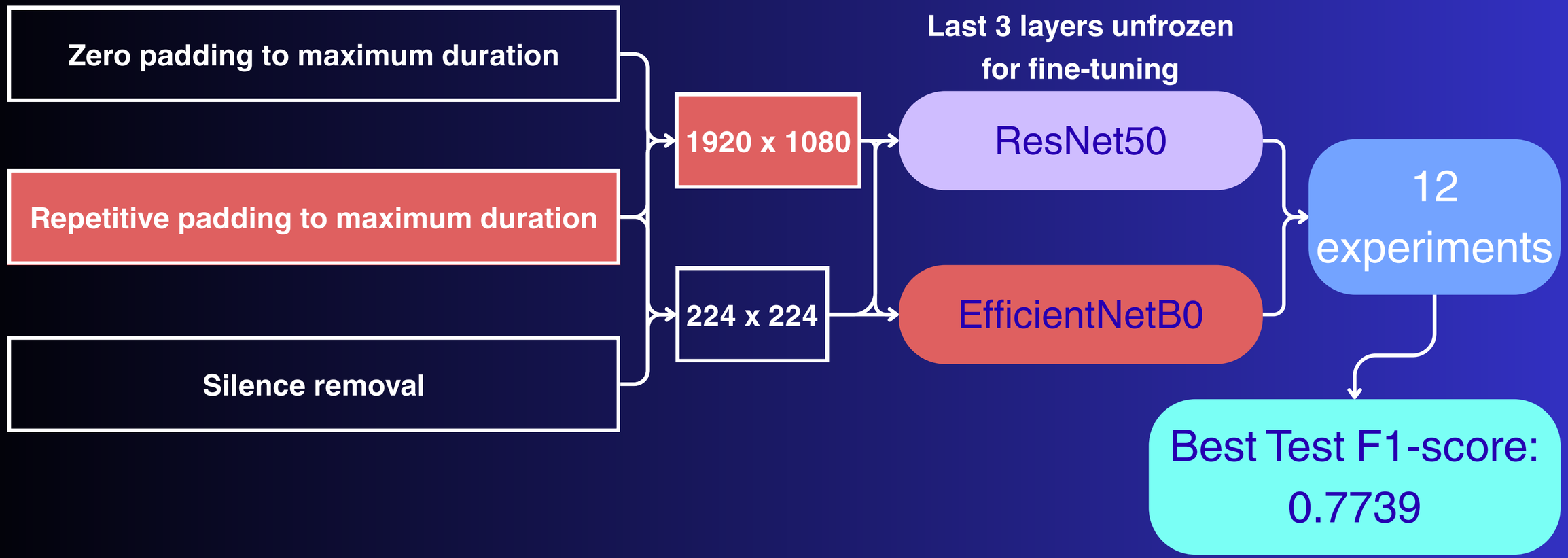
3 data preprocessing techniques x 2 image sizes x 2 models = 12 experiments



ANNAMALAI

CNNS: PRE-TRAINED CNNS

3 data preprocessing techniques x 2 image sizes x 2 models = 12 experiments



ANNAMALAI

CNNS: PRE-TRAINED CNNS

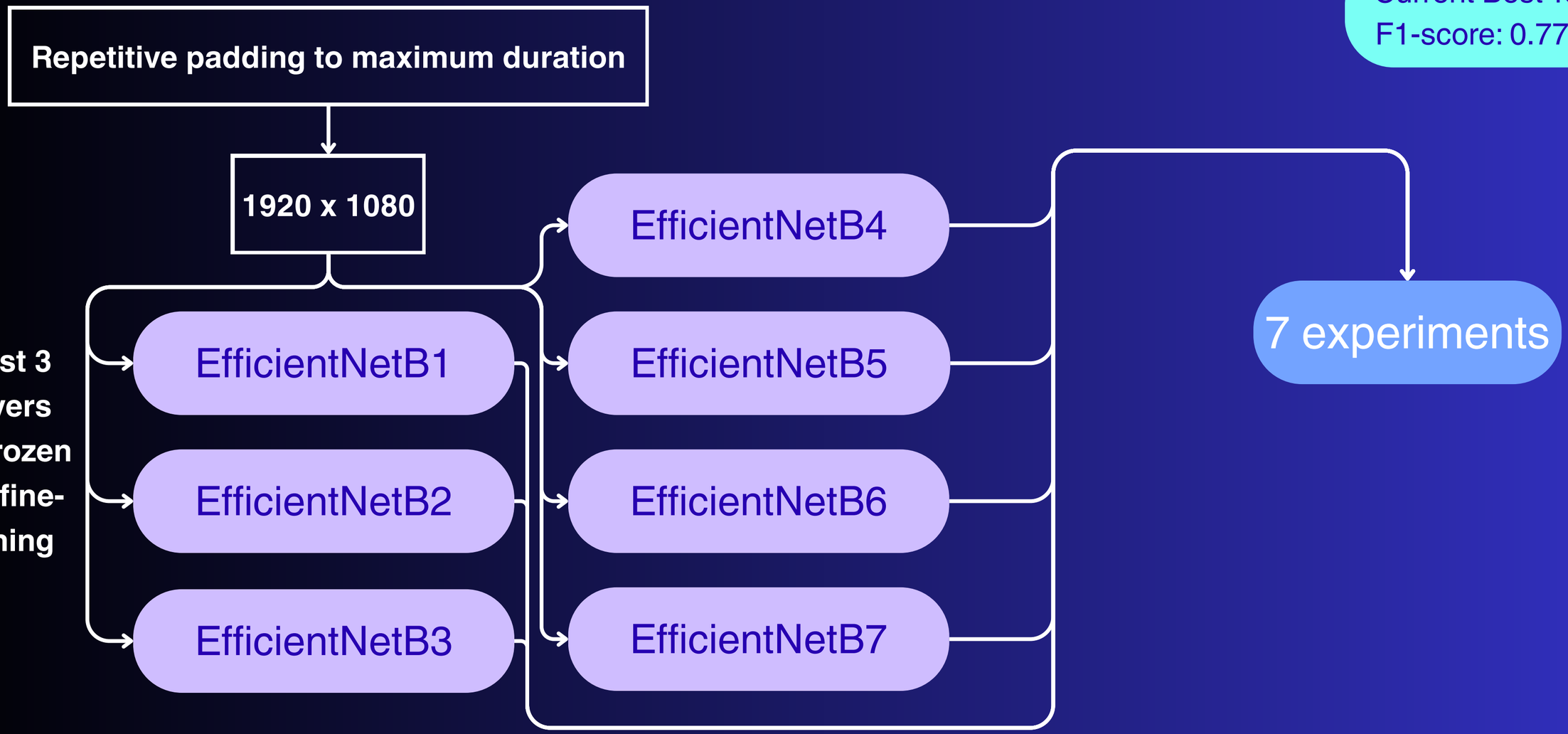
Current Best Test F1-score: 0.7739

Last 3 layers unfrozen for fine-tuning

ANNAMALAI

CNNS: PRE-TRAINED CNNS

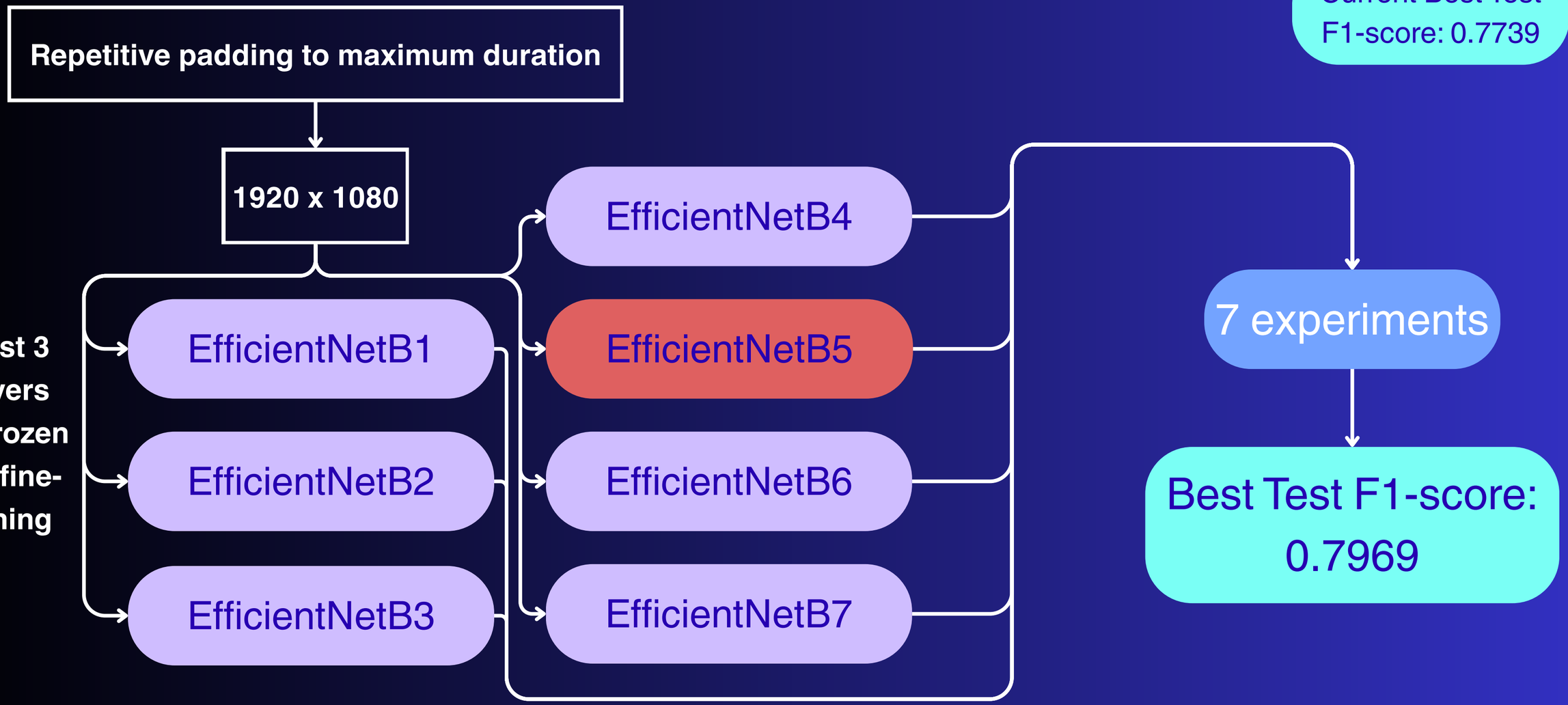
Current Best Test F1-score: 0.7739



ANNAMALAI

CNNS: PRE-TRAINED CNNS

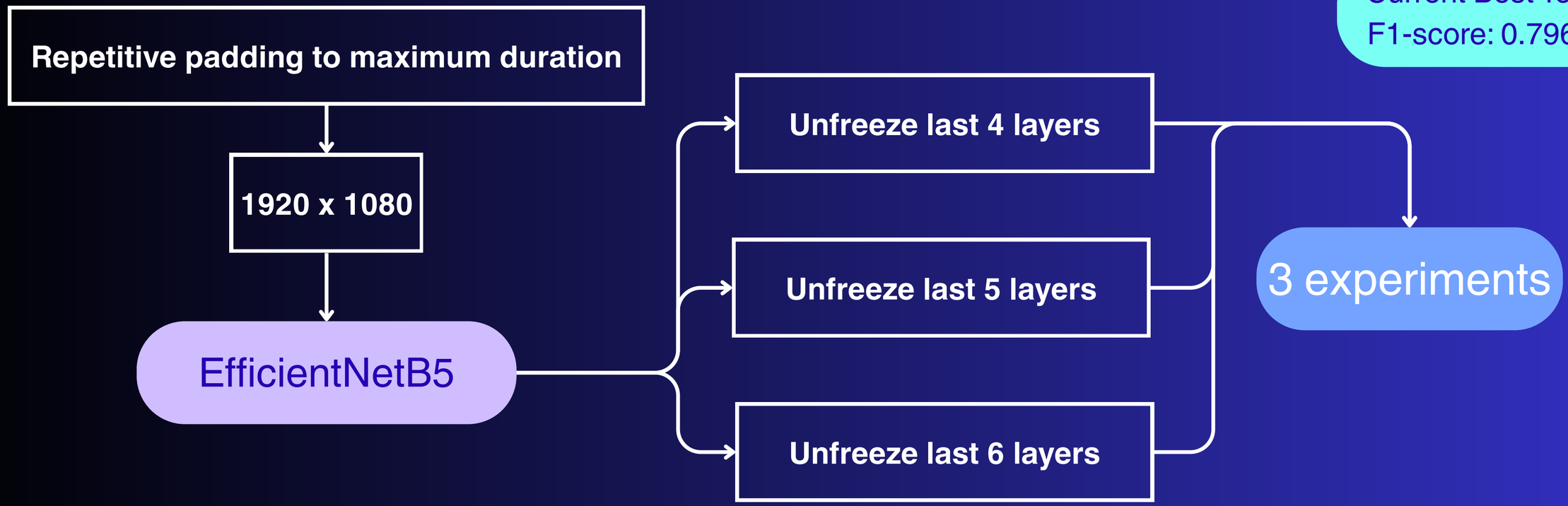
Current Best Test F1-score: 0.7739



ANNAMALAI

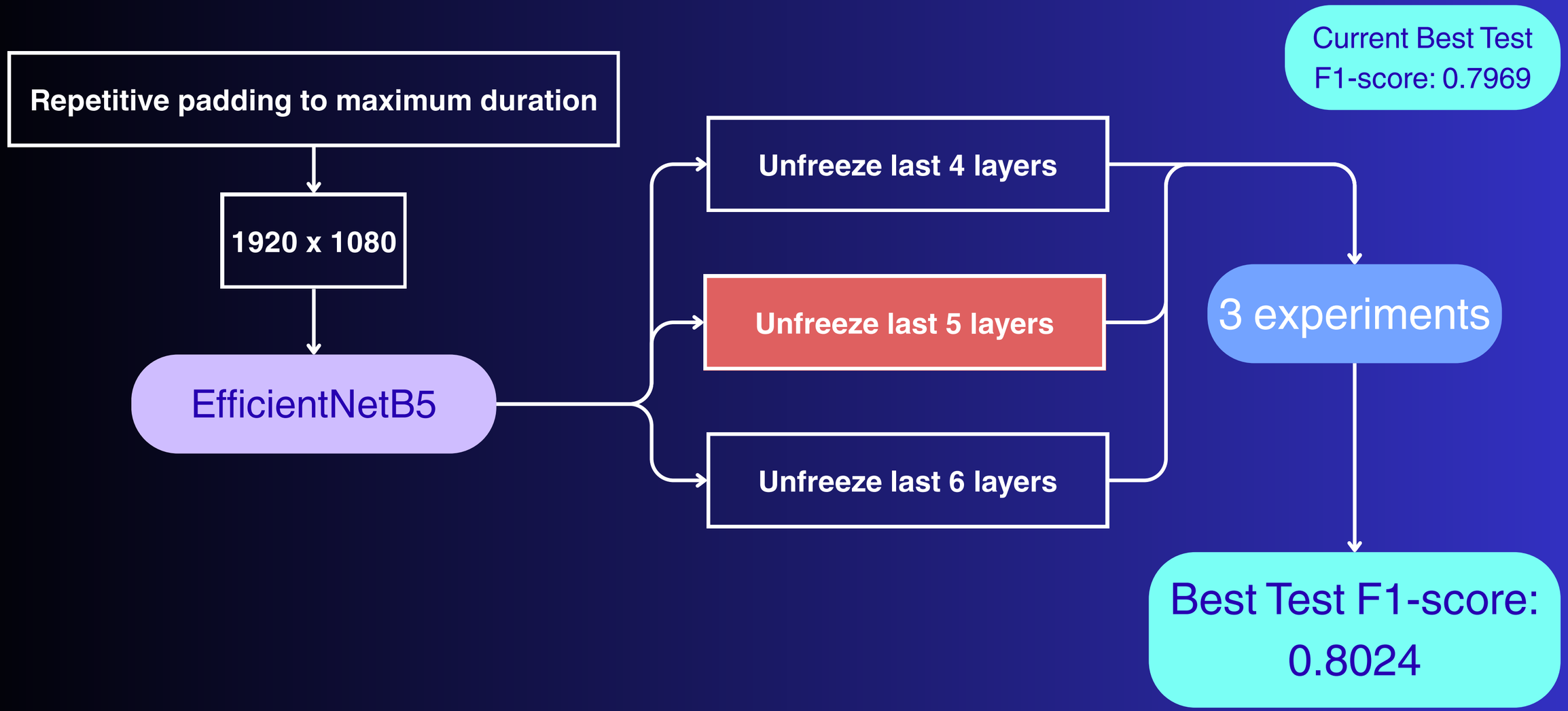
CNNS: PRE-TRAINED CNNS

Current Best Test
F1-score: 0.7969



ANNAMALAI

CNNS: PRE-TRAINED CNNS



ANNAMALAI

CNNS: PRE-TRAINED CNNS

Current Best Test F1-score: 0.8024

Dataset (63,174 samples)

Offline data augmentation:
1) Gaussian noise addition
2) Time stretching
3) Pitch shifting

Augmented dataset (195,837 samples)

Repetitive padding to maximum duration

1920 x 1080

EfficientNetB5

Unfreeze last 5 layers

ANNAMALAI

CNNS: PRE-TRAINED CNNS

Dataset (63,174 samples)

Offline data augmentation:
1) Gaussian noise addition
2) Time stretching
3) Pitch shifting

Augmented dataset (195,837 samples)

Repetitive padding to maximum duration

1920 x 1080

EfficientNetB5

Unfreeze last 5 layers

Current Best Test F1-score: 0.8024

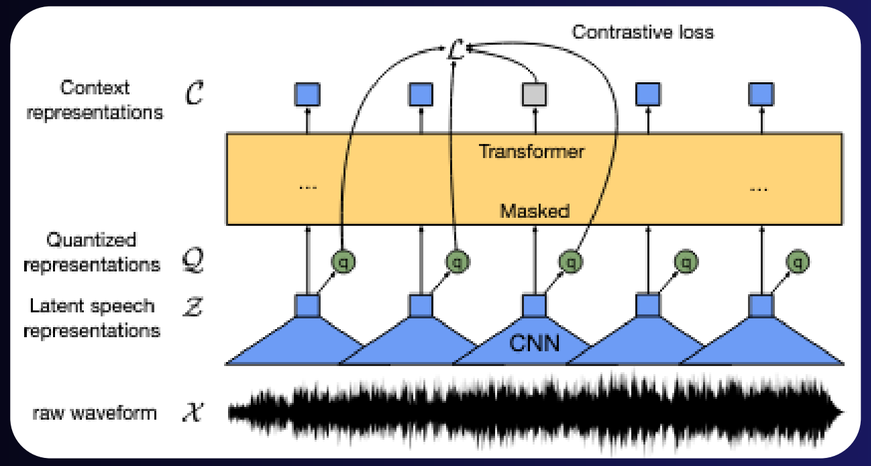
Best Test F1-score: 0.8260

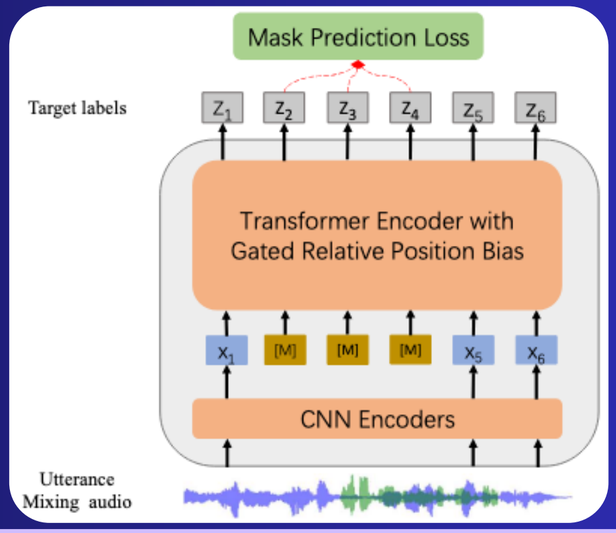
ANNAMALAI

TRANSFORMER MODELS

ANNAMALAI

TRANSFORMER MODELS





Wav2Vec2 developed by Facebook

WavLM developed by Microsoft

Speech models that accept raw speech signals as inputs

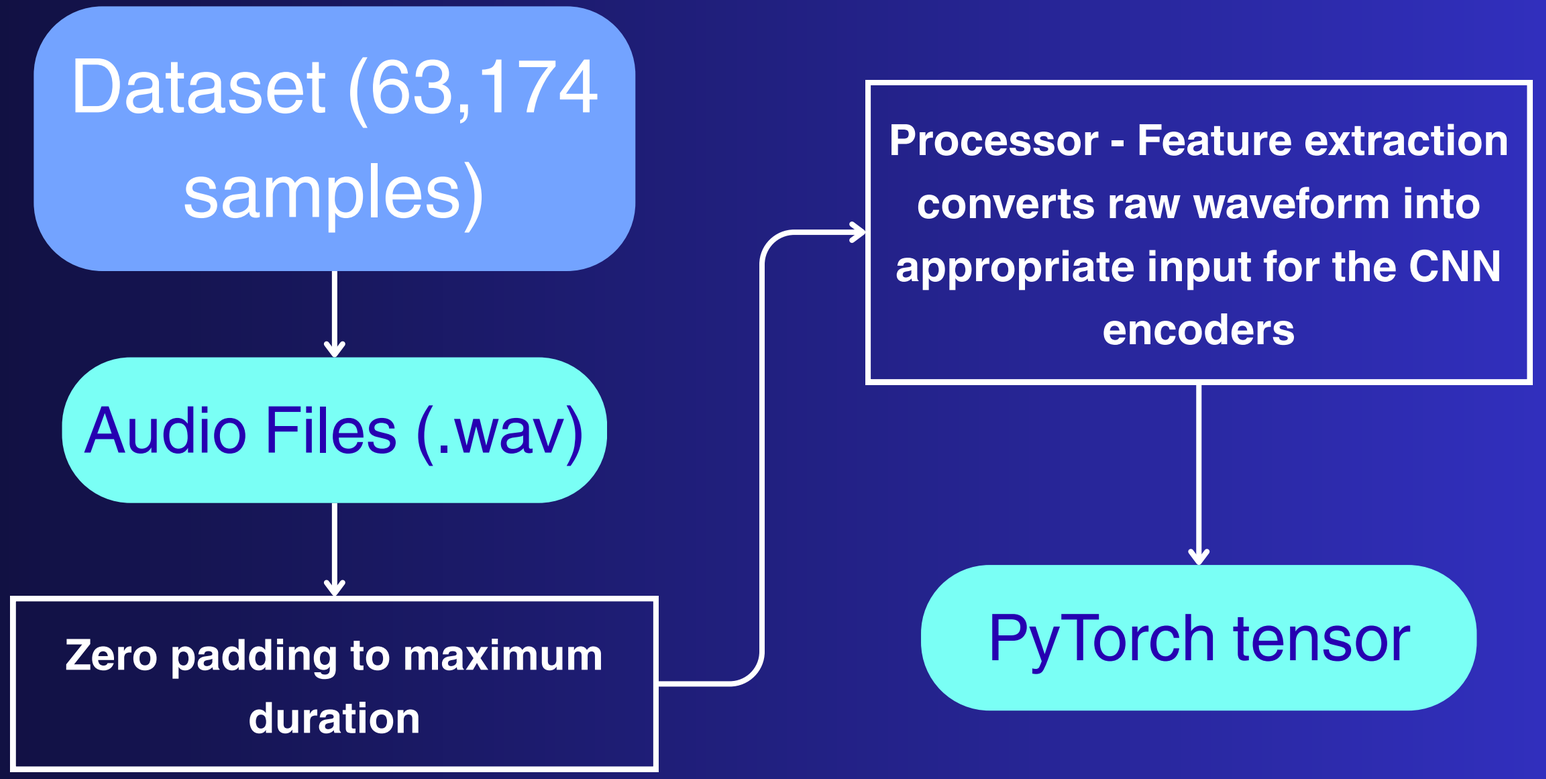
In our work, we use the base models, i.e wav2vec2-base and wavlm-base

ANNAMALAI

TRANSFORMER MODELS

Wav2Vec2

WavLM



ANNAMALAI

TRANSFORMER MODELS

Model	Test Accuracy	Test Precision	Test Recall	Test F1-score
Wav2Vec2	0.8538	0.8578	0.8358	0.8540
WavLM	0.8451	0.8492	0.8451	0.8449

ANNAMALAI

TRANSFORMER MODELS

Wav2Vec2

WavLM

Dataset (63,174 samples)

Offline data augmentation:
1) Gaussian noise addition
2) Time stretching
3) Pitch shifting

Augmented dataset (195,837 samples)

Audio Files (.wav)

Zero padding to maximum duration

Processor - Feature extraction converts raw waveform into appropriate input for the CNN encoders

PyTorch tensor

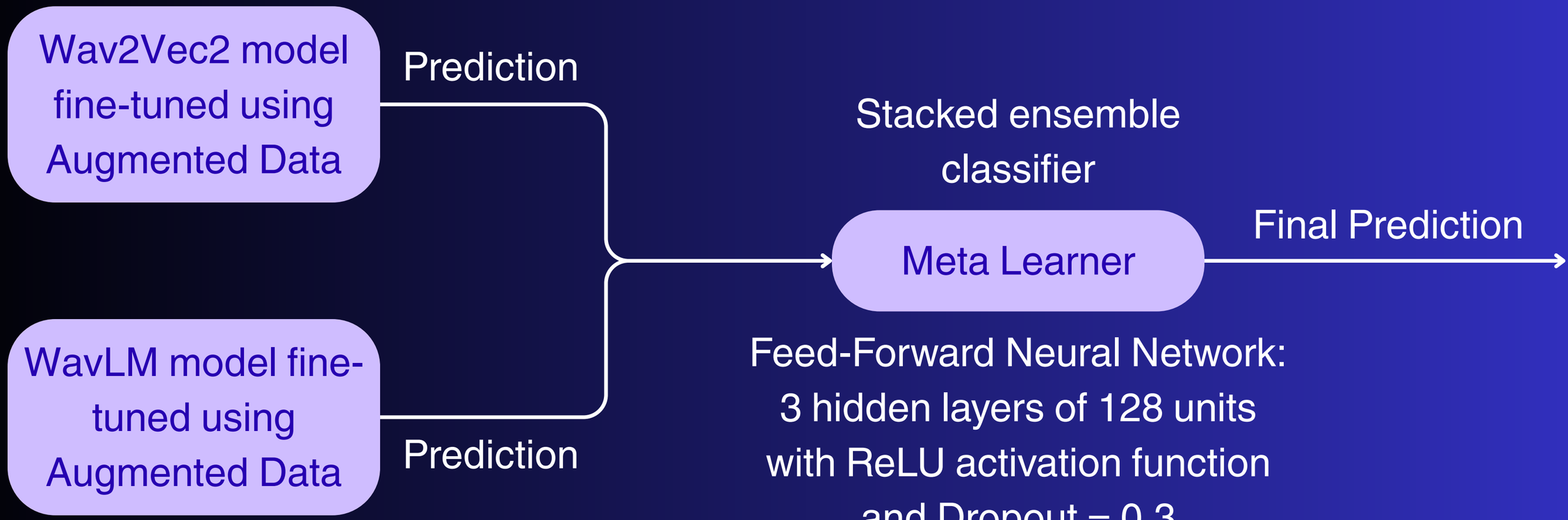
ANNAMALAI

TRANSFORMER MODELS

Model	Test Accuracy	Test Precision	Test Recall	Test F1-score
Wav2Vec2	0.8400	0.8469	0.8400	0.8407
WavLM	0.8566	0.8572	0.8566	0.8555

ANNAMALAI

TRANSFORMER MODELS



ANNAMALAI

TRANSFORMER MODELS

Model	Test Accuracy	Test Precision	Test Recall	Test F1-score
Wav2Vec2	0.8400	0.8469	0.8400	0.8407
WavLM	0.8566	0.8572	0.8566	0.8555
Ensemble	0.8611	0.8729	0.8611	**0.8641**

ANNAMALAI

DEMO

The theme of our demonstration revolves around SDG 16

SINGAPORE POLICE FORCE
SAFEGUARDING EVERY DAY

ANNAMALAI

DEMO

ANNAMALAI

THANK YOU

DO YOU HAVE ANY QUESTIONS FOR US?