Name : Aditi Baheti

Dataset :

- The ESC-50 dataset, of environmental audio clips across 50 categories, was used for our experiments.
- Specifically, we focused on the ESC-10 subset—marked by the ESC-10 flag—which narrows down the dataset to 10 categories.

Architecture 1:

- The **Sound Classifier** utilizes a series of convolutional layers that progressively increase in output channel size (from 16 to 64), each followed by max pooling to analyze audio signals, extracting key features essential for classification tasks.
- It employs layers of convolutional and pooling operations to capture and simplify the audio data's complex patterns, enhancing the focus on relevant audio characteristics.

1. Original Model: The initial setup employed convolutional layers for feature extraction from audio signals, without incorporating dropout, early stopping, or regularization techniques.
2. Enhanced Model: To address overfitting and improve generalization, the model was subsequently modified to include dropout for regularization, early stopping to prevent further training once validation performance plateaued, and additional regularization methods.

In our model, we employed a **k-fold cross-validation** approach where fold 1 was consistently reserved for validation across all model variations.
The remaining folds—2, 3, 4, and 5—were utilized for both training and testing phases.

Results with architecture 1 :

Key Observations with Hyperparameter Tuning:

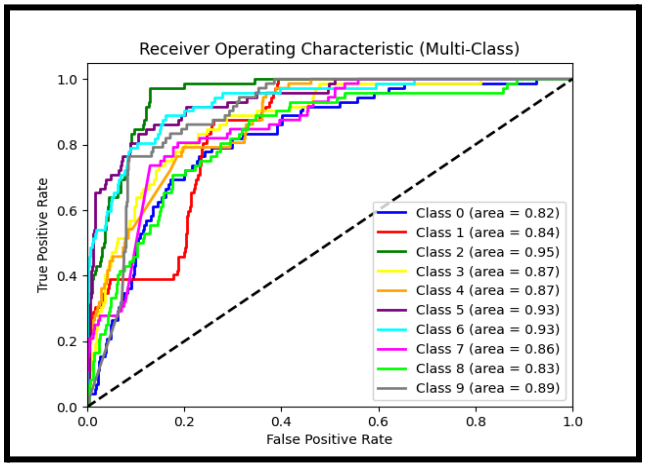
- For both training and validation phases, the "After Improvements" model (Model 2) consistently shows a strategic **reduction in overfitting**, as evidenced by lower test losses and higher test accuracies, F1 scores, and ROC AUC values across all folds when compared to the "Before Improvements" model (Model 1).

The table below shows the results for all the four folds with both the models.

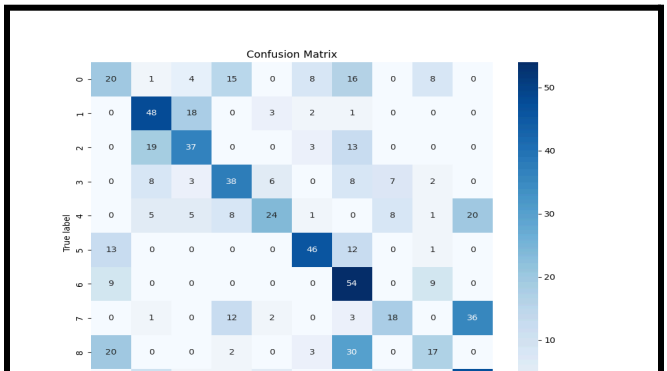
Metric	Fold 2 Original	Fold 2 Enhanced	Fold 3 Original	Fold 3 Enhanced	Fold 4 Original	Fold 4 Enhanced	Fold 5 Original	Fold 5 Enhanced

Avg. Training Accuracy	63.49%	51.15%	68.77%	43.67%	68.64%	55.49%	67.55%	49.84%
Avg. Validation Accuracy	34.25%	38.42%	39.17%	35.00%	38.54%	40.64%	38.44%	43.97%
Test Loss	7.7658	1.5196	4.4111	1.4217	4.1973	1.5789	4.4777	1.5499
Test Accuracy	30.42%	48.47%	42.08%	45.97%	40.56%	43.47%	45.69%	42.92%
F1 Score	0.26	0.47	0.40	0.45	0.41	0.42	0.44	0.39
ROC AUC	0.72	0.88	0.81	0.88	0.78	0.87	0.80	0.86

The Confusion Matrix and ROC curve for our best performing (Fold 2) is displayed:



The ROC curves indicate that the classifier performs well in distinguishing between most of the classes, with Class 2 having the highest area under the curve (AUC = 0.95), suggesting excellent classification performance for this class.
However, there is variability in the AUC values across classes, indicating some classes are more challenging to distinguish than others.

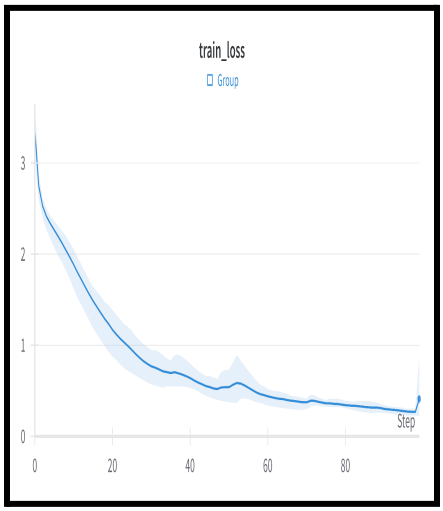


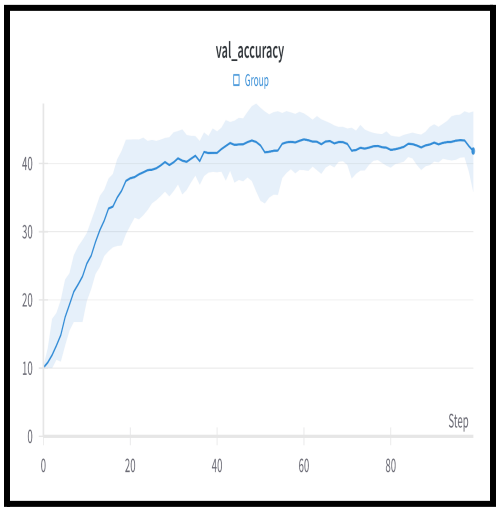
The confusion matrix shows that while many classes are predicted accurately (high values on the diagonal), there are notable instances

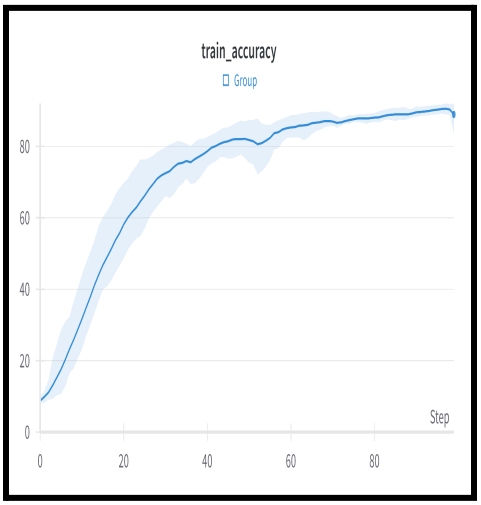
of misclassification, as seen by the off-diagonal numbers.
For instance, Class 1 appears to be frequently confused with Class 0, and Class 6 with Class 8, suggesting these pairs of classes may have similar features or there are inadequacies in the model.

For our **original CNN model**, after **averaging out the metrics across all folds** using Wandb.ai here are the results obtained .

- The model shows a steady decrease in both validation and training loss over time, indicating learning and model improvement with each epoch.
- However, there is a noticeable discrepancy between training and validation loss, suggesting **potential overfitting**.
- While training accuracy improves and stabilizes at a high level, **the validation accuracy plateaus much lower**, reinforcing the possibility of overfitting and highlighting the need for further model tuning to enhance generalization to unseen data.



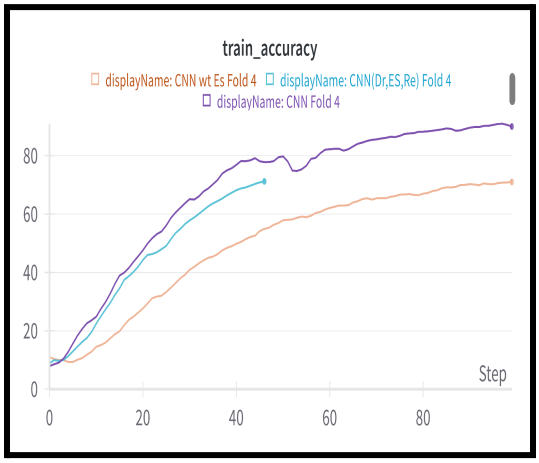


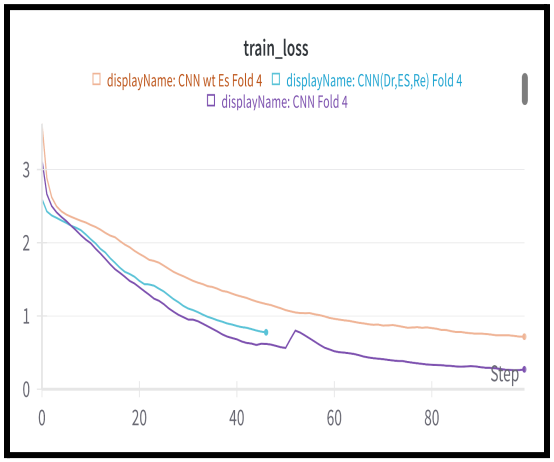


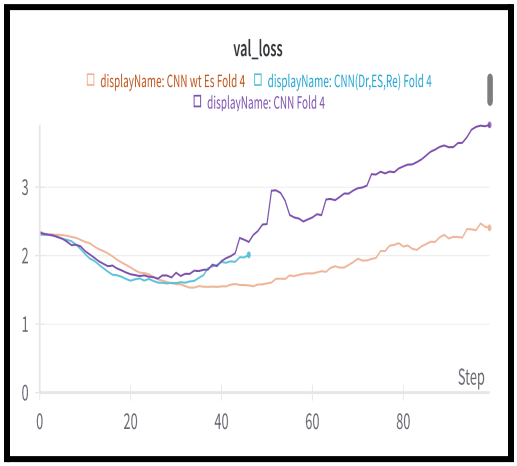
Additionally , some more experimentations were done mixing the base model and enhanced.

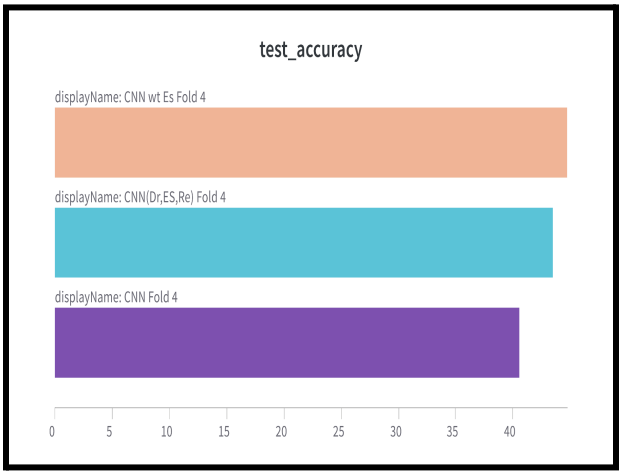
- The comparison charts for the fourth fold illustrate the performance metrics for three variations of the CNN model.
- They show that the model with dropout and early stopping (CNN(D,ES,Re)) has a more stable and consistent performance with lower training loss and higher validation accuracy compared to the other two variations.
- This suggests that the implementation of dropout and early stopping significantly aids in mitigating overfitting, leading to a more generalized model.

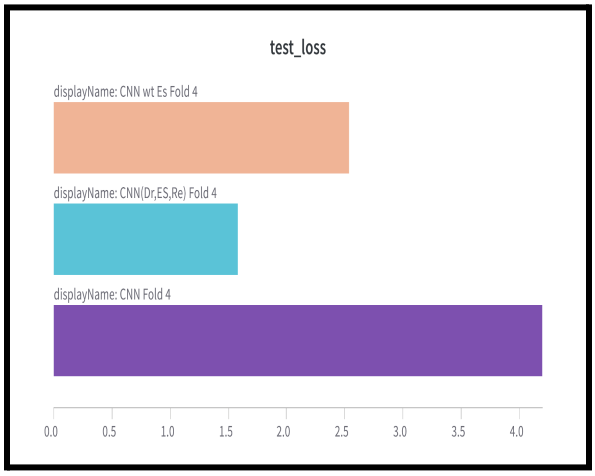
- Furthermore, the bar graphs for test accuracy and loss confirm the superior performance of the CNN(D,ES,Re) model on unseen data, reinforcing the effectiveness of these regularization and early stopping techniques in enhancing the model's reliability.











Architecture 2:

- The model first uses convolutional layers, which are like filters that pick up various features from the audio, such as tones or rhythms.
- At its core, the model employs transformer blocks that process the extracted features using attention mechanisms.
- This allows the model to dynamically focus on different parts of the audio sequence, capturing complex temporal relationships essential for understanding audio content.

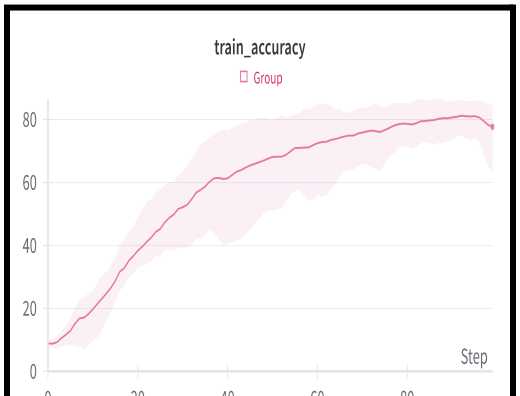
- The transformer blocks utilize multi-head attention mechanisms, splitting the attention process into multiple heads(1,2,4). T
- his structure allows the model to simultaneously attend to information from different representation subspaces at different positions, enriching the model's ability to interpret complex audio signals.
- To retain the sequential order of audio data, the model incorporates positional encodings which are learnable offering the model an opportunity to adapt positional information to the specific task at hand.
- The architecture concludes with a linear layer that translates the transformer's output, specifically the processed CLS token, into the desired number of classes, effectively categorizing the audio inputs.

Number of Heads	Train Accuracy	Valid Accuracy	Test Accuracy	Train Loss	Valid Loss	Test Loss	F1 Score	ROC AUC
1 Head	80.74%	46.39%	43.47%	0.5412	2.5903	2.9106	0.41	0.82
2 Heads	79.91%	49.86%	49.86%	0.5778	2.4115	2.4757	0.47	0.86
4 Heads	81.71%	44.86%	42.78%	0.5759	2.6297	2.4895	0.40	0.84

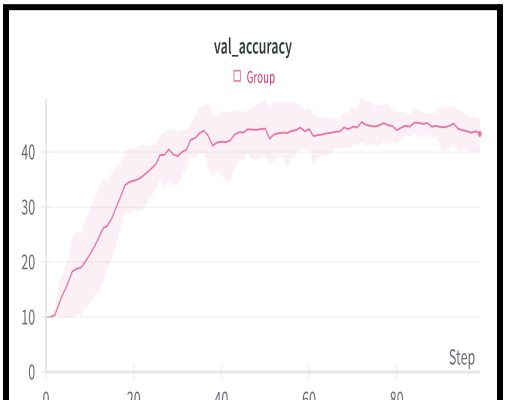
Key Observations:

- Effectiveness of Attention Heads: The model with 2 attention heads outperforms the others in terms of test accuracy and F1 score, suggesting a better balance between learning detailed features and generalizing across unseen data.
- Impact on Loss and Accuracy: While the model with 4 heads shows a slightly higher train accuracy, it does not translate into improved validation or test performance, indicating potential overfitting or less effective feature integration for classification.
- ROC AUC Consistency: The ROC AUC values indicate the model's capability to distinguish between classes. The 2 heads model exhibits superior performance, implying more effective utilization of attention mechanisms for classification tasks.

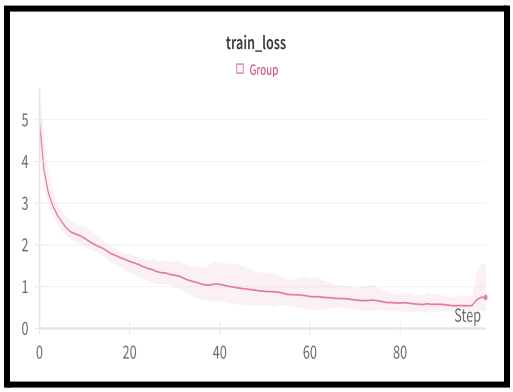
For Num Head = 1



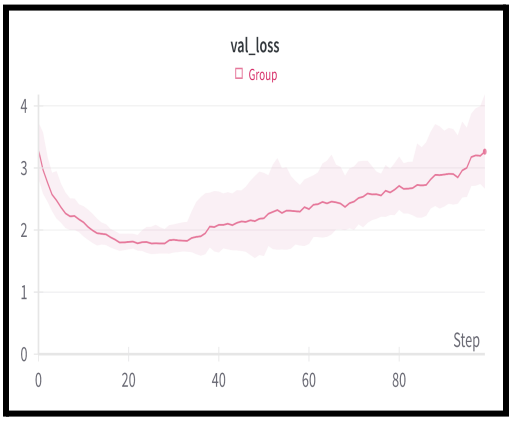
- The training accuracy graph shows a steady improvement as the number of steps increases, suggesting that the model is learning and adapting to the training data over time.

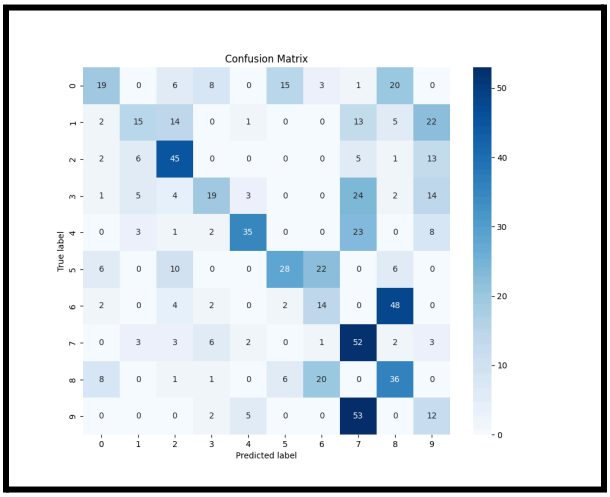


- The validation accuracy graph also shows improvement, but it plateaus earlier and at a lower value than the training accuracy.
- This plateau might indicate the model's limitations in generalizing from the training data to unseen validation data.

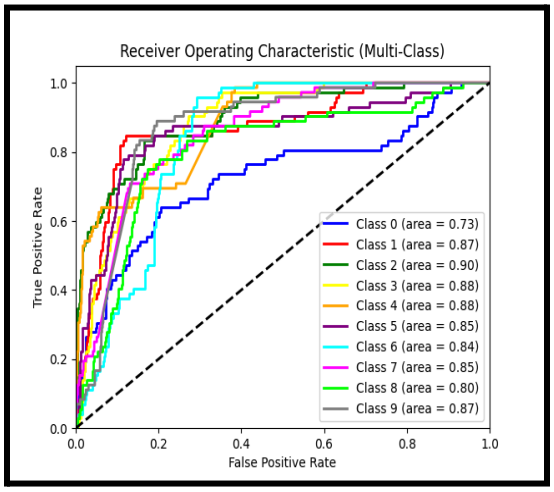


- The training loss decreases sharply initially and then gradually levels off, which is typical in model training as the model begins to converge on a solution.
- The validation loss decreases and then starts to fluctuate, which may suggest that the model is beginning to overfit to the training data

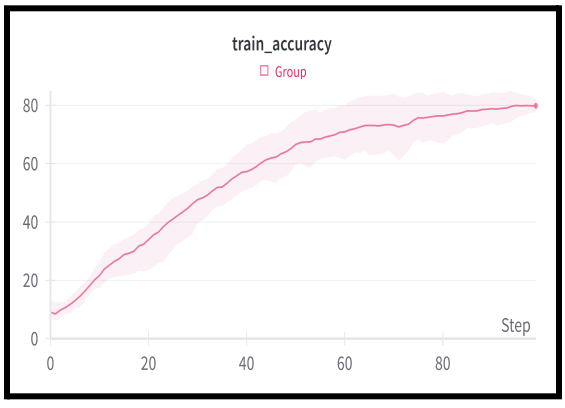




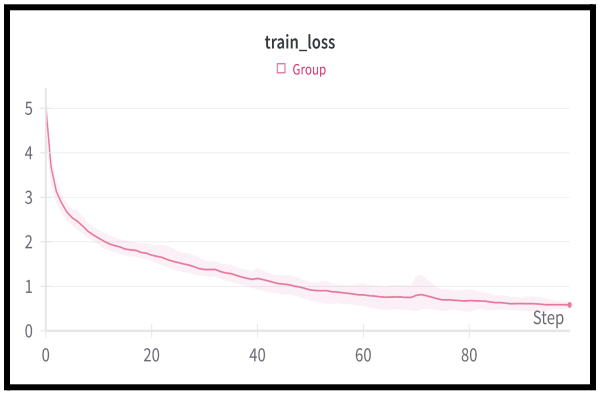
Some classes have a high number of correct predictions (along the diagonal), while others are more frequently confused with certain classes (off-diagonal elements), which indicates a more complex architecture required.



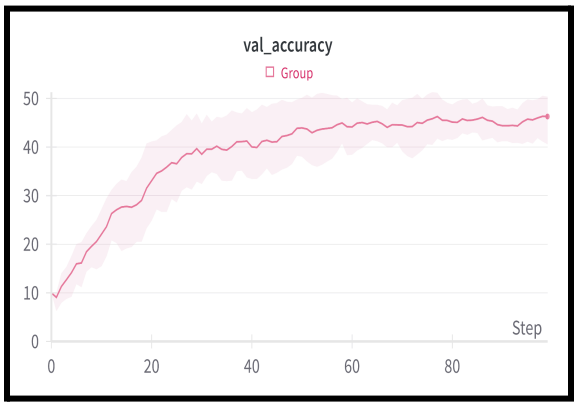
For Num Head = 2



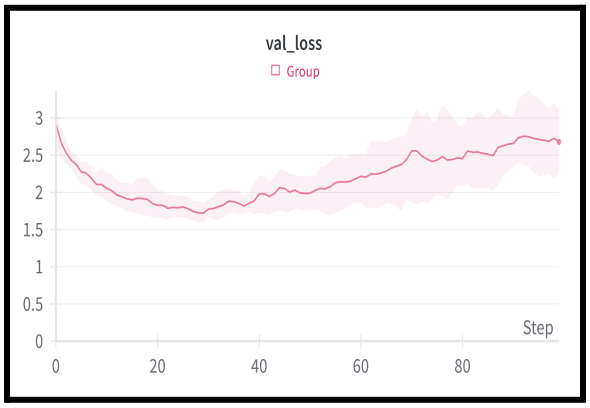
With two heads, we see a smoother convergence in training accuracy and loss, potentially reaching higher accuracy and lower loss values **more quickly than the one-head model.** This

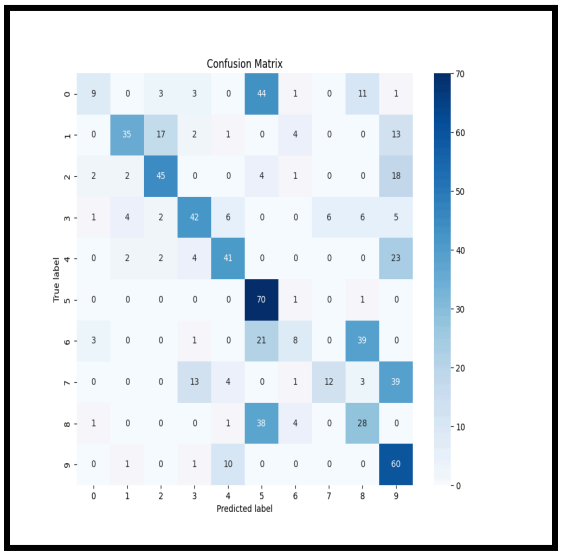


would suggest that the two-head model is able to learn more complex patterns in the data.

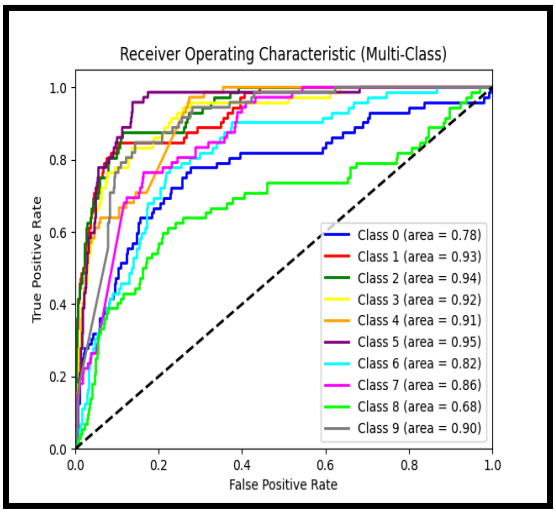


We can observe that the validation accuracy is higher than the one head model , indicating that it performs better over generalized data .





Certain classes such as Class 5 seem to be classified with high accuracy, while others, such as Class 0, show more significant confusion with oClass 5 with an AUC of 0.95 indicates excellent classification performance, whereas Class 8 with an AUC of 0.68 shows that the model struggles to distinguish this class from the others reliably.



For Num Head = 4

There are two models used :

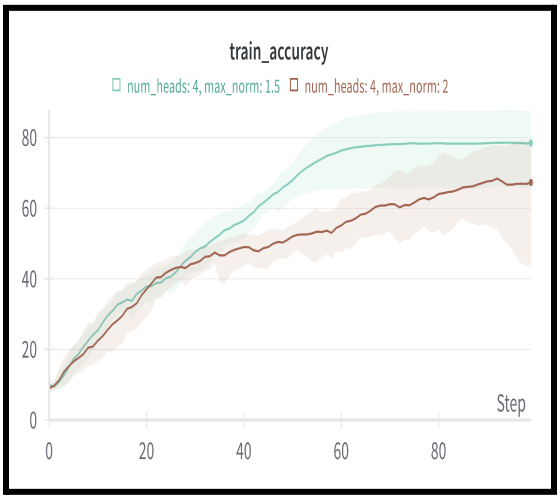
Base Model(Optimizer: Adam, Learning Rate: 0.001, Max Norm: 2):

- This model maintains a consistent learning rate throughout the training process and allows for a larger max norm constraint on the weights.
- The training and validation accuracy improvements are gradual, and the model achieves a modest performance.

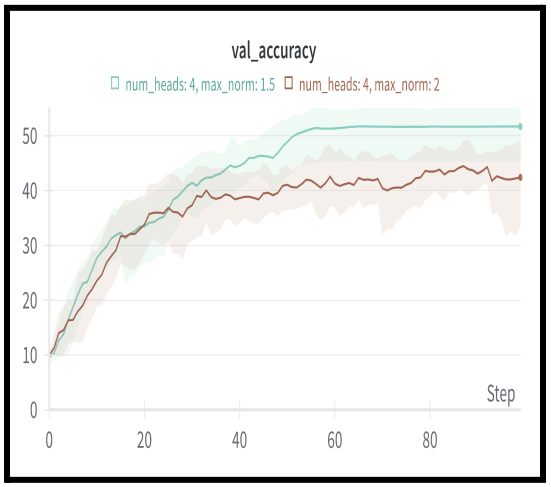
- However, the training loss reduction is steady, suggesting the model is learning effectively, albeit with potential for overfitting as indicated by a higher validation loss compared to Enhanced Model.
- In our training process for the AudioClassifierWithTransformer model, we encountered an issue with exploding gradients, where the loss escalated to 6000, disrupting the training.
- To address this, we applied gradient clipping, a technique that limits the size of gradients to prevent them from growing too large.
- By experimenting with clipping norms of 1, 1.5, and 2, we found that a norm of 1.5 yielded the best performance, effectively stabilizing the training process and enhancing the model's learning capability.

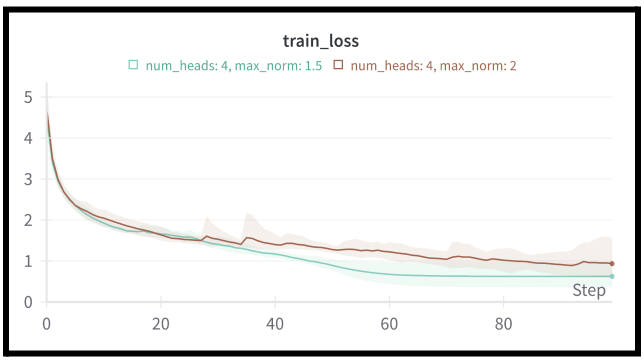
Enhanced Model (Optimizer: AdamW, Learning Rate Scheduler: ReduceLROnPlateau, Max Norm: 1.5):

- This model employs the AdamW optimizer, recognized for its weight decay regularization, coupled with a learning rate scheduler that reduces the learning rate when a plateau in performance is detected.
- The tighter max norm constraint potentially aids in regularizing the model further.
- The results show that this model generally achieves higher accuracy and lower loss on validation data, indicating better generalization capabilities.

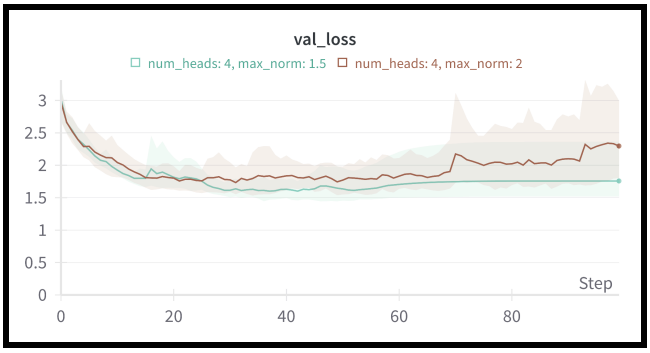


The enhanced model appears to have a slightly higher training accuracy throughout the training steps compared to the base model for num heads 4.The enhanced model demonstrates a higher validation accuracy compared to the base model. This suggests that the combined effect of the learning rate scheduler and the tighter constraint on weights (max norm of 1.5) might contribute to better generalization.

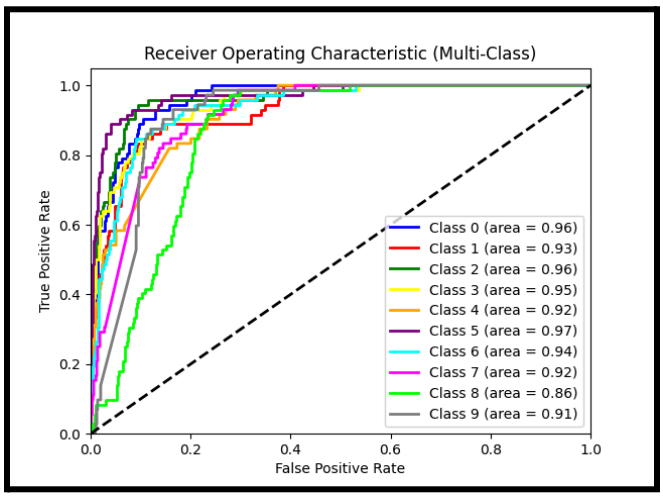




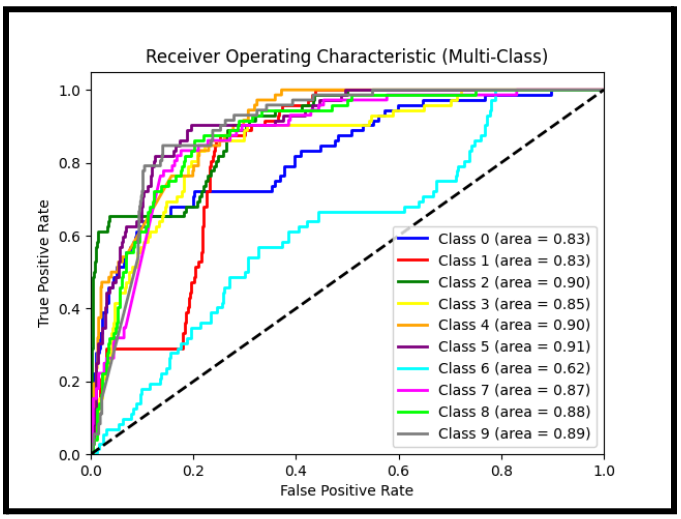
Both models show a similar pattern of decreasing training loss, which flattens as the steps increase, indicating convergence. However, the enhanced model shows lower loss, which could suggest a slightly better fitting to the training data.It's important to note that the validation accuracy for both models plateaus, but the plateau is achieved earlier and at a higher accuracy for the enhanced model .



The validation loss for the enhanced model is lower as compared to the base model, particularly in the later training steps. This could indicate that the enhanced model is less prone to overfitting and potentially generalizes better to unseen data.

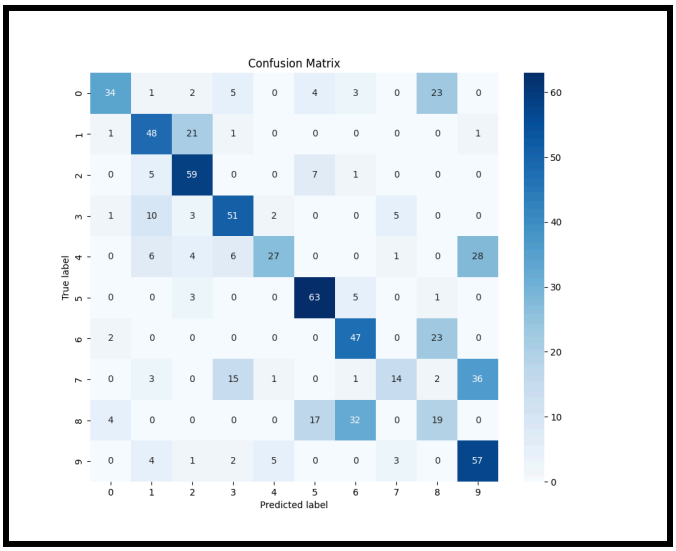


- Enhanced model : Shows higher Area Under the Curve (AUC) values for almost all the classes, indicating a strong ability to distinguish between positive class and negative classes.
- Classes such as 0, 2, 5, and 6 have particularly high AUC values (0.96, 0.96, 0.97, and 0.94, respectively), suggesting that the model is very effective at classifying these categories.
- The lowest AUC value is for Class 8 at 0.86, which, while lower than the others, still indicates good classification performance.

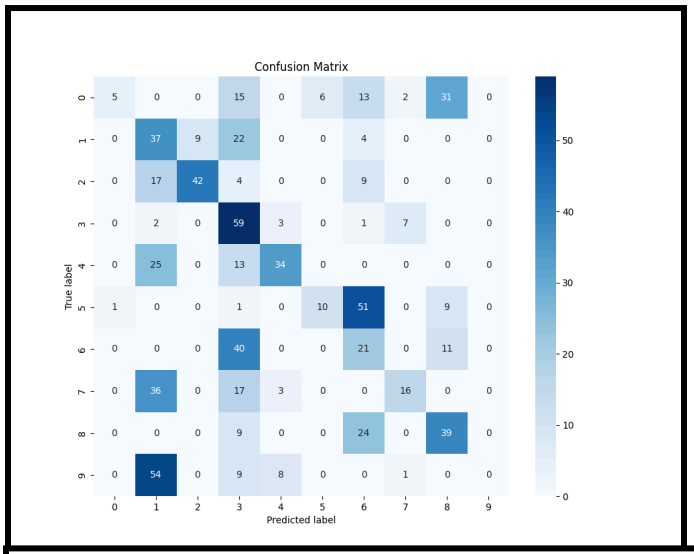


- Model:Exhibits lower AUC values for several classes when compared to the first model. For instance, Class 6's AUC drops significantly to 0.62, indicating a challenge in correctly classifying this particular class.

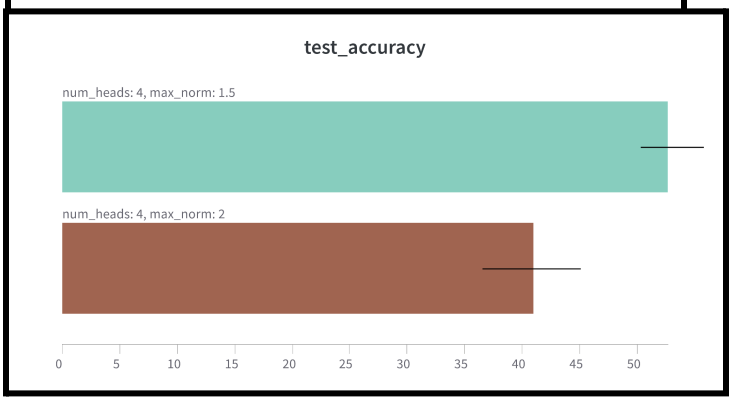
- Despite the overall lower performance, some classes like Class 2, 5, and 8 have high AUC values (0.90, 0.91, and 0.88, respectively), showing competencies in these classifications.

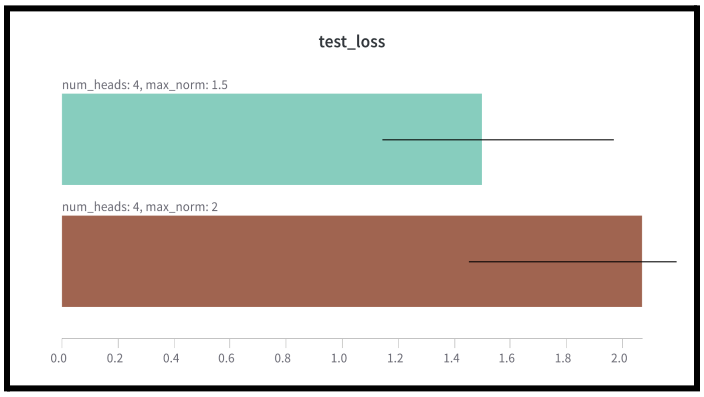


- This **enhanced model matrix** indicates a relatively balanced distribution of correct predictions across most classes, with particularly high accuracy for classes like 5 and 6.
- There are some misclassifications notable for classes 0 and 1, where they are confused with each other, and class 4 seems to be frequently misclassified as class 3.
- The darker shades in the diagonal suggest that the model is generally performing well in classifying most of the classes correctly.



- In this **base model matrix**, class 0 shows significant confusion with class 2, and class 1 with class 2 as well.
- Class 4 is often misclassified as class 3, similar to the first model, and there is a noticeable confusion between classes 6 and 8.
- The diagonal is less intense compared to the first matrix, indicating that while the model correctly predicts several classes, the rate of correct predictions is lower overall.





Metric	Enhanced Model - Best Fold
Train Accuracy	79.81%
Validation Accuracy	55.00%
Test Accuracy	**58.19%**
Train Loss	0.6030
Validation Loss	1.5191
Test Loss	1.1435
F1 Score	0.56
ROC AUC	0.93

Trainable Parameters:

- SoundClassifier: Has approximately 16.4 million(16,393,258) trainable parameters.
- AudioClassifierWithTransformer: Contains about 8.9(8,877,226) million trainable parameters.
- Both do not contain any non trainable parameters.