|  |  |
| --- | --- |
| **Name:** Syed Wamiq Ul Islam  **Batch:** BSCS 2024 | **Name:** Abdul Rehman Tahir  **Batch:** BESE 2023 |
| TUKL-DLL, NUST | **Date:** 23rd July 25 |
| **Supervisor:** Dr. Naseer Bajwa | |

**Progress Report**

**Problem:**

Using active learning to achieve best results, while using minimum number of samples, and finding out which active learning strategy is most efficient.

**Dataset Description**

**Shape:**

* X (No of Samples, Bands, Timesteps) = (20,000, 6, 28)
* y (No of Samples) = (20,000)

**Train/Val split:**

* 80/20 = 16000/4000

**Class distribution:**

* Class 0: 45.4%
* Class 1: 13%
* Class 2: 41.6%

**Hyperparameters**

* Dropout: 0.3
* Batch size: 32
* Learning rate: 1e-3
* LRScheduler: ReduceLROnPlateau(val\_loss)
* Weight decay: 1e-4
* Number of epochs: 30
* Optimizer: AdamW
* Loss function: CrossEntropyLoss with label smoothing and weights
* Label smoothing value: 0.1

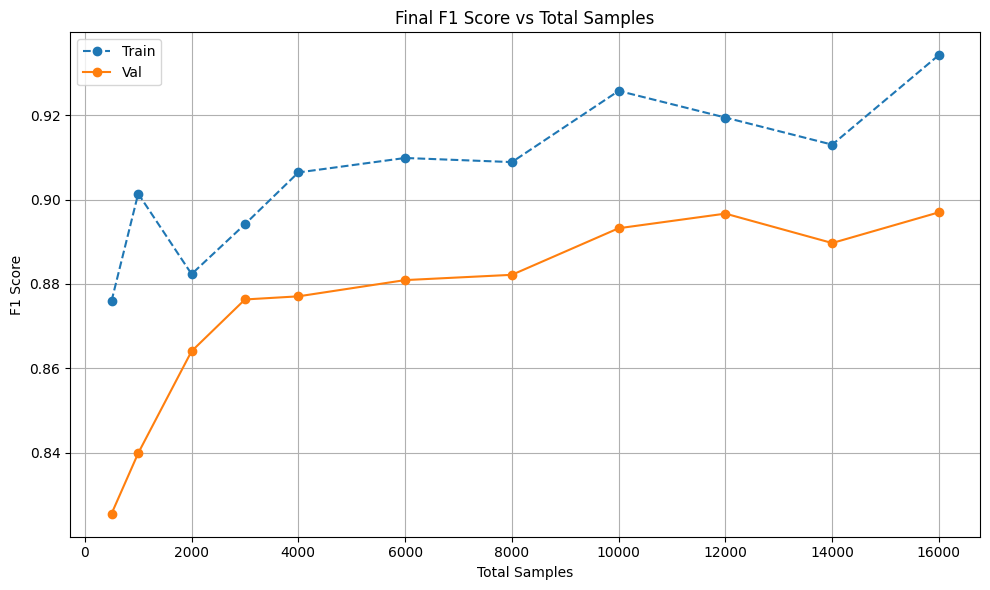
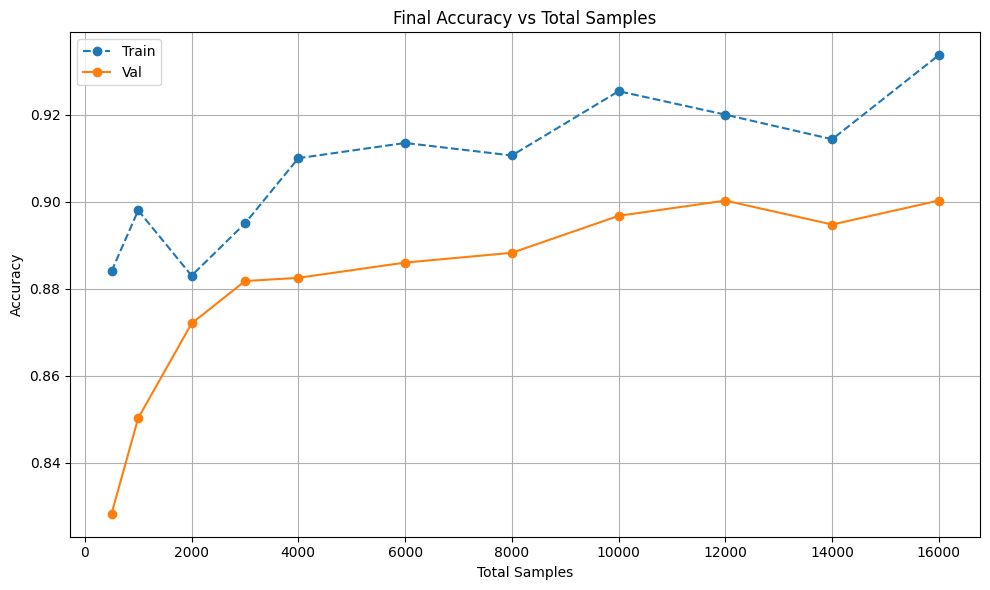
**Structure**

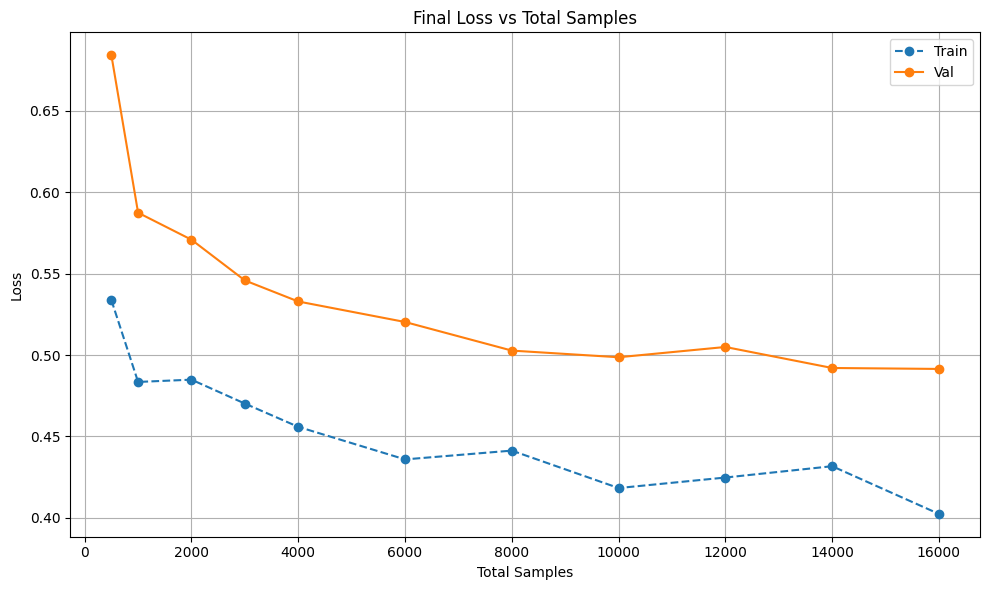
**Convblock**:Conv1D (kernel: 5) + BN + ReLU + Dropout(0.3)

|  |  |  |
| --- | --- | --- |
| **Block** | **Input Channels** | **Output Channels** |
| Conv1 | 6 | 64 |
| Conv2 | 64 | 128 |
| Conv3 | 128 | 256 |
| Conv4 | 256 | 512 |
| Conv5 | 512 | 1024 |
| AdaptiveAvgPool1d |  |  |
| FC layer | 1024 | 3 |

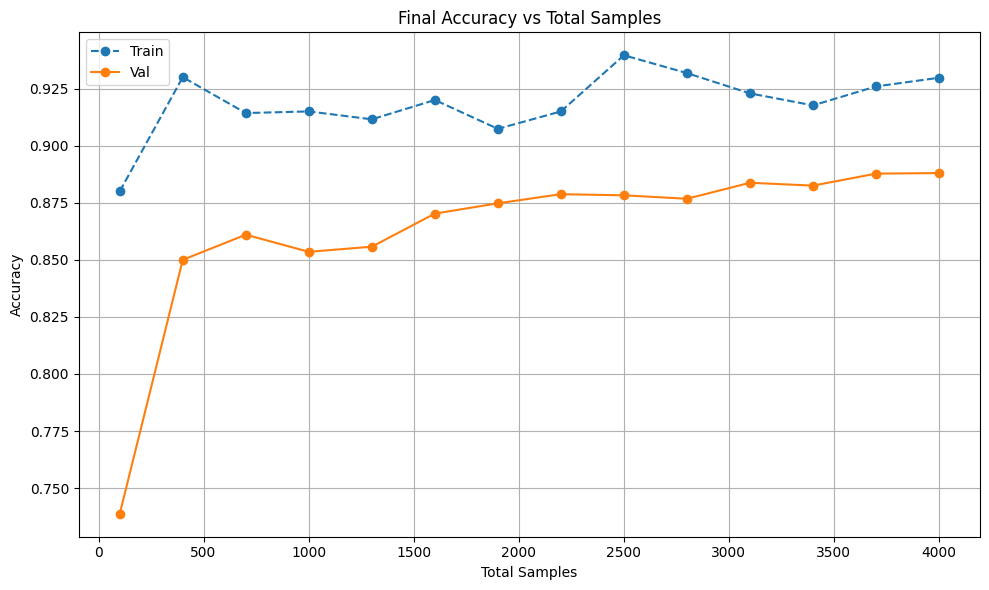
**Experiments**

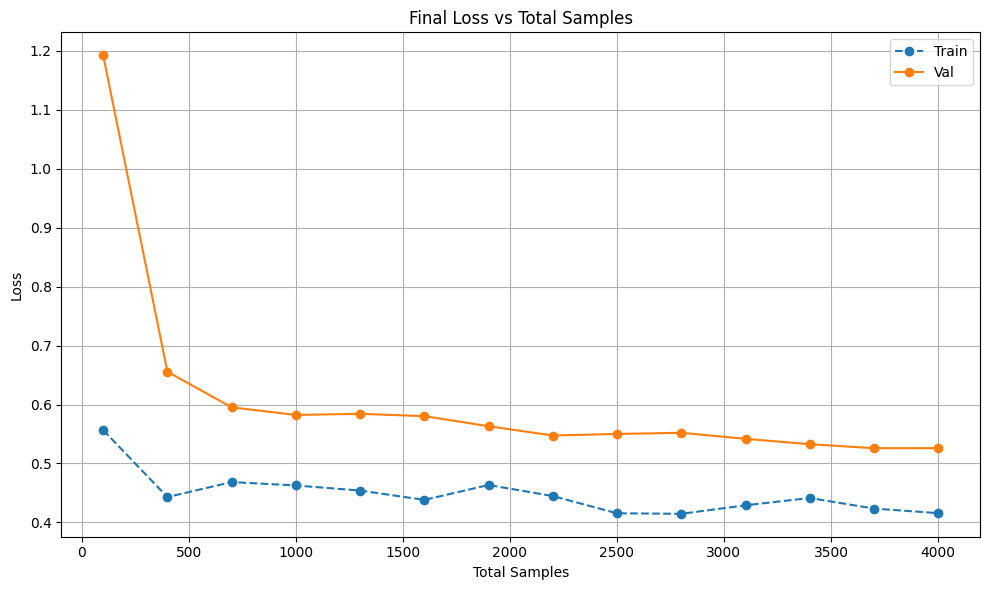
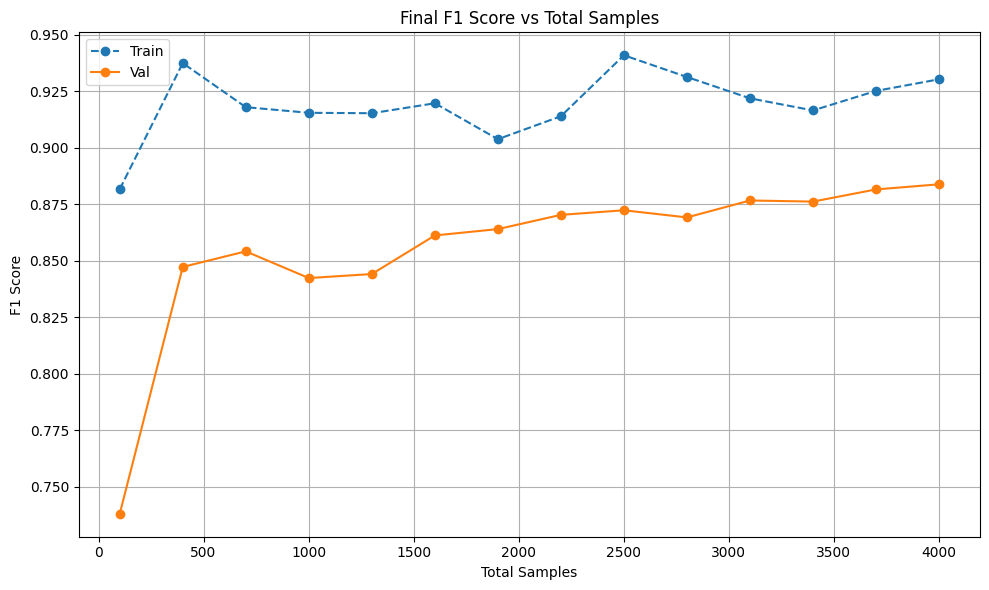
**Model metrics over training sample size 500-16000:**

* Trained the model on varying training sample sizes, from 500 to 16000 samples
* Model performance starts to saturate after around 4000 samples, and decreases significantly below 3000 samples
* Expanding 0-4000

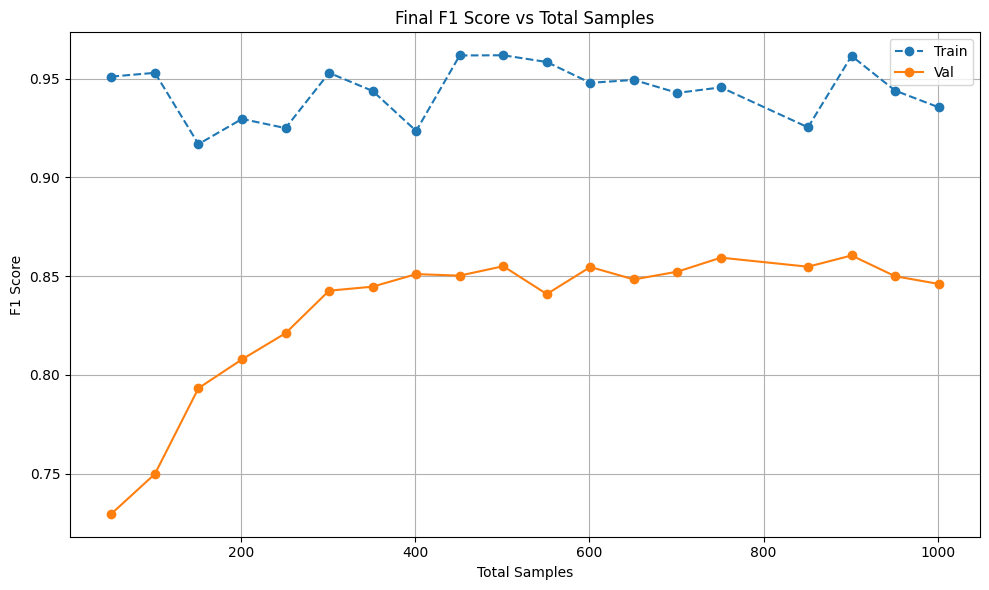
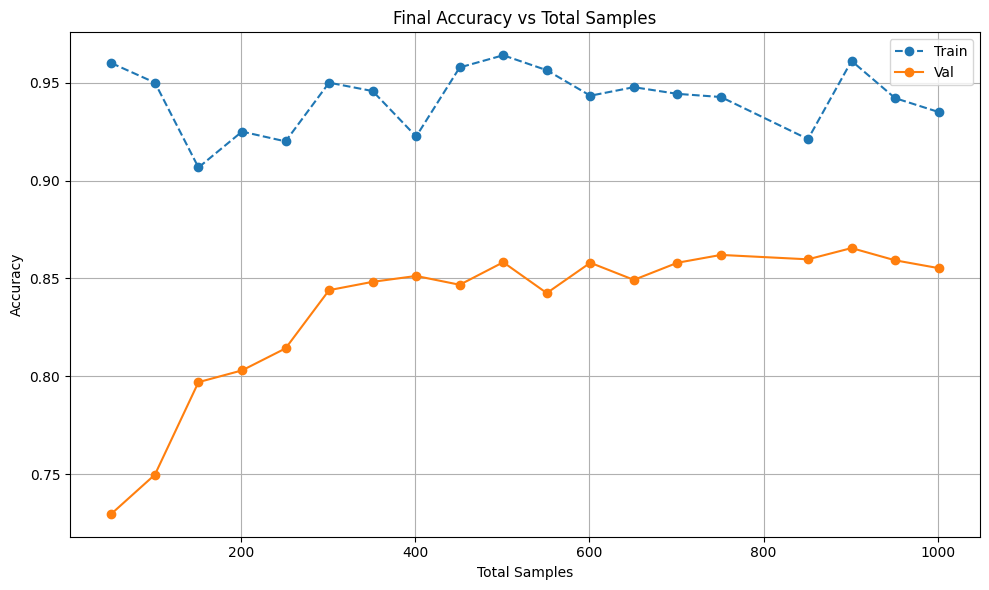


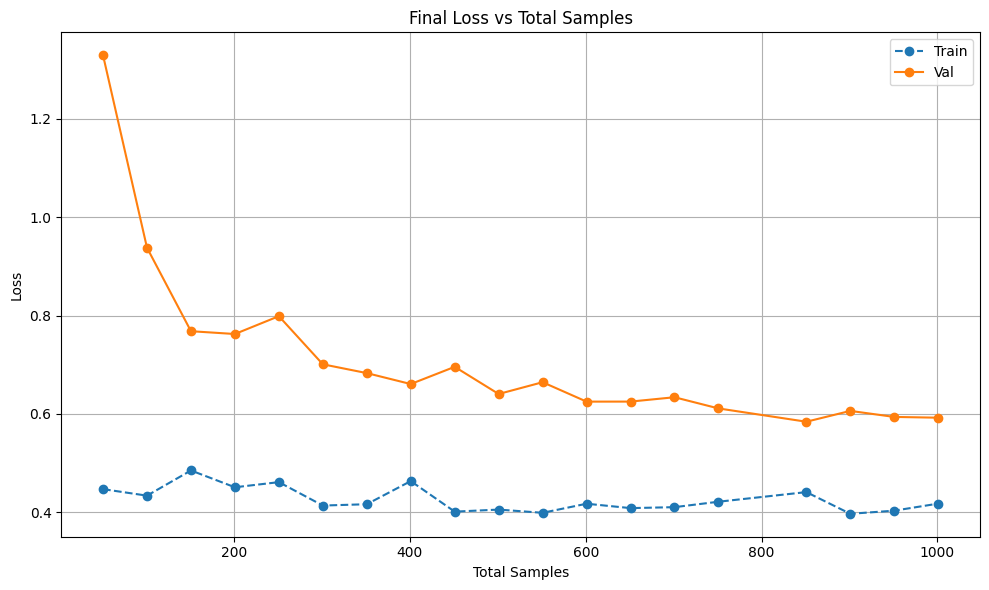
**Model metrics over sample size 100-4000:**

* Trained the model on varying data sizes, from 100 to 4000 samples
* Performance saturates at around 700 samples
* Accuracy is ~86.5% for 700 samples, and only increases to 88.5% for 4000 samples, and decreases rapidly below 400 samples
* Expanding 50-1000



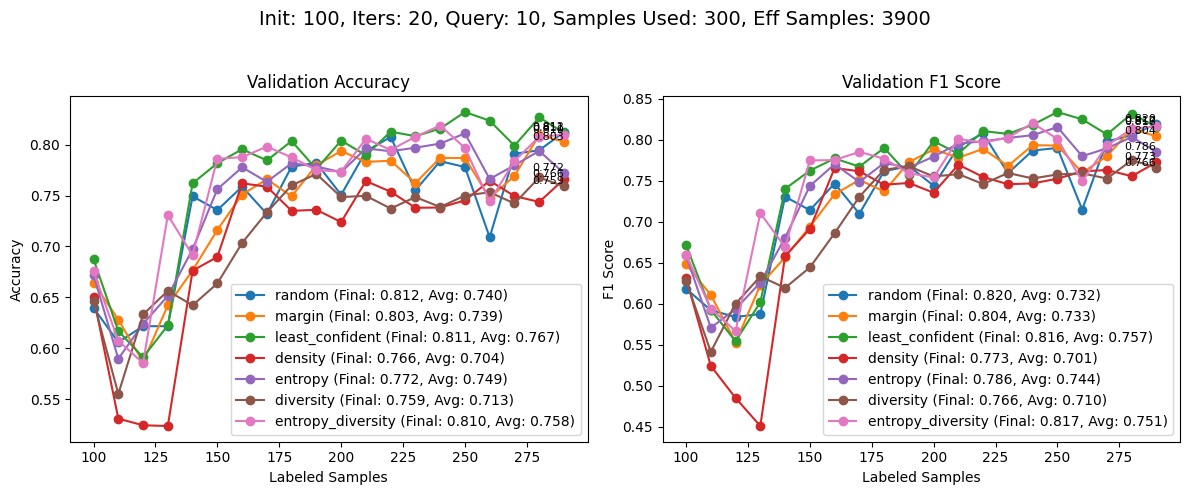
**Model metrics over sample size 50-1000:**

* Trained the model on varying data sizes, from 50 to 1000 samples
* Performance start to saturate at ~400 samples



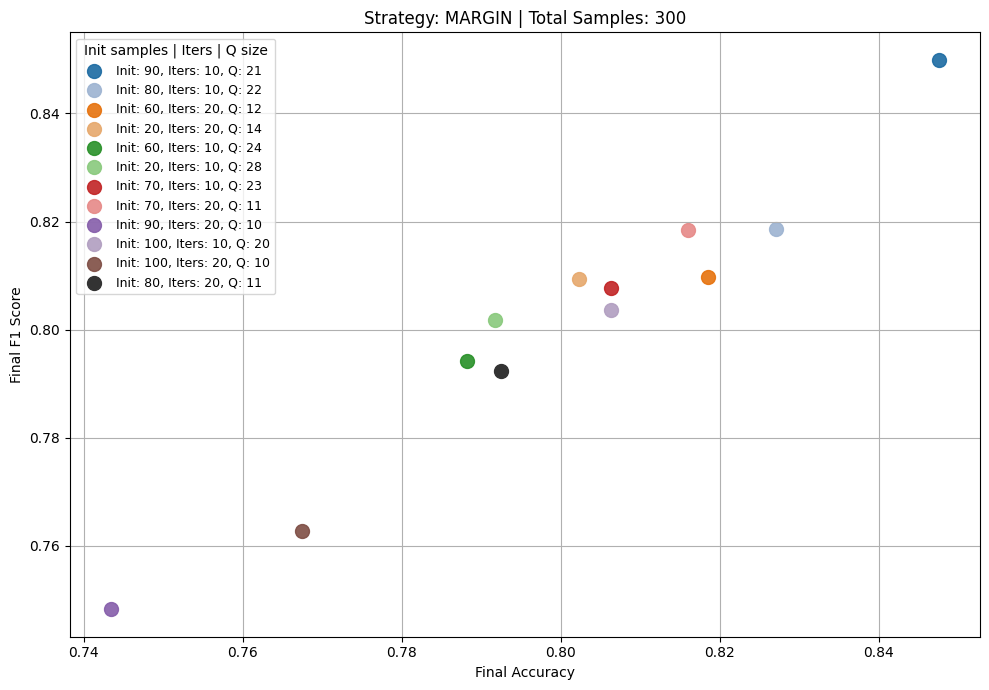
**Active Learning Methods over 300 samples:**

* Applied active learning to train the model on minimum samples to achieve respectable performance
* AL strategies: random, margin, least confident, entropy, density, diversity, entropy-diversity



**Evaluation over different initial sample sizes and iterations for 300 samples and margin AL:**

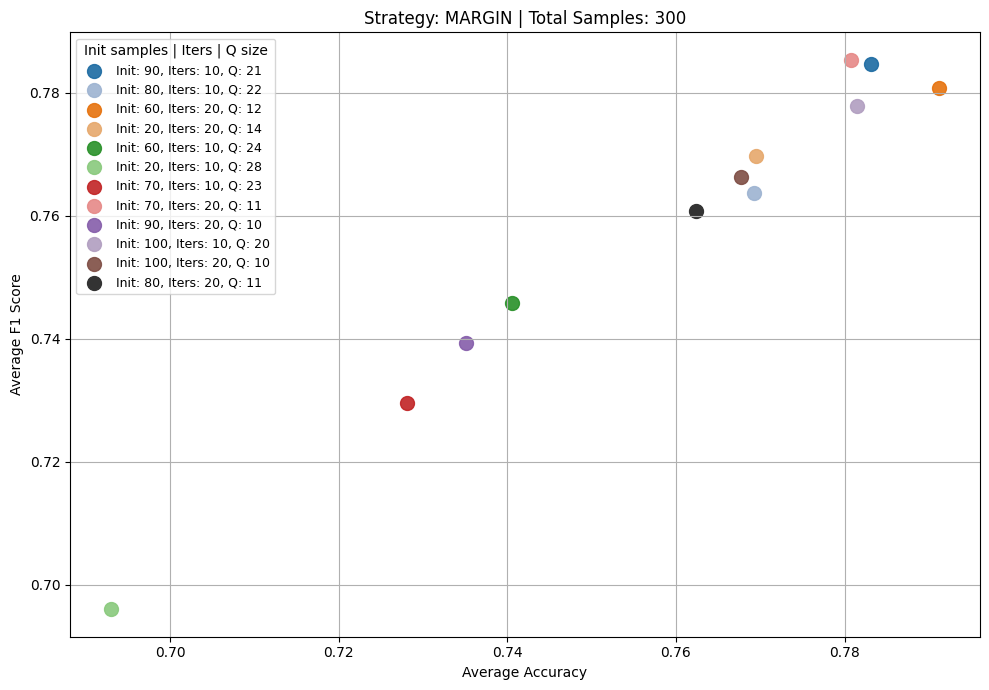
* Applied uncertainty margin active learning with 300 total samples over various initial samples, query sizes and iterations of AL loop
* 90 initial samples, with 10 iterations and 21 query samples yield best result



A screen shot of a graph

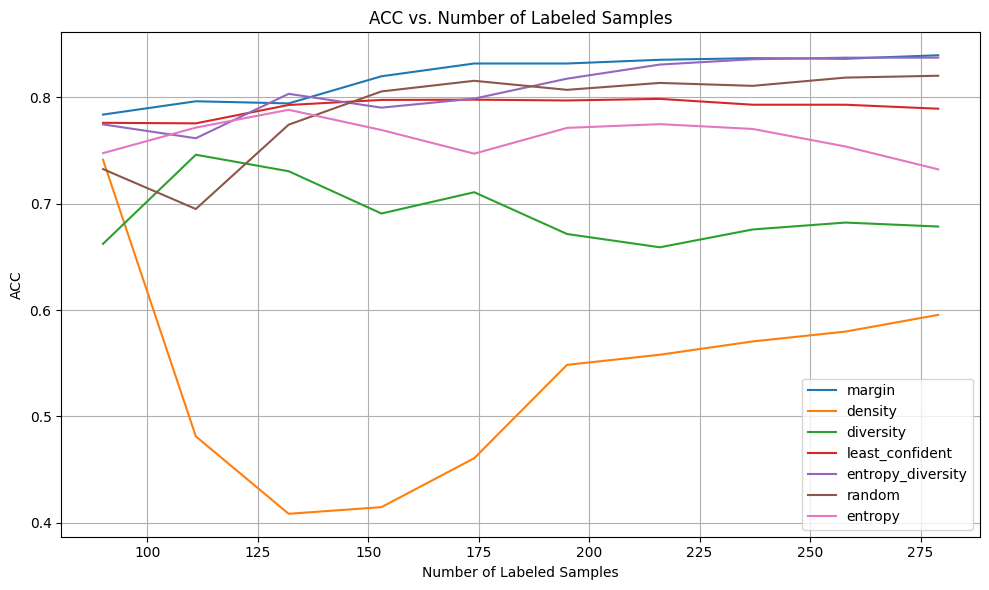
AI-generated content may be incorrect.

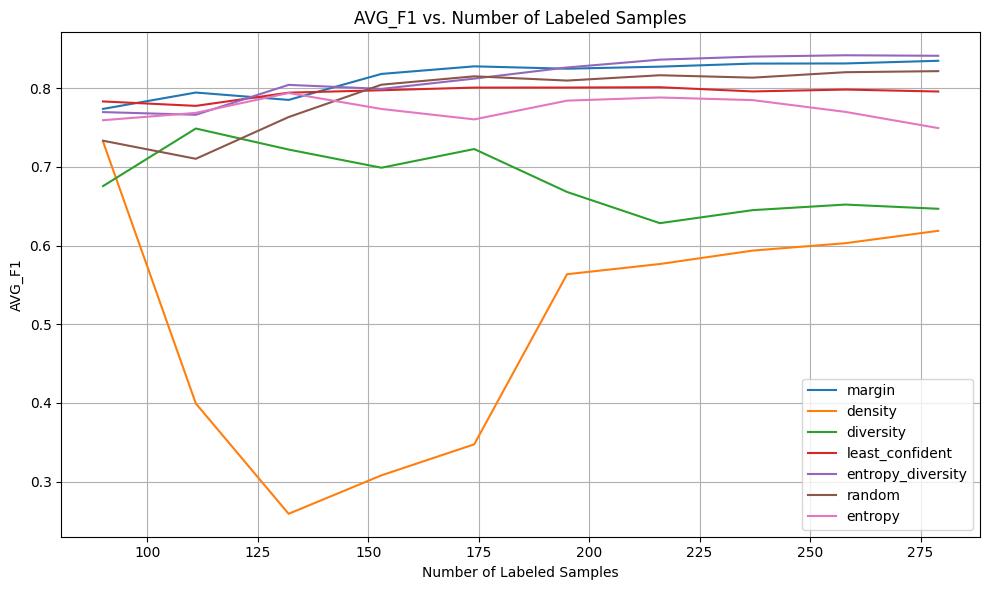
Avg f1/acc over complete active learning loop



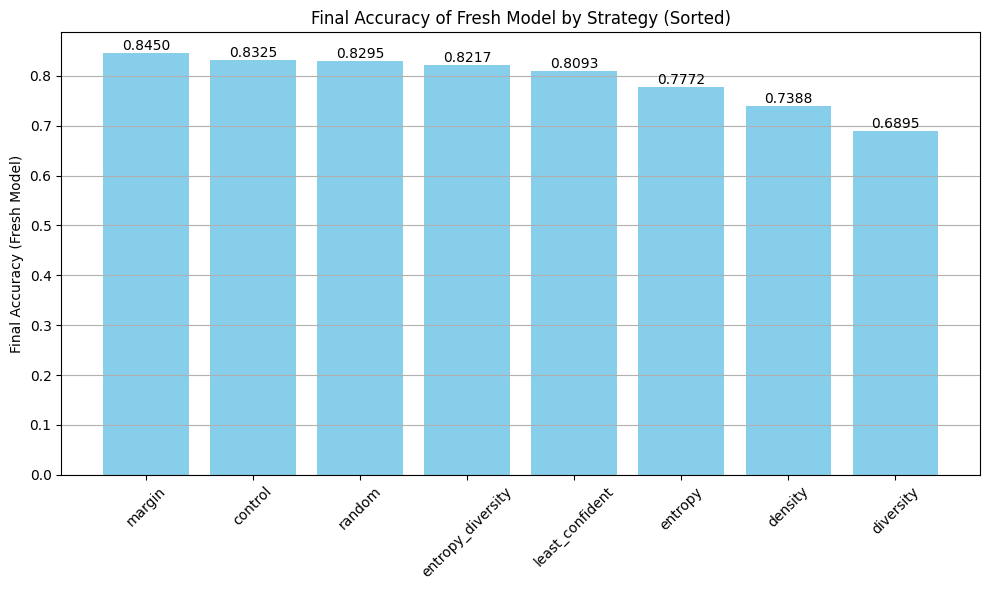
**Comparing different Active Learning methods:**

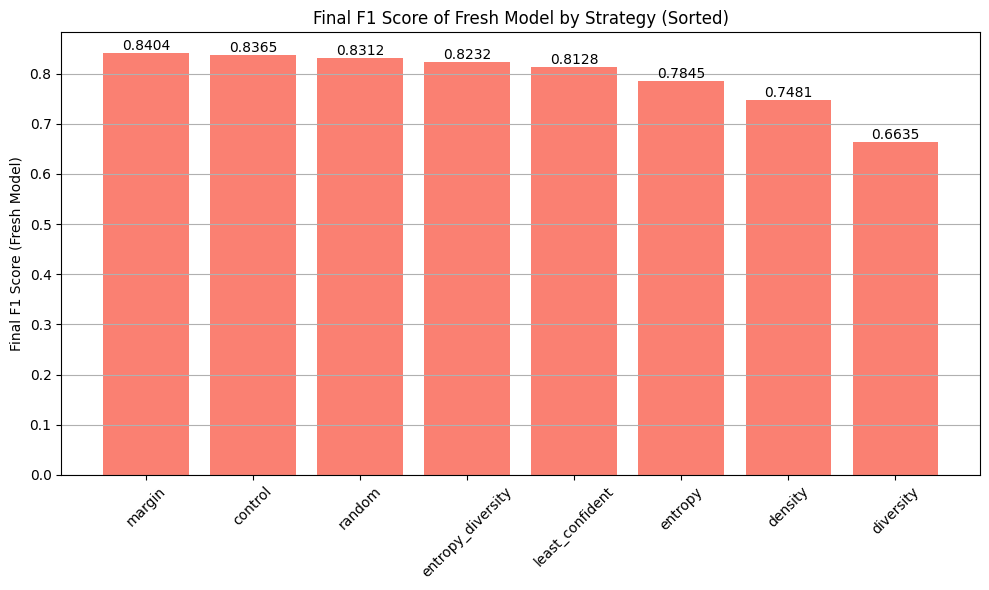
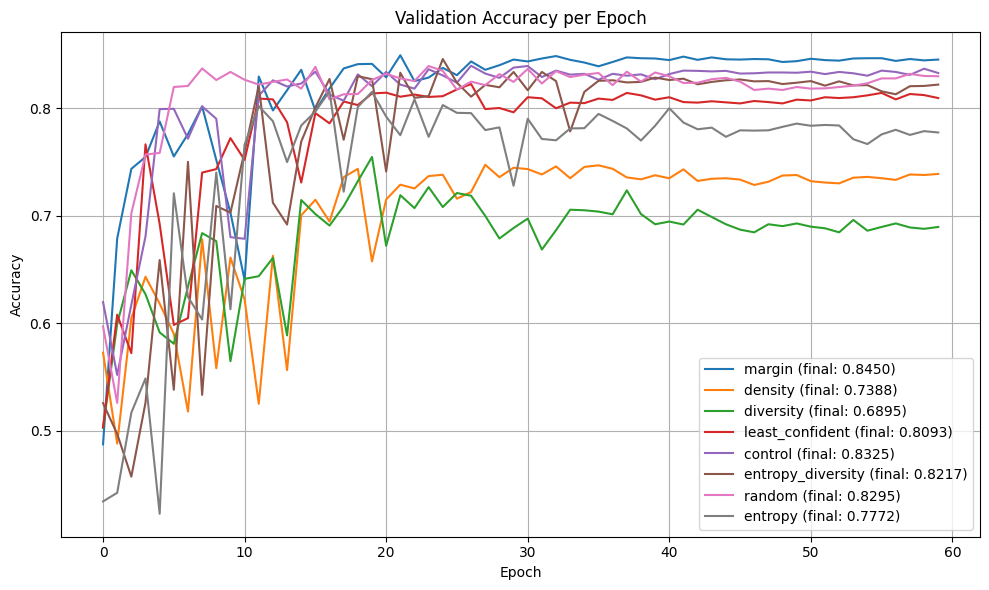
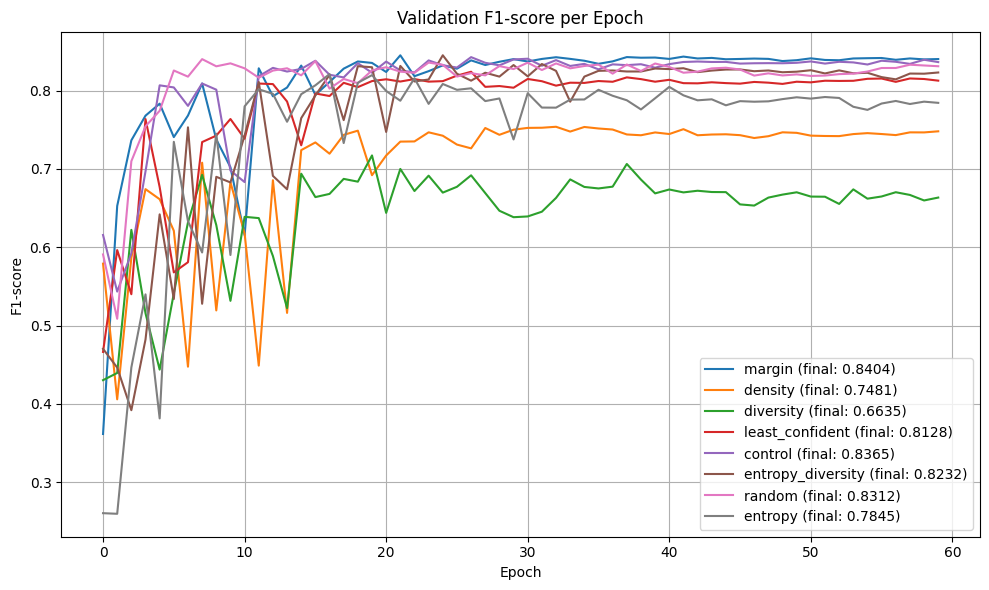
* AL methods include: Random, Margin, Least\_confident, Entropy, Density, Diversity, Entropy-Diversity
* It is apparent that margin and entropy-diversity are best methods





**Final Evaluation of fresh mode trained on dataset acquired form various AL methods**



**Conclusion**

* We were able to achieve 84.5% accuracy compared to 90% accuracy from 300/16000 samples ie. 94% of the result with 1.875% of the data.
* Comparatively we achieved 83.25% accuracy from using random samples, so we observed an increase of 1.25% by using active learning.
* This may suggest that the dataset was “easy” and thus using active learning did not have a major impact in the performance.