# **ONT | Task-5**

## **Distillation Techniques**

## **Introduction**

## The task is to apply knowledge distillation techniques by training a model on a European dataset (teacher model) and then using this model to guide the training of a smaller model (student model) for predicting GSNR (General Signal-to-Noise Ratio) on the same dataset. The same process is repeated for a USA dataset.

Knowledge distillation is a technique used to transfer knowledge from a large, complex model (teacher) to a smaller, simpler model (student). This process aims to leverage the teacher model's learned representations to improve the student model's performance. There are two main types of distillation:

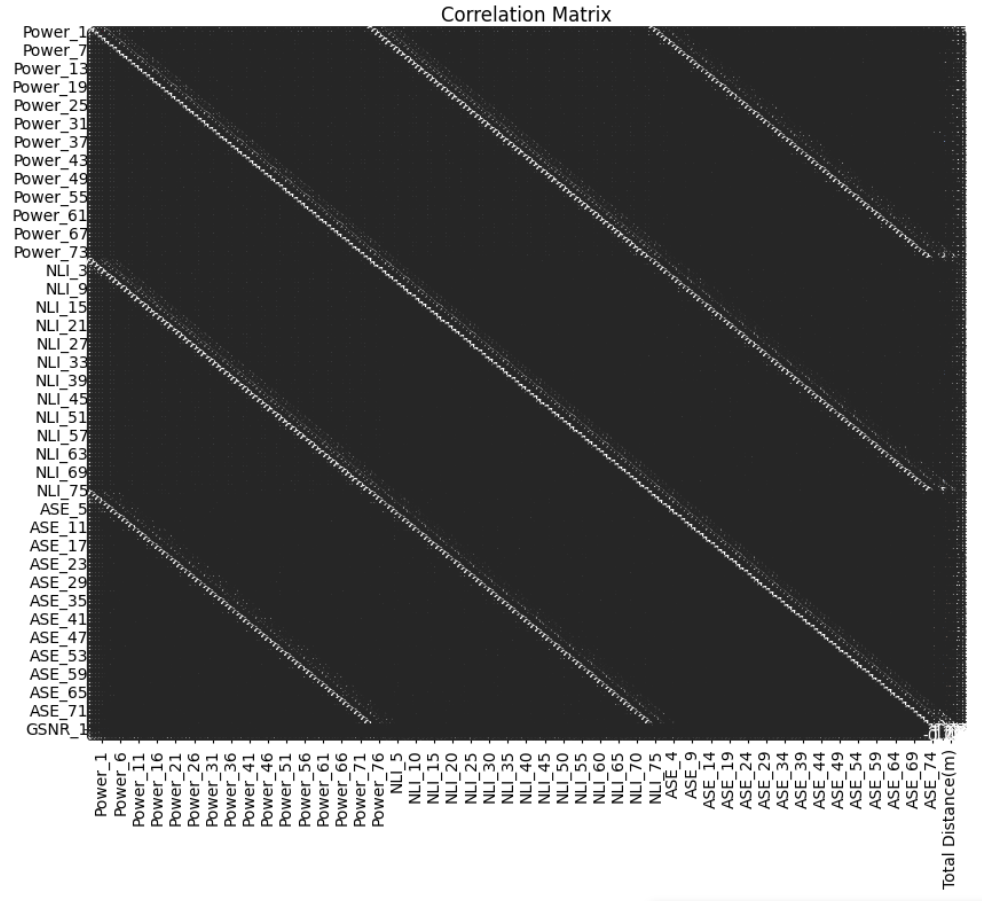
* **Hard Distillation:** The student model is trained on the same ground truth labels as the teacher model.
* **Soft Distillation:** The student model is trained using the soft labels (probability distributions) predicted by the teacher model. This often involves a temperature parameter to smooth the output probabilities.

## **Datasets**

Two datasets were utilized:

1. **European Dataset**: Served as the source dataset to train the initial model.
2. **USA Dataset**: Used for fine-tuning the pre-trained model to predict GSNR accurately in a different geographical context.

Both datasets contained features such as power, NLI (Nonlinear Interference), ASE (Amplified Spontaneous Emission) etc. which were essential for predicting GSNR.



## **Data Preparation**

**Preprocessing**

* **Missing Values**: Handled missing values by either dropping rows or filling them with mean values.
* **Feature Selection**: Dropped irrelevant columns such as 'frequency\_1 to frequency\_76' and 'GSNR\_2 to GSNR\_76', as well as 'Source', 'Destination', and 'Number of ON channels'.
* **Normalization**: MinMaxScaler was used to scale both the features and target variables to ensure uniformity and enhance model performance.
* **Custom Train-Test Split:** Data is split into training and testing sets using a custom function to ensure a balanced distribution of samples.

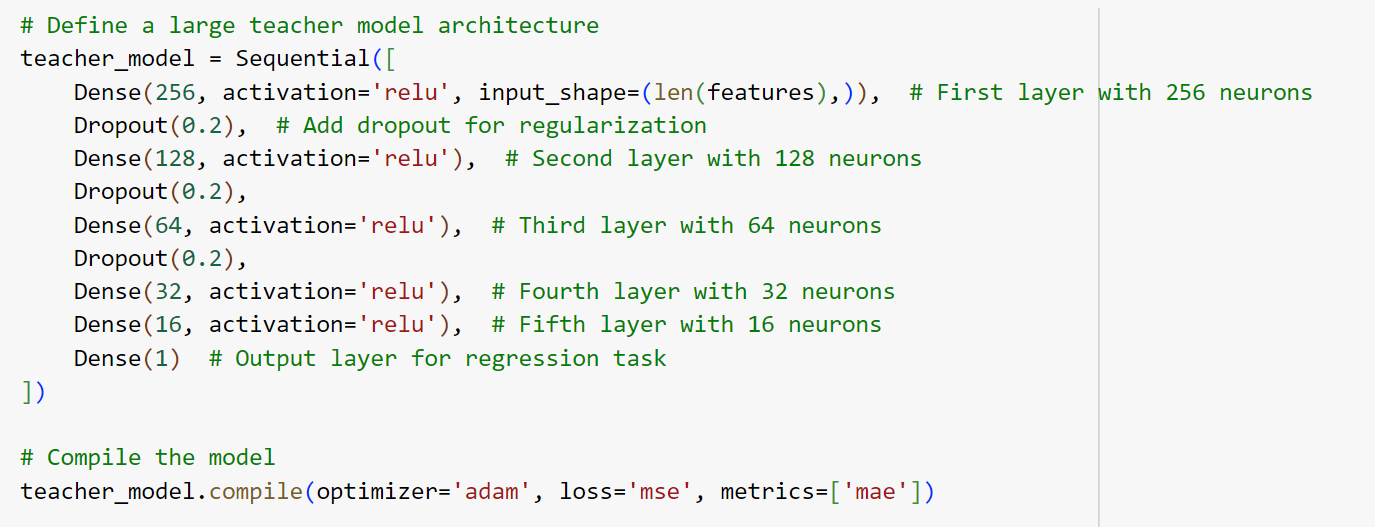
**Techniques and Libraries**

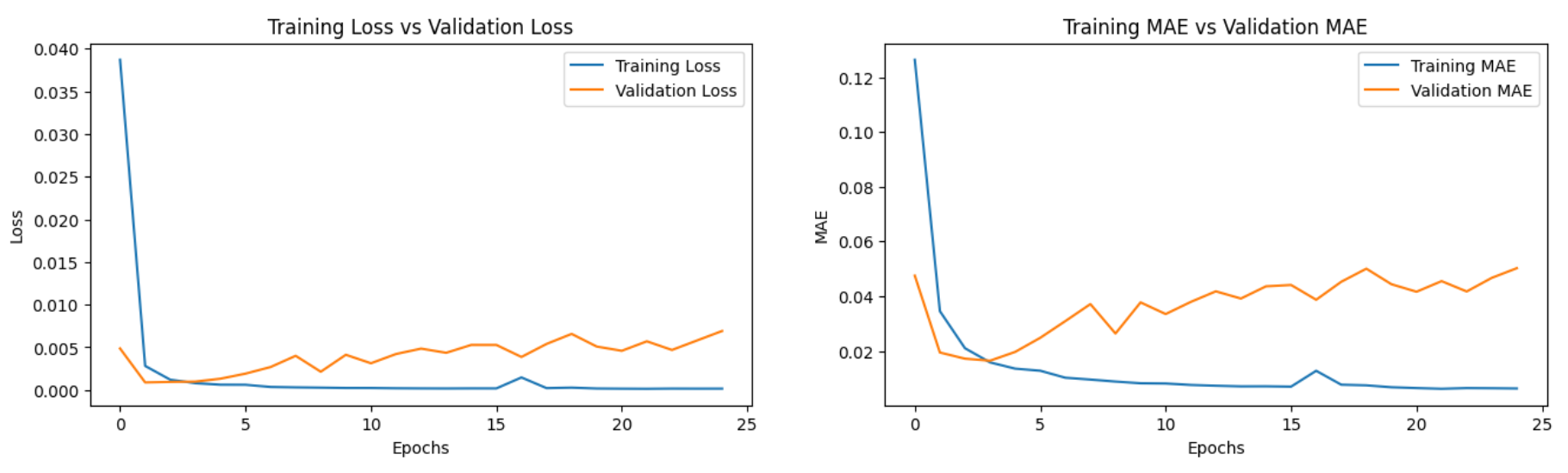
1. **Pandas**: For data manipulation and preprocessing.
2. **Scikit-Learn**: For data splitting, scaling, and evaluation metrics.
3. **TensorFlow and Keras**: For building, training, and fine-tuning deep learning models.
4. **Matplotlib and Seaborn**: For visualizing data distributions, correlations, and model performance.

## **European Dataset**

* **Model Training on European Dataset**

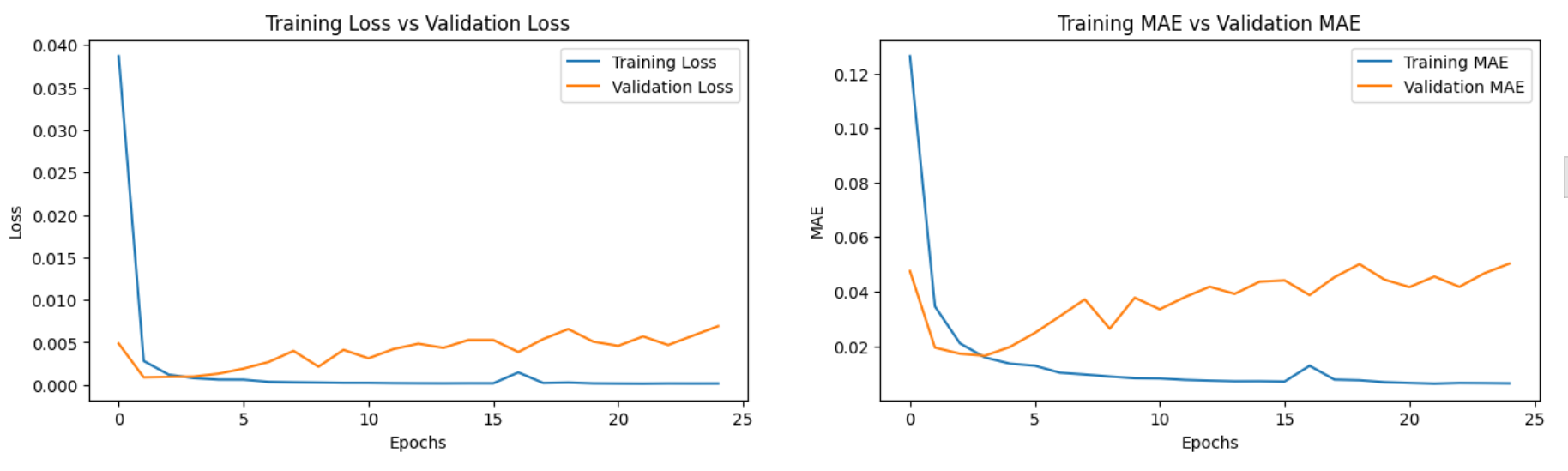
1. **Teacher Model Training:**
   * A deep neural network with multiple dense layers and dropout is defined and trained on the European dataset.
   * Training history (loss and MAE) is plotted to evaluate the model’s performance.



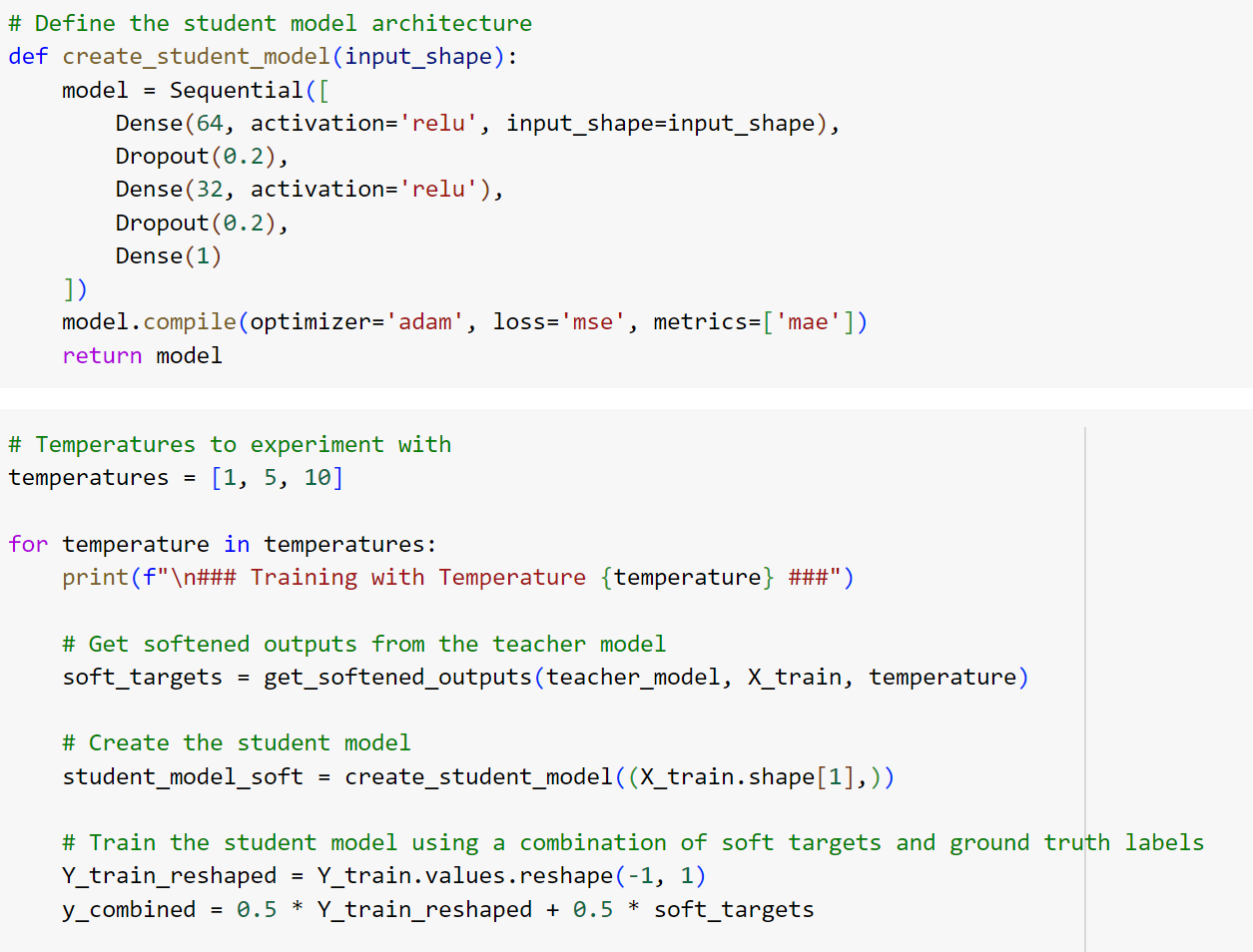


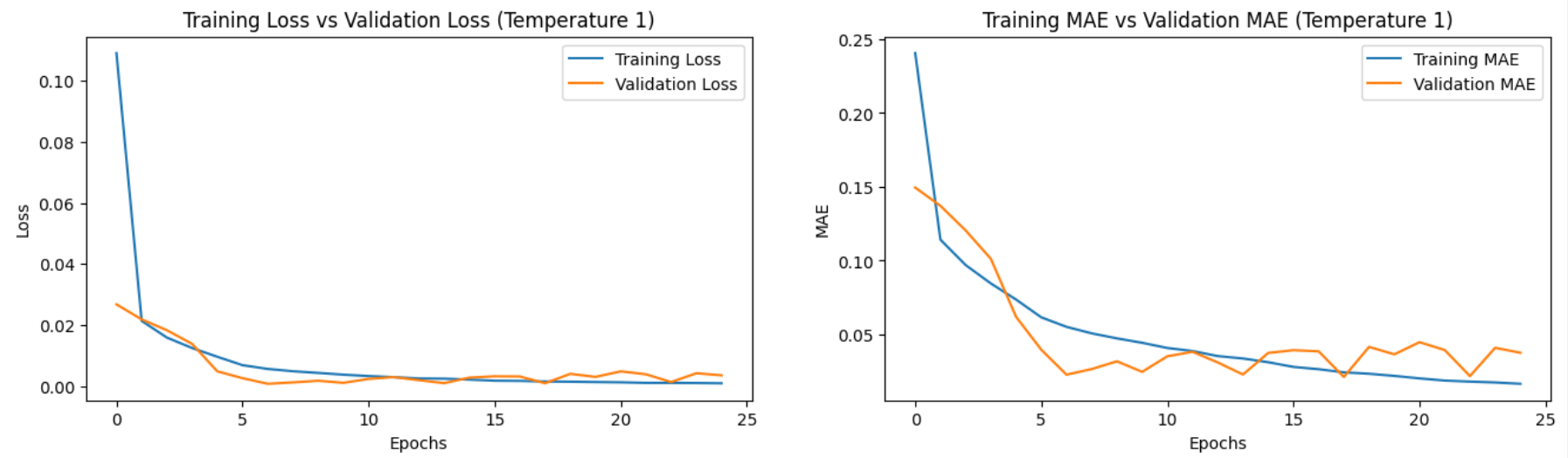
1. **Student Model Training (Hard Distillation):**
   * A simpler neural network architecture is defined for the student model.
   * The student model is trained using the same ground truth labels as the teacher model.
   * Training history is compared with the teacher model to assess performance.





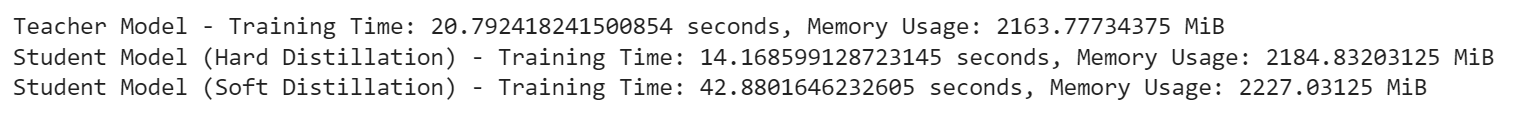
1. **Student Model Training (Soft Distillation):**
   * Softened outputs (probability distributions) from the teacher model are used to train the student model.
   * Various temperatures are experimented with to observe their effect on student model performance.
   * The student model is trained on a combination of ground truth and softened outputs.





* **Time and Memory Comparison**

1. **Performance Measurement:**
   * Training time and memory usage for both the teacher and student models are measured.
   * Comparisons are made to assess the efficiency of different models and distillation techniques.



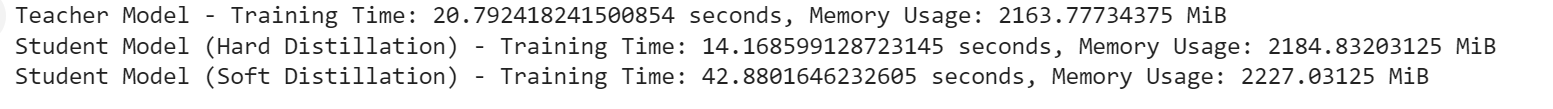
## **USA Dataset**

* **Model Training on USA Dataset**

1. **Teacher Model Training:**
   * A large neural network architecture is defined and trained on the USA dataset.
   * Training history (loss and MAE) is plotted.
2. **Student Model Training (Hard Distillation):**
   * A smaller neural network is trained on the USA dataset using hard distillation.
3. **Student Model Training (Soft Distillation):**
   * The student model is trained using softened outputs from the teacher model, with different temperatures tested.

* **Time and Memory Comparison**

1. **Performance Measurement:**
   * Training time and memory usage for both teacher and student models are measured and compared.



## **Conclusion**

The teacher model, being a larger and more complex architecture, takes longer for inference, especially with larger datasets. In contrast, both hard and soft distillation methods for the student model show relatively similar inference times. However, soft distillation requires additional time to calculate soft labels, making its overall process slightly longer. In terms of accuracy, the teacher model performs best as expected, followed by the hard distillation method which uses actual labels. Soft distillation, while least accurate, benefits from a smoothing parameter, where a temperature (T) value of 1 makes the function equivalent to softmax, and higher values of T smooth the label distribution but reduce accuracy. Training a student model with hard distillation is akin to training the teacher model if the architectures match, with the primary difference being the use of teacher model predictions to guide the student model.