Facial Expression Recognition Using Vision Transformers and Convolutional Neural Networks

by
Muhammet Hakan Taştan

T.C. YEDİTEPE ÜNİVERSİTESİ

Approved by **Prof. Dr. Duygun Erol Barkana**
(Advisor)

Faculty of Engineering
Department of Electrical and Electronics Engineering
Istanbul, *2020*

Introduction

- Facial expression recognition (FER) is categorizing human expressions from face images.
- FER has many practical applications in fields such as security, advertising, healthcare, and entertainment
- Recent advancements in deep learning have led to significant progress in FER
- Current methods are often resource-intensive and may not be suitable for real-time or mobile applications

T.C. YEDİTEPE ÜNİVERSİTESİ

Image Source: Generated using Playground-v2 bot on Poe (https://poe.com/Playground-v2)

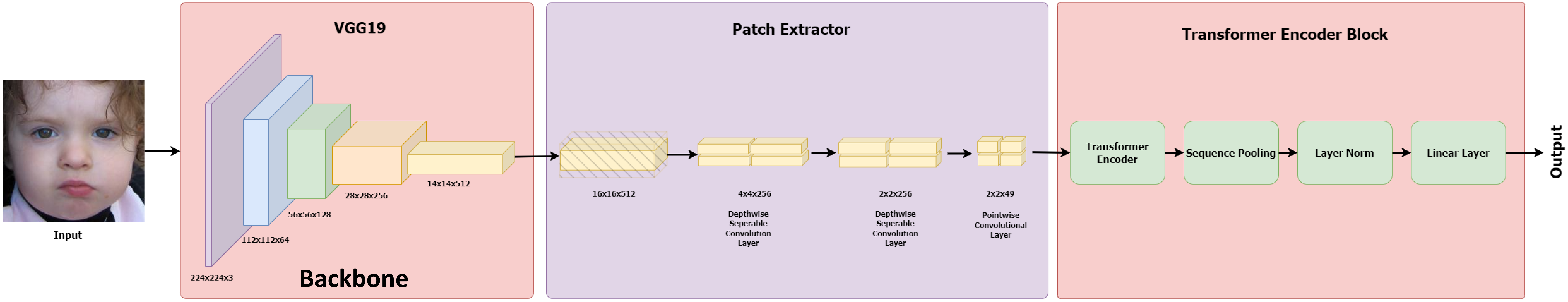
Cont'

- **Human-computer interaction**: used to improve the naturalness and effectiveness of interactions between humans and computers
- **Healthcare**: used to monitor patient progress and detect emotional states in mental health treatment
- **Entertainment**: used to create more engaging and personalized experiences.
- **Retail**: used to analyse customer satisfaction and improve customer service.
- **Advertising**: used to gauge customer engagement and measure the effectiveness of advertising campaigns
- **Education**: used to assess student engagement and performance in online learning environments.

T.C. YEDİTEPE ÜNİVERSİTESİ

Cont'

- This project tries to find a lightweight and efficient method to achieve high performance for the FER task.
- Proposed models combines CNNs and transformers for efficient facial expression recognition (FER)
- Achieves 55.54% accuracy on 8 classes with lightweight 5.67M parameters on AffectNet database



Emotion Modeling

<u>Dimensional Model:</u>

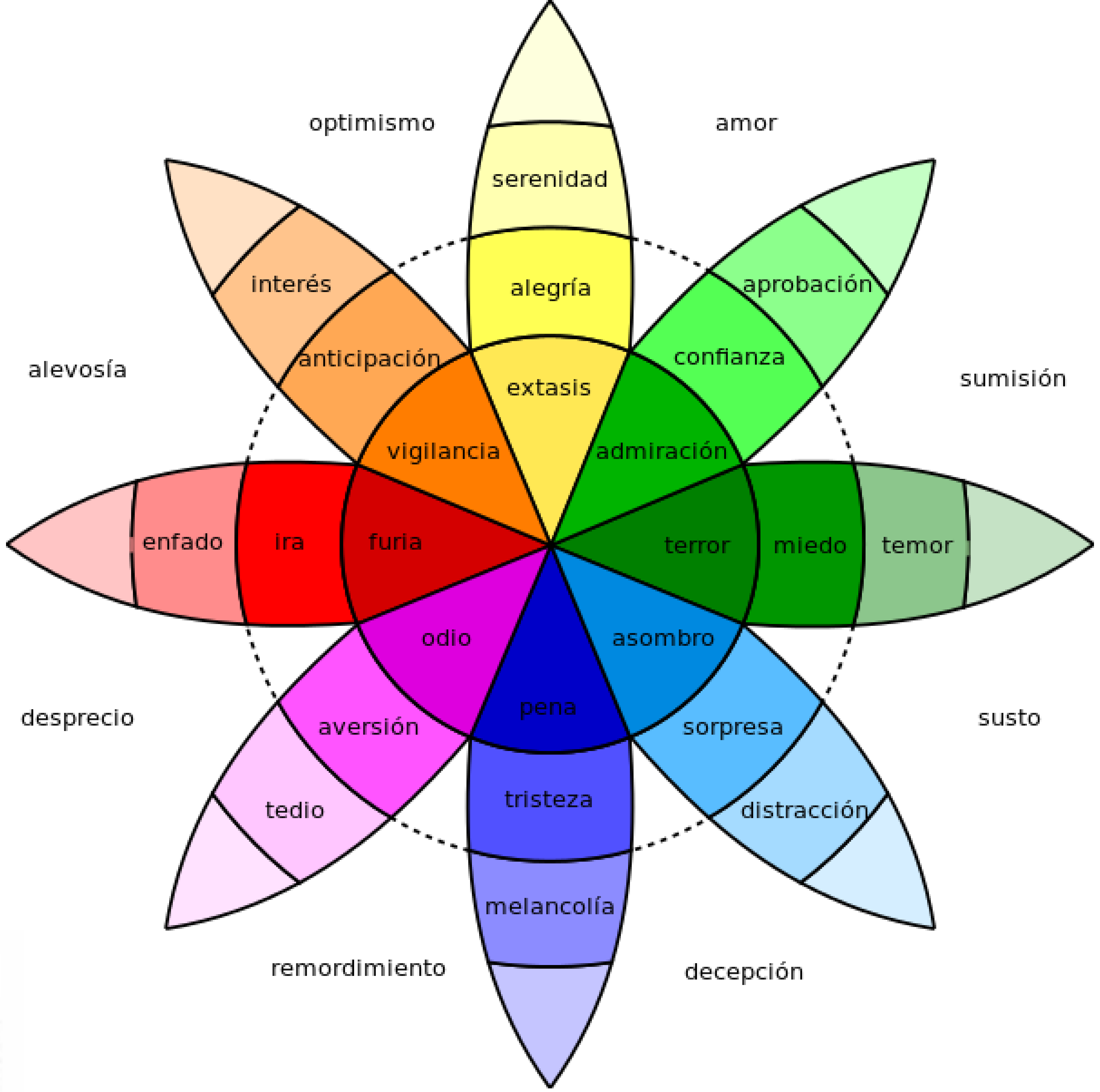
- The emotions are defined in a multidimensional continuous space.
- Two dimensional models use **valence** and **arousal** as dimensions
- Three dimensional models add also **dominance** to the **valence** and **arousal**.
- Valence signifies the intensity of the pleasantness
- Arousal indicates the level of physiological activity
- Dominance is recognized as attention.
- Exp. Circumplex Model, PAD

<u>Categorical Model:</u>

- The emotions are defined as distinct classes.
- Basic set of emotions or more complex set tailored to problem domain
- Basic emotions can be combined to create more complex emotions
- Widespread use for annotations of emotions in datasets.
- Exp. Ekman's 6 basic emotions, Plutchik's emotional wheel

Cont'

Circumplex Model



Plutchik's emotional wheel

Ekman's 6 Basic Emotions

Method

- CNNs were default choice for vision tasks for a considerable time
- Recently, vision transformers or attention-based approaches becoming more popular
- CNNs don't scale as well as transformers
- Transformers require large amounts of data and compute power
- Efforts are being made to combine CNNs and transformers

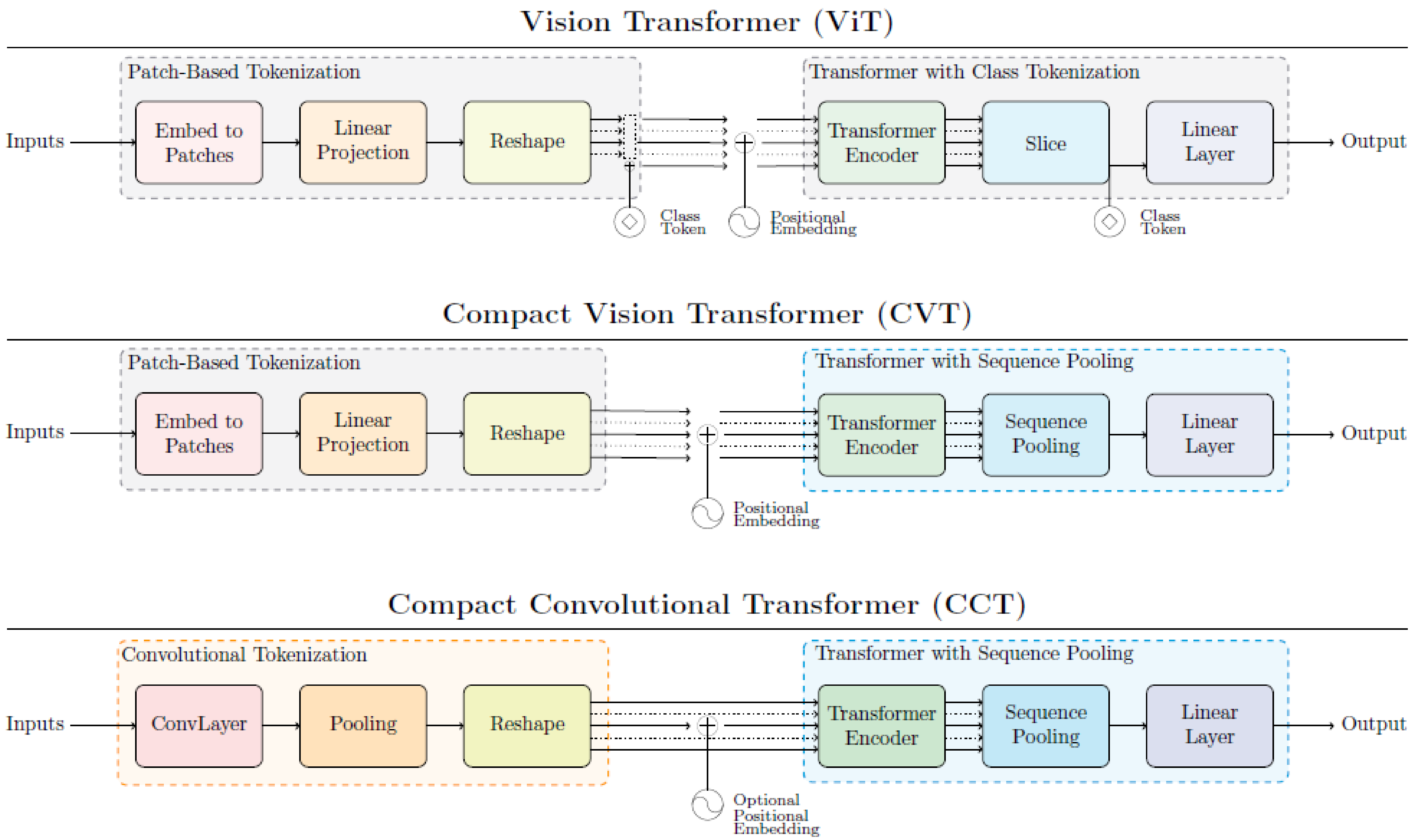
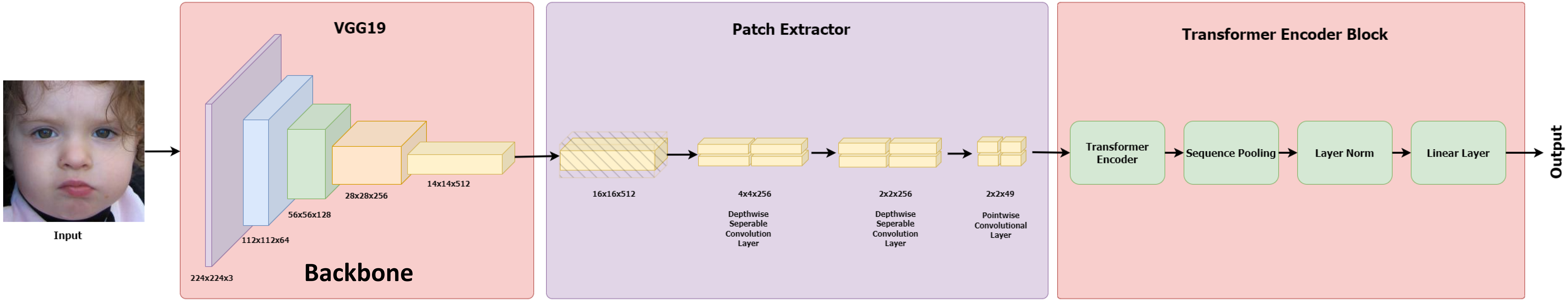


Image Source: A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li and H. Shi, "Escaping the Big Data Paradigm with Compact Transformers," *arXiv preprint arXiv:2104.05704*, 2022.

Cont'



- Backbone Network for extracting fine level features.
- MFN or Truncated VGG19

- Patch extractor is applied to extract relevant features from facial patches.
- The patch extractor consisted of two depthwise separable convolutions and one pointwise convolution

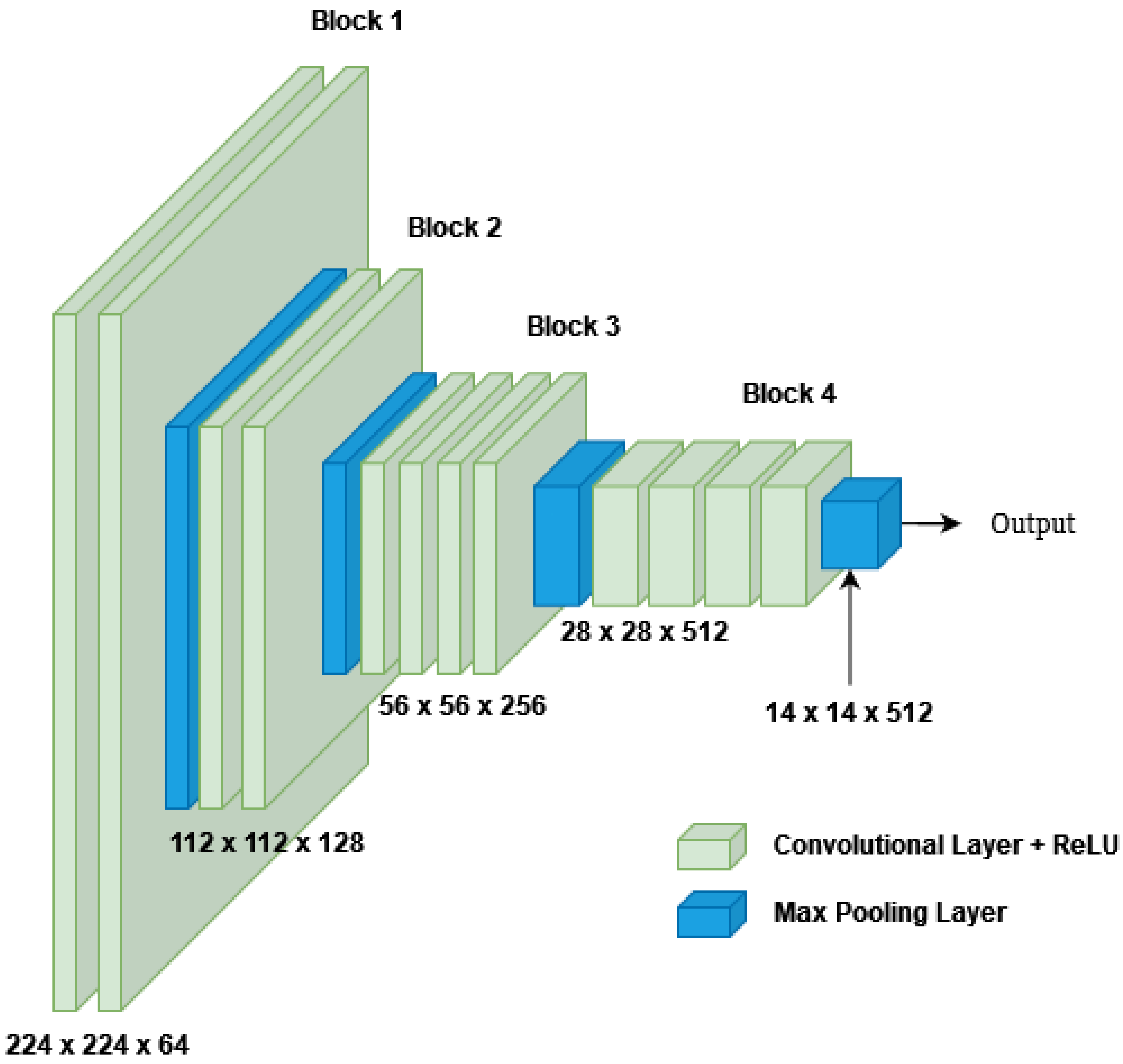
- Compact Convolutional Transformer
- Original Encoder layers followed by Sequence pooling.
- Sequence pooling allows the network to concentrate on specific segments of the input sequence after the encoder has processed it.

T.C. YEDİTEPE ÜNİVERSİTESİ

Cont'

VGG19:

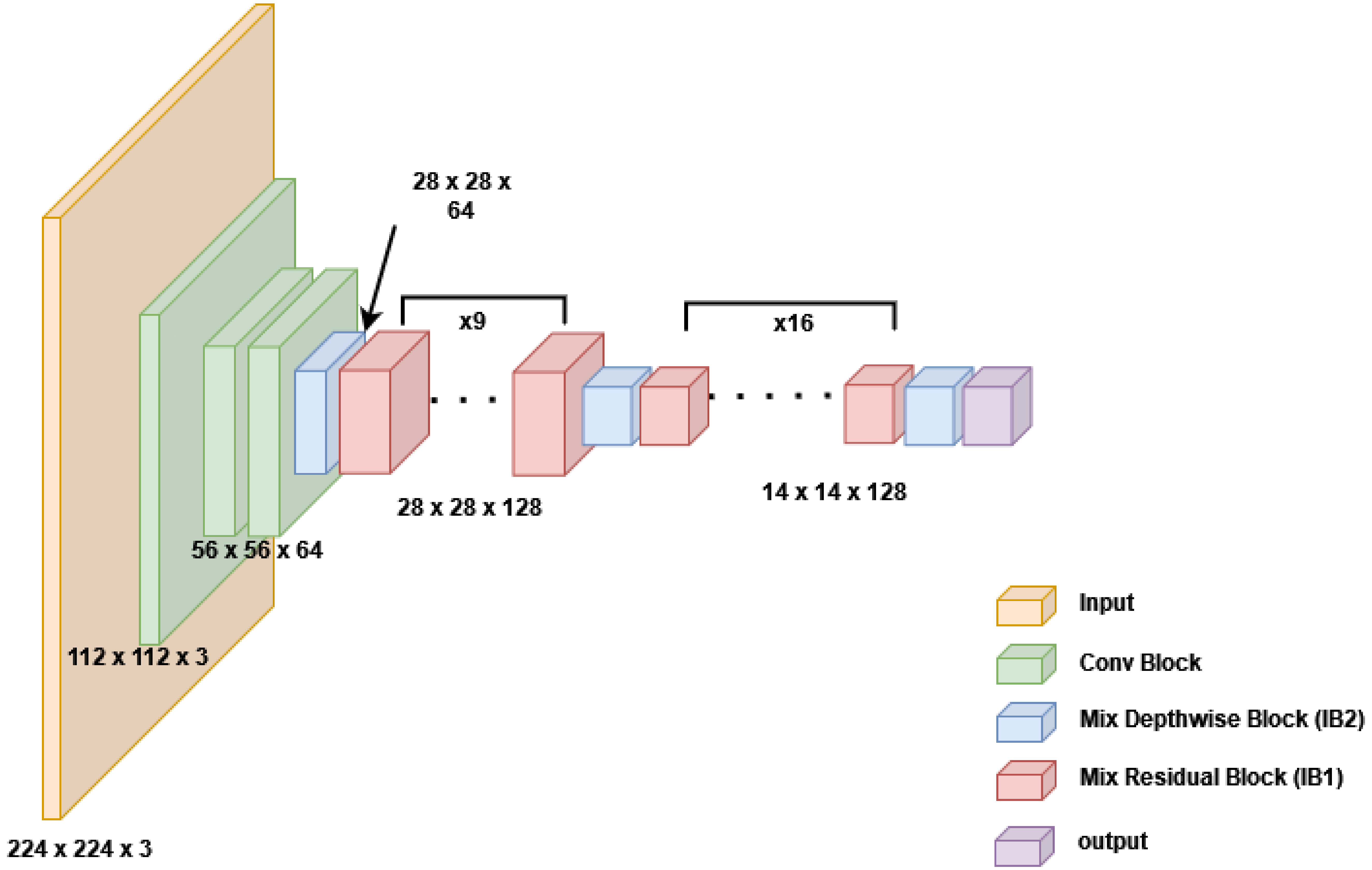
- Truncated after 4th pooling layer
- Simple and uniform architecture
- Applied various tasks and data sets



Cont'

Mixed Feature Network (MFN):

- Lightweight network
- Specifically designed to be suitable for face verification tasks



Data Pre-processing

- Training and evaluation performed on AffectNet database.
- Official small version of dataset employed
- All images collected by querying the web and manually annotated by professionals
- Officially provided validation set used for testing purposes
- 500 examples randomly sampled from each category for validation
- Channel-wise normalization and Random Augmentation is applied.

Labels	Training	Validation	Test
	Number		
Neutral	74374	500	500
Happy	133915	500	500
Sad	24959	500	500
Surprise	13590	500	500
Fear	5878	500	500
Disgust	3303	500	500
Anger	24382	500	500
Contempt	3250	500	499

T.C. YEDİTEPE ÜNİVERSİTESİ

Cont'


Original Image
RandAugment (n=2, m=9)
RandAugment (n=5, m=9)
RandAugment (n=2, m=16)
RandAugment (n=5, m=16)
Original Image
RandAugment (n=2, m=9)
RandAugment (n=5, m=9)
RandAugment (n=2, m=16)
RandAugment (n=5, m=16)
Original Image
Original Image (Normalized)
Normalized & RandAug (n=2, m=9)
Normalized & RandAug (n=5, m=9)
Normalized & RandAug (n=2, m=16)
Normalized & RandAug (n=5, m=16)

Cont'

- Dataset has a huge imbalance between classes.
- To address this problem oversampling is applied.
- N_c: Total number of examples in the dataset for category c

$$w_c = \frac{1}{N_c}$$

Class	Weight
Neutral	0.477
Happiness	0.265
Sadness	1.421
Surprise	2.609
Fear	6.032
Disgust	10.735
Anger	1.454
Contempt	10.91

Training Procedure

- Batch size: 32
- Optimizer AdamW with β_1=0.9 and β_2=0.999
- Learning rate = 0.001
- 12 Model is trained using 6 variants for at least 10 epoch
- First letter denotes the backbone network
- The number represents the number of encoder layers
- Models with a * trained without oversampling

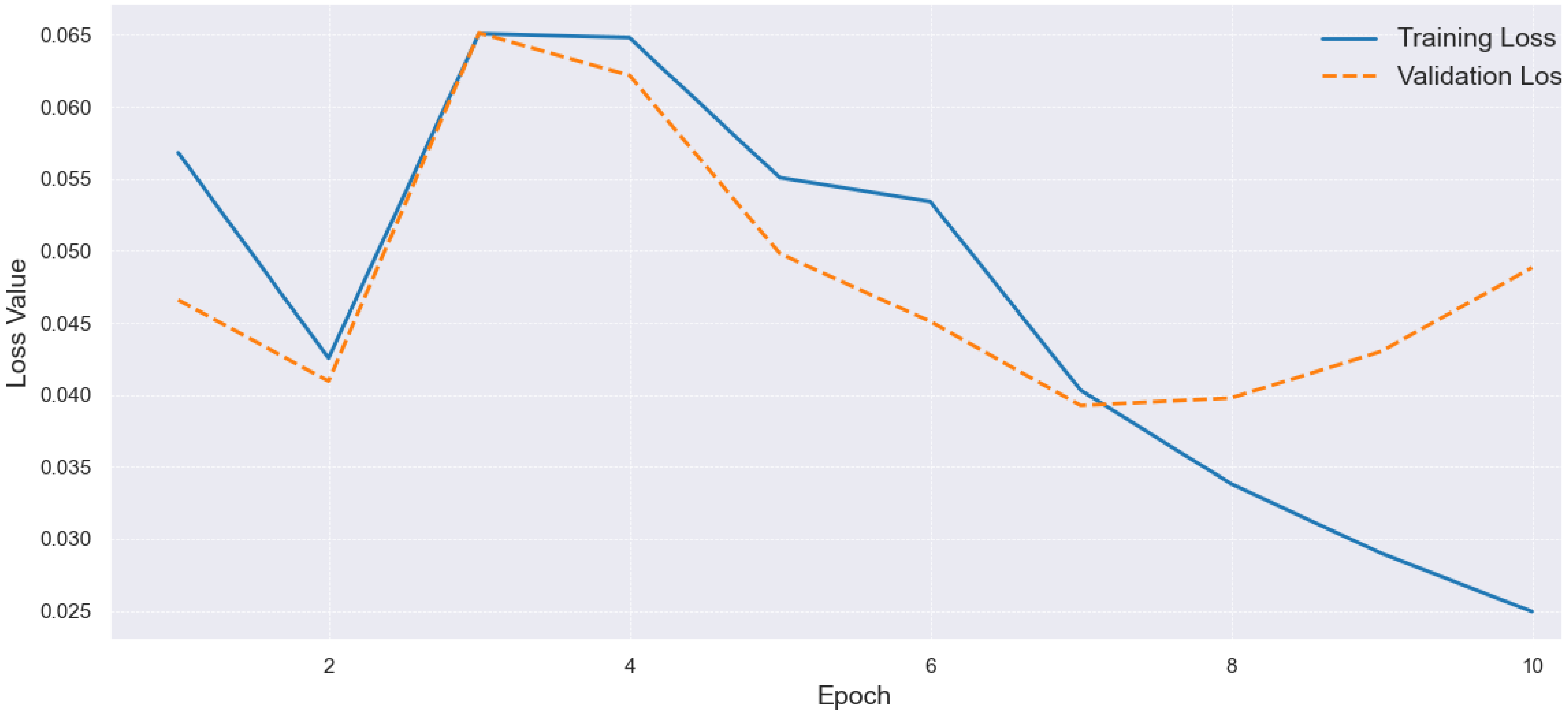
Results

Model	Performance	#Params
VCCT-1	%53.49	11.26M
VCCT-2	%53.41	11.71M
*MCCT-1	%48.19	3.00M
MCCT-1	%53.24	3.00M
*MCCT-2	%44.16	3.44M
MCCT-2	% 54.19	3.44M
*MCCT-3	%41.49	3.89M
MCCT-3	%51.21	3.89M
*MCCT-6	% 45.26	5.23M
MCCT-6	% 52.79	5.23M
*MCCT-7	% 40.79	5.67M
MCCT-7	% 55.54	5.67M

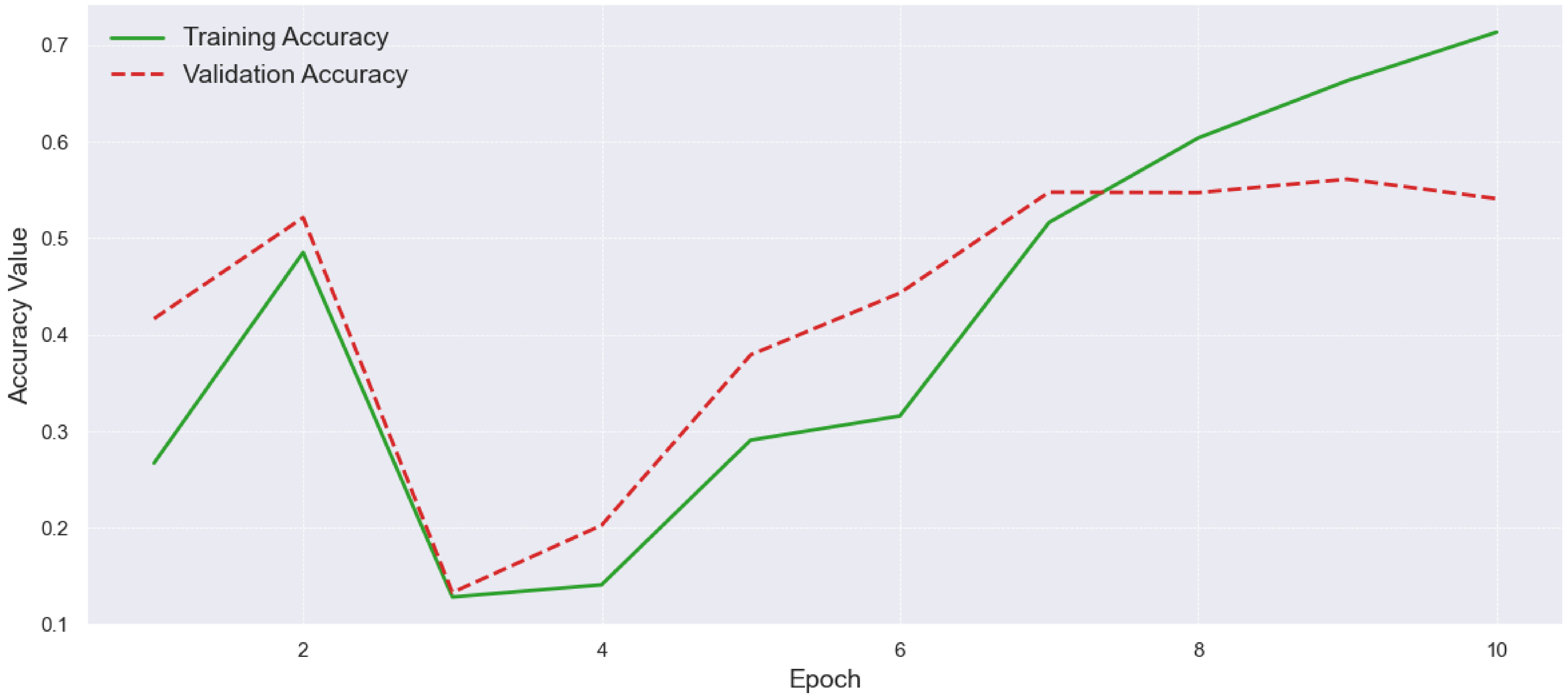
T.C. YEDİTEPE ÜNİVERSİTESİ

Cont'

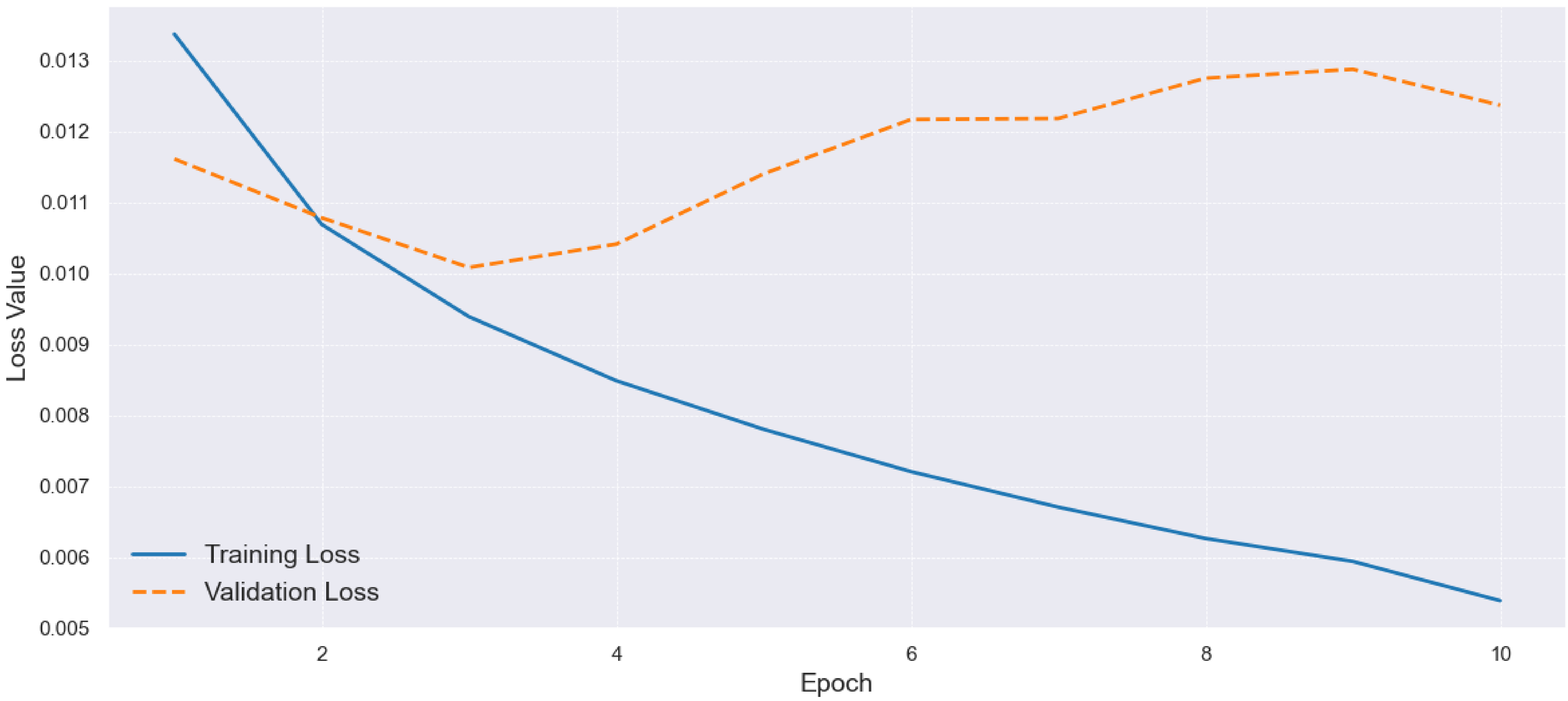
With Oversampling Training & Validation Loss



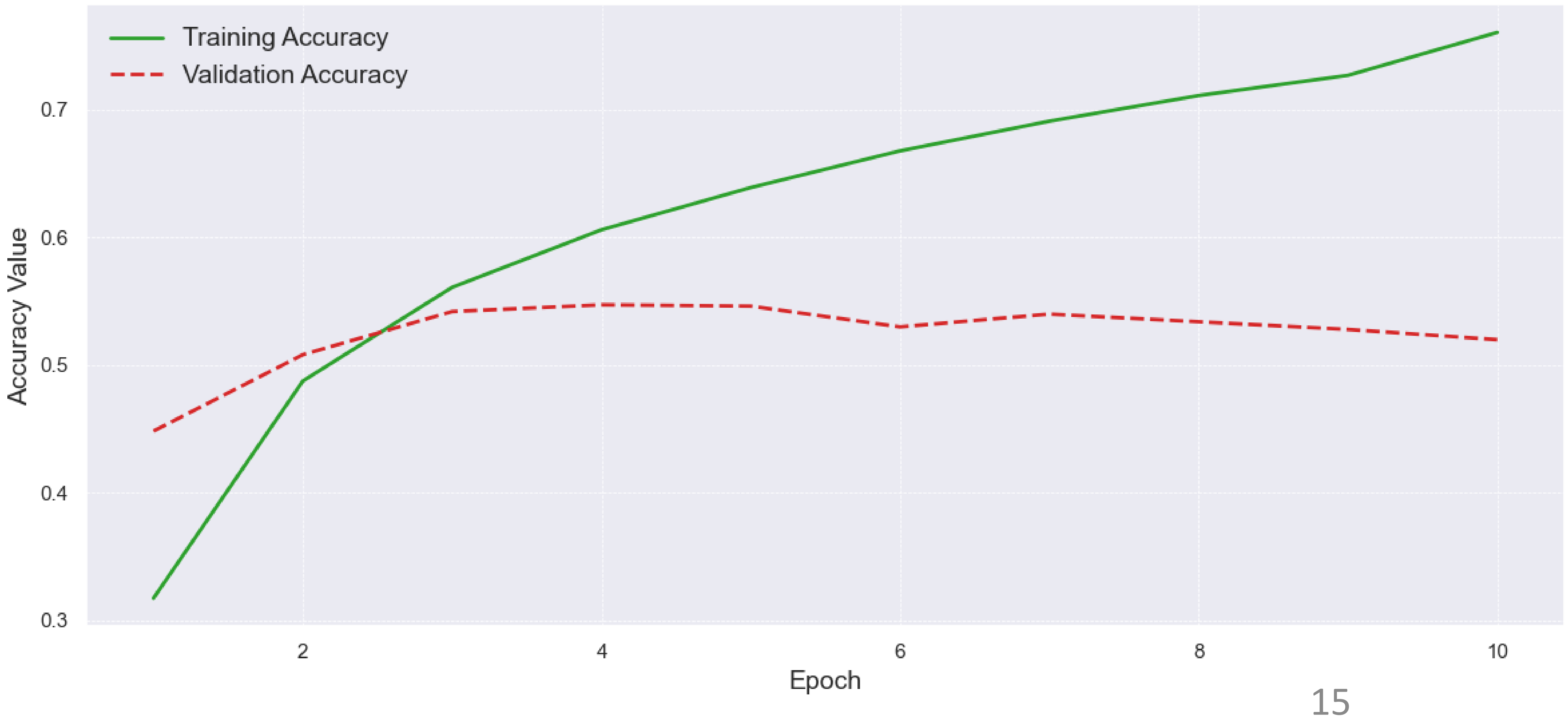
VCCT-1 With Oversampling Training & Validation Accuracy



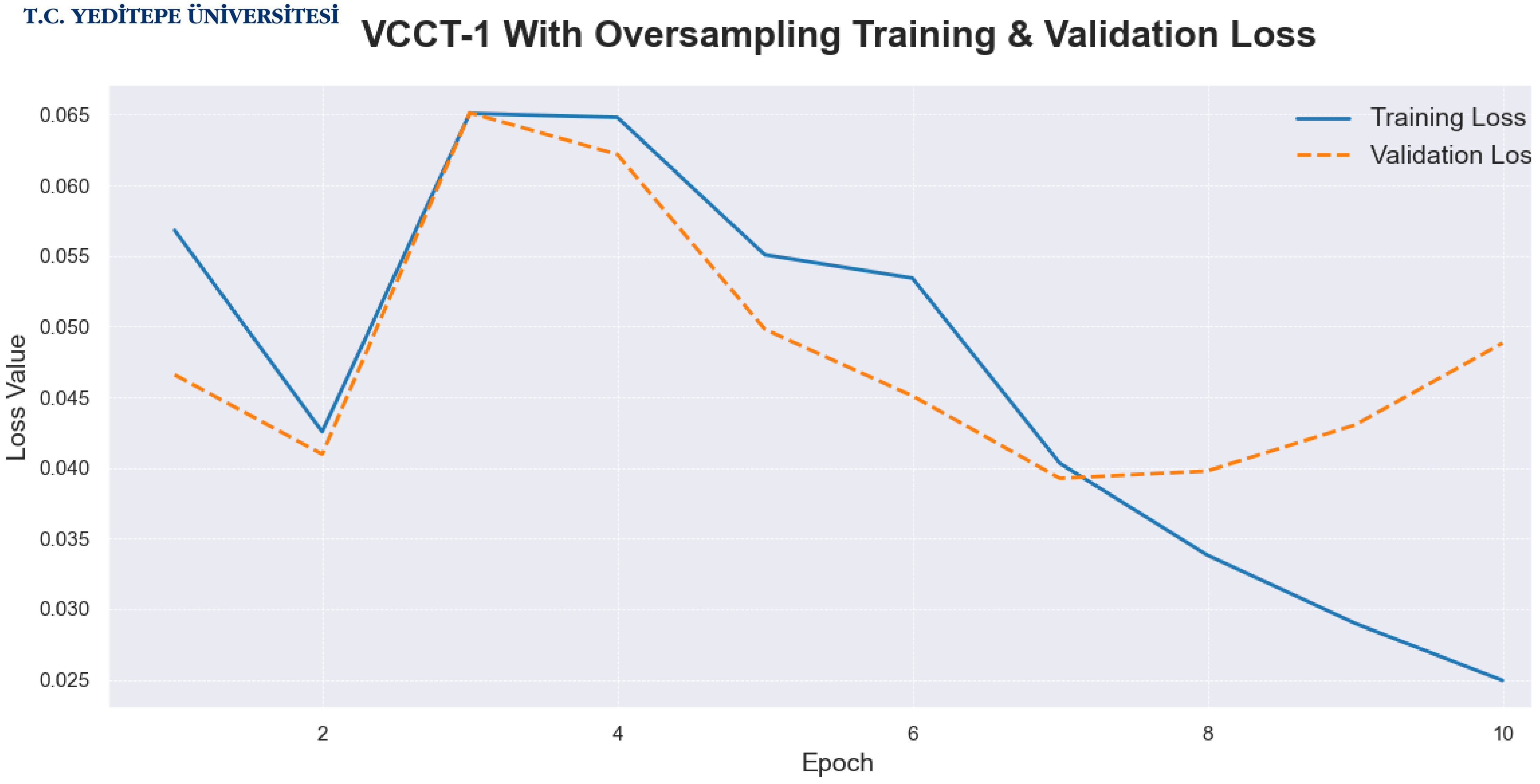
MCCT-1 With Oversampling Training & Validation Loss

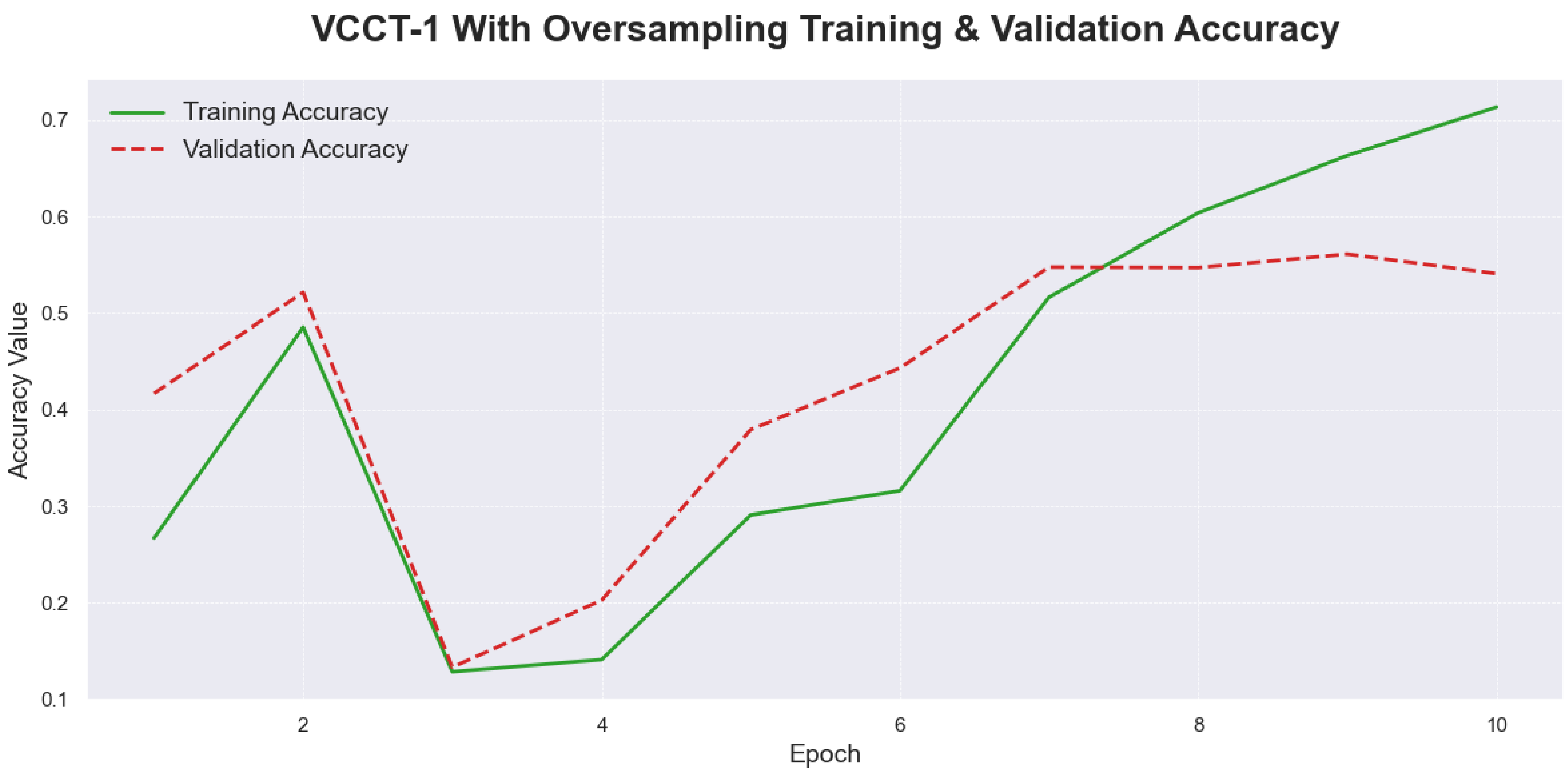


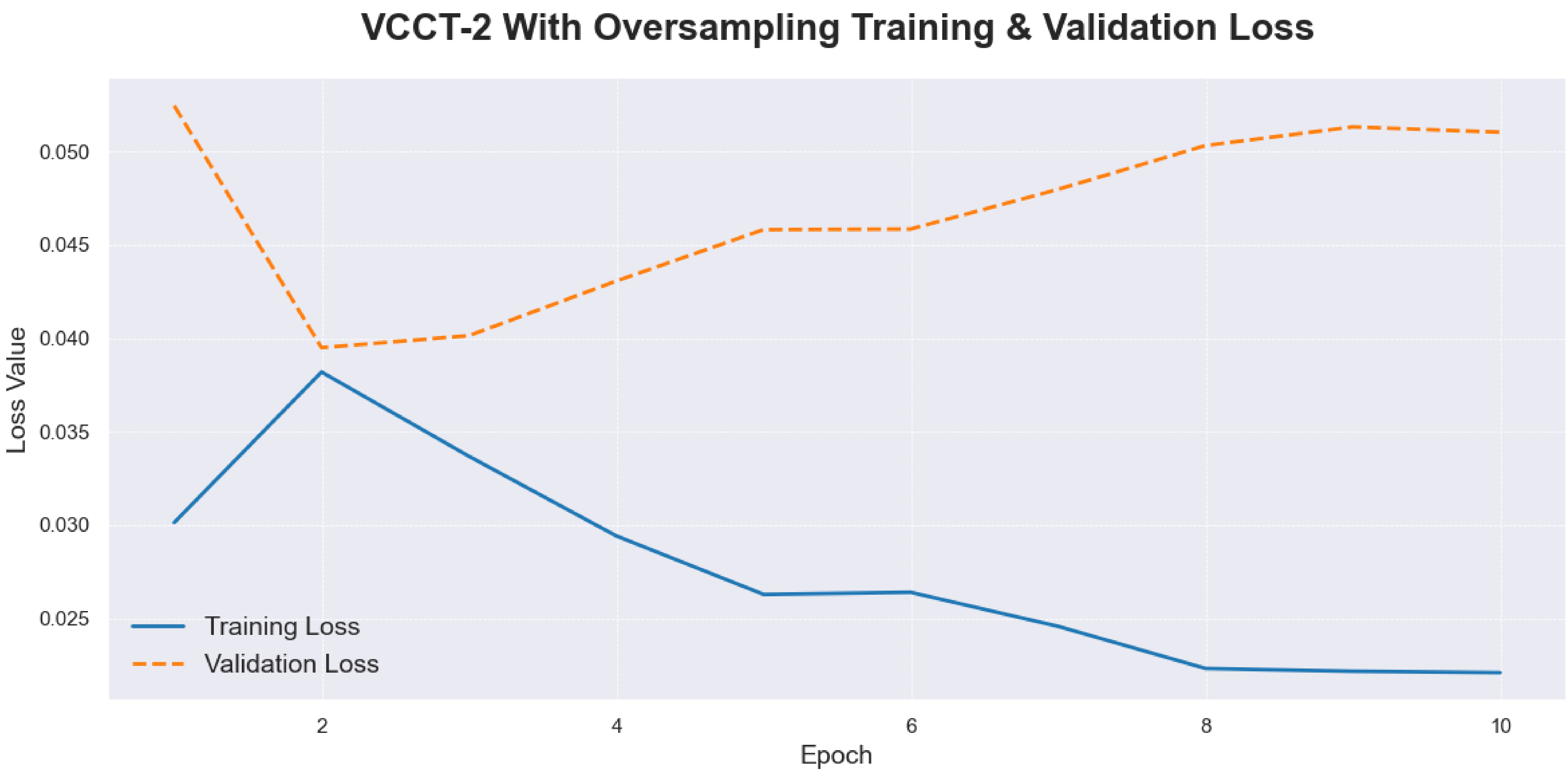
MCCT-1 With Oversampling Training & Validation Accuracy

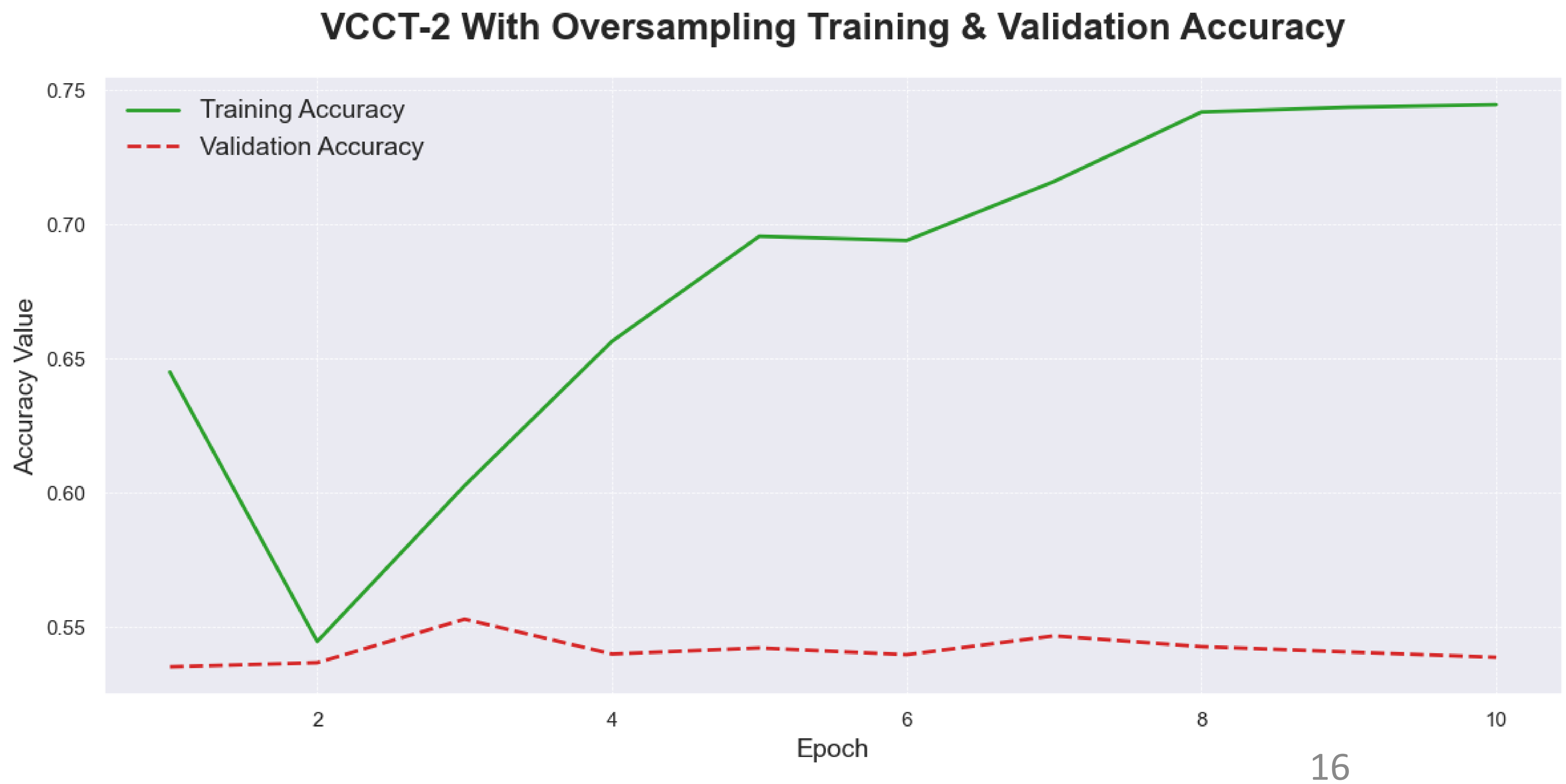


Cont'

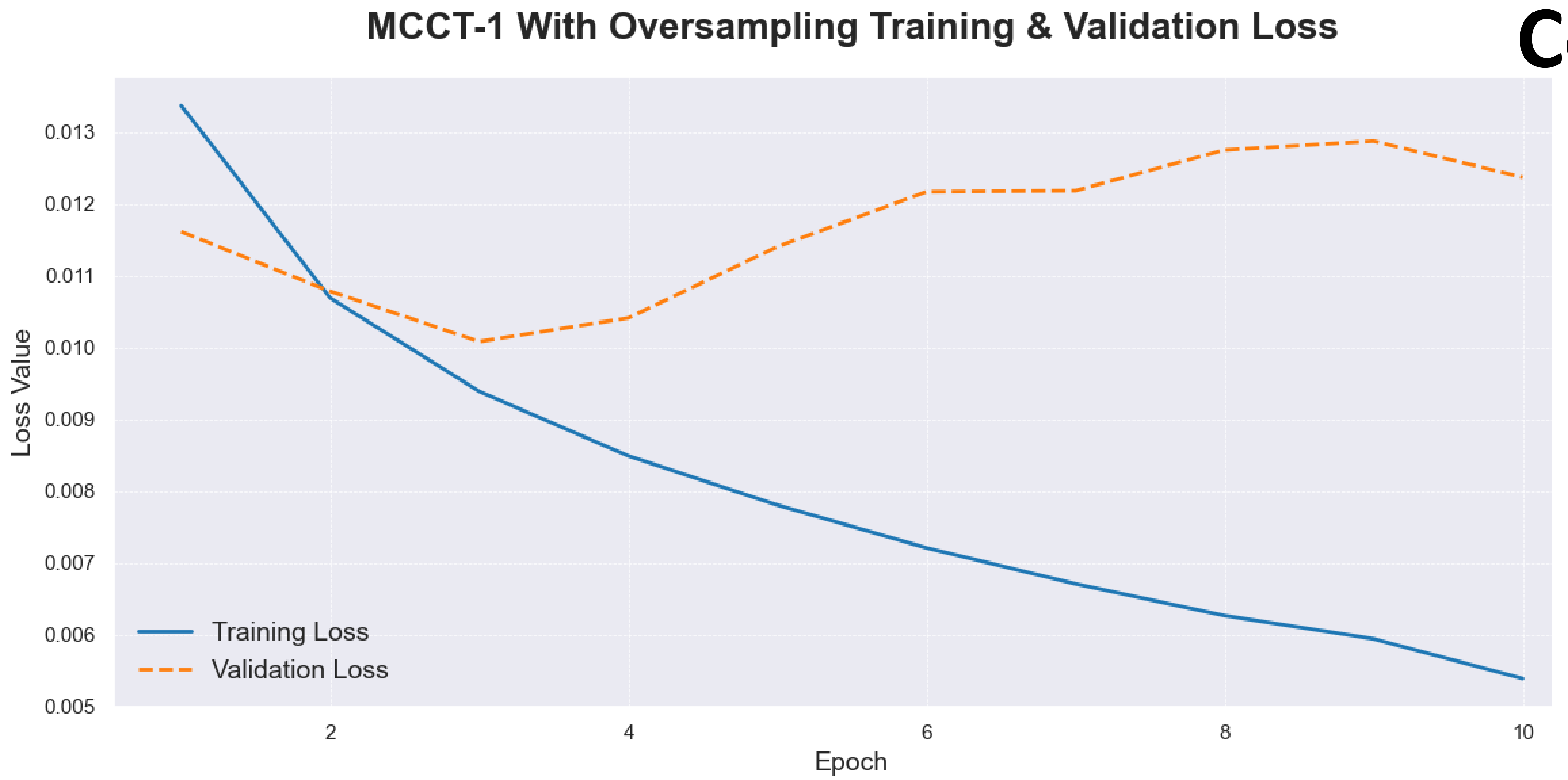

VCCT-1 With Oversampling Training & Validation Loss
Training Loss
Validation Los
Loss Value
Epoch

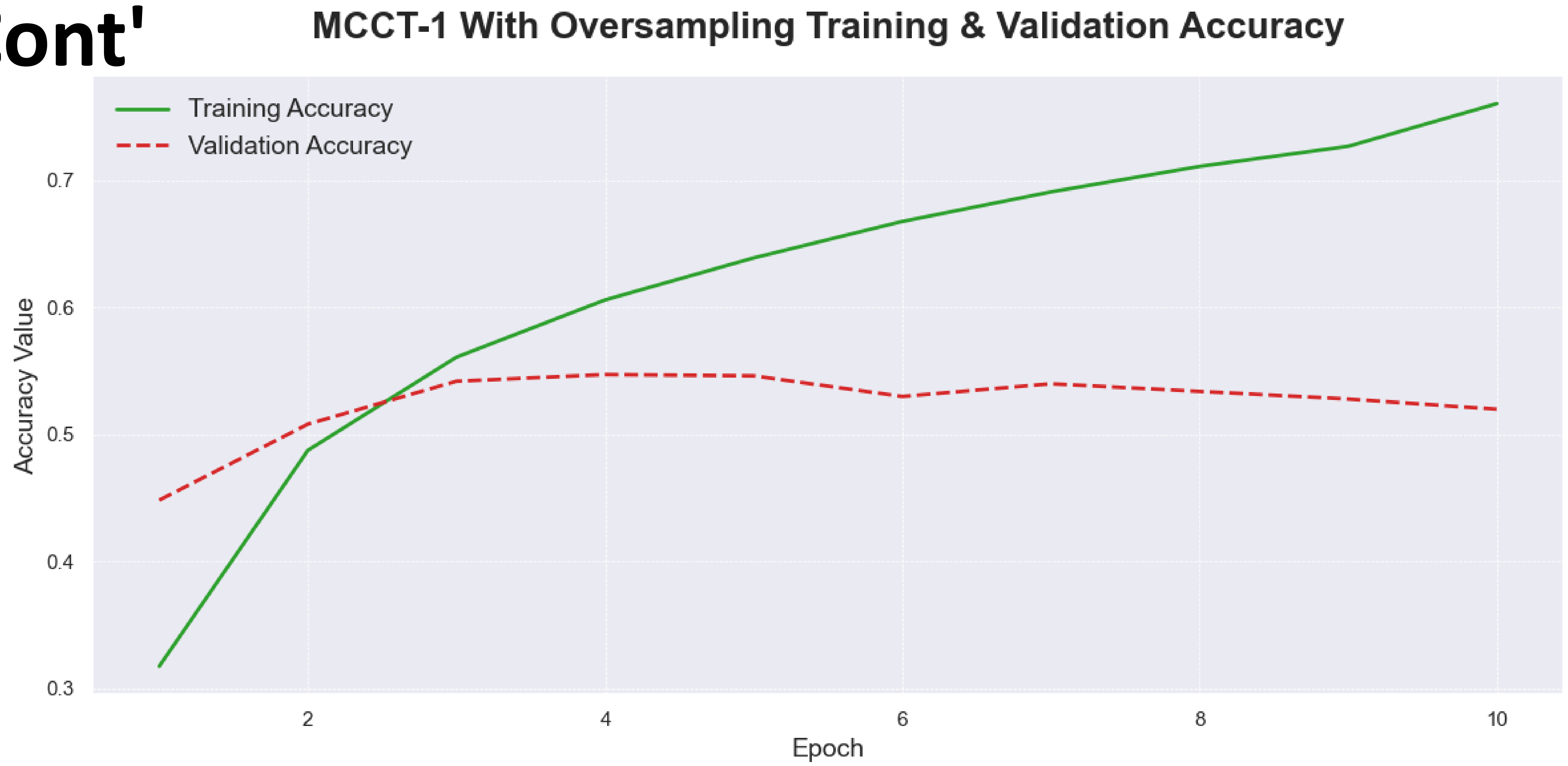

VCCT-1 With Oversampling Training & Validation Accuracy
Training Accuracy
Validation Accuracy
Accuracy Value
Epoch

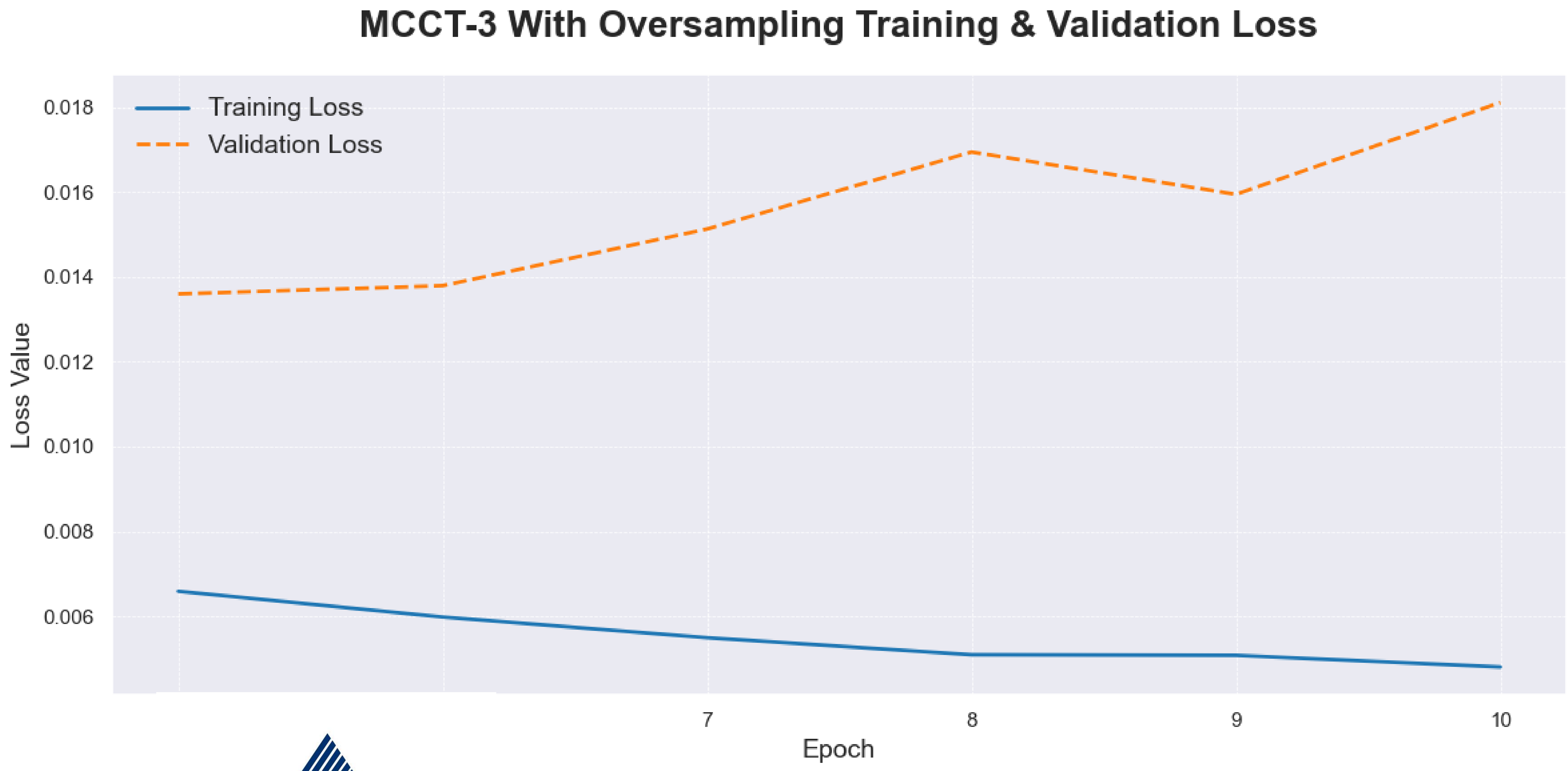

VCCT-2 With Oversampling Training & Validation Loss
Training Loss
Validation Loss
Loss Value
Epoch

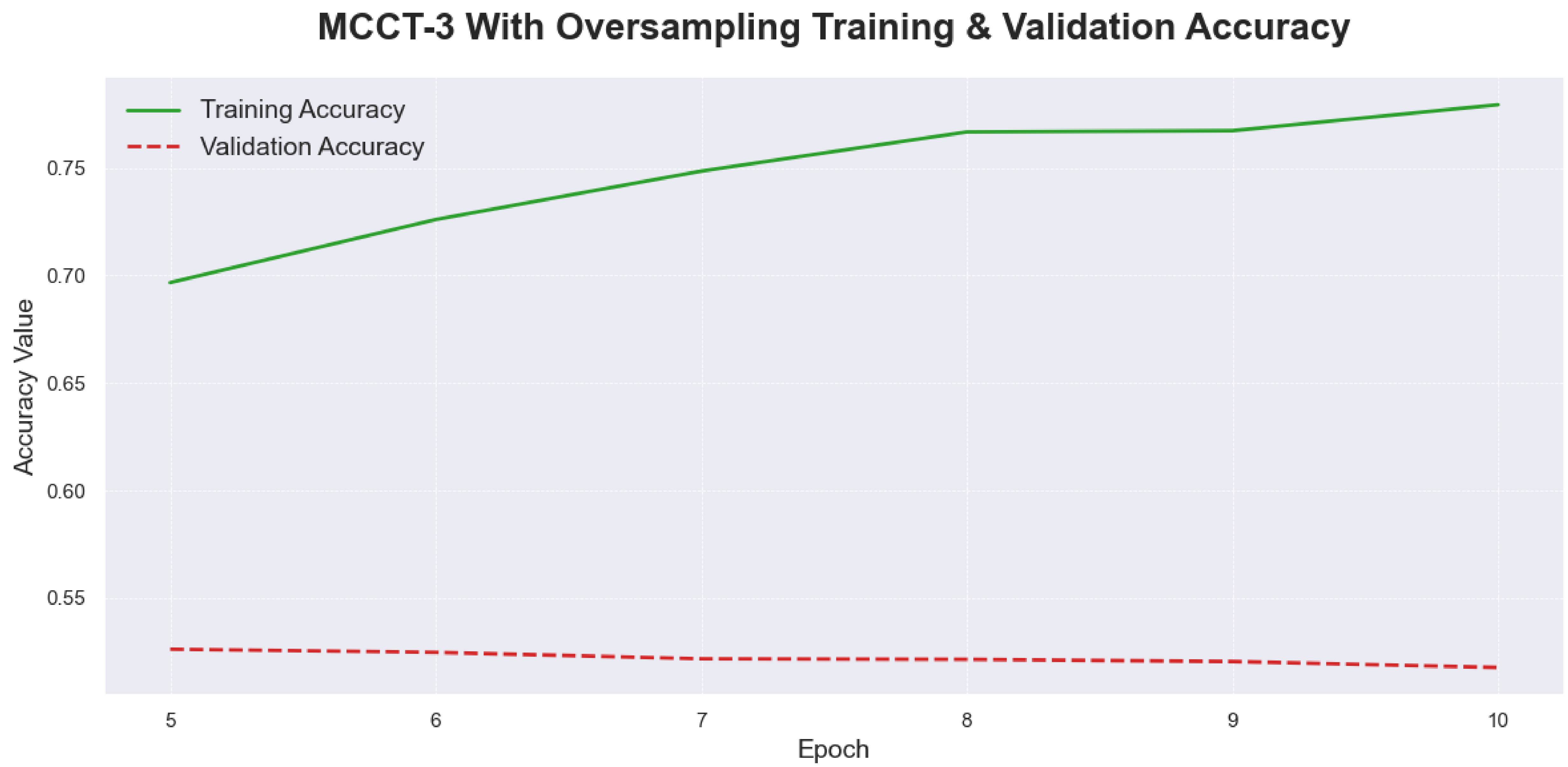

VCCT-2 With Oversampling Training & Validation Accuracy
Training Accuracy
Validation Accuracy
Accuracy Value
Epoch

Cont'



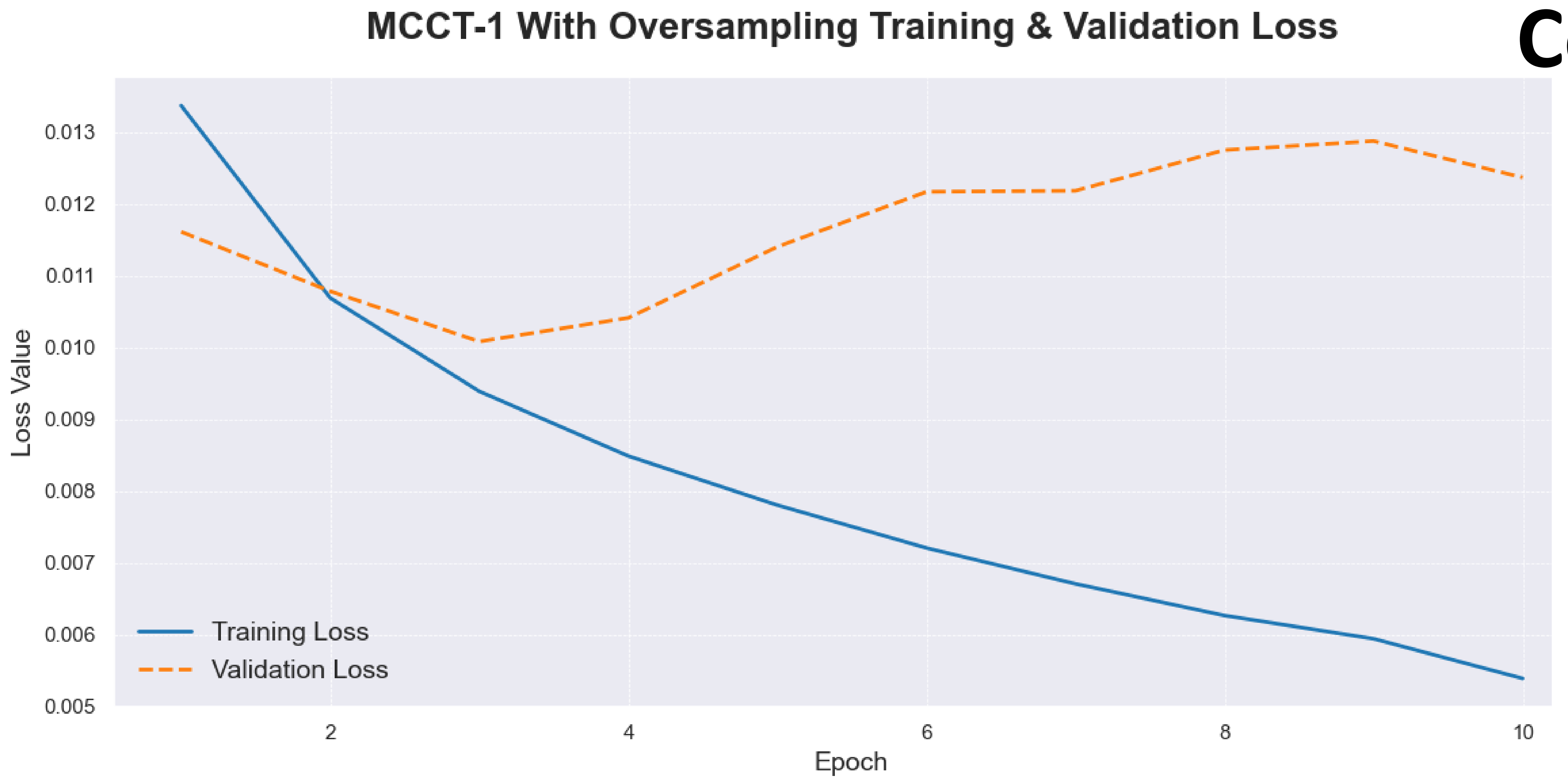


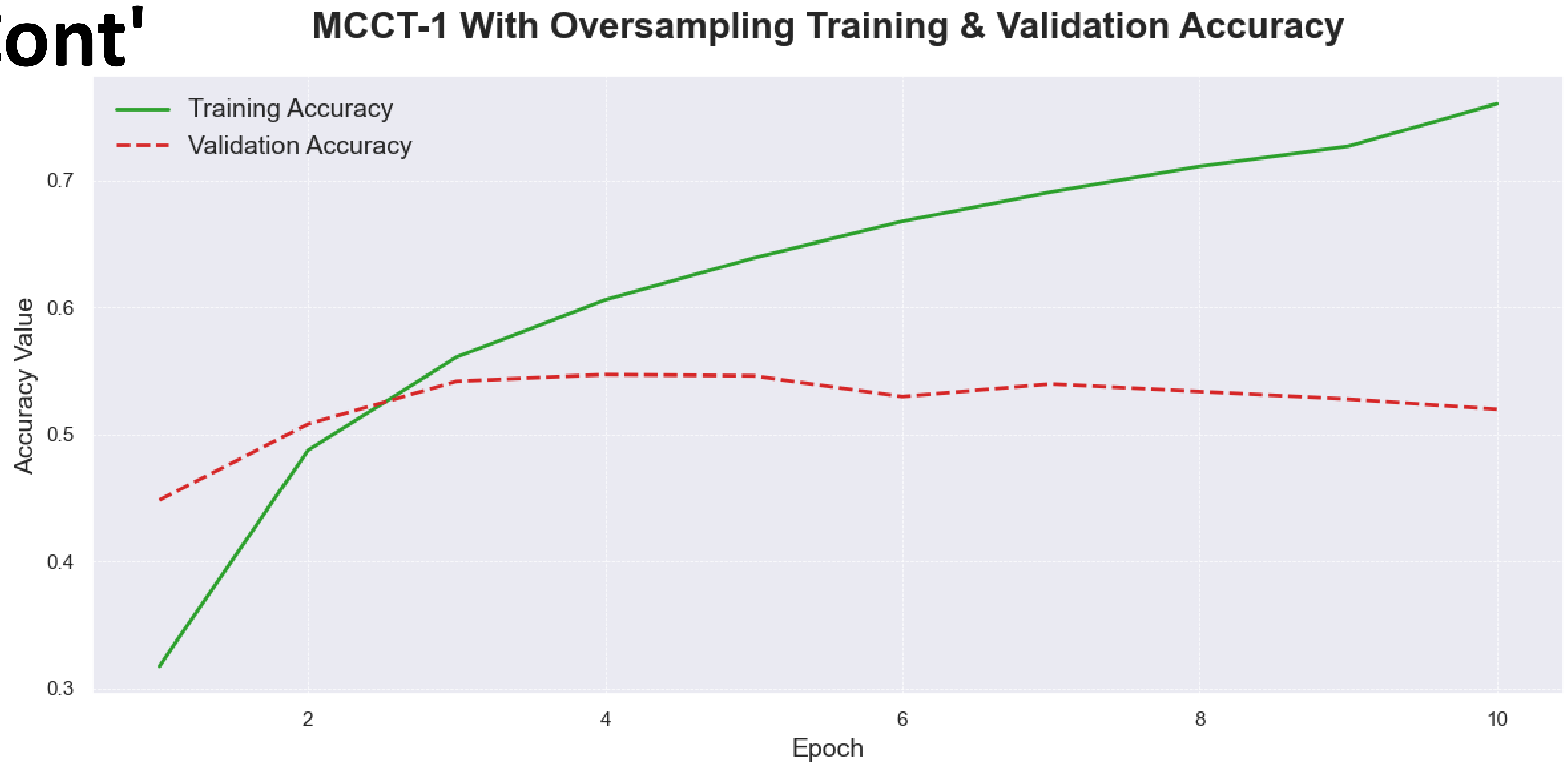


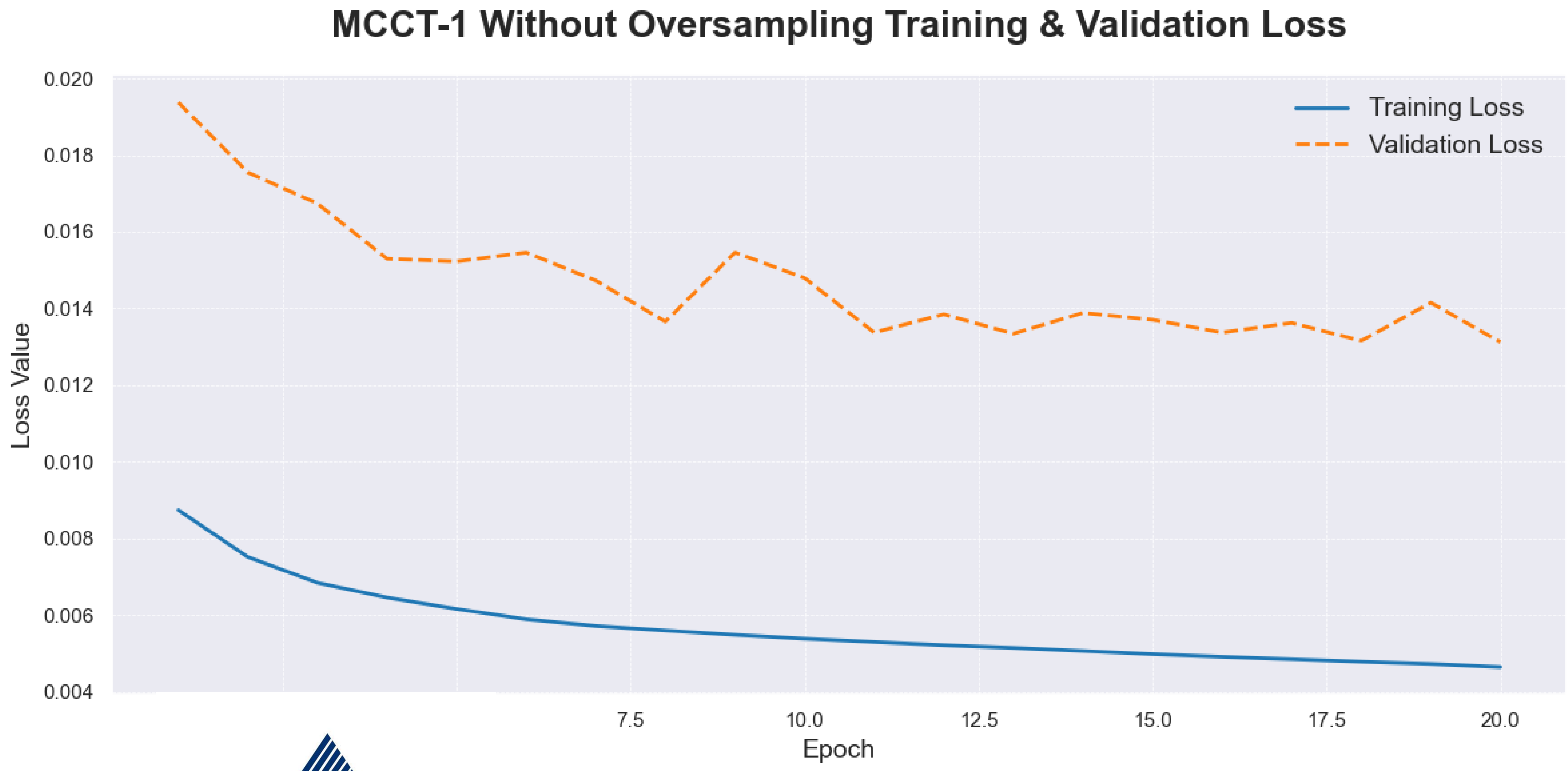


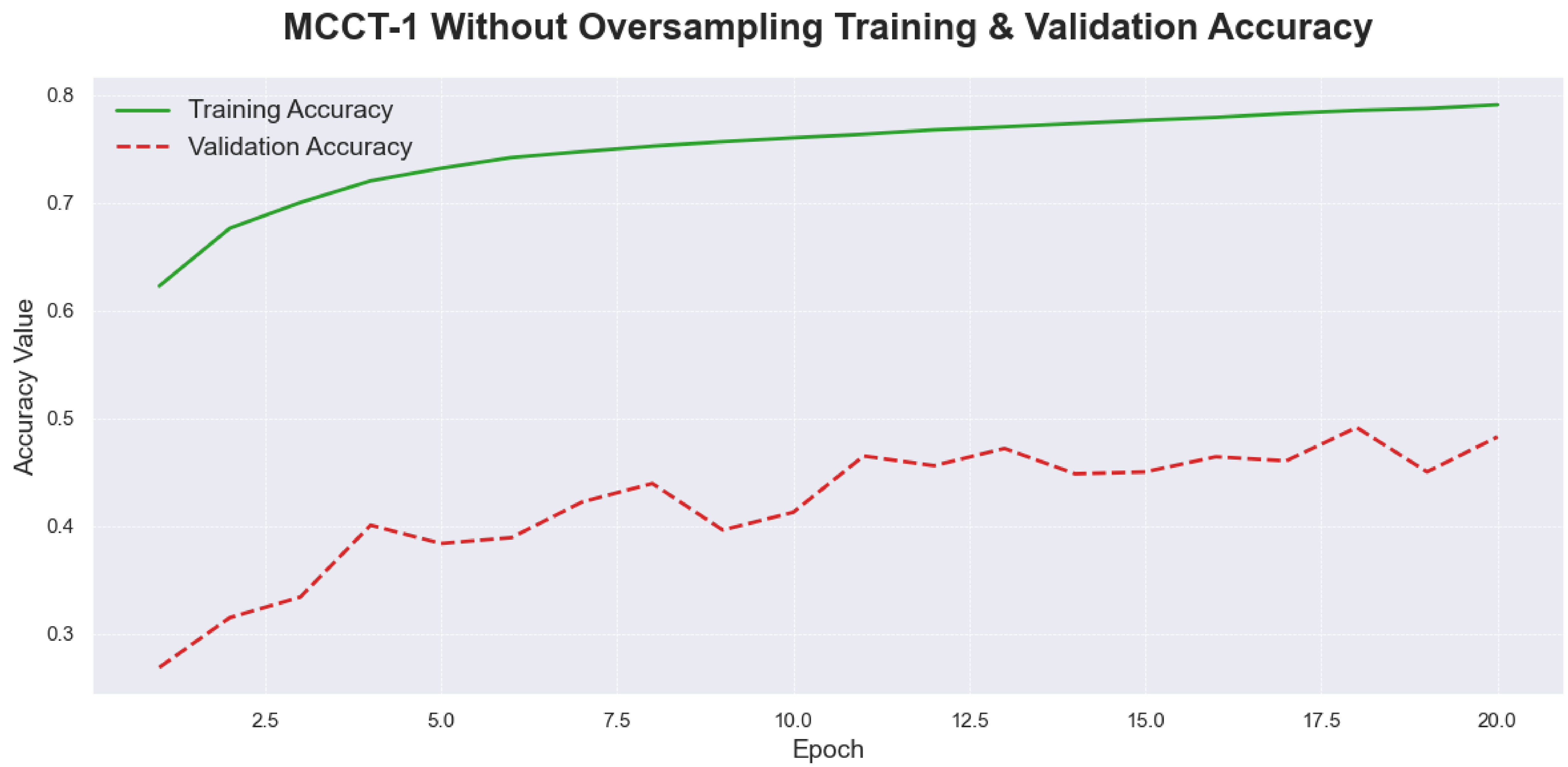
T.C. YEDİTEPE ÜNİVERSİTESİ

Cont'



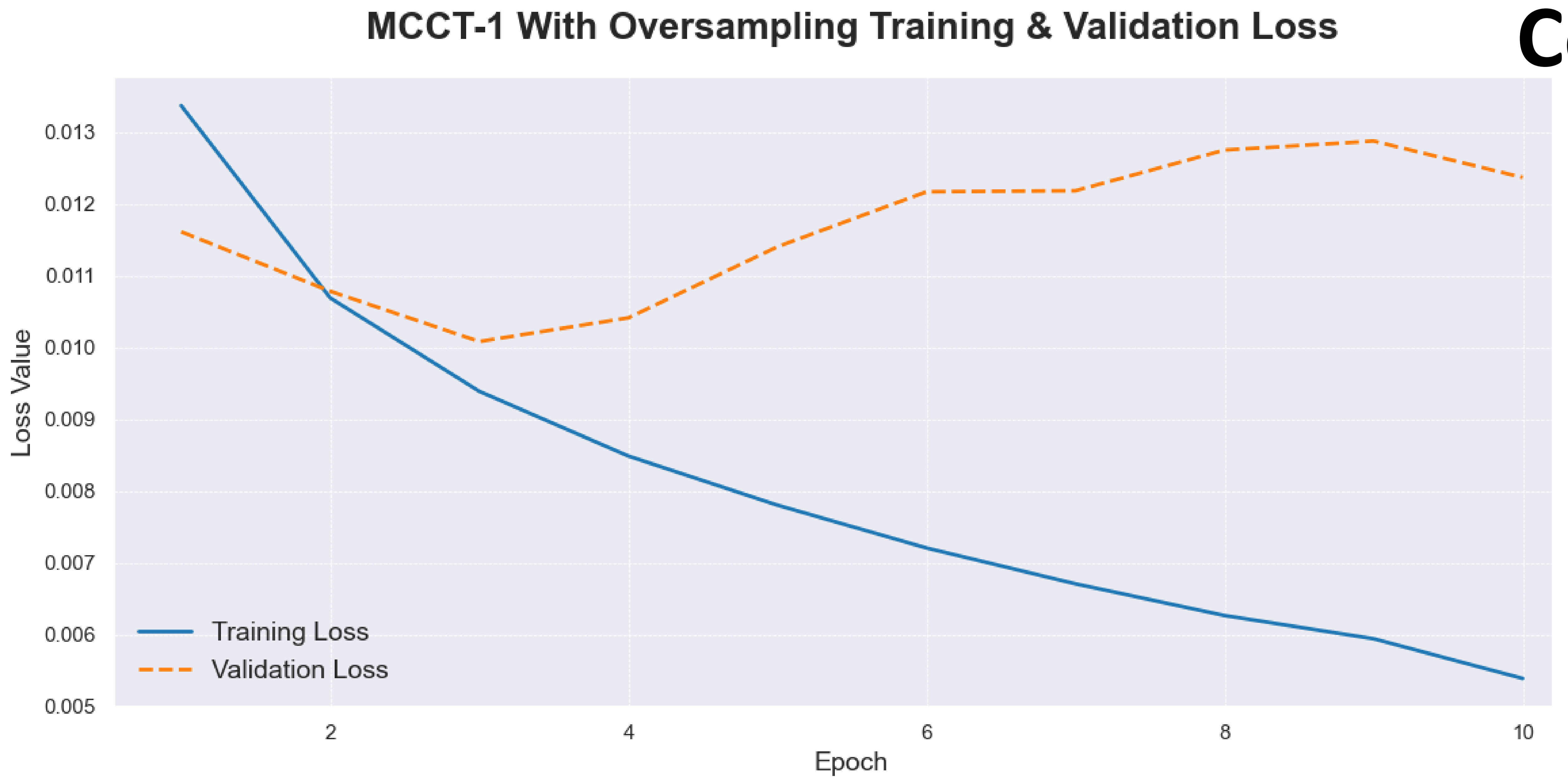


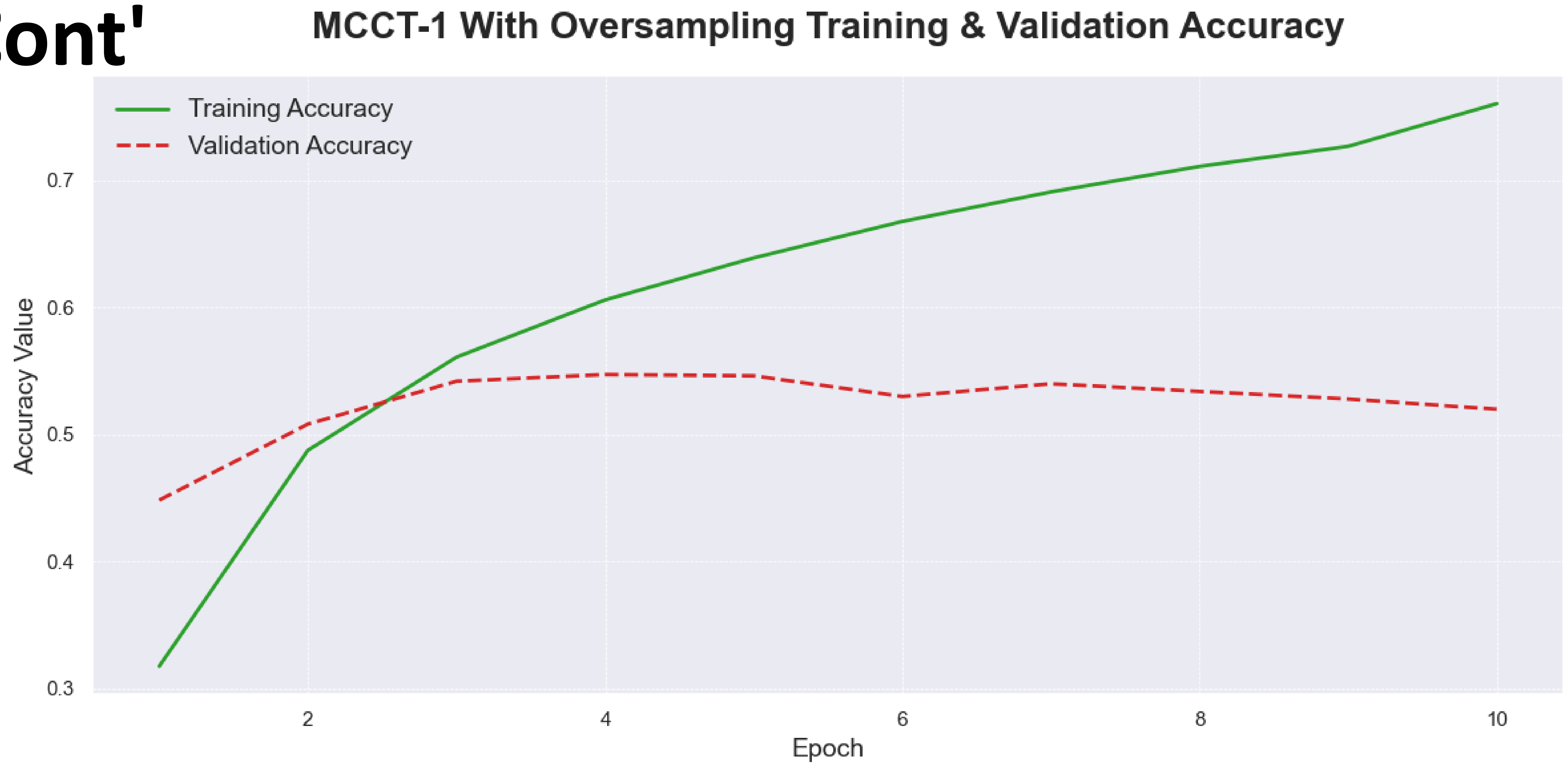


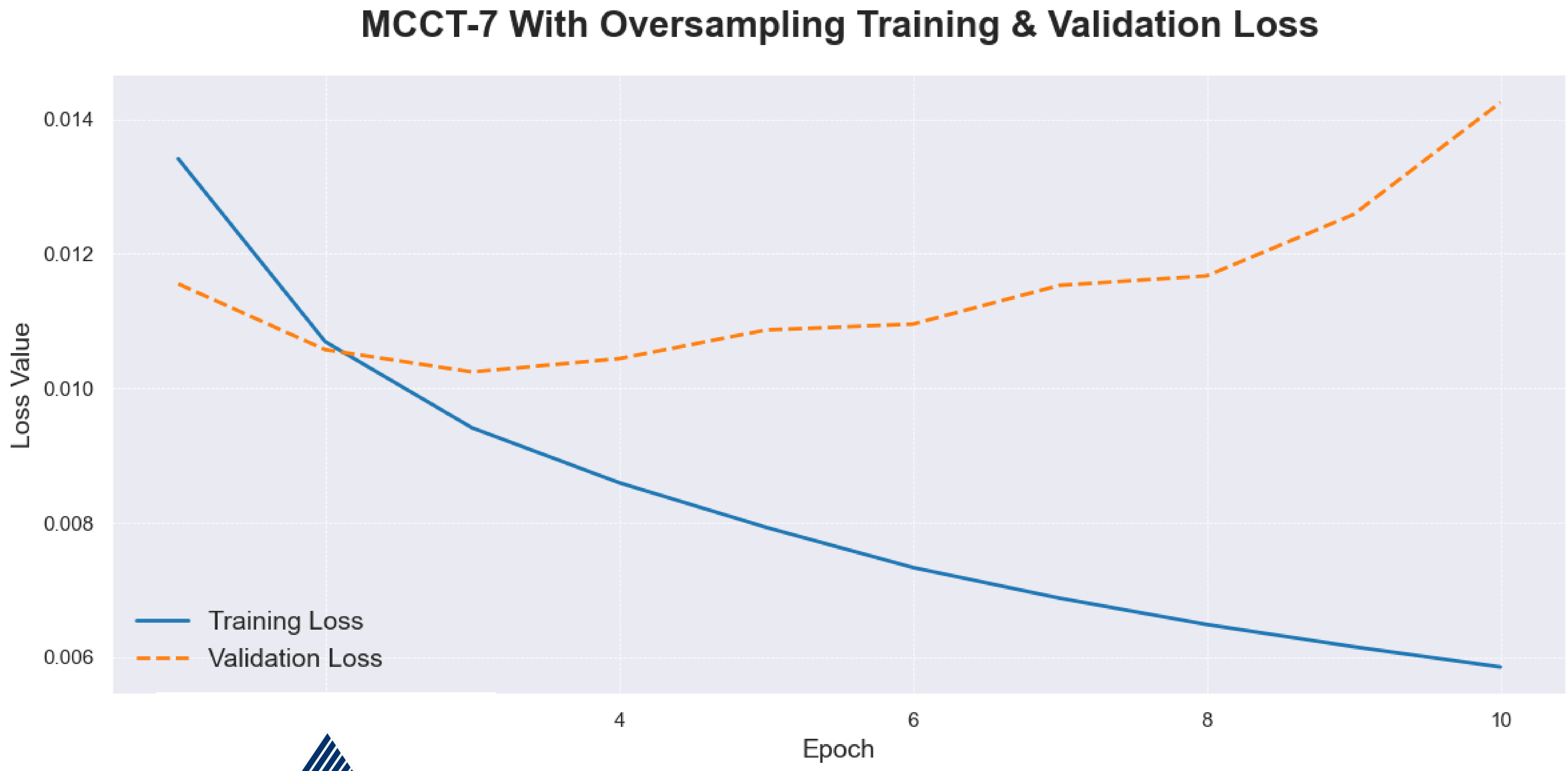


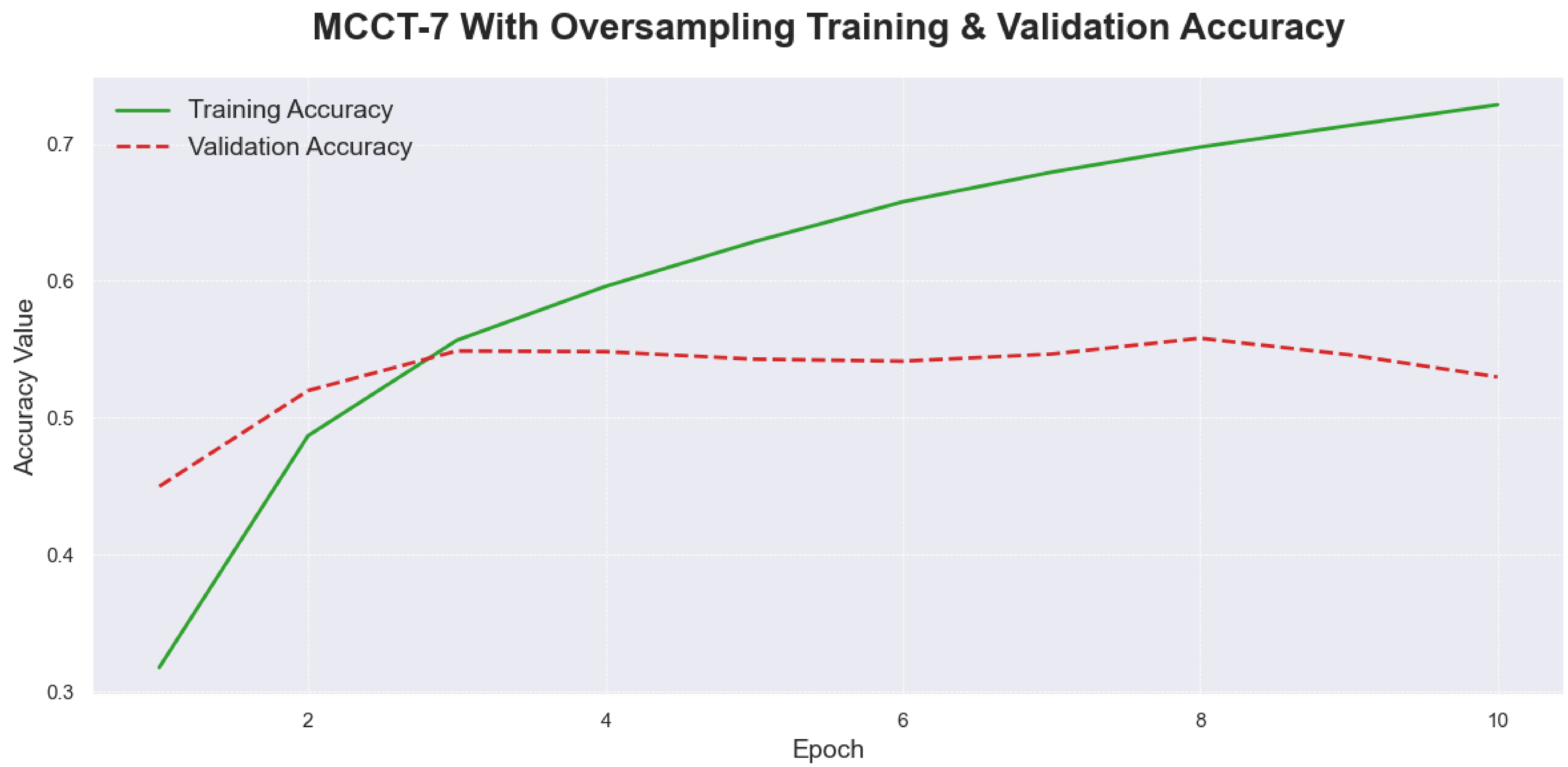
T.C. YEDİTEPE ÜNİVERSİTESİ

Cont'



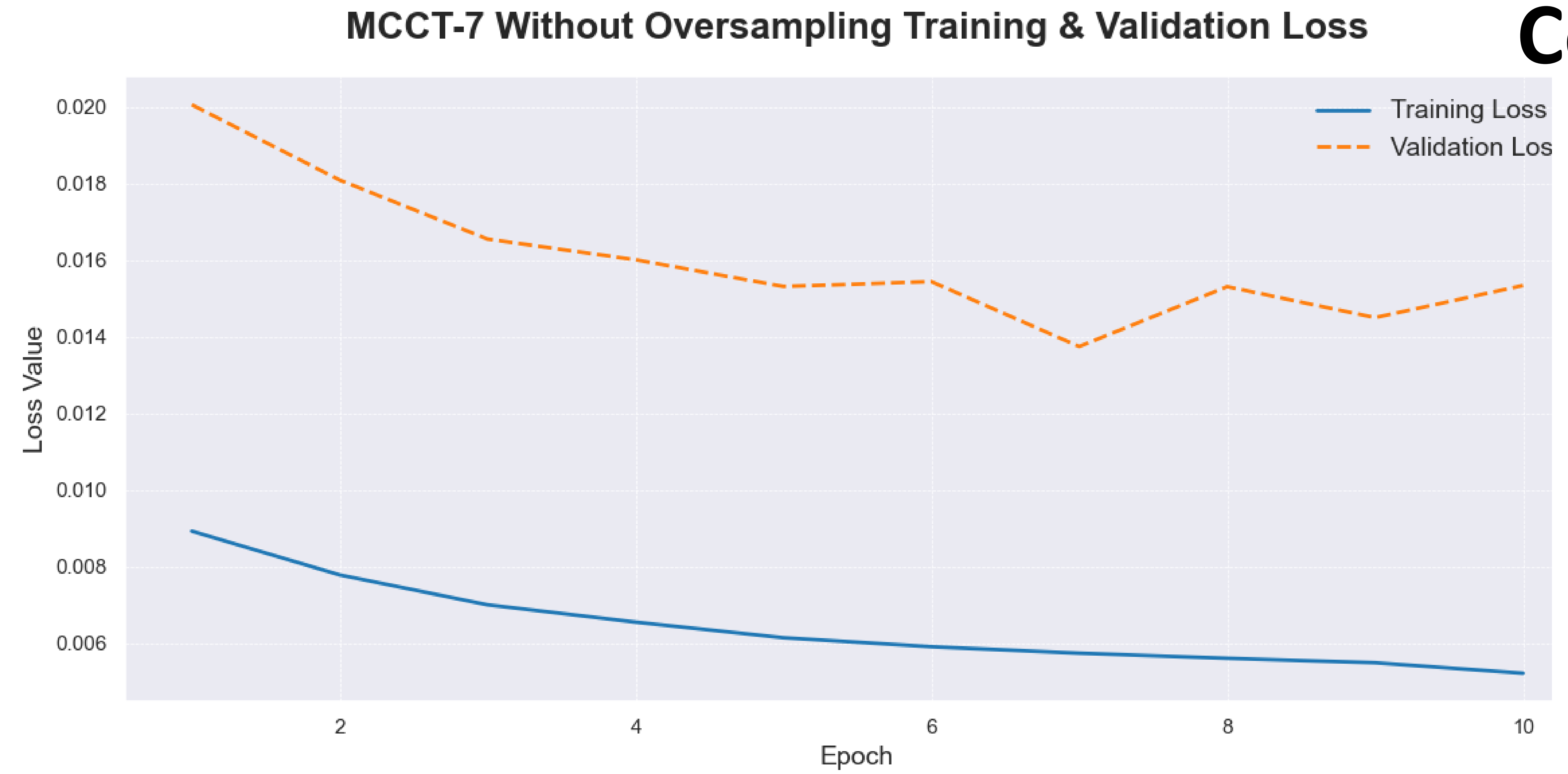


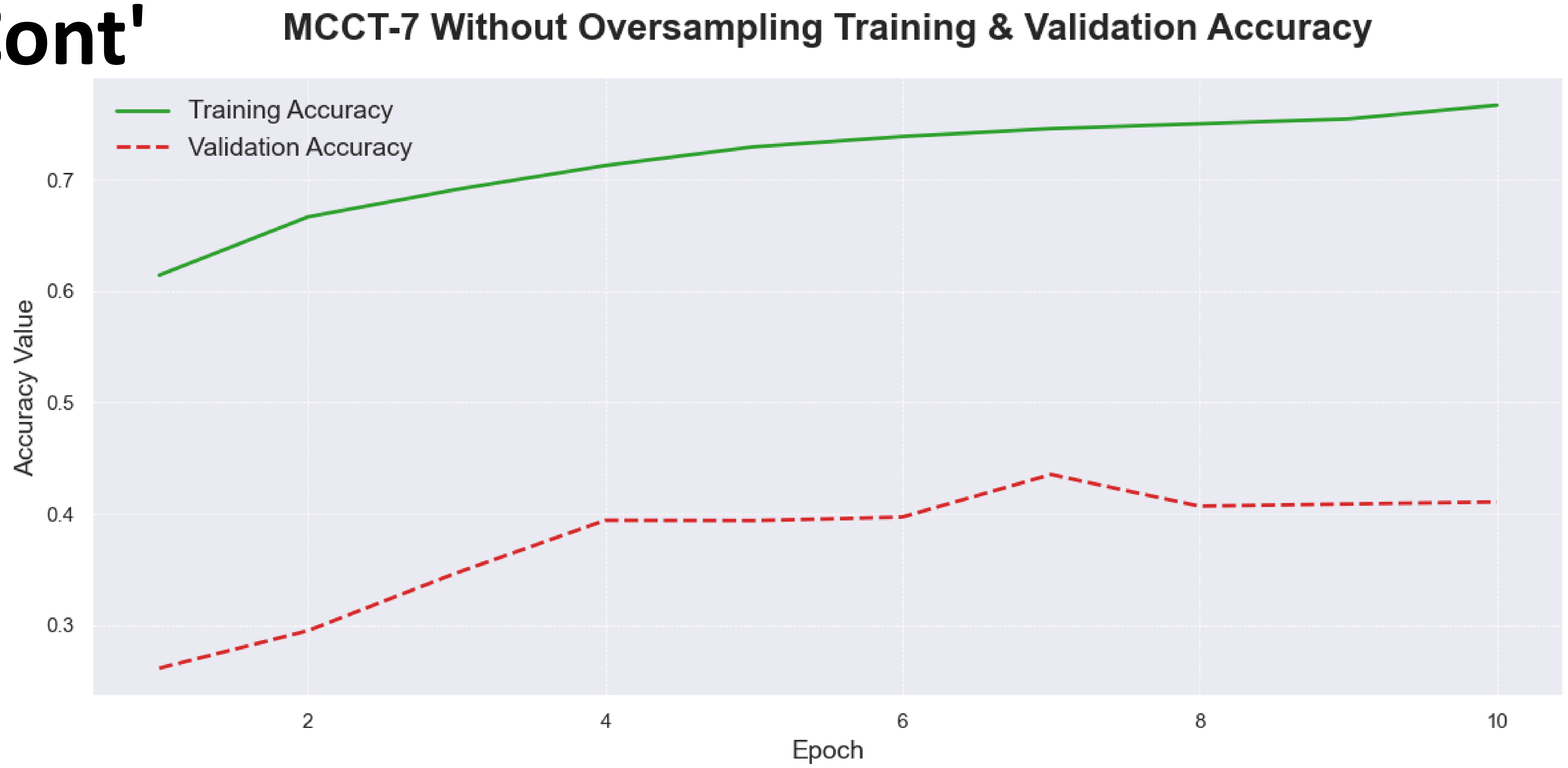


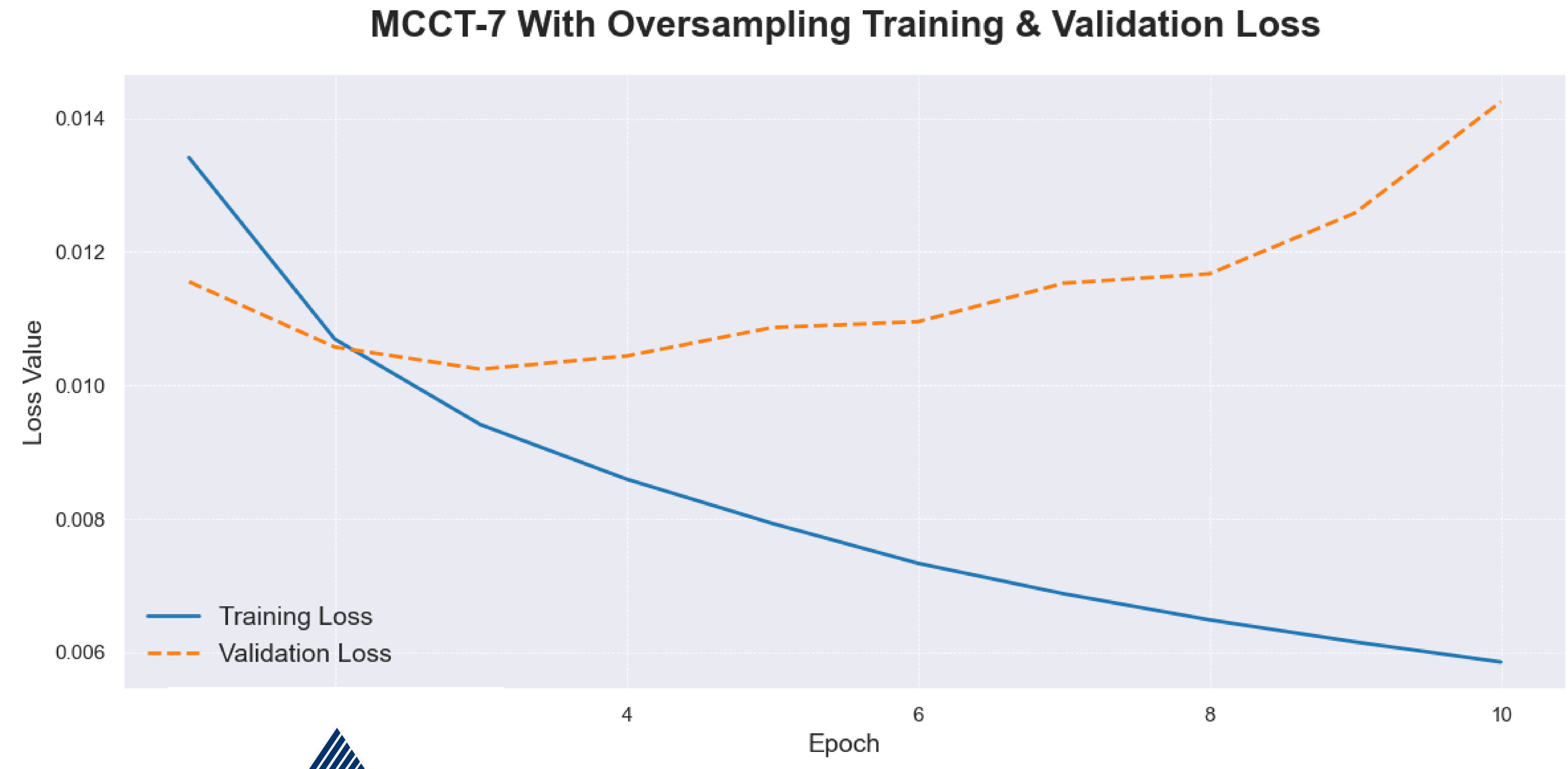


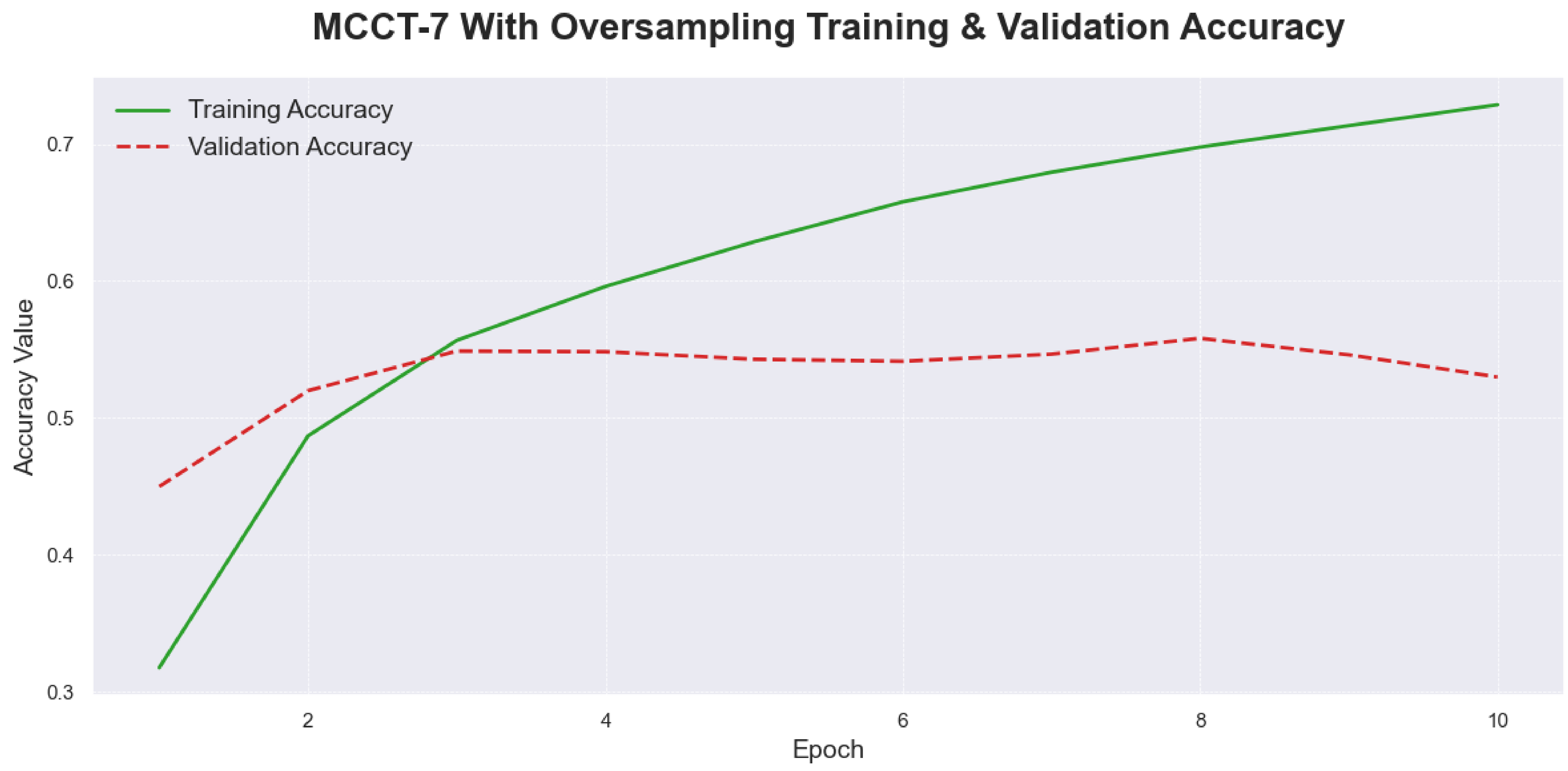
T.C. YEDİTEPE ÜNİVERSİTESİ

Cont'



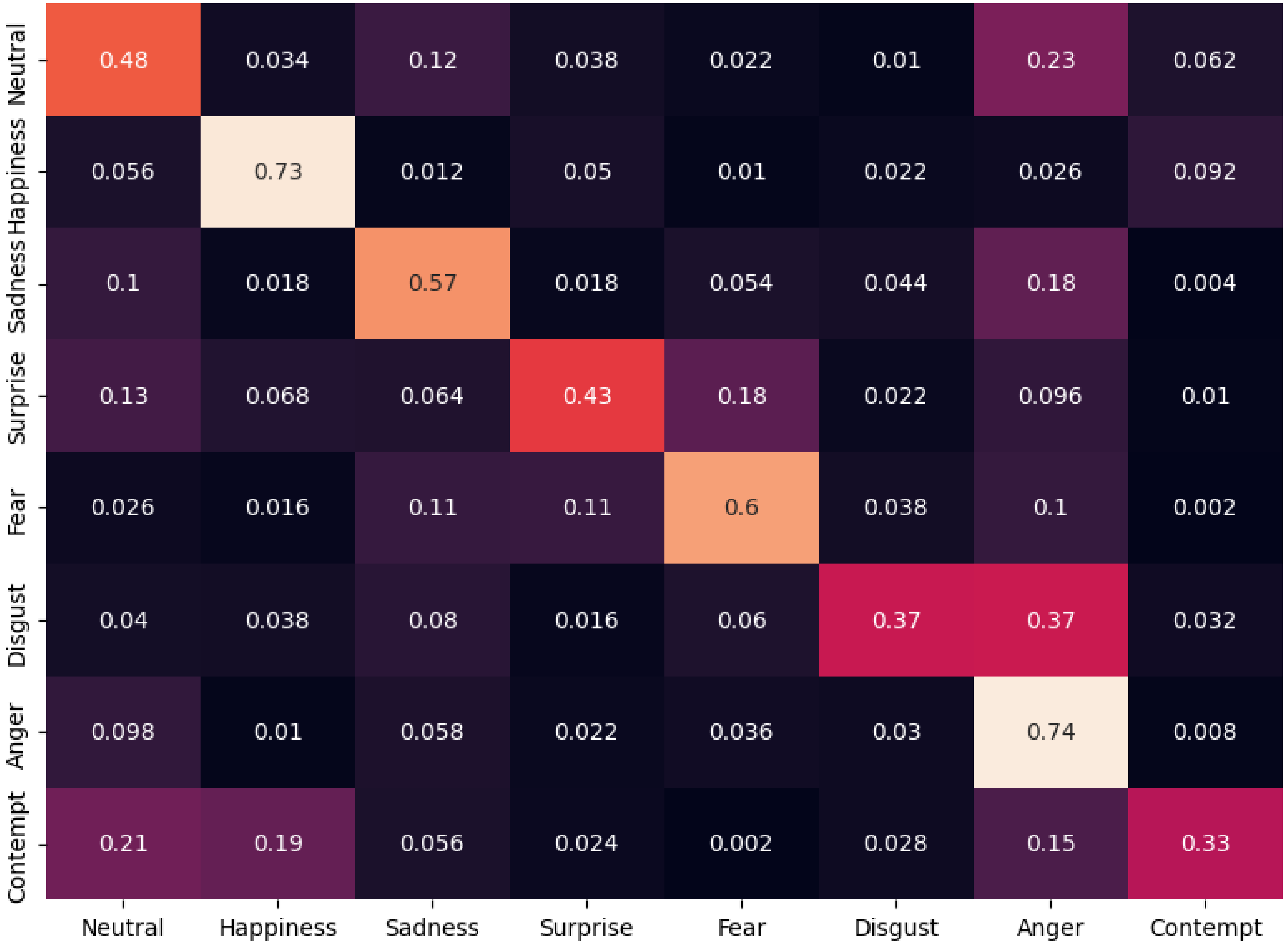




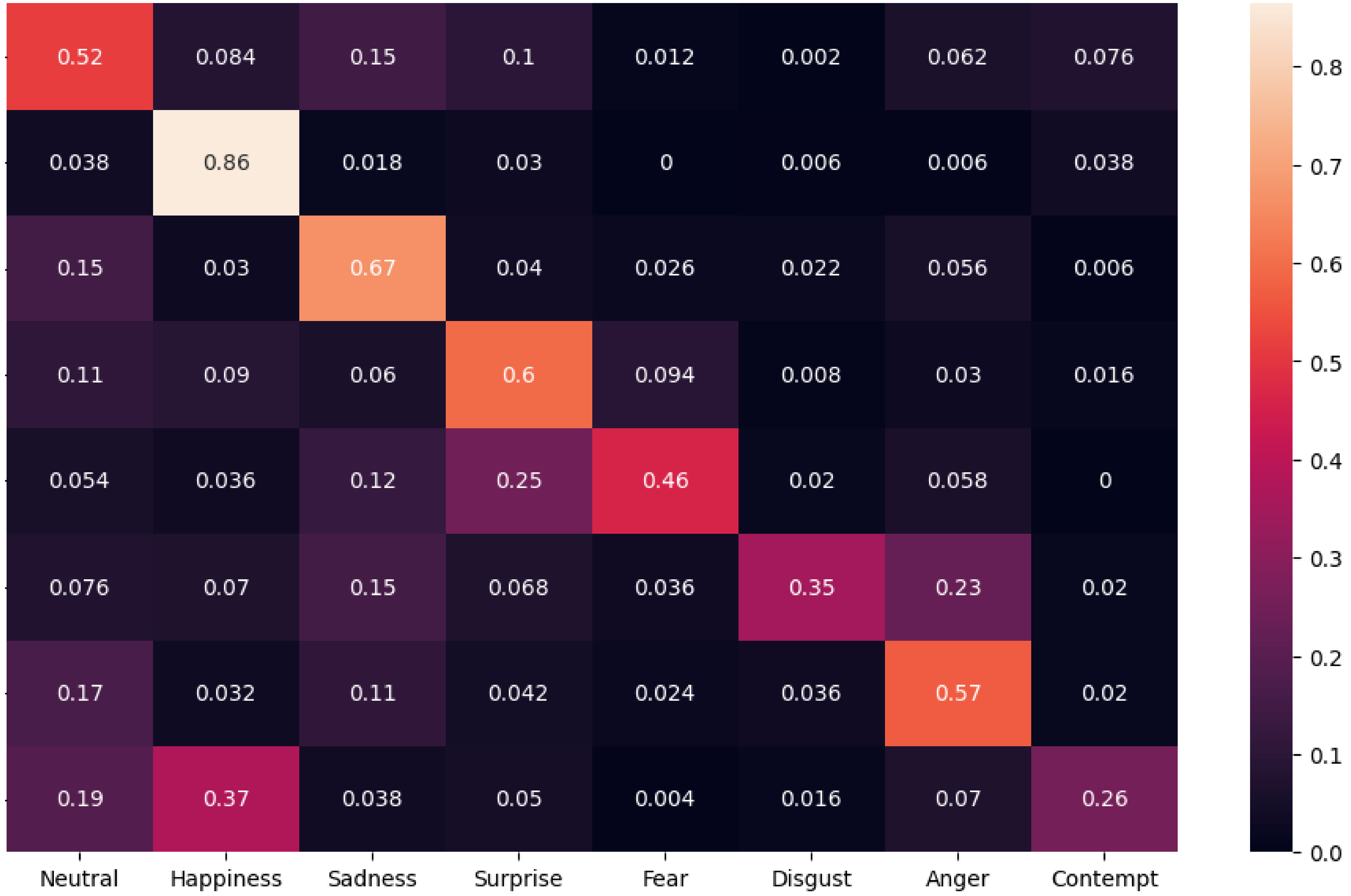


T.C. YEDİTEPE ÜNİVERSİTESİ

Cont'



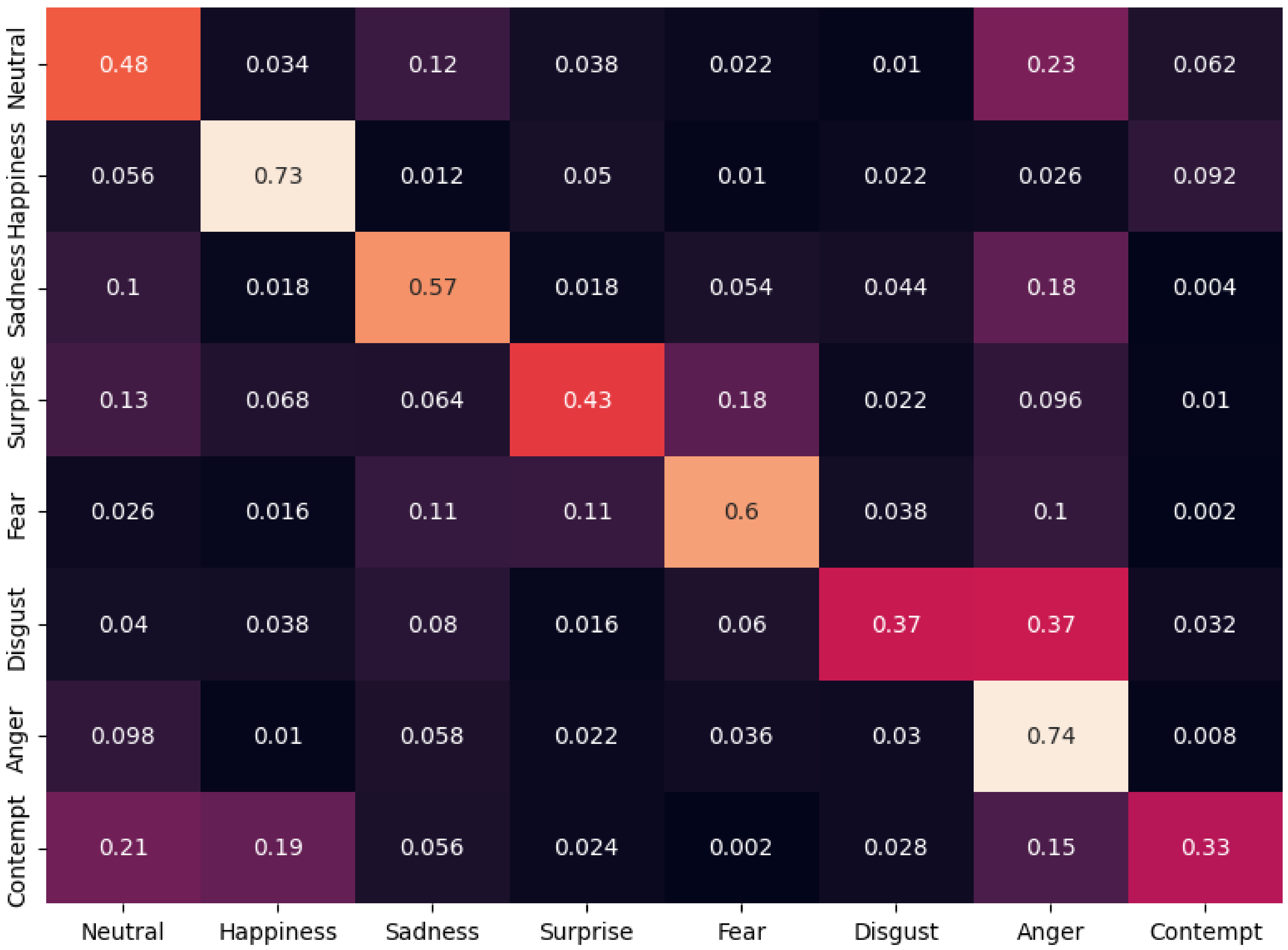
MCCT-1 (OS)



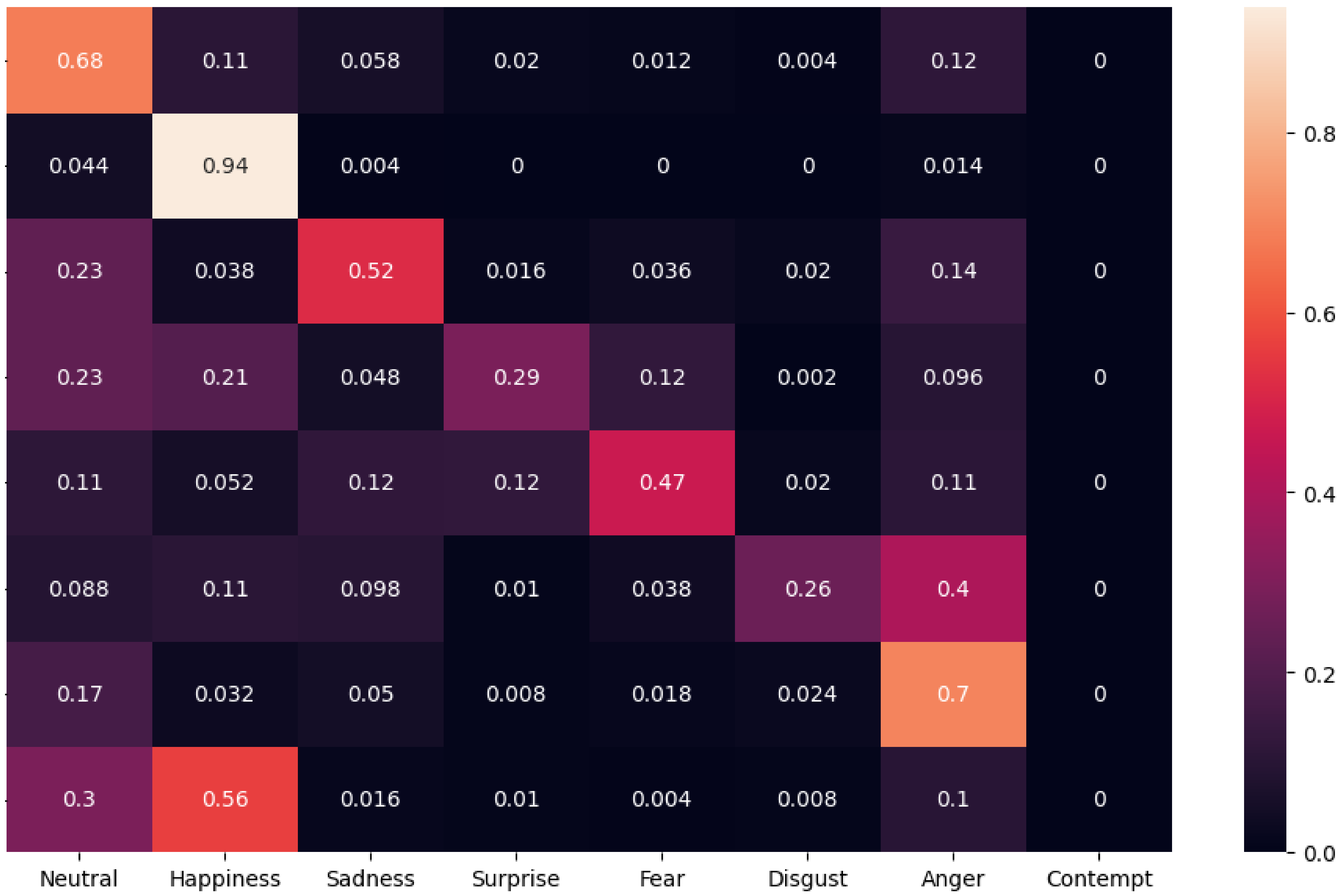
VCCT-1 (OS)

T.C. YEDİTEPE ÜNİVERSİTESİ

Cont'



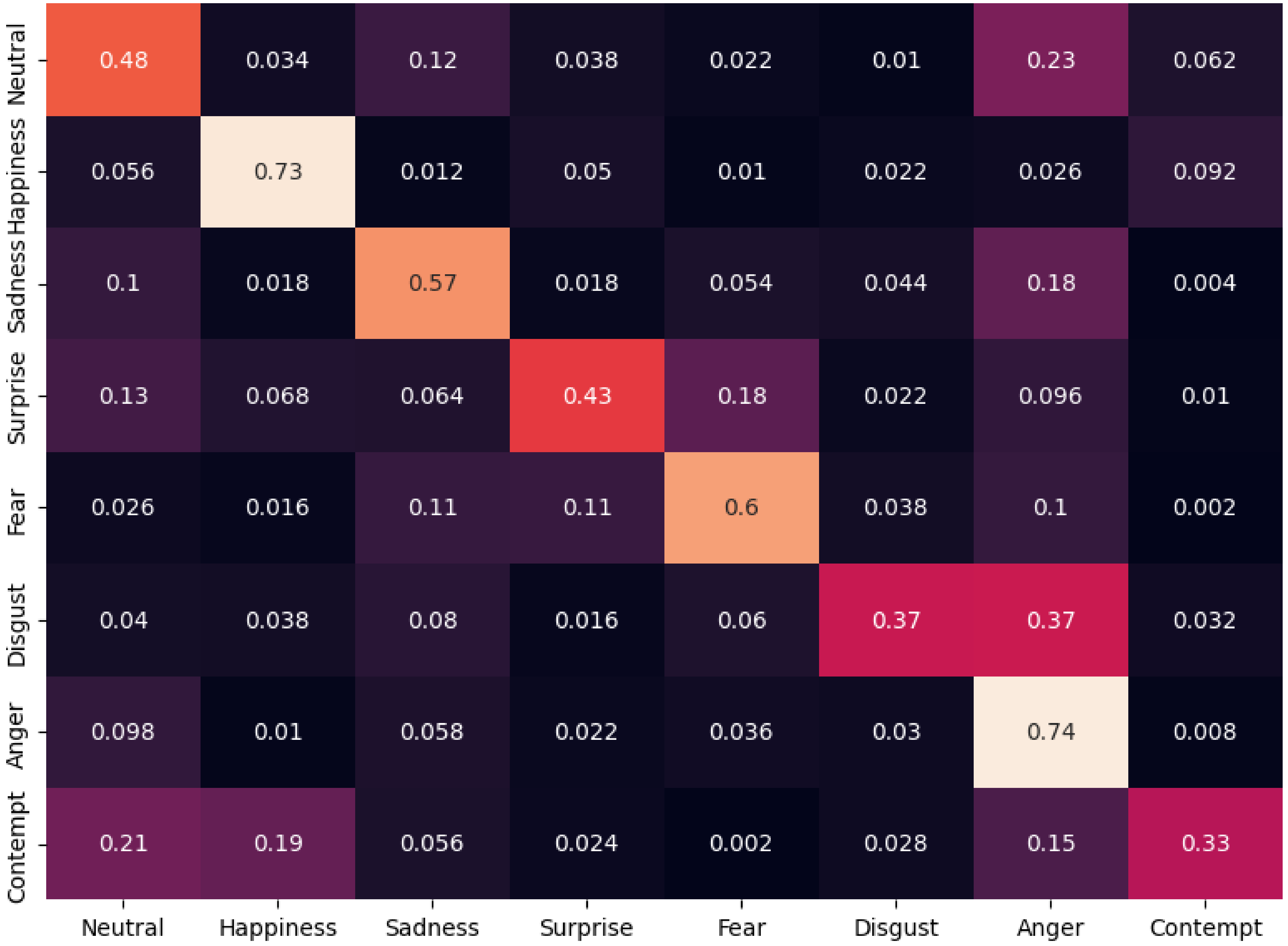
MCCT-1 (OS)



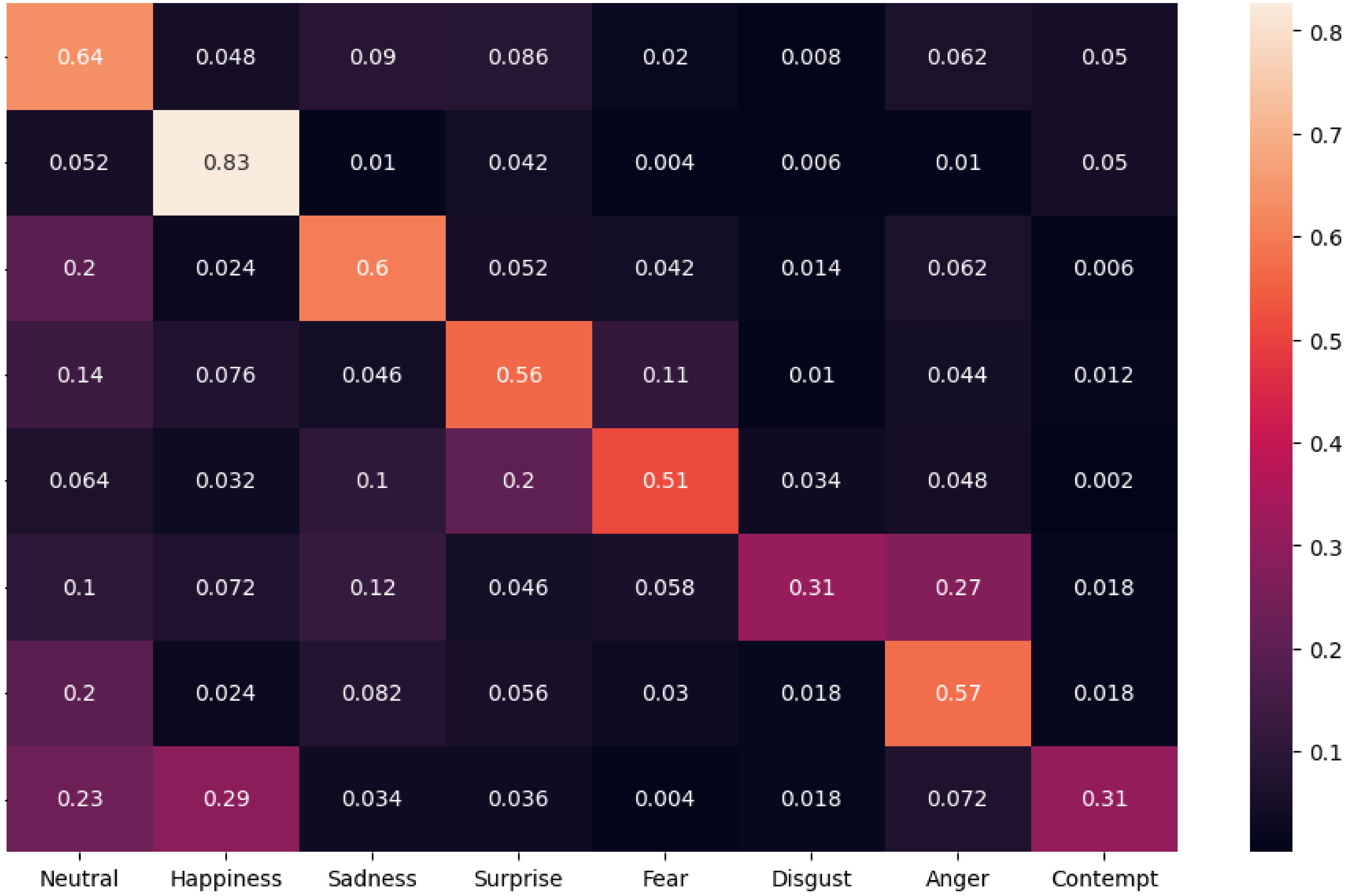
MCCT-1

T.C. YEDİTEPE ÜNİVERSİTESİ

Cont'



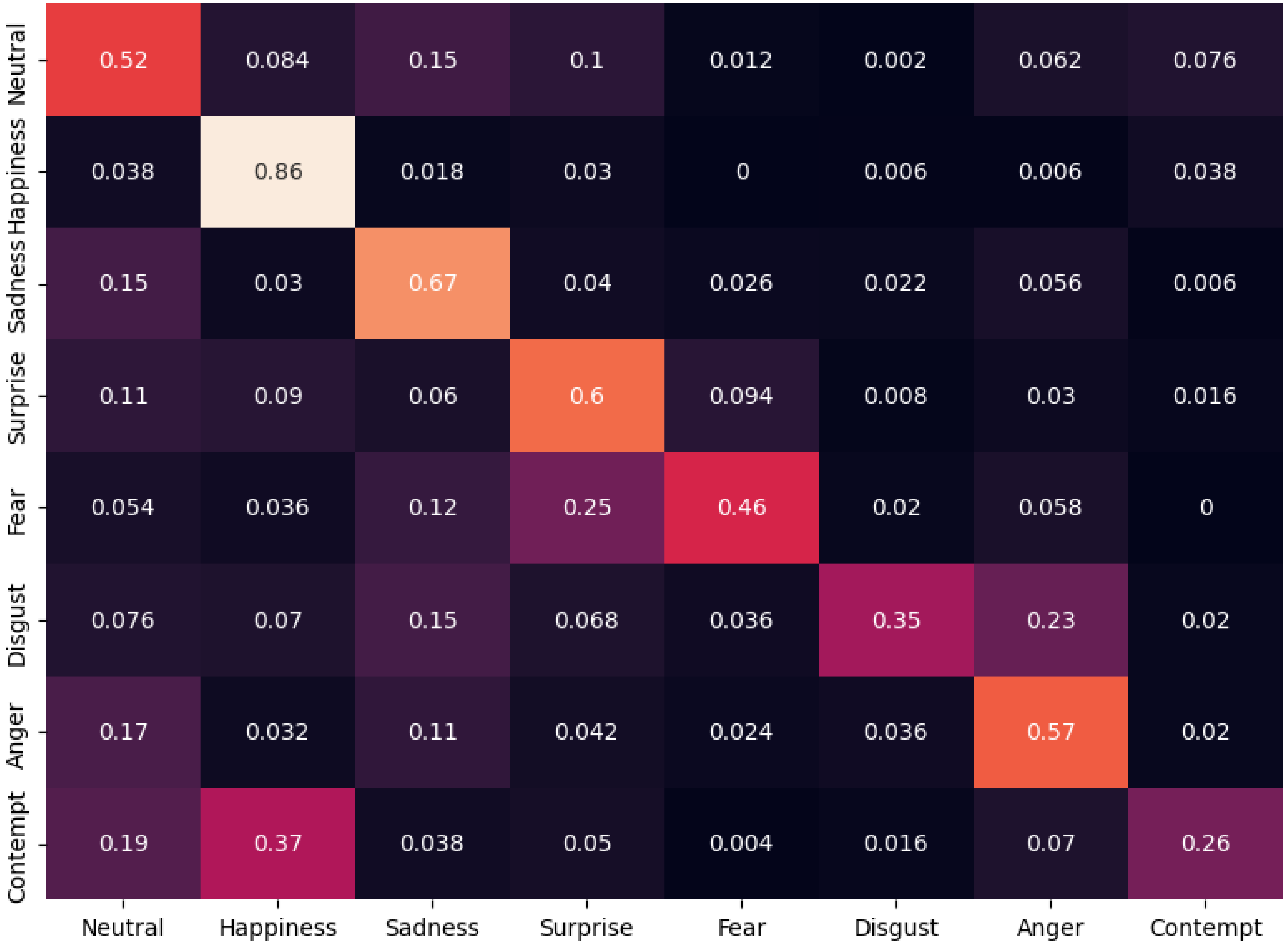
MCCT-1 (OS)



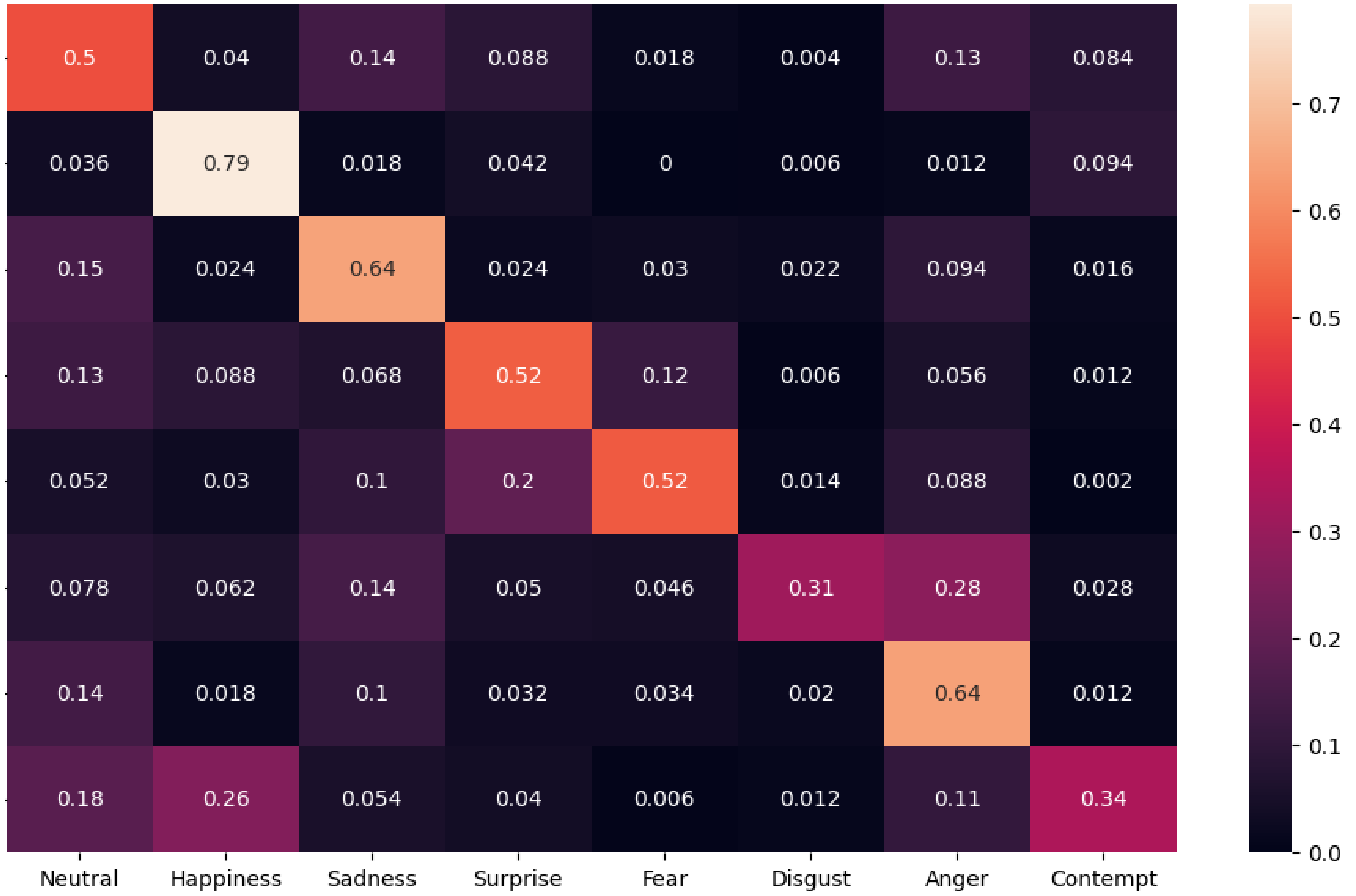
MCCT-2 (OS)

T.C. YEDİTEPE ÜNİVERSİTESİ

Cont'



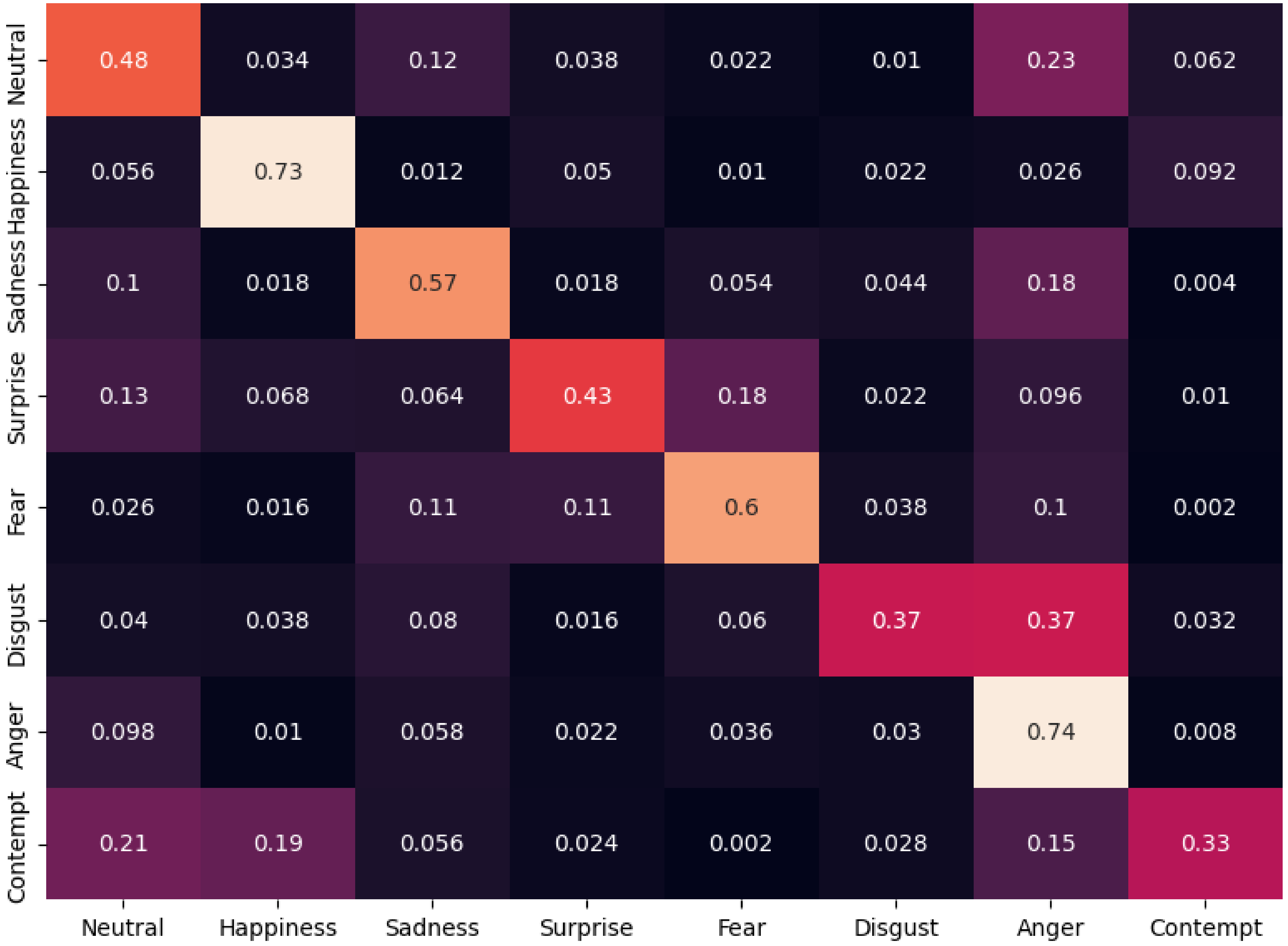
VCCT-1 (OS)



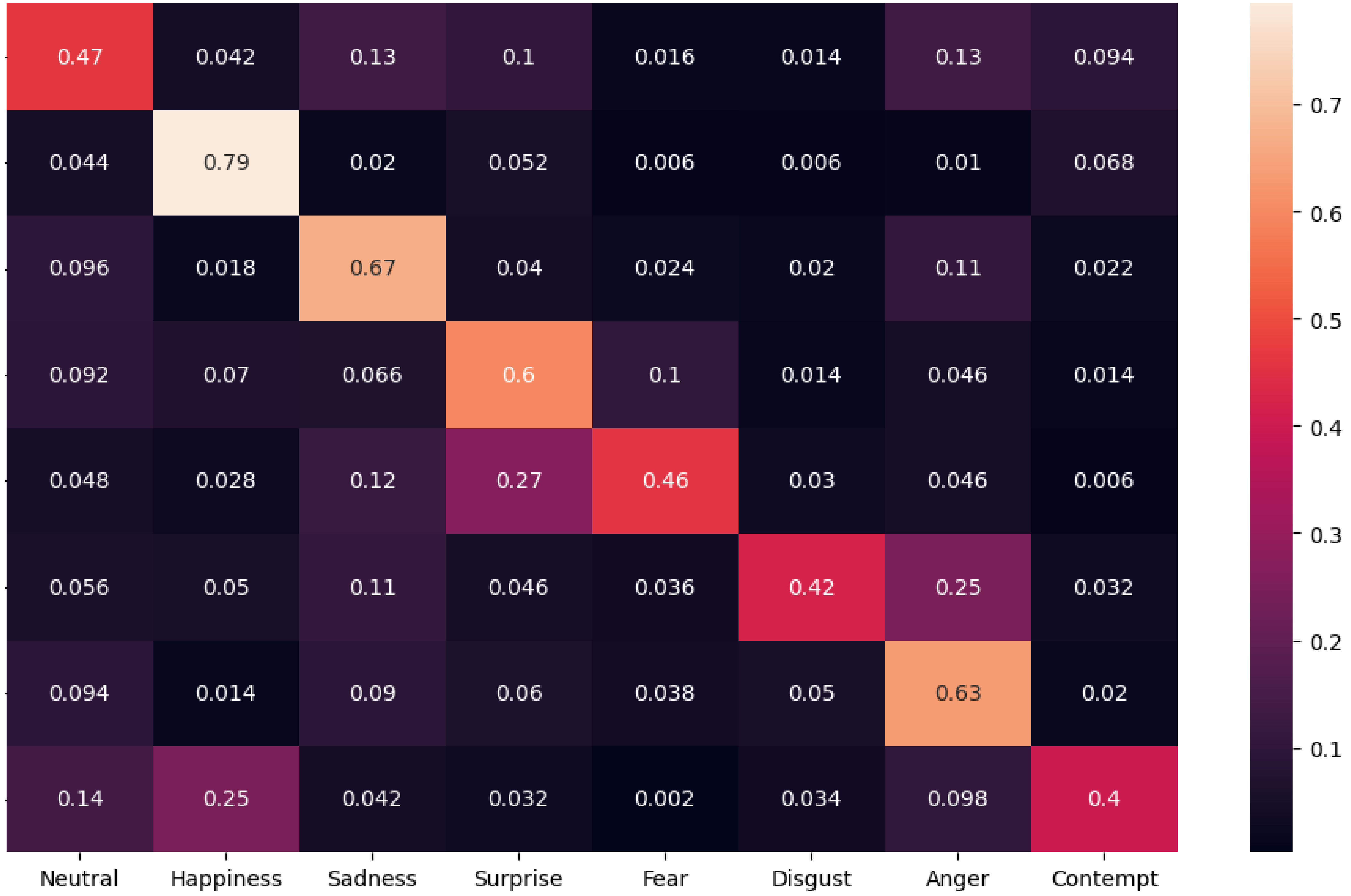
VCCT-2 (OS)

T.C. YEDİTEPE ÜNİVERSİTESİ

Cont'



MCCT-1 (OS)



MCCT-7 (OS)

T.C. YEDİTEPE ÜNİVERSİTESİ

Cont'

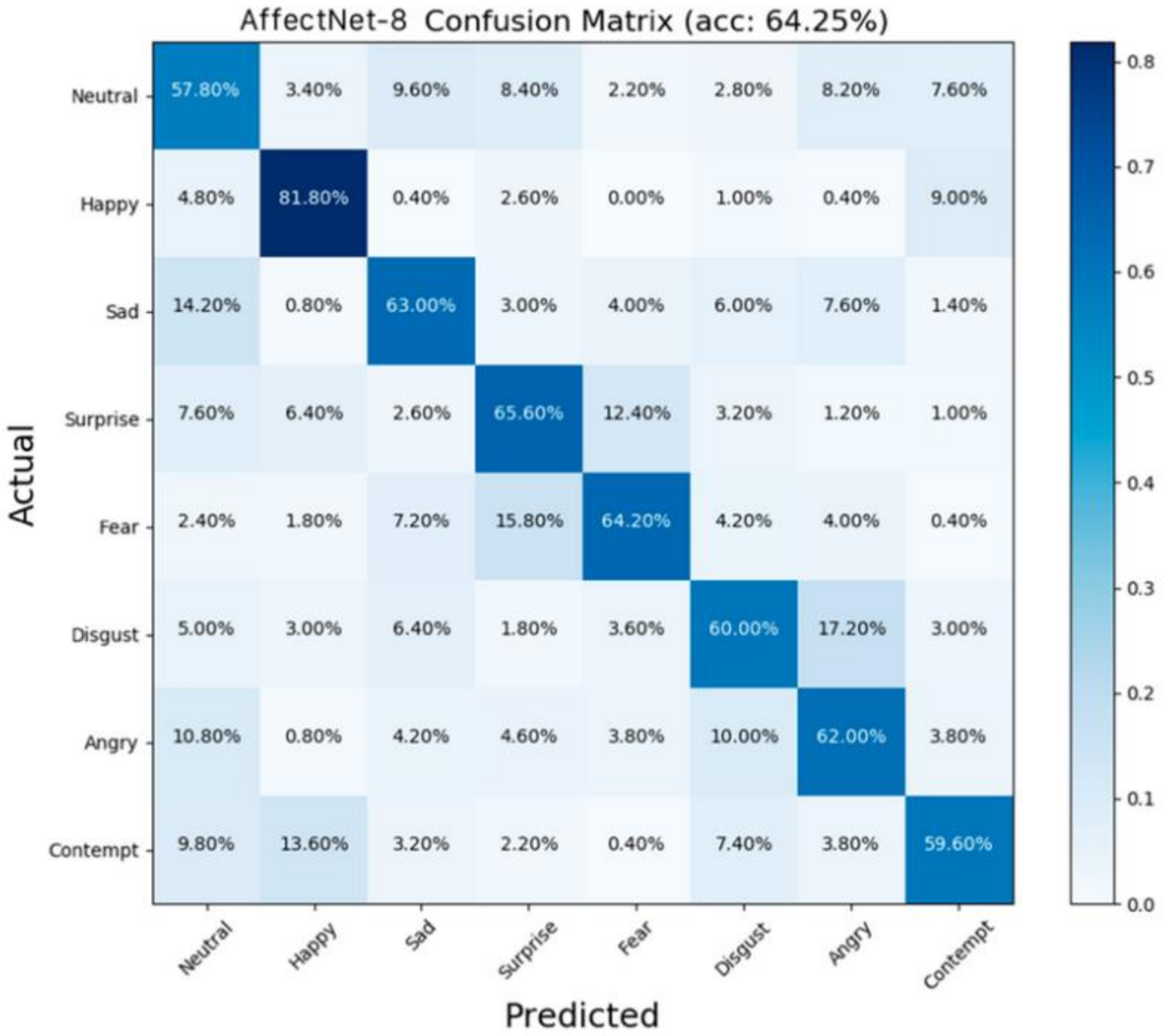


Image Source: S. Zhang, Y. Zhang, Y. Zhang, Y. Wang and Z. Song, "A Dual-Direction Attention Mixed Feature Network for Facial Expression Recognition," *Electronics*, vol. 12, no. 17, p. 3595, 2023.

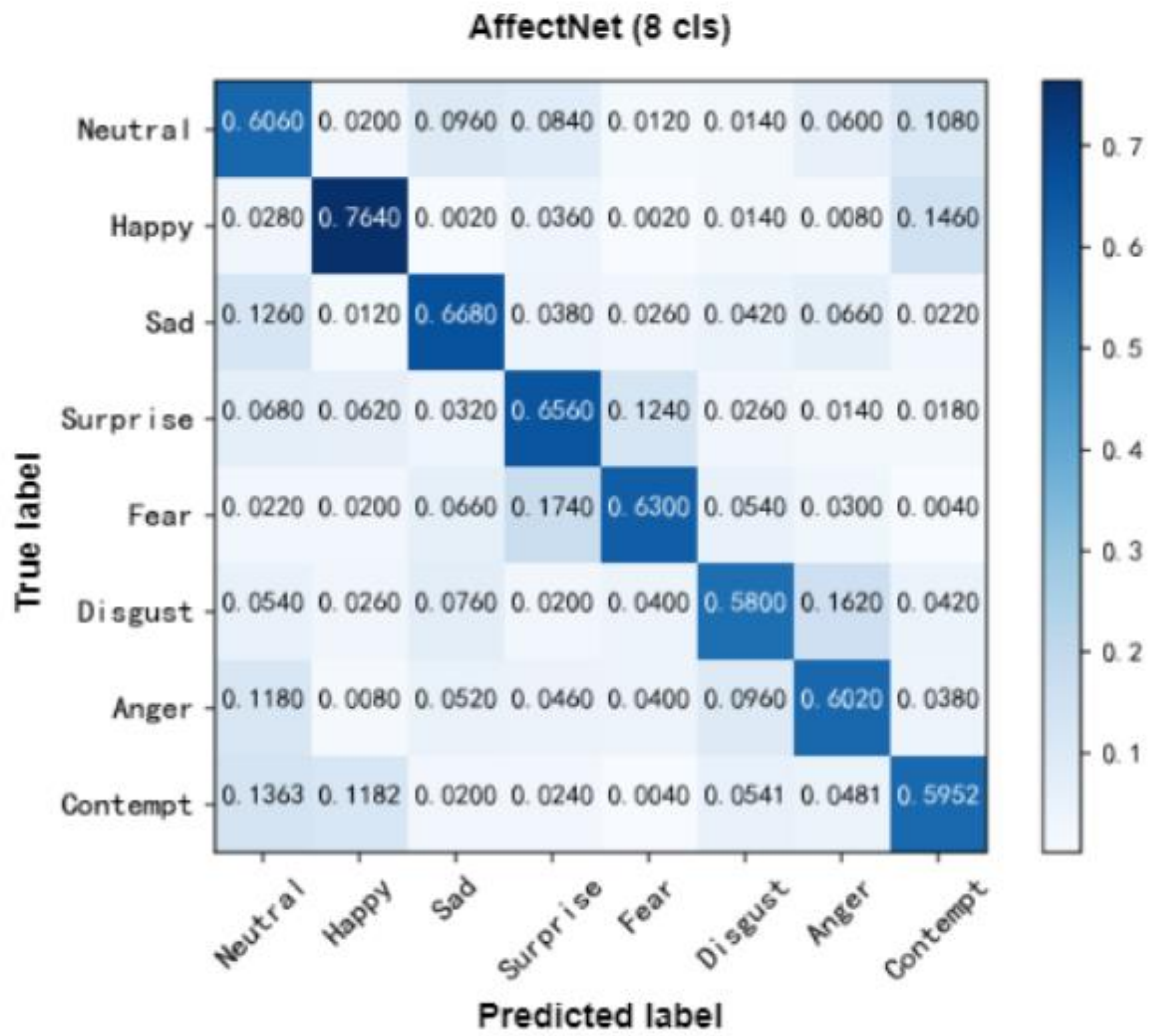
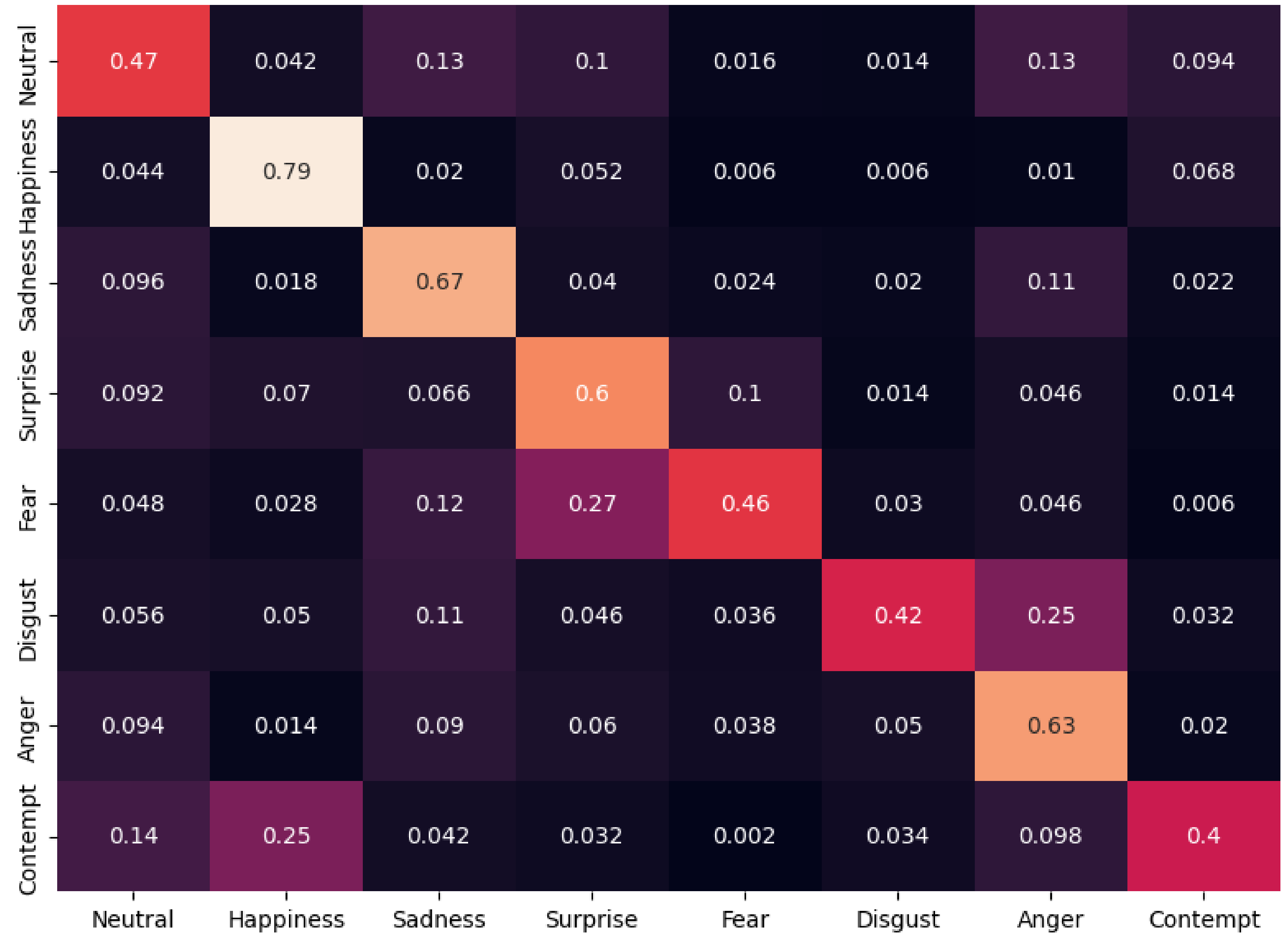


Image Source: J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie and A. Huang, "POSTER V2: A simpler and stronger facial expression recognition network," *arXiv preprint arXiv:2301.12149*, 2023.

MCCT-7 (OS)



T.C. YEDİTEPE ÜNİVERSİTESİ

Conclusion

- Best performing variant MCCT-7 achieved %55.54 on officially provided AffectNet validation set on 8 classes.
- Performing closer to or better than current best performing models in some categories
- Oversampling proved to enhance performance in all variants experimented with
- MCCT-1 showed better performance on 20 epochs compared to 10 epochs, even though it was overfitting.
- This suggests that MCCT-7 could also benefit from increased training epochs.

Future Work

- Explore use of different backbone networks to improve efficiency
- Explore use of different optimizers
- Conduct more rigorous search for hyperparameters of random augmentation or explore different data augmentation techniques
- Switch optimizer from AdamW to Stochastic Gradient Descent (SGD) to improve generalization ability

T.C. YEDİTEPE ÜNİVERSİTESİ

Thank You For Listening