Facial Emotion Recognition via Deep Learning Approaches using the FER-2013 Dataset

By: **Mohamed Adel**

Supervised by: Dr. **Amr S. Ghoneim**

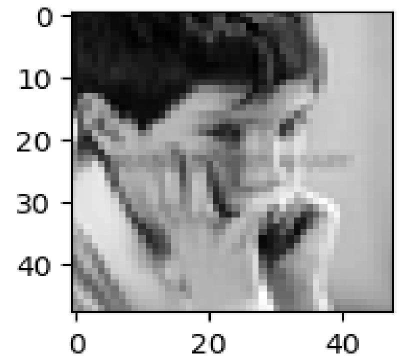
Problem Definition and Challenges:

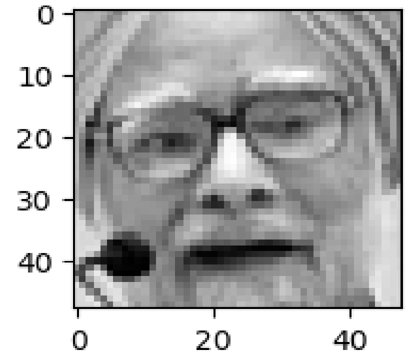
- Facial Emotion Recognition (FER) is a challenge by the ICML in 2013. The challenge was designed for competitors to design the best fitting model for recognizing face emotion from a picture [1].

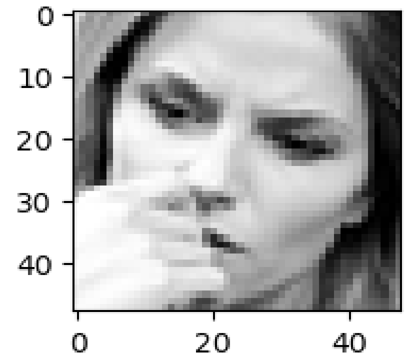
- This dataset was collected using Google image search API for images that included faces that matches a set of 184 emotion-related keyword.

- The images in the dataset are 48x48 pixels in size leading to difficulties in learning high-level features.

- Dataset includes different poses, different occlusions, have great intra-class similarity, and is characterized by imbalanced class problem.







Related Work (Selected Review of the Literature)

Title	Year	Objective of the study	Dataset type	Split, Pre-processing	Classifier Details	Accuracy
Facial Emotion recognition using Convolutional Neural Networks [5]	2019	Classify human faces into 7 face emotions	CSV	80-20 split	10-layers CNN	60.5%
Deep learning approaches for facial emotion recognition: A case study on FER-2013 [2]	2018	Examining performance of Deep learning approaches of facial expression recognition	JPG	75-25 split	AlexNet architecture	65%
Deep learning approaches for facial emotion recognition: A case study on FER-2013 [2]	2018	Examining performance of Deep learning approaches of facial expression recognition	JPG	Not mentioned	GoogLeNet architecture.	65.2%
Going Deeper in Facial Expression Recognition using Deep Neural Networks [6]	2015	Proposes a deep neural network architecture to address the FER-2013 dataset problem trying to achieve stat-of-the-art performance.	JPG	Not mentioned	Two convolutions each followed by max pooling, then four inception layers.	66.4%
Real-Time facial emotion recognition using deep learning [3]	2021	Analyse different emotions represented by the human face in real time.	CSV	Not mentioned	Xception model, implement pointwise conv followed by depthwise conv.	68.57%
Deep-Emotion: Facial Expression Recognition using attentional convolutional network [4]	2019	Propose a deep learning approach based on attentional CNN	JPG	Data Augmentation	CNN + attentional mechanism	70.02%
Facial expression recognition using convolutional neural networks: state of the art performance on FER2013 [8]	2016	Experiments aim to overcome the limitations of shallow and basic CNN architectures commonly used in FER.	JPG	Data Augmentation	Inception architecture	71.60%
Deep Learning using Linear Support vector Machines [9]	2015	Examines the effects of using SVM as an activation function for the last layer.	JPG	No	CNN using SVM as activation function.	71.2%
Facial Expression Recognition Using Convolutional Neural Networks: State of the Art [8]	2016	Experiments aim to overcome the limitations of shallow and basic CNN architectures commonly used in FER.	JPG	Data Augmentation	ResNet Architecture	72.4%
Facial Emotion Recognition: State of the Art Performance on FER2013 [7] **[8] VGG – 71.7%**	2021	Examining the performance of single-network research done on the FER-2013, comparing models created from scratch with predefined models.	JPG	Yes	VGG architecture	73.2%

What others did not consider?

~ Research is to see what everybody else has seen, and to think what nobody else has thought.

~ Albert Szent-Györgyi de Nagyrápolt

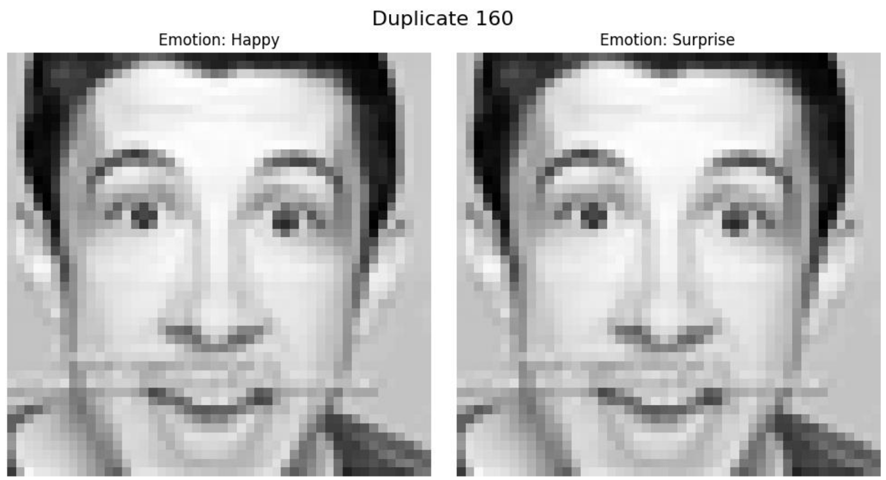
Methodology of the **Study .. *What other did not consider?***

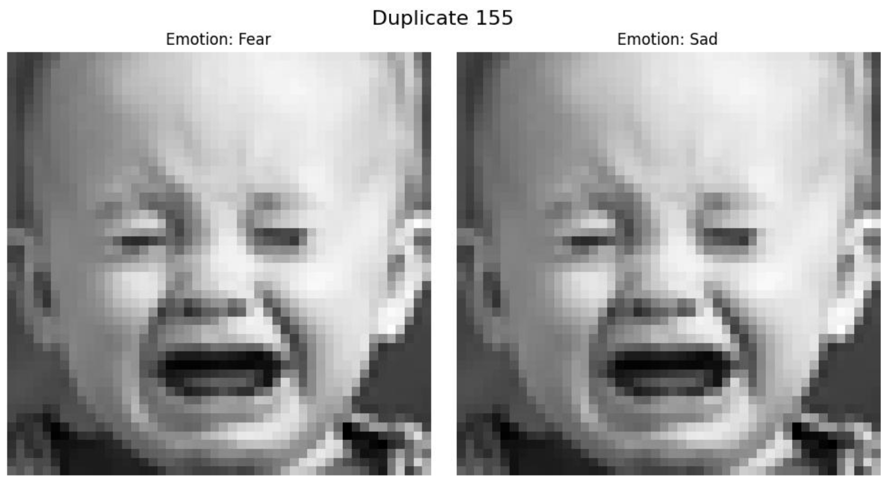
Data-Centric Approach .. *Our focus is on the data quality rather than the model's complexity.*

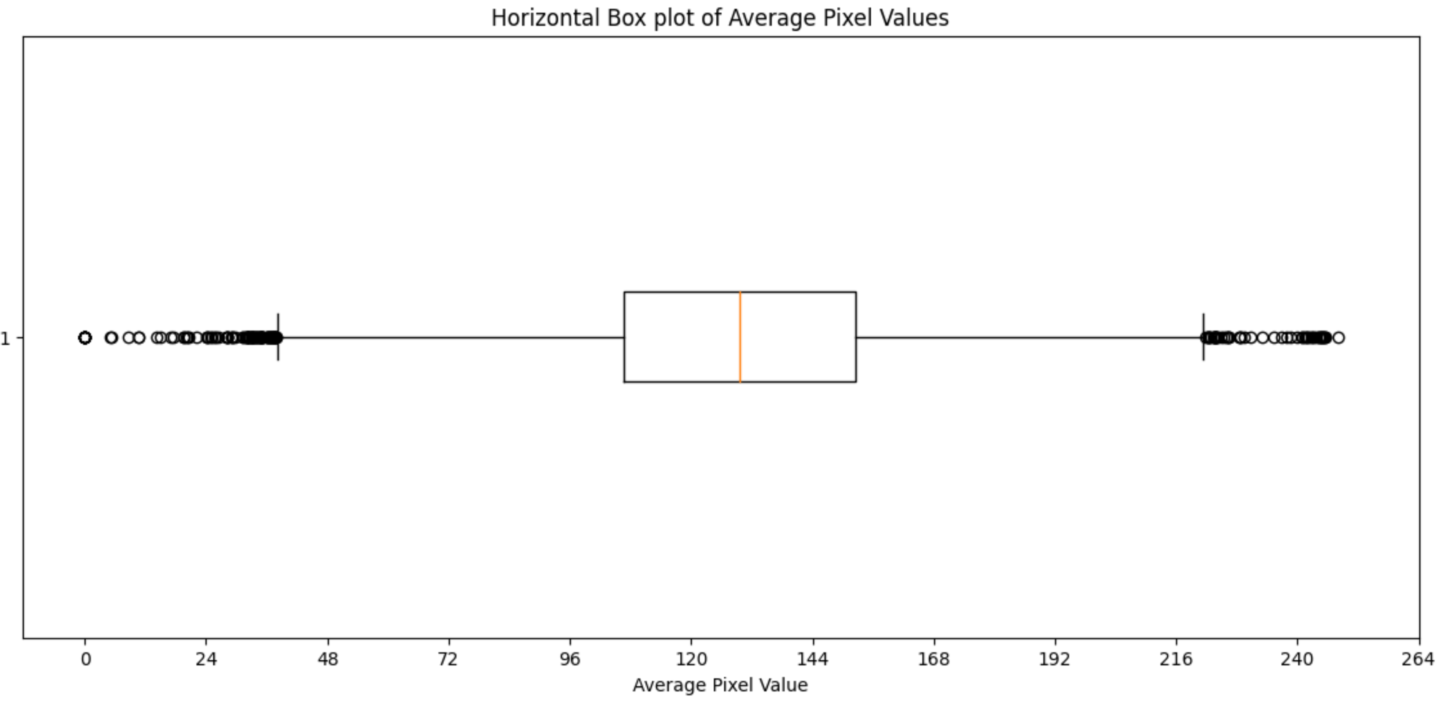
1. Identify any problems residing within the FER-2013 dataset using EDA.
2. Identify misleading images that affects models training.
3. Identify best sampling approach for the dataset.
   (Class_weight, ROS, SMOTE).
4. Finally, prepare a suitable model for the cleaned data
5. Fine-tune the model for increased performance and generalization.

Problem 1:
Duplicates with Non-Matching Emotions

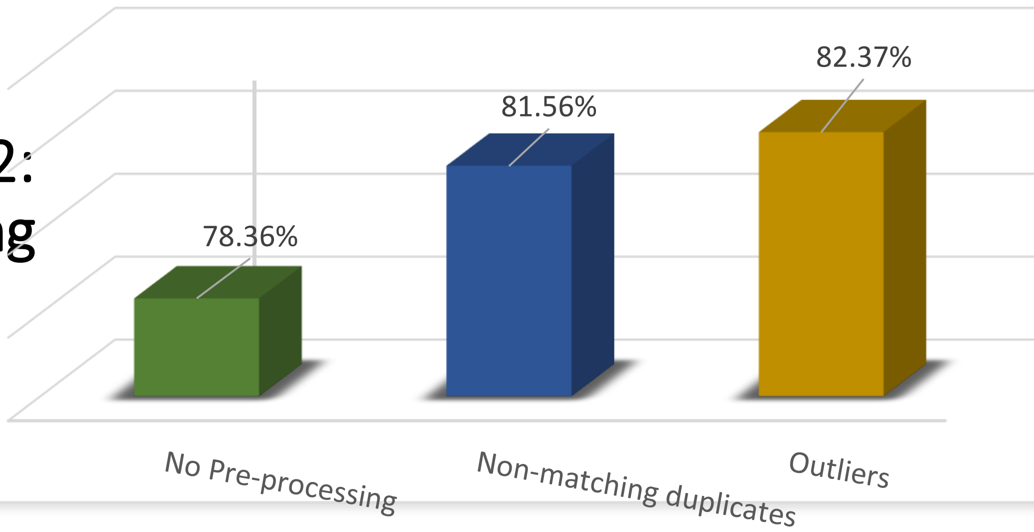
- Dataset:
  - 28709, and 7178 for training and testing respectively. (i.e., $(\frac{7178}{28709}) * 100 = 25\%$ test-split)
  - Images dimensions are (48, 48, 1).
  - Dataset available in CSV, JPG.
- Duplicates:
  - Dataset contained 1853 duplicates from which 282 are duplicates with non_matching emotions.

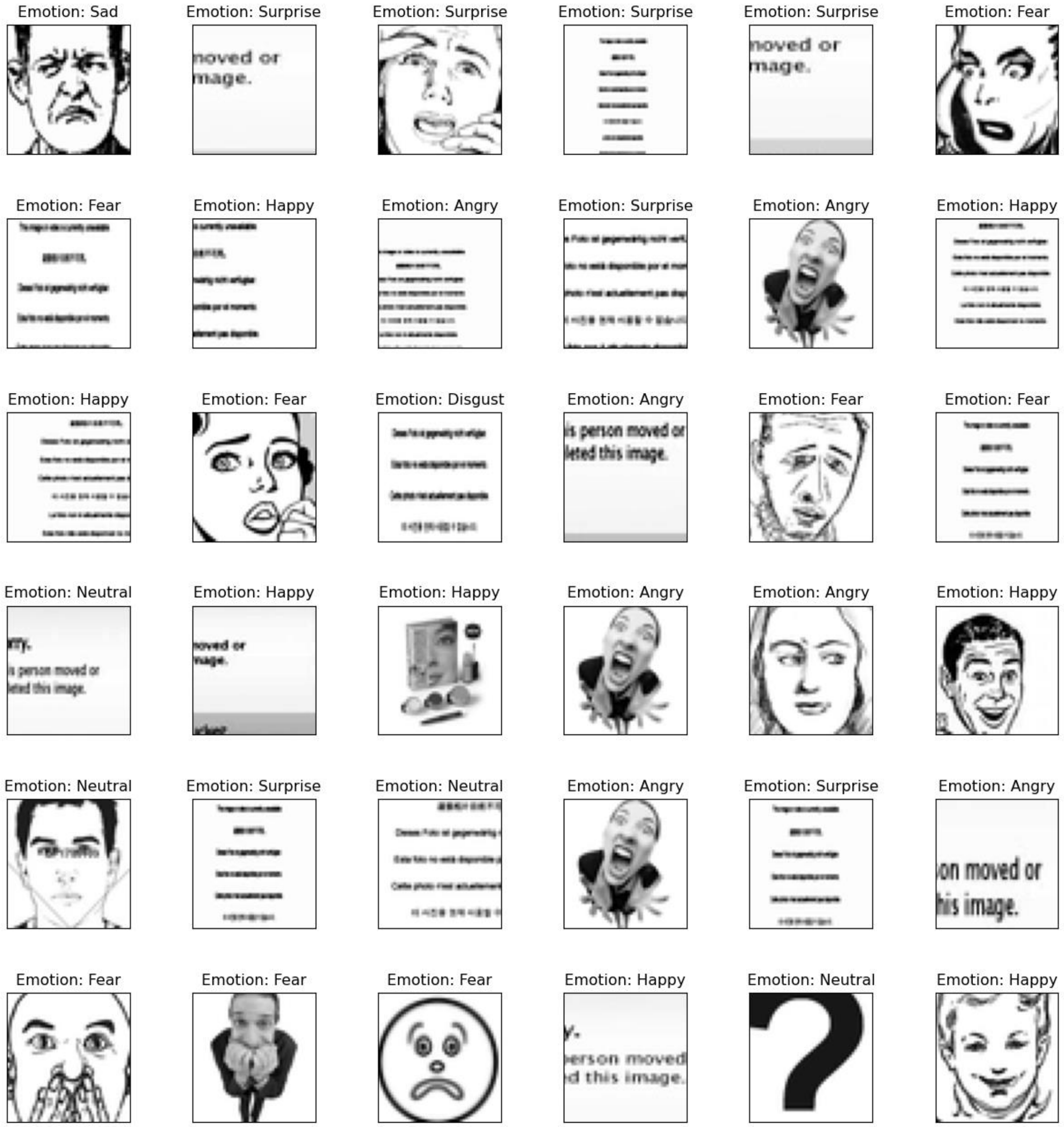




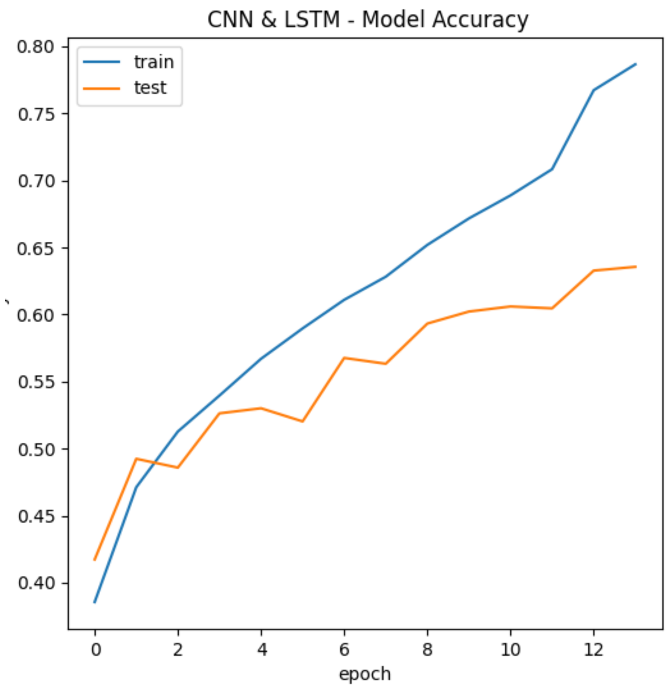


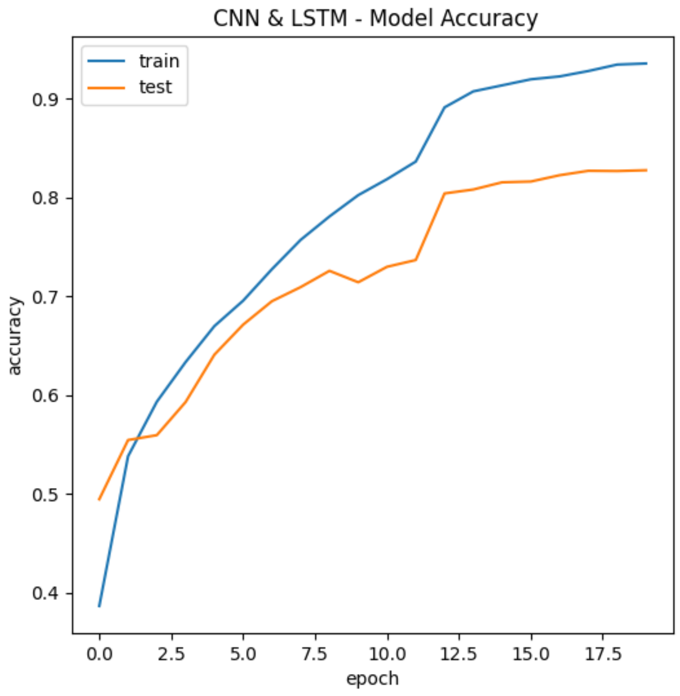
Problem 2: Misleading Outliers

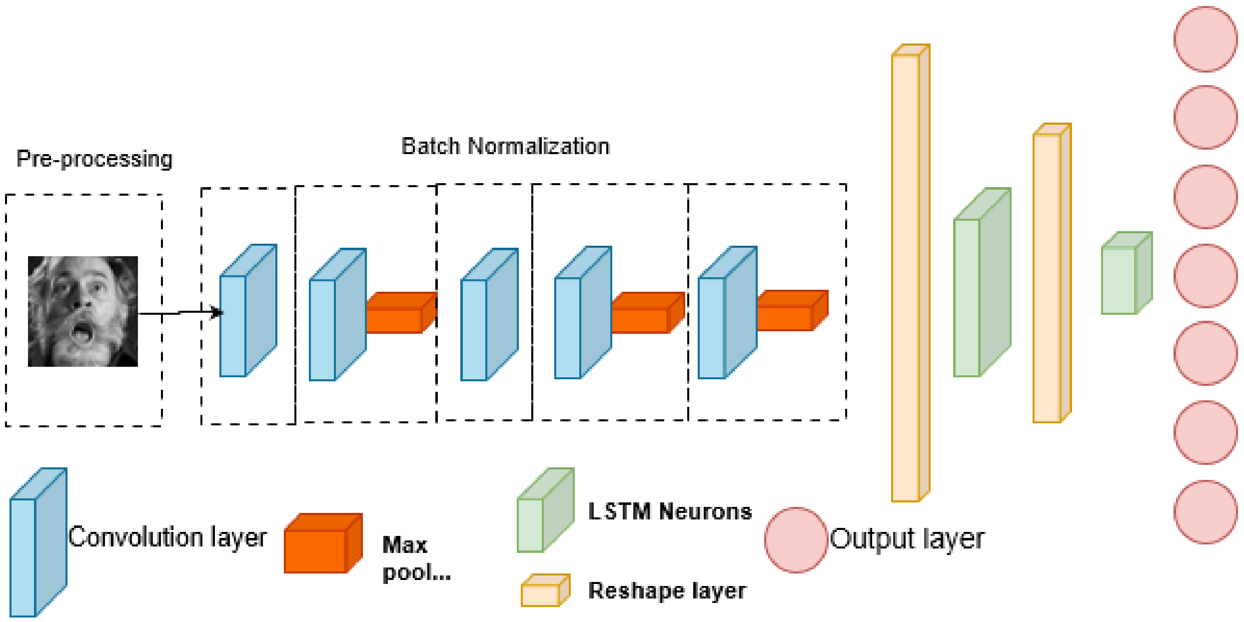




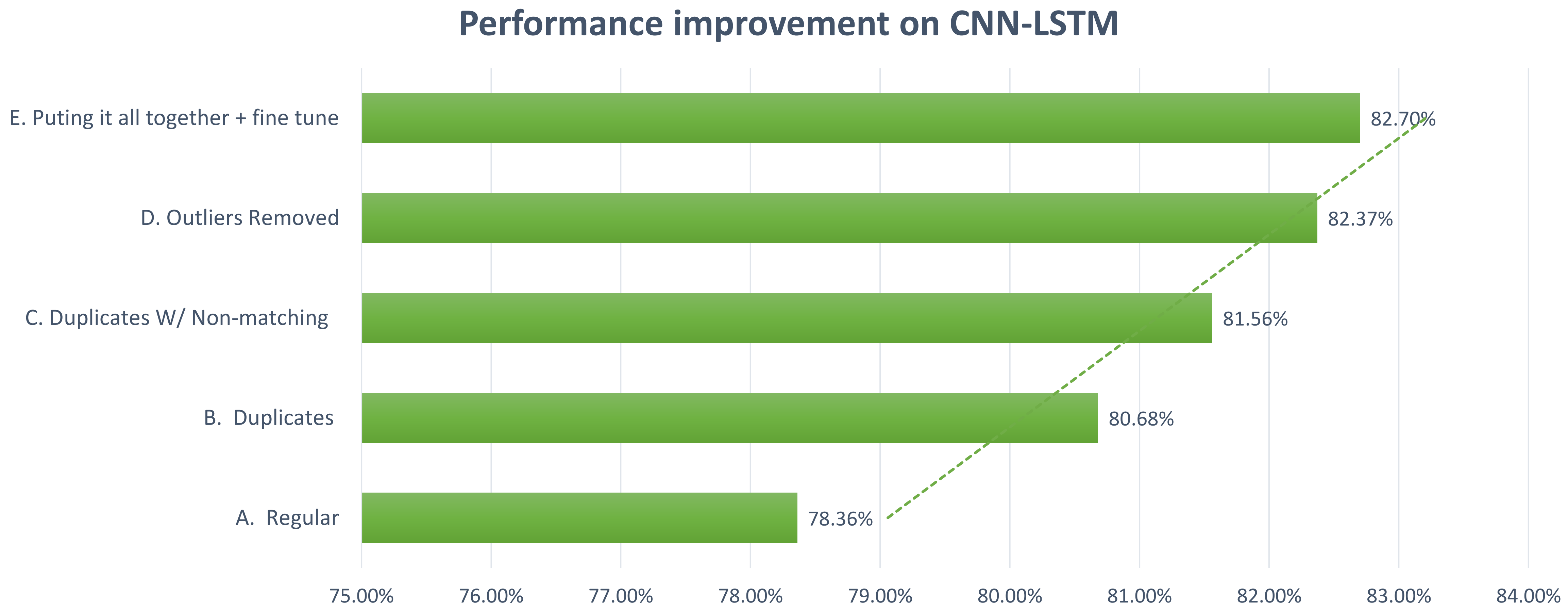
Intuition: CNN-LSTM





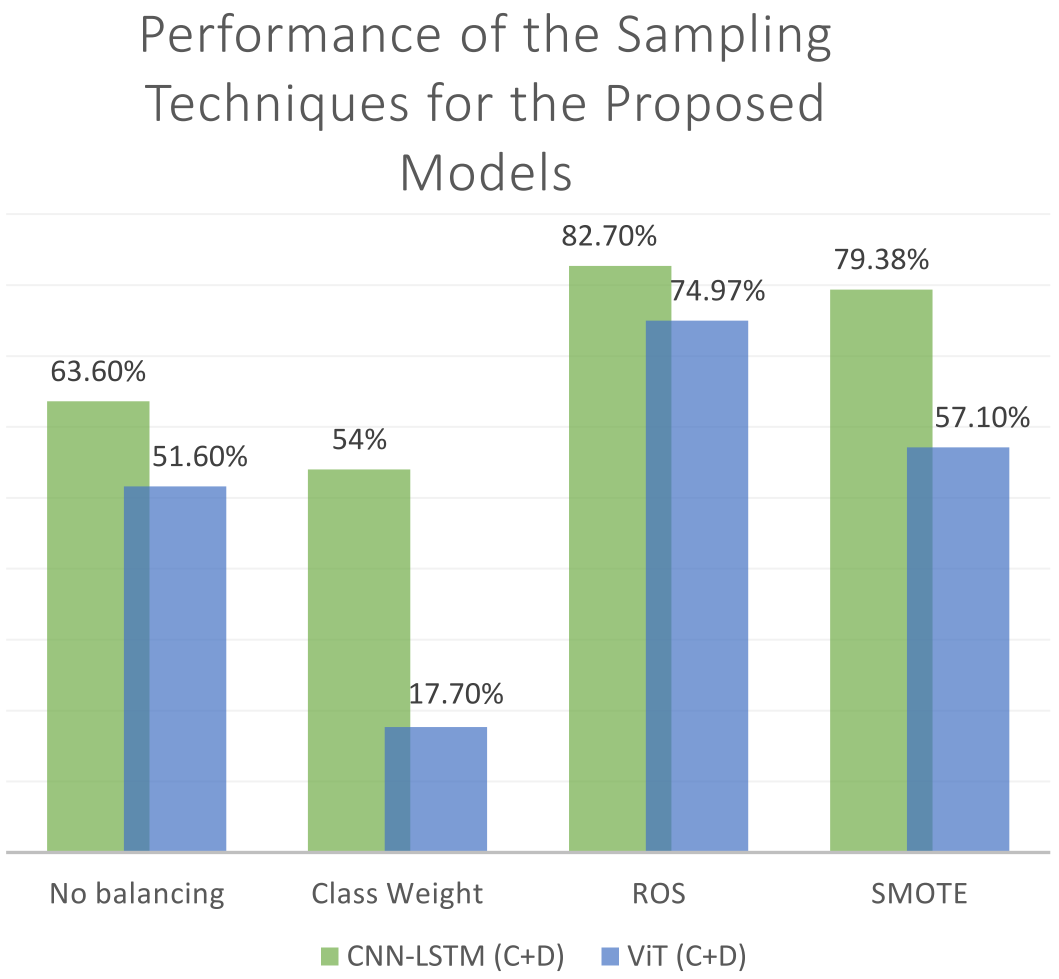


CNN-LSTM Variation Comparison



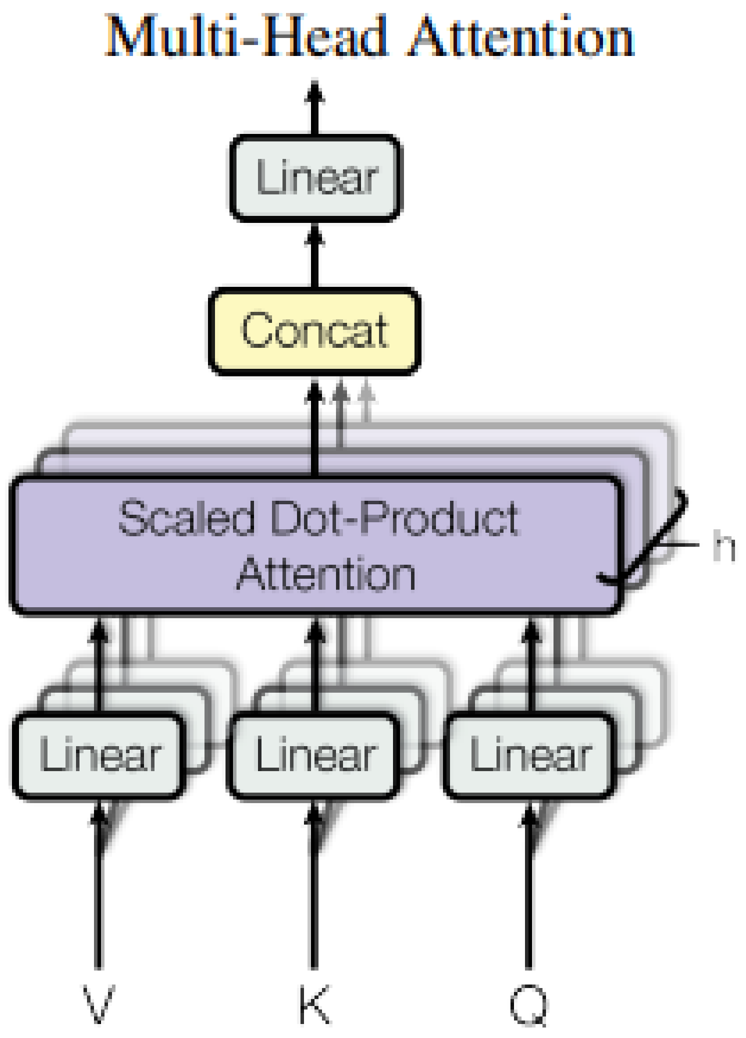
Investigating the Sampling Techniques

- **Class Weighting**: Method of balancing distribution of each class during training.
  - This is done by assigning different weights to each class, so that the model pays more attention to the minority classes.
- **Random Oversampling**: ROS is a technique for dealing with imbalanced datasets by duplicating minority class samples.
- **Synthetic Minority Oversampling Technique**: SMOTE is a technique for dealing with imbalanced datasets by creating new minority class samples.
  - This is done by finding the k-nearest neighbors of each minority class sample and then creating a new sample that is a linear combination of the k neighbors.



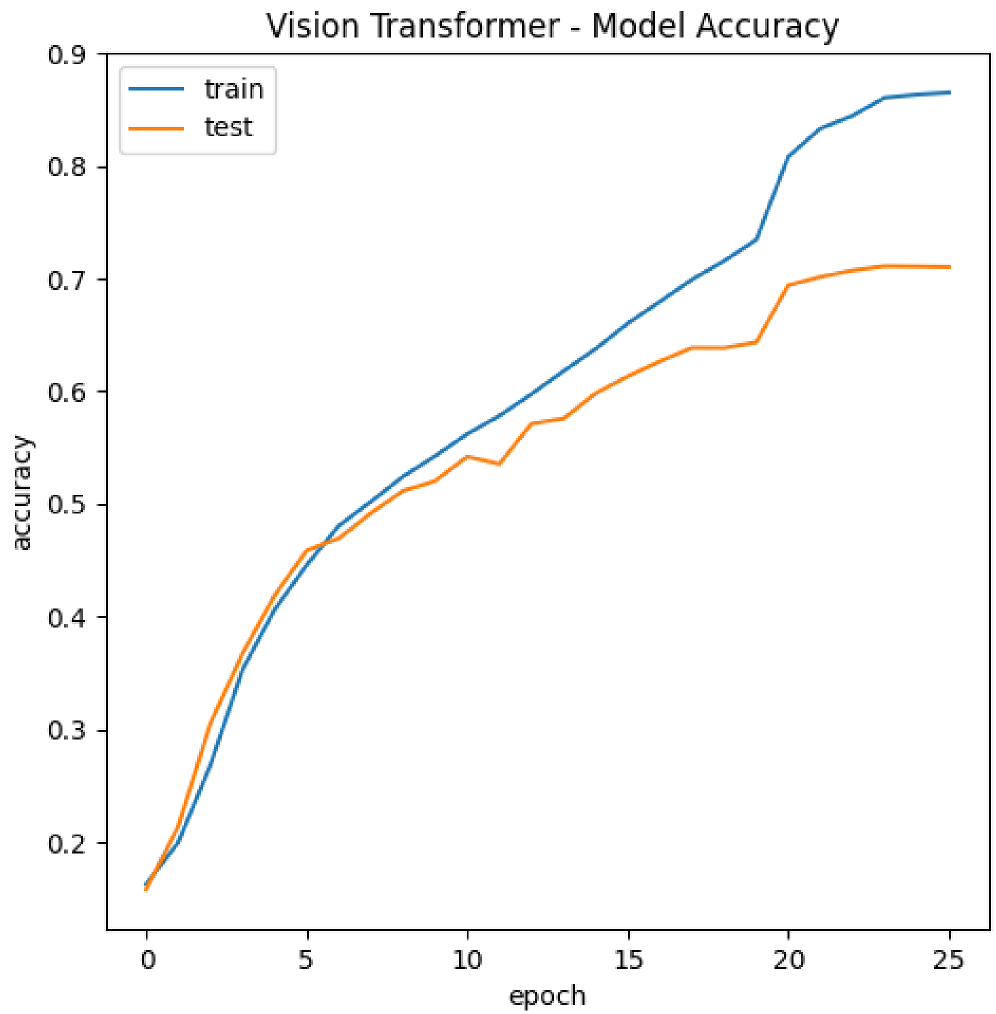
Vision Transformer
Multi-Headed Attention

- Which is a self-attention mechanism based on an encoder and a decoder architecture. The encoder aims to map an input sequence into a continuous representation that holds learned information about that input sequence. The decoder's goal is to take the continuous representation of the input and generate a single step-by-step output while feeding the decoder the previous outputs.
- An Encoder module consists of two sub-modules: MultiHeadAttention, a fully connected network, and a layer normalization. This multi-head attention mechanism is a self-attention mechanism. Self-attention allows the module to associate each image in the input with other images.



CNN + Self–Attention Mechanism

- The parameter num_heads in the function MultiHeadAttention() provided by Keras determines the number of parallel attention heads in the multi-head attention layer. Each head computes an independent attention mechanism, and the results are concatenated and linearly combined at the end. This allows the model to focus on different aspects of the input simultaneously, which can improve the quality of the linearly combined results.
- The parameter key_dim indicates the dimensionality of the keys; it determines the size of the dot product space, which impacts the expressiveness of the attention mechanism. A higher key dimension allows the model to learn more complex relationships between the queries and keys.
- In practice, these hyperparameters are often tuned through experimentation to find the best combination that maximizes performance on a validation set. Hence, choosing the hyper-parameters for the first model can be random. The initialized parameters for the num_head and key_dim was set to 8 for both parameters to test the initial performance of the model.



Conclusion

- In this work, we have not only achieved the highest attained accuracy of 82.7 % using a single network with no additional data but also, we have managed to introduce new issues that reside within the FER dataset that past authors still need to consider.

Future Work

- Fine-tune and improve the Vision Transformer model for better performance on the FER dataset.
- Explore the possibility of creating a cleaned FER dataset: Merge cleaned FER with auxiliary data to represent real-world scenarios of facial emotions & publish the cleaned dataset available for public use.
- Fine-tune ViT with ROS variation, to test the model's capacity for improvement.
- Investigate ensemble models for future studies, Choosing models that oppose each other in terms of emotion classifications (i.e., Binary-Tree classification on multi-class problem).

Selected References :

[1] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, “Challenges in representation learning: A report on three machine learning contests,” *arXiv.org*, 01-Jul-2013. [Online]. Available: https://arxiv.org/abs/1307.0414. [Accessed: 04-Apr-2023].

[2] I. Hatzilygeroudis and V. Palade, “Deep Learning Approaches for Facial Emotion Recognition: A Case Study on FER-2013,” in *Advances in hybridization of intelligent methods models, systems and applications*, Cham: Springer International Publishing, 2018, pp. 10–25.

[3] S. Chand, A. Singh, R. Bhatia, I. Kaur, and K. R. Seeja, “Real-time facial emotion recognition using Deep Learning,” *Algorithms for Intelligent Systems*, pp. 219–226, 2021.

[4] S. Minaee, M. Minaei, and A. Abdolrashidi, “Deep-emotion: Facial expression recognition using an attentional convolutional network," *Sensors*, vol. 21, no. 9, p. 3046, 2021.

[5] A. Saravanan, G. Perichetla, and D. K. S. Gayathri, “Facial emotion recognition using convolutional neural networks,” *arXiv.org*, 12-Oct-2019. [Online]. Available: https://arxiv.org/abs/1910.05602. [Accessed: 25-Nov-2022].

[6] A. Mollahosseini, D. Chan, and M. H. Mahoor, “Going deeper in facial expression recognition using deep neural networks,” *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[7] Y. Khaireddin and Z. Chen, “Facial emotion recognition: State of the art performance on FER2013,” *arXiv.org*, 08-May-2021. [Online]. Available: https://arxiv.org/abs/2105.03588. [Accessed: 04-Apr-2023].

[8] M. K. Pramerdorfer and Christopher, "Facial expression recognition using convolutional neural networks: state of the art," arXiv, vol. 1612.02903, 2016.

[9] Y. Tang, "Deep learning using linear support vector machines," arXiv, vol. 1306.0239, pp. 1-4, 2013.