Ocular Disease Recognition: Feature Fusion and Multi-Modality Analysis

Mary Nwangwu MCN8851@NYU.EDU
Center for Data Science, New York University
New York, NY, USA

Abstract

Ocular diseases, such as diabetic retinopathy, glaucoma, and age-related macular degeneration, are significant contributors to global blindness. As rates continue to escalate, the imperative for early detection and treatment becomes increasingly critical. While deep learning-based automated screening shows promise, addressing patients with multiple conditions remains a challenge. This study investigates multi-label ocular disease classification using the OIA-ODIR dataset, combining binocular fundus images with patient demographics. Various fusion techniques, including late and early fusion strategies, as well as methods like sum, product, and concatenation, are examined to optimize integration of heterogeneous data sources. ResNet-18 performance is evaluated, with late fusion methods, especially fundus image-only fusion using element-wise summation, demonstrating superior classification outcomes. These findings emphasize the need to tailor fusion techniques to dataset characteristics for effective automated ocular disease recognition.

Keywords: Ocular diseases, Multi-label classification, Feature fusion, Multimodal learning

1 Introduction

With over 2.2 billion people globally facing vision impairment, about half of which are preventable, early detection and treatment are crucial (WHO, 2023). Fundus diseases, including diabetic retinopathy (DR), age-related macular degeneration (AMD), glaucoma, and cataract, are leading causes of global blindness. Projections suggest a rise in blindness and low vision cases by 2030 (NEI, 2014). Manual screening by ophthalmologists, often time-consuming and reliant on expertise, emphasizes the necessity for large-scale automated screening to alleviate workload and prevent long-term vision damage.

In recent studies, significant progress has been made in the field of ocular disease classification and recognition through the application of deep learning techniques. In 2017, Choi et al. employed a multi-categorical deep learning neural network to classify retinal images, achieving enhanced performance when focusing on a subset of three integrated categories. In 2023, Abbas et al. introduced Deep-Ocular, a transfer learning architecture integrating self-attention and dense layers, achieving high accuracy in recognizing various ocular diseases from retinal fundus images.

Prior research typically focuses on single ocular diseases and multi-classification tasks, overlooking the complexity of patients with multiple conditions. To address this gap, the Ophthalmic Image Analysis-Ocular Disease Intelligent Recognition (OIA-ODIR) dataset was introduced (Li et al., 2021). It's a large-scale dataset for multi-disease detection based on binocular fundus images. The dataset's proposed network structure incorporates late feature fusion before final prediction. Benchmark assessments highlight the need for structured feature fusion methods, while the inclusion of patient demographics holds potential to enhance diagnostic accuracy. This study aims to advance automated ocular disease classification by evaluating changes in model architecture via fusion timing, fusion method, and feature inclusion on deep convolutional neural network (CNN) performance for multi-label classification of ocular diseases. The goal is to improve the model's performance on multi-label tasks, offering insights into ocular disease recognition.

2 Hypothesis

Drawing from insights into multi-modal machine learning and fusion methods, this study hypothesizes that early fusion of binocular fundus images with tabular patient data (age and sex), utilizing the concatenation method for image features, will enhance multi-label classification performance on the OIA-ODIR dataset compared to late fusion and the fundus image-only approach. This hypothesis is grounded in the understanding that early fusion integrates diverse data sources at the beginning, combining raw or minimally processed data to exploit cross-modal correlations effectively. By concatenating binocular fundus images while retaining separate tabular patient data, the model is anticipated to preserve the distinctive features of each modality, facilitating joint learning and capturing nuanced associations between them.

3 Data

The OIA-ODIR dataset is a comprehensive multi-disease fundus dataset (noa, 2019). It comprises 10,000 images from the left and right eyes of 5,000 patients, featuring binocular color fundus photographs labeled with eight categories: Normal (N), Diabetic Retinopathy (D), Glaucoma (G), Cataract (C), Age-Related Macular Degeneration (A), Hypertension (H), Myopia (M), and Other diseases (O). The images were sourced from 487 clinical hospitals across 26 provinces in China. Professional annotators and experienced ophthalmologists categorized images based on disease presence and severity, adhering to strict standards for accuracy and reliability.

The dataset also includes patient demographics (age and sex) and diagnostic keywords from ophthalmologists, distributed across training (3,500 patients), validation (500 patients), and test (1,000 patients) sets. Table 1 illustrates the distribution of images per disease category, revealing notable class imbalance, with 'Normal' cases predominating and 'Hypertension' cases as the minority. Patient demographics reveal a male count of 1,869 and a female count of 1,602, with about 75% of individuals being over the age of 40. OIA-ODIR's comprehensive annotations and patient data make it invaluable for ocular disease recognition algorithm development and evaluation, addressing real-world clinical complexities effectively.

Labels	N	D	G	C	A	H	M	O
Training Cases	1140	1105	200	203	161	99	168	952
Validation Cases	162	160	30	30	22	15	23	131
Testing Cases	324	320	56	63	45	27	46	269
All Cases	1626	1585	286	296	228	141	237	1352

Table 1: Proportion of images per category in training, validation, and test sets.

4 Materials & Methods

4.1 Data Pre-Processing

Several steps were implemented to refine the dataset for efficient analysis. One-hot encoded disease classifications were transformed into binarized vectors to accurately represent multi-label instances. Diagnostic keywords were removed from consideration as an input feature. Cases containing more than two disease labels were omitted for the training, validation and test sets to streamline multi-label prediction, resulting in the removal of 42 instances (29 from training, 5 from validation, and 8 from test). Furthermore, patient sex was encoded as binary values, with 1 representing male and 0 representing female. Patient age was left unaltered. The fundus images were transformed by resizing them to dimensions of 128x128 pixels and normalizing them using ImageNet mean and standard deviation values ([0.485, 0.456, 0.406] for mean and [0.229, 0.224, 0.225] for standard deviation). These transformations were applied to facilitate the use of pre-trained weights for model training.

4.2 Model Architecture and Setup

All models utilized the ResNet-18 architecture as the foundation for feature extraction and classification tasks. ResNet (Residual Neural Network) addresses the challenges posed by training very deep networks by introducing skip connections (He et al., 2016). These connections allow the network to directly learn residual mappings, thereby alleviating the vanishing gradient problem observed in deeper networks. ResNet-18, a specific implementation, has 18 layers. For multi-label classification, the ResNet-18 models were tailored to suit the task's demands. Inputs consisted of preprocessed fundus images, with the option to include patient age and sex. The labels, represented as binary vectors, were encoded using one-hot encoding to indicate the absence or presence of specific conditions. The model's outputs generated logits for eight classes, which were arranged into two columns representing the absence or presence of specific classes. This approach treated the task as eight separate two-class classifications, allowing for the calculation of confidence scores for each label.

4.2.1 Model Variations

Within the ResNet-18 framework, late fusion and early fusion strategies were explored alongside fusion methods (element-wise product, element-wise sum, and concatenation) to optimize multi-label classification. Late fusion merges features in later network stages, while early fusion integrates them at initial layers. Under each fusion strategy, one of the fusion methods - PROD (element-wise product), SUM (element-wise sum), or CONCAT (concatenation) - was applied to merge image feature tensors. The CONCAT method preserves individual representations within the binocular fundus images, while PROD and SUM methods emphasize interactions between the binocular fundus images. Integrating multi-modal inputs, comprising patient age and sex, involved concatenating them with image features either at the input layer (for early fusion) or fusion layer (for late fusion), thus facilitating the fusion process and augmenting classification performance.

4.2.2 Evaluation Metrics

When evaluating architecture performance, three primary metrics were considered: the kappa coefficient, F1 score, and Area Under the Curve (AUC). Kappa quantifies agreement between predicted and true class labels, accounting for chance agreement. F1 score, which balances precision and recall, is essential for accurately identifying positive cases. AUC assesses the model's ability to distinguish between classes. To ensure balanced performance across all classes, the macro F1 score was utilized. Additionally, both macro and per-class AUC were computed to provide a comprehensive assessment of architecture performance. Finally, the average of the kappa, macro F1, and macro AUC scores was calculated to yield a holistic single score, denoted as Final Score, for comparing the performance of different model architectures.

4.2.3 Hyperparameter Tuning & Model Training

Hyperparameter tuning was performed using random subsets of the full training, validation, and test sets, with each subset comprising 5% of the data from each class label. Batch size and learning rate were optimized specifically for the ResNet-18 Late Fusion SUM (LFS), fundus image-only architecture. This choice was based on findings from the benchmark paper, where the LFS architecture for ResNet-18 demonstrated superior performance compared to other fusion methods. To ensure a fair comparison, consistent hyperparameter values were maintained across all models tested. The primary objective of tuning was to maximize the macro F1 score while ensuring generalization to the validation set. The tuning process spanned 30 epochs, resulting in optimal values of 16 for batch size and 0.00001 for learning rate.

For training the model on the full dataset, the number of epochs was limited to 10 due to time constraints. The training and validation loss plots presented in Section 6: *Appendix* depict a consistent decrease towards zero for training loss throughout these epochs for each architecture type investigated. The loss function used was Binary Cross Entropy with logits, and gradient descent optimization was performed using the Adam Optimizer. The model outputs were transformed into probabilities using the sigmoid function for computing kappa and macro F1 scores. These probabilities were then thresholded to obtain binary predictions. For the calculation of macro and per-class AUC, the 'presence of condition' column was selected from both the labels and the probabilities.

5 Results

Following model training, the kappa coefficient, macro F1, macro AUC, and final score were computed for each of the twelve model architectures on the test set, as illustrated in Figure 1. The evaluation identified the late fusion image-only SUM architecture (Kappa = 0.724, F1 = 0.862, AUC = 0.751, Final = 0.779) as the top performer among the twelve variations tested, closely followed by the late fusion multi-modal input SUM architecture (Kappa = 0.726, F1 = 0.863, AUC = 0.744, Final = 0.778). Conversely, the worst performing architecture overall was the early fusion multi-modal CONCAT architecture (Kappa = 0.711, F1 = 0.855, AUC = 0.657, Final = 0.741), which was initially hypothesized to be the best performer.

Results for the per-class AUC scores among the top two performing architectures and the hypothesized best performing architecture are depicted in Figure 2. The late fusion image-only SUM architecture exhibited a high ability to distinguish Cataract (AUC = 0.96), Myopia (AUC = 0.95), and Glaucoma (AUC = 0.82) cases, but performed poorly in distinguishing Other diseases (AUC = 0.53), followed by AMD (AUC = 0.63). The late fusion multi-modal SUM architecture shows a similar trend, excelling in distinguishing Cataract (AUC = 0.97), Myopia (AUC = 0.94), and Glaucoma (AUC = 0.80) cases, but struggling in distinguishing Other diseases (AUC = 0.52), followed by AMD (AUC = 0.60). For the hypothesized best performing architecture (early fusion multi-modal CONCAT), the discriminatory ability among classes follows a similar order, but with lower AUC scores, especially for classes such as AMD and Other diseases, which exhibit scores at or below 0.5. The overall macro AUC is notably lower at 0.66, almost 10 points lower than the macro AUC for the late fusion image-only SUM architecture.

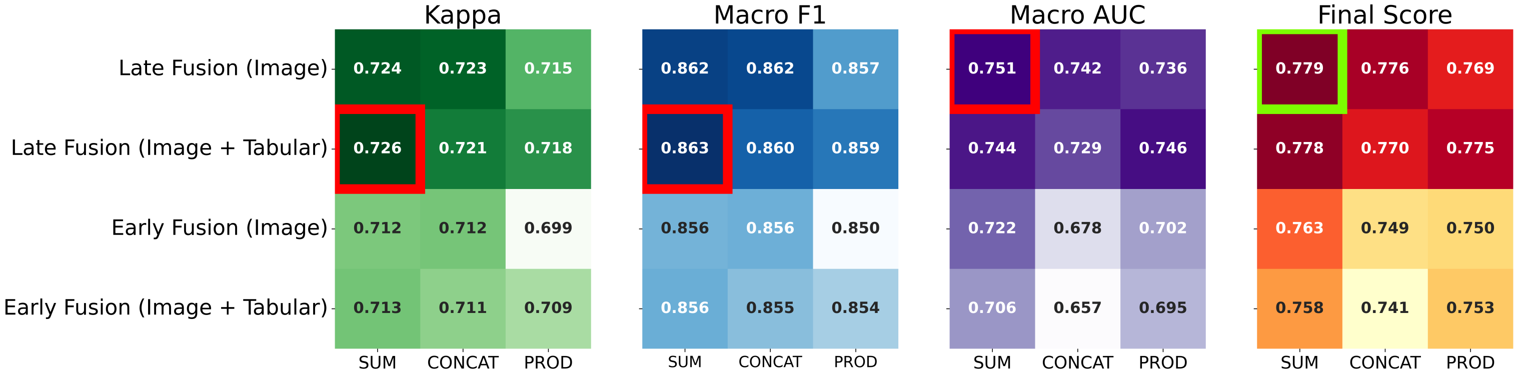


Figure 1: Model architecture comparisons using various test set performance metrics.

6 Discussion

This study hypothesized that early concatenation of binocular fundus image features, combined with patient age and sex, would outperform the late image-only feature summation architecture on the OIA-ODIR dataset. This assumption stemmed from early fusion's potential to integrate diverse data sources upfront, exploiting cross-correlations between image features, along with potential additional context from patient tabular data. However, contrary to expectations, the late fusion image-only SUM method yielded optimal classification performance. The emphasis on fine-tuning efforts primarily on the late fusion SUM method could account for the comparatively poorer performance of other models. Further architectural tuning might be necessary, but due to time constraints, this study was conducted using consistent hyperparameter values.

Late fusion architectures consistently outperformed their early fusion counterparts in terms of kappa, macro F1, and macro AUC scores on the test set, indicating superior model performance. This advantage potentially stems from late fusion's approach of processing each image independently, preserving their unique information before combining their extracted features. Despite its computational overhead, late fusion offers increased predictive accuracy and discrimination. Although there are trade-offs between late and early fusion strategies, the disparities in results are generally not that significant, particularly when employing an appropriate feature fusion method such as element-wise summation. Among image feature fusion methods, element-wise summation consistently produced superior results regardless of fusion timing. Summation combined information from both fundus images by adding their features, thus preserving all information and enhancing representation. Conversely, product fusion, which multiplied image features to capture interactions, may have amplified differences, cre-

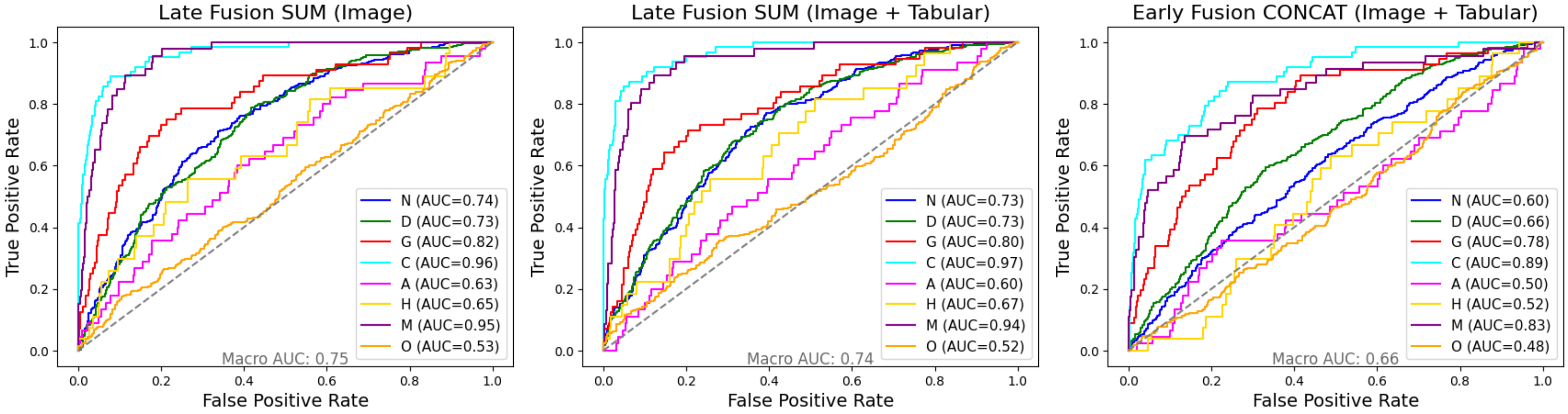


Figure 2: ROC curves for top models and hypothesized best performer based on final scores.

ating sensitivity to scale variations. Concatenation, which stacked image features to increase dimensionality, may face challenges related to data sparsity and complexity. However, within a specific fusion scheme (late or early), the choice of fusion method generally shows minimal difference, indicating comparable model performance overall. Effective multi-modal data fusion relies on the choice of technique used to construct the representation. While image-only fusion often benefits from sum and concatenation methods, element-wise product fusion of images has shown to enhance model performance when combined with tabular data. The integration of additional informative patient data (if accessible) could potentially improve model performance.

The top two performing architectures (LFS image-only and LFS multi-modal) struggled to distinguish the 'Other' class from the rest, which is understandable given its diverse makeup. Interestingly, despite being minority labels, both models excelled at differentiating Cataracts and Myopia, likely due to the distinct appearance of these conditions. The incorporation of age and gender into the model potentially aided in distinguishing certain classes like Hypertension and Cataract, as these conditions may exhibit age or gender-related patterns. However, considering that the majority of training and test data consists of patients over 40 years, this may not significantly impact AUC for these conditions or others. The models' performance on specific classes appears to be more influenced by image quality and inherent distinctiveness of certain conditions.

Critical considerations in model evaluation emphasize recognizing realistic sources of errors. Ensuring correct label columns during metric computation prevents misjudgment, while consistent data processing steps avoid discrepancies. Hyperparameter tuning may bias results, and dataset characteristics like image quality variations can introduce uncertainties. Addressing these pitfalls ensures robust experimental findings. The findings in this study stress the importance of selecting fusion techniques tailored to data characteristics, offering insights into optimizing model performance for multi-label classification of ocular diseases.

References

OIA-ODIR Dataset. https://github.com/nkicsl/OIA-ODIR?tab=readme-ov-file, 2019.

Can Cui, Haichun Yang, Yaohong Wang, Shilin Zhao, Zuhayr Asad, Lori A Coburn, Keith T Wilson, Bennett A Landman, and Yuankai Huo. Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. *Prog Biomed Eng (Bristol)*, 5(2), April 2023.

Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, pages 1–6. IEEE, July 2020.

T Haylat. INTRODUCTION TO DATA FUSION - haileleol tibebu - medium. https://medium.com/haileleol-tibebu/data-fusion-78e68e65b2d1, January 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE. ISBN 9781467388511. doi: 10.1109/CVPR.2016.90. URL http://ieeexplore.ieee.org/document/7780459/.

Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P Lungren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digital Medicine*, 3(1), December 2020.

Ning Li, Tao Li, Chunyu Hu, Kai Wang, and Hong Kang. A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection. February 2021.

Issa Memari. Precision, recall, accuracy, and F1 score for multi-label classification. https://medium.com/synthesio-engineering/precision-accuracy-and-f1-score-for-multi-label-classification-34ac6bdfb404, January 2021. Accessed: 2024-5-6.

Stewart Muchuchuti and Serestina Viriri. Retinal disease detection using deep learning techniques: A comprehensive review. *Journal of Imaging*, 9(4):84, April 2023.

National Eye Institute NEI. Eye disease statistics fact sheet. https://www.nei.nih.gov/sites/default/files/2019-04/NEI_Eye_Disease_Statistics_Factsheet_2014_V10.pdf, March 2014.

Sivarathri Susrutha and Robin Prakash Mathur. Review on ocular disease recognition using deep learning. In *2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, pages 316–321. IEEE, May 2023.

World Health Organization WHO. Blindness and vision impairment. https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment, August 2023.

Chao Zhang. Performance comparison between early fusion and late fusion in video analysis. https://chaozhangchn.medium.com/performance-comparison-between-early-fusion-and-late-fusion-5f9d88ffce66, May 2021. Accessed: 2024-5-6.

Appendix

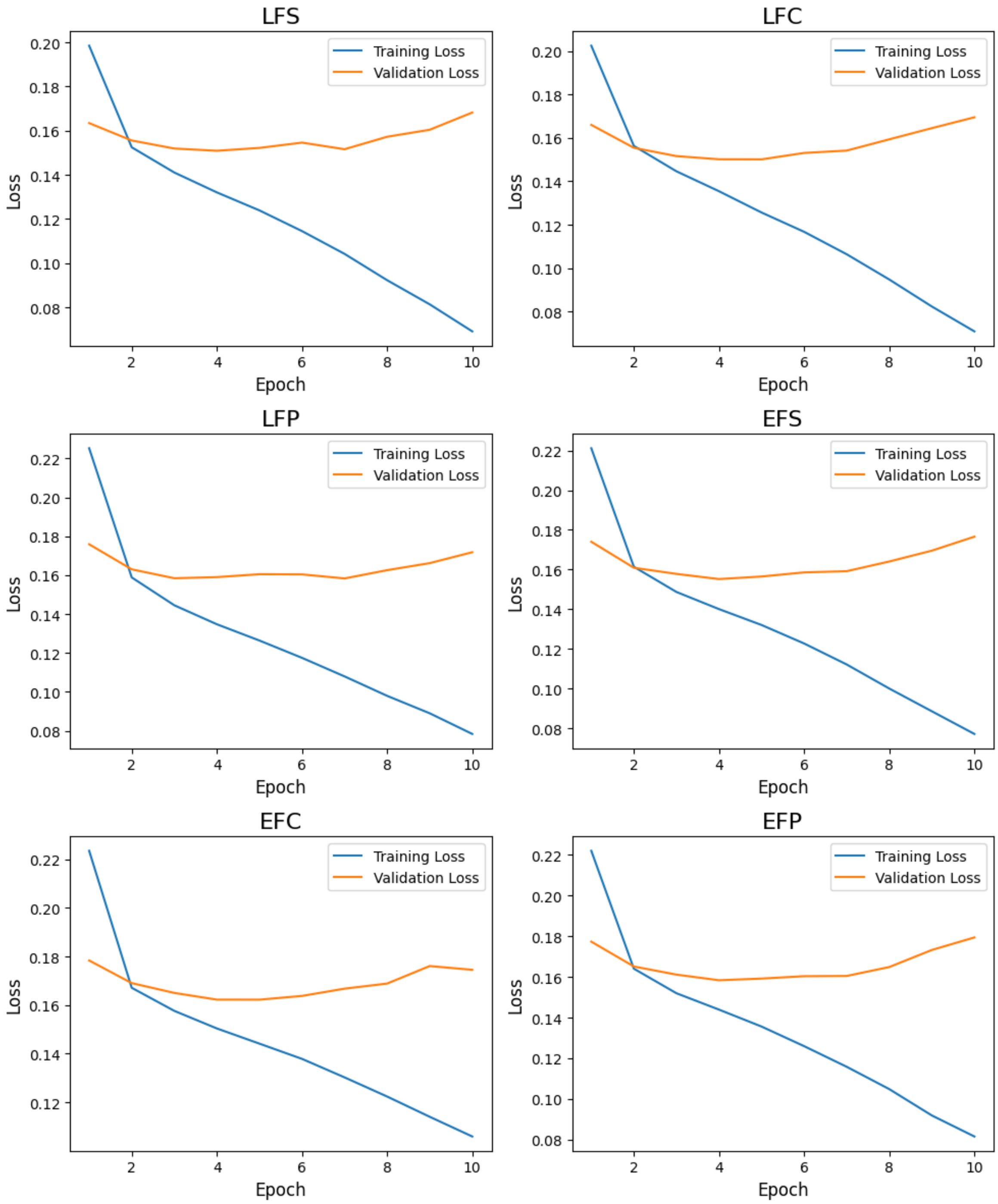
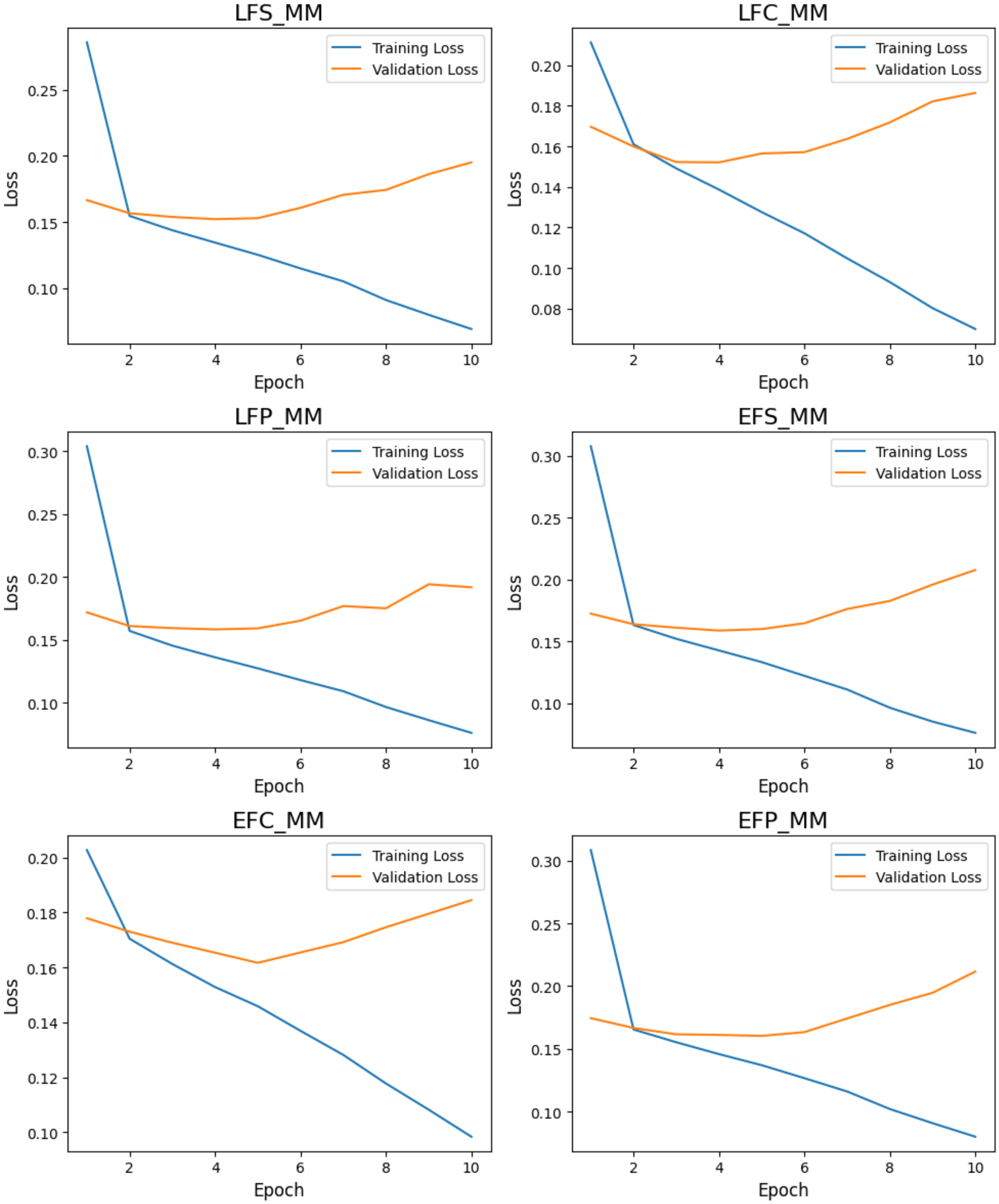


Figure 3: Training and validation losses for various image-only input architectures.



LFS_MM

LFC_MM

LFP_MM

EFS_MM

EFC_MM

EFP_MM

Figure 4: Training and validation losses for various multi-modal input architectures.