METAFUSION: A NOVEL METHOD FOR INTEGRATING CLINICAL METADATA WITH IMAGING MODALITIES FOR MEDICAL APPLICATIONS

Ananya Raghu, Anisha Raghu

ABSTRACT

Clinical history is the primary determinant of medical diagnosis and must be utilized with imaging data for accurate identification of medical conditions. We propose **MetaFusion**, a novel method for multi-modal data fusion that integrates clinical data with imaging data to improve the quality of diagnosis across a range of diseases using three distinct datasets. We demonstrate significant improvements over prior methods in skin cancer type detection using camera images, breast condition detection via mammograms, and glaucoma detection through retinal fundus images. Specifically, we observe improvements of 3.5% in balanced accuracy for skin cancer type detection, 9.7% for breast condition detection, and 7.1% for glaucoma detection compared to the top-performing prior methods.

***Index Terms*—** multimodal fusion, medical diagnosis, skin cancer, breast cancer, glaucoma, transfer learning

1. INTRODUCTION

Every year, hospitals worldwide generate over 3.6 billion medical images, including X-rays, MRIs, camera images, and CT scans. Recent advancements in artificial intelligence have led to a multitude of methods that leverage computer vision to assist healthcare professionals in analyzing image data. This has enabled the early detection of various medical conditions, including different types of cancer, heart disease, neurological disorders, and infectious diseases such as tuberculosis and pneumonia [1]. Although relevant clinical information is often provided as metadata, its availability in datasets is limited due to privacy considerations and challenges in selecting appropriate fusion methods [2].

In this study, we focus on three prevalent diseases: breast cancer, skin cancer, and glaucoma. Breast cancer is the most frequently diagnosed cancer worldwide and accounts for over 11% of all new cancer cases [3]. Skin cancer is also one of the most prevalent forms of cancer in the world. One in every 3 diagnosed cancers is skin cancer, and 132,000 new cases of melanoma occur each year [4]. In addition, glaucoma is a major cause of blindness worldwide, affecting over 120,000 people in the United States alone [5]. These diseases are typically diagnosed by obtaining specific imaging data—such as mammograms for breast cancer, dermoscopy images for skin cancer, and retinal fundus images for glaucoma—which are then reviewed by a specialist. Advances in image classification using deep learning techniques have been applied successfully to improve diagnosis of multiple diseases, including the three diseases of focus in this work [6–8]. However, all of these methods were not multi-modal and used only imaging data. Multi-modal fusion approaches combining clinical data with imaging have recently been proposed for skin cancer detection, showing the value of combining information from different modalities. [9–13]. In this paper we propose MetaFusion, a novel multi-modal fusion method which has several advantages in comparison to prior methods.

1.1. Novelty of MetaFusion approach

- MetaFusion is an innovative fusion method that shows superior performance for the detection of three diseases with different input modalities: skin cancer classification (from smartphone images), breast cancer detection (from X-ray mammograms) and glaucoma detection (from retinal fundus images). The metadata used in each of these cases is different.
- To our knowledge, MetaFusion is the first method to show consistent performance across three datasets (many prior methods focus solely on one disease such as skin cancer), demonstrating the robustness of our model in multimodal applications.
- We show that the proposed approach performs with higher accuracy compared to previous fusion strategies such as direct concatenation, MetaNet, and MetaBlock.
- In addition, our method is simple and highly parameter efficient, with the fusion block requiring only 1% of the number of parameters compared to methods that rely on cross-attention for feature fusion.

2. MATERIAL AND METHODS

2.1. Prior Work

The fusion of metadata, which often contains valuable information about the patient profile, with medical image information can lead to more accurate diagnosis of disease [11, 14, 15]. We focus on approaches that perform late fusion of embeddings of image and metadata, where the image embedding is generated by a feature extractor and the

metadata is mapped to an embedding via one-hot encoding followed by a linear layer. One common approach for the integration of metadata with medical images is through the direct concatenation of embeddings. While concatenation allows for the model to consider both image and metadata information, it has several limitations. Because medical images contain high-dimensional data of greater complexity compared to metadata, a concatenation-based approach may not be very effective in extracting relevant information from both the metadata and the image. A bigger limitation is that concatenation does not directly consider the potential effect of the metadata information on the image embeddings. For example, knowing a patient's skin tone can be useful in determining which visual features are important for skin cancer detection - in essence, metadata can help identify the most relevant parts of medical images. An alternative approach is the MetaNet architecture (Equation 1), which implements multiplication-based fusion for combining medical image information with metadata [9], where $\odot$ refers to the hadamard product.

$$\tilde{x}_{img} = x_{img} \odot \sigma(W x_{meta}) \tag{1}$$

Recently, the MetaBlock approach ((Equation 2) was proposed [10] for the integration of metadata in image classification. This method works similarly to an attention mechanism and helps the model concentrate on more important features by incorporating the metadata information into the image feature maps. The key idea of Metablock is to apply LSTM-like gates to select the most relevant features in the image embeddings.

$$\tilde{x}_{img} = \sigma\big(\big(\tanh(x_{img} \odot (W_1 x_{meta})\big) + W_2 x_{meta}\big) \tag{2}$$

Cross-attention based fusion was also proposed for fusing metadata with imaging data for skin cancer detection [11]. The idea of mutual attention to fuse information between the metadata and the image embeddings was introduced in [13]. The mutual attention block contains two multi-head cross attention modules to fuse information from images and metadata. The output of the cross attention block is used to modify the image and metadata embeddings using the residual-connection mechanism. Finally, the image and metadata embeddings are concatenated. In contrast, in [12] the image and metadata embeddings are concatenated into a longer sequence and this sequence is then passed through a self attention layer.

2.2. MetaFusion Algorithm

MetaFusion is a novel fusion method that combines intermediate representations (embeddings) prior to the final layer of a model, following the high level architecture shown in Figure 1.

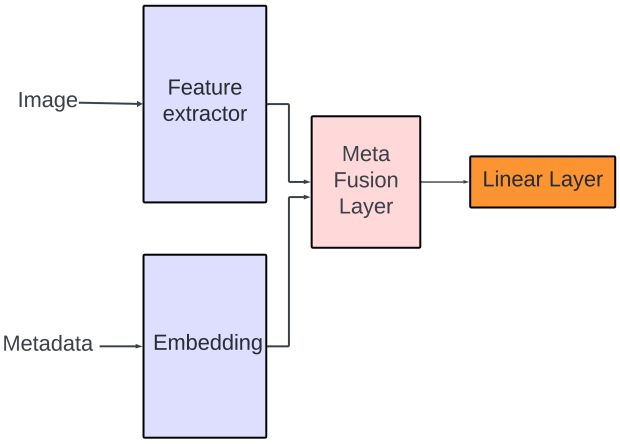

Fig. 1. High level pipeline of image and metadata fusion in machine learning.

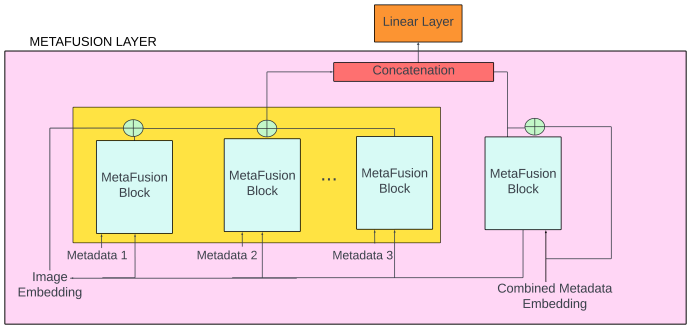

Fig. 2. MetaFusion layer consisting of multiple MetaFusion blocks with skip connections to modify the image based on metadata embeddings, and vice versa.

The MetaFusion layer, consists of multiple MetaFusion blocks as shown in Figure 2.

2.2.1. *MetaFusion Block*

We build the MetaFusion block based on the following ideas:

1. Apply a correction term to image and metadata embeddings prior to concatenation using the idea of residual connections.

$$x_1 = x_1 + f(x_1, x_2)$$
$$x_2 = x_2 + f(x_2, x_1)$$

2. We pick the correction term to be of the form $f(x_1, x_2) = x_1 \odot s(x_1, x_2)$ where $s(x_1, x_2)$ is a similarity function that ranges from -1 to 1.
3. We choose the similarity function to be $s(x_1, x_2) = tanh(x_1 \odot W x_2)$, inspired by the design of MetaBlock [10].

Putting it all together, we define the MetaFusion block as:

$$f(x_1, x_2) = x_1 \odot tanh(x_1 \odot (Wx_2)) \quad (3)$$

We obtain an image embedding $\mathbf{x_{image}}$ by passing the image through a pre-trained backbone. The metadata is converted into an embedding via a simple embedding look-up. Note that in earlier approaches such as MetaNet and MetaBlock, the metadata is simply used to modify the image embedding, but its features are not directly used. Only the concatenation and mutual attention based approaches use the metadata and image embeddings as inputs to the final layer of the respective models.

2.2.2. MetaFusion Layer

The MetaFusion layer consists of multiple MetaFusion blocks (Equation 4 and 5) in order to modify the image and metadata. Note that f refers to the MetaFusion block as shown in Equation 3.

$$\begin{aligned}\tilde{\mathbf{x}}_{\mathbf{img}} =& \mathbf{x}_{\mathbf{img}} + f(\mathbf{x}_{\mathbf{img}}, \mathbf{x}_{\mathbf{meta,1}}) + \\ & f(\mathbf{x}_{\mathbf{img}}, \mathbf{x}_{\mathbf{meta,2}}) + \cdots f(\mathbf{x}_{\mathbf{img}}, \mathbf{x}_{\mathbf{meta,n}}) \end{aligned} \quad (4)$$

$$\tilde{\mathbf{x}}_{\mathbf{meta}} = \mathbf{x}_{\mathbf{meta}} + f(\mathbf{x}_{\mathbf{meta}}, \mathbf{x}_{\mathbf{img}}) \quad (5)$$

2.3. Implementation Details

Image feature extraction is performed using a ResNet-18 backbone [16], which is initialized with weights pre-trained on ImageNet. For all experiments, models are trained for 50 epochs with a cosine learning rate scheduler and an Adam optimizer with an initial learning rate of 10^{-4}. Softmax cross entropy is used as the loss function in model training.

2.4. Datasets

Dataset	Number of Patients	Number of Samples	Classes	Clinical Features
PAD-UFES-20	1373	2298	6	6
CBIS-DDSM	892	2620	3	7
PAPILA	244	488	3	7

Table 1. High Level Dataset Summary

We make use of three different datasets (shown in Table 1) to demonstrate the robust performance of the MetaFusion model.

The PAD-UFES-20 dataset [17] consists of 2298 clinical images of skin lesions collected from different smartphone devices and patient clinical data related to each skin lesion. We split this dataset into 70% for the training set, 15% for the validation set and 15% for the testing set, and made use of the following clinical features: Skin cancer history, gender, fitzpatrick scale and three features describing the lesion: region in body, elevation and change. The task on this dataset is to perform six class classification between three skin diseases and three skin cancers. For more details on the clinical features, please see [17] .

The CBIS-DDSM [18] dataset contains 2620 mammography images along with metadata. We used the train-test split provided in the dataset. The task on this dataset is to classify the data into three classes: benign, malignant and benign with callback. We used six clinical features describing the abnormal breast tissue (known as a mass) provided as part of the dataset - mass margin, mass shape, number of abnormalities (identified through a segmentation algorithm, came with the dataset), breast density, radiologist difficulty in viewing abnormality and BI-RADS assessment.

The PAPILA dataset [19] contains 488 retinal fundus images along with clinical data. We split this dataset into 70% for the training set, 15% for the validation set and 15% for the testing set. The task on this dataset is to classify the data into three classes: normal, glaucoma and suspicious. We used all seven clinical features (age, gender, axial length, corneal thickness, intraocular pressure, refractive error, presence/absence of crystalline lens) provided as part of the dataset.

3. RESULTS

3.1. Metrics

Since the classification task is multi-class classification on datasets with class imbalance, we report balanced accuracy, precision, and F1 score metrics.

Standard accuracy can be misleading when the dataset is imbalanced, as the model can perform "well" by consistently predicting the majority class label which leads to a deceptively high standard accuracy measurement. Balanced accuracy gives a more realistic assessment by treating each class equally and taking the average of the recall for each class, working as a better measurement for imbalanced data. Precision is also reported, and F1 score provides an overall balanced evaluation by taking both the precision and recall metrics into account.

3.2. Key Results

Our key results are presented in Table 2, demonstrating the robustness and accurate performance of our model. Multiple metrics are used to quantify and compare the performance of the MetaFusion algorithm to prior fusion methods. We performed five-fold cross validation for all the experiments. The proposed MetaFusion method consistently outperforms other techniques in all three datasets, showing the most improvement for the CBIS-DDSM dataset.

	Balanced Accuracy			F1 score			Precision		
Fusion Method	**PAD UFES-20**	**CBIS DDSM**	**PAPILA**	**PAD UFES-20**	**CBIS DDSM**	**PAPILA**	**PAD UFES-20**	**CBIS DDSM**	**PAPILA**
Baseline	70.2 ± 2.6	48.1 ± 1.6	58.3 ±5.8	69.8 ± 2.8	47.4 ± 2.1	**58.8 ±5.8**	70.6 ± 3.0	49.7 ± 1.9	**60.5 ±5.5**
Concat	70.3 ± 3.0	48.4 ± 1.0	55.0 ± 4.1	71.0 ± 3.6	47.7 ± 1.2	55.3 ±4.4	72.4 ± 4.2	49.7 ±1.2	56.4 ± 4.8
MetaNet	**73.3 ± 1.5**	50.6 ± 1.5	**58.7 ± 4.0**	73.9 ±2.1	50.6 ± 1.8	58.6 ± 4.4	75.4 ±2.7	55.3 ± 4.6	59.2 ± 5.2
MetaBlock	72.9 ± 2.8	**58.7 ± 1.6**	50.2 ± 2.5	**74.9 ± 2.5**	**60.8 ± 1.9**	49.2 ±1.0	**78.3 ± 2.6**	**75.4 ± 1.4**	50.6 ± 2.8
MetaFusion	76.8 ± 1.1	68.4 ± 1.2	65.8 ± 4.1	77.1 ±1.4	70.1 ± 1.6	68.4 ± 3.7	78.4 ± 2.0	73.6± 2.7	73.4 ± 5.1

Table 2. Balanced accuracy, F1 score and Precision metrics, with standard deviation for fusion methods across datasets. MetaFusion is consistently the best performing method across all metrics shown.

3.3. Ablation Study

To determine the contribution of each component of the proposed MetaFusion approach to the model, we perform three ablation studies. In the first ablation study, we add an additional linear layer in the MetaFusion block:

$$x_1 = x_1 + x_1 \odot tanh\big(x_1 \odot W_2(ReLU(W_1 x_2))\big)$$

In the second ablation study, we remove the residual connection:

$$x_1 = x_1 \odot tanh\big(x_1 \odot W x_2\big)$$

In the third ablation study, we determine the importance of having a similarity function, by replacing it with a function of x_2, similar to MetaNet:

$$x_1 = x_1 + x_1 \odot tanh((W x_2))$$

The results from the performed ablation studies are presented in Table 3.

Method	**Balanced Accuracy**
MetaFusion	68.4 ± 1.2
+ Additional Linear Layer	66.7 ± 0.2
- Skip Connection	57.9 ± 1.3
- Hadamard Product Attention	64.1 ± 2.1

Table 3. Ablation studies on different variants of MetaFusion block on the CBIS DDSM dataset. Results show that skip connections are crucial for fusion, followed by the design of the similarity function.

As shown in Table 3, our method outperforms the ablation with no residual connection (Ablation 1), indicating the importance of skip connections in our model. Our method also performs better than Ablation 2, signifying that our current model architecture with just one linear layer outperforms a model consisting of two linear layers. MetaFusion also outperforms the Ablation 3 model, showing that the similarity function also provides performance gains. The ablation studies suggest that the residual connection and structure of the similarity functions are crucial to the performance of the MetaFusion block.

3.4. Complexity

Our method is highly parameter efficient, compared to transformer based approaches for fusion. Given an image embedding of length d_{img}, the number of parameters for MHA is $12d_{img}^2$. The complexity of our method is simply $Nd_{metadata}d_{img}$, where N is the total number of metadata features and $d_{metadata}$ is the dimensionality of the metadata. With $N = 6$, $d_{img} = 2048$ and $d_{metadata}$=16, the number of parameters in our approach is about one percent of the number of parameters in the multi-headed attention approach.

4. CONCLUSION

Our method is robust and consistently has the highest balanced accuracy, F1 score and precision among all three datasets while other techniques show a lot of variability across datasets. We note that the second best performing method for the balanced accuracy metric (MetaNet, MetaBlock, MetaNet) for each of the three datasets in our study is different as indicated in bold in Table 2. We attribute the robustness of our method to the design of the similarity function, skip connections and modifying both image and metadata embeddings prior to concatenation. As a next step, we plan to verify the performance of MetaFusion on larger datasets and more diseases with both different imaging modalities and more complex metadata. We note that more rigorous evaluation with diverse patient profiles is essential to validate the results observed in this paper. The success of this novel AI-based diagnostic approach offers the potential for more precise disease detection, paving the way for preventative healthcare strategies in the future.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available by open access datasets([17], [18] and [19]). Ethical approval was not required as confirmed by the license attached with the open access datasets.

6. ACKNOWLEDGMENTS

We thank Dr. Jillian Wise for her review and detailed feedback.

7. REFERENCES

[1] The White House, "Delivering on the promise of AI to improve health outcomes," Dec. 2023, Accessed: 2024-6-26.

[2] Sachin Kumar, Sita Rani, Shivani Sharma, and Hong Min, "Multimodality fusion aspects of medical diagnosis: A comprehensive review," *Bioengineering (Basel)*, vol. 11, no. 12, pp. 1233, Dec. 2024.

[3] Hangcheng Xu and Binghe Xu, "Breast cancer epidemiology, risk factors and screening.," *Chinese journal of cancer research*, 2023.

[4] Katelyn et al. Urban, "The global burden of skin cancer: A longitudinal analysis from the global burden of disease study, 1990-2017," *JAAD international*, 2021.

[5] "Glaucoma facts and stats - glaucoma research foundation," `glaucoma.org/articles/glaucoma-facts-and-stats`.

[6] Muhammad Azeem, "Skinlesnet: Classification of skin lesions and detection of melanoma cancer using a novel multi-layer deep convolutional neural network," *Cancers*, 2023.

[7] Selvakumar Thirumalaisamy, Kamaleshwar Thangavilou, Hariharan Rajadurai, Oumaima Saidani, Nazik Alturki, Sandeep Kumar Mathivanan, Prabhu Jayagopal, and Saikat Gochhait, "Breast cancer classification using synthesized deep learning model with metaheuristic optimization algorithm," *Diagnostics (Basel)*, vol. 13, no. 18, pp. 2925, Sept. 2023.

[8] Vijaya Kumar Velpula and Lakhan Dev Sharma, "Multi-stage glaucoma classification using pre-trained convolutional neural networks and voting-based classifier fusion," *Front. Physiol.*, vol. 14, pp. 1175881, June 2023.

[9] Weipeng Li, Jiaxin Zhuang, Ruixuan Wang, Jianguo Zhang, and Wei-Shi Zheng, "Fusing metadata and dermoscopy images for skin disease diagnosis," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. Apr. 2020, IEEE.

[10] Andre G C Pacheco and Renato A Krohling, "An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 9, pp. 3554–3563, Sept. 2021.

[11] Chubin Ou, Sitong Zhou, Ronghua Yang, Weili Jiang, Haoyang He, Wenjun Gan, Wentao Chen, Xinchi Qin, Wei Luo, Xiaobing Pi, and Jiehua Li, "A deep learning based multimodal fusion model for skin lesion diagnosis using smartphone collected clinical images and metadata," *Front. Surg.*, vol. 9, pp. 1029991, Oct. 2022.

[12] Theodor Cheslerean-Boghiu, Melia-Evelina Fleischmann, Theresa Willem, and Tobias Lasser, "Transformer-based interpretable multi-modal data fusion for skin lesion classification," Apr. 2023.

[13] Gan Cai, Yu Zhu, Yue Wu, Xiaoben Jiang, Jiongyao Ye, and Dawei Yang, "A multimodal transformer to fuse images and metadata for skin disease classification," *Vis. Comput.*, pp. 1–13, May 2022.

[14] Fumitoshi Fukuzawa, Yasutaka Yanagita, Daiki Yokokawa, Shun Uchida, Shiho Yamashita, Yu Li, Kiyoshi Shikino, Tomoko Tsukamoto, Kazutaka Noda, Takanori Uehara, and Masatomi Ikusaka, "Importance of patient history in artificial intelligence-assisted medical diagnosis: Comparison study," *JMIR Med. Educ.*, vol. 10, pp. e52674, Apr. 2024.

[15] M C Peterson, J H Holbrook, D Von Hales, N L Smith, and L V Staker, "Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses," *West. J. Med.*, vol. 156, no. 2, pp. 163–165, Feb. 1992.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[17] Andre G C Pacheco, Gustavo R Lima, Amanda S Salomão, Breno Krohling, Igor P Biral, Gabriel G de Angelo, Fábio C R Alves, Jr, José G M Esgario, Alana C Simora, Pedro B C Castro, Felipe B Rodrigues, Patricia H L Frasson, Renato A Krohling, Helder Knidel, Maria C S Santos, Rachel B do Espírito Santo, Telma L S G Macedo, Tania R P Canuto, and Luíz F S de Barros, "PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," *Data Brief*, vol. 32, no. 106221, pp. 106221, Oct. 2020.

[18] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Sci. Data*, vol. 4, no. 1, pp. 170177, Dec. 2017.

[19] Oleksandr Kovalyk, Juan Morales-Sánchez, Rafael Verdú-Monedero, Inmaculada Sellés-Navarro, Ana Palazón-Cabanes, and José-Luis Sancho-Gómez, "PAPILA: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment," *Sci. Data*, vol. 9, no. 1, pp. 291, June 2022.