Differential Diagnosis of Thyroid Tumors through Information Fusion from Multiphoton Microscopy Images using Fusion Autoencoder

Harshith Reddy Kethireddy[1*], Tejaswee A[1*], Lucian G. Eftimie[2], Radu Hristu[3], George A. Stanciu[3], and Angshuman Paul[1]

[1] Indian Institute of Technology Jodhpur, India
[2] Central University Emergency Military Hospital, Bucharest, Romania
[3] National University of Science and Technology Politehnica Bucharest, Romania
reddy.16@iitj.ac.in, tejaswee.1@iitj.ac.in, lucicaeftimie@yahoo.com, radu.hristu@upb.ro, stanciu@physics.upb.ro, apaul@iitj.ac.in

Abstract. Thyroid carcinomas are often diagnosed by histopathology, which is widely regarded as the most reliable method. However, alternative imaging modalities may also provide meaningful information about thyroid tumors. Multiphoton Microscopy (MPM) images may be one of them. MPM images include Second Harmonic Generation (SHG) and Two-Photon Excitation Fluorescence (TPEF) images. Nevertheless, the field of automated analysis of MPM images for the diagnosis of cancer is in its infancy. We propose a strategy for the differential diagnosis of thyroid tumors through information fusion from different types of MPM images. We introduce a novel fusion autoencoder (FAE) for this task. The fused information from the FAE is subsequently used by a classifier module for the differential diagnosis of thyroid tumors. Our method is one of the first approaches to look into the possibility of using MPM images for the diagnosis of thyroid tumors. Extensive experiments demonstrate the superiority of the proposed method compared to a number of state-of-the-art classification techniques. The code will be released upon the completion of the review process.

Keywords: Multiphoton Microscopy Images · Information Fusion · Fusion Autoencoder · Thyroid Tumor

1 Introduction

Thyroid tumors can be either benign (e.g., Follicular Adenoma (FA)) or malignant (e.g., Follicular Thyroid Carcinoma (FTC)). Histopathological analysis of the nodule is generally required for the identification of malignancies. Multiphoton Microscopy (MPM) imaging is an advanced imaging technique that utilizes nonlinear optical processes, such as Second Harmonic Generation (SHG) and Two-Photon Excited Fluorescence (TPEF), to provide high-resolution, three-dimensional images of biological tissues. SHG microscopy is useful for imaging

[*] These authors contributed equally to this work

collagen or myosin. TPEF is emitted by proteins in tissue which are autofluorescent. Depending on excitation and detection wavelengths, different tissue components can be imaged [1][2]. Thus, MPM can provide information about the tissues that may not be obtained using conventional histopathology. Backward-detected SHG (BSHG) imaging reveals a punctate pattern stemming from the presence of small-diameter, segmental collagen, facilitating the detection of fibrillogenesis in immature tissue. In contrast, both Forward-detected SHG (FSHG) and BSHG images of mature collagen fibrils display identical features. Thus, SHG microscopy specifically offers detailed visualization of collagen architecture inside a neoplasm [3], while TPEF imaging offers essential insights into intranodular details [4], making MPM a valuable tool in biomedical research, particularly for investigating tissue morphology and pathology. Thus, a combination of images from BSHG, FSHG, and TPEF images can provide complementary information about the sample under investigation.

Although SHG images are used for medical imaging applications [5], their use for the diagnosis of thyroid tumors is less explored. Currently, the thickness of the nodule capsule is the only significant histopathological characteristic that is associated with malignancy [6], as compared to benign nodules. It is crucial to distinguish thyroid carcinomas from adenomas and nodular goiters [7] since a misdiagnosis of this condition can have severe consequences, despite its low incidence rate. Considering corneal edema detection as mentioned in [5], SHG microscopy data has proven to be useful to identify seamlessly and monitor the architectural changes in the collagen of the cornea. It uses deep learning techniques to classify edematous corneal tissues using a combination of multiple models, and the model of such a combination has given better estimates than using stand-alone models, viz., ResNet-50 [8], InceptionV3 [9], and the Flexible Lightweight Model for Bioimage Analysis (FLIMBA). In [10], the authors show the benefits of using SHG images of ovarian tissues which get characterized using deep learning methods.

We propose a method for the differential diagnosis of thyroid tumors using MPM images of multiple modalities. We design a fusion autoencoder (FAE) that takes a stack of BSHG, FSHG and TPEF images for a Region of Interest (RoI) as inputs and provides feature maps with the fused information. Although many researchers have designed automated methods for the analysis of histopathology images from thyroid nodules [11], the use of MPM images in this context is relatively rare. Our primary contributions to this work are as follows:

- We design a method for the differential diagnosis of thyroid tumors using MPM images of three modalities.
- We propose a fusion autoencoder that can fuse information from BSHG, FSHG and TPEF images. The fused information is used for the classification of thyroid tumors.
- We experimentally demonstrate the utility of using MPM images from three different modalities.

The rest of the paper is organized as follows. We discuss the proposed method in Section 2 followed by experiments and results in Section 3. Finally, the paper is concluded in Section 5.

2 Methodology

We introduce a novel approach for the differential diagnosis of thyroid tumors based on MPM images, including BSHG, FSHG, and TPEF modalities. Our goal is to classify tumors into FA and FTC categories by fusing information from these images. These diverse modalities offer complementary insights into thyroid nodules [12]. Our objective is to integrate information extracted from BSHG, FSHG, and TPEF images of thyroid nodule capsules to classify tumors into FA and FTC categories. To achieve this, we propose a methodology comprising two main components: an Information Fusion Module (IFM) and a Classification Module (CM). The IFM, crucial for combining data from multiple modalities effectively, is devised around a novel fusion autoencoder architecture used for our specific diagnostic task. The fused information derived from the IFM feeds into the CM for the final diagnosis. A visual representation of our proposed method is illustrated in Figure 1. This schematic encapsulates the integration of BSHG, FSHG, and TPEF images through our fusion autoencoder.

2.1 The Information Fusion Module (IFM)

In designing the IFM, several key considerations are taken into account. Firstly, it is of vital importance that the fusion process effectively fuses salient information from BSHG, FSHG, and TPEF images, each corresponding to an RoI. Autoencoders have demonstrated efficacy in extracting significant features from data [13]. Therefore, we propose employing a Fusion Autoencoder (FAE) for this task, aimed at extracting relevant information from the different modalities.

The FAE architecture, as depicted in Figure 1, comprises of one input head and three output heads. Each RoI generates a stack of corresponding BSHG, FSHG, and TPEF images, which are then stacked as input to the FAE. The encoder within the FAE creates a latent space representation from the stacked images across these three modalities. Subsequently, the decoder consists of common layers, along with private layers dedicated to each of the three output heads. These private layers facilitate the preservation and propagation of modality-specific information. The private layers help to propagate modality-specific information. One output head is tasked to reconstruct the BSHG image solely from the latent space representation, while the other two output heads aim to reconstruct the FSHG and the TPEF images, respectively. Because of this, the fused feature maps derived from the latent space are expected to encapsulate a fusion of the salient information from all three imaging modalities.

Let $I_B(n)$, $I_F(n)$, $I_T(n)$ be the BSHG, FSHG and TPEF images, respectively corresponding to RoI n. We create a stack

$$I(n) = \{I_B(n), I_F(n), I_T(n)\} \tag{1}$$

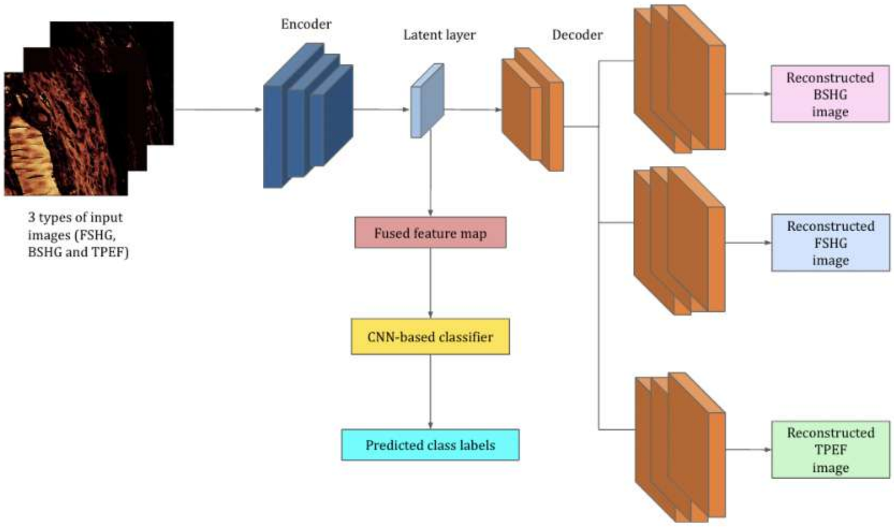


Fig. 1: A block diagram of the proposed method consisting of IFM (designed using FAE) and CM (designed using CNN-based classifier). A stack of BSHG, FSHG, and TPEF images of a single channel resulting in an image of 512 × 512 × 3-dimensional input is applied to the input layer of the encoder. The latent space representations are passed onto the decoder component that reconstructs the BSHG, FSHG, and TPEF images through three separate branches, respectively. L_B, L_F, L_T are losses from the reconstructed BSHG, FSHG, and TPEF images, respectively. The CNN-based classifier predicts the class label using the latent space representation (fused feature maps).

and apply $I(n)$ as input to the FAE. Let $I'_B(n)$, $I'_F(n)$ and $I'_T(n)$ represent the reconstructed BSHG, FSHG and TPEF images, respectively. We define the reconstruction loss for BSHG images (L_B) as the mean squared error between input BSHG images $I_B(\cdot)$ and the reconstructed BSHG images $I'_B(\cdot)$. Similar losses are defined for the FSHG images (L_F) and TPEF images (L_T). Therefore, the total loss for the FAE is

$$L = L_B + \alpha L_F + \beta L_T, \tag{2}$$

where α and β are the weightage of loss of the FSHG images and TPEF images relative to BSHG images. The weights for the different reconstruction terms are the hyperparameters of our model. These terms are chosen based on the validation performance. Our FAE is trained by minimizing L. The latent layer representations from the trained FAE is fed to the classification module.

2.2 The Classification Module (CM)

The CM is composed of a convolutional neural network-based classifier [14]. Using the latent space representations of the RoIs obtained from the IFM, the classifier undergoes training. This training process involves minimizing binary cross-entropy loss [15], which is designed for the task of two-class (FA and FTC) classification.

2.3 Inference

During the inference phase, a stack of BSHG, FSHG, and TPEF images corresponding to a test RoI is applied into the trained FAE. The FAE then generates a latent space representation from these images. This latent space representation is subsequently passed to the CM, where it is utilized to determine the final class labels for the given RoI.

2.4 Implementation Details

In our experimental setup, we conduct hyperparameter tuning for the FAE on a validation dataset comprising approximately 10% of the total data. The validation set is used to evaluate the performance of each combination. Through this process, we explored a range of hyperparameters and identified the set that achieved the highest mean AUROC score.

The proposed FAE has 25 convolutional layers, each followed by a GeLU activation function. Out of these, 12 are encoding layers, a latent layer and the other 12 are part of the decoder. The proposed architecture has 3 × 3 kernels in every convolutional layer. The decoder contains some common layers after the latent layer and gets branched out after a specific decoding layer (after the 6^{th} decoding layer in the experimental setup) to retain information about individual modalities as well as to have the fused information propagated further through these images. We use Max Pool layers of 2 × 2 dimensions after the first, and the sixth layers. In addition to these layers, we have two batch norm layers, one in the encoder and the other in the decoder components. The size of the feature maps after every convolutional layer starting from the first layer of encoder to the output of decoder are shown in Table 1.

The CM utilizes an EfficientNet-B4 architecture pretrained on ImageNet [16]. The FAE is trained for 30 epochs, while the classifier undergoes training for 50 epochs. We set the batch size for FAE input as 1 and 32 for the classifier. Both the FAE and the CNN-based classifier employ the Adam optimizer [17] with a learning rate of 0.0001. We use sigmoid activation function at the last classification layer. For a stack of input RoIs, we get class probabilities for the FA and FTC classes. The class with the highest probability score is considered to be the class of the input RoIs. For the ablation study, we maintain the same hyperparameter configurations, but find that utilizing LeakyReLU activation function yields better results for the fusion of BSHG with FSHG. Consequently, we adjust the learning rate for both the IFM and CM to 0.001. We tune the

Layer	Batch Size	Output Channels	Height	Width
1	1	12	502	502
2	1	14	500	500
3	1	16	248	248
4	1	20	246	246
5	1	20	244	244
6	1	32	242	242
7	1	64	240	240
8	1	128	118	118
9	1	128	116	116
10	1	256	114	114
11	1	512	112	112
12	1	512	110	110
13	1	3	108	108
14	1	256	112	112
15	1	128	114	114
16	1	64	116	116
17	1	32	118	118
18	1	24	120	120
19	1	12	122	122
20	1	1	124	124
21	1	1	126	126
22	1	1	128	128
23	1	1	257	257
24	1	1	512	512
25	1	1	512	512

Table 1: Feature Map Sizes After Each Convolutional Layer

hyperparameters for individual classifiers using the validation data, identifying an optimal learning rate of 0.001 and a batch size of 16. We take α = 0.5 and β = 0.05 in the loss function of (2).

3 Experiments and Results

3.1 Dataset

Our dataset contains different RoIs for 28 distinguishable tissue sections. Each RoI has FSHG, BSHG and TPEF images, each of 512 × 512 dimensions. We have 115 RoIs on a total of 8 tissue sections and 181 RoIs on 20 other tissue sections in FA and FTC categories, respectively. Further, each BSHG and FSHG images folder has raw data with ten linear polarization images captured at 0° to 180° with an interval of 20°. Sample images of BSHG, FSHG and TPEF of a particular RoI are shown in Figure 2. Tissue samples were obtained after partial or total thyroidectomy and were prepared according to standard histology protocols. Thin tissue sections stained with H&E were reviewed by a senior pathologist

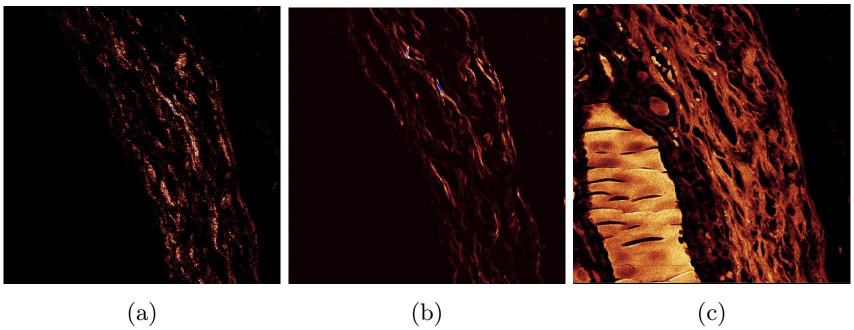


Fig. 2: Sample (a) BSHG, (b) FSHG and (c) TPEF images of linear polarization at 0° of a particular RoI

in order to place the diagnosis of either FA or FTC. Whole slide images were acquired from all the tissue sections. These virtual slides were annotated by the pathologist in order to highlight the nodule capsule surrounding the thyroid nodules, which were of interest in the present study. Using these annotated virtual slides as guidance, MPM images were collected on tissue slides around the thyroid nodule capsule. For each RoI, three images were collected: BSHG, FSHG and TPEF each having dimensions of 512×512 pixels. No postprocessing was applied to the images acquired by the nonlinear optical microscope.

3.2 Comparative Performances

To ensure that the train, validation, and test sets don't have any overlapping RoIs from the same tissue section, we adopt a dataset splitting strategy based on tissue sections for our experiments. This approach prevents any information leakage across splits. We use 70:10:20 split of tissue sections for training, validation, and testing. A random split of the dataset may result in an unequal number of training RoIs from FA and FTC. To deal with that, we perform splitting such that the number of training RoIs from the two classes are almost the same. Also, for each RoI, out of 10 polarization images of each modality, we average the pixel values of 9 polarization images to streamline the data representation (excluding the 180° image as it is the same as the 0° image) to form a 512×512-dimensional resultant image.

We compare the proposed method with various state-of-the-art classifiers including ResNet-18, ResNet-50, DenseNet-121 [18], EfficientNet-B0 (ENet-B0), and EfficientNet-B4 (ENet-B4) using individual modality images and evaluate their performances. All of these classifiers are pretrained on the ImageNet dataset. For each test data point, we compute the probabilities of belonging to classes FA and FTC. The data point is then assigned the class label with the higher probability. Based on these predicted class labels and the ground truth class labels, the values of the recall and precision are calculated. We calculate a recall and a precision value considering FA as the positive class. We do the same

Table 2: Various performance metrics obtained using different classifiers on the test dataset for individual modalities (BSHG, FSHG and TPEF images) are captured in the table. They are compared with the metric values derived from the proposed architecture where all three modalities are stacked together as input.

Modality	**Model**	**Class**	**Precision**	**Recall**	**F1 Score**	**AUROC**
BSHG	ResNet-18	FA	0.53 ± 0.5	0.10 ± 0.1	0.16 ± 0.2	0.55 ± 0.1
		FTC	0.72 ± 0.0	0.99 ± 0.0	0.83 ± 0.0	
	ResNet-50	FA	0.60 ± 0.2	0.25 ± 0.2	0.30 ± 0.1	0.57 ± 0.1
		FTC	0.74 ± 0.0	0.89 ± 0.1	0.81 ± 0.0	
	DenseNet-121	FA	0.63 ± 0.2	0.43 ± 0.2	0.46 ± 0.1	0.63 ± 0.0
		FTC	0.78 ± 0.0	0.84 ± 0.1	0.80 ± 0.1	
	EfficientNet-B0	FA	0.45 ± 0.1	0.47 ± 0.2	0.40 ± 0.1	0.58 ± 0.1
		FTC	0.76 ± 0.0	0.67 ± 0.2	0.70 ± 0.1	
	EfficientNet-B4	FA	0.42 ± 0.3	0.37 ± 0.4	0.23 ± 0.2	0.51 ± 0.0
		FTC	0.57 ± 0.3	0.66 ± 0.4	0.59 ± 0.3	
FSHG	ResNet-18	FA	0.41 ± 0.3	0.38 ± 0.4	0.26 ± 0.2	0.56 ± 0.1
		FTC	0.77 ± 0.1	0.73 ± 0.3	0.69 ± 0.2	
	ResNet-50	FA	0.71 ± 0.3	0.18 ± 0.1	0.26 ± 0.1	0.56 ± 0.0
		FTC	0.73 ± 0.0	0.94 ± 0.1	0.82 ± 0.0	
	DenseNet-121	FA	0.54 ± 0.2	0.56 ± 0.2	0.49 ± 0.1	0.63 ± 0.1
		FTC	0.80 ± 0.0	0.69 ± 0.3	0.70 ± 0.2	
	EfficientNet-B0	FA	0.59 ± 0.2	0.52 ± 0.2	0.47 ± 0.1	0.63 ± 0.1
		FTC	0.79 ± 0.0	0.75 ± 0.2	0.75 ± 0.1	
	EfficientNet-B4	FA	0.63 ± 0.2	0.40 ± 0.1	0.45 ± 0.1	0.63 ± 0.0
		FTC	0.77 ± 0.0	0.85 ± 0.1	0.80 ± 0.1	
TPEF	ResNet-18	FA	0.36 ± 0.0	0.38 ± 0.0	0.37 ± 0.0	0.54 ± 0.0
		FTC	0.72 ± 0.0	0.70 ± 0.1	0.71 ± 0.0	
	ResNet-50	FA	0.34 ± 0.1	0.39 ± 0.1	0.35 ± 0.1	0.53 ± 0.0
		FTC	0.72 ± 0.0	0.69 ± 0.1	0.70 ± 0.1	
	DenseNet-121	FA	0.63 ± 0.2	0.43 ± 0.2	0.46 ± 0.1	0.63 ± 0.0
		FTC	0.78 ± 0.0	0.84 ± 0.1	0.80 ± 0.1	
	EfficientNet-B0	FA	0.40 ± 0.0	0.63 ± 0.2	0.47 ± 0.1	0.61 ± 0.1
		FTC	0.83 ± 0.1	0.58 ± 0.2	0.64 ± 0.2	
	EfficientNet-B4	FA	0.39 ± 0.0	0.72 ± 0.3	0.50 ± 0.1	0.63 ± 0.1
		FTC	0.85 ± 0.1	0.54 ± 0.1	0.65 ± 0.1	
Proposed	EfficientNet-B4	FA	0.72 ± 0.2	0.47 ± 0.3	0.50 ± 0.2	0.66 ± 0.1
		FTC	0.74 ± 0.1	0.85 ± 0.1	0.78 ± 0.1	

considering FTC as the positive class. This enables us to get the class-wise recall and precision values. We run each method for ten times. For each run, the learnable parameters are initialized randomly. We take the best five runs out of ten runs to rule out the possibility of very poor initialization. The results for best five

Table 3: AUROC (mean ± sd) over ten runs for the proposed method. This table also shows the results using images of individual modalities and images from different combinations of two modalities.

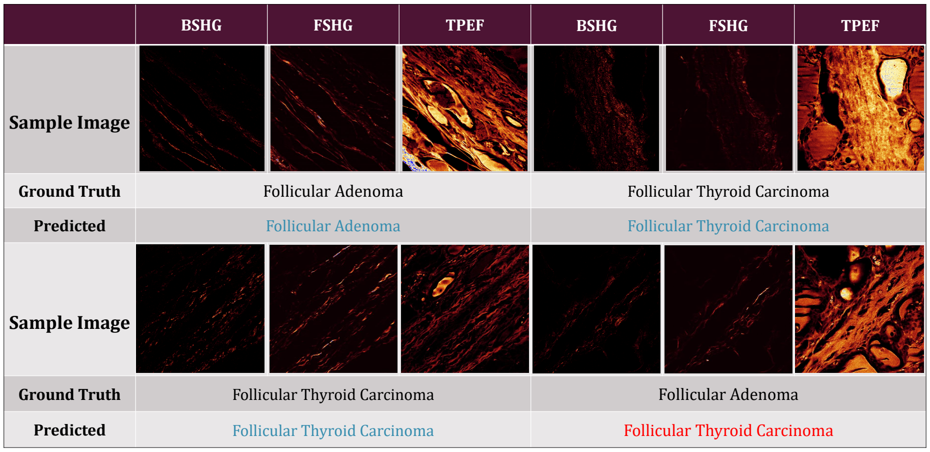
Modality	**Best Model**	**AUROC**
BSHG	DenseNet-121	0.57 ± 0.07
FSHG	EfficientNet-B4	0.56 ± 0.07
TPEF	DenseNet-121	0.58 ± 0.07
BSHG-FSHG	EfficientNet-B4	0.53 ± 0.04
BSHG-TPEF	EfficientNet-B4	0.53 ± 0.05
FSHG-TPEF	EfficientNet-B4	0.53 ± 0.04
Proposed	EfficientNet-B4	0.66 ± 0.10

runs (mean ± sd) are presented in Table 2. Notice that our method outperforms all competitors in terms of the mean AUROC. Subsequently, we look into the performances of competing methods when presented with stacked images from three modalities as input to the modified autoencoder. This is a type of early fusion [19]. This fusion helps to combine the information from the distinct sets of images and hence improves the capabilities of the model, thus leading to a more robust and efficient solution. Table 4 displays the metric values when computed on the images using early fusion technique and passing the latent layer feature maps to every competing classifier for comparison. It can be observed that our proposed method with ENet-B4 shows superior performances compared to its competitors. Results on sample images using the proposed method are presented in Figure 3.

Additionally, Table 3 shows the AUROC scores (mean ± sd) of the proposed method over ten runs. This table also contains the results over ten runs using images of individual modalities and a combination of images from two modalities. For these experiments, we take the classifiers that provided the best results when top five runs are considered.

3.3 Ablation Studies

We perform ablation studies to examine the impact of information fusion using images from three modalities. To this end, we perform experiments with different combinations of modalities to construct stacked images at the input



	BSHG	FSHG	TPEF	BSHG	FSHG	TPEF
Sample Image						
Ground Truth	Follicular Adenoma			Follicular Thyroid Carcinoma		
Predicted	Follicular Adenoma			Follicular Thyroid Carcinoma		
Sample Image						
Ground Truth	Follicular Thyroid Carcinoma			Follicular Adenoma		
Predicted	Follicular Thyroid Carcinoma			Follicular Thyroid Carcinoma		

Fig. 3: Sample BSHG, FSHG and TPEF images with their ground truth and predicted class labels using our method (blue: correct prediction, red: incorrect prediction).

of our method. Figure 4 illustrates the results obtained with various combinations of two modalities, namely, BSHG-FSHG, FSHG-TPEF, and BSHG-TPEF, alongside the results obtained using our proposed method. When utilizing two modalities, we incorporate two output heads for the ablation studies. It is evident from the results that using images from any two modalities leads to inferior performance compared to our proposed method. This shows the importance of information fusion using images from all three modalities in achieving optimal classification performance.

4 Discussion

As mentioned before, both BSHG and FSHG images primarily provide information on collagen or myosin. On the other hand, since TPEF is emitted by proteins in tissues which are autofluorescent, TPEF may provide more complementary information when combined with either BSHG or FSHG images.

Hence, our results shown in Table 3 indicate that the combination of BSHG and TPEF or FSHG and TPEF as input leads to better performance compared to using BSHG and FSHG alone. Specifically, the proposed method that integrates all three modalities—BSHG, FSHG, and TPEF—achieves the highest performance with a maximum mean AUROC score of 0.66.

Furthermore, the proposed method outperforms models that use any two modalities in combination. For example, the combination of BSHG and FSHG achieved a lower mean AUROC score compared to when TPEF was included. This combination hence enhances the capability of the model to distinguish between different classes of thyroid tumors.

Table 4: Performance metrics with different classifiers that take the latent space representation of our autoencoder as input.

Model	**Class**	**Precision**	**Recall**	**F1 Score**	**AUROC**
ResNet-18	FA	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.5 ± 0.00
	FTC	0.63 ± 0.00	1.00 ± 0.00	0.77 ± 0.00	
ResNet-50	FA	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.5 ± 0.00
	FTC	0.63 ± 0.00	1.00 ± 0.00	0.77 ± 0.00	
DenseNet-121	FA	0.27 ± 0.26	0.30 ± 0.38	0.24 ± 0.27	0.53 ± 0.06
	FTC	0.66 ± 0.05	0.77 ± 0.35	0.66 ± 0.22	
EfficientNet-B0	FA	0.28 ± 0.26	0.39 ± 0.38	0.33 ± 0.31	0.57 ± 0.08
	FTC	0.71 ± 0.09	0.75 ± 0.23	0.7 ± 0.07	
Proposed	FA	0.72 ± 0.19	0.47 ± 0.27	0.50 ± 0.22	0.66 ± 0.10
	FTC	0.74 ± 0.09	0.85 ± 0.12	0.78 ± 0.05	

Our ablation studies shown in Figure 4 also show the significance of using three modalities. This finding highlights the potential of multimodal approaches in medical imaging, where different imaging techniques can complement each other to provide a more comprehensive understanding of the tissue characteristics. The synergy between the network architecture and the new data modalities is a key factor driving the observed performance gains.

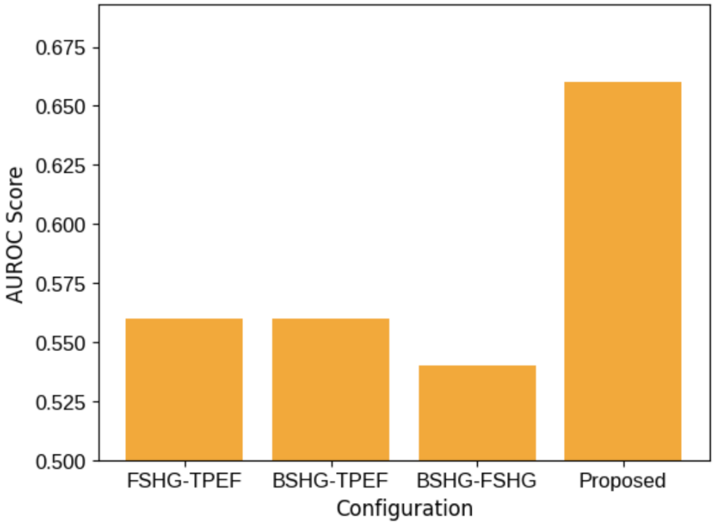


Fig. 4: Mean AUROC scores over five best runs using different combinations of modalities (FSHG with TPEF, BSHG with TPEF and BSHG with FSHG) alongside the proposed method (BSHG, FSHG, and TPEF as input).

5 Conclusion

We delve into the potential of information fusion from different types of MPM images for the differential diagnosis of thyroid tumors. To achieve this, we design a Fusion Autoencoder, aimed at integrating information from three distinct modalities. The latent space representation of the autoencoder is found to provide meaningful information through the fusion of MPM images of three different modalities. Rigorous experiments show that the proposed method can obtain a mean AUROC score of 0.66. However, the use of individual MPM images can achieve a maximum mean AUROC score of 0.63. This shows the impact of information fusion in our model. Furthermore, ablation studies show that the use of information fusion from images of any two modalities is significantly less effective compared to the proposed strategy of using images from three modalities. The present work is a proof-of-concept study to look into the utility of multiphoton microscopy images. This method of fusing information from images of multiple modalities and the use of a Fusion Autoencoder shows promising results for the diagnosis of thyroid tumors. In the future, we will explore the feasibility of integrating information from histopathological images with the information from MPM images to further enhance the accuracy of thyroid tumor diagnosis. We will also look into the possibility of utilizing larger datasets for our experiments. Moreover, we intend to extend the application of such information fusion strategies to incorporate non-image medical data, thereby broadening the scope of diagnostic capabilities.

References

1. L.Z. Li et al., *Two-photon autofluorescence imaging of fixed tissues: feasibility and potential values for biomedical applications*, *Oxygen Transport to Tissue XLI*, 375–381 (2020), *Springer.*
2. Y.W. Jun et al., *Addressing the autofluorescence issue in deep tissue imaging by two-photon microscopy: the significance of far-red emitting dyes*, *Chemical science*, **8**(11), 7696–7704 (2017), *Royal Society of Chemistry.*
3. P.J. Campagnola and C.Y. Dong, *Second harmonic generation microscopy: Principles and applications to disease diagnosis*, **5**, 13–26 (2011), *Laser & Photonics Reviews*, https://doi.org/10.1002/lpor.200910024.
4. S.J. Mulligan, B.D. Garrod, and M.A. Leake, *Two-photon fluorescence microscopy: Basic principles, advantages, and risks*, **86**, 105–129 (2007), *Methods in Cell Biology*, https://doi.org/10.1016/S0091-679X(06)86010-3.
5. S.R. Anton et al., *Automated detection of corneal edema with deep learning-assisted second harmonic generation microscopy*, *IEEE Journal of Selected Topics in Quantum Electronics*, **29**(6: Photonic Signal Processing), 1–10 (2023), https://doi.org/10.1109/JSTQE.2023.3149295.
6. M. Volante and M. Papotti, *A practical diagnostic approach to solid/trabecular nodules in the thyroid*, *Endocrine pathology*, **19**, 75–81 (2008), *Springer.*
7. D.S. Cooper et al., *Revised American Thyroid Association management guidelines for patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association (ATA) guidelines taskforce on thyroid nodules and differentiated thyroid cancer*, *Thyroid*, **19**(11), 1167–1214 (2009), *Mary Ann Liebert, Inc.*.

8. D. Sarwinda et al., *Deep Learning in Image Classification using Residual Network (ResNet) Variants for Detection of Colorectal Cancer*, *Procedia Computer Science*, **179**, 423–431 (2021), https://doi.org/https://doi.org/10.1016/j.procs.2021.01.025.
9. C. Szegedy et al., *Rethinking the inception architecture for computer vision*, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826 (2016), *IEEE*.
10. G. Wang et al., *Automated Ovarian Cancer Identification Using End-to-End Deep Learning and Second Harmonic Generation Imaging*, *IEEE Journal of Selected Topics in Quantum Electronics*, **29**(4: Biophotonics), 1-9 (2023), https://doi.org/10.1109/JSTQE.2022.3228567.
11. L.G. Eftimie et al., *Differential diagnosis of thyroid nodule capsules using random forest guided selection of image features*, *Scientific Reports*, **12**, 25788 (2022), https://doi.org/10.1038/s41598-022-25788-w.
12. R. Hristu et al., *PSHG-TISS: A collection of polarization-resolved second harmonic generation microscopy images of fixed tissues*, *Scientific Data*, **9**(1), 376 (2022), https://doi.org/10.1038/s41597-022-01201-3.
13. S. Petscharnig, M. Lux, and S. Chatzichristofis, *Dimensionality Reduction for Image Features using Deep Learning and Autoencoders*, *2017 IEEE International Conference on Systems, Man and Cybernetics*, 1-6 (2017), https://doi.org/10.1145/3095713.3095737.
14. D.R. Sarvamangala and R.V. Kulkarni, *Convolutional neural networks in medical image understanding: a survey*, *Evolutionary intelligence*, **15**(1), 1–22 (2022), https://doi.org/10.1007/s12065-021-00426-4.
15. U. Ruby and V. Yendapalli, *Binary cross entropy with deep learning technique for image classification*, *Int. J. Adv. Trends Comput. Sci. Eng*, **9**(10) (2020).
16. J. Deng et al., *Imagenet: A large-scale hierarchical image database*, *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (2009), *IEEE*.
17. S. Bock and M. Weiß, *A proof of local convergence for the Adam optimizer*, *2019 international joint conference on neural networks (IJCNN)*, 1–8 (2019), *IEEE*.
18. G. Huang, Z. Liu, and K.Q. Weinberger, *Densely Connected Convolutional Networks*, *CoRR*, abs/1608.06993 (2016).
19. L.M. Pereira, A. Salazar, and L. Vergara, *A Comparative Analysis of Early and Late Fusion for the Multimodal Two-Class Problem*, *IEEE Access*, **11**, 84283-84300 (2023), https://doi.org/10.1109/ACCESS.2023.3296098.
20. X. Chen, O. Nadiarynkh, S. Plotnikov, and P.J. Campagnola, *Second harmonic generation microscopy for quantitative analysis of collagen fibrillar structure*, **7**(4), 654–669 (2012), *Nature protocols*.
21. K. Tilbury, J. Hocker, B.L. Wen, N. Sandbo, V. Singh, and P.J. Campagnola, *Second harmonic generation microscopy analysis of extracellular matrix changes in human idiopathic pulmonary fibrosis*, **19**(8), 086014–086014 (2014), *Journal of biomedical optics*.
22. J. Adur, V.B. Pelegati, A.A. de Thomaz, M.O. Baratti, L.A.L.A. Andrade, H.F. Carvalho, F. Bottcher-Luiz, and C.L. Cesar, *Second harmonic generation microscopy as a powerful diagnostic imaging modality for human ovarian cancer*, **7**(1-2), 37–48 (2014), *Journal of biophotonics*.
23. A. Pak et al., *Comparative analysis of deep learning methods of detection of diabetic retinopathy*, *Cogent Engineering*, **7**(1), 1805144 (2020), https://doi.org/10.1080/23311916.2020.1805144.
24. M. Tan and Q.V. Le, *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*, *CoRR*, abs/1905.11946 (2019), http://arxiv.org/abs/1905.11946.

25. A.M. Pena et al., *In vivo multiphoton multiparametric 3D quantification of human skin aging on forearm and face*, *Scientific Reports*, **12**(1), 14863 (2022), https://doi.org/10.1038/s41598-022-07325-5.
26. S.G. Stanciu et al., *Toward next-generation endoscopes integrating biomimetic video systems, nonlinear optical microscopy, and deep learning*, *Biophysics Reviews*, **4**(2) (2023), https://doi.org/10.1007/s41048-023-00222-1.
27. H. Gunes and M. Piccardi, *Affect recognition from face and body: early fusion vs. late fusion*, *2005 IEEE International Conference on Systems, Man and Cybernetics*, 3437-3443 Vol. 4 (2005), https://doi.org/10.1109/ICSMC.2005.1571679.