2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-1179-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/EMBC46164.2021.9629739

End-to-End Neural Network for Feature Extraction and Cancer Diagnosis of In Vivo Fluorescence Lifetime Images of Oral Lesions*

Kayla Caughlin[1], Elvis Duran-Sierra[3], Shuna Cheng[3], Rodrigo Cuenca[2], Beena Ahmed[4], Jim Ji[5], Vladislav V. Yakovlev[3], Mathias Martinez[6], Moustafa Al-Khalil[6], Hussain Al-Enazi[7], Javier A. Jo[2], and Carlos Busso[1]

***Abstract*— In contrast to previous studies that focused on classical machine learning algorithms and hand-crafted features, we present an end-to-end neural network classification method able to accommodate lesion heterogeneity for improved oral cancer diagnosis using *multispectral autofluorescence lifetime imaging* (maFLIM) endoscopy. Our method uses an autoencoder framework jointly trained with a classifier designed to handle overfitting problems with reduced databases, which is often the case in healthcare applications. The autoencoder guides the feature extraction process through the reconstruction loss and enables the potential use of unsupervised data for domain adaptation and improved generalization. The classifier ensures the features extracted are task-specific, providing discriminative information for the classification task. The data-driven feature extraction method automatically generates task-specific features directly from fluorescence decays, eliminating the need for iterative signal reconstruction. We validate our proposed neural network method against *support vector machine* (SVM) baselines, with our method showing a 6.5%-8.3% increase in sensitivity. Our results show that neural networks that implement data-driven feature extraction provide superior results and enable the capacity needed to target specific issues, such as inter-patient variability and the heterogeneity of oral lesions.**

***Clinical relevance*— We improve standard classification algorithms for in vivo diagnosis of oral cancer lesions from maFLIm for clinical use in cancer screening, reducing unnecessary biopsies and facilitating early detection of oral cancer.**

I. INTRODUCTION

With increased research into targeted therapies, cancer has become increasingly viewed as a highly diverse problem, with a complex *tumor microenvironment* (TME) that leads to high inter- and intra- tumor heterogeneity [1]. While recent research has improved understanding of the TME and led to increases in targeted therapies, many aspects of cancer remain undefined [2]. Even a specific subtype of cancer, such as oral cancer, can be further separated into finer groupings based on differences in tissue composition of various oral tissues, risk factors (e.g., *human papillomavirus* (HPV)), and race [3]–[5]. Considering the complexity of oral cancer, automated classification algorithms could aid in early detection by providing care to patients without access to an expert clinician. However, lesion heterogeneity combined with the small size of datasets lead to overfitting that has been a key challenge limiting the use of automated classifiers for oral cancer diagnosis.

Recent efforts to automatically classify oral lesions have used *multispectral autofluorescence lifetime imaging* (maFLIM) endoscopy to assess molecular characteristics of tissue by measuring autofluorescence from cells associated with the structural and metabolic state of the lesion. To date, most maFLIM cancer classification or margin delineation for in vivo images use hand-derived features such as lifetime and intensity, in combination with classical machine learning methods such as *support vector machines* (SVM), *linear discriminant analysis* (LDA), or *quadratic discriminant analysis* (QDA) rather than neural networks in an effort to prevent overfitting [6]–[8]. Further, feature generation often uses a deconvolved signal reconstructed from an iterative fitting method [6], [7]. While these methods show promising preliminary results, high inter-patient variability and lesion heterogeneity may not be adequately modeled using signal estimation and pre-defined features. The use of data-driven feature extraction in a neural network classification architecture may provide a flexible approach that accommodates lesion heterogeneity if the tendency for overfitting is reduced. Therefore, we propose a joint autoencoder and classifier neural network to provide fit-free feature extraction and classification using raw fluorescence decays. In contrast to other fit-free methods such as Laguerre deconvolution [9], our network uses the classification label to ensure the generation of task-specific features. At the same time, the autoencoder structure reduces overfitting and retains the ability to use unlabeled data by optimizing the reconstruction loss.

In contrast to hand-crafted features, our end-to-end neural network uses a data-driven feature extraction method to generate discriminative features for cancer diagnosis. We successfully regularize the neural network to reduce overfitting, showing a 6.5%-8.3% increase in sensitivity compared to a baseline SVM trained with typical hand-crafted features.

II. BACKGROUND

In imaging of oral lesions, maFLIM endoscopy measures the autofluorescence in three spectral bands correspond-

*This work was supported by NIH, grant R01:5R01CA218739-04

[1]Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson, TX, USA
[2]School of Electrical and Computer Engineering, University of Oklahoma, Norman, OK, USA
[3]Department of Biomedical Engineering, Texas A&M University, College Station, TX, USA
[4]School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, Australia
[5]Department of Electrical and Computer Engineering, Texas A&M University at Qatar, Doha, Qatar
[6]Department of Cranio-Maxillofacial Surgery, Hamad Medical Corporation, Doha, Qatar
[7]Department of Otorhinolaryngology Head and Neck Surgery, Hamad Medical Corporation, Doha, Qatar

ing to collagen, reduced *nicotinamide adenine dinucleotide* (NADH), and *flavin adenine dinucleotide* (FAD), respectively. Following image acquisition, the measured fluorescence decays are often deconvolved by an iterative method using a fitting model, such as a bi-exponential decay [6], [8]. The deconvolved signal is then reconstructed from the fitting parameters and hand-crafted features are calculated. The standard features previously used consist of the parameters from the fit and variations on intensity and lifetime [6], [8]. The signal intensity (I_k) captures spectral characteristics, and is defined in Equation 1,

$$I_k = \int h_k(t)\ dt \tag{1}$$

where $h_k(t)$ is the deconvolved fluorescence decay signal at time t, and k specifies the channel. Equation 2 defines the signal lifetime (τ_k), which captures temporal characteristics.

$$\tau_k = \frac{\int t h_k(t)\ dt}{\int h_k(t)\ dt} \tag{2}$$

Two potential issues exist with this approach. First, the features are limited by signal reconstruction using a fitting model. For example, in a bi-exponential model, each channel in a single deconvolved fluorescence decay can be completely reconstructed using four parameters. A fitting model severely limits novel feature generation methods, as the signal is completely defined by the parameters from the fit. Fereidouni et al. [10] explored the Laguerre deconvolution and phasors as non-parametric deconvolution methods, but did not use these methods in classification. In skin cancer diagnosis using maFLIM endoscopy, Vasanthakumari et al. [11] showed promising performance with LDA and QDA classifiers trained with features calculated from the resulting phasors after frequency domain deconvolution. However, none of these methods incorporates the classification labels in the feature extraction process, so they cannot ensure that the features are task-specific. Second, an imposed fit and hand-crafted feature set may not adequately describe the heterogeneity in oral lesions. For example, Marsden et al. [8] found that a random forest classifier performed better on tonsil lesions than on tongue lesions although images from both locations were included in the training set. Perhaps due to the small size of datasets, few studies have documented any other approach for classification of oral lesions from in vivo maFLIM endoscopy images. For oral cancer classification, Jo et al. [6] deconvolved the signal using a bi-exponential model, then used the deconvolved signal to generate lifetime and intensity features. Marsden et al. [8] experimented with a 1-D *convolutional neural network* (CNN) for margin detection in oral lesions. However, they used a deconvolved signal reconstructed from Laguerre coefficients and found that neural network performance was inferior to random forest classification. Further, they noted issues with overfitting that required pre-training the network on synthetic examples of fluorescence lifetime estimation to obtain any reasonable result. CNNs have also been used for classification of oral photographic images using a network pre-trained on ImageNet [12]. However, this approach is not applicable to maFLIM endoscopy images trained on individual pixels. Like Fereidouni et al. [10] and Vasanthakumari et al. [11], our approach does not rely on a fitting model or iterative deconvolution. However, our end-to-end, joint neural network model ensures task-specific features while reducing overfitting without pre-training.

TABLE I
LOCATION OF ORAL LESIONS

Location	Benign	Dysplasia	SCC
Mucosa	10	3	9
Floor of Mouth	2	0	1
Gingiva	0	2	3
Lip	10	0	2
Mandible	0	0	1
Palate	1	0	0
Maxilla	0	0	1
Retromolar	1	0	0
Tongue	9	0	12
Total	33	5	29

III. METHODS

A. Data

The procedure for data collection included imaging, biopsy, and histopathological diagnosis (approved by the Institutional Review Boards at Hamad Medical Corporation in Doha, Qatar). All oral epithelial lesions were clinically identified as potentially cancerous or precancerous before imaging. A total of 67 in vivo images of oral lesions were collected using the maFLIM endoscope described by Cheng et al. [13]. The endoscope covers a circular area approximately 11 mm in diameter. A reference image of healthy oral tissue for each subject was also collected. Only the lesion images are used for this study. For classification purposes, we include cases of dysplasia with *squamous cell carcinoma* (SCC) lesions, formulating the task as a binary problem: benign vs. malignant (dysplasia + SCC) oral lesions.

The fluorescence decay from the lesion was measured at each pixel, with an image size of 160x160 pixels. Preprocessing at the image level consisted of signal inversion and median filtering. At the pixel-level, the measured decay was chopped or zero-padded to a length of 300 samples per channel. We use masks to exclude pixels with poor SNR or saturated pixels, which would affect the classifiers. As the gain used during acquisition was not constant, the concatenated decays were normalized to sum to 100. We directly use the decays without the deconvolution step. However, our baseline with SVM uses a bi-exponential model and iterative deconvolution method to estimate the deconvolved signal before feature calculation. Following deconvolution, the signal is given by the bi-exponential in Equation 3,

$$h_k = \alpha_{fast,k} e^{-t/\tau_{fast,k}} + \alpha_{slow,k} e^{-t/\tau_{slow,k}} \tag{3}$$

where h_k is the deconvolved signal, α is the weight of the exponential, τ is the rate of decay, and k denotes the channel. Table I shows the locations and classes of the 67 lesions.

B. Machine Learning Framework

We propose a joint neural network model that uses a shared encoder and separate paths to generate the signal

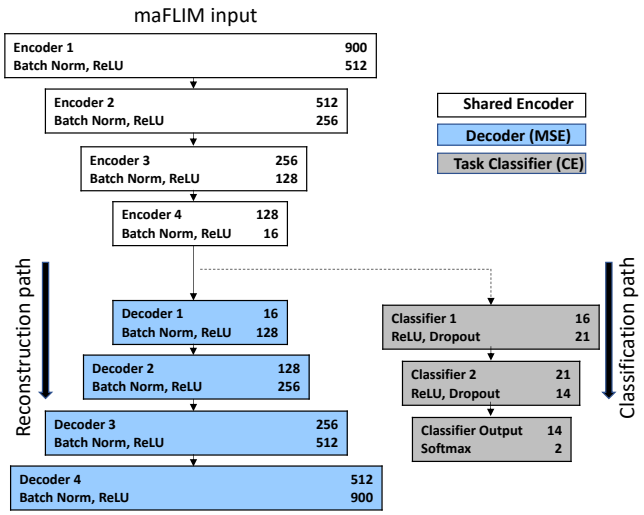

Fig. 1. Joint Neural Network Structure. The reconstruction path is trained with the *mean square error* (MSE) loss and the classification path is trained with the *cross entropy* (CE) loss.

reconstruction and the classification output. Figure 1 shows the proposed architecture, which carefully avoids overfitting, since deep learning approaches often rely on large databases that are not available in medical applications. The joint structure eliminates iterative signal reconstruction, generates task-specific features from the data, and retains the ability to use unlabeled data for domain adaptation and improved generalization in the future. Additionally, the reconstruction path serves as a form of regularization to reduce overfitting. The shared encoder has four contracting layers, reducing the dimension of the input signal from 900 to 16, creating a low dimension bottleneck embedding. The decoder mirrors the encoder, reconstructing the input signal by optimizing the *mean squared error* (MSE) between the input and reconstructed signal. The classifier is attached to the bottleneck embedding layer. It has two layers followed by a softmax layer to produce the classification output. The classifier is trained using the categorical *cross entropy* (CE) loss. We weight the reconstruction loss by a factor of two in comparison to the classification loss to cause the model to prioritize the reconstruction loss and reduce overfitting to the training classification labels. All layers are fully connected with *rectified linear unit* (ReLU) activations, except the final softmax layer in the classifier. All layers in the autoencoder use batch normalization. All layers in the classifier use dropout with a rate of $p = 0.5$ to regularize the network and avoid overfitting. The network is implemented using Keras with TensorFlow [14]. The architecture separately processes each of the $\sim$1.7M pixels in the 67 images, creating a prediction per pixel.

The evaluation uses a ten-fold cross validation approach creating train, development and test sets. One fold is used for testing the models. Then, we use two folds for the development set. The remaining seven folds are used in the train set. The partitions are selected on an image basis by randomly partitioning the images. Therefore, all the pixels from an image are included either in the train, development or test set. The sets did not have any overlap. Once randomly split, we used the same train, development, and test sets for our method and the baselines (Sec. III-C) to have fair comparisons. Although there is only one more malignant lesions than benign lesions (Table I), each image contains a different number of valid pixels, resulting in a different level of class imbalance based on the specific split of the data. To reduce the classifier bias from the training distribution, sample weights are generated using the *class weight* function from the *sci-kit learn* toolkit [15]. The model is optimized using Adam with a learning rate of 1e-5 and trained for 80 epochs with an early stopping on the development set with a patience of 10 epochs. The best model on the development set is used on the test set to increase generalization.

TABLE II
CLASSIFICATION PERFORMANCE - MEAN (SD)

	Sens.	Spec.	Prec.	F1	Acc.
SVM SFS	81.0(5.3)	67.3(4.8)	73.8(3.5)	74.9(3.6)	74.0(3.2)
SVM L1	79.2(5.1)	**73.3**(5.0)	**78.1**(4.9)	75.9(4.4)	76.4(3.4)
Neural Net	**87.5**(3.8)	67.6(6.5)	76.3(3.5)	**79.8**(2.6)	**77.6**(3.1)

C. Baseline

As a baseline, we use SVM classifiers with a total of 21 standard features used in other studies. Each channel contributes 1 intensity, 1 lifetime, and 3 features from the bi-exponential decay model (15 total channel-specific features). The weights given by the α's in Equation 3 sum to one, so only one weight per channel is used in the feature set. In addition to the channel-specific features, we include 3 ratios of intensities and 3 features from the sum of two intensities divided by the remaining intensity.

Although SVM is commonly used for classification of maFLIM endoscopy images, several variations exist. Marsden et al. [8] used SVM with a radial bias function kernel for margin delineation of in vivo maFLIM endoscopy images of oral lesions. Chen et al. [16] used a linear kernel SVM for classification of ex-vivo maFLIM endoscopy images of skin lesions. In consideration of the size of the dataset, we use a SVM with a linear kernel to minimize overfitting, implemented with Linear SVC using the *sci-kit learn* toolkit [15]. We set the value of the regularization parameter C to 1.0 (default) and use the *sequential forward selection* (SFS) from the Python library *MLxtend* [17] to find the optimal number of features based on the development performance.

To assess the effect of limited validation data on the model performance, we use the same linear kernel SVM as above, but replace SFS with L1 regularization to generate a sparse solution and implicitly perform feature selection. We optimized the regularization parameter, C, for each fold using the development performance ($C \in \{1e-5, 1e-4, 1e-3\}$). Using L1 regularization eliminates the need for feature selection, reducing the impact of a small development set and potentially improving generalization.

IV. RESULTS

Our data contains lesions from locations with only a single example (Table I). It also includes three examples of dysplasia. A random split may result in an unbalanced

test set, containing lesions from locations and classes that were not represented in the train set. Additionally, random initialization of the neural network can produce different results on separate runs. To give a fair assessment of each method, we run 10 trials with a different random ten-fold cross validation split for each trial. All classifiers are trained on a pixel level. A pixel receives the binary label of the image. However, the benign or malignant prediction for an image in the test set is computed by using the majority rule across the pixel predictions in an image. We assess performance with sensitivity and specificity rates.

Table II shows the mean and standard deviation for classification performance over ten trials. Our joint neural net model produced the best results, with a 6.5% increase in sensitivity over the SVM model with SFS and a 8.3% increase in sensitivity over the SVM model with L1 regularization. The specificity of our model was slightly higher than SVM with SFS. Considering the average of sensitivity and specificity, the neural network outperforms both SVM models. This result is promising as the features are directly extracted from the data, without relying on hand-crafted features used for the SVM models. As we collect more data, we expect the results to be even better. The SVM model with L1 regularization has the least gap between sensitivity and specificity which gives the highest specificity but lowest sensitivity of the three models. We attribute this more balanced result to the reduced reliance on the development set, which can fail to represent the testing data due to its small sample of a highly diverse dataset. In contrast to L2 regularization, L1 regularization typically does not select groups of correlated features [18]. Group selection is desirable when correlated features become discriminative when used in combination. However, oral lesions from different classes may not exhibit the same feature combinations, causing overfitting to specific subgroups of lesions in the train and development sets.

In the future, our neural network model can incorporate unlabeled data to expand both the training and development datasets without requiring labels using the reconstruction loss. The model will preserve complex relationships within subgroups of oral lesions, while preventing overfitting. Notably, our model allows the use of more data without requiring painful resection and time-consuming biopsy.

V. CONCLUSIONS

We proposed a new deep neural network method for feature extraction and classification of oral lesions. Our approach used joint training of a fully-connected feature extractor and classifier to diagnose oral cancer with up to 8.3% increase in sensitivity compared to baseline SVM models with standard features. We provide a second comparison using SVM with L1 regularization to reduce reliance on a small, limited development set for feature selection. The more balanced result indicates that our method may further improve by including additional, unlabeled data in training and development sets. In future work, our joint neural network can be improved using unlabeled images of a large number of oral lesions in the autoencoder path of the network. We can incorporate images of the contralateral side of each lesion by expanding the network to accept image pairs and contrast lesion and normal tissue from a single patient. In addition, we would like to assess the ability of our model to use unlabeled data to create a single classifier for images collected from multiple centers and from slightly different domains (e.g., skin lesions).

REFERENCES

[1] G. Stanta and S. Bonin, "Overview on clinical relevance of intra-tumor heterogeneity," *Frontiers in Medicine*, vol. 5, p. 85, April 2018.

[2] T. Haider, K. Sandha, V. Soni, and P. Gupta, "Recent advances in tumor microenvironment associated therapeutic strategies and evaluation models," *Materials Science and Engineering: C*, vol. 116, p. 111229, November 2020.

[3] G. Sarode, S. Sarode, J. Tupkari, and S. Patil, "Is oral squamous cell carcinoma unique in terms of intra-and inter-tumoral heterogeneity?" *Translational Research in Oral Oncology*, vol. 2, pp. 1–6, April 2017.

[4] M. Canning, G. Guo, M. Yu, C. Myint, M. Groves, J. Byrd, and Y. Cui, "Heterogeneity of the head and neck squamous cell carcinoma immune landscape and its impact on immunotherapy," *Frontiers in Cell and Developmental Biology*, vol. 7, p. 52, April 2019.

[5] R. Siegel, K. M. Kimberly, H. Fuchs, and A. Jemal, "Cancer statistics, 2021," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 1, pp. 7–33, January-February 2021.

[6] J. Jo, S. Cheng, R. Cuenca-Martinez, E. Duran-Sierra, B. Malik, B. Ahmed, K. Maitland, Y.-S. Cheng, J. Wright, and T. Reese, "Endogenous fluorescence lifetime imaging (FLIM) endoscopy for early detection of oral cancer and dysplasia," in *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2018)*, Honolulu, HI, USA, July 2018, pp. 3009–3012.

[7] R. Romano, R. Teixeira Rosa, A. Salvio, J. Jo, and C. Kurachi, "Multispectral autofluorescence dermoscope for skin lesion assessment," *Photodiagnosis Photodyn. Ther.*, vol. 30, p. 101704, June 2020.

[8] M. Marsden *et al.*, "Intraoperative margin assessment in oral and oropharyngeal cancer using label-free fluorescence lifetime imaging and machine learning," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 3, pp. 857–868, March 2021.

[9] J. Jo, Q. Fang, and L. Marcu, "Ultrafast method for the analysis of fluorescence lifetime imaging microscopy data based on the Laguerre expansion technique," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 11, no. 4, pp. 835–845, July-Aug. 2005.

[10] F. Fereidouni, D. Gorpas, D. Ma, H. Fatakdawala, and L. Marcu, "Rapid fluorescence lifetime estimation with modified phasor approach and Laguerre deconvolution: a comparative study," *Methods and applications in fluorescence*, vol. 5, no. 3, p. 035003, September 2017.

[11] P. Vasanthakumari, R. Romano, R. Rosa, A. Salvio, V. Yakovlev, C. Kurachi, and J. Jo, "Classification of skin-cancer lesions based on fluorescence lifetime imaging," in *SPIE Medical Imaging, 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging*, vol. 11317, Houston, TX, USA, February 2020.

[12] S. Camalan *et al.*, "Convolutional neural network-based clinical predictors of oral dysplasia: Class activation map analysis of deep learning results," *Cancers*, vol. 13, no. 6, p. 1291, March 2021.

[13] S. Cheng, R. Cuenca, B. Liu, B. Malik, J. Jabbour, K. Maitland, J. Wright, Y.-S. Cheng, and J. Jo, "Handheld multispectral fluorescence lifetime imaging system for in vivo applications," *Biomedical Optics Express*, vol. 5, no. 3, pp. 921–931, March 2014.

[14] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Symposium on Operating Systems Design and Implementation (OSDI 2016)*, Savannah, GA, USA, November 2016, pp. 265–283.

[15] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

[16] B. Chen, Y. Lu, W. Pan, J. Xiong, Z. Yang, W. Yan, L. Liu, and J. Qu, "Support vector machine classification of nonmelanoma skin lesions based on fluorescence lifetime imaging microscopy," *Analytical chemistry*, vol. 91, no. 16, pp. 10 640–10 647, August 2019.

[17] S. Raschka, "MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack," *The Journal of Open Source Software*, vol. 3, no. 24, p. 638, April 2018.

[18] L. Wang, J. Zhu, and H. Zou, "The doubly regularized support vector machine," *Statistica Sinica*, vol. 16, no. 2, pp. 589–615, April 2016.