






Pixel-level classification of pigmented skin cancer lesions using multispectral autofluorescence lifetime dermoscopy imaging

PRIYANKA VASANTHAKUMARI,¹ RENAN A. ROMANO,²  RAMON G. T. ROSA,² ANA G. SALVIO,³ VLADISLAV YAKOVLEV,¹  CRISTINA KURACHI,²  JASON M. HIRSHBURG,⁴ AND JAVIER A. JO^{5,*}

¹Texas A&M University, Department of Biomedical Engineering, College Station, TX, USA

²University of São Paulo, São Carlos Institute of Physics, São Paulo, Brazil

³Skin Department of Amaral Carvalho Hospital, São Paulo, Brazil

⁴University of Oklahoma Health Science Center, Department of Dermatology, Oklahoma City, OK, USA

⁵University of Oklahoma, School of Electrical and Computer Engineering, Norman, OK, USA

*javierjo@ou.edu

Abstract: There is no clinical tool available to primary care physicians or dermatologists that could provide objective identification of suspicious skin cancer lesions. Multispectral autofluorescence lifetime imaging (maFLIM) dermoscopy enables label-free biochemical and metabolic imaging of skin lesions. This study investigated the use of pixel-level maFLIM dermoscopy features for objective discrimination of malignant from visually similar benign pigmented skin lesions. Clinical maFLIM dermoscopy images were acquired from 60 pigmented skin lesions before undergoing a biopsy examination. Random forest and deep neural networks classification models were explored, as they do not require explicit feature selection. Feature pools with either spectral intensity or bi-exponential maFLIM features, and a combined feature pool, were independently evaluated with each classification model. A rigorous cross-validation strategy tailored for small-size datasets was adopted to estimate classification performance. Time-resolved bi-exponential autofluorescence features were found to be critical for accurate detection of malignant pigmented skin lesions. The deep neural network model produced the best lesion-level classification, with sensitivity and specificity of $76.84\% \pm 12.49\%$ and $78.29\% \pm 5.50\%$, respectively, while the random forest classifier produced sensitivity and specificity of $74.73\% \pm 14.66\%$ and $76.83\% \pm 9.58\%$, respectively. Results from this study indicate that machine-learning driven maFLIM dermoscopy has the potential to assist doctors with identifying patients in real need of biopsy examination, thus facilitating early detection while reducing the rate of unnecessary biopsies.

© 2024 Optica Publishing Group under the terms of the [Optica Open Access Publishing Agreement](#)

1. Introduction

Skin cancer is the most diagnosed type of cancer among fair-skinned populations [1]. The two most common types of skin cancer are malignant melanoma and non-melanoma skin cancer (NMSC). Basal cell carcinoma (BCC) and squamous cell carcinoma (SCC) are the two major NMSC categories. As of 2019, the incidence rates per 100,000 persons for melanoma, SCC and BCC were 17, 262, and 525 respectively [2]. Although NMSC represents the majority of skin cancer cases, most deaths due to skin cancer are accountable to malignant melanoma. In 2023, the number of new cases of melanoma was estimated to be 97,610 in the US [3]. Almost 78% of melanoma cases are diagnosed at the early stages with no metastases, however, 9% and 5% of the cases are diagnosed with regional and distant metastases respectively. The five-year survival rate of patients diagnosed with melanoma at early stages is about 100%; unfortunately, when diagnosed at advanced stages with regional and distant metastases, the five-year survival

rate dramatically decreases to 71% and 32% respectively [3]. The gold standard for skin lesion diagnosis is visual inspection (often assisted with a dermoscope) followed by biopsy resection and histopathological evaluation as recommended by the dermatologist. The main challenge with this method, however, is the inability to discriminate visually similar benign and malignant lesions, e.g., malignant melanoma lesions are very similar to benign pigmented seborrheic keratosis (pSK) lesions.

Clinical diagnosis of skin melanoma with the unaided eye is ~60% accurate [4]. Although dermoscopy is a widely used prescreening technique, the diagnosis result is highly subjective. Only those lesions which appear suspicious to the dermatologist are referred for biopsy resection and histopathological analysis. Several metrics based on visual and morphological characteristics of the lesion, such as the Menzies method, CASH metric, ABCD rule, and seven-point checklist, are used to determine if a lesion is suspicious or not [5,6]. It has been reported that melanoma diagnosis accuracy by non-experts using a dermoscope is similar to the clinical diagnosis accuracy by experts without a dermoscope [4]. Ahadi et. al [7], assessed 4,123 skin cancers samples collected over 3 consecutive years and obtained a false negative rate of 9.53% and a false positive rate of 17.14%. This means that 9.53% of malignant skin lesions were missed and did not undergo biopsy, while 17.14% of benign skin lesions were unnecessarily biopsied. Therefore, objective image-guided techniques that accurately discriminate clinically similar malignant from benign skin lesions can potentially aid in the highly subjective skin cancer prescreening stage.

The current treatment strategy followed by dermatologists is to completely remove the cancerous lesion from the skin surrounded by a rim or margin of healthy tissue. Histopathological evaluation of the resected tissue is one of the first margin assessment methods adopted during tumor excision surgery [8,9]. However, this approach evaluates only a few sections along the surgical margin, leading to missing out tumor extensions and subclinical tumor spread. According to the American Academy of Dermatology guidelines, a standard 5 mm excision margin is recommended for melanoma in situ or lentigo melanoma [10,11], while an excision margin of 1-2 cm is recommended for invasive melanoma [12]. The accessibility of the lesion and the functional importance of the affected tissue vary depending on the lesion's anatomical location. For example, a head and neck melanoma lesion may be functionally and cosmetically more critical than a melanoma lesion located at the thigh. Therefore, it is crucial to facilitate complete tumor removal while preserving healthy or unaffected surrounding tissue [9]; unfortunately, it is challenging to determine a standard for skin lesion excision margin. One study has reported that almost 60% of melanoma cases needed an excision margin greater than 5 mm to facilitate the complete removal and minimize recurrence rates [13]. The recommended surgical excision margins range between 5 to 20 mm depending on the depth of invasion of the lesion [14,15]. However, in some anatomical locations, it is very difficult to take the recommended margin and can be very cosmetically/functionally destructive [16]. Mohs microscopic surgery (MMS) has become a popular technique to ensure complete tumor removal [10,16,17]. Although MMS is a single visit surgical procedure, it is a tedious and time-consuming procedure. During MMS, the patient must wait until the analysis of the excised lesion is completed using frozen section evaluation [18]. Depending on the histopathology result, the doctor determines the need to remove another layer of tissue. The entire procedure, which can include multiple rounds of excision, can take several hours to be completed. In summary, all the above-mentioned margin assessment approaches have several limitations, including 1) incomplete tumor removal, 2) unwanted removal of functionally important healthy tissue, 3) tedious and time-consuming, 4) painful procedures that may require longer healing time for the patients.

In addition to dermoscopy, other optical imaging and sensing approaches have been explored for skin cancer screening, diagnosis, and surgical margin assessment [19,20]. Some of these optical techniques explored as potential skin cancer prescreening tools are reflectance confocal microscopy (RCM) [21], optical coherence tomography (OCT) [22], Raman spectroscopy [23],

Brillouin spectroscopy [24], photoacoustic imaging (PAI), hyperspectral imaging (HSI), and autofluorescence imaging [25,26]. RCM produces images of the horizontal sections of the lesions to view the cellular structures of the skin at varying depths [21]. OCT is a high-resolution imaging technique that generates 2D or 3D cross-sectional views of the tissue [27]. Although RCM and OCT monitor the structural or morphological changes in the skin non-invasively, the interpretation requires expert knowledge to understand the histological structures. In addition, both techniques do not provide information on the biochemical changes in the skin. Raman spectroscopy [28] is a vibrational spectroscopy technique that monitors the molecular signatures of the tissue and has been shown to be capable of identifying neoplastic progressions in the cells. It is currently being extensively investigated for discriminating malignant skin lesions; however, due to the slow image acquisition, it might not be practical for routine clinical evaluations [20,23,28]. PAI combines ultrasonic and optical imaging by irradiating the tissue with short laser pulses of different wavelengths [29]. The difference in light absorption by several tissue chromophores are then detected by an ultrasonic transducer. Currently this technique has been primarily used for investigating melanocytic lesions with melanin as the target chromophore [30]. HSI is a non-invasive imaging technique where each spatial point contains spectral information from several narrow bands across the electromagnetic spectrum. The use of HSI technique for melanoma detection is primarily based on the differences in the absorption spectrum of melanin and hemoglobin in the spectral bands 530-570 nm and 600-700 nm respectively [31]. However, capturing the spectral information over hundreds of narrow bands increases the complexity and cost of the instrument [32]. Although not many works have reported the diagnosis accuracy of skin cancer using PAI, some works [33,34] have used HSI for evaluating the diagnosis accuracy of skin cancer. Liu et. al. [33], utilized random forests models for identifying the staging of SCC and the influence of the selected region of interest on the performance using hyperspectral microscopic imaging. Huang et.al. [34], classified hyperspectral images of skin cancer into BCC, SCC and seborrheic keratosis using the 'you only look once' (YOLO) version 5 machine learning model and compared its performance with models trained on RGB images of the lesions. Leon et. al. [31], classified HSI images of benign and malignant pigmented skin lesions using machine learning methods such as support vector machine (SVM), random forests and neural networks.

Fluorescence based techniques for skin cancer diagnosis can be performed with the help of exogenous fluorophores as well as with skin endogenous fluorophores [26]. Exogenous fluorophore-based techniques pose several drawbacks, including invasiveness, the required waiting time between administration and imaging, and limited specificity. Autofluorescence techniques that monitor the fluorescence emission from the endogenous fluorophores in the tissue have also been explored for cancer diagnosis [35–37]. Such methods have the advantage of monitoring the biochemical changes in the cells as they undergo malignant transformations [36,38,39]. Nicotinamide adenine dinucleotide (NADH) and flavin adenine dinucleotide (FAD) are metabolic cofactors involved during the three main steps of cellular respiration: glycolysis, Krebs cycle, and oxidative phosphorylation [40,41]. As the cells transform from normal/healthy to full-blown malignancy, their metabolic rate increases. According to the Warburg hypothesis, the rate of anaerobic respiration through glycolysis dominates oxidative phosphorylation to enable energy production in the deeply situated tumor cells [42]. This can cause changes in the relative concentrations of NADH and FAD, which are reflected in their steady-state fluorescence intensities. In addition, the deviations in the metabolic pathways can induce variations in the relative concentrations of free and protein bound NADH and FAD and their protein binding sites [43]. These alterations can modulate the fluorescence decay lifetimes of these molecules. The relative amounts of collagen present in the extracellular matrix of malignant lesions compared to that in benign lesions is another important biomarker for malignant transformations [44]. Therefore, monitoring the autofluorescence emission of NADH, FAD, and collagen, can serve

as important indications of malignancy that can help discriminate benign and malignant skin lesions.

Several studies have reported the potential of steady-state intensity autofluorescence measurements to discriminate malignant from normal or benign skin lesions [45–47]. The excitation wavelengths used in the studies that explore skin autofluorescence ranged from 260 nm to 1000 nm [45]. Lohmann et. al, conducted in-vivo steady-state autofluorescence measurements on human skin lesions under 365 nm excitation and 470 nm emission. Considerable differences in the fluorescence intensities collected from melanoma and nevi lesions was observed [46]. Fast et. al., developed a multiphoton imaging system for in-vivo and ex-vivo autofluorescence measurement of human skin at 780 nm two-photon excitation and two emission channels at 535 nm and 720 nm. The imaging system was able to visualize melanocytic dendrites in actinically damaged skin, which is useful in distinguishing melanoma and other pigmented skin conditions [47].

Time-resolved techniques, such as multispectral autofluorescence lifetime imaging (maFLIM), have the inherent advantage of monitoring the fluorescence decay dynamics in addition to the steady-state intensities, reducing the false positive rate associated with highly pigmented and inflammatory conditions [48,49]. Several studies have explored using the time-resolved autofluorescence from the skin to discriminate skin cancer lesions [26,50–52]. Miller et. al, characterized the time-resolved autofluorescence response from SCC lesions and healthy skin in a mice model, using 480 nm excitation and 535 nm fluorescence emission, reporting a smaller short lifetime component in SCC lesions relative to the healthy skin [26]. Pires et. al, conducted in-vivo time-resolved fluorescence measurements in a murine model of cutaneous melanoma, using 378 nm and 445 nm excitations, and 440 nm and 514 nm emission detection, specifically targeting NADH and FAD fluorescence. Melanoma lesions were discriminated from healthy tissue using the long and short lifetime components of the bi-exponential fluorescence response [50]. Pastore et. al., employed multiphoton FLIM to detect metabolic changes on syngeneic melanoma mouse models, using two-photon excitation at 740 nm and 900 nm, and emission detection at 447 nm and 540 nm, to specifically measure the fluorescence lifetimes of free and bound NADH and FAD. A significant difference in the ratio of bound and free NADH between melanoma and healthy tissues was observed, while the short and long lifetime components of NADH did not vary significantly [51]. De Beule et. al, reported that average fluorescence lifetime from 390 and 600 nm emission bands under 355 and 440 nm excitation are useful in discriminating ex-vivo human biopsy samples from BCC and healthy skin tissue [52].

Several machine learning and deep learning techniques [53–61] had been implemented in previous publications to distinguish skin cancer lesions from either healthy tissue or other benign skin lesions. Some of the machine learning techniques that have been explored are k-nearest neighbors (kNN) [56], support vector machine (SVM) [57], logistic regression [58], and AdaBoost [57]. Many pre-trained convolutional neural networks (CNN) models such as VGG16 [59], VGGNet [60], AlexNet [59] and ResNet-152 [61], have also been used in combination with transfer learning to classify the dermoscope images collected from skin lesions. It is interesting to note that most of the previous works used publicly available dermoscopic datasets, such as ISBI, PH2, ISIC archive or Atlas [54,55]. Therefore, such works mainly make use of the morphological or textural features of the lesions, rather than its biochemical mechanisms.

In this work, we aim to classify pigmented skin lesions using the autofluorescence characteristics of intrinsic fluorophores in the tissue, which in turn reflects the underlying biochemical changes during malignant transformations [53]. This work develops an objective image-guided maFLIM based strategy to discriminate benign from malignant pigmented skin lesions using pixel-level features. The maFLIM images collected from 30 patients exhibiting benign or malignant skin lesions are employed in developing random forests and deep neural-network based classification models. The models generate prediction probability maps that could potentially be used to classify the lesion as either benign or malignant for screening and diagnosis purposes. In

addition, these maps could be used to determine the tumor margins in the acquired field of view; therefore, this technique could also potentially help delineating tumor margins during skin cancer excision surgery. Our previous work [62] introduced a feature extraction strategy based on frequency-domain analysis of fluorescence decay signals and utilized image-level global features for classifying benign and malignant lesions. The focus of the previous work was to classify lesions at the image level with the intention of assisting doctors in early diagnosis of skin cancer, and employed extensive feature selection techniques to select a feature subset for training Quadratic Discriminant Analysis (QDA) based machine learning models. In contrast, this work explores features extracted using time-domain deconvolution to discriminate benign and malignant lesions at the pixel-level. Pixel level classification potentially aids in identifying regions of malignancy and thereby assisting doctors in margin assessment. Moreover, this work investigates two more complex machine learning models – random forests and deep neural networks without explicit feature selection for pixel-level classification.

2. Methods

2.1. *maFLIM dermoscopy imaging of skin lesions*

The data used in this work is the same data used in a previous publication from our group [62]. A total of 30 patients ($n_{\text{patients}} = 30$) from the Dermatology Department of the Amaral Carvalho Cancer Hospital (Jahu, Sao Paulo, Brazil) were recruited for this study, following a human study protocol approved by the Internal Review Board of that institution (CAAE: 71208817.5.00005434). Only patients presenting at least one pigmented skin lesion undergoing biopsy examination for skin cancer diagnosis were recruited. The pigmented skin lesions considered in this work were solar lentigo, pSK, pigmented superficial BCC, pigmented nodular BCC, and melanoma.

Clinical maFLIM images were acquired *in vivo* from the patient's clinically suspicious lesions using an in-house developed time-domain maFLIM dermoscope previously described [35]. Figure S1 in the [Supplement 1](#) shows the clinical photograph of a melanoma skin lesion, and a photograph of the handheld maFLIM dermoscope imaging the forearm of a patient. In this maFLIM dermoscope, the excitation wavelength is 355 nm, and the skin tissue autofluorescence is simultaneously imaged at the three emission bands of 390 ± 20 nm, 452 ± 22.5 nm, and >496 nm, preferentially targeting collagen, NADH, and FAD autofluorescence emission, respectively. The maFLIM dermoscope was operated with a temporal resolution of 0.4 ns, field-of-view (FOV) of 8.65×8.65 mm², and lateral resolution of 120 μ m. For the rest of the paper, these three emission spectral bands of will be more conveniently referred to as 390 nm, 452 nm, and 500 nm. All images were acquired with an average laser excitation power of 10 mW measured at the sample, 140×140 pixels per image, and at a pixel rate of 10 kHz. These image acquisition parameters corresponded to an acquisition time of 1.96 s per image and an excitation energy exposure of 1.96 mJ at the sample, which is significantly lower than the maximum permissible exposure (MPE) levels for skin based on guidelines from the American National Standards Institute – ANSI [63].

After signing the corresponding institutional IRB-approved written informed consent form, each patient underwent the following imaging protocol right before the scheduled biopsy examination procedure. First, the lesion was gently cleaned with a gauze soaked in a saline solution. Then, the tip of the maFLIM dermoscope, previously disinfected using a gauze soaked in 70% ethanol, was placed in contact with the lesion, and an maFLIM image was acquired. Right after maFLIM imaging, lesion tissue biopsy was performed following standard procedures. Each maFLIM image was labeled based on the histopathological evaluation of the lesion biopsy, which was considered the gold standard in this study. A total of 60 skin lesions (i.e., $n_{\text{lesions}} = 60$) were imaged from the 30 patients recruited for this study.

2.2. maFLIM data pre-processing

The maFLIM data measured at each image pixel location (x, y) is composed of three concatenated fluorescence intensity temporal decay signals $s_{m,\lambda}(x, y, t)$ measured at the three targeted emission spectral bands (λ). The preprocessing steps applied to each pixel maFLIM temporal signal are as follows. First, offset and background subtraction was applied to the raw maFLIM signal, $s_{m,\lambda}(x, y, t)$, followed by spatial averaging (order 5×5) to increase the signal-to-noise ratio (SNR) of the time-dependent signal. Second, pixels presenting either signal saturation or low SNR (<15 dB) were detected and masked. Third, the duration of the temporal decay signals for all emission bands was adjusted to 149 temporal samples (59.6 ns) by applying zero padding. The pre-processed maFLIM decay signals at each spectral emission channel are represented as: $s_{\lambda_1}(x, y, t)$, $s_{\lambda_2}(x, y, t)$, and $s_{\lambda_3}(x, y, t)$, where $\lambda_1 = 390$ nm, $\lambda_2 = 452$ nm, and $\lambda_3 = 500$ nm.

2.3. Time domain deconvolution and feature extraction

In the context of time-domain maFLIM data analysis [64], the fluorescence decay $s_{\lambda}(x, y, t)$ measured at each emission spectral band (λ) and spatial location (x, y) can be modeled as the convolution of the fluorescence impulse response (FIR) $h_{\lambda}(x, y, t)$ of the sample and the instrument response function (IRF) $u_{\lambda}(t)$ measured at each λ :

$$s_{\lambda}(x, y, t) = u_{\lambda}(t) * h_{\lambda}(x, y, t) \quad (1)$$

The standard method for time-domain maFLIM data analysis proceeds by first deconvolving the IRF of each $u_{\lambda}(t)$ from the corresponding measured time-resolved fluorescence signal $s_{\lambda}(x, y, t)$ to estimate the sample FIR for each image pixel, $h_{\lambda}(x, y, t)$, which is usually modeled as a multi-exponential decay. The model order (number of exponential components) can be selected by analyzing the model-fitting mean squares error (MSE) as a function of the model order. For the maFLIM data of this study, a model order of two was selected, since the addition of a third component did not reduce the model-fitting MSE. Thus, the skin tissue FIR was modeled as:

$$h_{\lambda}(x, y, t) = \alpha_{fast,\lambda} e^{\frac{-t}{\tau_{fast,\lambda}(x,y)}} + \alpha_{slow,\lambda} e^{\frac{-t}{\tau_{slow,\lambda}(x,y)}} \quad (2)$$

In (2), $\tau_{fast,\lambda}$ and $\tau_{slow,\lambda}$ represent the time-constant (lifetime) of the fast and slow decay components, respectively; while $\alpha_{fast,\lambda}$ and $\alpha_{slow,\lambda}$ represent the contribution of the fast and slow decay components, respectively. The parameters of the bi-exponential decay model were estimated for each pixel by nonlinear least squares iterative reconvolution [64]. After deconvolution, the bi-exponential parameters estimated at each pixel were used as features representing the temporal dynamics of the fluorescence decays at each emission spectral band: $\alpha_{fast,\lambda}(x, y)$, $\alpha_{slow,\lambda}(x, y)$, $\tau_{fast,\lambda}(x, y)$, $\tau_{slow,\lambda}(x, y)$. The component contributions were normalized so that the sum of $\alpha_{fast,\lambda}(x, y)$ and $\alpha_{slow,\lambda}(x, y)$ is equal to one, so that $\alpha_{fast,\lambda}(x, y)$ and $\alpha_{slow,\lambda}(x, y)$ are in the range [65,1]. Since the normalized values of $\alpha_{fast,\lambda}(x, y)$ and $\alpha_{slow,\lambda}(x, y)$ are not independent, only $\alpha_{fast,\lambda}(x, y)$ was kept as a feature. One more bi-exponential feature considered was the average fluorescence lifetime for each spectral band ($\tau_{avg,\lambda}(x, y)$) computed at each pixel location as follows:

$$\tau_{avg,\lambda}(x, y) = \frac{\int t h_{\lambda}(x, y, t) dt}{\int h_{\lambda}(x, y, t) dt} \quad (3)$$

In addition, the following spectral intensity features were also estimated from the deconvolved FIR, $h_{\lambda}(x, y, t)$. Absolute fluorescence intensities $I_{\lambda}(x, y)$ for each emission spectral bands were

simply computed by time integrating the FIR $h_\lambda(x, y, t)$:

$$I_\lambda(x, y) = \int h_\lambda(x, y, t) dt \quad (4)$$

The normalized fluorescence intensities $I_{\lambda,n}(x, y)$ were then computed from the multispectral absolute fluorescence intensities $I_\lambda(x, y)$ as follows:

$$I_{\lambda,n}(x, y) = \frac{I_\lambda(x, y)}{\sum_\lambda I_\lambda(x, y)} \quad (5)$$

Lastly, the ratio of absolute intensities from the three spectral channels were computed at each pixel location resulting in three additional spectral intensity features:

$$\frac{I_{390,n}}{I_{452,n}}(x, y) = \frac{I_{390,n}(x, y)}{I_{452,n}(x, y)} \quad (6)$$

$$\frac{I_{452,n}}{I_{500,n}}(x, y) = \frac{I_{452,n}(x, y)}{I_{500,n}(x, y)} \quad (7)$$

$$\frac{I_{500,n}}{I_{390,n}}(x, y) = \frac{I_{500,n}(x, y)}{I_{390,n}(x, y)} \quad (8)$$

Altogether, the feature extraction approach based on time-domain deconvolution of the maFLIM data generated a total of six spectral intensity and twelve bi-exponential maFLIM features for each pixel location, as summarized in Table 1. In this paper, classification models were independently built on three different feature pools: spectral intensity feature pool ($n_{\text{features}} = 6$), bi-exponential feature pool ($n_{\text{features}} = 12$), and combined feature pool produced by combining both intensity and bi-exponential features ($n_{\text{features}} = 18$). The rationale for evaluating the classification performance using the three different feature pools independently is to gain understanding on how the different feature families (intensity vs. bi-exponential) contribute and complement each other to the specific classification problem (i.e., discriminating benign vs. malignant pigmented skin lesions).

Table 1. Feature set showing spectral intensity and bi-exponential maFLIM features at each pixel location.

Spectral Intensity features		Bi-exponential features		
$I_{390,n}$	$\frac{I_{390,n}}{I_{452,n}}$	$\alpha_{fast,390}$	$\alpha_{fast,452}$	$\alpha_{fast,500}$
$I_{452,n}$	$\frac{I_{452,n}}{I_{500,n}}$	$\tau_{slow,390}$	$\tau_{slow,452}$	$\tau_{slow,500}$
$I_{500,n}$	$\frac{I_{500,n}}{I_{390,n}}$	$\tau_{fast,390}$	$\tau_{fast,452}$	$\tau_{fast,500}$
		$\tau_{avg,390}$	$\tau_{avg,452}$	$\tau_{avg,500}$

In the machine learning strategies applied to this study, the classification models were trained at the pixel level (i.e., each image pixel was considered a training data with a corresponding feature vector). The pixel-level ground truth label was not available, however, since the lesion histopathological diagnosis was taken as the ground truth label for all the pixels of the lesion. This will introduce pixel mislabeling, as not all pixels within the image field of view correspond to lesion pixels, and not all lesion pixels correspond to the same lesion histopathological class. Nevertheless, the benefits of increasing the sample size by switching to a pixel-level classification model training can outweigh the disadvantage of pixel mislabeling.

2.4. Data splitting strategy

The dataset was randomly split patient-wise into five folds, each fold composed of data from six patients. Since there were 17 patients with benign lesion and 13 patients with malignant lesions,

four of the five folds contained maFLIM images from three patients with benign lesions and from three patients with malignant lesions, while the last fold contained maFLIM images from five patients with benign lesions and from one patient with malignant lesions. Since some patients had multiple lesions, the number of lesions in each fold was variable. When the dataset is split into five folds in this manner, it forms a partition. This random splitting of the dataset allows us to produce many partitions with different benign and malignant lesions combinations in each fold. To obtain a reasonable estimate of the model performance and to minimize any dependency on the data splitting, the entire process was repeated on 10 different random partitions of the data. The same 10 different random partitions were used to develop and estimate the performance of all classification models explored. Figure 1(a) shows the data splitting strategy. Each fold within a partition becomes the test set during a given iteration; thus, the number of iterations equals the number of folds within a partition, which is five. In this way, every lesion data gets a chance to be part of the test set. Out of the remaining four folds, one is randomly chosen as the validation set, and the remaining three folds forms the training set. Although the number of patients (30) and the number of imaged lesions (60) are small, the size of the training set is around 650,000 pixels and that of the testing and validation sets are each around 250,000 pixels. This is because the sample size increases multifold when the models are trained at the pixel level. It should be noted, however, that the final classification performance was estimated at the lesion level, as described in section 2.5. Since the dataset is unbalanced (41 benign lesions vs. 19 malignant lesions), all the classification models implemented in this paper used a class weight, whereby the ratio of weights assigned to the malignant and benign classes during classification was 2:1 [66].

2.5. Classification model with random forests

The workflow used to train and estimate the classification performance of the random forests model is depicted in Fig. 1(b). The random forest model was trained using 50 trees, and the Gini impurity was used as the criterion for measuring the quality of the split [67]. The tree was expanded until all leaves contain less than 2 samples. The number of features considered when looking for the best split is the square root of the number of features. The performance of the model was evaluated independently on the three feature pools mentioned in section 2.3, with consistent data splits across all the partitions (i.e., the splits of the data were kept the same when evaluating the intensity, bi-exponential, or combined feature pools). A random forests classifier trained at the pixel-level generated a prediction probability for every pixel location in the input lesion image. Thus, the prediction probabilities of all pixels from a given lesion image were combined to produce a prediction probability map of the lesion.

Since the ground truth annotation is only available at the lesion level and not at the pixel level, a lesion-level classification needed to be obtained from the prediction probability maps by applying two thresholds. The first threshold (Th_{pix}) was applied on the pixel-level prediction probabilities to generate a binary classification map. Every pixel in the classification map was set to '0' or '1' depending on the pixel prediction probability (i.e., 0: prediction probability $< Th_{pix}$; 1: prediction probability $> Th_{pix}$). The second threshold (Th_{per}) was applied on the proportion of pixels classified as malignant (or '1') in each lesion image. After applying Th_{per} to the proportion of pixels classified as malignant, the entire lesion image was classified as either benign or malignant.

The two thresholds Th_{pix} and Th_{per} were optimized using the training and validation sets before being applied to the test set as follows. For each fold in a partition, the model was optimized on the training set, and prediction probability maps were generated for every lesion image in the validation set. A receiver operator characteristics curve (ROC) was then constructed using these validation prediction probability maps, and a threshold Th_{pix} was chosen from the operating point in the ROC curve closest to the ideal operating point (0,1). The chosen Th_{pix} was then applied on the prediction probability maps to generate classification maps of the lesion images in

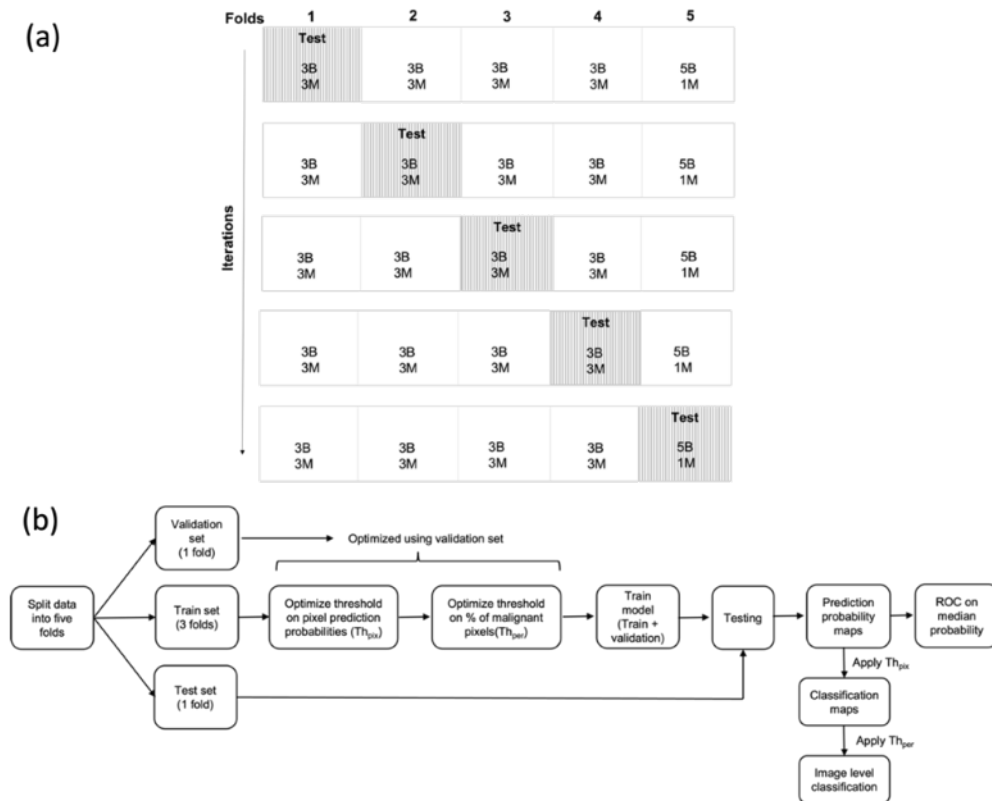


Fig. 1. (a) Data-splitting strategy in a sample partition showing the distribution of patients with either benign (B) and malignant (M) lesions in the five fold for each iteration. The test set (shaded box) is sequentially varied in each iteration. A total of 10 of such random partitions are generated. (b) Workflow used to train and estimate the classification performance of the random forest models. For each of the 10 random partitions generated, this workflow was repeated until all five folds were used as test set.

the validation set. Proportion of pixels classified as malignant pixels were computed for every classification map, and another ROC was constructed on the computed percentage. Subsequently, the threshold Th_{per} was chosen as the operating point on the ROC curve closest to the ideal operating point (0,1).

After optimizing Th_{pix} and Th_{per} using the validation set, a new random forests model was retrained on the combined training and validation sets. The retrained model was then applied to the test set to generate prediction probability maps to each lesion included in the test set. Finally, the thresholds Th_{pix} and Th_{per} , previously optimized using the training and validation sets from the corresponding fold in the partition, were applied to produce a lesion-level classification of each lesion included in the test set. The test set classification performance was finally quantified either directly from the generated prediction probability maps, or after applying the thresholds Th_{pix} and Th_{per} , as described as follows.

Test set performance estimation from prediction probability maps: Upon completion of all the five folds in a partition, every lesion had become a part of a test set. An ROC curve was constructed on the median pixel value of the prediction probability maps from the test lesions in the partition, and a partition-level AUC was computed. This process was repeated for the 10

different random partitions generated as described in section 2.4. Thus, for each model explored, the mean and standard deviation the AUC of the ten partitions were computed.

Test set performance estimation after thresholding: Classification maps were generated after applying the previously optimized Th_{pix} on the test set prediction probability maps. An ROC curve was constructed on the proportion of pixels classified as malignant of the classification maps from the test lesions in the partition, and a partition-level AUC is computed. In addition, lesion-level classification was obtained after applying the previously optimized Th_{per} on the classification maps. Confusion matrices are constructed at the partition-level by pooling together all the test lesions within each partition. Sensitivity, specificity, accuracy, misclassification rate, F-score and precision were finally computed from the confusion matrices. This process was repeated for the 10 different random partitions generated as described in section 2.4. Thus, for each model explored, the mean and standard deviation of each performance metric of the ten partitions were computed.

The random forests algorithm allows monitoring the importance of features during the training process. Feature importance is measured using the Gini index which is the sum over the number of splits that uses the feature, over all the trees, calculated proportionally to the number of samples it splits [67]. Therefore, the importance of each feature in all the feature pools were estimated during the training processes, which can provide insight on the most relevant features of each pool.

2.6. Classification model with deep neural networks

A neural network classification model was also explored independently using the three different feature pools described in section 2.3. To facilitate reliable comparison, the splits of the data into train, validation, and test sets across all the ten partitions were kept the same as in section 2.5. Figure 2(a) outlines the diagram of the deep learning model showing the input layer, hidden layers, and output layer. The activation function at each hidden layer was rectified linear unit (ReLU), while that at the output layer was sigmoid. This caused the model output value to have a range between 0 and 1. To prevent overfitting, dropout regularization was applied between each hidden layer with a dropout rate of 0.3. In Keras deep learning framework [68], the dropout rate is defined as the fraction of units to drop. Adam optimizer [69] was selected for optimizing the model using a loss function computed from the ground truth label and the prediction probability with a learning rate of 0.001. A binary cross-entropy loss function was used, as it performs best for binary classification problems [70]. The loss function is defined as:

$$loss = \frac{1}{N} \sum_{i=1}^N -(q_i \log(pr_i) + (1 - q_i) \log(1 - pr_i)) \quad (9)$$

In (9), N is the number of data points, pr_i is the predicted probability, and q_i is the truth label. If the truth label is 1, the second part of the equation becomes zero and activates the first part. If the truth label is 1 and the predicted probability is close to 1, the loss function becomes closer to zero. Similarly, if the predicted probability is close to 0, when the truth label is 0, the loss function is again close to zero.

The number of hidden layers and the number of units in each layer are the model hyperparameters, which were tuned with the training and validation sets using the ‘Keras tuner’ functionality provided by the Keras library. Keras tuner is an easy-to-use hyperparameter optimization framework that defines a search space and utilizes built-in algorithms to find the best hyperparameter values. The hyperband algorithm is chosen as an effective way to simultaneously tune multiple hyperparameter configurations in the model [71]. More information about the hyperband algorithm is provided in the Supplement 1. The following hyperparameters were tuned by varying them over a range using a specific step size.

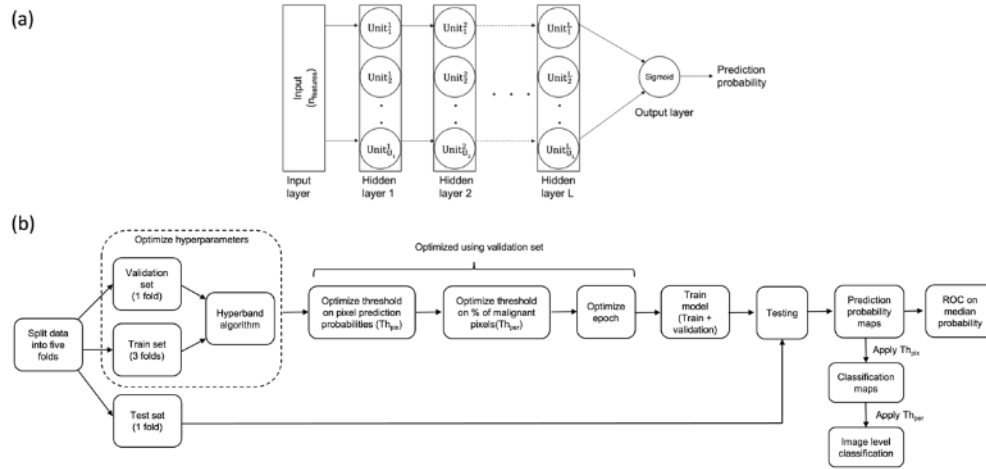


Fig. 2. (a) Schematic of the neural network model, (b) Workflow used to train and estimate the classification performance of the neural network models.

1. Number of hidden layers 'L' – min value = 3, max value = 7, step size = 1
2. Number of units in each hidden layer 1, 'U₁' – min value = 20, max value = 100, step size = 10

The workflow used to train and estimate the classification performance of the neural network model is depicted in Fig. 2(b). The hyperparameters are tuned using the hyperband algorithm. The thresholds Th_{pix} and Th_{per} are optimized in a similar fashion as explained in section 2.5. Subsequently, the model is trained on the train set for fifty epochs using the optimum hyperparameters, and the validation accuracy is monitored simultaneously at each epoch to check for overfitting. The epoch at which the validation accuracy is the highest is chosen as the best epoch. After optimizing the epoch, the model is trained on a set obtained by combining the training and validation sets using the optimized hyperparameters and the optimum epoch. The test set is then tested on the trained model to generate prediction probability maps and classification maps corresponding to each lesion included in the test set. Thresholds Th_{pix} and Th_{per} used on the test set are the ones optimized using the train and validation sets from the same fold in the partition. The test set classification performance can be computed from either the generated prediction probability maps or the classification maps. The performance estimation from the prediction probability maps and the classification maps are the same as explained in section 2.5.

Since each partition has five iterations, the total number of test sets is 50, corresponding to the ten partitions. This means that 50 sets of optimum hyperparameter configurations are generated. A histogram of the tuned hyperparameters was generated to determine the most frequent values found as optimal for each of the hyperparameters. Figure. S4. in the Supplement 1 shows the histogram distributions of all the hyperparameters.

3. Results

3.1. *maFLIM dermoscopy clinical imaging of skin lesions*

The distribution of patients ($n_{patients} = 30$) and lesions ($n_{lesions} = 60$) imaged in this study showing benign and malignant conditions is provided in Table 2. Benign lesions included solar lentigo and pigmented seborrheic keratosis (pSK). Malignant lesions included pigmented superficial

BCC, pigmented nodular BCC, and melanoma. The maFLIM feature maps of a representative melanoma lesion are shown in Fig. 3. For comparison purposes, the scales of the feature maps across the three spectral wavelengths are kept consistent.

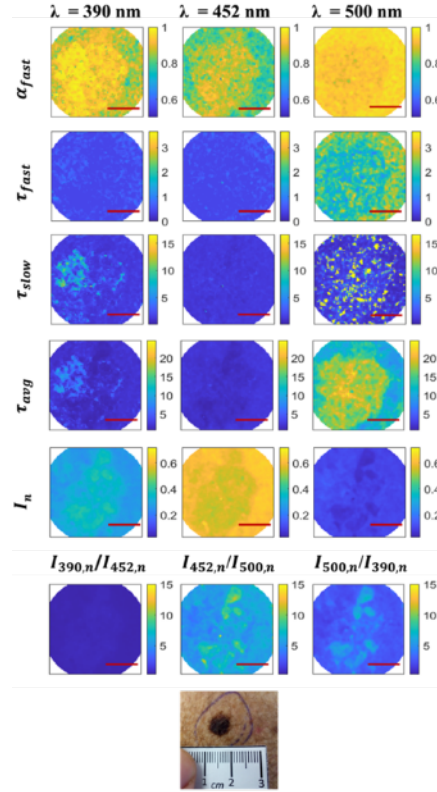


Fig. 3. Representative maFLIM feature maps of a melanoma lesion. The columns show the feature maps corresponding to the three emission spectral channels. Rows correspond to i) maps of the weight of the fast decay, ii) fast lifetime maps, iii) slow lifetime maps, iv) average lifetime maps, v) normalized integrated intensity maps, vi) maps of the ratios of the intensities, vii) while light image of the lesion. Scale bar: 3 mm.

Table 2. Distribution of imaged benign and malignant lesions.

	Type	No. patients	No. lesions
Benign	Solar lentigo	2	10
	Pigmented seborrheic keratosis	15	31
Malignant	Pigmented superficial BCC	2	6
	Pigmented nodular BCC	5	5
	Melanoma	6	8

3.2. Random forests classification model

3.2.1. Threshold optimization

Thresholds Th_{pix} and Th_{per} are optimized using the training and validation sets and then applied to the testing set during each of the five iterations or folds for a given partition (Fig. 1(b)). This means that after executing the experiments on all the 10 partitions, 50 different values of Th_{pix} and Th_{per} are optimized. Figure S2 in the [Supplement 1](#) shows the distribution of the 50 optimized Th_{pix} and Th_{per} threshold values for the three feature pools independently evaluated. For the three feature pools, the optimized Th_{pix} values were mostly less than 0.2, while the optimized Th_{per} values were mostly around 0.5.

3.2.2. Test set performance estimation from prediction probability maps

The predictions probability maps generated from the test sets in a representative data partition using the random forest model trained on the combined feature pool are shown in Fig. 4(a). The maps are generated from the test set images in the five folds of the representative partition, and therefore contains every lesion image in the dataset. The labels 'Benign' and 'Malignant' are from the image-level ground truth of the lesions. Mean value of the prediction probabilities can be computed from the prediction probability map of each lesion image in the test set from all the 5 folds in the partition. The violin plots showing the distribution of mean prediction probabilities from the test sets in the sample partition for the three feature pools are shown in Figs. 4(b), 4c, and 4d. The labels 'Benign' and 'Malignant' are from the image level ground truth. Non-parametric Mann-Whitney U test was conducted on the mean prediction probabilities between benign and malignant test set lesions for each feature pool, and the p-values were found to be statistically significant ($p\text{-value} < 0.01$). The p-values obtained for this sample partition are: $7.59e-05$ for intensity feature pool, $1.59e-06$ for bi-exponential feature pool, and $6.04e-07$ for the combined feature pool. Table S1 in the [Supplement 1](#) shows the results of the Mann-Whitney U test obtained for all the 10 data partitions for random forest models trained on the three feature pools, with all p-values obtained being < 0.01 .

Table 3 shows the AUCs of the ROCs constructed on the mean prediction probabilities from the prediction probability maps in the test sets, using random forests classifier trained on the three explored feature pools. The ROC curves are constructed by combining the lesions from all the five folds of a given partition, e.g., one ROC curve is constructed for each partition. The last row shows the mean and standard deviation values of the ten ROC-AUC values obtained from the ten different partitions. As it can be seen, the ROC-AUC values with the combined feature pool were higher than for the other two feature pools, with a mean ROC-AUC of 0.88 ± 0.01 .

3.2.3. Test set performance estimation after thresholding

The classification maps obtained after applying Th_{pix} on the prediction probability maps of a representative data partition using the random forest model trained on the combined feature pool are shown in Fig. 5(a). Figures 4(a) and 5(a) are from the same data partition. The labels 'Benign' and 'Malignant' at the top of the figure separating the two groups are from the image level ground truth of the lesions. The color-coded labels indicate pixels marked as malignant (purple - label '1') or benign (black - label '0') after applying Th_{pix} on the prediction probability maps. Proportion of pixels marked as malignant can be computed from each classification map in a partition. Figures 5(b), 5(c) and 5(d) shows the violin plots indicating the distribution of the proportion of malignant pixels from all the test sets in the sample partition for the three feature pools. The plots are separated into benign and malignant groups based on the image level ground truth values. Non-parametric Mann-Whitney U test conducted on the proportion of malignant pixels between benign and malignant test set lesions shows statistically significant p-values for the three feature pools ($p < 0.01$). The obtained p-values for this sample partition are:

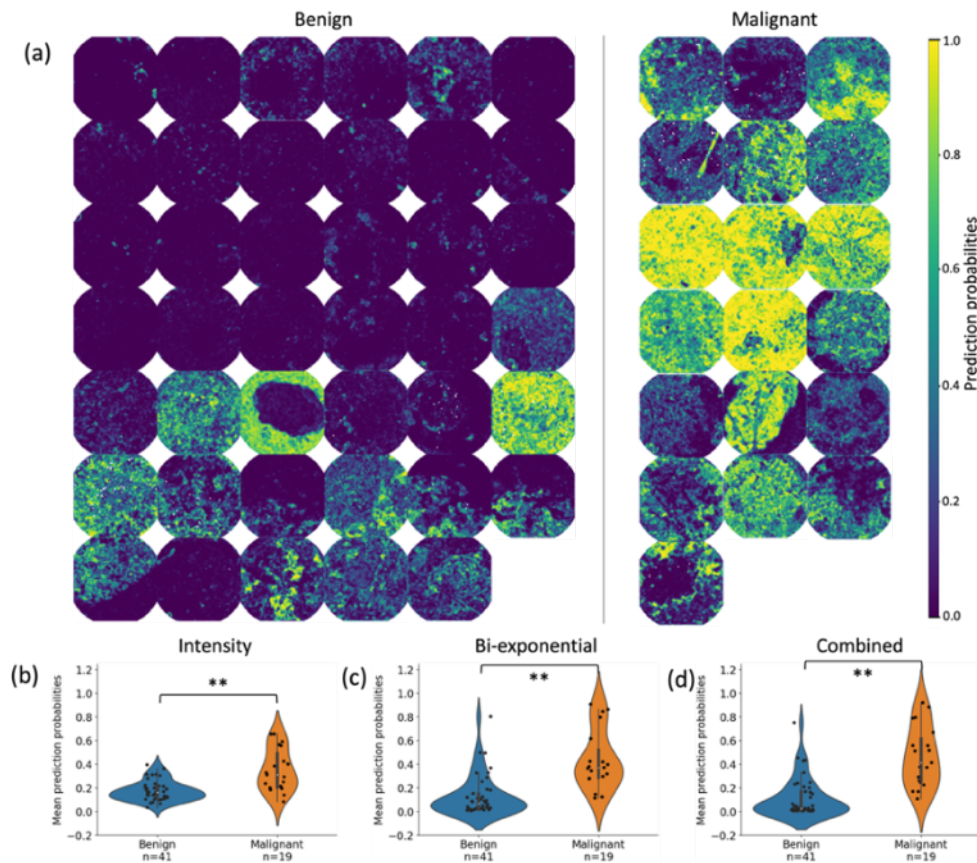


Fig. 4. (a) Prediction probability maps generated with random forests classifier for a sample data partition trained on the combined feature pool. Labels 'Benign' and 'Malignant' are from the image-level ground truth of the lesions. Violin plots showing the mean prediction probabilities from the test sets within the sample partition for: (b) intensity (p-value = 7.59×10^{-5}), (c) biexponential (p-value = 1.59×10^{-6}), and (d) combined feature pools (p-value = 6.04×10^{-7}). ** above the plots indicate that the two groups are statistically significant with a p-value < 0.01.

4.74×10^{-5} for intensity feature pool, 4.37×10^{-6} for bi-exponential feature pool, and 9.89×10^{-7} for the combined feature pool. Table S2 in the [Supplement 1](#) shows the results of the Mann-Whitney U test obtained for all the 10 data partitions for random forest models trained on the three feature pools, with all p-values obtained being < 0.01.

Table 4 shows the AUCs of the ROCs constructed on the proportion of malignant pixels from the classification maps in the test sets, using random forests classifier trained on the three explored feature pools. The last row shows the mean and standard deviation values of the ten ROC-AUC values. The ROC-AUC values with the combined feature pool were higher than for the other two feature pools, with a mean ROC-AUC of 0.87 ± 0.02 .

To estimate the lesion-level classification performance, we need to generate confusion matrices for the test sets within each data partition. Applying Th_{per} on the classification maps classifies each test set lesion at the image-level. Confusion matrices are generated at the partition-level by including all test images from the five folds of that partition. Table 5 shows the mean and

Table 3. The AUCs of the ROCs constructed from the mean prediction probabilities of the test sets in each partition using the random forest classifier for the three feature pools.

Partitions	Spectral Intensity feature pool	Bi-exponential feature pool	Combined feature pool
1	0.80	0.84	0.87
2	0.80	0.83	0.86
3	0.81	0.86	0.89
4	0.84	0.87	0.89
5	0.84	0.87	0.89
6	0.79	0.86	0.88
7	0.81	0.85	0.87
8	0.82	0.89	0.90
9	0.81	0.86	0.88
10	0.84	0.87	0.89
Mean \pm Standard Deviation	0.82 ± 0.01	0.86 ± 0.01	0.88 ± 0.01

Table 4. The AUCs of the ROCs constructed from the proportion of malignant pixels of the test sets from each partition using the random forest classifier for the three feature pools.

Partitions	Spectral Intensity feature pool	Bi-exponential feature pool	Combined feature pool
1	0.85	0.84	0.85
2	0.82	0.86	0.86
3	0.75	0.87	0.86
4	0.84	0.84	0.86
5	0.79	0.88	0.89
6	0.82	0.87	0.87
7	0.87	0.80	0.91
8	0.83	0.87	0.90
9	0.85	0.81	0.85
10	0.80	0.88	0.89
Mean \pm Standard deviation	0.82 ± 0.03	0.85 ± 0.03	0.87 ± 0.02

standard deviation of the sensitivity, specificity, accuracy, misclassification rate, F-score and precision from all the 10 partitions.

3.2.4. Feature importance

The feature importance metric (Gini index) obtained for each feature during all the 50 different test sets (5 iterations x10 partitions) are used to compute the mean and standard deviation shown in the Fig. 6. This, in turn, reveals the most important features in each feature pool. The three most significant features from the spectral intensity feature pool are $I_{390,n}$, $\frac{I_{390,n}}{I_{452,n}}$, and $\frac{I_{500,n}}{I_{390,n}}$. The

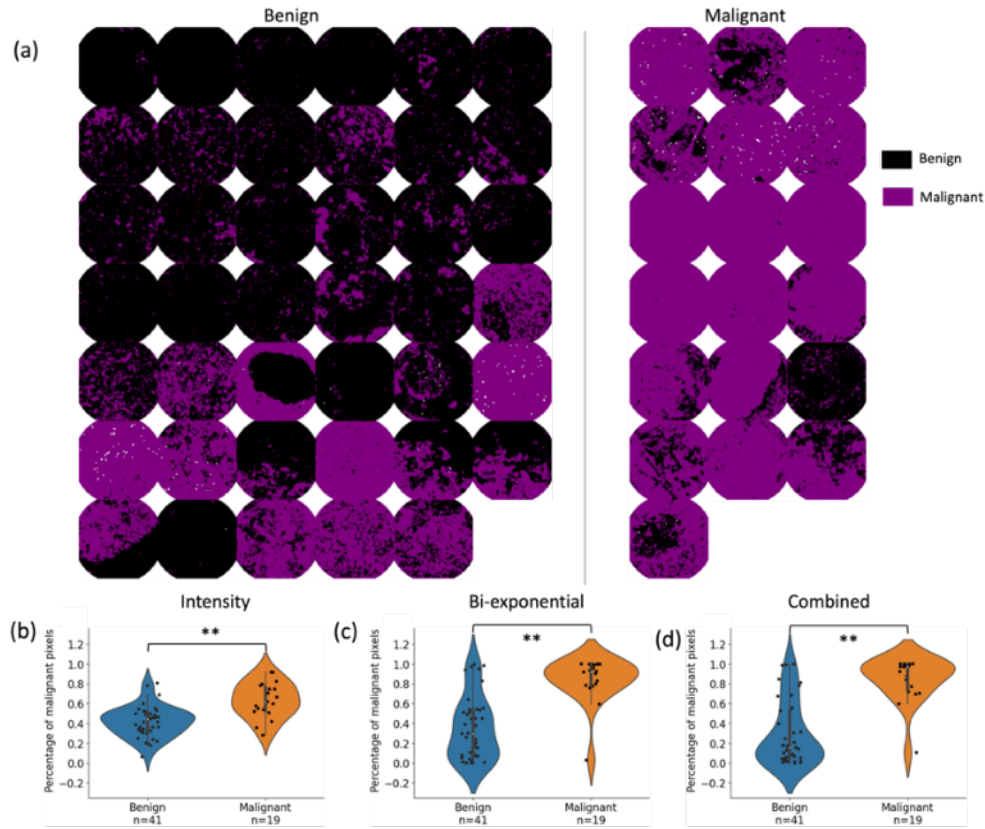


Fig. 5. (a) Classification maps generated with random forests classifier for a sample data partition trained on the combined feature pool. Violin plots showing the proportion of malignant pixels from the test sets within the sample partition for: (b) intensity (p-value = 4.74e-05), (c) biexponential (p-value = 4.37e-06), and (d) combined feature pools (p-value = 9.89e-07). ** above the plots indicate that the two groups are statistically significant with a p-value < 0.01.

Table 5. Performance metrics calculated from the test sets over all the data partitions when trained on a random forest classifier – these metrics are computed at the partition level.

Feature pool	Sensitivity (%)	Specificity (%)	Accuracy (%)	Misclassification rate (%)	F-score (%)	Precision (%)
Intensity	64.21 ± 11.95	77.07 ± 14.18	73.00 ± 7.37	27.00 ± 7.37	60.36 ± 5.31	61.78 ± 16.32
Bi-exponential	72.63 ± 10.73	76.09 ± 5.95	75.00 ± 4.01	25.00 ± 4.01	64.58 ± 6.12	58.81 ± 5.47
Combined feature pool	74.73 ± 14.66	76.83 ± 9.58	76.17 ± 5.00	23.83 ± 5.00	66.19 ± 7.11	61.46 ± 8.83

three most important features from the bi-exponential feature pool are $\tau_{fast,452}$, $\tau_{fast,390}$, and $\alpha_{fast,390}$. In the combined feature pool, the three most important features are $I_{390,n}$, $\alpha_{fast,390}$, and $I_{452,n}$.

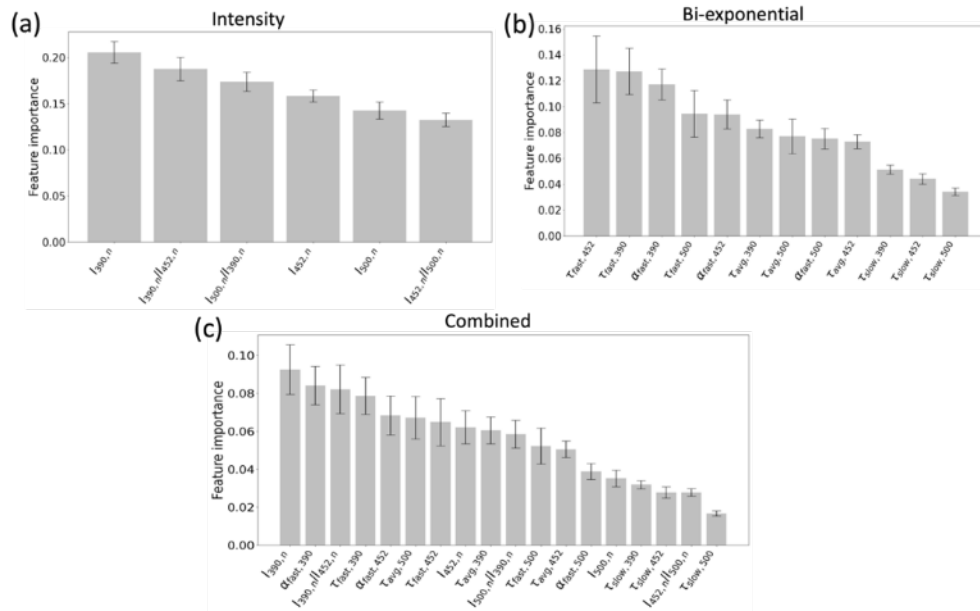


Fig. 6. Mean and standard deviation of feature importance metrics (Gini index) obtained for each feature during all the 50 test sets (5 iterations x 10 partitions) computed for the: (a) intensity, (b) biexponential, and (c) combined feature pools.

3.3. Neural networks classification model

3.3.1. Threshold optimization

Thresholds Th_{pix} and Th_{per} are optimized using the training and validation sets during each iteration of the workflow (Fig. 2(b)). The thresholds applied to the test sets are the ones optimized during that iteration using the train and validation sets from the same partition. This means that after executing the experiments on all the 10 partitions, 50 different values of Th_{pix} and Th_{per} are obtained. Figure S3 in the [Supplement 1](#) shows the distribution of Th_{pix} and Th_{per} for the three feature pools respectively. For all the feature pools, the Th_{pix} values are mostly centered around 0.2, while the Th_{per} values are mostly centered around 0.5.

3.3.2. Test performance estimation from prediction probability maps

The predictions probability maps generated from the test sets in a sample data partition using the neural network model trained on the combined feature pool are shown in Fig. 7(a). The sample data partition is the same as that was used for the random forests model in Figs. 4 and 5. The maps shown in the figure are generated from the test set images in the five folds of the partition, and therefore contains every lesion image in the dataset. The labels ‘Benign’ and ‘Malignant’ are from the image-level ground truth of the lesions. Mean value of the prediction probabilities can be computed from the prediction probability map of each lesion image in the test sets from all the five folds in the partition. Figures 7(b), 7(c) and 7(d) contains the violin plots showing the distribution of mean prediction probabilities from the test sets in the sample partition for the three feature pools. The labels ‘Benign’ and ‘Malignant’ are from the image level ground truth. Mann-Whitney U test was conducted on the two groups for each feature pool. The p-values obtained for this sample data partition are: $4.72e-04$ for intensity feature pool, $5.80e-05$ for bi-exponential feature pool, and $5.12e-07$ for the combined feature pool. Table S3 in the

Supplement 1 shows the results of the Mann-Whitney U test obtained for all the 10 data partitions for neural network models trained on the three feature pools, with all calculated p-values < 0.01.

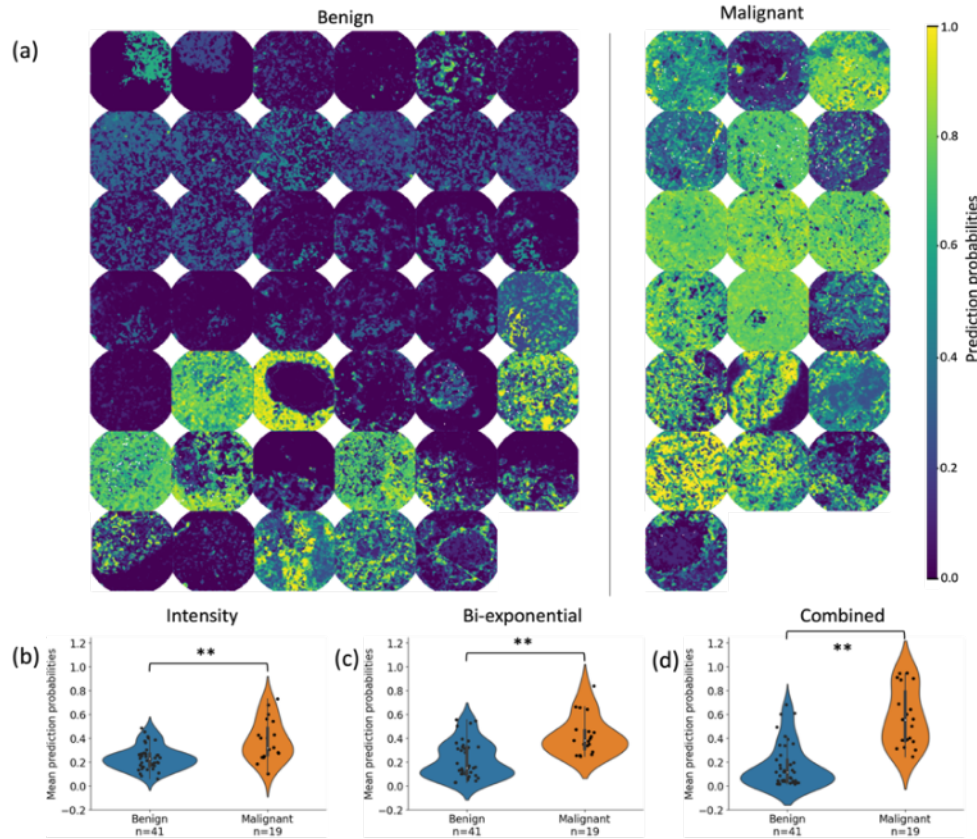


Fig. 7. (a) Prediction probability maps generated with neural network classifier for a sample data partition trained on the combined feature pool. Labels ‘Benign’ and ‘Malignant’ are from the image-level ground truth of the lesions. Violin plots showing the mean prediction probabilities from the test sets within the sample partition for: (b) intensity (p-value = 4.72e-04), (c) biexponential (p-value = 5.80e-05), and (d) combined feature pools (p-value = 5.12e-07). ** above the plots indicate that the two groups are statistically significant with a p-value < 0.01.

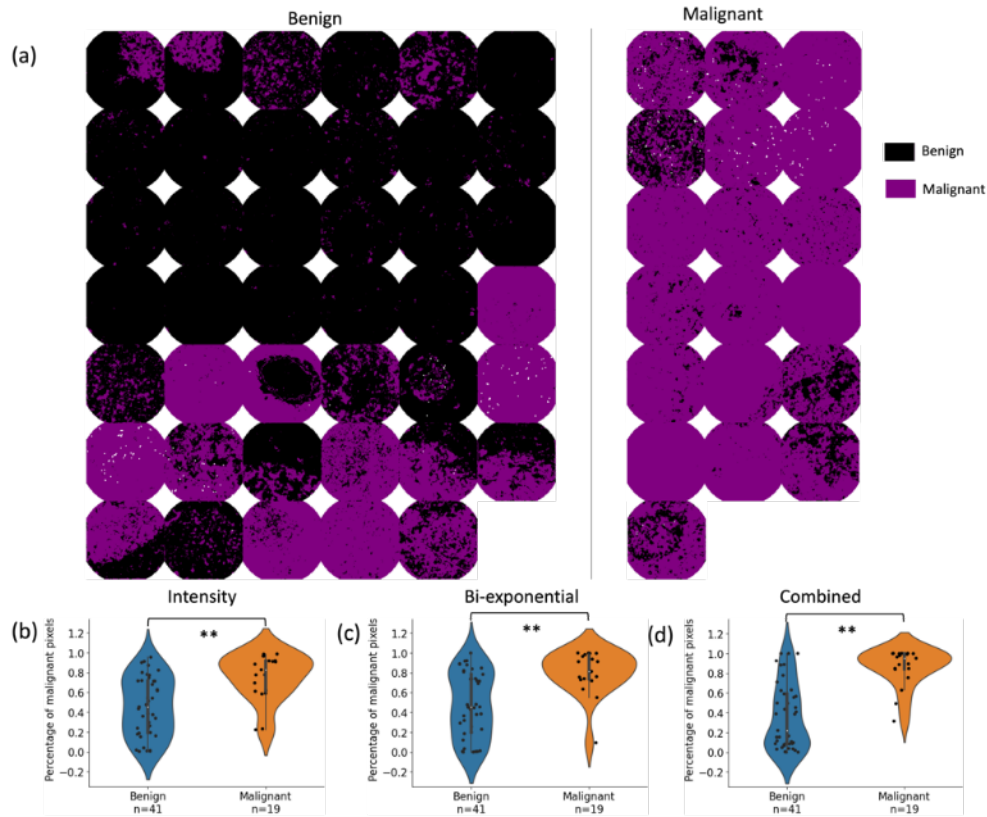
Table 6 shows the AUCs of the ROCs constructed on the mean prediction probabilities at the partition level from the prediction probability maps in the test sets, using neural network classifier trained on the three explored feature pools. The ROC-AUC values with the combined feature pool were higher than for the other two feature pools, with a mean ROC-AUC of 0.88 ± 0.02 .

3.3.3. Test performance estimation after thresholding

The classification maps obtained after applying Th_{pix} on the prediction probability maps of a sample data partition using the neural network model trained on the combined feature pool are shown in Fig. 8(a). Figures 7 and 8 are obtained from the same data partition. The labels ‘Benign’ and ‘Malignant’ at the top of Fig. 8(a) separating the two groups are from the image level ground truth of the lesions. The color-coded labels indicate pixels marked as malignant (purple - label ‘1’) or benign (black - label ‘0’) after applying Th_{pix} on the prediction probability maps. Proportion of pixels marked as malignant can be computed from each classification map

Table 6. The AUCs of the ROCs constructed from the mean prediction probabilities of the test sets in each partition using the neural network classifier for the three feature pools.

Partitions	Spectral Intensity feature pool	Bi-exponential feature pool	Combined feature pool
1	0.77	0.88	0.86
2	0.76	0.83	0.90
3	0.78	0.87	0.89
4	0.83	0.83	0.90
5	0.81	0.89	0.89
6	0.77	0.88	0.85
7	0.78	0.81	0.88
8	0.78	0.83	0.91
9	0.80	0.80	0.87
10	0.81	0.81	0.89
Mean \pm Standard deviation	0.79 ± 0.02	0.84 ± 0.03	0.88 ± 0.02

**Fig. 8.** (a) Classification maps generated with neural network classifier for a sample data partition trained on the combined feature pool. Violin plots showing the proportion of malignant pixels from the test sets within the sample partition for: (b) intensity (p-value = 3.01×10^{-4}), (c) biexponential (p-value = 2.73×10^{-4}), and (d) combined feature pools (p-value = 4.05×10^{-6}). ** above the plots indicate that the two groups are statistically significant with a p-value < 0.01.

in a partition. Figures 8(b), 8(c) and 8(d) show the violin plots indicating the distribution of the proportion of malignant pixels from the test sets in the sample partition for the three feature pools. The plots are separated into benign and malignant groups based on the image level ground truth values. Mann-Whitney U test was conducted on the proportion of malignant pixels, and p-values for the three feature pools were statistically significant ($p < 0.01$). The obtained p-values for this sample data partition are: 3.10×10^{-4} for intensity feature pool, 2.73×10^{-4} for bi-exponential feature pool, and 4.05×10^{-6} for combined feature pool. Table S4 in the [Supplement 1](#) shows the results of the Mann-Whitney U test obtained for all the 10 data partitions for neural network models trained on the three feature pools, with all calculated p-values < 0.01 .

Table 7 shows the AUCs of the ROCs constructed on the proportion of malignant pixels from the classification maps in the test sets, using neural network classifier trained on the three explored feature pools. The ROC-AUC values with the combined feature pool were higher than for the other two feature pools, with a mean ROC-AUC of 0.86 ± 0.02 .

Table 7. The AUCs of the ROCs constructed from the proportion of malignant pixels of the test sets using the neural network classifier for the three feature pools.

Partitions	Spectral Intensity feature pool	Bi-exponential feature pool	Combined feature pool
1	0.85	0.87	0.82
2	0.80	0.88	0.88
3	0.72	0.83	0.83
4	0.82	0.80	0.88
5	0.83	0.86	0.86
6	0.80	0.82	0.81
7	0.84	0.80	0.89
8	0.79	0.79	0.87
9	0.75	0.74	0.86
10	0.83	0.81	0.89
Mean \pm Standard deviation	0.80 ± 0.03	0.82 ± 0.04	0.86 ± 0.02

To estimate the lesion-level classification performance, we need to generate confusion matrices for the test sets within each data partition. Applying Th_{per} on the classification maps classifies each test set lesion at the image-level. Confusion matrices are generated at the partition-level by including all test images from the five folds of that partition. Table 8 shows the mean and standard deviation of the sensitivity, specificity, accuracy, misclassification rate, F-score and precision from all the 10 partitions.

Table 8. Performance metrics calculated from the test sets over all the data partitions when trained on a neural network classifier

Feature pool	Sensitivity (%)	Specificity (%)	Accuracy (%)	Misclassification rate (%)	F-score (%)	Precision (%)
Intensity	63.15 ± 13.31	72.43 ± 14.39	69.50 ± 7.49	30.50 ± 7.49	56.74 ± 6.05	54.21 ± 9.42
Bi-exponential	74.21 ± 15.86	75.60 ± 10.17	75.16 ± 52.94	24.83 ± 52.94	65.01 ± 7.46	59.95 ± 85.13
Combined feature pool	76.84 ± 12.49	78.29 ± 5.50	77.83 ± 3.50	22.16 ± 3.50	68.31 ± 6.59	62.39 ± 4.48

3.3.4. Neural network hyperparameter optimization

Since the classifier trained on each feature pool is evaluated on 10 data partitions with five iterations per partition, 50 different configurations of hyperparameters are obtained. The histogram of the hyperparameter values that are selected during the 10 data partitions are shown in Figure S4 in the [Supplement 1](#). The most frequent number of hidden layers selected were three, six, and four for the intensity, bi-exponential, and combined feature pools, respectively. The number of units for each of the hidden layers is also a hyperparameter that was tuned. Since the number of layers is variable over all the iterations of the ten partitions, the average number of units for the hidden layers were estimated for the three feature pools, showing similar distributions peaking around 60 to 70 units for the three feature pools. Finally, a small value of epoch, between 1 and 5, was frequently selected for all the three feature pools.

4. Discussion

Previous studies from our group have demonstrated the feasibility of label-free biochemical and metabolic imaging of skin lesions using maFLIM dermoscopy [35,72,73]. In this study, we demonstrate the potential of using pixel-level maFLIM features to develop classification models capable of generating prediction probability maps and classification maps that could potentially be used to not only detect malignant skin lesion, but also delineate lesion margins. Random forest and deep neural networks classification models were explored in this work, as they do not require explicit feature selection [74]. Random forest models select features at every node in a decision tree to determine the feature for splitting the node. Neural networks learn the relevant input features and optimizes the weights accordingly during the training process. In addition to these two classification models, three different feature pools were also explored. The autofluorescence emission of skin lesions can be quantified in terms of spectral intensity and time-resolved bi-exponential fluorescence features extracted at the pixel-level from the maFLIM dermoscopy imaging data. Thus, feature pools with either spectral intensity or bi-exponential features, and a combined feature pool were independently evaluated with each classification model.

To identify which classification model and feature pool performs the best out of the six model/feature-pool combinations, a rigorous cross-validation strategy was adopted. First, the data was partitioned at the patient level, i.e., the lesion images that belong to one patient were all assigned to either the training, validation, or test sets. Second, 10 different partitions were used to minimize potential bias associated to a specific random data splitting. Finally, the same 10 partitions were used to validate each of the evaluated model/feature-pool combinations to provide a fair comparison of their performance. Such rigorous validation strategy allows for unbiased classification performance estimations using the limited available dataset (30 patients, 60 skin lesions), which is a common limitation of medical imaging datasets.

The performance estimated from the prediction probability maps with each feature pool was consistent in both models. The combined spectral intensity and bi-exponential feature pool resulted in the highest performance in the ROCs constructed from mean prediction probabilities (mean ROC-AUC: 0.88 ± 0.01 for both models). Also interestingly, the bi-exponential feature pool (mean ROC-AUC: 0.86 ± 0.01 for random forest and 0.84 ± 0.03 for neural networks) outperformed the spectral intensity feature pool (mean ROC-AUC: 0.82 ± 0.01 for random forest and 0.79 ± 0.02 for neural networks). These observations highlight the importance of time-resolved fluorescence features for discriminating malignant from benign pigmented skin lesions. The violin plots in Figs. 4 and 7 also show that the mean prediction probabilities from each lesion prediction probability map are statistically significant between benign and malignant lesions for all the three feature pools, suggesting its potential as a novel imaging biomarker of skin cancer.

The performance estimated from the classification maps are also consistent with the performance trends observed from the prediction probability maps for both the models. The combined spectral intensity and bi-exponential feature pool resulted in the highest performance in the ROCs constructed from the proportion of malignant pixels (mean ROC-AUC: 0.87 ± 0.02 for random forest and 0.86 ± 0.02 for neural networks). Also, the bi-exponential feature pool (mean ROC-AUC: 0.85 ± 0.03 for random forest and 0.82 ± 0.04 for neural networks) outperformed the spectral intensity feature pool (mean ROC-AUC: 0.82 ± 0.03 for random forest and 0.80 ± 0.03 for neural networks). Image level classification performances obtained after applying Th_{per} also indicate that the combined feature pool produce the best classification performances for both the models. Sensitivity and specificity for the combined feature pool are $74.73\% \pm 14.66\%$ and $76.83\% \pm 9.58\%$ respectively for the random forests model, while the sensitivity and specificity for the neural network model with the combined feature pool are $76.84\% \pm 12.49\%$ and $78.29\% \pm 5.50\%$ respectively. Neural network model provides $\sim 2\%$ improvement in both sensitivity and specificity in comparison to the random forest model. The random forest classifier (Table 5) yielded similar specificity for both intensity and bi-exponential features, while the bi-exponential features yielded superior sensitivity than the intensity features. And when both feature families were used, both sensitivity and specificity increased with respect to models using only one family of features. Similarly, with the neural network classifier (Table 8), bi-exponential feature pool is superior to intensity feature pool in both sensitivity and specificity, while combined feature pool performed best among the three feature pools.

This study also identified specific spectral intensity and bi-exponential fluorescence features that are important for malignant skin lesion discrimination (Fig. 6). Some of these features are associated to the contribution of collagen by itself ($I_{390,n}$) or in relation to the contribution of NADH ($\frac{I_{390,n}}{I_{452,n}}$) to skin lesion autofluorescence [44,75]. In malignant lesions, both collagen degradation and epidermis thickening result in decreased excitation of and detected emission from collagen in the dermis. Increased metabolic activity of neoplastic epithelial cells, on the other hand, results in higher mitochondrial concentration of NADH [44]. Collagen has longer fluorescence lifetime (>3 ns) than NADH (<3 ns). Due to their overlapping emission spectra, collagen and NADH fluorescence emissions are expected in both the 390 ± 20 nm and 452 ± 22.5 nm spectral channels [75]. Thus, faster autofluorescence decays are expected in malignant skin lesions, resulting in specific changes in the values of $\tau_{fast,390}$, $\tau_{fast,452}$, and $\alpha_{fast,390}$.

The performance of the neural network model is better than the random forests model by $\sim 2\%$ in terms of sensitivity and specificity (Tables 5 and 8). The advantage of the random forest models, however, is that they are significantly less complex and provide direct interpretation of important features (Fig. 6). It is worth noting that, in this study, the input of both models was the maFLIM features extracted from the time-resolved fluorescence data; thus, extraction of standard fluorescence features (normalized spectral intensities, bi-exponential decay model parameters) was required to train both models. A potential advantage of neural network models not explored in this study is that they can accept the time-resolved fluorescence data directly as input; thus, maFLIM feature extraction would no longer be required before training.

In this study, we also validated an efficient approach for tuning the hyperparameters of a neural network model by integrating the rigorous cross-validation strategy discussed before with the use of the hyperband algorithm. Through this approach, it was possible to generate histograms of optimal hyperparameters estimated from multiple realizations of the test set (50 in this study). Inspection of these histograms (Figure. S4) can guide the design of an optimal neural network architecture for further validation of the model in a more extensive dataset.

In a previous study, we showed the potential of extracting global maFLIM features for machine-learning based discrimination of malignant from benign pigmented skin lesions [62]. The previous study explored machine learning models that required explicit feature selection prior to training, whereas in this study the models explored do not require feature selection.

The sensitivity and specificity obtained while training ensemble classifiers using a combined set of intensity and bi-exponential global features are 84.21% and 65.85% respectively. In this study, we explored pixel-level maFLIM features which enable developing classification models capable of generating prediction probability maps and classification maps. Compared to the previous work, this work shows an improvement in specificity ($78.29\% \pm 5.50\%$) and a decrease in sensitivity ($76.84\% \pm 12.49\%$). However, unlike the previous publication, this work provides the ability to generate probability maps and classification maps, which can in turn help in identifying regions of malignancy. Overall classification performance in this work may be affected by the lack of pixel-level ground truth labels. For diagnosis purposes (the focus of this study), the spatial information contained in the prediction probability maps can be disregarded, as only a positive or negative skin lesion classification is needed. In such cases, only lesion-level classification results from the classification maps can be considered. On the other hand, for lesion margin detection purposes, the prediction probability maps can potentially provide a direct visualization of the lesion real boundaries. In this study, however, pixel-level ground truth information was not available, which introduces pixel mislabeling in the available data. The effect of pixel mislabeling can be mitigated by using the statistics of the prediction maps when the goal is to provide a lesion-level diagnosis. For lesion margin detection, however, pixel-level ground-truth information will be required to develop models that will generate prediction probability maps reflecting the accurate boundaries of the imaged lesions. Pixel-level ground-truth can be obtained by annotating the resected tissue by a pathologist to identify regions of malignancy following sectioning and histopathology staining [76]. This can be done by superimposing the stained slide with an ex-vivo image to identify landmarks and orientations of the tissue. In this way, ground truth information of several regions within the image can be assigned to the pixels corresponding to those regions within the FOV. Future efforts will explore the applications of maFLIM for lesion margin detection.

4.1. Study limitations and future work

Although this work demonstrates the potential to classify benign and malignant pigmented skin lesions using pixel-level maFLIM features, several limitations are recognized. First, the database of maFLIM images is limited in both the type of benign and malignant skin conditions, and the number of samples per condition. A more comprehensive database is needed to fully develop accurate enough classification methods for skin lesion discrimination, and to rigorously quantify their performance in prospective studies. Second, the lack of histopathology-based assessment of the maFLIM imaging data at the pixel-level prevented to specifically quantify the capabilities of maFLIM dermoscopy as a tool for not only detecting malignant skin lesions, but also determining their true extension and margins. The current implementation processes maFLIM data to generate classification maps of imaged lesions, as well as lesion-level classification. For the classification maps to reflect the accurate mapping of a malignant lesion, these models would need to be retrained and optimized on maFLIM data with pixel-level labeling using the same framework. Third, the current maFLIM dermoscopy system provides nonspecific excitation and spectral detection of skin autofluorescence component emission. Finally, the current implementation of the machine-learning classification models does not allow for real-time processing of maFLIM data. Ongoing research efforts aiming to overcome these limitations include collecting maFLIM dermoscopy images from a plurality of nonpigmented and pigmented skin lesions from patients of various skin tones, performing accurate pixel-level registration between the lesion maFLIM imaging data and histopathology tissue sections, developing improved maFLIM dermoscope systems with multiwavelength excitation capabilities, and implementing real-time maFLIM data processing, pixel-level classification, and tissue mapping visualization. One approach to conduct real-time pixel-level classification is to employ sequence models such as long short-term memory

networks (LSTM) [72,76] to process the fluorescence decay signals directly without explicit feature extraction.

4.2. Clinical perspective

Providing adequate treatment strategies to patients is of utmost importance, especially to those who do not have access to a dermatologist. To this date, there is no objective device that primary care physicians or dermatologists could use to identify suspicious skin cancer lesions. Such a device could help doctors identify patients that do not require biopsy resection and histopathology evaluation, thereby minimizing the number of unwanted painful biopsy procedures. In addition to diagnostics, the pixel-level prediction maps generated using this technique can potentially help identify the tumor extensions and determine adequate margins for tumor excision surgeries. Finally, extending this work with fast processing systems capabilities can enable (near) real-time objective i) lesion diagnosis to assess the need of biopsy and resection, and ii) margin detection to enable complete tumor removal, thus minimizing the chances of recurrence.

5. Conclusions

This study demonstrates the use of pixel-level features extracted from maFLIM dermoscopy data for objective discrimination of malignant from benign pigmented skin lesions. Specific spectral steady-state intensity and bi-exponential autofluorescence were identified as relevant for malignant skin lesion discrimination. Time-resolved bi-exponential autofluorescence features were found to be critical for accurate detection of malignant pigmented skin lesion. The deep neural network model produced the best lesion-level classification, producing a partition-level sensitivity and specificity of $76.84\% \pm 12.49\%$ and $78.29\% \pm 5.50\%$ respectively. Pixel-level maFLIM enables developing classification models capable of generating prediction probability maps and classification maps, which were successfully used to provide objective diagnosis of pigmented skin lesions. Future efforts will explore the use of pixel-level maFLIM features for malignant skin lesion margin detection.

Funding. National Institutes of Health (1R01CA218739, 1R21CA269099, 5P20GM135009, R01GM127696, R01GM152633, R21GM142107); Cancer Prevention and Research Institute of Texas (RP180588); Fundação de Amparo à Pesquisa do Estado de São Paulo (2013/07276-1 (CEPOF), 2014/50857-8 (INCT)); Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CNPq-PQ (305795/2016-3), CNPq-PVE (401150/2014-3 and 314533/2014-1)).

Acknowledgements. This study was supported by the National Institutes of Health (grants 1R01CA218739, R01GM127696, R01GM152633, R21GM142107, 1R21CA269099) and the Cancer Prevention and Research Institute of Texas (grant RP180588). This study was also supported by the following Brazilian funding agencies: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001; CNPq-PVE (401150/2014-3 and 314533/2014-1); CNPq-PQ (305795/2016-3) and São Paulo Research Foundation (FAPESP) grants: 2013/07276-1 (CEPOF); 2014/50857-8 (INCT). This work was also partially funded by the Air Force Office of Scientific Research (AFOSR) (FA9550-20-1-0366, FA9550-20-1-0367, FA9550-23-1-0599), and by NASA, BARDA, NIH, and USFDA, under Contract/Agreement No. 80ARC023CA002. Research reported in this publication was also supported in part by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences from NIH, under grant number P20 GM135009.

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Supplemental document. See [Supplement 1](#) for supporting content.

References

1. M. Perez, J. A. Abisaad, K. D. Rojas, *et al.*, "Skin cancer: Primary, secondary, and tertiary prevention. Part I," *J. Am. Acad. Dermatol.* **87**(2), 255–268 (2022).
2. P. Aggarwal, P. Knabel, and A. B. Fleischer, "United States burden of melanoma and non-melanoma skin cancer from 1990 to 2019," *J. Am. Acad. Dermatol.* **85**(2), 388–395 (2021).
3. R. L. Siegel, K. D. Miller, N. S. Wagle, *et al.*, "Cancer statistics, 2023," *Ca-Cancer J. Clin.* **73**(1), 17–48 (2023).

4. H. Kittler, H. Pehamberger, K. Wolff, *et al.*, "Diagnostic accuracy of dermoscopy," *Lancet Oncol.* **3**(3), 159–165 (2002).
5. B. Rosado, S. Menzies, A. Harbauer, *et al.*, "Accuracy of computer diagnosis of melanoma: A quantitative meta-analysis," *Arch. Dermatol.* **139**(3), 361–367 (2003).
6. M. E. Celebi and A. Zornberg, "Automated quantification of clinically significant colors in dermoscopy images and its application to skin lesion classification," *IEEE Syst J* **8**(3), 980–984 (2014).
7. M. Seyed Ahadi, A. Firooz, H. Rahimi, *et al.*, "Clinical Diagnosis has a High Negative Predictive Value in Evaluation of Malignant Skin Lesions," *Dermatol Res Pract* **2021**, 1 (2021).
8. M. P. Lee, J. F. Sobanko, T. M. Shin, *et al.*, "Evolution of Excisional Surgery Practices for Melanoma in the United States," *JAMA Dermatol* **155**(11), 1244–1251 (2019).
9. A. Kimyai-Asadi, T. Katz, L. H. Goldberg, *et al.*, "Margin involvement after the excision of melanoma in situ: The need for complete en face examination of the surgical margins," *Dermatol. Surg.* **33**(12), 1434–1441 (2007).
10. M. E. Dawn, A. G. Dawn, and S. J. Miller, "Mohs surgery for the treatment of melanoma in situ: A review," *Dermatol. Surg.* **33**(4), 395–402 (2007).
11. J. H. Kunishige, D. G. Brodland, and J. A. Zitelli, "Surgical margins for melanoma in situ," *J. Am. Acad. Dermatol.* **66**(3), 438–444 (2012).
12. C. G. Ethun and K. A. Delman, "The importance of surgical margins in melanoma," *J. Surg. Oncol.* **113**(3), 339–345 (2016).
13. N. Agarwal-Antal, G. M. Bowen, and J. W. Gerwels, "Histologic evaluation of lentigo maligna with permanent sections: Implications regarding current guidelines," *J. Am. Acad. Dermatol.* **47**(5), 743–748 (2002).
14. G. Tchernev, V. Malev, J. W. Patterson, *et al.*, "A novel surgical margin (1 cm) might be from benefit for patients with dysplastic nevi, thin melanomas, and melanoma in situ: Analysis based on clinical cases," *Dermatol Ther* **33**(2), 1 (2020).
15. S. M. Swetter, J. A. Thompson, M. R. Albertini, *et al.*, "NCCN Guidelines® insights: melanoma: cutaneous, version 2.2021: featured updates to the NCCN guidelines," *J. Natl. Compr. Cancer Network* **19**(4), 364–376 (2021).
16. S. Cheraghlou, S. R. Christensen, G. O. Agogo, *et al.*, "Comparison of Survival after Mohs Micrographic Surgery vs Wide Margin Excision for Early-Stage Invasive Melanoma," *JAMA Dermatol* **155**(11), 1252–1259 (2019).
17. C. L. F. Temple and J. P. Arlette, "Mohs micrographic surgery in the treatment of lentigo maligna and melanoma," *J. Surg. Oncol.* **94**(4), 287–292 (2006).
18. V. G. Prieto, Z. B. Argenyi, R. L. Barnhill, *et al.*, "Are en face frozen sections accurate for diagnosing margin status in melanocytic lesions?" *Am J Clin Pathol* **120**(2), 203–208 (2003).
19. L. Rey-Barroso, S. Peña-Gutiérrez, C. Yáñez, *et al.*, "Optical technologies for the improvement of skin cancer diagnosis: A review," *Sensors* **21**(1), 1–31 (2021).
20. M. A. Calin, S. V. Parasca, R. Savastru, *et al.*, "Optical techniques for the noninvasive diagnosis of skin cancer," *J Cancer Res Clin Oncol* **139**(7), 1083–1104 (2013).
21. P. Calzavara-Pinton, C. Longo, M. Venturini, *et al.*, "Reflectance confocal microscopy for in vivo skin imaging," *Photochem. Photobiol.* **84**(6), 1421–1430 (2008).
22. A. Levine, K. Wang, and O. Markowitz, "Optical Coherence Tomography in the Diagnosis of Skin Cancer," *Dermatol. Clin.* **35**(4), 465–488 (2017).
23. X. Feng, M. C. Fox, J. S. Reichenberg, *et al.*, "Superpixel Raman spectroscopy for rapid skin cancer margin assessment," *J Biophotonics* **13**(2), 1–7 (2020).
24. M. Troyanova-Wood, Z. Meng, and V. V. Yakovlev, "Differentiating melanoma and healthy tissues based on elasticity-specific Brillouin microspectroscopy," *Biomed. Opt. Express* **10**(4), 1774 (2019).
25. R. A. Romano, R. G. T. Rosa, J. A. Jo, *et al.*, "Label-free multispectral lifetime fluorescence to distinguish skin lesions," *Proc. SPIE* **10890**, 93 (2019).
26. J. P. Miller, L. Habimana-Griffin, T. S. Edwards, *et al.*, "Multimodal fluorescence molecular imaging for in vivo characterization of skin cancer using endogenous and exogenous fluorophores," *J. Biomed. Opt.* **22**(6), 066007 (2017).
27. S. A. Alawi, M. Kuck, C. Wahrlich, *et al.*, "Optical coherence tomography for presurgical margin assessment of non-melanoma skin cancer - a practical approach," *Exp Dermatol* **22**(8), 547–551 (2013).
28. C. C. Horgan, M. S. Bergholt, M. Z. Thin, *et al.*, "Image-guided Raman spectroscopy probe-tracking for tumor margin delineation," *J. Biomed. Opt.* **26**(03), 1 (2021).
29. B. Park, C. H. Bang, C. Lee, *et al.*, "3D wide-field multispectral photoacoustic imaging of human melanomas in vivo: a pilot study," *Journal of the European Academy of Dermatology and Venereology* **35**(3), 669–676 (2021).
30. M. T. Stridh, J. Hult, A. Merdasa, *et al.*, "Photoacoustic imaging of periorbital skin cancer ex vivo: unique spectral signatures of malignant melanoma, basal, and squamous cell carcinoma," *Biomed. Opt. Express* **13**(1), 410 (2022).
31. R. Leon, B. Martinez-Vega, H. Fabelo, *et al.*, "Non-invasive skin cancer diagnosis using hyperspectral imaging for in-situ clinical support," *J. Clin. Med.* **9**(6), 1662 (2020).
32. T. H. Johansen, K. Møllersen, S. Ortega, *et al.*, "Recent advances in hyperspectral imaging for melanoma detection," *Wiley Interdiscip Rev Comput Stat* **12**(1), 1 (2020).
33. L. Liu, M. Qi, Y. Li, *et al.*, "Staging of Skin Cancer Based on Hyperspectral Microscopic Imaging and Machine Learning," *Biosensors* **12**(10), 790 (2022).
34. H. Y. Huang, Y. P. Hsiao, A. Mukundan, *et al.*, "Classification of Skin Cancer Using Novel Hyperspectral Imaging Engineering via YOLOv5," *J. Clin. Med.* **12**(3), 1 (2023).

35. R. A. Romano, R. G. Teixeira Rosa, A. G. Salvio, *et al.*, "Multispectral autofluorescence dermoscope for skin lesion assessment," *Photodiagn. Photodyn. Ther.* **30**, 1134 (2020).
36. H. Zeng, D. I. McLean, C. E. MacAulay, *et al.*, "Autofluorescence properties of skin and applications in dermatology," in *Biomedical Photonics and Optoelectronic Imaging* (SPIE, 2000), 4224, pp. 366–373.
37. Y. Wu and J. Y. Qu, "Autofluorescence spectroscopy of epithelial tissues," *J. Biomed. Opt.* **11**(5), 054023 (2006).
38. O. I. Kolenc and K. P. Quinn, "Evaluating cell metabolism through autofluorescence imaging of NAD(P)H and FAD," *Antioxid. Redox Signaling* **30**(6), 875–889 (2019).
39. R. Na, I.-M. Stender, L. Ma, *et al.*, "Autofluorescence spectrum of skin: component bands and body site variations," *Skin Research and Technology* **6**(3), 112–117 (2000).
40. A. J. Walsh, R. S. Cook, H. C. Manning, *et al.*, "Optical metabolic imaging identifies glycolytic levels, subtypes, and early-treatment response in breast cancer," *Cancer Res* **73**(20), 6164–6174 (2013).
41. M. C. Skala, K. M. Riching, A. Gendron-Fitzpatrick, *et al.*, "In vivo multiphoton microscopy of NADH and FAD redox states, fluorescence lifetimes, and cellular morphology in precancerous epithelia," *Proc. Natl. Acad. Sci. U.S.A.* **104**(49), 19494–19499 (2007).
42. M. G. vander Heiden, L. C. Cantley, and C. B. Thompson, "Understanding the Warburg Effect: The Metabolic Requirements of Cell Proliferation," *Science* **324**(5930), 1029–1033 (2009).
43. A.-M. Pena, M. Boulade, S. Brizion, *et al.*, "Multiphoton FLIM imaging of NADH and FAD to analyze cellular metabolic activity of reconstructed human skin in response to UVA light," in *(SPIE-Intl Soc Optical Eng)*, 2019, p. 9.
44. I. Georgakoudi, B. C. Jacobson, M. G. Müller, *et al.*, "NAD(P)H and Collagen as in Vivo Quantitative Fluorescent Biomarkers of Epithelial Precancerous Changes," *Cancer Res* **62**, 682–687 (2002).
45. I. Giovannacci, C. Magnoni, P. Vescovi, *et al.*, "Which are the main fluorophores in skin and oral mucosa? A review with emphasis on clinical applications of tissue autofluorescence," *Arch. Oral Biol.* **105**, 89–98 (2019).
46. W. Lohmann and R. H. Bodeker, "In Situ Differentiation Between Nevi and Malignant Melanomas by Fluorescence Measurements," *Naturwissenschaften* **78**(10), 456–457 (1991).
47. A. Fast, A. Lal, A. F. Durkin, *et al.*, "Fast, large area multiphoton exoscope (FLAME) for macroscopic imaging with microscopic resolution of human skin," *Sci. Rep.* **10**(1), 18093 (2020).
48. L. Marcu, "Fluorescence lifetime techniques in medical applications," *Ann. Biomed. Eng.* **40**(2), 304–331 (2012).
49. V. Huck, C. Gorzelanny, K. Thomas, *et al.*, "From morphology to biochemical state - intravital multiphoton fluorescence lifetime imaging of inflamed human skin," *Sci. Rep.* **6**(1), 22789 (2016).
50. L. Pires, M. S. Nogueira, S. Pratavieira, *et al.*, "Time-resolved fluorescence lifetime for cutaneous melanoma detection," *Biomed. Opt. Express* **5**(9), 3080 (2014).
51. M. N. Pastore, H. Studier, C. S. Bonder, *et al.*, "Non-invasive metabolic imaging of melanoma progression," *Exp Dermatol* **26**(7), 607–614 (2017).
52. P. A. A. de Beule, C. Dunsby, N. P. Galletly, *et al.*, "A hyperspectral fluorescence lifetime probe for skin cancer diagnosis," *Rev. Sci. Instrum.* **78**(12), 1 (2007).
53. M. A. Kassem, K. M. Hosny, R. Damaševičius, *et al.*, "Machine learning and deep learning methods for skin lesion classification and diagnosis: A systematic review," *Diagnostics* **11**(8), 1390 (2021).
54. M. Dildar, S. Akram, M. Irfan, *et al.*, "Skin cancer detection: A review using deep learning techniques," *Int J Environ Res Public Health* **18**(10), 5479 (2021).
55. A. G. C. Pacheco and R. A. Krohling, "Recent advances in deep learning applied to skin cancer detection," (2019).
56. K. Ramlakhan and Y. Shang, "A mobile automated skin lesion classification system," in *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence. IEEE* (IEEE, 2011), pp. 138–141.
57. T. Y. Sathesha, D. Satyanarayana, M. N. G. Prasad, *et al.*, "Melanoma Is Skin Deep: A 3D Reconstruction Technique for Computerized Dermoscopic Skin Lesion Classification," *IEEE J. Transl. Eng. Health Med.* **5**, 1–17 (2017).
58. Y. A. Khristoforova, I. A. Bratchenko, D. N. Artemyev, *et al.*, "Optical diagnostics of malignant and benign skin neoplasms," *Procedia Eng.* **201**, 141–147 (2017).
59. J. Amin, A. Sharif, N. Gul, *et al.*, "Integrated design of deep features fusion for localization and classification of skin cancer," *Pattern Recognit Lett* **131**, 63–70 (2020).
60. A. Romero Lopez, X. Giro-I-Nieto, J. Burdick, *et al.*, "Skin lesion classification from dermoscopic images using deep learning techniques," in *017 13th IASTED International Conference on Biomedical Engineering (BioMed). IEEE* (International Association of Science and Technology for Development-IASTED, 2017), pp. 49–54.
61. M. F. Jojoa Acosta, L. Y. Caballero Tovar, M. B. Garcia-Zapirain, *et al.*, "Melanoma diagnosis using deep learning techniques on dermoscopic images," *BMC Med. Imaging* **21**(1), 6 (2021).
62. P. Vasanthakumari, R. A. Romano, R. G. T. Rosa, *et al.*, "Discrimination of cancerous from benign pigmented skin lesions based on multispectral autofluorescence lifetime imaging dermoscopy and machine learning," *J. Biomed. Opt.* **27**(06), 1–26 (2022).
63. R. Henderson and K. Schulmeister, *Laser Safety* (Taylor & Francis, 2004), 53(9).
64. J. R. Lakowicz, *Principles of Fluorescence Spectroscopy* (Springer, 2006).
65. J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J Big Data* **6**(1), 27 (2019).
66. B. H. Menze, B. M. Kelm, R. Masuch, *et al.*, "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data," *BMC Bioinformatics* **10**(1), 213 (2009).
67. J. Moolayil, *Learn Keras for Deep Neural Networks* (Apress, 2019).

68. D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," [arXiv](#), (2014).
69. U. Ruby, P. Theerthagiri, J. Jacob, *et al.*, "Binary cross entropy with deep learning technique for Image classification," [International Journal of Advanced Trends in Computer Science and Engineering](#) **9**(4), 5393–5397 (2020).
70. L. Li, K. Jamieson, A. Rostamizadeh, *et al.*, "Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization," *Journal of Machine Learning Research* **18**, 1–52 (2018).
71. P. Vasanthakumari, R. A. Romano, R. G. T. Rosa, *et al.*, "AI-driven discrimination of benign from malignant pigmented skin lesions based on multispectral autofluorescence lifetime dermoscopy imaging," in *Proc. SPIE Photonics in Dermatology and Plastic Surgery*, B. Choi and H. Zeng, eds. (SPIE, 2022), p. 1193408.
72. P. Vasanthakumari, R. A. Romano, R. G. Teixeira Rosa, *et al.*, "Classification of skin-cancer lesions based on Fluorescence Lifetime Imaging," in *Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging* (International Society for Optics and Photonics, 2020), p. 34.
73. R. Prasanna Kumar, D. Melcher, P. Buttolo, *et al.*, "Vehicle Seat Occupancy Detection and Classification Using Capacitive Sensing," in *SAE Technical Papers* (SAE International, 2024). doi:10.4271/2024-01-2508
74. E. Duran-Sierra, S. Cheng, R. Cuenca-Martinez, *et al.*, "Clinical label-free biochemical and metabolic fluorescence lifetime endoscopic imaging of precancerous and cancerous oral lesions," *Oral Oncol.* **105**, 104635 (2020).
75. M. Marsden, B. W. Weyers, J. Bec, *et al.*, "Intraoperative Margin Assessment in Oral and Oropharyngeal Cancer Using Label-Free Fluorescence Lifetime Imaging and Machine Learning," *IEEE Trans. Biomed. Eng.* **68**(3), 857–868 (2021).
76. R. Prasanna Kumar, D. Melcher, P. Buttolo, *et al.*, "Tracking Occupant Activities in Autonomous Vehicles Using Capacitive Sensing," *IEEE Trans. Intell. Transport. Syst.* **24**(7), 6800–6819 (2023).