Nearest Neighbor-Based Strategy to Optimize Multi-View Triplet Network for Classification of Small-Sample Medical Imaging Data

Phawis Thammasorn, Wanpracha A. Chaovalitwongse, *Senior Member, IEEE*, Daniel S. Hippe, Landon S. Wootton, Eric C. Ford, Matthew B. Spraker, Stephanie E. Combs, Jan C. Peeken, *Member, IEEE*, and Matthew J. Nyflot

***Abstract*— Multi-view classification with limited sample size and data augmentation is a very common machine learning (ML) problem in medicine. With limited data, a triplet network approach for two-stage representation learning has been proposed. However, effective training and verifying the features from the representation network for their suitability in subsequent classifiers are still unsolved problems. Although typical distance-based metrics for the training capture the overall class separability of the features, the performance according to these metrics does not always lead to an optimal classification. Consequently, an exhaustive tuning with all feature–classifier combinations is required to search for the best end result. To overcome this challenge, we developed a novel nearest-neighbor (NN) validation strategy based on the triplet metric. This strategy is supported by a theoretical foundation to provide the best selection of the features with a lower bound of the highest end performance. The proposed strategy is a transparent approach to identify whether to improve the features or the classifier. This avoids the need for repeated tuning. Our evaluations on real-world medical imaging tasks (i.e., radiation therapy delivery error prediction and sarcoma survival prediction) show that our strategy is superior to other common deep representation learning baselines [i.e., autoencoder (AE) and softmax]. The strategy addresses the issue of feature's interpretability which enables more holistic feature creation such that the medical experts can focus on specifying relevant data as opposed to tedious feature engineering.**

***Index Terms*— Medical data classification, multi-view learning, representation learning, transfer learning metric learning.**

I. Introduction

CLINICIANS often consider medical information from all data sources available (e.g., blood tests, patient history, clinical physiology) before making clinical decisions. Similarly, machine learning (ML) systems designed for medical decision support also need to incorporate data from various clinical sources. The sources range from medical scans of different modalities [1] to a list of expert-defined variables [2] in combination with other patient data and profiles. In ML research, the context of using such diverse information from complex sources and data definitions is commonly studied under the topic of multi-view learning [36], [37]. An increasing number of widely used approaches lie in the utilization of deep learning architectures to engineer and enrich fusional features. A major difficulty for the approaches under the clinical settings is a lack of the available training samples due to several factors such as restrictive patient privacy laws and procedures in data acquisition/access, extreme heterogeneity of clinical settings across institutions and patient cohorts, high medical imaging study costs, and low numbers of patient enrolment in the study. Moreover, traditional data augmentation and generation techniques are ineffective because the distributions of the patient data are often unknown or hard to verify. Under such scarcity, transfer learning via representation is an alternative strategy to train the architecture for an actual prediction task.

Three approaches are commonly used for representation transfer: end-to-end transfer training, autoencoders (AEs), and metric learning. While all the approaches have found successes in diverse domains [3]–[5], previous studies suggested that metric learning has the potential for the small-sample problem [6], [7]. Triplet network is an applicable architecture under the metric approach trained for extracting class-separable features. Training the network for the best classification features is, however, a tedious task. Although the training for the class-separable features is generally sensible, there is a lack of effective strategies to translate the metric loss to an achievable

Manuscript received September 29, 2019; revised April 10, 2020, October 3, 2020, and December 26, 2020; accepted February 2, 2021. This work was supported by the National Science Foundation (NSF) Award # 1734913. *(Corresponding author: Phawis Thammasorn.)*

Phawis Thammasorn and Wanpracha A. Chaovalitwongse are with the Department of Industrial Engineering, University of Arkansas, Fayetteville, AR 72701 USA (e-mail: pthammas@uark.edu; artchao@uark.edu).

Daniel S. Hippe is with the Department of Radiology, University of Washington, Seattle, WA 98195 USA (e-mail: dhippe@uw.edu).

Landon S. Wootton, Eric C. Ford, and Matthew J. Nyflot are with the Department of Radiation Oncology, University of Washington, Seattle, WA 98195 USA (e-mail: landsw@uw.edu; eford@uw.edu; nyflot@uw.edu).

Matthew B. Spraker is with the Department of Radiation Oncology, Washington University at St. Louis, St. Louis, MO 63130 USA (e-mail: mspraker@wustl.edu).

Stephanie E. Combs and Jan C. Peeken are with the Department of Radiation Oncology, Technical University of Munich (TUM), 81675 Munich, Germany, and also with the Department of Radiation Sciences (DRS), Institute of Radiation Medicine (IRM), Helmholtz Zentrum München, 85764 Neuherberg, Germany (e-mail: stephanie.combs@tum.de; jan.peeken@tum.de).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2021.3059635.

Digital Object Identifier 10.1109/TNNLS.2021.3059635

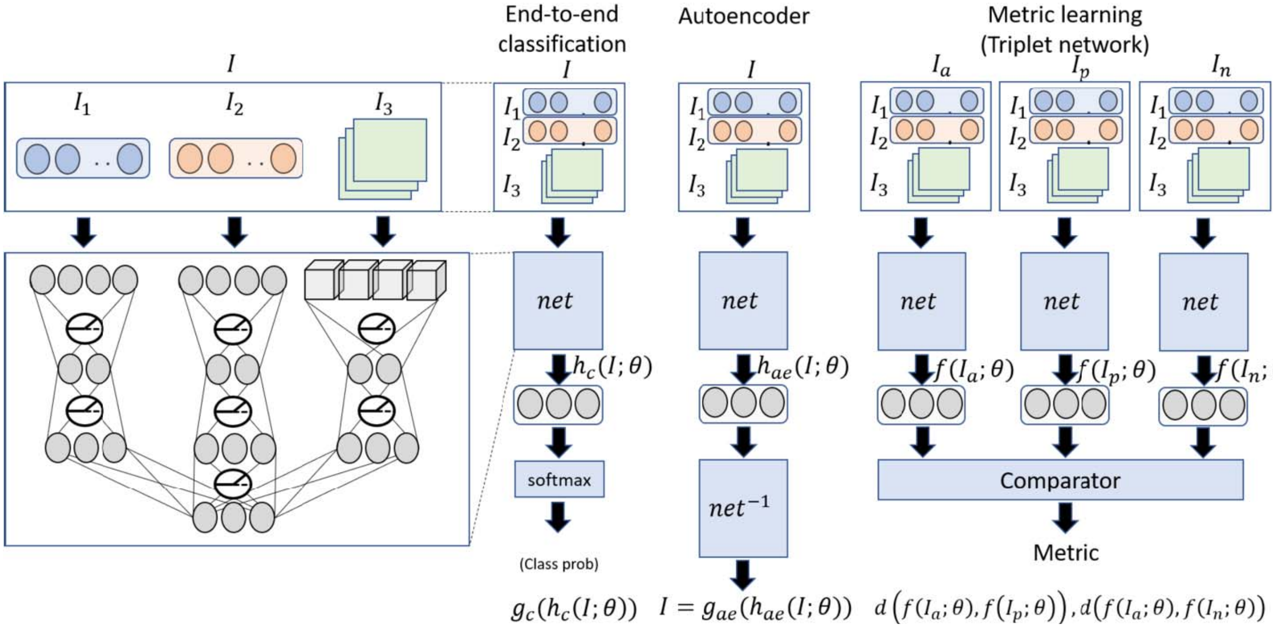

Fig. 1. Comparison of the three commonly used deep learning architectures as the embedding network for the two-stage multi-view representation learning strategy. From left to right: end-to-end classification network, AE network, and triplet network for metric learning.

classification result such that similar losses among the training epochs may provide features with drastically different classification performances. One needs to exhaustively tune and compare all possible feature–classifier combinations to decide the best classification architecture. Consequently, it is also difficult to determine whether underperformance was caused by the trained features, the classifier, or both.

This article presents a novel strategy to optimize the training of the triplet network with theoretical support to guarantee the quality of training. The strategy uses key information from the separation metric loss to fine-tune hyperparameters of an adaptive nearest-neighbor (NN) validation, which evaluates a lower bound on the achievable end performance. The validation guarantees that the features with the best classification potential would be selected while providing an achievable target result. This approach can overcome the need of the repeated tunings on both the features and the subsequent final classifiers. To demonstrate the effectiveness of the proposed strategy, this study applied it to two real-world medical data classification problems: quality assurance of radiation therapy delivery and sarcoma patient survival. Both real-world data sets have small sample sizes (e.g., <200 sample/class) and multi-view inputs.

II. Related Works

A. Representation Learning Strategies

Given multi-source raw input data $I = (I^1, I^2, \ldots, I^s)$ where s is the number of information sources, the purpose of representation learning is to extract a fusional feature embedding $X = f(I)$ for subsequent prediction tasks. In the past, the features were often obtained through time-consuming feature engineering. Recent approaches, however, use some forms of parameterized deep networks trained with related data or surrogate tasks. Training and using the networks are often organized as two-stage operations. Intuitively, the first stage trains the network for surrogate tasks or helper tasks. Through surrogate training, the network learns to extract X with some inductive bias [40] useful for solving the target tasks in the second stage. Fig. 1 illustrates a conceptual difference between the architectures of the three commonly used representation learning strategies.

The end-to-end transfer learning organizes the feature extraction as a by-product of an end-to-end training (e.g., using output from the penultimate layer as features). Specifically, the strategy uses $f(I) = h_{\mathrm{c}}(I;\theta)$ where θ is a set of parameters and $h_{\mathrm{c}}(I;\theta)$ is from the end-to-end network $g_{\mathrm{c}}(h_{\mathrm{c}}(I);\theta)$ tuned for the surrogate task. To train for the second stage, $g_{\mathrm{c}}(X;\theta)$ is subsequently fine-tuned (e.g., soft-max) or replaced by other classifiers. The end-to-end strategy is common in many domains, especially for Computer Vision where pre-trained networks from large-scale image data sets [8], [9] are available for surrogate learning. There are also some adaptations to the feature fusion problems in the medical domain [10]–[12].

The AE approach trains an architecture for a latent representation using an unsupervised self-reconstruction as the surrogate task. An intermediate layer of the trained network is used as the low-dimensional representation extractor. The main intuition for using the layer is to capture a manifold of compressed patterns with reduced complexity to ease the second-stage training. Specifically, an AE network is trained to reconstruct input $I = g_{\mathrm{ae}}(h_{\mathrm{ae}}(I;\theta))$. After training, $g_{\mathrm{ae}}(X;\theta)$ for the reconstruction is removed. Then, the latent representation extractor $f(I) = h_{\mathrm{ae}}(I;\theta)$ is designated from its hidden layer $h_{\mathrm{ae}}(I;\theta)$ to extract features from the multi-format I for the second-stage classification task. The approach has found some success in signal processing and classification in the medical domain [13]–[15]. However, training the AE networks to reduce the reconstruction loss can be less effective with diverse forms of input (e.g., more loss from larger inputs).

In contrast to the first two approaches, the metric learning approach trains the feature extraction architecture to satisfy a suitability metric (e.g., loss function) such that the extracted features can ease the subsequent tasks. Triplet network is an example of this approach that trains for general class separation within the feature space. The class separability distance metric was introduced in [16] and [17]. The distance function $d(f(I_{\rm i};\theta), f(I_{\rm j};\theta))$ is maximized if $I_{\rm i}$ and $I_{\rm j}$ belong to different classes and is minimized if $I_{\rm i}$ and $I_{\rm j}$ belong to the same class. After training, $X = f(I;\theta)$ is used for subsequent tasks. The approach has been applied generally to image retrieval [19], [20] and adapted for medical classification [21].

The literature suggests that X with better surrogate performance implies better target results. However, the suitability of X in terms of the target performance has not been quantified during the first-stage training. To train the system efficiently, prior quantification is necessary for identifying whether the prediction with X should be improved at the first stage (features) or the second stage (classifiers). Otherwise, all possible feature–classifier candidates have to be tested. We propose to avoid the repetition such that only the features with the best potential are selected for target task training.

B. Triplet Network

In this work, we develop the validation based on information during the triplet network training. Triplet network is an architecture under the metric learning approach. The network consists of three identical extractors with a shared parameter set θ and a comparator network [7]. The choice of the extracting architecture is selected based on the input data characteristics, such as feed-forward networks for 1-D vector data and convolutional neural networks (CNNs) for 2-D data. The extractors create feature vectors for an anchor input $I_{\rm a}$, a same = class or positive input $I_{\rm p}$, and a different-class or negative input $I_{\rm n}$, which results in $X_{\rm a}$, $X_{\rm p}$, and $X_{\rm n}$, respectively. Afterward, the comparator evaluates the vectors and backpropagates the error gradient to adjust θ. The general idea of the metric loss criteria was originally proposed in [16], where the features of same-class samples should be clustered together and positioned away from that of the different classes in the feature space. The idea is simplified in subsequent works as fixed-margin triplet loss

$$\mathcal{L}_{\rm tri} = \max\big(0, D_{\rm a,p} - D_{\rm a,n} + m\big) \tag{1}$$

where $D_{\rm a,p} = \|X_{\rm a} - X_{\rm p}\|^2$, $D_{\rm a,n} = \|X_{\rm a} - X_{\rm n}\|^2$, and m is the fixed margin. Minimizing the loss results in a separation of at least m distance among samples of different classes.

With this architecture, the input data need to be organized into a set of triplets. Given $\mathbf{I} = \{I_1, I_2, \ldots\}$ and $\mathbf{C} = \{C_1, C_2, \ldots\}$ as the input data set and its corresponding class, a triplet is organized as $T = (I_{\rm a}, I_{\rm p}, I_{\rm n})$ where $I_{\rm a}$ is any sample in $\mathbf{I}$, and $I_{\rm p}$ and $I_{\rm n}$ are sampled based on their class relationship with $I_{\rm a}$. After training with metric loss, one of the extractors is then used to calculate for $X = f(I;\theta)$. Triplet sampling results in a larger input data set. For a two-class example, let n be the number of input sample, p be the number of positive among the n inputs, and $n - p$ be the number of negative. Let class 1 be positive class. For each $I_{\rm a}$ from class 1, the number of triplet combinations is $T_{\rm p} = p \times p \times (n - p)$. For class 2, the number is $T_{\rm n} = (n - p) \times (n - p) \times p$. Thus, the total number in the triplet data set is $T_{\rm p}+T_{\rm n} = n^2p - np^2$ or $O(n^2)$ expansion from the original size. Note that the increase in the inputs for the network does not increase the number of the network parameters due to parameter sharing. Therefore, it is also an alternative to data augmentation, which is limited in the medical context as it increases the risk of learning with invalid data. Applying the expansion also leads to better classification [6], [7], [18] for small-sample setting.

C. Nearest Neighbor Estimation and Classification

Our validation criteria are formulated based on NN estimation. NN is a classical approach to estimate the local distribution of relevant value $p(V \mid X)$, where V is either a discrete class label C for classification or a continuous response R for regression. Given known feature–value pairs $\{(X_1, V_1), (X_2, V_2), \ldots\}$ and a query with unknown value $(X_{\rm q}, V_{\rm q})$, $p(V_{\rm q} \mid X_{\rm q})$ can be estimated as

$$p\big(V_{\rm q} = v \mid X_{\rm q}\big) = \sum_{{\rm i}\in\mathcal{N}_{\rm q,r_q}} \frac{\mathbb{1}(V_{\rm i} = v)}{|\mathcal{N}_{\rm q,r_q}|}. \tag{2}$$

Once $p(V_{\rm q} = v \mid X_{\rm q})$ is estimated, $V_{\rm q}$ can be decided for a classification task in which $V_{\rm q} = C_{\rm q}$ and

$$C_{\rm q} = \operatorname*{argmax}_{v}\, p\big(C_{\rm q} = v \mid X_{\rm q}\big) \tag{3}$$

or for a regression task in which $V_{\rm q} = R_{\rm q}$ and

$$R_{\rm q} = \sum_{{\rm i}\in\mathcal{N}_{\rm q,r_q}} v_{\rm i}\, p\big(V_{\rm i} \mid X_{\rm q}\big) \tag{4}$$

where $r_{\rm q}$ and $\mathcal{N}_{\rm q,r_q}$ are, respectively, a radius and a set of sample indices belonging to the neighborhood in the estimation such that any known sample $X_{\rm i}$ located within the distance $r_{\rm q}$ away from $X_{\rm q}$ is considered a neighbor of $X_{\rm q}$. i belongs to the set of indices $\mathcal{N}_{\rm q,r_q}$. $r_{\rm q}$ is also a hyperparameter that defines the neighborhood boundary and indirectly determines the number of neighbors for estimation. Note that $r_{\rm q}$ can also be set as $d_{\rm q,k}$ or the distance to kth NN of $X_{\rm q}$. The estimation under the setting is called k-nearest-neighbor (KNN) estimation. In theory, increasing value of k for $r_{\rm q} = d_{\rm q,k}$ leads to more confidence in the estimation as the upper bound estimation error decreases according to the law of large number [24]. In practice, however, it also detrimentally increases the chance of including samples that do not belong to $p(V_{\rm q} \mid X_{\rm q})$. Determining the best value of $r_{\rm q}$ or k remains largely open research. The typical ways involve cross-validation or separate optimization [22]–[23], [34]–[35]. The NN is also a suitable nonparametric tool for non-linear analysis of deep learning features [21], [25].

III. Proposed Method

This article proposes a new strategy to optimize the training of the triplet network for classification. While the typical metric loss roughly reflects the degree of class separation, the difficulty in translating the first-stage loss to classification performance leads to ambiguity whether improvement should be emphasized on representation learning or the classifier training. Due to the need for translation, we introduce a new adaptive NN criterion as the validation in the triplet network training. The proposed strategy can be used for comparison of candidate features prior to the second stage.

A. Using NN Performance as Metric Loss in Triplet Training

The key idea of the strategy is using the empirical performance of NN classification with an appropriate hyper-parameter as a quality measure of features and an achievable baseline. NN classification has close ties with Bayes error rate E_{bayes} or the minimum error possible for a distribution of input [24]. With a sizable K in KNN, it can be established that

$$E_{\text{bayes}} \leq \hat{E}_{\text{KNN}} \leq E_{\text{KNN}} \leq E_{\text{NN}} \leq 2E_{\text{bayes}} \tag{5}$$

or equivalently

$$A_{\text{NN}} \leq A_{\text{KNN}} \leq \hat{A}_{\text{KNN}} \leq A_{\text{bayes}} \tag{6}$$

where E_{KNN} is the upper bound of KNN error, E_{NN} is the upper bound for 1-NN, $A = 1 - E$ is the lower bound of accuracy rate, and $\hat{E}_{\text{KNN}}$ and $\hat{A}_{\text{KNN}}$ are the empirical error and accuracy, respectively.

Asserting $\hat{A} \approx A_{\text{bayes}}$ using the validation data set, the empirical measure suggests that the features with better $\hat{A}$ may perform better when optimally classified. Asserting $\hat{A} \leq A_{\text{bayes}}$ using the testing data set, the empirical measure suggests that there is some room for improving the classification result with an appropriate subsequent process. Thus, the measure can select features from a first-stage training epoch that attains the best $\hat{A}$. Then, the subsequent classifiers should be trained with the selected features for end performance. The second-stage classifier and the validated features with the best performance can then be designated for the final classification framework. These validation steps can replace the training with exhaustive combinations of feature–classifier candidates.

B. Adaptive Neighbor Scope for Validation

The quality of the empirical estimation $\hat{A}$ depends on the appropriate number of neighbors. However, setting such a hyper-parameter is often done by repeated tuning processes which we try to avoid. The issue can be alleviated by setting the hyper-parameter according to information from the representation learning step. A naive method, instead of setting a fixed k value, is to set the neighbor radius $r_{\text{q}} = m$. However, the setting may detrimentally lead to finding less or no neighbor, which will be discussed in Section III-C.

To overcome this challenge, we propose a new adaptive neighborhood scope to determine the radius r_{q} for each query point using a closed-form evaluation on the first-stage metrics. After completing the representation training, our approach calculates an adaptive neighbor radius for each represented X_{a} in the training set. Then, each radius is used to approximate the radius for each query. The key idea for finding the radius on each X_{a} is to start the NN estimation from a large radius and then reduce the search radius based on class distribution within the larger neighborhood. The reduction of search scope is done such that the new neighborhood is more homogeneous. All steps of the proposed strategy are summarized in Algorithm 1.

The first step is to record an arbitrarily large value for an initial search radius r_{a} of each X_{a}. We set $r_{\text{a}} = d_{\text{a},k}$ where k is a sizable value such that the initial neighborhood area contains sufficient samples. The second step is to use the radius to find all positive and negative samples within the neighborhood of X_{a} in the validation set. The third step is to collect local statistics $\bar{D}_{\text{a,p}}$ and $\bar{D}_{\text{a,n}}$, which are the means of distances to the positive samples and the negative samples within the neighborhood defined by $\mathcal{N}_{\text{a,r}_{\text{a}}}$. The fourth step is to use the statistics to suggest a better boundary distance r^* as

$$r^* = \frac{\bar{D}_{\text{a,p}} + \sqrt{\left(\bar{D}_{\text{a,p}}\right)^2 + 8\bar{D}_{\text{a,p}}\bar{D}_{\text{a,n}}}}{4}. \tag{7}$$

Details on deriving r^* will be discussed in Section III-C. The fifth step is to compare the neighborhoods of the old and new radius values. If the probability of having the same-class sample within the new radius $p(C_{\text{a}} = C_{\text{i}}|X_{\text{a}}, i \in \mathcal{N}_{\text{a},r^*})$ is greater or equal to that of the old $p_{\text{same}} = p(C_{\text{a}} = C_{\text{i}}|X_{\text{a}}, i \in \mathcal{N}_{\text{a,r}_{\text{a}}})$, then $r_{\text{a}} = r^*$ such that the new radius is accepted. Both probability terms are calculated empirically. Note that the terms of the new radius $p(C_{\text{a}} = C_{\text{i}}|X_{\text{a}}, i \in \mathcal{N}_{\text{a},r^*})$ and that of the old radius p_{same} are calculated with the training data, whereas the radius r^* is calculated with local statistics from the validation data to avoid diverging too much from the actual distribution. The process repeats until no new r^* can be calculated (e.g., no samples within new $\mathcal{N}_{\text{a,r}^*}$, only samples of the same class or different classes are present in the validation neighborhood, $\bar{D}_{\text{a,p}} > \bar{D}_{\text{a,n}}$, r^* is equal to previous r_{a}, etc.) It is worth noting that every new r^* is smaller or equal to the previous candidates. Also, every replaced radius r_{a} is not discarded but recorded and used in the query step.

After the search radius values are prepared for all X_{a}, they are used for the classification of the query data set. Note that the query data set can be either a validation data set for treating the proposed method as the validation step to gauge the potential performance or a testing data set when considering the proposed method as the classifier for a performance baseline. To do the inference, 1NN is applied for a query sample first to find the training NN with pre-recorded radius values. The goal of 1NN is to estimate the best neighbor radius for a query point such that $r_{\text{q}} \approx r_{\text{a}}$. Then, the smallest r_{q} is used to estimate the class label from the training set. In the case of no neighbors attained from the radius, the larger radius recorded for the same training sample is used instead. A default KNN search is applied when the recorded values yield no neighbor.

The importance of the proposed method is that the triplet loss metric provides a guideline on how to adaptively tune a hyperparameter for NN classification. Specifically, $\bar{D}_{\text{a,p}}$ and $\bar{D}_{\text{a,n}}$ can be simply calculated from the comparator part of the network for each available X_{a}. Thus, the NN classification lower bound can be quantified prior to the second stage.

Algorithm 1 Adaptive Neighbor Scope for the NN Validation

```
Input: $\theta$ trained triplet network parameter set,
  $R_a$ list of $r_a$ the radius distances for $X_a$,
  $I_i \epsilon \mathfrak{I}$ training inputs from the training data set,
  $I_v \epsilon \mathcal{V}$ validating inputs from the validation data set,
  $I_t \epsilon \mathfrak{T}$ validating inputs from the query data set
  $k$ initial number of neighbors for setting the initial radius
Begin:
  \\ extracting the features and NN search model
  $\mathcal{X}_i \leftarrow f(\mathfrak{I};\theta), \mathcal{M}_i \leftarrow$ NN model from the training features $\mathcal{X}_i$
  $\mathcal{X}_v \leftarrow f(\mathcal{V};\theta), \mathcal{M}_v \leftarrow$ NN model from the validation features $\mathcal{X}_v$
  $\mathcal{X}_t \leftarrow f(\mathfrak{T};\theta)$
  For each $X_a \in \mathcal{X}_i$ do:
    \\ setting the initial search radius
    $r_a \leftarrow \mathcal{M}_v.distance\_to\_kth\_neighbor(X_a, k)$
    $R_a.append(r_a)$
    \\ calculate $p_{same}$ with the training set
    $\mathcal{N}_{a,r_a} \leftarrow \mathcal{M}_i.neighbors(X_a, r_a)$
    $p_{same} \leftarrow sameClassProb(\mathcal{N}_{a,r_a}, \mathcal{X}_i)$
    Repeat:
      \\ find the positive and the negative for radius suggestion
      $r_a \leftarrow R_a.get\_last\_element()$
      $\mathcal{N}_{a,r} \leftarrow \mathcal{M}_v.neighbors(X_a, r_a)$
      If $\mathcal{N}_{a,r}$ contains both positive and negative samples then:
        $\bar{D}_{a,p}, \bar{D}_{a,n} \leftarrow local\_statistics(\mathcal{N}_{a,r})$ \\ get the local statistics
        $r \leftarrow calculate\_r^*(\bar{D}_{a,p}, \bar{D}_{a,n})$ \\ calculated the new radius
        \\ calculate and compare the current $p_{same}$ of the training set
        $\mathcal{N}_{a,r_a} \leftarrow \mathcal{M}_i.neighbors(X_a, r)$
        If $p_{same} \leq sameClassProb(\mathcal{N}_{a,r_a}, \mathcal{X}_i)$ then:
          $p_{same} \leftarrow sameClassProb(\mathcal{N}_{a,r_a}, \mathcal{X}_i)$
          $R_a.append(r)$
        End if
      End if
    Until no new $r^*$ candidate
  End for
  \\ end of training phase
  For each $X_q \in \mathcal{X}_t$ do:
    $X_a \leftarrow \mathcal{M}_i.get\_nearest\_neighbor(X_q)$
    Repeat:
      $r \leftarrow R_a.remove\_last\_element()$
      $\mathcal{N}_{q,r} \leftarrow \mathcal{M}_i.neighbors(X_q, r)$
      If $\mathcal{N}_{q,r}$ is not $\emptyset$ then:
        $C_q \leftarrow NNClassify(\mathcal{N}_{q,r}, \mathcal{X}_i)$
      End if
    Until $\mathcal{N}_{q,r}$ is not $\emptyset$ or $R_a$ is $\emptyset$
    \\ apply default KNN classification if no neighbors
    If $C_q$ is undetermined then:
      $C_q \leftarrow KNNClassify(X_i, k)$
    End if
  End for
End
Return all $C_q$
```

We assert that the NN performance is the translation result from the suitability metric such that the class-separable features from the metric learning step push the lower bound up to that performance. Thus, a more sophisticated classifier should be able to use the features to achieve better performance.

C. Theoretical Insights From Adaptive Neighbor Scope

To obtain theoretical insights from the proposed adaptive neighbor scope, we make the three following assumptions.

In the NN step, each feature vector X_q in the query set is independent and identically distributed (iid) to the feature vector X_a of the training and validation sets. This assumption is very common in ML research as it is often assumed that data samples in training, validation, and testing sets are iid. It also suggests that an outlier query of sparse or null neighborhoods is rather rare. Thus, the frequency of the null neighborhood is negligible if it is not present in the training set. We exclude outlier cases out of the scope of our study.

The correct classification of X_q is more probable if the neighborhood of X_q has a larger probability of having samples from the same class than that of having samples from the different class. This assumption is intuitively an extension of the first assumption such that the neighborhood with a more homogeneous distribution is less susceptible to the NN errors.

The best radius for X_q is $r_q^* \approx r_a^*$ for X_a that is closest to X_q. This assumption presumes that the neighbor distribution surrounding X_q is very similar to that of its NN X_a. This is in line with other known literature of the NN estimation.

Based on the second assumption, r_q should be set to retrieve samples from the neighborhood containing only one class. If triplet loss in (1) is minimized, samples of different classes are separated at least m distance. However, the following propositions and theorem show that simply setting $r_q = m$ does not always lead to the correct results even if the loss is minimized.

Proposition 1.1: Given $\mathcal{N}_{a,r_a}$, $D_{a,i} \leq r_a$ is true for any known sample X_i where $i \in \mathcal{N}_{a,r_a}$. Otherwise, $D_{a,i} > r_a$ and $i \notin \mathcal{N}_{a,r_a}$.

Proposition 1.2: Given a query with unknown class X_q which is closest to a known sample X_a, X_q is either a sample within the area covered by a radius r_a from X_a such that $D_{a,q} \leq r_a$ or an outlier such that $D_{a,q} > r_a$.

These propositions define obvious conditions for a feature point to be a neighbor of X_a such that the sample coordinate is covered by the defined radius around X_a to be considered one of the neighbors. The same condition is also applied to query a sample of unknown class closest to X_a.

Theorem 1: If the triplet fixed-margin loss $\mathcal{L}_{tri}$ is minimized for all X_a, then $r_q \leq m - D_{a,q}$ can be used to form a neighborhood $\mathcal{N}_{q,r_q}$ of which all the samples X_j where $j \in \mathcal{N}_{q,r_q}$ have the same class label as that of X_a for any X_q closest to X_a.

Proof: Consider minimized $\mathcal{L}_{tri} = 0$. Then, the following statement can be rearranged from (1):

$$D_{a,n} \geq D_{a,p} + m > m. \tag{8}$$

Consider setting $r_a = m$, (8) implies that no negative sample is in $\mathcal{N}_{a,m}$. For any X_q closest to X_a and fall within r_a radius from X_a, the largest radius from X_q that does not cover an

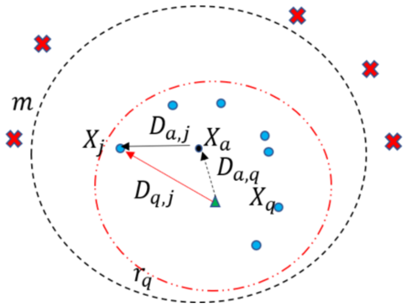

Fig. 2. Example of the neighborhood surrounding X_{a}, X_{j}, and X_{q} when $\mathfrak{L}_{\mathrm{tri}}$ is minimized. Under the loss condition, all the negative samples are outside the radius of m, and r_{q} is the potential radius from X_{q} such that the known samples covered by the radius are of the same class corresponding to the label of X_{a}.

area outside that of $\mathfrak{N}_{\mathrm{a,m}}$ is $m - D_{\mathrm{a,q}}$, and the neighborhood $\mathfrak{N}_{\mathrm{q,m}-D_{\mathrm{a,q}}} \subseteq \mathfrak{N}_{\mathrm{a,m}}$. Then, $D_{\mathrm{q,j}} \leq m - D_{\mathrm{a,q}}$ is true according to Proposition 1.2 for any X_{j} where $j \in \mathfrak{N}_{\mathrm{q,m}-D_{\mathrm{a,q}}}$. From Proposition 1.1, $D_{\mathrm{a,j}} \leq m$ is true as all j are also contained in $\mathfrak{N}_{\mathrm{a,m}}$. From Proposition 1.2, $D_{\mathrm{a,q}} \leq m$ is true as X_{q} is the closest to X_{a}. Subsequently, consider a triangle formed by $X_{\mathrm{j}}, X_{\mathrm{q}}$, and X_{a} as shown in Fig. 2. The following triangle inequality holds:

$$D_{\mathrm{a,j}} \leq D_{\mathrm{a,q}} + D_{\mathrm{q,j}} \leq D_{\mathrm{a,q}} + m - D_{\mathrm{a,q}} \leq m. \tag{9}$$

We can then prove this theorem by contradiction. Assuming that one of X_{j} belongs to a negative class denoted by $n(j)$, where $j \in \mathfrak{N}_{\mathrm{q,m}-D_{\mathrm{a,q}}}$. Then, according to (9), $D_{\mathrm{a,n}(j)} \leq m$ contradicts the condition in (8), which asserts that the distance from X_{a} to any negative sample is greater than m. ■

The implication of Theorem 1 is that m has less utility in specifying the radius to the neighborhood boundary. The best neighborhood area with no probability of having a different-class sample after optimizing the loss for a query is confined in $r_{\mathrm{q}} \leq m - D_{\mathrm{a,q}}$, which is smaller than m. Thus, setting m as the radius may not guarantee the single-class neighborhood. Although the theorem encourages setting a smaller radius, it introduces the risk of having no sample within the smaller radius when fewer samples exist in the limited training set and when $m \ll \mathrm{E}[D_{\mathrm{a,p}}]$ as there is no limit on how large $D_{\mathrm{a,p}}$ can increase. It is also possible that the theoretical radius can approach 0. Thus, the constant m alone is insufficient for specifying the radius r_{q} for NN classification even if the network optimally separates the features by the fixed constant.

Corollary 1: If triplet fixed-margin loss $\mathfrak{L}_{\mathrm{tri}}$ is minimized for all X_{a}, then $E[D_{\mathrm{a,n}}] \geq E[D_{\mathrm{a,p}}]$ for any X_{a}.

Equation (8) in Theorem 1 also establishes that $D_{\mathrm{a,n}} \geq D_{\mathrm{a,p}}$ is true when the loss is minimized. It also implies that the inequality from applying expectation on both sides of the statement is true for any positive m value. The corollary suggests that attaining a smaller number of different-class samples than that of the same class is likely when the search radius r_{q} is sufficiently small. It gives an opportunity for setting $r_{\mathrm{q}} \geq m - D_{\mathrm{a,q}}$ if a small portion of the different-class samples in the neighborhood can be tolerated. Under such a scenario, a good neighborhood area should be alternatively defined as an area containing same-class samples as the majority as opposed to a single-class neighborhood.

To search for the alternative best r_{q}, the following propositions and theorems lay the foundation for our method.

Proposition 2.1: Given X_{a}, $\mathfrak{N}_{\mathrm{a,r_a}}$, and $\mathfrak{N}_{\mathrm{a,r_a}}$ containing indices of both the same-class and different-class samples, there exist $L^s_{\mathrm{a,r_a}}$ and $L^d_{\mathrm{a,r_a}}$, which are random variables of distances to the same-class sample and a different-class sample within the neighborhood radius r_{a} from X_{a}.

The proposition expresses the existence of the samples in terms of the distances to the same-class and different-class samples from an anchor point. Specifically, any X_{i} in the neighborhood can be used to calculate $L^s_{\mathrm{a,r_a}} = D_{\mathrm{a,i}}|X_{\mathrm{a}}, \mathfrak{N}_{\mathrm{a,r_a}}$ if $C_{\mathrm{a}} = C_{\mathrm{i}}$, or $L^d_{\mathrm{a,r_a}} = D_{\mathrm{a,i}}|X_{\mathrm{a}}, \mathfrak{N}_{\mathrm{a,r_a}}$ if $C_{\mathrm{a}} \neq C_{\mathrm{i}}$.

Proposition 2.2: Given an alternative radius $r \leq r_{\mathrm{a}}$, the lower bound of $p(i \in \mathfrak{N}_{\mathrm{a,r}}, C_{\mathrm{a}} = C_{\mathrm{i}}|X_{\mathrm{a}}, i \in \mathfrak{N}_{\mathrm{a,r_a}})$ can be defined by

$$S\left(r; \mu^s_{\mathrm{a,r_a}}\right) = \begin{cases} \left(1 - \frac{\mu^s_{\mathrm{a,r_a}}}{r}\right) p\left(C_{\mathrm{a}} = C_{\mathrm{i}}|X_{\mathrm{a}}, i \in \mathfrak{N}_{\mathrm{a,r_a}}\right), & \text{if } r > \mu^s_{\mathrm{a,r_a}} \\ 0, & \text{if } r \leq \mu^s_{\mathrm{a,r_a}} \end{cases} \tag{10}$$

where $\mu^s_{\mathrm{a,r_a}} = \mathrm{E}[L^s_{\mathrm{a,r_a}}]$.

Proof: Consider that $p(i \in \mathfrak{N}_{\mathrm{a,r}}, C_{\mathrm{a}} = C_{\mathrm{i}}|X_{\mathrm{a}}, i \in \mathfrak{N}_{\mathrm{a,r_a}})$ is equal to $p(D_{\mathrm{a,i}} \leq r, C_a = C_i | X_{\mathrm{a}}, i \in \mathfrak{N}_{\mathrm{a,r_a}})$

$$= p\left(D_{\mathrm{a,i}} < r | C_{\mathrm{a}} = C_{\mathrm{i}}, X_{\mathrm{a}}, i \in \mathfrak{N}_{\mathrm{a,r_a}}\right) p\left(C_{\mathrm{a}} = C_{\mathrm{i}}|X_{\mathrm{a}}, i \in \mathfrak{N}_{\mathrm{a,r_a}}\right)$$
$$= p(L^s_{\mathrm{a,r_a}} \leq r) p(C_{\mathrm{a}} = C_{\mathrm{i}}|X_{\mathrm{a}}, i \in \mathfrak{N}_{\mathrm{a,r_a}}).$$

Then, consider Markov inequality for random variable $L^s_{\mathrm{a,r_a}}$

$$p\left(L^s_{\mathrm{a,r_a}} \geq r\right) \leq \frac{\mu^s_{\mathrm{a,r_a}}}{r}. \tag{11}$$

Equation (11) can be rearranged as

$$p\left(L^s_{\mathrm{a,r_a}} \leq r\right) \geq 1 - \frac{\mu^s_{\mathrm{a,r_a}}}{r}. \tag{12}$$

Regardless of the r value, $p(L^s_{\mathrm{a,r_a}} \leq r)$ must be nonnegative and upper bounded at 1. Thus, $p(L^s_{\mathrm{a,r_a}} \leq r) \geq 0$ if $r \leq \mu^s_{\mathrm{a,r_a}}$. From (12), the lower bound from Markov equality leads to

$$p\left(L^s_{\mathrm{a,r_a}} \leq r\right) p\left(C_{\mathrm{a}} = C_{\mathrm{i}}|X_{\mathrm{a}}, i \in \mathfrak{N}_{\mathrm{a,r_a}}\right) \geq \left(1 - \frac{\mu^s_{\mathrm{a,r_a}}}{r}\right) p\left(C_{\mathrm{a}} = C_{\mathrm{i}}|X_{\mathrm{a}}, i \in \mathfrak{N}_{\mathrm{a,r_a}}\right). \tag{13}$$

Therefore, the term on the right-hand side becomes the lower bound probability $S(r; \mu^s_{\mathrm{a,r_a}})$ in the proposition. ■

Proposition 2.3: Given a radius distance $r \leq r_{\mathrm{a}}$, the upper bound of $p(i \in \mathfrak{N}_{\mathrm{a,r}}, C_{\mathrm{a}} \neq C_{\mathrm{i}}|X_{\mathrm{a}}, i \in \mathfrak{N}_{\mathrm{a,r_a}})$ is defined by

$$Q\left(r; \mu^d_{\mathrm{a,r_a}}, \sigma^d_{\mathrm{a,r_a}}\right) = \begin{cases} \frac{\left(\sigma^d_{\mathrm{a,r_a}}\right)^2}{\left(\mu^d_{\mathrm{a,r_a}} - r\right)^2} p\left(C_{\mathrm{a}} \neq C_{\mathrm{i}}|X_{\mathrm{a}}, i \in \mathfrak{N}_{\mathrm{a,r_a}}\right), & \text{if } \mu^d_{\mathrm{a,r_a}} - r > \sigma^d_{\mathrm{a,r_a}} \\ p\left(C_{\mathrm{a}} \neq C_{\mathrm{i}}|X_{\mathrm{a}}, i \in \mathfrak{N}_{\mathrm{a,r_a}}\right), & \text{if } \mu^d_{\mathrm{a,r_a}} - r \leq \sigma^d_{\mathrm{a,r_a}} \end{cases} \tag{14}$$

where $r < \mu^d_{\mathrm{a,r_a}}$, $\mu^d_{\mathrm{a,r_a}} = \mathrm{E}[L^d_{\mathrm{a,r_a}}]$, $(\sigma^d_{\mathrm{a,r_a}})^2$ is variance of $L^d_{\mathrm{a,r_a}}$.

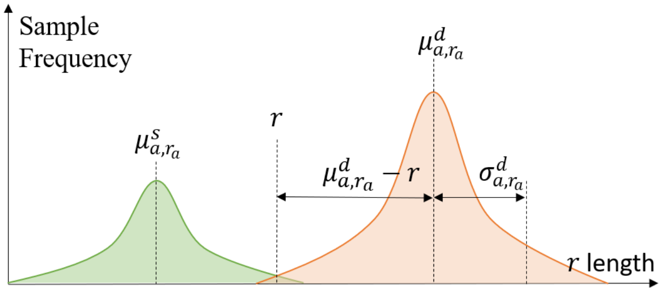

Fig. 3. Example of the distributions of the distances to the same-class and different-class samples from X_{a}. Setting the value of r determines which portions of the samples are included in the neighborhood defined by $\mathcal{N}_{\mathrm{a,r}}$.

Proof: Similar to Proposition 2.2, consider that probability $p(i \in \mathcal{N}_{\mathrm{a,r}}, C_{\mathrm{a}} \neq C_{\mathrm{i}} | X_{\mathrm{a}}, i \in \mathcal{N}_{\mathrm{a,r_a}})$ is equal to

$$\begin{aligned} & p\big(D_{\mathrm{a,i}} \leq r, C_a \neq C_i | X_{\mathrm{a}}, i \in \mathcal{N}_{\mathrm{a,r_a}}\big) \\ & = p\big(D_{\mathrm{a,i}} \leq r | C_a \neq C_i, X_{\mathrm{a}}, i \in \mathcal{N}_{\mathrm{a,r_a}}\big) p\big(C_{\mathrm{a}} \neq C_{\mathrm{i}} | X_{\mathrm{a}}, i \in \mathcal{N}_{\mathrm{a,r_a}}\big) \\ & = p\big(L^{d}_{\mathrm{a,r_a}} \leq r\big) p\big(C_{\mathrm{a}} \neq C_{\mathrm{i}} | X_{\mathrm{a}}, i \in \mathcal{N}_{\mathrm{a,r_a}}\big). \end{aligned} \tag{15}$$

Then, the Chebyshev inequality for random variable $L^{d}_{\mathrm{a,r_a}}$ is

$$p\big(\big|L^{d}_{\mathrm{a,r_a}} - \mu^{d}_{\mathrm{a,r_a}}\big| \geq k\sigma^{d}_{\mathrm{a,r_a}}\big) \leq \frac{1}{k^2}. \tag{16}$$

Let $k = (\mu^{d}_{\mathrm{a,r_a}} - r)/\sigma^{d}_{\mathrm{a,r_a}}$. Then, the inequality can be expressed as

$$p\big(\big|L^{d}_{\mathrm{a,r_a}} - \mu^{d}_{\mathrm{a,r_a}}\big| \geq \mu^{d}_{\mathrm{a,r_a}} - r\big) \leq \frac{\big(\sigma^{d}_{\mathrm{a,r_a}}\big)^2}{\big(\mu^{d}_{\mathrm{a,r_a}} - r\big)^2}.$$

Fig. 3 depicts the line distance from X_{a} to $\mu^{d}_{\mathrm{a,r_a}}$ according to (16). Any $L^{d}_{\mathrm{a,r_a}} < r$ must have its difference from $\mu^{d}_{\mathrm{a,r_a}}$ larger than $\mu^{d}_{\mathrm{a,r_a}} - r$. Thus, the left-side probability term $p(|L^{d}_{\mathrm{a,r_a}} - \mu^{d}_{\mathrm{a,r_a}}| \geq \mu^{d}_{\mathrm{a,r_a}} - r)$ covers a fraction of $L^{d}_{\mathrm{a,r_a}}$ population that is less than r. Then, we can posit that

$$p\big(L^{d}_{\mathrm{a,r_a}} \leq r\big) \leq p\big(\big|L^{d}_{\mathrm{a,r_a}} - \mu^{d}_{\mathrm{a,r_a}}\big| \geq \mu^{d}_{\mathrm{a,r_a}} - r\big) \tag{17}$$

and

$$p\big(L^{d}_{\mathrm{a,r_a}} \leq r\big) \leq \frac{\big(\sigma^{d}_{\mathrm{a,r_a}}\big)^2}{\big(\mu^{d}_{\mathrm{a,r_a}} - r\big)^2}. \tag{18}$$

Regardless of the r value, $p(L^{d}_{\mathrm{a,r_a}} \leq r)$ must not exceed 1. Thus, $p(L^{d}_{\mathrm{a,r_a}} \leq r) \leq 1$ if $\mu^{d}_{\mathrm{a,r_a}} - r \leq \sigma^{d}_{\mathrm{a,r_a}}$. Then, the upper bound from Chebyshev inequality leads to

$$\begin{aligned} & p\big(L^{d}_{\mathrm{a,r_a}} \leq r\big) p\big(C_{\mathrm{a}} \neq C_{\mathrm{i}} | X_{\mathrm{a}}, i \in \mathcal{N}_{\mathrm{a,r_a}}\big) \\ & \qquad \leq \frac{\big(\sigma^{d}_{\mathrm{a,r_a}}\big)^2}{\big(\mu^{d}_{\mathrm{a,r_a}} - r\big)^2} p\big(C_{\mathrm{a}} \neq C_{\mathrm{i}} | X_{\mathrm{a}}, i \in \mathcal{N}_{\mathrm{a,r_a}}\big). \end{aligned} \tag{19}$$

Consequently, the term on the right-hand side becomes the upper bound probability $Q(r; \mu^{d}_{\mathrm{a,r_a}}, \sigma^{d}_{\mathrm{a,r_a}})$ in the proposition. ■

Propositions 2.2 and 2.3 define limits on the local distributions of the same-class and different-class samples as functions of the radius distance r from X_{a}. If a new neighborhood of X_{a} is to be re-defined using r instead of r_{a}, then the limits can provide information on the portion of same-class and different-class samples within the new neighborhood.

Theorem 2: Given X_{a}, r_{a}, and $\mu^{d}_{\mathrm{a,r_a}} \geq \mu^{s}_{\mathrm{a,r_a}}$, according to the local distributions in $\mathcal{N}_{\mathrm{a,r_a}}$, there exists r^{*}_{a} that results in the highest lower bound on probability of having same-class samples when $\mathcal{N}_{\mathrm{a,r^*}}$ define a new neighborhood where

$$r^{*}_{\mathrm{a}} = \frac{\mu^{s}_{\mathrm{a,r_a}} + \sqrt{\big(\mu^{s}_{\mathrm{a,r_a}}\big)^2 + 8\mu^{s}_{\mathrm{a,r_a}}\mu^{d}_{\mathrm{a,r_a}}}}{4} \tag{20}$$

and

$$\mu^{s}_{\mathrm{a,r_a}} \leq r^{*}_{\mathrm{a}} \leq \mu^{d}_{\mathrm{a,r_a}} \tag{21}$$

$$r^{*}_{\mathrm{a}} \leq r_{\mathrm{a}}. \tag{22}$$

Proof: Consider using the radius r to define $\mathcal{N}_{\mathrm{a,r}} \subseteq \mathcal{N}_{\mathrm{a,r_a}}$ where $r \leq r_{\mathrm{a}}$. Fractions of the same-class and different-class populations would be contained in the new neighborhood. Then, the lower bound on probability of having the same-class samples in the new neighborhood is $\lambda(r)$ defined as

$$\lambda(r) = \frac{S\big(r; \mu^{s}_{\mathrm{a,r_a}}\big)}{S\big(r; \mu^{s}_{\mathrm{a,r_a}}\big) + Q\big(r; \mu^{d}_{\mathrm{a,r_a}}, \sigma^{d}_{\mathrm{a,r_a}}\big)}. \tag{23}$$

Equation (23) implies the worst case that the least amount same-class population and the largest amount different-class population specified by $S(r; \mu^{s}_{\mathrm{a,r_a}})$ and $Q(r; \mu^{d}_{\mathrm{a,r_a}}, \sigma^{d}_{\mathrm{a,r_a}})$ are included in the new neighborhood of $\mathcal{N}_{\mathrm{a,r}}$. To find r^{*}_{a} with the largest value of $\lambda(r)$, we arrange a derivative $(\partial\lambda(r))/(\partial r) = 0$ to solve for r^{*}_{a}, which results in

$$r^{*}_{\mathrm{a}} = \frac{\mu^{s}_{\mathrm{a,r_a}} \pm \sqrt{\big(\mu^{s}_{\mathrm{a,r_a}}\big)^2 + 8\mu^{s}_{\mathrm{a,r_a}}\mu^{d}_{\mathrm{a,r_a}}}}{4}. \tag{24}$$

The derivative suggests two values of r^{*}_{a} candidates. However, any value of $r \leq \mu^{s}_{\mathrm{a,r_a}}$ would result in $S(r; \mu^{s}_{\mathrm{a,r_a}}) = 0$. Thus, r^{*}_{a} value in (20) is true. Proving that $\mu^{s}_{\mathrm{a,r_a}} \leq r^{*}_{\mathrm{a}} \leq \mu^{d}_{\mathrm{a,r_a}}$ in (21) and $r^{*}_{\mathrm{a}} \leq r_{\mathrm{a}}$ in (22) can be done by careful consideration of r^{*}_{a}, $L^{s}_{\mathrm{a,r_a}}$, and $L^{d}_{\mathrm{a,r_a}}$ values. Considering $\mu^{d}_{\mathrm{a,r_a}} \geq \mu^{s}_{\mathrm{a,r_a}}$, $\mu^{s}_{\mathrm{a,r_a}}$ can be rewritten as $\mu^{s}_{\mathrm{a,r_a}} = \mu^{d}_{\mathrm{a,r_a}} - \tau$ where τ is a non-negative constant. Then, r^{*} can be expressed as

$$\begin{aligned} r^{*}_{\mathrm{a}} &= \frac{\big(\mu^{d}_{\mathrm{a,r_a}} - \tau\big) + \sqrt{\big(\mu^{d}_{\mathrm{a,r_a}} - \tau\big)^2 + 8\mu^{d}_{\mathrm{a,r_a}}\big(\mu^{d}_{\mathrm{a,r_a}} - \tau\big)}}{4} \\ r^{*}_{\mathrm{a}} &= \frac{\big(\mu^{d}_{\mathrm{a,r_a}} - \tau\big) + \sqrt{9\big(\mu^{d}_{\mathrm{a,r_a}}\big)^2 - 10\mu^{d}_{\mathrm{a,r_a}}\tau - \tau^2}}{4} \\ r^{*}_{\mathrm{a}} &\leq \frac{\mu^{d}_{\mathrm{a,r_a}} + \sqrt{9\big(\mu^{d}_{\mathrm{a,r_a}}\big)}^{2}}{4} \leq \frac{\mu^{d}_{\mathrm{a,r_a}} + 3\mu^{d}_{\mathrm{a,r_a}}}{4} \leq \mu^{d}_{\mathrm{a,r_a}}. \end{aligned}$$

Given that $r^{*}_{\mathrm{a}} \leq \mu^{d}_{\mathrm{a,r_a}}$, then $r^{*}_{\mathrm{a}} \leq \mathrm{E}[L^{d}_{\mathrm{a,r_a}}] \leq \max(L^{d}_{\mathrm{a,r_a}}) \leq r_{\mathrm{a}}$. It is easy to see that $\mu^{s}_{\mathrm{a,r_a}} \leq r^{*}_{\mathrm{a}} \leq \mu^{d}_{\mathrm{a,r_a}}$ and $r^{*}_{\mathrm{a}} \leq r_{\mathrm{a}}$. ■

The implication of Theorem 2 is that for any neighborhood defined by $\mathcal{N}_{\mathrm{a,r_a}}$ in which $\mu^{s}_{\mathrm{a,r_a}} \leq \mu^{d}_{\mathrm{a,r_a}}$ is true, there exists a better neighborhood $\mathcal{N}_{\mathrm{a,r^*_a}}$ formed as a subset of $\mathcal{N}_{\mathrm{a,r_a}}$. The formulation of $\mathcal{N}_{\mathrm{a,r^*_a}}$ from $\mathcal{N}_{\mathrm{a,r_a}}$ is useful for defining a new neighborhood for NN classification such that the neighborhood radius for a query does not have to be restricted by the optimal condition in Theorem 1, or some fixed value of last resort. It is also noteworthy that the calculation of r^{*}_{a} is also applicable even if $\mathcal{L}_{\mathrm{tri}}$ is not completely minimized. As long as the conditions $E[D_{\mathrm{a,n}}] \geq E[D_{\mathrm{a,p}}]$ or $\mu^{s}_{\mathrm{a,r_a}} \leq \mu^{d}_{\mathrm{a,r_a}}$ in the neighborhood hold, r^{*}_{a} can still be derived. Using Theorem 2, the value of r^*_{a} can be estimated with local statistics calculated using samples within a neighborhood area. Replacing $\mu^s_{\mathrm{a,r_a}}$ and $\mu^d_{\mathrm{a,r_a}}$ with $\bar{D}_{\mathrm{a,p}}$ and $\bar{D}_{\mathrm{a,n}}$ in (20), r^*_{a} calculation is the same as in (7).

Corollary 2: Given r^*_{a}, the value of $\lambda(r^*_{\mathrm{a}})$ is larger when $\mu^s_{\mathrm{a,r^*_a}}$ and $\sigma^d_{\mathrm{a,r^*_a}}$ are smaller, and $\mu^d_{\mathrm{a,r^*_a}}$ is bigger.

According to the definition of $\lambda(r^*_{\mathrm{a}})$, a smaller $\mu^s_{\mathrm{a,r^*_a}}$ leads to a larger value of $S(r^*_{\mathrm{a}}; \mu^s_{\mathrm{a,r^*_a}})$. A smaller $\sigma^d_{\mathrm{a,r^*_a}}$ and a bigger $\mu^d_{\mathrm{a,r^*_a}}$ result in a decrease in $Q(r^*_{\mathrm{a}}; \mu^d_{\mathrm{a,r^*_a}}, \sigma^d_{\mathrm{a,r^*_a}})$. However, $\mu^s_{\mathrm{a,r^*_a}}$, $\mu^d_{\mathrm{a,r^*_a}}$, $\sigma^d_{\mathrm{a,r^*_a}}$, and $\lambda(r^*_{\mathrm{a}})$ are constants calculated from the features once the training iteration ends. This corollary encourages controls over the values of these constants during training. A simple way is to follow regularization with the global loss as proposed in [26]. Thus, the triplet loss with regularization is expressed as

$$L_{\mathrm{tri}} = w_{fm} \sum_{\mathrm{a}} \sum_{p} \sum_{n} \max\big(0, D_{\mathrm{a,p}} - D_{\mathrm{a,n}} + m\big) + w_{ms}\mu^s - w_{md}\mu^d + w_{sd}\big(\sigma^d\big)^2 \tag{25}$$

where w_{fm} is the weight value for the triplet loss, μ^s and μ^d are the average distances to the same-class and different-class samples, respectively, $(\sigma^d)^2$ is the variance of the distance to different-class samples, and w_{ms}, w_{md}, and w_{sd} are the weights for the corresponding means and variances. The result in this corollary provides a theoretical motivation to apply the regularization terms in the triplet network training that is used in this study.

D. Computational Complexity

Let F and n be the number of maximum training epochs for the triplet network and the total number of training data, respectively. Without the strategy, the training has to go back and forth between the feature candidates and classifiers tuning. Normally, there are $O(F)$ feature candidates from the first-stage training as each epoch produces one candidate. Let Z be the worst case number of operations taken to train and test the second-stage classifiers, then the total operations for the typical two-stage training are at most $O(F \times Z)$.

With the proposed validation strategy, the number of candidates is reduced to $O(1)$, or one to a few candidates with the best lower bound performance, as opposed to $O(F)$ from the first stage. The method takes overall $n \times O(n) \times (k-1) = O(n^2)$ worst case operations for each of the validation round. In other words, the calculation for each of the n samples needs $O(n)$ initial KNN search and computes at most $k-1$ times before stopping when at most one sample is in the new radius. Thus, the overall complexity is $O(F) \times O(n^2) + O(1) \times O(Z)$ where $O(F) \times O(n^2)$ is the first-stage validation on the $O(F)$ candidates and $O(1) \times O(Z) = O(Z)$ is the second-stage training after the selection for best feature epoch.

Generally, the strategy suits our setting where large F and Z are desirable. As F is often set according to the number of data definitions or the problem difficulty, it is often the case for any deep learning approach that $F > n^2$ in our small-sample setting. Comparatively, n^2 is approximately the size of the expanded triplet data set. While the larger F improves loss minimization, the larger Z implies that more classifiers may be tuned for best performance. Not only does Z grow according to the data and difficulty but also it depends on the number and complexity of the end classifiers. The real advantage is the decoupling of F and Z on the complexity terms. The validation allows the flexibility for end classification without incurring too much overall complexity.

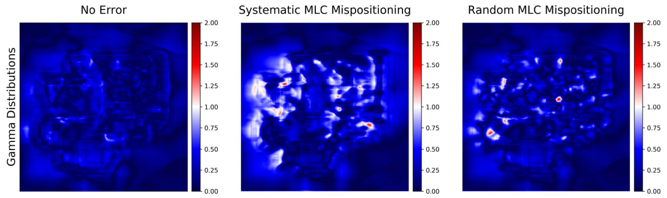

Fig. 4. Three examples of the resulting EPID gamma images in the experiment in which MLC mispositioning errors can be observed. From left to right: the EPID gamma images are categorized as no error, systematic error, and random error. The high image intensity indicates a larger deviation from the radiation therapy plan.

IV. Computational Experiments

In this study, we applied the proposed strategy to two real-world data sets with the small sample size problem. Both data sets were acquired for clinical imaging research in which medical imaging data, expert-defined features, and clinical factors were captured as the medical information.

A. Data Sets, Clinical Problems, and Multi-View Features

1) EPID Gamma Images: The goal of the classification task is to classify whether patient-specific quality assurance images of radiotherapy treatments contain errors. The clinical motivation is described in [7], [41]. Briefly, in radiation therapy delivery, the electronic portal imaging device (EPID) is used to capture the radiation beam to form 2-D images. The images of the patient treatment are compared with the intended treatment to ensure the safety and quality of the radiation treatment delivery by trained personnel. The decision on whether the plans are clinically suitable is based on gamma maps derived from the images. In clinical practice, the gamma values >1.0 are considered failing (e.g., indicate that there may be a problem with the patient treatment) and 90% of the total number of pixels in the gamma maps must pass to ensure the integrity of the treatment.

The data set consisted of 558 2-D gamma maps (256 × 256 pixels) collected for radiation therapy quality assurance. The data were simulated as in [7] from 23 patient treatment plans using 186 intensity-modulated radiation therapy (IMRT) beams. For the gamma maps, one-third of the images had no introduced errors, one-third had a random mechanical error [random mispositioning of the multileaf collimator (MLC)], and one-third had a systematic mechanical error (systematic misplacement of the MLC). Fig. 4 illustrates examples from each type of the two errors whose patterns can be non-intuitive. To keep the semantics, no scaling was applied similar to [7].

Two types of image features were extracted from the gamma maps. The first type was a set of radiomic features extracted using the PORTS software [27], which have been widely used as expert-engineered features for medical imaging

TABLE I

LIST OF THE RADIOMICS FEATURES INCLUDED IN THE EPID DATA SET

Histogram	Mean, Variance, Skewness, Kurtosis, Energy, Entropy
Zone Size	Small Zone Size Emphasis, Large Zone Size Emphasis, Low Gray Level Zone Emphasis, High Gray Level Zone Emphasis, Small Zone Low Gray Level, Small Zone High Gray Level, Large Zone Low Gray Level, Large Zone High Gray Level, Gray Level Nonuniformity, Zone Size Nonuniformity, Zone Size Percentage

tasks including our problem [2], [7], [28], [29]. A total of 17 radiomic features were calculated as per [41]. The radiomic features that were selected in this study are shown in Table I. The second type was from deep networks pretrained with the ImageNet data set for large-scale image recognition. The InceptResnetv2 architecture [9] was used for the extraction of the features after resampling the gamma image to 224 × 224, which resulted in a 1536-D feature vector for each image.

A total of 558 images along with their extracted data were randomly divided into two sets of 303 and 255 cases. Thirty image cases were selected randomly from the former set for validation leaving 273 cases for training. The latter set became the out-of-sample images for testing.

2) 3-D MR Images of Sarcoma: The clinical question of this data set is to determine whether patients with soft-tissue sarcoma (STS) would survive longer than 1096 days (three years). STS is a malignancy that represents about 1% of all cancers and presents many challenges for clinical management. The clinical motivation is further described in [29], [42], [43].

This data set included a set of magnetic resonance imaging (MRI) scans of patients with sarcoma soft-tissue cancer. We acquired pretreatment contrast-enhanced T1-weighted 3-D MRI scans from two independent cohorts of patients diagnosed with biopsy-proven STS from two different institutes of the University of Washington (UW cohort) and the Technische Universität Munich (Munich cohort). Images were accessed from the institutional picture archiving and communication system (PACS). All patients who were less than 18 years old or were diagnosed with Kaposi or primary bone sarcomas were excluded. The included patients had sarcomas of various histologies of the extremity, trunk, or retroperitoneum. This study focused on the American Joint Committee on Cancer (AJCC) version 7 stage II–III patients only, which encompasses non-metastatic patients with large (i.e., >5 cm) and/or higher grade (i.e., >1) tumors. The patients with image artifacts due to multiple MRI acquisitions were also excluded. The total patients in the two cohorts were 200 and 72 for UW and Munich cohorts, respectively.

In both sarcoma cohorts, radiologist and radiation oncologist experts evaluated each image for quality and manually segmented the gross tumor as the region of interest (ROI), which was defined as all enhancing tumor on contrast-enhanced T1 MRI. This was completed using MIM software (version 6.6, MIM Software Inc, Cleveland, OH) for the UW cohort and iPlan RT (version 4.1.2, Brainlab, Munich, Germany) for the Munich cohort. Fig. 5 visualizes some samples in the data set.

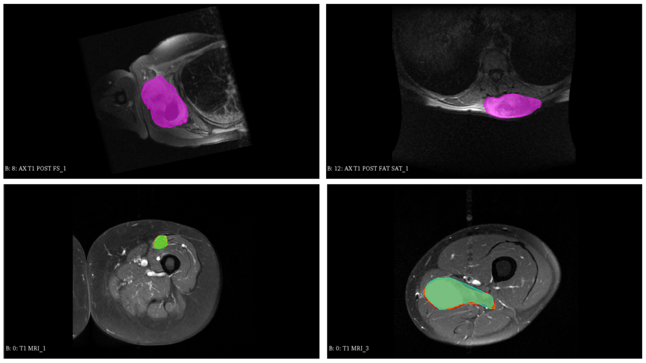

Fig. 5. Example visualization of samples in the Sarcoma MRI data set with the tumor volume ROI defined by the experts. Top: samples from the UW cohort patients. Bottom: samples from the Munich cohort patients.

TABLE II

LIST OF CLINICAL VARIABLES INCLUDED IN THE SARCOMA DATA SET. THE VARIABLES ARE LATER PRE-PROCESSED INTO A 27-D FEATURE VECTOR

Binary variables	- Receive chemotherapy - Size (>5cm) - sex
Ordinal variables	- Histology (10 levels) - FNCLCC grade (3 levels) - AJCC clinical stage (5 levels) - Margin (2 levels) - locations (3 levels)
Continuous variable	Age (z-normalized)

Each scan contained one ROI which was resampled to the fixed resolution of 1 × 1 × 1 mm^3. All the image ROIs were then resampled again into fixed-size bounding rectangles of 64 × 64 × 64 voxels. The bounded data were normalized with the simple strategy as in [38] because the tumor data were already contained within expert-delineated ROIs. Specifically, the voxels' intensity values were clipped at the 99th percentile value before subtraction with minimum intensity values and scaled to the range of [0–1].

As before, the features were extracted using the PORTS software package. In this case, 45 features were used. Additionally, clinical variables such as whether the patient received chemotherapy were defined by experts as additional inputs. The complete list of the variables is provided in Table II. The variables were preprocessed into 27-D features where each ordinal feature value was transformed into a binary level indicator and each continuous value was z-normalized. In this study, the UW cohort patients were used in the training and validation steps. Out of the 200 cases, 180 cases were included in training and 20 cases were randomly sampled to be included in validation. The 72 data cases from the Munich cohort were used as the testing set.

B. Experimental Setups

We evaluated our strategy with two experiments. The purpose was to demonstrate the ability of the method to provide a competitive final classification performance and to show its utility as the feature selector for subsequent tasks. In both experiments, we compared the testing performance of the proposed method against four baseline approaches, namely, triplet network with triplet loss validation, AE, end-to-end softmax, and expert-defined features. The chosen deep learning baselines are widely applied transfer learning approaches capable of solving medical imaging tasks using multiple data domains [40], which is in line with our problem where the available data include but are not limited to visual images. All the baselines were evaluated in three settings: (1) Three-class classification (no error versus systematic error versus random error) on EPID data; (2) two-class classification (no error versus error) on EPID data; and (3) two-class classification (survive versus not survive after 1096 days) on Sarcoma data. In addition, we compared the performance of our strategy in the survival regression setting to predict the survival time of subjects in the Sarcoma data.

Experiment I: Our Strategy to Optimize the Training in the Validation Step. In this experiment, the features chosen from our validation are evaluated both qualitatively and quantitatively with the expectation that the features have better class separation and yield better end performance. We trained the triplet network and used our adaptive neighbor scope. The epoch with the best validation performance was chosen for evaluation with other state-of-the-art baselines, namely, the typical triplet network validated with triplet loss, AE validated using reconstruction loss, and the end-to-end network validated using softmax cross-entropy loss.

For qualitative evaluation, we visually compared the homogeneity of the validated features with that of the other baselines. Prior to visualization, we used t-SNE [39] for dimensional reduction due to its widely regarded advantage in preserving local proximity between the feature points after reduction from high dimensions. The reduction is useful for inspecting local class distribution. All the features were reduced using T-SNE to two dimensions with the perplexity parameter set to five for the EPID and ten for the Sarcoma data sets. Then, scatter plots of the features from the testing set were created. To capture regions relevant to decision boundaries, the lower and upper limits on both the horizontal and the vertical axes were set to the minimum and maximum values of the features of each baseline. Homogeneous regions were overlaid onto plots of features from all the baselines. We define the homogeneous region as the area whose two-third majority of KNN within the plot belong to the same class. As the homogeneity implies less difficulty for classification, better features should result in larger overlays for all classes. Similar plots and overlays were created for radiomic features in both data sets for comparison such that the feature plots of successful baselines should be more homogenous than that of the existing features.

For quantitative evaluation, the representation baselines were compared using the classification performance. To further cope with the small medical data sets, these features were reduced using principal component analysis (PCA) retaining 99% of the original variance. Compared with t-SNE, PCA does not try to preserve the local neighborhood structure. However, it is widely used for its ability to retain large variance in the original feature space using projection to a few principal components (PCs). It also presents the possibility of producing independent features after projection as all the PCs are orthogonal. Low-dimensional independent features are desirable for subsequent classifiers.

After reduction, we used four commonly used ML algorithms as the final classifiers in the testing step to classify the medical data sets. All the algorithms, including support vector machine (SVM) with linear kernel, decision tree (DT), KNN, and multi-layer perceptron (MLP) with a 24-dimension hidden layer, were implemented using Scikit-learn package [31] with python. The reduced features from all the baselines were subsequently fed into the four ML classifiers for comparison of the medical prediction results. We also compared the performance with the strategy from our previous study, which used repetitive tuning [7].

To demonstrate that our strategy is applicable to other related tasks, we also used validated features from the Sarcoma data set for survival regression. It has been shown that features from deep learning models for survival classification can also be used for survival regression [30]. For this setting, the features were used to train a Cox proportional hazard model [32], which outputs a relative hazard value based on the recorded survival time. The results were compared using the concordance index (C-Index).

Experiment II: Our Strategy as the End Classification. In this experiment, we evaluated the performance of our adaptive scope strategy as the final classifier in the testing step. We compared its effectiveness with a standard KNN as the final classifier. We used two transfer learning approaches for features in the training and validation steps: the fixed-margin triplet network validated with triplet loss in (1) and the adaptive scope, and the AE network validated mean-squared-error (MSE) reconstruction loss. We also compared the result with an end-to-end softmax classifier validated with an accuracy measure. The classification accuracies of all the approaches were obtained from the testing data set for performance evaluation.

C. Computational Settings

In both experiments, triplet network, AE, and end-to-end softmax models were trained using the same architecture for the extractor network, the encoding network, and the network prior to softmax to ensure a fair comparison among different representation learning approaches. The organization of the dedicated and combined representation networks was similar to the rough description in Fig. 1. For 2-D and 3-D image inputs, four consecutive convolutional and max-pooling layers followed by a feed-forward network were used to process the data down to a low-dimensional vector. For each kind of vector input (e.g., the features from the pre-trained network, the radiomic features, and the expert variables), a feed-forward neural network was used to process the data into the same dimension as that of the image results. Regardless of the input types, a leaky rectified linear unit (one-ReLu) defined as $\max(0.01x, x)$ was used as the activation function in all convolutions and feed-forward layers. On top of the dedicated networks, a feed-forward neural network processed and combined the vectors of different kinds into unified features for the training of each specific representation learning strategy.

The numbers of network parameters in our models are different in both data sets. For the EPID data, the network for image data was (32-32-32-32) which referred to the number of 5 × 5 convolutional filters for each layer according to [7]. The subsequent feed-forward network was (8192-1024-128), which means that the network takes in an image input of 8192 dimensions and reduces down to 1024 and 128 dimensions at the end of the network. The feed-forward networks for the features from the pre-trained InceptResnetV2 and the radiomics features were (1536-128) and (17-128), respectively. At the final encoding part, a feed-forward network of (384-128) was used to digest the concatenated vector from the three types of features down to a unified feature vector of 128-dimension. Similarly, for the Sarcoma data, the filter size used in the 3-D image network was 3 × 3 × 3. The convolutional network was (16-32-64-128), and the subsequent feed-forward network was (8192-1024-128). The feed-forward networks corresponding to the radiomics features and the clinical variables were (45-128) and (27-128), respectively. The final encoding architecture of (384-128) was also used for the Sarcoma data.

All the triplet-based models in the experiment were subject to regularization in (25). The regularization weights were set to $w_{fm} = 1.000$, $w_{ms} = w_{md} = 1$, and $w_{sd} = 0$ for training with both the EPID and Sarcoma data sets. We found in prior experiments that setting a larger w_{sd} easily led to overfitting in training with our small data. Thus, we reduced it to 0 in our experiments. The margin value m in the training of fixed-margin triplet was set to 1 000 000. The end-to-end softmax and AE networks were trained under cross-entropy loss and MSE reconstruction loss, respectively. Unlike the triplet network and the end-to-end softmax approaches that backpropagate the error signal back to the rest of the organized network, the AE network was trained incrementally. Starting from a dedicated network, the network of each input type was trained using a decoding network with the same parameter setting as the encoding network but in reverse order. Deconvolution layers were used instead of convolutional layers in the decoding process. After dedicated training, the dedicated decoding network was removed. Then, the combined encoding and decoding networks were used to train for combined representation, which reconstructed the reduced vectors of all input types. After training, the combined decoding was removed leaving the final layer of the encoding network as the representation extractor.

All networks were trained until convergence using Adams optimizer [33] with the learning rate set to 0.0001. Batch sizes of all networks were set to 30 for the EPID and 20 for the Sarcoma data. For all networks, the maximum number of epochs was set to 1000. The training with the EPID data was done without data augmentation. For the Sarcoma data, simple augmentations of random flipping vertically and horizontally were applied to the image inputs. In all experiments, $k = \lceil \sqrt{n} \rceil$ was set for the simple KNN where n is the size of the training set. Specifically, $k = 17$ and 15 for the EPID and Sarcoma data sets, respectively. The same k values were set for the adaptive neighbor scope for both the initial search radius and the last-resort KNN when the method failed to find a neighbor.

TABLE III

COMPARISON OF THE TWO-CLASS ACCURACY (%) ON THE EPID DATA USING THE PROPOSED METHOD AGAINST THE DIFFERENT FEATURE VALIDATION CONFIGURATIONS. THE FEATURES WERE REDUCED WITH PCA RETAINING 99% VARIANCE AND FED INTO DIFFERENT ML ALGORITHMS AS THE FINAL CLASSIFIER

Configurations (# of features)		**End Classifier in Testing**			
Training	**Validation**	**SVM**	**KNN**	**DT**	**MLP**
Fixed Margin Triplet	Adaptive Scope (26)	**78.04**	73.73	74.12	65.88
	Triplet Loss (32)	77.65	73.73	72.94	73.33
AE (78)		54.17	69.80	69.80	61.18
Softmax (65)		52.78	70.59	74.12	62.75
Radiomics Features (10)		66.27	66.27	65.88	66.67
Best Repetitive Tuning [7]		74.51			

D. Computational Results

1) Experiment I: Our Strategy to Optimize the Triplet in the Validation Step: Visualization of the features from the proposed method and the other baselines is presented in Fig. 6. The plot overlays reflect the degree of difficulty in demarcating the decision boundary. Overall, the triplet network created more separable features for all classes. The figure illustrates larger homogeneous regions formed by triplet-based features compared with the smaller regions formed by those of the AE, softmax, and radiomics features. Some baseline plots were also overly dominated by a single class. The plots of the radiomic features largely presented non-homogeneous areas without the overlay suggesting that classification can be done more easily with deep representation approaches. However, the difference between the proposed and the typical validation was not clearly observed without quantitative evaluation.

The quantitative results of this experiment are summarized in Tables III–V. For the EPID data, the features that were trained, validated, and selected by the proposed approach outperformed the other baseline algorithms in almost all settings. The proposed strategy achieved the best testing accuracy of 78.04% and 69.02% on the two-class and three-class settings of the EPID data using SVM as the final classifier. Similarly, our strategy achieved the best testing accuracy of 66.67% using a simple KNN on the sarcoma data. When compared with the current state-of-the-art approach by our group [7], the strategy proposed in this article outperformed our previous strategy which involved careful manual tuning with repetitive selection. Most of the feature–classifier combinations from the softmax and AE baselines underperformed compared with the triplet-based approach. Nevertheless, the performances among the four classifiers in different strategies are consistent with the qualitative results as the better separated features from the proposed method achieved better results than that of the softmax and AE with less homogeneity.

The results in Table VI also showed superior performance of the proposed validation strategy in survival regression achieving the highest C-Index of 0.6590 using the Cox model. Some other representation models achieved C-Index near 0.5, which means the models learned almost nothing for regression. It is noted that the underlying regression context here is an extremely hard problem such that the sources and settings (sarcoma subtypes, locales, data collection standards) were very diverse. Moreover, the features were not derived specifically

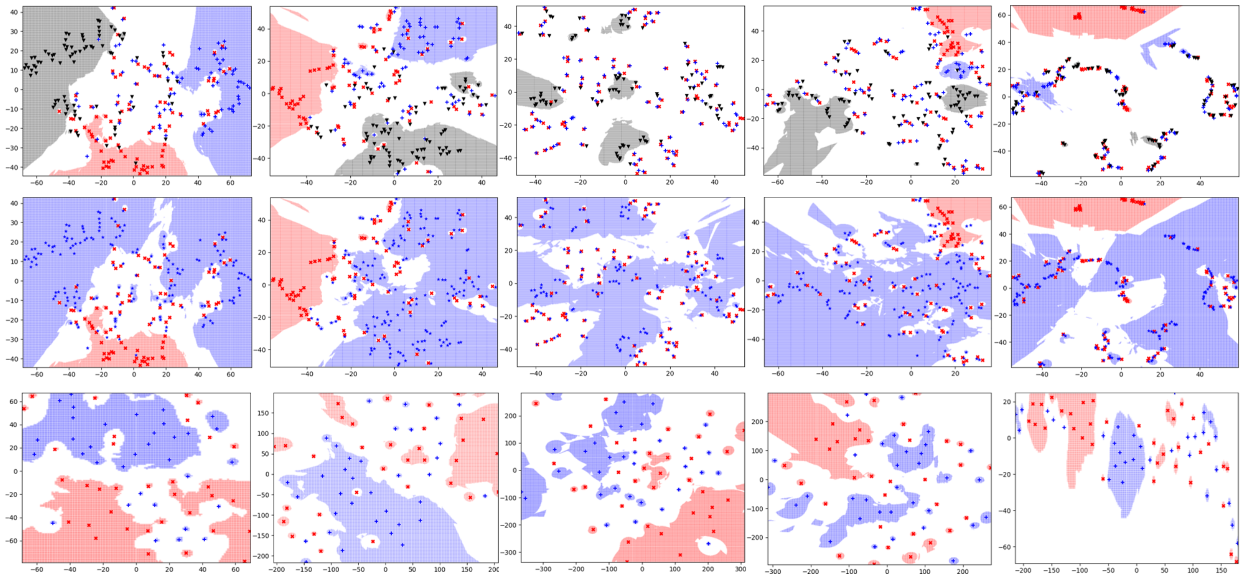

Fig. 6. Visualization of the validated testing features in experiment I after reduced by T-SNE to two dimensions along with additional plots of the expert-defined radiomics features reduced in the same manner. From top to bottom: the visualizations in each row are the plots from the EPID three-class, EPID two-class, and Sarcoma experiments. From left to right: the plots are from the triplet features validated with the adaptive scope, the triplet features validated with the triplet loss, the AE features, and the features from the end-to-end softmax network. For the EPID, "x," "+," "∇," and "*" denote the feature points from the no error, random error, systematic error, and combined error classes, respectively. For the Sarcoma, "x" and "+" denote feature points for the survive and non-survive classes, respectively. The scatter plots are overlayed with a class-based neighborhood such that the space in the neighborhood has two-third majority of KNN from the same plot belonging to one class. The larger neighborhoods imply a more homogeneous space and suitability for classification, whereas the larger white space implies less homogeneity and difficulty in classification. Overall, the triplet-based features are more homogeneous compared with that of the AE, softmax, and radiomics features.

TABLE IV

Comparison of the Three-Class Accuracy (%) on the EPID Data Using the Proposed Method Against the Different Feature Validation Configurations. The Features Were Reduced With PCA Retaining 99% Variance and Fed Into Different ML Algorithms as the Final Classifier

Configurations (# of features)		**End Classifier in Testing**			
Training	**Validation**	**SVM**	**KNN**	**DT**	**MLP**
Fixed Margin Triplet	Adaptive Scope (26)	**69.02**	66.27	63.53	52.94
	Triplet Loss (32)	67.06	63.92	61.18	59.61
AE (78)		54.17	61.57	54.12	45.49
Softmax (65)		52.78	52.55	54.51	41.57
Radiomics Features (10)		66.27	47.84	42.75	54.90
Best Repetitive Tuning [7]		67.78			

TABLE V

Comparison of the Two-Class Accuracy (%) on the Sarcoma Data Using the Proposed Method Against the Different Feature Validation Configurations. The Features Were Reduced With PCA Retaining 99% Variance and Fed Into Different ML Algorithms as the Final Classifier

Configuration (# of features)		**End Classifier in Testing**			
Training	**Validation**	**SVM**	**KNN**	**DT**	**MLP**
Fixed Margin Triplet	Adaptive Scope (19)	59.72	**66.67**	63.89	65.28
	Triplet Loss (17)	61.11	61.11	63.89	62.50
AE (65)		54.17	56.94	51.39	54.17
Softmax (49)		52.78	55.55	47.22	59.72
Radiomics Features (20)		51.39	54.17	52.78	44.44

for regression in the transferred models. Nevertheless, the performance trend in survival regression followed closely with that of the classification task, demonstrating that the proposed strategy can be an effective approach in related learning tasks.

2) *Experiment II: Our Strategy as the End Classification:*

The results of comparing classification models on both data

TABLE VI

Comparison of C-Index on the Sarcoma Data Using the Proposed Method Against the Different Feature Validation Configurations. The Features Were Reduced With PCA Retaining 99% Variance and Fed Into the Cox Proportional Hazard Regression

Configurations (# of features)		**C-Index**
Training	**Validation**	
Fixed Margin Triplet	Adaptive Scope (19)	**0.6590**
	Triplet Loss (17)	0.6026
AE (65)		0.4932
End-to-end Softmax (49)		0.5103

sets are presented in Table VII. Using the proposed adaptive scope on both the validation and testing outperformed all the other configurations using representation learned from fixed-margin triplet loss, AE, and softmax trained with the same architecture. The proposed method achieves the best accuracies of 73.73% and 67.45% on the two-class and three-class settings of the EPID data and 65.28% on the Sarcoma data. However, experiments showed varying results when the proposed approach was used only as the final classifier in the testing step. It can be observed that the proposed approach, as a classifier, gave only minor improvements for the representation from the fixed margin approach validated with the triplet loss, while the performance of the representation from the AE network declined, possibly due to overfitting. The results suggested that the approach has the capability to select good features among many training epochs and yields good end performance. However, it performs similar to a simple KNN and may overfit when it plays no role in feature validation.

Overall, the proposed validation strategy performs well as in the feature validation step. Although our strategy does

TABLE VII

COMPARISON OF ACCURACY (%) ON THE EPID DATA USING THE PROPOSED VALIDATION AND ADAPTIVE SCOPE CLASSIFICATION AGAINST THE OTHER VALIDATION CLASSIFICATION CONFIGURATIONS. THE FEATURE UNDERWENT NO REDUCTION PRIOR TO KNN OR OUR ADAPTIVE SCOPE CLASSIFICATION

Algorithm Configurations			**EPID**		**Sarcoma**
Training	**Validation**	**Testing**	**3-Class Acc (%)**	**2-Class Acc (%)**	**2-Class Acc (%)**
Fixed Margin Triplet	Adaptive Scope	Adaptive Scope	**67.45**	**73.73**	**65.28**
	Triplet Loss	Adaptive Scope	64.31	**73.73**	62.50
	Triplet Loss	KNN	63.92	73.73	61.11
AE	Reconstruction loss	Adaptive Scope	45.49	70.59	43.06
	Reconstruction loss	KNN	54.50	70.59	56.94
End-to-end Softmax			50.98	72.54	59.72

not directly affect typical fixed-margin training, it helps increase the performance of the classification without having to perform repeated validation and testing steps. Thus, our strategy may be best applied to the triplet network, whose metric loss can tune the NN hyperparameter. In general, the features from the fixed-margin triplet network achieved better performance compared with the other baselines in our small multi-view data sets. However, the results showed some limitations of our strategy on the feature embedding of AE and end-to-end softmax. We speculate that the underperformance of the AE was due to the complexity of the heterogeneous forms of inputs. Such inferior results compared with that of the class-based triplet training suggested that the encoded patterns represented by the latent features were not simplified enough to aid the classifier. It is also worth noting that the end-to-end softmax overfit our small data sets and underperformed compared with the triplet-based models; despite using the label data such that it had a quick convergence during training, the testing performance was poor. Nevertheless, the poor results (∼50%) in some settings do not mean the AE and softmax approaches learned nothing and gave random results. For the EPID data, the performances were comparable to the traditional threshold-based approach done by the experts in clinical setting (∼42%–49% [7]). For the Sarcoma data, softmax with end-to-end training result outperformed AE and was close to the typical triplet approach with KNN in Table VII. The results suggested that the softmax approach may learn better with end-to-end training. However, it may not be suitable for transfer learning in the multi-view sarcoma problem with the small sample size.

Our strategy is designed to investigate whether overall classification improvement should be emphasized on better representation or better subsequent classifier. Comparing the proposed measure against a baseline (e.g., the end-to-end softmax or the radiomic feature), outperformance means the features are suitable and can achieve better performance at the second stage. Otherwise, superior performance is uncertain. The representation network should then be reevaluated for improvement. For example, In Table VII scenario, the second-stage tuning was worth pursuing due to the superior lower bound performance than that of end-to-end softmax. The achievable lower bound in Table VII can then judge the second-stage classifiers such that the best classifiers from the results in Tables III–V should outperform the baseline performances in Table VII. Otherwise, more effort should be put into the classification stage rather than the feature stage.

The triplet results also present the drawback of triplet loss as a validation measure. In Table VII, the achievable lower bound results of features selected with triplet loss were inferior to that of the proposed method in the three-class EPID and Sarcoma tasks. The conclusion was also supported by the same trends of the second stage in Tables III–V.

V. CONCLUSION

We successfully developed a novel NN-based strategy for evaluating the unified features from the triplet-based classification training of multi-view medical data. The strategy provides a theoretical lower bound on classification performance of the features to aid training for the final classification. By comparing the lower bound, the strategy can be used to validate which training epoch generates the features with better potential for classification prior to the classification stage. The lower bound also determines whether the classifier drives better performance which reduces the burden of the repetitive tuning between the feature networks and the classifier for end performance. Our experimental results show that the triplet network has the potential to outperform the end-to-end softmax and AE networks in the classification tasks while retaining similar utility for transferring to other related tasks, such as survival regression. Our strategy may be useful in a setting where a limited sample size is available and data augmentation is limited or infeasible.

The ability to transfer the representation for other related tasks is particularly in line with modern medical research of which the patient's data along with the features are recorded and processed in a common workflow (e.g., radiomic workflow [18], [28]) such that the information can be archived for future study. In the bigger picture, our proposed strategy also suggests that the feature development for classification is neither data-specific nor dependent on the available expert knowledge. Our multi-view experiments demonstrated that relevant data sources could be used integrally in a unified framework. Thus, experts should focus on defining and curating the relevant data rather than tedious feature engineering.

This study also reveals challenges and opportunities for the interpretability of deep representation learning networks, which have been commonly criticized. The proposed method indicates which phase of the learning to improve but does not explain how to improve. There are many interesting expansions to tackle this challenge such as exploiting the first-stage information for classifiers which leads to tighter lower bounds, investigating how the adaptive methods cope with the outliers, and exploring more sophisticated end classifiers. These and other problems are subjects for future work.

REFERENCES

[1] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[2] N. Beig *et al.*, "Radiogenomic analysis of hypoxia pathway reveals computerized MRI descriptors predictive of overall survival in glioblastoma," *Proc. SPIE*, vol. 10134, Mar. 2017, Art. no. 101341U.

[3] C. Lynch, K. Aryafar, and J. Attenberg, "Images don't lie: Transferring deep visual semantic features to large-scale multimodal learning to rank," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 541–548.

[4] K. Somandepalli, V. Martinez, N. Kumar, and S. Narayanan, "Multimodal representation of advertisements using segment-level autoencoders," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, Oct. 2018, pp. 418–422.

[5] J. Lee, S. Abu-El-Haija, B. Varadarajan, and A. Natsev, "Collaborative deep metric learning for video understanding," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 481–490.

[6] C. Zhang, W. Tavanapong, G. Kijkul, J. Wong, P. C. de Groen, and J. Oh, "Similarity-based active learning for image classification under class imbalance," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2018, pp. 1422–1427.

[7] M. J. Nyflot, P. Thammasorn, L. S. Wootton, E. C. Ford, and W. A. Chaovalitwongse, "Deep learning for patient-specific quality assurance: Identifying errors in radiotherapy delivery by radiomic analysis of gamma images with convolutional neural networks," *Med. Phys.*, vol. 46, no. 2, pp. 456–464, Feb. 2019.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[9] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, Feb. 2017, pp. 4278–4284.

[10] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan, "Chest pathology detection using deep learning with non-medical training," in *Proc. IEEE 12th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2015, pp. 294–297.

[11] L. Rampasek and A. Goldenberg, "Learning from everyday images enables expert-like diagnosis of retinal diseases," *Cell*, vol. 172, no. 5, pp. 893–895, Feb. 2018.

[12] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Sep. 2017, pp. 603–611.

[13] A. B. Said, A. Mohamed, T. Elfouly, K. Harras, and Z. J. Wang, "Multimodal deep learning approach for joint EEG-EMG data compression and classification," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2017, pp. 1–6.

[14] N. Jaques, S. Taylor, A. Sano, and R. Picard, "Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Oct. 2017, pp. 202–208.

[15] W. Zhu *et al.*, "Neural multi-scale self-supervised registration for echocardiogram dense tracking," 2019, *arXiv:1906.07357*. [Online]. Available: http://arxiv.org/abs/1906.07357

[16] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. Int. Workshop Similarity-Based Pattern Recognit.* Cham, Switzerland: Springer, Oct. 2015, pp. 84–92.

[17] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk, "Learning local feature descriptors with triplets and shallow convolutional neural networks," in *Proc. BMVC*, Sep. 2016, vol. 1, no. 2, p. 3.

[18] P. Thammasorn, "Deep-learning derived features for lung nodule classification with limited datasets," *Proc. SPIE*, vol. 10575, Feb. 2018, Art. no. 105751F.

[19] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.

[20] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognit.*, vol. 48, no. 10, pp. 2993–3003, 2015.

[21] T. Peng, M. Boxberg, W. Weichert, N. Navab, and C. Marr, "Multi-task learning of a deep K-nearest neighbour network for histopathological image classification and retrieval," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2019, pp. 676–684.

[22] S. Sun and R. Huang, "An adaptive K-nearest neighbor algorithm," in *Proc. 7th Int. Conf. Fuzzy Syst. Knowl. Discovery*, vol. 1, Aug. 2010, pp. 91–94.

[23] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1774–1785, May 2018.

[24] R. O. Duda, P. E. Hart, and D. G. Stork, "Nonparametric techniques," in *Pattern Classification*, vol. 2, 2nd ed. New York, NY, USA: Wiley, 2012, ch. 7. Sec. 4, p. 163.

[25] N. Papernot and P. McDaniel, "Deep K-nearest neighbors: Towards confident, interpretable and robust deep learning," 2018, *arXiv:1803.04765*. [Online]. Available: http://arxiv.org/abs/1803.04765

[26] B. G. Vijay Kumar, G. Carneiro, and I. Reid, "Learning local image descriptors with deep siamese and triplet convolutional networks by minimizing global loss functions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5385–5394.

[27] P. Kinahan, L. Pierce, and M. Nyflot, *An Open-Science Toolkit for Image Texture Analysis of Pet Oncology Images*. Oak Brook, IL, USA: RSNA, 2015.

[28] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, Feb. 2016.

[29] M. B. Spraker *et al.*, "MRI radiomic features are independently associated with overall survival in soft tissue sarcoma," *Adv. Radiat. Oncol.*, vol. 4, no. 2, pp. 413–421, Apr. 2019.

[30] P. F. Christ *et al.*, "SurvivalNet: Predicting patient survival from diffusion weighted magnetic resonance images using cascaded fully convolutional and 3D convolutional neural networks," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 839–843.

[31] F. Pedregosa, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 10, pp. 2825–2830. 2011.

[32] J. Fox and S. Weisberg, "Cox proportional-hazards regression for survival data in R," in *An R Companion to Applied Regression*, 2nd ed. Thousand Oaks, CA, USA: SAGE Publications, Inc., Nov. 2010. Accessed: Feb. 21, 2021. [Online]. Available: https://socialsciences.mcmaster.ca/jfox/Books/Companion-2E/appendix/Appendix-Cox-Regression.pdf

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[34] A. O. Finley and R. E. McRoberts, "Efficient K-nearest neighbor searches for multi-source forest attribute mapping," *Remote Sens. Environ.*, vol. 112, no. 5, pp. 2203–2211, May 2008.

[35] Y.-C. Liaw, M.-L. Leou, and C.-M. Wu, "Fast exact k nearest neighbors search using an orthogonal search tree," *Pattern Recognit.*, vol. 43, no. 6, pp. 2351–2358, Jun. 2010.

[36] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," 2013, *arXiv:1304.5634*. [Online]. Available: http://arxiv.org/abs/1304.5634

[37] A. Serra, P. Galdi, and R. Tagliaferri, "Multiview learning in biomedical applications," in *Artificial Intelligence in the Age of Neural Networks and Brain Computing*. New York, NY, USA: Academic, 2019, pp. 265–280.

[38] Q. Dou *et al.*, "Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1182–1195, May 2016.

[39] L. Van Der Maaten and G. Hinton, "Visualizing data using T-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[40] V. Cheplygina, M. de Bruijne, and J. P. W. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Med. Image Anal.*, vol. 54, pp. 280–296, May 2019.

[41] L. S. Wootton, M. J. Nyflot, W. A. Chaovalitwongse, and E. Ford, "Error detection in intensity-modulated radiation therapy quality assurance using radiomic analysis of gamma distributions," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 102, no. 1, pp. 219–228, Sep. 2018.

[42] J. C. Peeken *et al.*, "CT-based radiomic feature predict tumor grading and have prognostic value in patients with soft tissue sarcomas treated with neoadjutant radiation therapy," *Radiotherapy Oncol.*, vol. 135, pp. 187–196, Jun. 2019.

[43] J. C. Peeken *et al.*, "Tumor grading of soft tissue sarcomas using MRI-based raiomics," *EBioMedicine*, vol. 48, pp. 332–340, Oct. 2019.

Phawis Thammasorn received the B.E. degree in computer engineering from Chiang Mai University, Chiang Mai, Thailand, in 2013, and the M.S. degree in computer science from the University of Southern California, Los Angeles, CA, USA, in 2016. He is currently pursuing the Ph.D. degree with the Department of Industrial Engineering, University of Arkansas, Fayetteville, AR, USA.

His primary research interest includes deep learning approach for predictive analytics on multi-modal data. His research interests include but are not limited to machine learning, artificial intelligence, data mining, computer vision, and multimedia processing.

Wanpracha A. Chaovalitwongse (Senior Member, IEEE) received the Ph.D. degree in industrial and systems engineering from the University of Florida, Gainesville, FL, USA, in 2003.

He was a Professor of industrial and systems engineering and radiology (joint) with the University of Washington, Seattle, WA, USA. He served as the Associate Director of the Integrated Brain Imaging Center, University of Washington Medical Center, Seattle. He currently holds a research professor position at the Department of Industrial Engineering, University of Arkansas, Fayetteville, AR, USA. His research interests include optimization and machine learning in neurophysiological and imaging data. He holds several patents for novel optimization techniques adopted in the development of seizure prediction system.

Dr. Chaovalitwongse was a recipient of several awards for his research, such as the NSF CAREER Award in 2005 and the William Pierskalla Best Paper for Research Excellence in Operations Research and Health Care Applications by the Institute of Operations Research and the Management Sciences twice in 2004 and 2008.

Daniel S. Hippe received the B.S. degree in electrical engineering and the M.S. degree in statistics from the University of Washington, Seattle, WA, USA, in 2004 and 2011, respectively.

He is currently a Statistician with the UW Department of Radiology with numerous collaborations within Radiology, including with the Vascular Imaging Laboratory and the Quantitative Breasting Imaging Laboratory, Seattle, WA, USA—and with research groups and departments outside Radiology, including I-LABS, Seattle, WA, USA, the Nghiem Laboratory, Seattle, WA, USA, Radiation Oncology, Dermatology, Cardiology, and Pediatrics. He enjoys working with investigators across a wide variety of disciplines and continually learning about different areas of medicine. The use of imaging is the common theme which connects the wide range of projects he works on.

Landon S. Wootton was born in Austin, TX, USA, in 1987. He received the B.S. degree in physics from The University of Texas at Austin, Austin, in 2008, and the Ph.D. degree in medical physics from the Graduate School of Biomedical Sciences, The University of Texas at Austin, in 2014.

From 2015 to 2017, he was a Medical Physics Resident at the University of Washington, Seattle, WA, USA, where he stayed on as a Faculty Member until 2020. He is currently the Medical Physicist for Baylor Scott and White Health in Round Rock, TX, USA. His research interests include scintillation dosimetry, neutron therapy, automation, and advanced image analysis.

Eric C. Ford received the B.S. degree in physics from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 1992, and the M.A. degree in physics and the Ph.D. degree in astrophysics from Columbia University in 1994 and 1997, respectively.

He is a Professor and Director of Medical Physics at the Department of Radiation Oncology, University of Washington School of Medicine, Seattle, WA, USA. His research interests are on patient safety and quality improvement, technology and techniques for radiobiology, and global radiation oncology. He has published more than 147 publications on the topics and authored the textbook *Primer on Radiation Oncology Physics*.

Dr. Ford currently serves in leadership capacities in the American Association of Physicists in Medicine (AAPM) and the American Society of Radiation Oncology (ASTRO).

Matthew B. Spraker received the bachelor's degree in biology from Indiana University in 2005 and the M.D. and Ph.D. degrees from the University of Illinois at Chicago, Chicago, IL, USA, in 2013 and 2009, respectively.

He joined the Department of Radiation Oncology, School of Medicine, Washington University at St. Louis, St. Louis, MO, USA, as an Assistant Professor, in 2018, after completing a residency in radiation oncology at the University of Washington, Seattle, WA, USA. His clinical research interests include sarcoma, clinical informatics, biomedical imaging, patient safety, and quality improvement.

Stephanie E. Combs was born in 1976. She received the M.D. degree with preclinical work in field of neuroanatomy, where she studied the sympathoadrenal system and the impact of growth factors on the development and formation of the neuronal and non-neuronal networks.

After her graduation and promotion in 2003, she has worked as a Research Associate in Heidelberg. Following her Post-Doctoral Lecture qualification in 2009, she was promoted to the Vice Chair of the Radiation Oncology Department, Heidelberg, in 2011. In 2014, she was appointed as a Professor and the Chair of the Department of Radiation Oncology, Technical University of Munich (TUM), Munich, Germany. In 2015, she also took over the Institute of Radiation Medicine, Helmholtz Zentrum. Since 2019, she heads the TUM senate. Her key expertise is highly conformal radiation therapy (stereotactic treatment, intensity-modulated radiation therapy (IMRT)/image-guided radiation therapy (IGRT)/adaptive radiation therapy (ART), protons, and carbon ions). Her scientific research interests include treatment optimization for brain and skull base tumors, biomarkers in radiation oncology, pediatric oncology, gastrointestinal oncology, uro-oncology, gynecological oncology, radiochemotherapy, and radioimmunotherapy.

Dr. Combs has been serving as a Board Member of the Neurooncological Working Group of the German Cancer Society (DKG). She received several scientific awards such as the Hermann Holthusen Prize of the German Radiation Oncology Society.

Jan C. Peeken (Member, IEEE) studied medicine at the University of Freiburg, Freiburg im Breisgau, Germany, with intermittent studies abroad at The University of Nice, Nice, France, and the Harvard Medical School, Boston, MA, USA. He received the M.D. degree for investigation in the pathophysiology of myeloproliferative disorders in molecular hematology, in 2018, and the Habilitation degree from the Medical Faculty, Technical University of Munich, Munich, Germany, in 2020. He is currently pursuing the Residency in Radiation Oncology with the "Klinikum rechts der Isar" University Hospital, Technical University of Munich. He is also serving as the Junior Group Leader of the Institute of Radiation Medicine, Helmholtz Zentrum Munich. In 2019, he visited the Department of Radiation Oncology, University of Washington, Seattle WA, USA, as a Research Fellow. His research interests include the applications of artificial intelligence techniques in radiation oncology.

Dr. Peeken's awards and honors include an ASH Abstract Achievement Award, a Young Investigator Travel Award, and the Award for Science and Research 2018 for his M.D. thesis funded by the ROMIUS Foundation.

Matthew J. Nyflot received the Ph.D. degree in medical physics from the University of Wisconsin in 2011 and was certified as a Diplomate of the American Board of Radiology in therapeutic medical physics in 2014.

He is currently the Medical Physicist and an Associate Professor with the Department of Radiation Oncology, with an adjunct appointment with the Department of Radiology, University of Washington, Seattle, WA, USA. His current research interests include the application of data science for precision cancer therapy and safety and quality in radiation therapy.