Center-Focused Affinity Loss for Class Imbalance Histology Image Classification

Taslim Mahbub, Ahmad Obeid, Sajid Javed, Jorge Dias, *Senior Member, IEEE*, Taimur Hassan, and Naoufel Werghi, *Senior Member, IEEE*

***Abstract*—Early-stage cancer diagnosis potentially improves the chances of survival for many cancer patients worldwide. Manual examination of Whole Slide Images (WSIs) is a time-consuming task for analyzing tumor-microenvironment. To overcome this limitation, the conjunction of deep learning with computational pathology has been proposed to assist pathologists in efficiently prognosing the cancerous spread. Nevertheless, the existing deep learning methods are ill-equipped to handle fine-grained histopathology datasets. This is because these models are constrained via conventional softmax loss function, which cannot expose them to learn distinct representational embeddings of the similarly textured WSIs containing an imbalanced data distribution. To address this problem, we propose a novel center-focused affinity loss (CFAL) function that exhibits 1) constructing uniformly distributed class prototypes in the feature space, 2) penalizing difficult samples, 3) minimizing intra-class variations, and 4) placing greater emphasis on learning minority class features. We evaluated the performance of the proposed CFAL loss function on two publicly available breast and colon cancer datasets having varying levels of imbalanced classes. The proposed CFAL function shows better discrimination abilities as compared to the popular loss functions such as ArcFace, CosFace, and Focal loss. Moreover, it outperforms several SOTA methods for histology image classification across both datasets.**

***Index Terms*—Data imbalance learning, fine-grained classification, histology image classification, supervised clustering, and whole slide image analysis.**

I. Introduction

CANCER is one of the primary causes of death worldwide and a significant obstacle to greater life expectancy in every country throughout the world [1]. According to the National Cancer Institute reports [2], approximately 6 billion USD was available for funding cancer research in 2020 alone, and this number is likely to rise as the demand for improved cancer diagnosis and treatment approaches increases. Histopathological diagnosis, which involves a pathologist reviewing high-resolution multi-gigapixel Whole Slide Images (WSIs), is a gold standard for diagnosing cancer [3]. Pathologists examine WSIs stained with Hematoxylin and Eosin (HE) dyes to perform cancer diagnosis, prepare a treatment plan, and evaluate key prognostic characteristics. If cancer is detected early, it can considerably affect the patient's mortality rate [4], [5].

However, the visual examination of WSIs requires an expert pathologist, and many hospitals and clinics lack such specialists [6]. In addition, manual diagnosis of these slides is time-consuming, prone to error, and subject to inter-observer variance [7]. Detecting cancer early utilizing automatic procedures by applying machine learning methods has been a critical development in recent years [7], [8], [9], [10], [11]. Typically, these techniques use deep convolutional neural networks (CNNs) to learn features from ground-truth data that expert pathologists label. While various variations of CNN models and networks have been developed to propose automated cancer detection systems, the issue of fine-grained image categorization and data imbalance remain understudied. Consequently, several existing classification frameworks are vulnerable concerning classification performance for under-represented groups [12].

While the majority of existing deep learning methods use cross-entropy/Softmax loss functions for training deep classifiers, various shortcomings of this loss have been explored, such as its lack of robustness to noisy labels [14], and its poor margin-based penalty [15], [16], [17], which can result in decreased generalization performance [18]. In the domain of histopathology classification, the increased generalization power of deep learning models is substantial [19]. Moreover, Softmax loss is less suitable for handling data imbalances that often occur in computational pathology datasets [20], [21]. "Softmax loss" refers to the softmax activation in the output layer of a neural network, followed by normalized cross-entropy loss. To overcome these limitations, in the current work, we propose a novel loss function, center-focused affinity loss (CFAL), that enhances the intra-class compactness of the learned features and efficiently learns characteristics from minority classes. Our proposed algorithm increases the discriminative performance of the vanilla affinity loss, which clusters and classifies data

Manuscript received 13 February 2023; revised 10 June 2023 and 6 October 2023; accepted 20 November 2023. Date of publication 24 November 2023; date of current version 6 February 2024. This work was supported in part by ASPIRE Institute, UAE, Ref: AARE20-279, and in part by the Terry Fox Foundation Ref: I1037. *(Corresponding author: Taslim Mahbub.)*

Taslim Mahbub, Ahmad Obeid, Sajid Javed, and Jorge Dias are with the Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, UAE (e-mail: 100060836@ku.ac.ae; ahmad.obeid@ku.ac.ae; sajid.javed@ku.ac.ae; jorge.dias@ku.ac.ae).

Taimur Hassan is with the Department of Electrical and Computer Engineering, Abu Dhabi University, Abu Dhabi, UAE (e-mail: taimur.hassan@adu.ac.ae).

Naoufel Werghi is with the Center for Secure Cyber-Physical Systems (C2PS) at Khalifa University, Abu Dhabi, UAE (e-mail: naoufel.werghi@ku.ac.ae).

The paper code is publicly available at https://github.com/Taslim-M/CFAL-max-margin-keras.

Digital Object Identifier 10.1109/JBHI.2023.3336372

© 2023 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

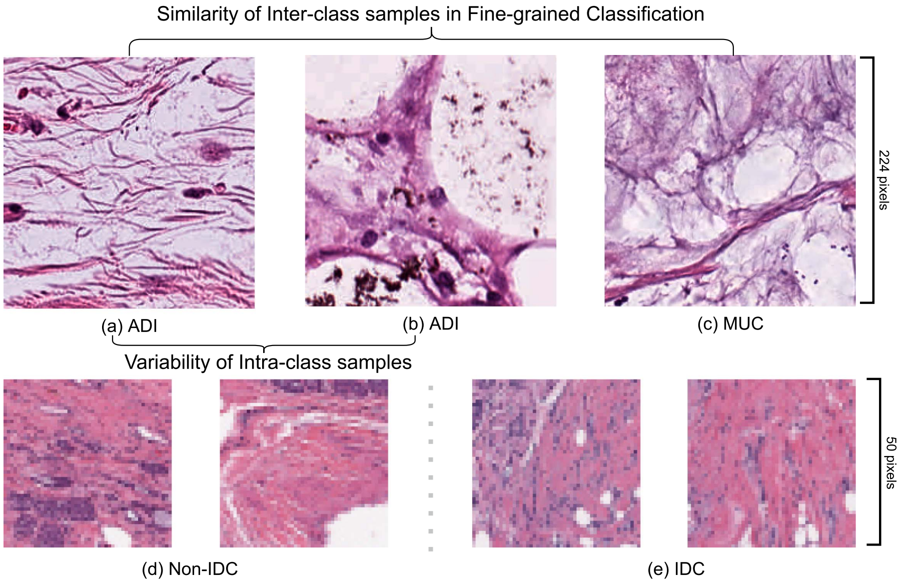

Similarity of Inter-class samples in Fine-grained Classification
224 pixels
(a) ADI
(b) ADI
(c) MUC
Variability of Intra-class samples
50 pixels
(d) Non-IDC
(e) IDC

Fig. 1. (a)-(c) Two of the nine classes from the colorectal cancer dataset [13] are represented by three random image patches where ADI refers to Adipose tissue, and MUC is Smooth Muscle tissue. (d) Breast cancer classification dataset showing non Invasive Ductal Carcinoma (IDC) patches and (e) IDC patches. They illustrate the challenge in fine-grained histopathology image classification because of the broad variability of image appearance within the same class and the high similarity between other classes.

using a single objective function. By applying CFAL, CNNs can extract more robust discriminative features from fine-grained histopathology datasets. We demonstrate this by benchmarking the classification performance on two distinct publicly available histology image classification datasets.

We summarize our main contributions as follows:

1) We propose a novel loss function termed center-focused affinity loss which enhances neural networks' intra-class compactness and discriminative power. Therefore, it improves the classification performance on fine-grained imbalanced histopathology datasets.
2) We evaluate the proposed loss function against three angular-margin-based loss functions (ArcFace, CosFace, SphereFace) and Focal Loss and assess their compatibility in the computational pathology domain.
3) We propose an ensemble network that employs feature extractor models trained with different loss functions and demonstrates improved classification.
4) A rigorous validation on two independent datasets demonstrates the superiority of our proposed algorithm compared to existing SOTA methods.

The rest of the paper is organized as follows. Section II describes the related work. Section III explains our proposed algorithm in detail including loss function, datasets, and the ensemble network. Section IV presents the experimental settings and Section V explains results and discussions, followed by the conclusion and future directions in Section VI.

II. Related Works

A. Histopathology Image Classification

Digital pathology is the process of capturing histology slides to create multi-gigapixel whole slide images (WSI), which has enabled the use of machine learning algorithms for detection, segmentation, and classification problems in the histopathology domain [22]. A WSI file can be gigabytes in filesize, making it difficult to load the entire WSI into memory for training ML models. These WSIs are commonly divided into smaller image patches, which are subsequently fed into the computer's memory for training ML models for various tasks. Modern methods based on deep learning (DL) have outperformed both traditional handcrafted feature methods, and traditional machine learning methods for histology image classification [8], [23]. In DL-based image classification tasks, CNNs are the most popular method for extracting features from an input image using convolution filters. These features are then passed through dense layers that are fully interconnected in order to capture the relationship between the extracted features without considering spatial information. Typically, the last layer of a CNN model is a softmax classifier, where each node output indicates the probability that an input image belongs to a specific class.

In this study, we employ labeled datasets at the patch level (see Fig. 1) to examine the impact of our proposed loss and ensemble network. One of these datasets is the histopathology image dataset of Invasive Ductal Carcinoma (IDC), the most prevalent phenotypic subtype of all breast cancers [23]. Each image patch is labeled as IDC negative (class 0) or IDC positive (class 1). The WSI naturally includes the majority of class 0 (healthy) patches. Cruz-Roa et al. [23] use two layers of convolutional and pooling layers with a tanh activation to extract features, followed by a fully-connected layer with a softmax activation for output. The cross-entropy loss was used, and the F1 score produced was 0.718 with a balanced accuracy of 84.23%, outperforming handcrafted image features with a Random Forest classifier. The authors of [22] achieve an F1-score of 0.7648 and an accuracy of 84.68 percent by applying the AlexNet model architecture. Similar CNN-based approaches were also applied to the 100,000 HE-stained colorectal cancer (CRC) dataset containing image patches from 9 tissue categories [24]. Kather et al. [24] demonstrate that the VGG19 model outperforms other models, namely AlexNet, SquezeNet, GoogLeNet, and Resnet50, by obtaining an accuracy of 94.3 percent on an external 7 k testing dataset. These studies rely on softmax activation followed by cross-entropy loss, which neither addresses the imbalance in data nor the fine-grained categorization challenge that naturally exists in these datasets. Parallel endeavors have been undertaken in the field of lung cancer pathology, wherein an Inception V3 network was employed to achieve an AUC score of 0.97, thereby substantiating the effectiveness of CNNs in cancer subtyping [25]. Similarly, in prostate cancer pathology classification, researchers [26] applied augmentation techniques to the NASNetLarge CNN architecture and attained a tumor detection accuracy of over 98%.

B. Imbalanced Learning

Several Face Recognition (FR) community researchers realized that softmax loss is insufficient for learning discriminative features. They advocated for novel loss functions that increase NN models' generalization capabilities [27]. This field of research has been one of the hottest subjects in the deep FR community, and the histopathology domain shares a similar challenge in which intra-class variances and inter-class similarities exist in the dataset. One method for learning discriminative feature embeddings is to learn a center for each class in Euclidean space and penalize the distance between a feature embedding and its class center, as proposed by Center loss [28]. Alternatively, angular-margin-based loss functions implement a cosine angular margin penalty to make the learned features more separable. Two notable examples are the CosFace [15], and ArcFace [16] loss functions, which overcame the optimization challenge of previously proposed angular-margin loss functions and obtained promising results in facial recognition tasks. Even though angular-margin-based loss can add discriminative constraints to a hypersphere manifold to improve decision boundaries, it has been demonstrated that it is sensitive when the datasets contain noisy data points [29], which can be observed in histopathology datasets [30]. Therefore, in computational pathology classification research, the strength of these strategies is yet to be investigated.

Recent literature also investigates various techniques to solve class-imbalanced learning because different real-world datasets are frequently skewed, with the majority of data belonging to a few dominant/over-represented classes and minority classes having relatively few data points. In the machine learning literature, two general strategies are used to address the problem of class data imbalance: data-level and algorithmic approaches. We can re-sample the data using the data-level method by either over-sampling the minority class or under-sampling the majority class. Under-sampling the majority class can result in losing vital information from some unique samples. Under-sampling is not viable in the medical domain because it can result in the loss of critical information required to generalize the ML model.

Meanwhile, oversampling can cause overfitting [31] and significantly increase training time. Regarding algorithmic solutions to the class-imbalance problem, we can implement cost-sensitive learning by assigning weights to each class [21], [31], or modify the loss function [20], [28], [32]. Modifying the loss function essentially changes how we train the model and provides a better theoretical foundation for solving specific learning issues with the softmax loss. Despite improvements in performance in natural image datasets, more research is needed to investigate the impact of these algorithmic solutions on fine-grained histopathology datasets.

III. Methodology

This paper proposes a loss function that alters the last layer activation function and the objective function to learn discriminative features from imbalanced fine-grained datasets. The block diagram of the proposed framework is shown in Fig. 2. In the subsequent paragraphs, we explain the mathematical formulation of the proposed CFAL.

A. Loss Function Formulation

To enhance the classification performance of fine-grained imbalanced pathology datasets, we propose an approach that integrates cost-sensitive learning and max-margin learning. We use the class-balanced loss function paradigm [31] to add cost-sensitivity into the learning process. By weighting class samples, this loss function aims to remedy the class imbalance. In addition, in order to implement max-margin learning, we apply improvements upon affinity loss that satisfy the criterion of enhancing margins across classes in an imbalanced dataset [20]. In the following paragraphs, we explain these two principles concisely.

1) Class-Balanced Loss Function: The authors of [31] define "effective number of samples" as the smallest subset of a given class that provides the most information. The effective number can be interpreted as the number of truly unique examples in the class, i.e., excluding duplicate samples or those exhibiting minor differences. For instance, naive data augmentation approaches that add tiny noise or do slight translations do not provide additional meaningful data for the model to learn. Therefore the effective number of samples does not rise due to data augmentation. In [31], the "effective number of samples" of a specific class y is approximated to:

$$E_{n_y} = \frac{1-\beta}{1-\beta^{n_y}} \tag{1}$$

where n_y is the number of samples in the ground-truth class y. $\beta \in [0, 1]$ is a parameter that regulates the rate at which the effective number grows as the number of samples increases. Normally, β should be unique to each class; however, for simplicity in parameter tuning, it is assumed that β is the same across all classes. In this work, we experiment with a range of values for β to determine the optimal value for a particular dataset. The class balanced loss function is therefore weighted by E_{n_y} rather than the number of samples in the class y, as is the case with

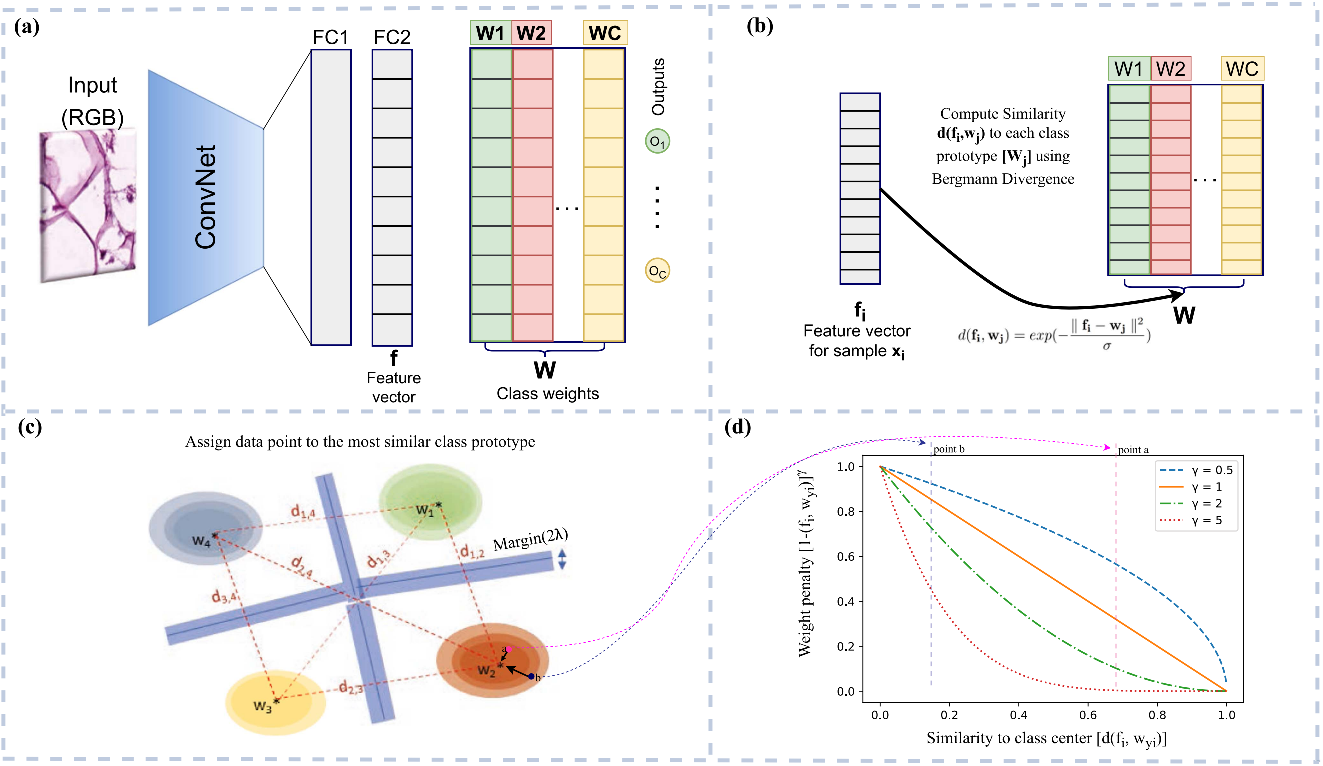

(a)
Input
(RGB)
ConvNet
FC1
FC2
W1
W2
WC
Outputs
O_1
O_C
f
Feature vector
W
Class weights
(b)
Compute Similarity
$\mathbf{d(f_i,w_j)}$ to each class
prototype $[\mathbf{W_j}]$ using
Bergmann Divergence
W1
W2
WC
$\mathbf{f_i}$
Feature vector
for sample $\mathbf{x_i}$
W
$d(\mathbf{f_i}, \mathbf{w_j}) = exp(-\frac{\| \mathbf{f_i} - \mathbf{w_j} \|^2}{\sigma})$
(c)
Assign data point to the most similar class prototype
w_1
w_2
w_3
w_4
$d_{1,4}$
$d_{2,4}$
$d_{1,3}$
$d_{1,2}$
$d_{3,4}$
$d_{2,3}$
Margin(2λ)
a
b
(d)
point b
point a
Weight penalty $[1-(f_i, w_{yi})]^\gamma$
Similarity to class center $[d(f_i, w_{yi})]$
0.0
0.2
0.4
0.6
0.8
1.0
γ = 0.5
γ = 1
γ = 2
γ = 5

Fig. 2. Overview of the proposed framework. (a) During the training phase, the weights (W) representing class prototypes are learned, where C is the number of classes. The learned hidden features (FC2) in Euclidean space are similar to its class prototype, where similarity is measured using (3). (b) The feature vector is assigned to an output class by using the similarity measure in (3) instead of the vector dot product commonly used in softmax. (c) By using the center-focused affinity loss, we ensure intra-class compactness and inter-class separation of these class prototypes. Unlike the vanilla affinity loss, the margin penalty for misclassification of minor classes is more strict in comparison to the major classes. (d) The penalty term to promote intra-class compactness. A point is very similar to its class-prototype center if the similarity value is close to 1 (e.g., point a) and away from the center if it is close to 0 (e.g., point b). Best viewed in color.

traditional class weighting approaches, as illustrated below:

$$\mathscr{CB}(\mathbf{p}, y) = \frac{1}{E_{n_y}} \mathscr{L}(\mathbf{p}, y) \tag{2}$$

where $\mathscr{CB}$ denotes the class-balanced loss function, $\mathscr{L}$ represents any employed loss function, and $\mathbf{p}$ is the prediction probability for the specific class y.

2) Affinity Loss Function: The affinity loss function proposed by [20] adopts the max-margin paradigm with Gaussian affinity and simultaneously learns classification and clustering. In contrast to Softmax activation, which employs the inner vector product $\langle \mathbf{w}, \mathbf{f} \rangle$ between the class weights $\mathbf{w} \in \mathbb{R}^{d \times C}$ and the feature vectors $\mathbf{f} \in \mathbb{R}^{d}$ at the last fully connected layer, the affinity loss function quantifies the similarity using the Gaussian similarity measure in terms of the Bergman divergence:

$$d(\mathbf{f_i}, \mathbf{w_j}) = \exp\left(-\frac{\| \mathbf{f_i} - \mathbf{w_j} \|^2}{\sigma}\right) \tag{3}$$

where σ refers to a weighting parameter. Contrary to Softmax activation, the above divergence measure enables margin maximization constraints to be enforced, hence enabling the loss function to reduce intra-class variability and raise inter-class distance by margin enforcement between each class. In addition, the authors use diversity regularization and multi-centered partitioning factors in the loss function for intra-class centroid delocalization and class imbalance mitigation. Let $\{X_i, Y_i\}$ denote the input-output pairs, C and N the number of classes and training samples, respectively. The feature space representation from the input samples is denoted by: $f_i, i = 1 : N$ and the class weights by $w_j, j = 1 : C$, is the class vectors. The affinity loss function is expressed as follows:

$$\mathscr{L}_a = \mathscr{L}_{mm} + R(w) \tag{4}$$

$$\mathscr{L}_{mm} = \sum_{j} \max\left(0, \lambda + d\left(\mathbf{f_i}, \mathbf{w_j}\right) - d\left(\mathbf{f_i}, \mathbf{w_{y_i}}\right)\right), j \neq y_i \tag{5}$$

$$R(w) = E\left[\left(\| \mathbf{w_j} - \mathbf{w_k} \|^2 - \mu\right)^2\right], j < k \tag{6}$$

where $i \in [1, N]$, $j \in [1, C]$, $d(\mathbf{f_i}, \mathbf{w_{y_i}})$ is the similarity with its true class, $d(\mathbf{f_i}, \mathbf{w_j})$ is the similarity of the sample with other classes, and λ denotes the enforced margin. The $R(w)$ is a 'diversity regularizer' term that enforces the class centers to spread out in the feature space, which ensures the learned features converge to a class prototype center (the idea behind center loss [28]), but the prototypes are equally spaced out to give a fair representation to all class samples. The μ is the mean distance between all class prototypes.

3) Proposed Center-Focused Affinity Loss (CFAL): We propose a synergic integration between the max-margin constraints with Gaussian affinity, a loss-agnostic modulating factor, and a local-penalty term to promote intra-class compactness to yield the center-focused affinity loss function, abbreviated as $\mathscr{L}_{cfal}$, expressed as follows:

$$\mathscr{L}_{cfal} = \frac{1}{E_{n_y}}(1 - d(\mathbf{f_i}, \mathbf{w_{y_i}}))^{\gamma} L_a(\mathbf{p}, y) \tag{7}$$

The class-balancing term ($\frac{1}{E_{n_y}}$) serves to penalize any misclassification of minor class samples more severely, hence focusing more on rare data examples during the training phase. This serves as a global objective to place greater emphasis on learning discriminative features from minority class samples. Meanwhile, the local penalty term ($(1 - d(\mathbf{f_i}, \mathbf{w_{y_i}})^{\gamma}$) enables each mini-batch to focus on difficult examples that are further away from the correct class center w_{y_i}. In turn, this enables our loss function to encourage intra-class compactness in order to learn robust and discriminative feature embeddings for each class in fine-grained datasets. γ is a focusing parameter to adjust the rate at which samples closer to the class-center are weighted down.

In addition to Softmax loss, we benchmark our loss function against focal loss, SphereFace, ArcFace, and CosFace loss functions. In the Focal loss, Lin et al. [32] re-shape the cross entropy loss to lessen the impact of easily classified samples and magnify the loss for hard, misclassified samples. They also add a class-wise weight α to increase the importance of the minority class. We replace the binary value of alpha with class-balanced weights for $\alpha \in \mathbb{R}^{1xC}$ using (2), where C is the number of unique classes. Similarly, SphereFace, ArcFace and CosFace loss functions were modified with the additional class-balanced term to provide a fair comparison to the proposed center-focused affinity loss function.

B. Model Compatibility

The proposed loss function ($\mathscr{L}_{cfal}$) is easily pluggable as a differentiable building piece into any deep neural network design. As illustrated in Fig. 2, the only network segment affected is the final output layer, where a custom layer block is inserted with the center-focused affinity loss as the loss function for training the model parameters. The final layer weights learn class prototypes for each class present in the dataset. The outputs are computed using the Bergman divergence (3) rather than the vector dot product employed by Softmax loss. To conduct the comparative study, all other parameters of the CNNs from the previous studies were maintained. The backbone CNN feature extractor was chosen based on the proposed models from previously published literature for each dataset outlined in the next section.

C. Ensemble Network

In an ensemble network, many networks with distinct or identical architectures are combined to function as one giant network. Typically, an ensemble will outperform a single network within the pack because, collectively, it is unlikely that the network will make the same specific mistake [33]. Several network designs, loss functions, and starting conditions can be utilized to achieve network diversity in an ensemble network. Nevertheless, employing different initialization conditions for the same network trained with the same loss function can only guarantee a varied representation of the same hidden features that are not explicitly optimized to be of different shapes [34]. In this work, we consider a diverse ensemble network that employs a homogenous feature extractor model but is trained with different loss functions to acquire heterogeneous feature embeddings. While previous research has attempted to use diverse ensemble networks by modifying the objective function [33], to our knowledge, this is the first study in this domain to use a single architecture for feature extraction followed by different learning paradigms to train the model to learn diverse features. By maintaining a common feature extractor for the backbone, we can decouple the problem of determining the optimal architecture for feature extractor networks and/or utilize a pretrained pathology feature extractor. Specifically, three max-margin-based loss functions and Focal-loss are employed to train four distinct networks that learn varied feature projections. This idea is represented in Fig. 3, where we can see that each model has learned something about a specific class that the others have not. The ensemble output was derived by applying three distinct aggregating procedures to the posterior classification probabilities of each individual network: (1) averaging, (2) maximum confidence, and (3) majority voting. The class with the highest confidence among the four networks is predicted as the final label in max confidence. The majority voting strategy permits each of the four networks to cast one vote for the class output, and the class with the most votes is predicted to be the final ensemble output.

TABLE I
INVASIVE DUCTAL CARCINOMA (IDC) DATASET DISTRIBUTION [22]

Class	Number of Samples	Percentage (%)
IDC Negative	198,738	71.61
IDC Positive	78,786	28.39
Total	*277,524*	*100.00*

IV. EXPERIMENTAL SETUP

This section presents a detailed description of the datasets we used in this study, along with the training and implementation details.

A. Dataset Description

As benchmarks, we validate our proposed loss function ($\mathscr{L}_{cba}$) using two distinct public-access histopathology datasets. First is the Invasive Ductal Carcinoma (IDC) dataset introduced by [23]. The dataset is heavily skewed towards IDC negative (non-cancerous) cases (Table I). The dataset contains a total of 277,524 patches measuring 50 × 50 pixels that were extracted from the WSI of 162 patients. We use the study in [22] to serve as a baseline to compare our results. AlexNet is used for classification with an F1-score of 0.7648 and a balanced accuracy of 84.68%. In addition to the benchmark results, AlexNet

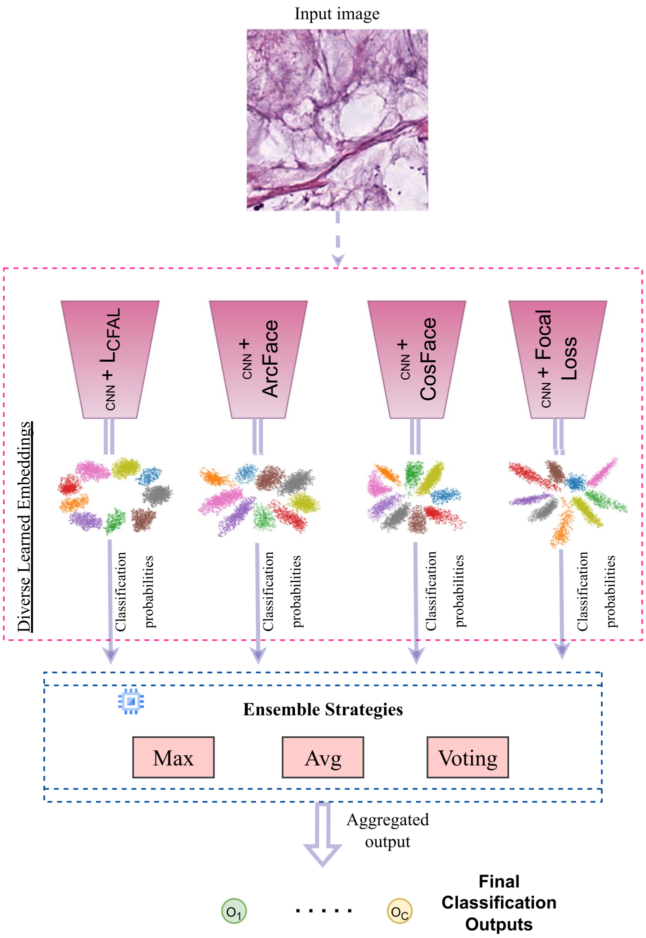

Input image
CNN + LCFAL
CNN + ArcFace
CNN + CosFace
CNN + Focal Loss
Diverse Learned Embeddings
Classification probabilities
Classification probabilities
Classification probabilities
Classification probabilities
Ensemble Strategies
Max
Avg
Voting
Aggregated output
O1
.....
OC
Final Classification Outputs

Fig. 3. Schematic illustrating the trained ensemble network with four different loss functions. To ensure diversity, each of the hidden features is optimized uniquely. Three distinct methodologies are employed to produce a comprehensive analysis in aggregating data from individual networks. The final aggregated output is used to evaluate the whole ensemble network. The 2D feature projections depicted above are from the CRC dataset, where $\mathscr{L}_{cfal}$ appears to have more separable learned 2D projections.

offers a lightweight feature extractor, which facilitates a broader exploration of additional hyperparameters (see Table III). This attribute is particularly advantageous in our project, as searching for an optimal hyperparameter range is crucial. Furthermore, the efficacy of AlexNet in pathology has been substantiated by previous research. For instance, the authors in [8] employed the base model of AlexNet to create a variant that has recently attained state-of-the-art classification accuracy on several pathology datasets. We split of the full IDC dataset to randomly assign 70% of data to the training set and 30% to the testing set, in accordance with the benchmark study.

Second, we utilize the colorectal cancer (CRC) histology slides dataset [13], which consists of 9 different tissue classes (Table II). The CRC dataset contains a total of 100,000 training image patches collected from 86 whole-slide images (WSI), with each image patch measuring 224 by 224 pixels. During the training epochs, 15% of the data from the training set was used for validation [13]. The training WSIs were extracted from the NCT biobank and the UMM pathology archive. The two datasets have varying degrees of data imbalance, allowing us to evaluate the performance of the suggested loss function against varying levels of data imbalance. Testing results for the CRC dataset are performed on an independent test dataset of 7,180 images, created using an additional 25 HE WSIs from the DACHS study from the NCT biobank. Using five prominent CNN models, researchers in [13] conclude that the VGG19 model trained using ImageNet weights performs the best with a test accuracy of 94.3%, which is the baseline. Beyond the superior performance of VGG19 in the task of colorectal cancer (CRC) classification, as compared to more recent models like ResNet [13], it also presents an opportunity to test the efficacy of our proposed loss function across models with diverse architectural depths. This is facilitated by the denser architecture of VGG19, which stands in contrast to the AlexNet model selected for the IDC dataset.

TABLE II
Colorectal Cancer (CRC) Dataset Distribution [13]

Tissue Class	Number of Samples	Percentage (%)
ADI (adipose tissue)	10,407	10.41
BACK (background)	10,566	10.57
DEB (debris)	11,512	11.51
LYM (lymphocytes)	11,557	11.56
MUC (mucus)	8,896	8.90
MUS (smooth muscle)	13,536	13.54
NORM (normal colon mucosa)	8,763	8.76
STR (cancer-associated stroma)	10,446	10.45
TUM (colorectal adenocarcinoma epithelium)	14,317	14.32
Total	*100,000*	*100.00*

B. Training and Implementation Details

The proposed loss function is implemented in TensorFlow 2.10.0 and Python 3.9.13. The CNNs were trained on an NVIDIA Quadro RTX 6000 GPU. Due to the large number of images in both datasets, 32-image batches were loaded using generators. The images in the IDC dataset were resized to 32 x 32 pixels to correspond with the experiment described in [22]. We empirically determine that λ (which denotes the enforced margin in the $\mathscr{L}_{cfal}$), σ (weighting parameter in the Gaussian similarly measure), γ (focusing parameter), and β (controls growth of the effective number of samples) are additional hyperparameters that can influence the performance of the model trained with the proposed $\mathscr{L}_{cfal}$). These hyperparameters are application-dependent (based on image properties) and must be fine-tuned based on the application domain and the dataset imbalance level. Likewise, SphereFace, ArcFace, and CosFace have two hyperparameters: the feature scaling parameter (s) and the angular margin (m). The values and ranges for each hyperparameter in the tuning search space are displayed in Table III. Early stopping is used with a patience of 5 to prevent over-fitting when the validation loss plateaus.

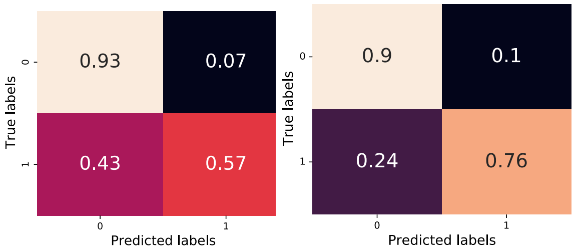
TABLE III
HYPERPARAMETER RANGES IN THE PARAMETER SEARCH

Loss	Hyperparameter	Values
All	Beta (β)	0.9, 0.99, 0.999, 0.9999, 0.99999
$\mathscr{L}_{cfal}$	Lambda (λ)	0.1 – 0.9 (increments of 0.1)
$\mathscr{L}_{cfal}$	Sigma (σ)	80 – 430 (increments of 50)
$\mathscr{L}_{cfal}$	Gamma (γ)	0.5, 1.0, 2.0, 5.0
ArcFace/ CosFace/ SphereFace	Scaling (s)	1.0, 5.0, 10.0, 15.0, 20.0, 30.0, 64.0
ArcFace/ CosFace/ SphereFace	Angular margin (m)	0.1, 0.35, 0.5, 0.7, 0.9

TABLE IV
CLASSIFICATION RESULTS WITH IDC DATASET WITH DIFFERENT LOSS FUNCTIONS

Model: AlexNet	Precision	Recall	F1-Macro	Accuracy (%)
Cross-entropy Loss [22]	0.80	0.74	0.76	84.7
Class-Balanced SphereFace	0.80	0.79	0.80	83.7
Center-Focused Affinity Loss ($\mathscr{L}_{cfal}$)	0.83	0.83	**0.83**	**86.0**
Class-Balanced ArcFace	0.83	0.81	0.82	85.7
Class-Balanced CosFace	0.83	0.82	0.82	85.7
Class-Balanced Focal Loss	0.81	**0.84**	0.82	84.9
Ensemble: Average Strategy	0.83	0.83	0.83	86.2
Ensemble: Max Strategy	0.83	0.82	0.82	86.0
Ensemble: Voting Strategy	0.83	0.83	0.83	**86.3**

Top metrics from individual models (black) and ensemble networks (brown) are highlighted in bold.



(a) CCE (b) $\mathscr{L}_{cfal}$

Fig. 4. IDC Dataset confusion matrix normalized over true(row) values. Class 1 is the IDC Positive (cancerous) minor class. (a) AlexNet with Categorical cross-entropy (CCE) loss (F1-score 0.77, Accuracy 83.5%). (b) AlexNet with $\mathscr{L}_{cfal}$ (F1-score 0.83, Accuracy 86%).

V. RESULTS AND DISCUSSION

In this section, we present a detailed evaluation of the proposed framework and its comparison with state of the art works. Section V-A presents the results for the benchmark datasets, followed by a comparison with SOTA works in Section V-B. In Section V-C, we present an analysis of the selected hyperparameters, followed by a series of additional ablation studies to to display our proposal's versatility further.

A. Comparative Study

On the IDC dataset, Table IV illustrates how the suggested center-focused affinity loss improves classification performance in contrast to the AlexNet with categorical cross-entropy loss (CCE) and softmax activation employed by [22]. The best model from their research is utilized as a benchmark. The hyperparameter tuning results are presented in Table IX. Since accuracy is not a comprehensive evaluation criterion, the F1 score is the most crucial statistic in imbalanced datasets. With the implementation of the proposed center-focused affinity loss ($\mathscr{L}_{cfal}$), the F1-score increases by 7% (from 76% to 83%), indicating an improvement in the recall and precision of minor class samples. This is further illustrated in the confusion matrix plot (normalized over true/row values) in Fig. 4. In addition, our proposed loss ($\mathscr{L}_{cfal}$) outperforms SphereFace, ArcFace, CosFace, and Focal Loss for the same network concerning F1-Score and accuracy metrics. Since SphereFace is unable to outperform the benchmark CCE study, we omit this model from the ensemble networks. We conducted a paired t-test to analyze the significance of the improvements. At a p-value of 0.05, we identified significant advancements in the precision metric over CCE, SphereFace, and Focal Loss. Likewise, we observed notable improvements in the recall metric relative to CCE, SphereFace, and ArcFace. The F1 Score also exhibited significant enhancement compared to CCE and SphereFace, and the top-1-accuracy showed marked improvement over CCE, SphereFace, and Focal Loss. The ensemble networks can leverage the unique feature representations learned by the 4 models and further improve the performance, with the voting strategy yielding the highest overall top-1 accuracy (86.3%).

The results for the comparative study with the CRC dataset are summarized in Table V. This subset of experiments underwent an independent hyperparameter tuning, and the optimal parameters are listed in Table IX. The F1-macro and AUROC scores were determined using the trained VGG19 model and code as provided by [13]. As seen in Table V, a 2.0% increase in the accuracy metric is observed by utilizing the class-balanced affinity loss ($\mathscr{L}_{cfal}$) instead of softmax loss. In addition, the F1-macro score has significantly increased from 0.92 to 0.95 (a 3% increase), indicating the general improvement of precision and recall measures across all classes. In particular, the STR (cancer-associated stroma) class which performs very poorly with softmax loss, is shown to significantly improve by the center-focused affinity loss, as seen in the confusion matrix in Fig. 5. Moreover, Table V demonstrates that LCFA is superior to SphereFace, ArcFace, CosFace, and Focal loss. Utilizing the

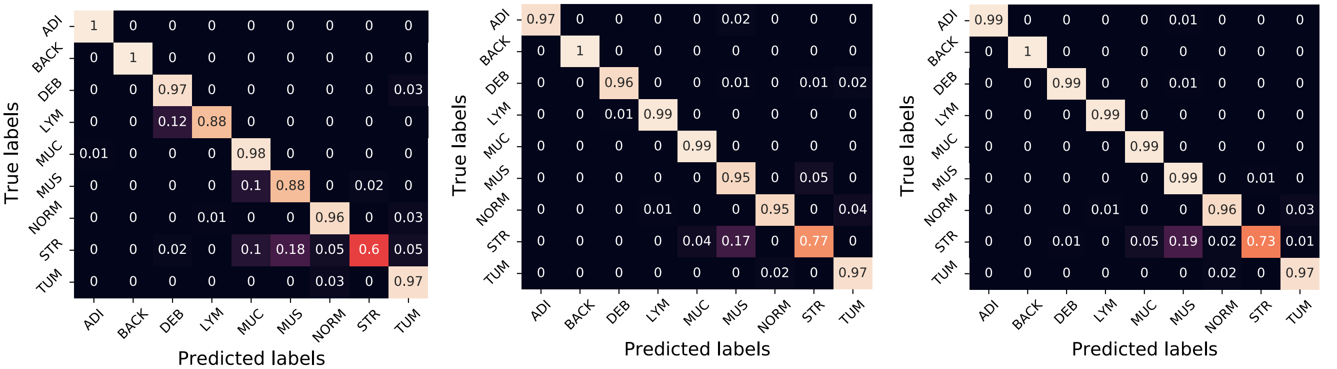


Fig. 5. CRC dataset confusion matrix normalized over true (row) values with (a) VGG19 with CCE (F1-score 0.92, Accuracy 94.3%). (b) VGG19 with $\mathscr{L}_{cfal}$ (F1-score 0.95, Accuracy 96.3%). (c) Ensemble network with Voting aggregation (F1-score 0.96, Accuracy 97.0%).

TABLE V
CLASSIFICATION RESULTS WITH COLORECTAL CANCER (CRC) DATASET WITH DIFFERENT LOSS FUNCTIONS

Model: VGG19	Precision	Recall	F1-Macro	Accuracy (%)
Cross-entropy Loss [13]	0.93	0.92	0.92	94.3
Class-Balanced SphereFace	0.94	0.93	0.93	94.0
Center-Focused Affinity Loss ($\mathscr{L}_{cfal}$)	0.95	**0.95**	0.95	**96.3**
Class-Balanced ArcFace	0.95	0.94	0.95	95.8
Class-Balanced CosFace	0.95	0.94	0.95	95.9
Class-Balanced Focal Loss	0.95	0.94	0.95	95.9
Ensemble: Average Strategy	0.96	0.96	0.96	96.8
Ensemble: Max Strategy	0.95	0.95	0.95	96.2
Ensemble: Voting Strategy	**0.97**	0.96	0.96	**97.0**

Top metrics from individual models (black) and ensemble networks (brown) are highlighted in bold.

TABLE VI
CLASS-WISE AUROC FOR COLORECTAL CANCER (CRC) DATASET WITH DIFFERENT LOSS FUNCTIONS

Class	Cross entropy [13]	$\mathscr{L}_{cfal}$	Ensemble: Voting Strategy
ADI	0.997	0.984	0.993
BACK	1.000	1.000	1.000
DEB	0.979	0.977	0.996
LYM	0.941	0.993	0.996
MUC	0.983	0.993	0.995
MUS	0.936	0.966	0.987
NORM	0.976	0.965	0.977
STR	0.800	0.885	0.863
TUM	0.981	0.984	0.983

t-test (p-value 0.05), we discerned significant enhancements in the precision metric of LCFAL relative to CCE and SphereFace, the recall metric over CCE, SphereFace, ArcFace, and CosFace, the F1-Score in comparison to CCE and SphereFace, and the accuracy metric over CCE, SphereFace, and ArcFace. In line with prior results, SphereFace did not surpass the benchmark study in terms of accuracy. Consequently, this model was excluded from the ensemble networks. Nevertheless, by combining the four networks into a single ensemble network, the overall performance is often superior to that of the individual networks, demonstrating significant improvements across all metrics (p-value < 0.05). In both datasets, the ensemble network's averaging and majority voting strategies outperform the maximum confidence strategy. This indicates that one model is more likely to predict the incorrect class label, but the error decreases when all models contribute to the final probabilities. In terms of accuracy and F1-macro, the majority voting ensemble is the most effective, with a 1% improvement over the top-performing individual model. With respect to class-wise metrics, Table VI shows that $\mathscr{L}_{cfal}$ outperforms CCE loss in 5 of the 9 classes, whereas the majority-voting ensemble network outperforms the CCE loss-based network in 6 of the 9 classes (with BACK class having the same score across all three losses).

B. Comparison With State-of-the-Art Works

The proposed method is compared to previous studies performed to classify the same benchmark datasets. In terms of F1-score, Table VII demonstrates that our work outperforms numerous SOTA methods in the IDC dataset

TABLE VII
COMPARISON OF RECENT SOTA STUDIES ON THE IDC DATASET VERSUS THE PROPOSED METHODS

Method	Accuracy (%)	F1-Score
Custom CNN [23]	84.2	0.72
AlexNet [22]	84.7	0.76
VGG16 [8]	81.1	-
AlexNet-BC [8]	86.3	-
Proposed - AlexNet ($\mathscr{L}_{cfal}$)	86.0	0.83
Proposed - Ensemble Network	86.3	0.83

TABLE VIII
COMPARISON OF RECENT SOTA STUDIES ON THE CRC DATASET VERSUS THE PROPOSED METHODS

Method	Accuracy (%)	F1-Score
VGG19 [13]	94.3	0.92
DenseNet [35]	91	0.90
Xception [35]	94	0.93
Bespoke CNN [35]	92	0.92
Ensemble (DenseNet, Xception, CNN, Inception-ResNetV2) [35]	96.2	-
Proposed - VGG19 ($\mathscr{L}_{cfal}$)	96.3	0.95
Proposed - Ensemble Network	97.0	0.96

and achieves the maximum level of precision with a recent study. In Table VIII, a comprehensive evaluation of the CRC dataset reveals that it outperforms heavier models and a more recent ensemble network with different backbones.

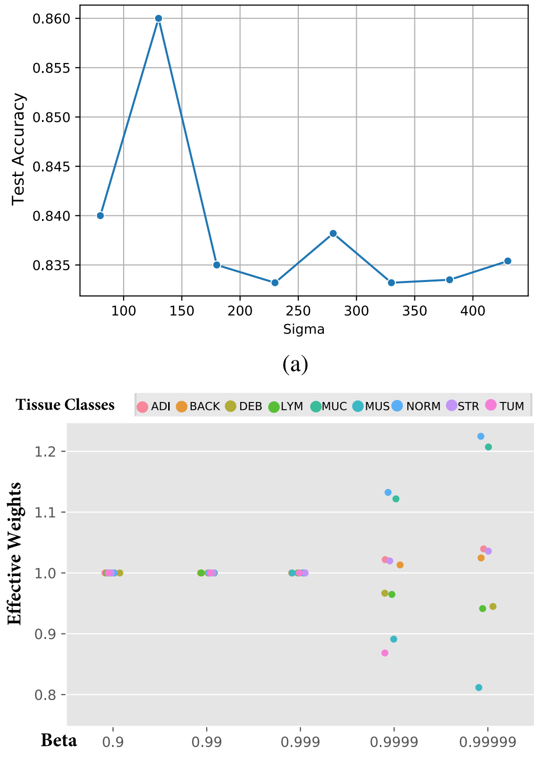
C. Selected Hyperparameters and Training Time

In addition to the already extensive search space in the neural network models, our proposed balanced affinity loss function adds three additional hyperparameters: sigma (σ), lambda (λ), gamma (γ), and beta (β). Each of these hyperparameters' ranges must be empirically defined based on the application domain (Table III). Table IX summarizes the optimal hyperparameters for each model across the datasets. The lambda and sigma values corresponding to the proposed loss were identical to those of the AlexNet model on the IDC dataset, indicating that the aforementioned hyperparameters for $\mathscr{L}_{cfal}$ training may be applied to a variety of histopathological image datasets.

Fig. 6(a) depicts the effect of the (σ) hyperparameter on the maximum test accuracy for the IDC dataset using the AlexNet model and $\mathscr{L}_{cfal}$. Fig. 6(b) illustrates how the beta value affects the effective weights allocated to each class in the CRC dataset. For values 0.999 and lower, the variation is negligible and close to 1 (standard weight). As the beta value increases, the learning process pays more attention to under-represented classes, as these data samples are likely to be unique samples with crucial characteristics that do not reappear in the dataset.

TABLE IX
BEST PERFORMANT MODEL HYPERPARAMETERS

Loss	Hyperparameter	IDC	KCRC
$\mathscr{L}_{cfal}$	Beta (β)	0.999	0.9999
$\mathscr{L}_{cfal}$	Lambda (λ)	0.1	0.1
$\mathscr{L}_{cfal}$	Sigma (σ)	130	130
$\mathscr{L}_{cfal}$	Gamma (γ)	2.0	1.0
SphereFace	Scaling (s)	30.0	20.0
SphereFace	Angular margin (m)	0.35	0.35
ArcFace	Scaling (s)	30.0	15.0
ArcFace	Angular margin (m)	0.5	0.1
CosFace	Scaling (s)	10.0	20.0
CosFace	Angular margin (m)	0.5	0.35



(a)

(b)

Fig. 6. (a) Impact of sigma values on the test accuracy of the IDC dataset with AlexNet. (b) Impact of beta values on the effective weights for each class in the CRC dataset. A higher effective weight results in more attention to the particular class during training.

The proposed loss function ($\mathscr{L}_{cfal}$) expands the search space for model tweaking during the training phase by adding more hyperparameters. Further, our investigations demonstrate that sigma and lambda values might vary significantly between datasets. For instance, a low value of sigma (10) can work well with the MNIST dataset, but for the histopathology datasets,

TABLE X
RUNTIME COMPARISON OF THE PROPOSED LOSS VERSUS BENCHMARK LOSS FUNCTIONS FOR EVERY EPOCH IN SECONDS

Loss	Average Time (SD) per model	
	AlexNet (IDC)	VGG19 (CRC)
CCE	53 (1.1)	214 (0.4)
$\mathscr{L}_{cfal}$	87 (5.3)	248 (1.6)
ArcFace	56 (5.3)	217 (3.3)
CosFace	57 (0.6)	215 (0.3)
Focal Loss	83 (1.2)	224 (0.5)

a higher value (130) yields better results. Beta governs the growth of effective number as the sample size increases, and the optimal value of Beta relies on the degree of imbalance in the dataset. Thus, with greater control over the training phase of a network, $\mathscr{L}_{cfal}$ introduces a bigger search space for the optimal hyperparameters. We achieve the same optimal values for sigma (130) and lambda (0.1) for both histopathology dataset studies, which can be used as a baseline for subsequent research in the same domain.

Furthermore, the additional penalty term in our proposed loss function, the training time per batch within an epoch will increase due to the additional comparisons. This is proven by the runtime comparison presented in Table X. Experimental outcomes indicate that angular-margin loss functions exhibit comparatively faster performance with minimal computational overhead relative to CCE. Conversely, both CFAL and focal loss demonstrate a pronounced computational expense, underscoring the significant contribution of the local penalty term to the computational overhead.

D. Ablation Study

1) Loss Function: We experimentally analyze the effect of adding the class-balancing and the penalty term to the vanilla affinity loss. Simply adding the class-balanced term yields a class-balanced affinity loss, expressed as follows:

$$\mathscr{L}_{cba} = \frac{1}{E_{n_y}} \mathscr{L}_a(\mathbf{p}, y) \tag{8}$$

As seen in Tables XI and XII, our proposed loss ($\mathscr{L}_{cfal}$) outperforms the vanilla implementation as well as the class-balanced version of the loss across both the datasets. Preliminary results of the class-balanced variant of the loss function were introduced in our ICPR workshop paper [36]. The best-determined hyperparameters (Table IX) were reutilized in this investigation.

TABLE XI
ABLATION STUDY WITH IDC DATASET COMPARING VANILLA AFFINITY LOSS ($\mathscr{L}_a(\mathbf{P}, y)$) WITH $\mathscr{L}_{cba}$ AND $\mathscr{L}_{cfal}$

Model: AlexNet	F1-Macro	Accuracy (%)
Vanilla Affinity Loss ($\mathscr{L}_a$)	0.80	84.0
Class-Balanced Affinity Loss ($\mathscr{L}_{cba}$) [36]	0.81	84.5
Center-Focused Affinity Loss ($\mathscr{L}_{cfal}$)	0.83	86.0

TABLE XII
ABLATION STUDY WITH CRC DATASET COMPARING VANILLA AFFINITY LOSS ($\mathscr{L}_a(\mathbf{P}, y)$) WITH $\mathscr{L}_{cba}$ AND $\mathscr{L}_{cfal}$

Model: VGG19	F1-Macro	Accuracy (%)
Vanilla Affinity Loss ($\mathscr{L}_a$)	0.94	95.4
Class-Balanced Affinity Loss ($\mathscr{L}_{cba}$) [36]	0.95	96.0
Center-Focused Affinity Loss ($\mathscr{L}_{cfal}$)	0.95	96.3

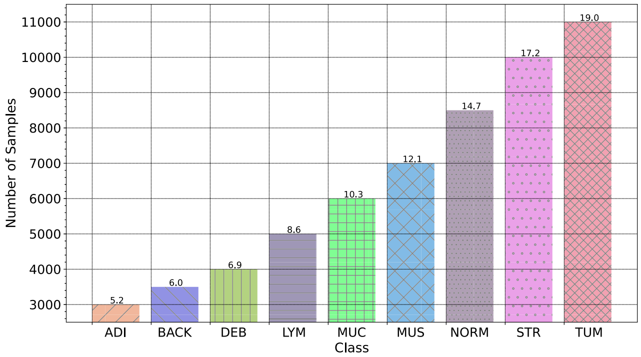


Fig. 7. Data distribution inside the skewed long-tail CRC dataset. The bar-labels show the percentages, where ADI's part of the data is 5.2% and TUM's share is 19.0%.

2) Long-Tailed CRC Dataset: We transform the CRC dataset into a long-tailed distribution to determine whether our suggested loss function is more robust against a more skewed data distribution. After randomly sampling a specific number of images from each category, the skewed dataset has 58,000 images, shown in Fig. 7. The best models (Table IX) were trained with the skewed dataset, and the testing dataset remains unchanged. Table XIII demonstrates that our proposed loss ($\mathscr{L}_{cfal}$) performs better than cross-entropy, ArcFace, CosFace, and Focal Loss across metrics like Precision, Recall, and F1-Macro. Given the long-tailed distribution of the data, accuracy may not serve as a representative metric for holistic performance evaluation. Consequently, we have excluded it from the results table. Using the t-test (p-value 0.05), significant improvements in Recall and F1-Score were observed over benchmark loss functions. Precision metric showed significant enhancement against all loss functions, barring ArcFace.

3) Validating With Transformer-Based Model: We demonstrate that our proposed loss function is compatible with

TABLE XIII
CLASSIFICATION RESULTS WITH THE LONG-TAILED COLORECTAL CANCER (CRC) DATASET WITH DIFFERENT LOSS FUNCTIONS

Model: VGG19	Precision	Recall	F1-Macro
Cross-entropy Loss	0.91	0.90	0.90
Center-Focused Affinity Loss ($\mathscr{L}_{cfal}$)	0.94	**0.94**	**0.94**
Class-Balanced ArcFace	0.92	0.91	0.91
Class-Balanced CosFace	0.94	0.92	0.93
Class-Balanced Focal Loss	0.92	0.90	0.91

Top metrics from individual models are highlighted in bold.

TABLE XIV
CLASSIFICATION RESULTS WITH COLORECTAL CANCER (CRC) DATASET WITH DIFFERENT LOSS FUNCTIONS AND TRANSFORMER-BASED MODEL

Model: CCT	Precision	Recall	F1-Macro	Accuracy (%)
Cross-entropy Loss	0.91	0.90	0.90	92.0
Center-Focused Affinity Loss ($\mathscr{L}_{cfal}$)	0.94	**0.95**	**0.95**	**95.5**
Class-Balanced ArcFace	0.94	0.94	0.94	95.0
Class-Balanced CosFace	0.94	0.93	0.93	94.6
Class-Balanced Focal Loss	0.93	0.94	0.93	94.5

Top metrics from individual models are highlighted in bold.

transformer-based models, which have recently gained popularity for medical image classification [37]. In the tokenization step, we employ a convolutional-based projection since applying convolutions captures spatial and low-level data and has improved numerous applications [37], [38]. In particular, the Compact Convolutional Transformer (CCT) model is applied, which uses a sequence pooling layer following the transformer attention layer [39]. The CCT model has outperformed the ViT transformer in many image classification tasks, especially for smaller datasets [40], [41]. Additionally, CCT models uniquely facilitate training from scratch on datasets by integrating the robust feature extraction capabilities of convolutional layers with the attention mechanisms of transformers, a distinctive attribute not typically exhibited by other vision transformers [39]. Furthermore, the CCT model is opted due to its lightweight architectural design. This characteristic renders it particularly suitable for our project, as it facilitates a more extensive exploration of hyperparameters, as detailed in Table III. In our experimental setup, we employed two transformer layers coupled with two convolutional layers and a projection dimension of 128. This configuration resulted in a total of 400 k trainable parameters, thereby enabling rapid training and the testing of numerous parameter variations.

The classification performance of the CCT model with various loss functions is presented in Table XIV. Compared to softmax-based cross-entropy loss, the performance of margin-based penalty losses and focal loss is superior. The proposed $\mathscr{L}_{cfal}$ loss has the highest performance with an F1-score of 0.95 and an Accuracy of 95.5, which are 3 and 3.5 percentage points higher than the CCE loss, respectively.

VI. CONCLUSION

This paper presented a novel center-focused affinity loss ($\mathscr{L}_{cfal}$) function that improves the performance of imbalanced data distribution in fine-grained histopathology image classification. The key idea is to utilize the proposed penalty term and class-balanced factor to improve the effectivity of the affinity loss function, enabling it to learn the uniform-sized equidistant clusters in the feature space, hence enhancing class separability and minimizing intra-class disparities. These characteristics of the loss function are essential for addressing the fine-grained and imbalanced data aspects of histopathology datasets, which conventional softmax loss cannot overcome. We validated our proposed method across two publicly available datasets, demonstrating a superior performance compared to the existing literature. Compared to ArcFace, CosFace, and Focal loss, our loss function displays an overall better classification performance. To the best of our knowledge, this is the first attempt to use a max-margin-based loss function paradigm to address the fine-grained and imbalanced characteristics of histopathology datasets. We also propose an ensemble network to leverage the unique feature embeddings obtained by training the models with various loss functions. The caveat of the proposed loss function includes an extensive search space due to the four new hyperparameters that require tuning based on the application context and the level of data imbalance. In future works, the proposed methodology could be evaluated within the context of weakly supervised pathology classification tasks, wherein only slide-level labels are available. Given that the process of labeling each tile or patch is resource-intensive, learning at the slide-level could capitalize on a broader spectrum of data [42], [43]. By harnessing the attributes of the CFAL, the model could concentrate on learning discriminative features from slides belonging to a minority class.

REFERENCES

[1] F. Bray, M. Laversanne, E. Weiderpass, and I. Soerjomataram, “The ever-increasing importance of cancer as a leading cause of premature death worldwide,” *Cancer*, vol. 127, no. 16, pp. 3029–3030, 2021.

[2] P. Beidler, M. Nguyen, and J. Kang, “Extracting knowledge of NCI research directions from funding data using language processing,” *J. Clin. Oncol.*, vol. 39, no. 15, May 2021, doi: 10.1200/jco.2021.39.15_suppl.e13547.

[3] S. R. Lakhani, I. O. Ellis, S. Schnitt, P. H. Tan, and M. v. d. Vijver, *WHO Classification of Tumours of the Breast*. Lyon, France: International Agency for Research on Cancer, 2012. [Online]. Available: https://publications.iarc.fr/Book-And-Report-Series/Who-Classification-Of-Tumours/WHO-Classification-Of-Tumours-Of-The-Breast-2012

[4] A. N. Giaquinto, K. D. Miller, K. Y. Tossas, R. A. Winn, A. Jemal, and R. L. Siegel, "Cancer Statistics for African American/Black People 2022," *CA: Cancer J. Clinicians*, vol. 72, no. 3, pp. 202–229, 2022.

[5] R. A. Smith et al., "American Cancer Society guidelines for the early detection of cancer," *CA: A Cancer J. Clinicians*, vol. 52, no. 1, pp. 8–22, Feb. 2002.

[6] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, and S. Li, "Breast cancer multi-classification from histopathological images with structured deep learning model," *Sci. Rep.*, vol. 7, no. 1, Jun. 2017, Art. no. 4172.

[7] M. Sapkota, X. Shi, F. Xing, and L. Yang, "Deep convolutional hashing for low-dimensional binary embedding of histopathological images," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 805–816, Mar. 2019.

[8] M. Liu et al., "A deep learning method for breast cancer classification in the pathology images," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 10, pp. 5025–5032, Oct. 2022.

[9] R. Alkadi, F. Taher, A. El-baz, and N. Werghi, "A deep learning-based approach for the detection and localization of prostate cancer in T2 magnetic resonance images," *J. Digit. Imag.*, vol. 32, no. 5, pp. 793–807, Oct. 2019, doi: 10.1007/s10278-018-0160-1.

[10] T. Hassan, S. Javed, A. Mahmood, T. Qaiser, N. Werghi, and N. Rajpoot, "Nucleus classification in histology images using message passing network," *Med. Image Anal.*, vol. 79, 2022, Art. no. 102480.

[11] S. Javed, A. Mahmood, J. Dias, and N. Werghi, "Multi-level feature fusion for nucleus detection in histology images using correlation filters," *Comput. Biol. Med.*, vol. 143, Apr. 2022, Art. no. 105281. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482522000737

[12] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[13] J. N. Kather et al., "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study," *PLoS Med.*, vol. 16, no. 1, 2019, Art. no. e1002730.

[14] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 8792–8802.

[15] H. Wang et al., "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5265–5274.

[16] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4685–4694.

[17] G. Elsayed, D. Krishnan, H. Mobahi, K. Regan, and S. Bengio, "Large margin deep networks for classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 850–860.

[18] P. Khosla et al., "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 18661–18673.

[19] K. Stacke, G. Eilertsen, J. Unger, and C. Lundström, "Measuring domain shift for deep learning in histopathology," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 2, pp. 325–336, Feb. 2021.

[20] M. Hayat, S. Khan, S. W. Zamir, J. Shen, and L. Shao, "Gaussian affinity for max-margin class imbalanced learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6469–6479.

[21] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5375–5384.

[22] A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *J. Pathol. Informat.*, vol. 7, no. 1, p. 29, 2016.

[23] A. Cruz-Roa et al., "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks," *Proc SPIE*, vol. 9041, 2014, Art. no. 904103.

[24] J. N. Kather et al., "Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer," *Nature Med.*, vol. 25, no. 7, pp. 1054–1056, 2019.

[25] N. Coudray et al., "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature Med.*, vol. 24, no. 10, pp. 1559–1567, 2018.

[26] Y. Tolkach, T. Dohmgörgen, M. Toma, and G. Kristiansen, "High-accuracy prostate cancer pathology using deep learning," *Nature Mach. Intell.*, vol. 2, no. 7, pp. 411–418, 2020.

[27] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, 2021.

[28] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 499–515.

[29] F. Wang et al., "The devil of face recognition is in the noise," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 765–780.

[30] Z. Li et al., "Deep learning methods for lung cancer segmentation in whole-slide histopathology images–The ACDC, lunghp challenge 2019," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 2, pp. 429–440, Feb. 2021.

[31] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9268–9277.

[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[33] J. M. Noothout et al., "Knowledge distillation with ensembles of convolutional neural networks for medical image segmentation," *J. Med. Imag.*, vol. 9, no. 5, pp. 052407–052407, 2022.

[34] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4320–4328.

[35] S. Ghosh, A. Bandyopadhyay, S. Sahay, R. Ghosh, I. Kundu, and K. Santosh, "Colorectal histology tumor detection using ensemble deep neural network," *Eng. Appl. Artif. Intell.*, vol. 100, 2021, Art. no. 104202.

[36] T. Mahbub, A. Obeid, S. Javed, J. Dias, and N. Werghi, "Class-balanced affinity loss for highly imbalanced tissue classification in computational pathology," in *Proc. Int. Conf. Pattern Recognit.*, 2022, pp. 499–513.

[37] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, 2022.

[38] S. Liu et al., "Capturing time dynamics from speech using neural networks for surgical mask detection," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 8, pp. 4291–4302, Aug. 2022.

[39] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the Big Data paradigm with compact transformers," 2021, *arXiv:2104.05704*.

[40] M. A.-E. Zeid, K. El-Bahnasy, and S. Abo-Youssef, "Multiclass colorectal cancer histology images classification using vision transformers," in *Proc. 10th Int. Conf. Intell. Comput. Inf. Syst.*, 2021, pp. 224–230.

[41] A. I. Jajja et al., "Compact convolutional transformer (CCT)-based approach for whitefly attack detection in cotton crops," *Agriculture*, vol. 12, no. 10, 2022, Art. no. 1529.

[42] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomed. Eng.*, vol. 5, no. 6, pp. 555–570, 2021.

[43] C.-W. Wang, H. Muzakky, Y.-C. Lee, Y.-J. Lin, and T.-K. Chao, "Annotation-free deep learning-based prediction of thyroid molecular cancer biomarker BRAF (V600E) from cytological slides," *Int. J. Mol. Sci.*, vol. 24, no. 3, 2023, Art. no. 2521.