Neurocomputing 491 (2022) 206–216

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

CS-AF: A cost-sensitive multi-classifier active fusion framework for skin lesion classification

Di Zhuang *, Keyu Chen, J. Morris Chang

Department of Electrical Engineering, University of South Florida, Tampa, FL 33620, United States

ARTICLE INFO

Article history:
Received 9 July 2020
Revised 4 February 2022
Accepted 20 March 2022
Available online 23 March 2022

Keywords:
Deep Neural Networks
Multi-classifier fusion
Active fusion
Ensemble learning
Cost-sensitive classification
Skin lesion analysis

ABSTRACT

Convolutional neural networks (CNNs) have achieved the state-of-the-art performance in skin lesion analysis. Compared with single CNN classifier, combining the results of multiple classifiers via fusion approaches shows to be more effective and robust. Since the skin lesion datasets are usually limited and statistically biased, while designing an effective fusion approach, it is important to consider not only the performance of each classifier on the training/validation dataset, but also the relative discriminative power (e.g., confidence) of each classifier regarding an individual sample in the testing phase, which calls for an active fusion approach. Furthermore, in skin lesion analysis, the data of certain classes (e.g., the benign lesions) is usually abundant which makes them an over-represented majority, while the data of some other classes (e.g., the cancerous lesions) is deficient which makes them an underrepresented minority. It is more crucial to precisely identify the samples from an underrepresented (i.e., in terms of the amount of data) but more important minority class (e.g., cancerous skin lesions). In other words, misclassifying a more severe skin lesion to a benign or less severe skin lesion should have relative more cost (e.g., money, time and even lives). To address such challenges, we present CS-AF, a cost-sensitive multi-classifier active fusion framework for skin lesion classification. In the experimental evaluation, we prepared 96 base classifiers (of 12 CNN architectures) on the ISIC Challenge 2019 research dataset. Our experimental results show that our framework consistently outperforms both the static and the active fusion competitors in terms of the accuracy and total costs.

1. Introduction

Deep learning (DL) has achieved great success in many applications related to skin lesion analysis. For instance, Zhang et al. [1] has shown that convolutional neural networks (CNNs) have achieved the state-of-the-art performance in skin lesion classification. Also, as the development of various deep learning techniques, numerous different designs of classifiers, that might have different CNN architectures, use different sizes of the training data, use different subsets or classes distributions of the training data or use different feature sets, were proposed to tackle the skin lesion classification problem. For instance, as shown in the ISIC Challenges [2–4], several CNN architectures have been used in skin lesion analysis, including ResNet, Inception, DenseNet, PNASNet, etc. Because of such difference (i.e., CNN architectures, subset of the training data, feature sets, etc.), those classifiers tend to have distinct performance under different conditions (e.g, different subsets or classes distributions of different datasets). There is no one-size-fits-all solution to design a single classifier for skin lesion classification. It is necessary to investigate multi-classifier fusion techniques to perform skin lesion classification under different conditions.

Designing an effective multi-classifier fusion approach for skin lesion classification needs to address two challenges. First, since the datasets are usually limited and statistically biased [2–4], while conducting multi-classifier fusion, it is necessary to consider not only the performance of each classifier on the training/validation dataset, but also the relative discriminative power (e.g., confidence) of each classifier regarding an individual sample in the testing phase. This challenge requires the researchers to design an active fusion approach, that is capable of tuning the weight assigned to each classifier dynamically and adaptively, depending on the characteristics of given samples in the testing phase. Second, since in most of the real-world skin lesion datasets [2–4] the data of certain classes (e.g., the benign lesions) is abundant which makes them an over-represented majority, while the data of some other classes (e.g., the cancerous lesions) is deficient which makes them an underrepresented minority, it is more crucial to

* Corresponding author.
E-mail addresses: dizhuang@usf.edu (D. Zhuang), keyu@usf.edu (K. Chen), chang5@usf.edu (J.M. Chang).

https://doi.org/10.1016/j.neucom.2022.03.042
0925-2312/

precisely identify the samples from an underrepresented (i.e., in terms of the amount of data) but more important minority class (e.g., cancerous skin lesions). For instance, a deadly cancerous skin lesion (e.g., melanoma) that rarely appears during the examinations should be barely misclassified as benign or other less severe lesions (e.g., dermatofibroma). Specifically, misclassifying a more severe lesion to a benign or less severe lesion should have relative more cost (e.g., money, time and even lives). Hence, it is also important to enable such "cost-sensitive" feature in the design of an effective multi-classifier fusion approach for skin lesion classification.

In this work, we propose CS-AF, a cost-sensitive multi-classifier active fusion framework for skin lesion classification, where we define two types of weights: the objective weights and the subjective weights. The objective weights are designed according to the classifiers' reliability to recognize the particular skin lesions, which is determined by the prior knowledge obtained through the training phase. The subjective weights are designed according to the relative confidence of the classifiers while recognizing a specific previously "unseen" sample (i.e., individuality), which are calculated by the posterior knowledge obtained through the testing phase. While designing the objective weights, we also utilize a customizable cost matrix to enable the "cost-sensitive" feature in our fusion framework, where given a sample, different outputs (i.e., correct predictions or all kinds of errors) of a classifier should result in different costs. For instance, the cost of misclassifying melanoma as benign should be much higher than misclassifying benign as melanoma. In the experimental evaluation, we trained 96 base classifiers as the input of our fusion framework, utilizing twelve CNN architectures on the ISIC Challenge 2019 research dataset for skin image analysis [2–4]. We compared our approach with two static fusion baseline approaches (i.e., max voting and average fusion) and two state-of-the-art active fusion approaches (i.e., MCE-DW [5] and DES-MI [6]). Our experimental results show that our CS-AF framework consistently outperforms the static fusion baseline approaches and the state-of-the-art competitors in terms of accuracy, and always achieves lower total cost.

To summarize, our work has the following contributions:

- We present a novel and effective multi-classifier active fusion framework, where the proposed multi-classifier weight assignment not only leverages the "reliability" (i.e., the objective weights) extracted from the prior knowledge of the training/-validation dataset, but also take advantages of the "individuality" (i.e., the subjective weights) computed from the posterior knowledge of the testing dataset.
- We propose an approach to enable the "cost-sensitive" feature of our multi-classifier active fusion framework, where the proposed multi-classifier weight assignment can easily actively adapt to different customized cost matrices.
- To the best of our knowledge, our work is the first one that attempts to apply active fusion for skin lesion analysis, and demonstrates its advantages over the conventional static fusion and existing active fusion approaches. Specifically, a comprehensive experimental evaluation using twelve popular and effective CNN architectures has been conducted on the most popular skin lesion analysis benchmark dataset, ISIC Challenge 2019 research datasets [2–4]. For the sake of reproducibility and convenience of future studies about fusion approaches in skin lesion analysis, we have released our prototype implementation of CS-AF, information regarding the experiment datasets and the code of our comparison experiments. [1]

[1] https://github.com/keyu07/CS-AF

The rest of this paper is organized as follows: Section 2 presents the related literature review. Section 3 presents the notations of cost-sensitive active fusion, and describes our proposed framework. Section 4 presents the experimental evaluation. Section 5 makes the conclusion.

2. Related work

2.1. Multi-classifier fusion

Fusion approaches have been widely applied in numerous applications, such as skin lesion analysis [7–9], human activity recognition [10,11], active authentication [12], facial recognition [13–15], botnet detection [16–18], domain generalization [19] and community detection [20,21]. In terms of whether the weights are dynamically/adaptively assigned to each classifier, the multi-classifier fusion approaches are divided into two categories: (i) static fusion, where the weight assigned to each participating classifier will be fixed after its initial assignment, and (ii) active fusion, where the weights are adaptively tuned depending on the characteristics of given samples in the testing phase. Many conventional approaches, such as the bagging [22], boosting [23,24] and stacking [25], are static fusion approaches.

To date, a few methods attempting to conduct active fusion were also proposed [5,26,6,27,28]. For instance, Chen et al. [27] propose to use an attention model to fuse the weights of different CNNs trained on different scaled input images. Fang et al. [28] propose to use an U-shape pyramid neural network structure to facilitate the need of training multi-scale CNNs on multiple partially labeled datasets. Both approaches [27,28] present some interesting ideas in adaptively fusing multiple CNNs trained on different datasets. However, both solutions mainly focus on multi-scale image datasets, rather than imbalanced or cost-sensitive datasets. META-DES [26] defines five distinct sets of meta-features to measure the level of competence of a classifier for the classification of input samples, and proposes to train a meta-classifier to determine the rank or weight of a base classifier while facing input samples. However, those meta-classifiers were trained on the same dataset (i.e., the training dataset) as the base classifiers, which would make the meta-classifiers be less effective or generalized to the "unseen" dataset (i.e., the testing dataset). Also, META-DES has only been evaluated on several small sample size datasets, which didn't demonstrate its effectiveness, scalability and generalizability towards more complex datasets or problems, e.g., skin lesion analysis. DES-MI [6] propose an active fusion approach where the weights are determined via emphasizing more on the classifiers that are more capable of classifying examples in the region of underrepresented area among the whole sample distribution. However, DES-MI only focuses on identifying the most competent classifiers on the training dataset, which cannot provide enough adaptivity towards the "unseen" testing dataset. MCE-DW [5] proposes to use the decision credibility that is evaluated by fuzzy set theory and cloud model, to determine the real-time weight of a base classifier regarding the current sample in the testing phase. However, both DES-MI and MCE-DW are designed to work on imbalanced dataset, rather than providing the adaptivity and flexibility to a cost-sensitive dataset with customized cost matrices, e.g., skin lesion analysis.

In our work, we propose a novel multi-classifier active fusion framework, that leverages the "reliability" (Section 3.3) and the "individuality " (Section 3.4) of the base classifiers to assign the weights dynamically and adaptively. Also, we propose an approach to enable the "cost-sensitive" feature of our framework, where the proposed multi-classifier weight assignment can easily actively adapt to different customized cost matrices.

2.2. Fusion of CNNs for skin lesion analysis

Convolutional neural networks (CNNs) have achieved the state-of-the-art performance [2–4] in skin lesion analysis since 2016 (i.e., ISIC 2016 Challenge [2]), where nearly all the teams employed CNNs in either feature extraction or classification procedure. Recently, several approaches attempting to apply fusion on CNNs to tackle the skin lesion classification problems are proposed [29,30,7]. For instance, Marchetti et al. [29] presents a fusion of CNNs framework for the classification of melanomas versus nevi or lentigines, where five fusion approaches were implemented to fuse 25 different CNN classifiers trained on the same dataset of the same problem to make a single decision. Bi et al. [30] proposes another CNNs fusion framework to tackle the classification of melanomas versus seborrheic keratosis versus nevi, where three ResNet classifiers were trained for three different classification problems via fine-tuning pretrained ImageNet CNNs: the original three-class problem and two binary classifiers (i.e., melanoma versus both other lesion classes and seborrheic carcinoma versus both other lesion classes). Perez et al. [7] conducts a comparison study between two fusion strategies for melanoma classification: selecting the classifiers at random (i.e., among 125 models over 9 CNNs architectures), and selecting the classifiers depending on their performance on a validation dataset.

To summarize, most of the existing approaches use static fusion approaches for skin lesion analysis. However, as discussed in Section 1, since the skin lesion datasets are usually limited and statistically biased [2–4], it is necessary to enable active fusion in such problem. To the best of our knowledge, our work is the first to design, apply and evaluate active fusion approaches in the skin lesion classification problems.

2.3. Cost-sensitive machine learning

A variety of cost-sensitive machine learning approaches have been proposed to tackle the class imbalance issue in pattern classification and learning problems. Mollineda et al. [31], a comprehensive study on the class imbalance issue, divides most of the cost-sensitive machine learning approaches into two categories: the data-level and the algorithmic-level. The data-level approaches usually manipulate the data distribution via over-sampling the samples of the minority classes or under-sampling the samples of the majority classes. For instance, SMOTE [32] is an over-sampling technique proposed to address the over-fitting problem via synthesizing more of the samples of the minority classes. Several variants of the SMOTE approach [33–36] are also proposed to solve this issue.

The algorithmic-level approaches directly re-design the machine learning algorithms to minimize a customizable loss function, that enables the "cost-sensitive" feature, of the classifier on certain dataset (e.g., improving the sensitivity of the classifier towards minority classes). For instance, Importance-weighted risk minimization has been proposed in many machine learning algorithms and implementations, such as LibSVM [37], weighted cross entropy loss functions [38,39]. However, as pointed out by [40], the weighted cross entropy is only more effective in the early stage (e.g., a few epochs) of training CNNs, and its impact diminishes quickly as the number of epochs successively increasing. Therefore, it calls for alternative solutions, compared with directly applying weighted cross entropy, for the fusion of CNNs for imbalanced or cost-sensitive datasets. Zhang et al. [41] proposes an extreme learning machine (ELM) based evolutionary cost-sensitive classification approach, where the cost matrix would be automatically identified given a specific task (i.e., which error cost more). Iranmehr et al. [42] extends the standard loss function of support vector machine (SVM) to consider both the class imbalance (i.e., the cost) and the classification loss. Khan et al. [43] proposes a cost-sensitive deep neural network framework that could automatically learn the "cost-sensitive" feature representations for both the majority and minority classes, where during the training phase, the proposed framework would perform a joint-optimization on the class-dependent costs and the deep neural network parameters. In this work, we enable the "cost-sensitive" feature in the process of multi-classifier fusion, and employ it in the skin lesion classification problem.

3. Methodology

3.1. Multi-classifier fusion

In multi-classifier fusion, we define a classification space, as shown in Fig. 1, where there are m classes and k classifiers. Let $\mathscr{M} = \{M_1, M_2, \ldots, M_k\}$ denote the set of base classifiers and $\mathscr{C} = \{C_1, C_2, \ldots, C_m\}$ denote the set of classes. Let p_{kj}^m denote the posterior probability of given sample j identified by classifier M_k as belonging to class C_m, where $P_{kj} = \{p_{kj}^1, p_{kj}^2, \ldots, p_{kj}^m\}$ and $\sum_{l=1}^{m} p_{kj}^l = 1$. Hence, all the posterior probabilities form a $k \times m$ decision matrix as follows:

$$P_j = \begin{bmatrix} p_{1j}^1 & p_{1j}^2 & \cdots & p_{1j}^m \\ p_{2j}^1 & p_{2j}^2 & \cdots & p_{2j}^m \\ \vdots & \vdots & \ddots & \vdots \\ p_{kj}^1 & p_{kj}^2 & \cdots & p_{kj}^m \end{bmatrix} \tag{1}$$

Since the importance of different classifiers might be different, we assign a wight w_i to the decision vector (i.e., posterior probabilities vector) of each classifier C_i, where $i \in \{1, 2, \ldots, k\}$. Let $P_m(j)$ denote the sum of the posterior probabilities, that sample j belonging to class m, of all the classifiers. Then, we have

$$P_m(j) = \sum_{i=1}^{k} w_i \cdot p_{ij}^m \tag{2}$$

The final decision (i.e., class) $D(j)$ of sample j is determined by the maximum posterior probabilities sum:

$$D(j) = \max_i P_i(j), \ i \in \{1, 2, \ldots, m\} \tag{3}$$

Conventional multi-classifier fusion approaches either use the same weight for all the classifiers (i.e., average fusion) or use static weights that will not be changed after its initial assignment during the training phase. As illustrated in Fig. 1, our weights (i.e., $w_k = \frac{O_k + S_k}{2}$) contains two components: (i) the objective weight O_k that is static and determined by the prior knowledge obtained through the training phase (Section 3.3ii) the subjective weight S_k that is dynamic and calculated by the posterior knowledge obtained through the testing phase (Section 3.4). For different applications, we can assign different weights toward O_k and S_k. To be simplified and for demonstration purposes, in this work, we assign the same weight, i.e., 0.5, on both O_k and S_k, while combining them together.

3.2. Cost-sensitive problem formulation

As discussed in Section 1, given a sample, different outputs (i.e., the correct prediction or all kinds of errors) of a classifier should result in different costs. For instance, misclassifying a more severe lesion to a benign or less severe lesion should have relative higher cost. Let c_{pq} denote the cost of classifying an instance belonging to class p into class q. Then, we obtain a cost matrix as follows:

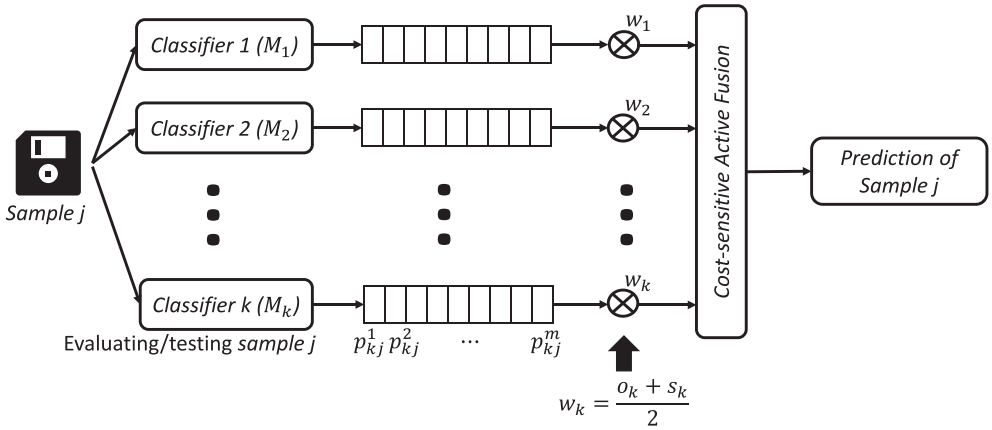


Fig. 1. The Overview of CS-AF Framework.

$$CM = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mm} \end{bmatrix} \tag{4}$$

Let $W = \{w_1, w_2, \ldots, w_k\}$ be a fusion weight vector, and $\mathscr{W}$ be the fusion weight vector space, where $W \in \mathscr{W}$. The goal of cost-sensitive multi-classifier fusion is to find the $W^* \in \mathscr{W}$, that can minimize the average cost of the fusion approach's outcomes over all the testing samples.

3.2.1. Examples of the design of cost matrix

In this section, we would like to show examples of the design of cost matrices. To demonstrate the "cost-sensitive" feature in our CS-AF framework, here we design two different cost matrices for the application of skin lesion analysis. There are eight classes, i.e., melanoma (MEL), squamous cell carcinoma (SCC), basal cell carcinoma (BCC), melanocytic nevus (NV), actinic keratosis (AK), dermatofibroma (DF), vascular lesion (VASC), benign keratosis (BKL) in our skin lesion classification problem, where MEL, SCC and BCC are cancerous, and the rest are benign. We would like to demonstrate our work by designing two cost matrices: Cost Matrix A, which emphasizes on the identification of cancerous skin lesions (i.e., the cost of misclassifying a cancerous skin lesion is much more than a benign one); and Cost Matrix B (the opposite of Cost Matrix A), which emphasizes on the identification of benign skin lesions.

We propose to follow the principles below to design our experimental cost matrices:

- All the costs should be positive, since it will be item-wise multiplied with the confusion matrix. As such, it will not result in negative values in the cost-sensitive confusion matrix.
- The cost of the correct predictions should depend on the relative severeness of the corresponding disease. For instance, it should be more valuable (i.e., less cost) to classify a more severe disease (i.e., melanoma) correctly. To figure out the relative severeness relationships among all eight skin lesion classes and design our cost matrix (i.e., Cost Matrix A) in a better way, we referred to the American Academy of Dermatology Association's guidance [44]. To be simplified and enable the evaluation of our work, based on the reference, we heuristically ordered the severeness of the 8 skin lesion classes (from the most severe one to the least severe one) as follows: melanoma (MEL), squamous cell carcinoma (SCC), basal cell carcinoma (BCC), melanocytic nevus (NV), actinic keratosis (AK), dermatofibroma (DF), vascular lesion (VASC), benign keratosis (BKL). It is worth noting that the absolute cost (i.e., quantitative evaluation) for each disease is non-trivial to decide, but the relative severeness (i.e., qualitative evaluation) is able to determine.
- The relative costs of different incorrect predictions should be based on their relative severeness. For instance, misclassifying melanoma (i.e., a deadly cancerous skin lesion) as benign keratosis should result in much more cost than the opposite scenario.
- The maximum cost of correct predictions should be no more than the minimum cost of incorrect predictions.

Fig. 4 illustrates the Cost Matrix A and Cost Matrix B that we utilized to evaluate our framework in the experimental evaluation. Let us take the design of Cost Matrix A as an example. Firstly, we assign the cost of correct prediction of each skin lesion class, i.e., $c_{ii}, i = 1, 2, \ldots, m$ (as defined in Section 3.2), according to the relative disease severeness, where predicting a more severe skin lesion class correctly should result in less cost. For instance, we set the cost of correct prediction of MEL (i.e., the most severe one) as 1, and the cost of correct prediction of BKL (i.e., the least severe one) as 8. Secondly, to calculate the relative cost of each incorrect prediction, we follow the equation below:

$$c_{ij} = \left(\frac{c_{jj}}{c_{ii}}\right)^2, \ i \neq j \tag{5}$$

where as defined in Section 3.2, c_{ij} denote the cost of classifying an instance belonging to class i into class j. For instance, if the cost of correct prediction of MEL is 1 and the cost of correct prediction of BKL is 8, the cost of misclassifying an instance belonging to MEL into BKL would be $(\frac{8}{1})^2 = 64$. Last but not least, to ensure the cost of correct predictions are always no more than the cost of incorrect predictions, without loss of generality, we normalized the costs of misclassifications to integers between 16 and 200, using min–max scaling. Fig. 4a shows the final result of our designed Cost Matrix A.

To evaluate our framework under different cost matrices, we also designed a Cost Matrix B (as shown in Fig. 4b), that emphasizes on benign lesions (i.e., the cost of misclassifying a benign lesion is much more than a cancerous lesion). Cost Matrix B follows the same design steps as Cost Matrix A, other than considering an exactly reverse order of the severeness. For instance, in the design

of Cost Matrix B, melanoma became the "least severe" one while benign keratosis became the "most severe" one.

3.3. Computing the objective weights

The objective weights are designed according to the classifiers' reliability to recognize the particular skin lesions, which is determined by the prior knowledge obtained through the training phase. In the training phase, we separate all the labelled data into two parts: training dataset and validation dataset. The training dataset will be used to train/build the base classifiers, while the validation dataset will be used to evaluate the performance of each base classifier. The reliability of each base classifier depends on its performance on the validation dataset. Therefore, in order to get an effective and unbiased reliability of each base classifier, the training dataset and validation dataset cannot have overlapped data. Specifically, as shown in Fig. 2, computing the objective weights in our framework contains three steps:

- Classifier build. We prepare a set of base classifiers, where all the classifiers might have different CNN architectures, use different size of the training data, or use different subset or classes distributions of the training data. In this step, we trained 96 base classifiers, more details are introduced in Section 4.2.
- Reliability validation. Let r_i denote the reliability of a base classifier M_i, that is designed to describe the average recognition performance of the classifier on the validation data. Higher accuracy and less error on the validation data usually means higher reliability. Hence, we use the confusion matrix result of each base classifier on the same validation dataset as its reliability, where a confusion matrix [45] is a table that is often used to describe the performance of a classifier on a set of validation data for which the true values are known. It allows easy identification of confusion between classes, e.g., one class is commonly mislabeled as the other. Many performance measures could be computed from the confusion matrix (e.g., F-scores). As such, we use r_i^{pq} to denote the probability of a base classifier M_i classifying an instance belonging to class p into class q.
- Cost-sensitive adjustment. As described in Section 3.2, we would like to enable the "cost-sensitive" feature in the design of our objective weights. As shown in Fig. 3, for each classifier M_i, we use an element-wise multiplication between its reliability r_i (confusion matrix) and the customized cost matrix (Section 3.2) to formulate a cost-sensitive confusion matrix, where all the results/errors in the confusion matrix have been adjusted based on the cost matrix. Then, we use the micro-average F1-score [46] of the cost-sensitive confusion matrix to define the objective weight of each base classifier, and all the object weights are automatically normalized to $(0, 1]$.

3.4. Computing the subjective weights

The subjective weights are designed according to the relative confidence of the classifiers while recognizing a specific previously "unseen" image (i.e., individuality), which are calculated by the posterior knowledge obtained through the testing phase. The individuality of each base classifier is dynamically computed from its discriminant confidence towards each previously "unseen" testing data, to capture the posterior knowledge that a base classifier cannot obtain from the training and validation datasets. Specifically, as shown in Fig. 5, computing the subjective weights in our framework contains three steps:

- Sample evaluating/testing. Each testing data is evaluated/tested through all the classifiers to obtain the corresponding decision vectors (i.e., the soft labels).
- Individuality calculation. We consider the individuality of a classifier as its relative class discriminative power regarding a given testing data. A classifier can easily identify the class of a given testing data in the testing phase, if its posterior probabilities of the corresponding decision vector is highly concentrated in one class, and the misclassification rate would also be low. On the contrary, if the distribution of the posterior probabilities is close to uniform, the classifier shows its difficulty in discriminating the class of the given testing data. Also, different classifiers would present different distribution of the posterior probabilities in the decision vectors while testing the same testing data. Hence, we define the individuality i_k of a classifier M_k using the posterior probabilities distribution as follows:

$$i_k = \frac{1}{m-1}\sum_{l=1}^{m}(p_{kj}^{*} - p_{kj}^{l}) = \frac{m \cdot p_{kj}^{*} - 1}{m-1} \tag{6}$$

where p_{kj}^{*} is the largest posterior probability value in P_{kj}. Based on Eq. (6), the individuality of each base classifier on a given testing data depends on its output probability of the most probable class. For each testing data, the base classifier that has the highest output probability of the most probable class achieves the highest individuality.

- Normalization. Since the subjective weights are relative values among all the base classifiers, we normalize each individuality i_k to the subjective weight $S_k \in [0, 1]$ as follows:

$$S_k = \frac{i_k - i^{min}}{i^{max} - i^{min}} \tag{7}$$

where $i^{min} = \min_{j \in 1,2,\ldots,k} i_j$ (i.e., the minimum individuality among all base classifiers), and $i^{max} = \max_{j \in 1,2,\ldots,k} i_j$ (i.e., the maximum individuality among all base classifiers).

4. Experimental evaluation

We conducted our experiments on the ISIC Challenge 2019 dataset [47,48,3] and utilized 12 CNN architectures to evaluate the performance of our proposed CS-AF framework. Two examples of cost matrices, that emphasize on different skin lesion classes (i.e., cancerous lesion classes vs. benign lesion classes), have been designed to evaluate the effectiveness of the "cost-sensitive" feature of our proposed CS-AF framework. Furthermore, extensive comparisons have been made among two static fusion approaches (i.e., Max Voting Fusion and Average Fusion), two state-of-the-art active fusion approaches (i.e., DES-MI [6] and MCE-DW [5]), AF (i.e., active fusion without the "cost-sensitive" feature) and our CS-AF framework. The presented results show that our approach consistently outperforms both the static and the active fusion approaches in terms of the overall accuracy and the total cost, is more adaptive to the customized cost matrices than the other two active fusion competitors, and consistently better than AF in terms of the total cost under different conditions.

4.1. Experiment dataset

In our experiments, we utilized the well known ISIC Challenge 2019 dataset [47,48,3]. Since the ground truth of the original testing data was not available, we only employed the original training data without meta-data in our experimental evaluation. This dataset (i.e., the original training data of the ISIC Challenge 2019) contains 25,331 images in total, coming from 3 source datasets: BCN_20000 [47], HAM10000 [48] and MSK [3]. It depicts 8 skin lesion diseases (i.e., 8 classes): melanoma (MEL, 4,522 images), melanocytic nevus (NV, 12,875 images), basal cell carcinoma

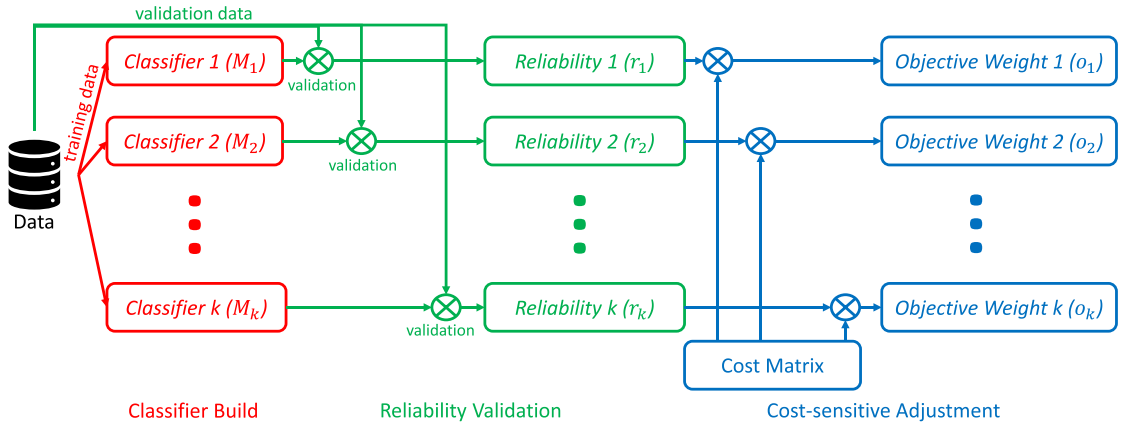


Fig. 2. The calculation of objective weights.

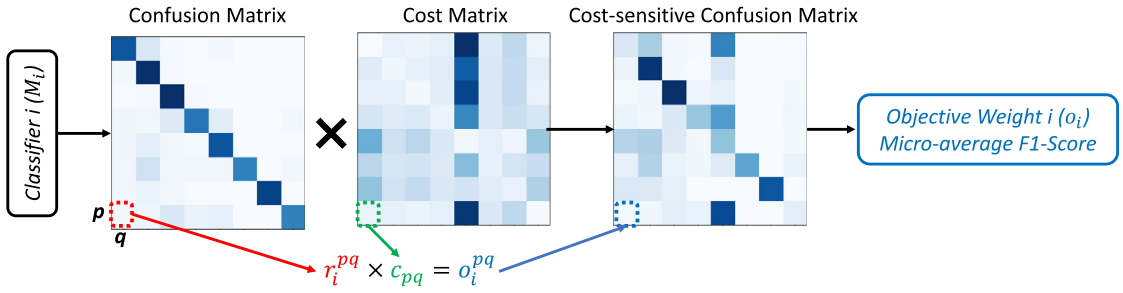


Fig. 3. The calculation of cost-sensitive confusion matrix.

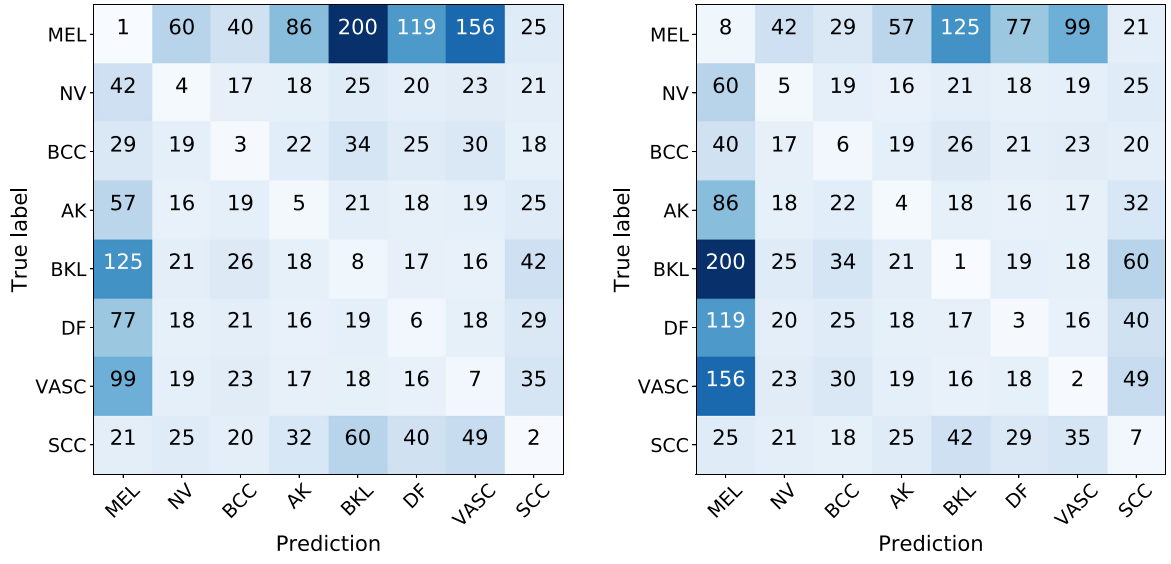


Fig. 4. Two examples of cost matrices: (a) Cost Matrix A (emphasizing on cancerous skin lesions); (b) Cost Matrix B (emphasizing on benign skin lesions).

(BCC, 3,323 images), actinic keratosis (AK, 867 images), benign keratosis (BKL, 2624 images), dermatofibroma (DF, 239 images), vascular lesion (VASC, 253 images) and squamous cell carcinoma (SCC, 628 images). We split the entire 25,331 images into training (80%), validation (5%) and testing (15%) datasets.

To evaluate the performance of our proposed CS-AF framework using the base classifiers that are trained from the datasets with different classes distributions, we designed 4 training datasets that have different classes distributions. For instance, one training dataset could have balanced classes distribution, and the other training datasets could have unbalanced classes distributions in different ways. The details (i.e., classes distributions) of each training dataset are shown in Table 1 and described as below:

- Dist-1: This training dataset follows the classes distribution of the original training dataset of the ISIC Challenge 2019 dataset.
- Dist-2: This training dataset contains evenly distributed number of samples for all classes.

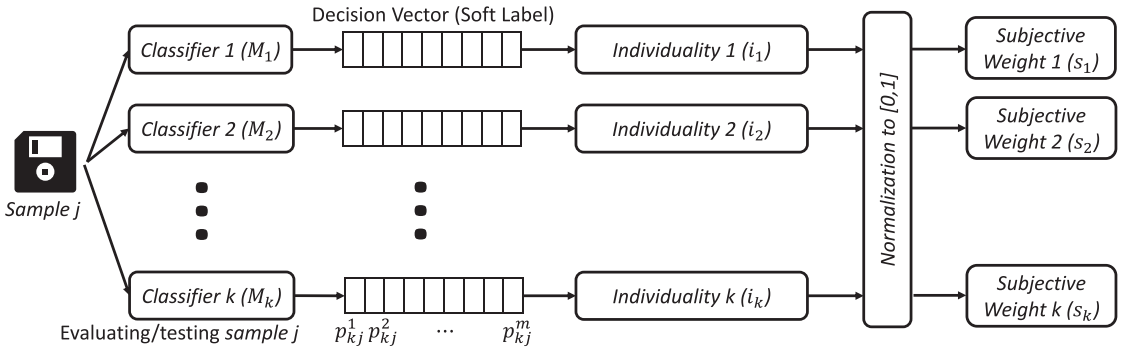


Fig. 5. The calculation of subjective weights.

Table 1
The number (ratio) of samples of each skin lesion classes of different split training datasets.

Skin Lesion	Dist-1	Dist-2	Dist-3	Dist-4
MEL	3,662 (18.1%)	2,509 (12.4%)	5,052 (22.5%)	604 (2.7%)
SCC	502 (2.5%)	2,510 (12.4%)	4,331 (19.3%)	1,200 (5.4%)
BCC	2,670 (13.2%)	2,494 (12.4%)	3,781 (16.9%)	1,812 (8.1%)
NV	10,235 (50.5%)	2,512 (12.4%)	3,032 (13.5%)	2,529 (11.3%)
AK	705 (3.5%)	2,564 (12.5%)	2,463 (11.0%)	3,150 (14.1%)
DF	188 (1%)	2,444 (12.3%)	1,871 (8.3%)	3,702 (16.9%)
VASC	194 (1%)	2,522 (12.5%)	1,262 (5.6%)	4,334 (19.3%)
BKL	2,099 (10.4%)	2,612 (13.0%)	626 (2.8%)	5,056 (22.6%)

- Dist-3: This training dataset contains more samples for cancerous lesion classes, and less samples for benign lesion classes.
- Dist-4: This training dataset contains less samples for cancerous lesion classes, and more samples for benign lesion classes (i.e., the opposite order of classes distributions as in Dist-3).

To generate different training datasets satisfying different classes distributions described above, we utilized data augmentation techniques to generate more images for the skin lesion classes lacking of images, and randomly sampled smaller portions from the classes with superfluous images. The main data augmentation techniques utilized are rotation (for 45, 90, 135, 180, 225, 270 and 315 degrees, respectively), horizontal flip or the combination of both. We also utilized the same strategy to generate the validation and testing datasets, to ensure the numbers of samples of all classes are equal, where there are approximate 200 samples of each class in the validation dataset, and approximate 500 samples of each class in the testing dataset.

In addition, to evaluate the performance of our proposed CS-AF framework using the base classifiers that are trained from different subsets of the training dataset, for each of those four training datasets that have different classes distributions, we randomly select 70% of its data to produce another sub-dataset, namely, Sub-70. Therefore, in our experimental evaluation, there are 8 different split training datasets in total (i.e., Dist-1, Dist-1 Sub-70, Dist-2, Dist-2 Sub-70, Dist-3, Dist-3 Sub-70, Dist-4 and Dist-4 Sub-70).

4.2. Base classifiers preparation

We chose 12 different CNN architectures to evaluate the fusion approaches performance. These 12 CNN architectures were popular and have been shown to have good transfer learning performance on the skin lesion analysis [7]. All the 12 CNN architectures were trained on the 8 different split training datasets as we mentioned in the previous section. Therefore, we obtained a pool of $12 \times 8 = 96$ base classifiers, the corresponding accuracy of each base classifier is shown in Table 2. Notably, different CNN architecture requires different size of input images as below.

- 331×331: PNASNet-5-Large, NASNet-A-Large.
- 320×320: ResNeXt101-32×16d.
- 299×299: Xception, Inception-V4, Inception-V3, InceptionResNet-V2.
- 224×224: SENet154, Dual Path Net-107, SE-ResneXt101-32×4d, EfficientNet-b7, ResNet152.

It is also worth noting that in our experiments, we treat CNNs with different size of input images equally during the weight assignment, since our objective weight assignment depends on the overall performance of each base classifier evaluated on the given validation dataset, which already considered the specification differences of different CNNs.

All the base classifiers were fine-tuned by stochastic gradient decent (SGD) with learning rate 10^{-3} and momentum 0.9. The learning rate degraded in 20 epochs by 0.1. We stopped the training process either after 40 epochs or while the validation accuracy was failed to improve for 7 consecutive epochs. Our experiments were implemented using Pytorch, running on a server with 4 GTX 1080Ti 11 GB GPUs. To keep the same batch size 32 in each evaluation, and due to the memory constraint of single GPU, certain CNN architectures were trained with more GPUs:

- 2 GPUs: SENet154, EfficientNet-b7, Dual Path Net-107.
- 4 GPUs: PNASNet-5-Large, ResNeXt101-32×16d, NASNet-A-Large.

4.3. Experimental procedure

As described in Section 4.2, we have prepared 96 base classifiers. To evaluate the effectiveness of our active fusion approach extensively, each time we perform the fusion approaches on a randomly selected subset (i.e., N classifiers) of those 96 base classifiers, where $N = 8, 16, 24, 32, 40, 48, 56, 64, 72, 80, 88, 96$. For each

Table 2
The performance (accuracy in %) of the base classifiers of twelve CNN architectures trained on 8 different split training datasets.

CNN Architectures	Dist-1	Dist-1 Sub-70	Dist-2	Dist-2 Sub-70	Dist-3	Dist-3 Sub-70	Dist-4	Dist-4 Sub-70
PNASNet-5-Large [49]	78.48	76.53	81.14	**80.73**	77.01	75.21	78.76	75.34
NASNet-A-Large [50]	78.35	76.71	79.80	78.31	76.00	75.36	76.12	74.32
ResNeXt101-32×16d [51]	79.47	76.96	**83.09**	80.08	**80.18**	**77.14**	**79.47**	**77.47**
SENet154 [52]	**80.31**	**77.72**	81.19	76.33	79.04	74.04	78.76	76.43
Dual Path Net-107 [53]	76.61	74.51	79.10	77.92	76.07	70.88	77.24	74.80
Xception [54]	74.63	74.07	78.53	75.19	76.46	72.93	75.82	74.30
Inception-V4 [55]	76.76	74.22	80.11	77.45	77.09	75.37	75.89	74.10
InceptionResNet-V2 [55]	77.58	76.64	70.81	77.39	77.77	76.12	76.48	74.01
SE-ResneXt101-32×4d [52]	77.45	77.21	79.87	78.41	75.38	74.68	75.60	74.33
ResNet152 [56]	75.69	73.23	79.27	75.01	76.00	74.96	75.77	74.45
Inception-V3 [57]	75.16	73.82	79.52	78.83	73.69	72.41	75.62	72.07
EfficientNet-b7 [58]	67.31	63.07	74.10	71.28	71.81	68.75	71.48	67.37

N, we repeat the random selection experiments for 100 times, and use the averaged performance as the final results.

4.4. Evaluate the effectiveness of CS-AF

To demonstrate the effectiveness of our approach, a comparison between two CS-AF implementations (using two different cost matrices to compute the objective weights) and five other fusion approaches has been conducted. The competitors are:

- Max Voting Fusion is a static approach, where the predictions are combined from multiple base classifiers and only the predicted class with the highest votes will be included in the final prediction.
- Average Fusion is another static approach, where it averages the decision vectors of multiple base classifiers and uses the averaged decision vector to make the final prediction.
- DES-MI [6] filters the base classifiers by assigning weight to each of them, where the weight is based on the performance of k-nearest neighbors in validation set regarding the current test sample.
- MCE-DW [5] determines the classifier reliability by fuzzy set theory, and combines decision credibility of each test sample to make the final decision.
- AF is a baseline active fusion approach by removing the cost-sensitive adjustment step from CS-AF while calculating the objective weights.
- CS-AF (Cost Matrix A) is our approach while computing its objective weights using Cost Matrix A (Section 3.2.1).
- CS-AF (Cost Matrix B) is our approach while computing its objective weights using Cost Matrix B (Section 3.2.1).

Given a competitor fusion approach, we evaluate its effectiveness in terms of (i) its averaged accuracy on our testing dataset (as shown in Fig. 6a), ii) its total cost on our testing dataset specified by Cost Matrix A (as shown in Fig. 6b), and Cost Matrix B (as shown in Fig. 6c). The total costs are calculated by the sum of the item-wise product between the confusion matrix resulted from our testing dataset and the specified cost matrix. Better fusion approach usually leads to higher accuracy on the testing dataset and lower total cost specified by certain cost matrix. From the results illustrated in Fig. 6, we obtain the observations below:

- Compared with the best performed base classifier, ResNeXt101-32×16d, as shown in Table 2, our two implementations of CS-AF and AF consistently achieve over 2%-5% higher accuracy on the same testing dataset.
- For all the fusion approaches, as more base classifiers involved, the accuracy tends to increase and the total cost tends to decrease.
- As illustrated in Fig. 6a, in terms of the accuracy, our two implementations of CS-AF and AF consistently outperform the static fusion approaches (i.e., Max Voting and Average). Compared with the active fusion competitors (i.e., MCE-DW and DES-MI), our CS-AF (Cost Matrix B) consistently obtains the highest accuracy.
- As illustrated in Fig. 6b and Fig. 6c, in terms of the total cost, CS-AF consistently outperforms the other fusion competitors (i.e., Max Voting, Average MCE-DW, DES-MI and AF). For instance, while calculating the total cost using Cost Matrix A, CS-AF (Cost Matrix A) always achieves the lowest total cost, and while calculating the total cost using Cost Matrix B, CS-AF (Cost Matrix B) always obtains the lowest total cost. Thus, it demonstrates that our proposed cost-sensitive active fusion approach could adapt to different customized cost matrices and is optimized to achieve the lowest total cost.
- DES-MI is more sensitive to the number of base classifiers. For instance, it always has the worst performance with few classifiers involved, and finally has close performance with our CS-AF as shown in Fig. 6. This is because in the implementation

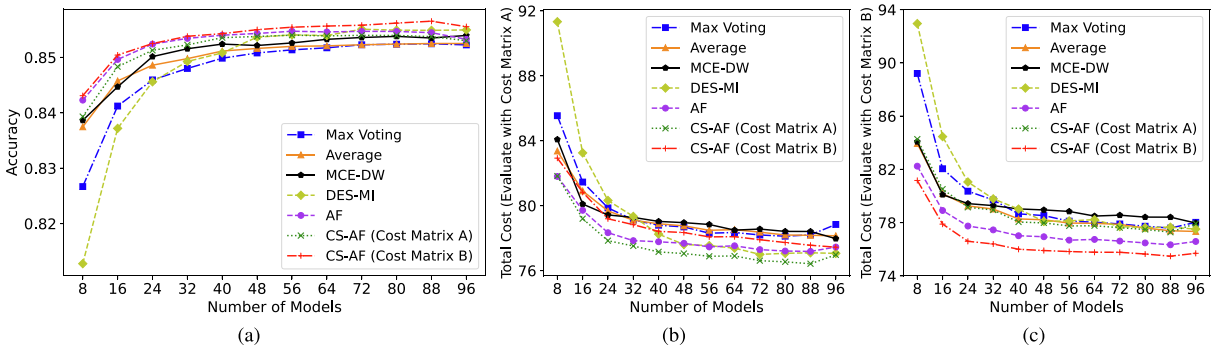


Fig. 6. Evaluate the Effectiveness of CS-AF. (a) The overall accuracy on our testing dataset; (b) The total cost calculated by Cost Matrix A; (c) The total cost calculated by Cost Matrix B.

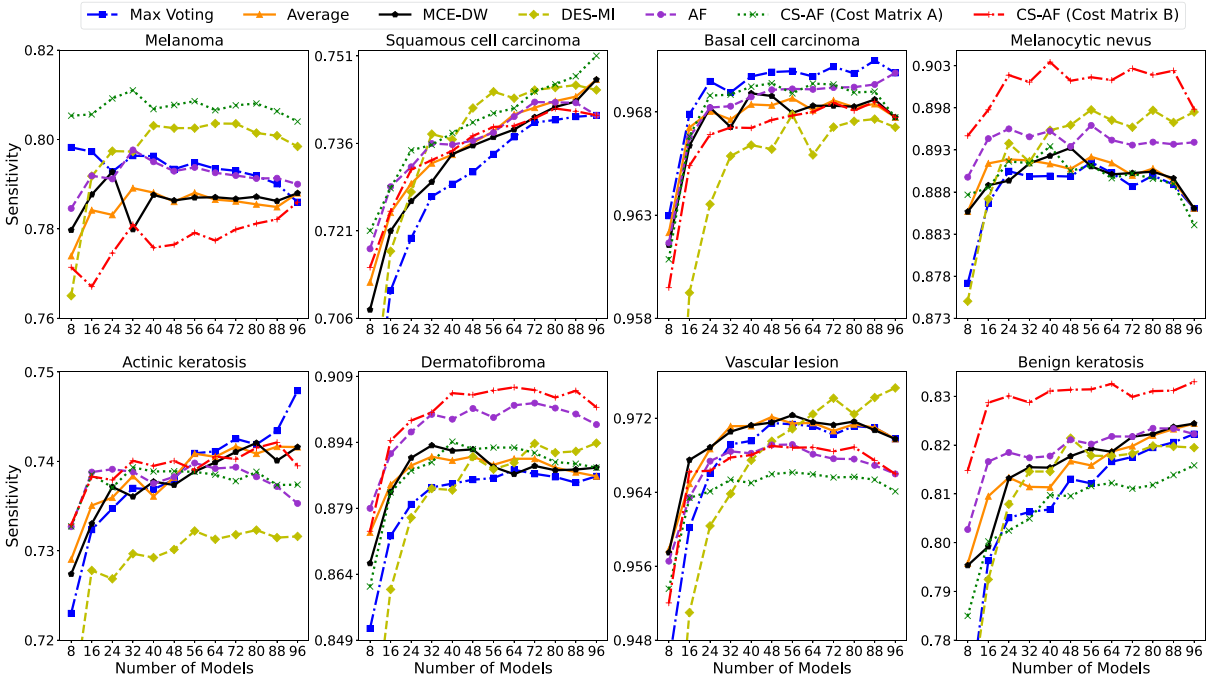


Fig. 7. The sensitivity results of each single class of CS-AF (Cost Matrix A) and CS-AF (Cost Matrix B).

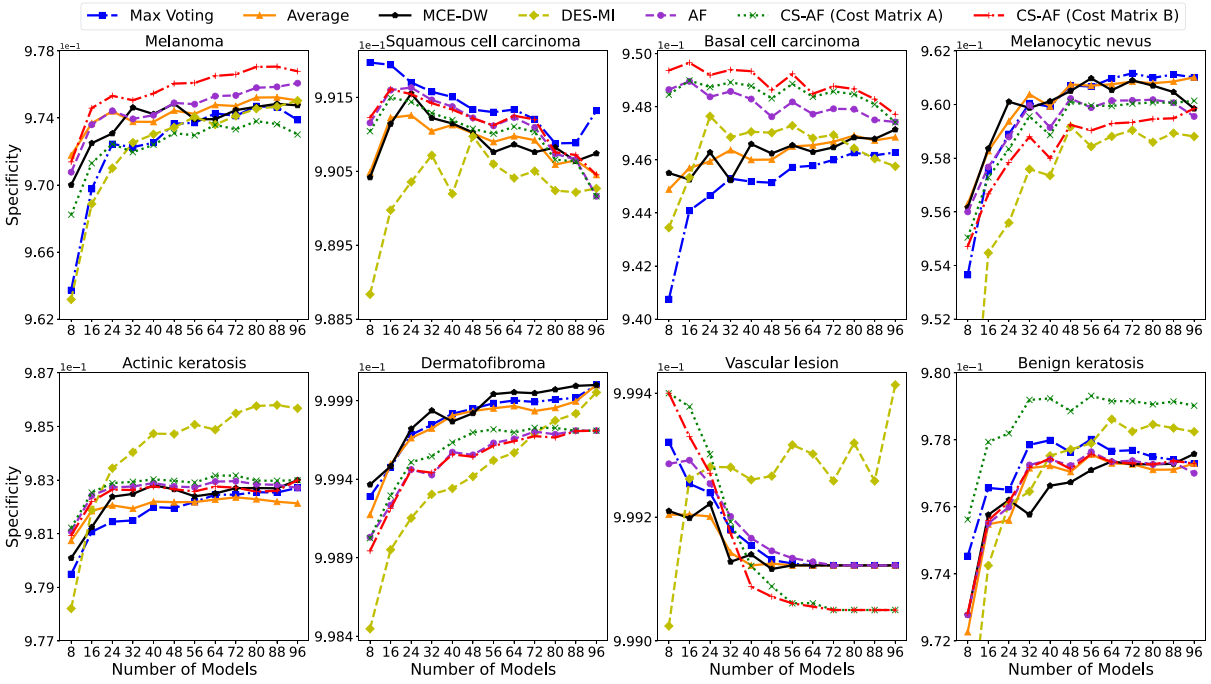


Fig. 8. The specificity results of each single class of CS-AF (Cost Matrix A) and CS-AF (Cost Matrix B).

of DES-MI, it only remains the most confident (i.e., top 40%) base classifiers among all given base classifiers. So, there are only very few base classifiers to be considered to make the final decision while the number of given base classifiers is low.

4.5. Analyze the "Cost-sensitive" of CS-AF

As discussed above, our proposed CS-AF could adapt to different cost matrices and is optimized to achieve the lowest total cost under a specified cost matrix, namely "cost-sensitive". In this section, we would like to analyze how such "cost-sensitive", while using certain customized cost matrices, influences the performance of CS-AF on certain single skin lesion classes, thus reducing the total cost. We evaluate single class performances using sensitivity and specificity, defined as below:

$$sensitivity = \frac{TP}{TP + FN} \tag{8}$$

where *TP* denotes the number of true positives and *FN* denotes the number of false negatives.

$$specificity = \frac{TN}{TN + FP} \tag{9}$$

where *TN* denotes the number of true negatives and *FP* denotes the number of false positives.

Fig. 7 and Fig. 8 illustrate the sensitivity and specificity results of each single class of CS-AF (Cost Matrix A), CS-AF (Cost Matrix B) and all the other competitors, respectively. We can observe that:

- Compared with CS-AF (Cost Matrix B), CS-AF (Cost Matrix A) tends to achieve higher sensitivity on more severe cancerous skin lesion classes (i.e., melanoma, squamous cell carcinoma and basal cell carcinoma), and lower sensitivity on less severe benign skin lesion classes (i.e., benign keratosis, vascular lesion, dermatofibroma, actinic keratosis and melanocytic nevus).
- Compared with CS-AF (Cost Matrix A), CS-AF (Cost Matrix B) tends to achieve higher specificity on those more severe cancerous skin lesion classes, and lower specificity on those less severe benign skin lesion classes.
- Compared with the other competitors, our approaches, CS-AF (Cost Matrix A) obtains the highest sensitivity on the most severe cancerous skin lesion (melanoma), and CS-AF (Cost Matrix B) consistently obtains the highest sensitivity on the most benign skin lesion class (benign keratosis).
- Compared with the other competitors, our approaches, CS-AF (Cost Matrix B) consistently achieves the highest specificity on melanoma, and CS-AF (Cost Matrix A) consistently achieves the highest specificity on benign keratosis.

As described in Section 3.2.1, Cost Matrix A emphasizes on the cancerous skin lesions (i.e., the cost of misclassifyinga a cancerous skin lesion is much more than a benign skin lesion), while Cost Matrix B emphasizes on the benign lesions (i.e., the cost of misclassifyinga a benign skin lesion is much more than a cancerous skin lesion). While using Cost Matrix A to compute the objective weights of our CS-AF implementation (i.e., CS-AF (Cost Matrix A)), it tends to increase the TP and FP of cancerous skin lesion classes and decrease the FN and TN of benign skin lesion classes, thus resulting in higher sensitivity and lower specificity for cancerous skin lesion classes. CS-AF (Cost Matrix B) also works in such way accordingly. Therefore, the Fig. 7 and Fig. 8 demonstrate that our proposed CS-AF is "cost-sensitive", where its performance on certain single skin lesion classes could be adapted to certain customized cost matrices.

5. Conclusion

In this paper, we propose CS-AF, a cost-sensitive multi-classifier active fusion framework for skin lesion classification, where we define two types of weights: the objective weights that are designed according to the classifiers' reliability to recognize the particular skin lesions, and the subjective weights that are designed according to the relative confidence of the classifiers while recognizing a specific previously "unseen" image (i.e., individuality). We also enable the "cost-sensitive" feature in our framework, via incorporating a customizable cost matrix in the design of the objective weights. In the experimental evaluation, we trained 96 classifiers of 12 CNN architectures as the base classifiers, and compared our CS-AF framework with two static fusion approaches (i.e., Max Voting Fusion and Average Fusion), two active fusion competitors (i.e., DES-MI [6] and MCE-DW [5]), and a baseline active fusion approach, AF. Our experimental results show that our CS-AF framework consistently outperforms the static and active fusion competitors in terms of accuracy, and always achieves the lowest total cost. We also demonstrated our "cost-sensitive" feature by using two examples of cost matrices. In the future work, we plan to (i) investigate and incorporate other metrics (i.e., other than F1-score) in the design of the objective weights; (ii) design learning-based approach to determine the subjective weights; and (iii) employ and evaluate our CS-AF framework in other medicine-related applications.

CRediT authorship contribution statement

Di Zhuang: Conceptualization, Methodology, Software, Validation, Investigation, Writing-original-draft, Writing-review-editing, Visualization, Supervision, Project-administration. **Keyu Chen:** Software, Validation, Investigation, Visualization, Data-curation. **J. Morris Chang:** Writing-review-editing, Funding-acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

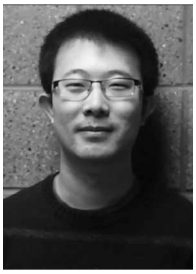
Effort sponsored in part by United States Special Operations Command (USSOCOM), under Partnership Intermediary Agreement No. H92222-15–3-0001–01. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon. [2]

References

[1] J. Zhang, Y. Xie, Y. Xia, C. Shen, Attention residual learning for skin lesion classification, IEEE Trans. Med. Imaging 38 (9) (2019) 2092–2103.
[2] D. Gutman, N.C. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, A. Halpern, Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic), arXiv preprint arXiv:1605.01397.
[3] N.C. Codella, D. Gutman, M.E. Celebi, B. Helba, M.A. Marchetti, S.W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, et al., Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic), in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE, 2018, pp. 168–172.
[4] N. Codella, V. Rotemberg, P. Tschandl, M.E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, et al., Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), arXiv preprint arXiv:1902.03368.
[5] F. Ren, Y. Li, M. Hu, Multi-classifier ensemble based on dynamic weights, Multimedia Tools Appl. 77 (16) (2018) 21083–21107.
[6] S. García, Z.-L. Zhang, A. Altalhi, S. Alshomrani, F. Herrera, Dynamic ensemble selection for multi-class imbalanced datasets, Inf. Sci. 445 (2018) 22–37.
[7] F. Perez, S. Avila, E. Valle, Solo or ensemble? choosing a cnn architecture for melanoma classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.
[8] N.N. Di Zhuang, K. Chen, J.M. Chang, Saia: Split artificial intelligence architecture for mobile healthcare systems, arXiv preprint arXiv:2004.12059.
[9] B. Zeinali, D. Zhuang, J.M. Chang, Esai: Efficient split artificial intelligence via early exiting using neural architecture search, arXiv preprint arXiv:2106.12549.
[10] D. Tao, L. Jin, Y. Yuan, Y. Xue, Ensemble manifold rank preserving for acceleration-based human activity recognition, IEEE Trans. Neural Networks Learn. Syst. 27 (6) (2014) 1392–1404.
[11] D. Zhuang, J.M. Chang, Utility-aware privacy-preserving data releasing, arXiv preprint arXiv:2005.04369.
[12] P.-Y. Wu, C.-C. Fang, J.M. Chang, S.-Y. Kung, Cost-effective kernel ridge regression implementation for keystroke-based active authentication system, IEEE Trans. Cybern. 47 (11) (2016) 3916–3927.
[13] C. Ding, D. Tao, Trunk-branch ensemble convolutional neural networks for

[2] The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the United States Special Operations Command.

video-based face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 1002–1014.
[14] H. Nguyen, D. Zhuang, P.-Y. Wu, M. Chang, Autogan-based dimension reduction for privacy preservation, Neurocomputing.
[15] D. Zhuang, S. Wang, J.M. Chang, Fripal: Face recognition in privacy abstraction layer, in: 2017 IEEE Conference on Dependable and Secure Computing, IEEE, 2017, pp. 441–448.
[16] L. Mai, D.K. Noh, Cluster ensemble with link-based approach for botnet detection, J. Netw. Syst. Manage. 26 (3) (2018) 616–639.
[17] D. Zhuang, J.M. Chang, Peerhunter: Detecting peer-to-peer botnets through community behavior analysis, in: 2017 IEEE Conference on Dependable and Secure Computing, IEEE, 2017, pp. 493–500.
[18] D. Zhuang, J.M. Chang, Enhanced peerhunter: Detecting peer-to-peer botnets through network-flow level community behavior analysis, IEEE Trans. Inf. Forensics Secur. 14 (6) (2018) 1485–1500.
[19] K. Chen, D. Zhuang, J.M. Chang, Discriminative adversarial domain generalization with meta-learning based cross-domain validation, Neurocomputing 467 (2022) 418–426.
[20] A. Tagarelli, A. Amelio, F. Gullo, Ensemble-based community detection in multilayer networks, Data Min. Knowl. Disc. 31 (5) (2017) 1506–1543.
[21] D. Zhuang, M.J. Chang, M. Li, Dynamo: Dynamic community detection by incrementally maximizing modularity, IEEE Trans. Knowl. Data Eng.
[22] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.
[23] R.E. Schapire, The strength of weak learnability, Mach. Learn. 5 (2) (1990) 197–227.
[24] Y. Freund, R.E. Schapire, A desicion-theoretic generalization of on-line learning and an application to boosting, in: European conference on computational learning theory, Springer, 1995, pp. 23–37.
[25] D.H. Wolpert, Stacked generalization, Neural Networks 5 (2) (1992) 241–259.
[26] R.M. Cruz, R. Sabourin, G.D. Cavalcanti, T.I. Ren, Meta-des: A dynamic ensemble selection framework using meta-learning, Pattern Recogn. 48 (5) (2015) 1925–1935.
[27] L.-C. Chen, Y. Yang, J. Wang, W. Xu, A.L. Yuille, Attention to scale: Scale-aware semantic image segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3640–3649.
[28] X. Fang, P. Yan, Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction, IEEE Trans. Med. Imaging 39 (11) (2020) 3619–3629.
[29] M.A. Marchetti, N.C. Codella, S.W. Dusza, D.A. Gutman, B. Helba, A. Kalloo, N. Mishra, C. Carrera, M.E. Celebi, J.L. DeFazio, et al., Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images, J. Am. Acad. Dermatol. 78 (2) (2018) 270–277.
[30] L. Bi, J. Kim, E. Ahn, D. Feng, Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks, arXiv preprint arXiv:1703.04197.
[31] R. Mollineda, R. Alejo, J. Sotoca, The class imbalance problem in pattern classification and learning, in: II Congreso Espanol de Informática (CEDI 2007). pp. 978–84, ISBN, 2007.
[32] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[33] K.-J. Wang, B. Makond, K.-H. Chen, K.-M. Wang, A hybrid classifier combining smote with pso to estimate 5-year survivability of breast cancer patients, Appl. Soft Comput. 20 (2014) 15–24.
[34] H. Han, W.-Y. Wang, B.-H. Mao, Borderline-smote: a new over-sampling method in imbalanced data sets learning, in: International conference on intelligent computing, Springer, 2005, pp. 878–887.
[35] C. Bunkhumpornpat, K. Sinapiromsaran, C. Lursinsap, Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem, in: Pacific-Asia conference on knowledge discovery and data mining, Springer, 2009, pp. 475–482.
[36] T. Maciejewski, J. Stefanowski, Local neighbourhood extension of smote for mining imbalanced data, in: 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2011, pp. 104–111.
[37] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM transactions on intelligent systems and technology (TIST) 2 (3) (2011) 1–27.
[38] S. Panchapagesan, M. Sun, A. Khare, S. Matsoukas, A. Mandal, B. Hoffmeister, S. Vitaladevuni, Multi-task learning and weighted cross-entropy for dnn-based keyword spotting., in: Interspeech, Vol. 9, 2016, pp. 760–764.
[39] T.H. Phan, K. Yamamoto, Resolving class imbalance in object detection with weighted cross entropy losses, arXiv preprint arXiv:2006.01413.
[40] J. Byrd, Z. Lipton, What is the effect of importance weighting in deep learning?, International Conference on Machine Learning, PMLR (2019) 872–881
[41] L. Zhang, D. Zhang, Evolutionary cost-sensitive extreme learning machine, IEEE Trans. Neural Networks Learn. Syst. 28 (12) (2016) 3045–3060.
[42] A. Iranmehr, H. Masnadi-Shirazi, N. Vasconcelos, Cost-sensitive support vector machines, Neurocomputing 343 (2019) 50–64.
[43] S.H. Khan, M. Hayat, M. Bennamoun, F.A. Sohel, R. Togneri, Cost-sensitive learning of deep feature representations from imbalanced data, IEEE Trans. Neural Networks Learn. Syst. 29 (8) (2017) 3573–3587.
[44] Types of skin cancer, url: https://www.aad.org/public/diseases/skin-cancer/types/common, [Online; Accessed 02-04-2022] (2022).
[45] S. Visa, B. Ramsay, A.L. Ralescu, E. Van Der Knaap, Confusion matrix-based feature selection, MAICS 710 (2011) 120–127.
[46] V. Van Asch, Macro-and micro-averaged evaluation measures [[basic draft]], Belgium: CLiPS 49.
[47] M. Combalia, N.C. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, A.C. Halpern, S. Puig, J. Malvehy, Bcn20000: Dermoscopic lesions in the wild, arXiv preprint arXiv:1908.02288.
[48] P. Tschandl, C. Rosendahl, H. Kittler, The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, Sci. Data 5 (2018) 180161.
[49] C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. Yuille, J. Huang, K. Murphy, Progressive neural architecture search, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 19–34.
[50] B. Zoph, Q.V. Le, Neural architecture search with reinforcement learning, arXiv preprint arXiv:1611.01578.
[51] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492–1500.
[52] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
[53] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, J. Feng, Dual path networks, in: Advances in neural information processing systems, 2017, pp. 4467–4475.
[54] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.
[55] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-first AAAI conference on artificial intelligence, 2017.
[56] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[57] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.
[58] M. Tan, Q.V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, arXiv preprint arXiv:1905.11946.



Di Zhuang received his Ph.D. degree in electrical engineering, and B.E. degree in computer science and information security from University of South Florida, Tampa and Nankai University, China, respectively. His research interests include cyber security, social network science, privacy enhancing technologies, machine learning and deep learning. He is a member of IEEE.

Keyu Chen received his master degree in electrical engineering from the University of South Florida, Tampa. He is currently working toward the PhD degree in electrical engineering at the University of South Florida, Tampa. His research interests include machine learning, deep learning, natural language processing and data analytics.

J. Morris Chang is a professor in the Department of Electrical Engineering at the University of South Florida. He received the Ph.D. degree from North Carolina State University. His past industrial experiences include positions at Texas Instruments, Microelectronic Center of North Carolina and AT&T Bell Labs. He received the University Excellence in Teaching Award at Illinois Institute of Technology in 1999. His research interests include: cyber security, wireless networks, and energy efficient computer systems. In the last six years, his research projects on cyber security have been funded by DARPA. Currently, he is leading a DARPA project under Brandeis program focusing on privacy-preserving computation over the Internet. He is a handling editor of Journal of Microprocessors and Microsystems and an editor of IEEE IT Professional. He is a senior member of IEEE.