Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

A fuzzy distance-based ensemble of deep models for cervical cancer detection

Rishav Pramanik[a], Momojit Biswas[b], Shibaprasad Sen[c,*], Luis Antonio de Souza Júnior[d,e], João Paulo Papa[e,f], Ram Sarkar[a]

[a] Department of Computer Science and Engineering, Jadavpur University, 188 Raja S C Mallick Rd, Kolkata, 700032, West Bengal, India
[b] Department of Metallurgical and Material Engineering, Jadavpur University, 188 Raja S C Mallick Rd, Kolkata, 700032, West Bengal, India
[c] Department of Computer Science and Technology, University of Engineering and Management, Kolkata, 700160, West Bengal, India
[d] Department of Computing, São Carlos Federal University-UFScar, São Carlos, São Paulo, Brazil
[e] Regensburg Medical Image Computing (ReMIC), Ostbayerische Technische Hochschule Regensburg (OTH Regensburg), Regensburg, Bavaria, Germany
[f] Department of Computing, São Paulo State University, Av. Eng. Luiz Edmundo Carrijo Coube, 14-01, Bauru, São Paulo, Brazil

ARTICLE INFO

Article history:
Received 8 January 2022
Revised 22 March 2022
Accepted 23 March 2022

Keywords:
Cervical cancer
Computer-aided detection
Deep learning
Fuzzy logic
Ensemble learning

ABSTRACT

Background and Objective: Cervical cancer is one of the leading causes of women's death. Like any other disease, cervical cancer's early detection and treatment with the best possible medical advice are the paramount steps that should be taken to ensure the minimization of after-effects of contracting this disease. PaP smear images are one the most effective ways to detect the presence of such type of cancer. This article proposes a fuzzy distance-based ensemble approach composed of deep learning models for cervical cancer detection in PaP smear images.
Methods: We employ three transfer learning models for this task: Inception V3, MobileNet V2, and Inception ResNet V2, with additional layers to learn data-specific features. To aggregate the outcomes of these models, we propose a novel ensemble method based on the minimization of error values between the observed and the ground-truth. For samples with multiple predictions, we first take three distance measures, i.e., Euclidean, Manhattan (City-Block), and Cosine, for each class from their corresponding best possible solution. We then defuzzify these distance measures using the product rule to calculate the final predictions.
Results: In the current experiments, we have achieved 95.30%, 93.92%, and 96.44% respectively when Inception V3, MobileNet V2, and Inception ResNet V2 run individually. After applying the proposed ensemble technique, the performance reaches 96.96% which is higher than the individual models.
Conclusion: Experimental outcomes on three publicly available datasets ensure that the proposed model presents competitive results compared to state-of-the-art methods. The proposed approach provides an end-to-end classification technique to detect cervical cancer from PaP smear images. This may help the medical professionals for better treatment of the cervical cancer. Thus increasing the overall efficiency in the whole testing process. The source code of the proposed work can be found in github.com/rishavpramanik/CervicalFuzzyDistanceEnsemble.

1. Introduction

Cervical cancer occurs in the cells of the cervix at the lower section of the uterus. Cervical cancer cases are mostly related to infection with high-risk human papillomaviruses (HPV) transmitted through sexual contact. Statistically, cervical cancer alone is the second leading cause of death due to malignancy amongst women [1]. Unfortunately, most of the 2018 recorded cases alone were from low and middle-income countries [2]. As with most of the other diseases, regular checkups and early detection can significantly reduce the chance of fatality [3], with cytological tests (e.g., PaP smear test) one of the most effective ways to detect the occurrence of cervical cancer [3]. PaP smear test is a screening procedure to find the presence of cancerous or precancerous

* Corresponding author.
E-mail addresses: rishavpramanik@gmail.com (R. Pramanik), mb16biswas@gmail.com (M. Biswas), shibubiet@gmail.com (S. Sen), luis.souza.unesp@gmail.com (L.A.d. Souza Júnior), papa.joaopaulo@gmail.com (J.P. Papa), ram.sarkar@jadavpuruniversity.in (R. Sarkar).

https://doi.org/10.1016/j.cmpb.2022.106776
0169-2607/© 2022 Elsevier B.V. All rights reserved.

cells in the cervix (opening of the uterus). During this process, a sample is gently scraped from the cervix and is further examined to know the growth of abnormal cells. Computer-aided Detection (CAD) techniques are becoming very popular nowadays as an alternative to manual diagnosis, for they are less prone to human-like errors [4]. Typically, such methods consider an image as the input and process them for predicting input instances as injured or healthy. Several research articles on the early detection of cervical cancer have been published in the past few years using machine learning [5–7]. These methods are generally trained on a particular dataset to extract specific features for the classification purpose.

Often described as discriminative architecture [8], Convolutional Neural Networks (CNN) have always been the go-to choice considering their astonishing outcomes in comparison to classic handcrafted feature engineering. Initially proposed for document recognition and later extended to multiple fields such as image, video, speech, and audio processing [9], CNNs have been immensely successful while overcoming multidisciplinary research challenges over the past few decades. A typical CNN architecture consists of two parts: a feature extractor and a classifier, which are trained during the learning algorithm. Notably, the use of CNN does not answer every other challenge. CNNs usually require support from other methods to improve performance. One can note numerous approaches developed specifically to aggregate decisions produced by different CNNs. Recently, [10] proposed a weighted Support Vector Machine (SVM) trainable aggregator function to learn an aggregation of multiple CNN architectures concerning handwritten music symbol classification. [11] employed the Fuzzy Choquet integral-based aggregation method for human action recognition. Such aggregator function can be classified as a tunable aggregator, since its performance relies upon a tuning step. However, to get rid of any training or tuning, researchers like [12] used ensemble methods based on the sum and product rules to aggregate the outcomes of CNNs. However, such methods have their pros and cons. For example, in some cases, trainable aggregator functions could be helpful if the number of classes is high and with conflicting predictions given by the base classifiers. In this scenario, such functions may learn the aggregation of features (i.e., confidence scores) to classify a sample correctly. On the other hand, tunable and non-parametric aggregators can work well when the number of target classes is low, and the base classifiers' outcomes are not that conflicting [13].

We propose an ensemble of deep learning models to detect cervical cancer from PaP Smear images in this work. Initially, we resize and augment the images using some popular augmentation techniques described further. Next, these images feed three transfer learning models (each pre-trained on ImageNet dataset) with additional layers of convolution and max pooling, followed by a fully connected network. If the base classifiers contradict themselves to give multiple predictions (multiple base classifiers give multiple predictions), we process their confidence scores through a fuzzy distance-based ensemble mechanism that does not need prior training or fine-tuning. The ensemble method considers three distance measures (i.e., Euclidean, Manhattan, and Cosine). The proposed method is trained and tested on a benchmark dataset, namely the SipakMed PaP Smear dataset [14]. Additional experiments on two other publicly available datasets namely Herlev and Mendeley LBC show the robustness of the proposed method. Experimental results confirm that the proposed method outperforms state-of-the-art methods.

The primary contributions of this paper are listed above:

- We design an ensemble of CNN models to detect cervical cancer from PaP Smear Images. Three transfer learning models and additional layers are used to learn data-specific features that are considered base learners.
- We propose a novel fuzzy distance-based aggregator function that minimizes the difference between the observed and ground-truth (ideal solution) samples.
- The proposed ensemble technique considers distances from the ideal solution in three different spaces. It can successfully aggregate confidence scores generated by base learners so that the ensemble's performance can be improved.
- The proposed model outperforms many state-of-the-art methods when evaluated on three standard and publicly available cervical cancer datasets.

The remainder of the paper is organized as follows: Section 2 surveys deep learning-based methods proposed to cope with cervical cancer detection from Pap smear images. Section 3 presents a detailed description of the proposed approach, and Section 4 discusses methodology and experimental results. Section 5 provides an analysis and discussion of the methodology. Finally, Section 6 states concluding remarks.

2. Related research

Park et al. [15] provided a comparative study of some popular machine-learning-based architectures and a deep learner. The comparison concluded that deeper layers help to learn high-level features, which was reflected in the classification results. At the same time, machine learning-based methods require additional human experts to select relevant feature sets for more effective training. Bora et al. [16] employed AlexNet (a popular and one of the oldest CNN architectures) for feature extraction followed by a Maximal Information Compression Index (MICI)-based unsupervised feature selection. Finally, two classifiers, i.e., Softmax Regression and Least Square Support Vector Machines (LSSVM), were used to monitor the model's performance. The authors found out that the use of feature selection could significantly boost performance. Indeed, MICI has been widely used for clustering/classification tasks, and it uses k-NN to group similar samples. However, k-NN uses distances to evaluate the similarities, and it is highly prone to errors in higher dimensions. The work of [17] stands as one of the very first that used deep transfer learning in the context of cervical cancer detection from PaP smear images. They considered a CNN pretrained on the ImageNet dataset for image classification purposes. Before feeding the network, a nucleus center was required to extract the data patch, which was accomplished by a ground-level segmentation mask. Notice this structure may not be straightforward to find for an abnormal cell.

Nanni et al. [18] extracted deep-features using CNN for further feeding an SVM classifier. The sum rule was then used to classify images. Similarly, [19] used texture-based features such as Grey Level Co-occurrence Matrix and Gabor, along with features extracted by a CNN model. They used the Principal Component Analysis for dimensionality reduction and SVM classification purposes. In both cases, the authors showed significant performance improvement. Ghoneim et al. [20] used an extreme learning machine (ELM) classifier with CNN for cervical cancer classification. The proposed approach first extracts fine-tuned features from a CNN and then uses them to feed an ELM classifier. Notably, the use of Graph Convolution Networks (GCNs) has also been explored recently in the context of cervical cancer by Shi et al. [21]. Both CNN and GCNs were used for feature extraction purposes. However, GCNs may also require additional data for validation purposes, as stated by Li et al. [22]. The authors concluded that a drop in performance might occur when we do not consider a validation set when training GCNs.

Manna et al. [23] used three transfer learning models and a fuzzy-rank-based ensemble method for cervical cytology classification. Lin et al. [24] employed morphological and appearance infor-

Table 1
Description of the dataset used in the experiments.

Index	Class	Category	Number of images
0	Dyskeratotic	Abnormal	813
1	Koilocytes	Abnormal	825
2	Metaplastic	Benign	793
3	Parabasal	Normal	787
4	Superficial-Intermediate	Normal	831

mation to train four CNN models with explicitly augmented data and further used fully connected networks for classification purposes. A fused domain-level approach has been used to combine the models' outcomes. On the other hand, the results revealed that the proposed approach performed somewhat biased to some particular class and, therefore, could not perform similarly for images of all types. Elakkiya et al. [25] used a Generative Adversarial Network (GAN) for cervical cancer detection where the generator and discriminator are composed of a a Region-based Convolution Neural Network (R-CNN). Interestingly, in an article by Wang et al. [26], a pruning strategy was applied on a transfer learning-based model for the classification of PaP smear images. However, the method required much computational effort for training, as pruning the CNN model was iteratively learned for a certain number of epochs.

In another research, [27] fused color, morphological and texture-based features with CNN-based features for classification purposes. The authors did not use any data augmentation technique in their work, and hence the CNN-based features might be incompatible with dealing with rotated and translated images, which is usually required in a real-life scenario. A recent article by Zhang et al. [28] employed quantitative features using cell DNA Index (DI) and rectified DI computed from images and used a fine-tuned Fully Connected Network Long Short-term Memory (FCN-LSTM) for classification. A fuzzy non-linear regression-based rectifying method was used for cell qualitative analysis. This method performed satisfactorily with most types of cervical cancer. However, specific varieties of dysplasia (e.g., lesions in the cervical canal) are very likely to be missed, leading to undesirable outputs.

Cao et al. [29] used a patch-based attention network consisting of a feature pyramid network and an R-CNN with DenseNet 169 as the backbone for cervical cancer detection. The experimental protocol comprised a binary classification, which might not be helpful for clinicians for treatment purposes. The specific presence of certain types of cells would help clinicians for better treatment purposes.

3. Methods and materials

This section presents the proposed approach for cervical cancer detection using PaP smear images. First, training samples are resized, and then the images are augmented using online augmentation techniques that include random zooming, shifting, flipping, and rotation. Augmentation is performed by taking into consideration that CNN models become more competent to handle translation and rotation-related problems [30]. These augmented images are then fed to three ImageNet pre-trained CNN models with additional layers for further fine-tuning on the dataset. The additional layers consist of convolution, max-pooling, and fully connected layers, as depicted in Fig. 1. It is often inconceivable to have large (and annotated) datasets in biomedical image analysis. On the other hand, transfer learning techniques come to deal with dataset dependencies [31]. For such a reason, we employed a CNN pre-trained on ImageNet dataset for the current experiment.

We considered three different CNN architectures from diverse backgrounds so that proper evaluations could be performed. Confidence scores are extracted from each trained CNN, and their aggregation is performed using a fuzzy distance-based ensemble method. For conflicting samples (i.e., multiple base classifiers output different predictions), a particular class is finally predicted using three different distance measures. Fig. 2 illustrates the entire pipeline of the proposed approach.

3.1. Dataset description

The proposed model is evaluated on a publicly available dataset named SIPaKMeD[1] provided by Plissiti et al. [14]. The dataset contains Pap smear images from five categories: (i) Superficial-Intermediate, (ii) Parabasal, (iii) Koilocytes, (iv) Dyskeratotic, and (v) Metaplastic. The dataset comprises 4,049 images, out of which 966 slides of cluster cell images were cropped. This dataset suffers from mild class imbalance, as detailed in Table 1.

3.2. Inception V3

Inception V3 is the third generation of Google's Inception-based CNN architecture Proposed by Szegedy et al. [32]. Typically Inception architectures mainly focus on using less computational resources to train deep CNN architectures. Unlike typical models like VGG-Net, Inception V3 uses smaller convolutions (smaller filter sizes), which proved to be effective and computationally cheaper. Compared to trivial convnets like the VGG-Net family, it uses a constant filter size (although small) for convolution. In contrast, Inception V3 uses asymmetric convolutions, which capture the spatial dependency amongst the features. The output features are concatenated together and processed onto the next layer. The architecture employs an auxiliary classifier inspired by earlier versions of Inception-based architecture, being a shallow CNN model used within layers during the training phase.

3.3. MobileNet V2

MobileNet V2 is a portable implementation of a state-of-the-art CNN architecture proposed by Sandler et al. [33], which is based on its predecessor MobileNet V1 that adopts depth-wise separable convolutions. The authors were able to showcase its effectiveness along with reductions in computational efforts significantly. Besides, linear bottleneck layers were introduced to overcome ReLU's limitation regarding preservation of complete information. Inverted residual blocks were also introduced, which are very similar to residual connections [34].

3.4. Inception ResNet V2

Two mainstream image recognition architectures, i.e., the residual models [34] and the inception models [32], can be combined to obtain better models [35]. For such a purpose, less computationally expensive inception blocks are employed, followed by a 1×1 convolutional layer (for filter expansion) to scale up the convolutions as compensation for the dimensional reduction of inception blocks. Inception-based models are usually computationally inexpensive, and the fusion of residual features with higher-level features could successfully scale up the optimization process without thinking much about the vanishing gradient problem [34]. Deep residual models can quickly scale up to 100 layers and encode higher-level features. However, prohibitive to VGG-like Nets that can barely support 13 to 16 convolutional layers. A major difference with Inception models is that batch normalization is added to the top of convolutional layers rather than to the top of residual blocks. Theoretically, the use of batch normalization would be

[1] https://www.cse.uoi.gr/~marina/sipakmed.html.

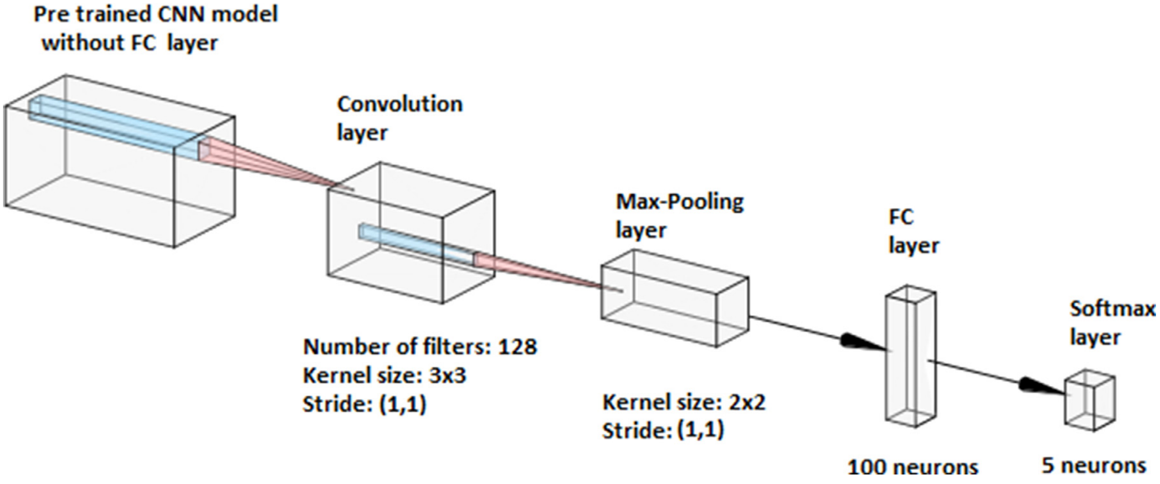

Fig. 1. The overall pipeline of the deep model. The Rectified Linear Unit (ReLU) activation function is used for all layers, and no padding has been adopted concerning the convolutional layers. The number of output neurons fits the number of classes (Table 1).

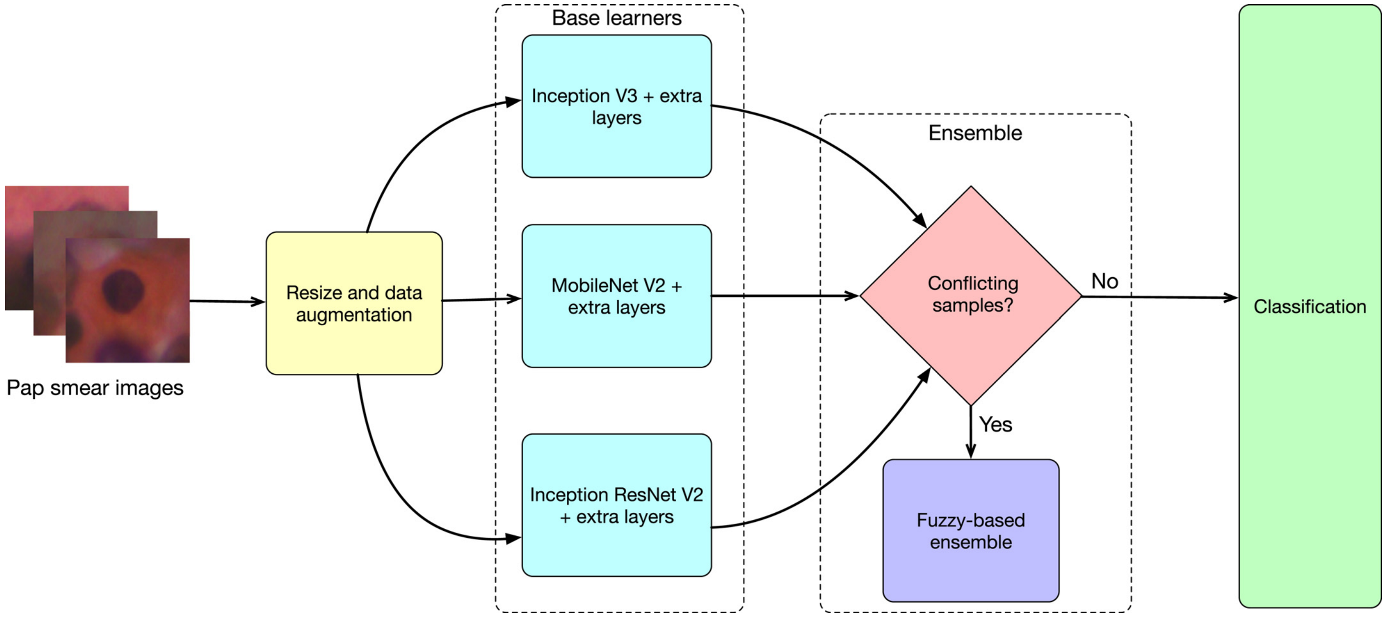
Base learners
Inception V3 + extra layers
MobileNet V2 + extra layers
Inception ResNet V2 + extra layers
Pap smear images
Resize and data augmentation
Ensemble
Conflicting samples?
No
Yes
Fuzzy-based ensemble
Classification

Fig. 2. Overall pipeline of the proposed approach for cervical cancer detection.

advantageous. Experimentally, the latter resulted in high memory usage. These modifications could successfully scale up the performance when compared to both residual network-based [34] and inception-based architectures [32]. In this work, we use Inception ResNet V2, a computationally costlier (though effective) version of Inception ResNet.

3.5. Fuzzy distance-based ensemble

The proposed fuzzy distance-based ensemble approach is designed based on the principle that the difference between observed and ideal solutions should be minimum. In our case, the observed solution is the confidence score generated by a base classifier about the sample's label under consideration. The ideal solution is '1', i.e., the highest possible score generated by a base classifier for a given class. This difference is calculated as a consensus formed by the product rule among three differences determined by Euclidean, Manhattan, and Cosine distances. The constituted agreement helps to form a robust decision-making process, thus making the entire method more reliable. At first, we determine if all base learners output targets a particular class. In that case, the corresponding class is treated as the final prediction. Otherwise, we proceed with the fuzzy distance-based ensemble to make the final prediction, where the confidence scores from each classifier and its corresponding class labels are considered for the final outcome.

Let $\mathcal{X} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$ be a dataset such $\boldsymbol{x}_i \in \mathbb{R}^n$ stands for a given sample and $y_i \in \{1, 2, \ldots, c\}$ denotes its corresponding class label. Let $S_z^j(\boldsymbol{x}_i)$ be the confidence score of sample $\boldsymbol{x}_i$ assigned by the zth classifier corresponding to the jth class, $z = 1, 2, \ldots, M$. For each sample $\boldsymbol{x}_i$ and class label j, we can build the ensemble by computing the distance between the "ideal” solution vector $\mathbf{1} = \{1\}_{i=1}^{M}$ and the confidence scores of each classifier, i.e., $P_j(\boldsymbol{x}_i) = \left(\mathbf{1} - S_1^j(\boldsymbol{x}_i), \mathbf{1} - S_2^j(\boldsymbol{x}_i), ..., \mathbf{1} - S_M^j(\boldsymbol{x}_i)\right)$. For we consider three distance measures, we have one ensemble for each, i.e., $P_j^E(\boldsymbol{x}_i)$, $P_j^M(\boldsymbol{x}_i)$, and $P_j^C(\boldsymbol{x}_i)$, which stand for the ensembles concerning Euclidean, Manhattan and Cosine distances, respectively. We can gather them all in a single variable $P_j^*(\boldsymbol{x}_i) = (P_j^E(\boldsymbol{x}_i), P_j^M(\boldsymbol{x}_i), P_j^C(\boldsymbol{x}_i))$ concerning the jth class label. Last but not least, if we take into account all class labels

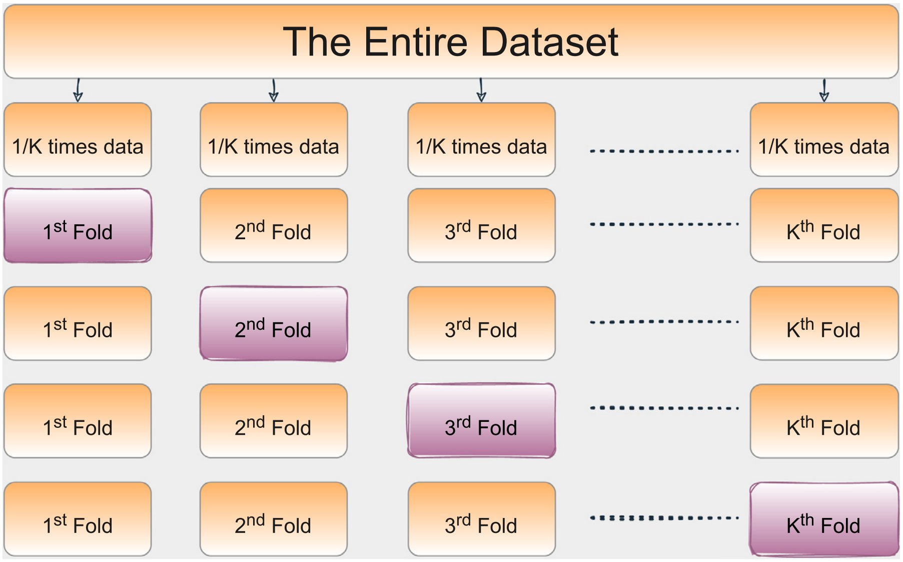

Fig. 3. Illustration of k-fold cross-validation. Each time, the pink fold is used for the testing and the remaining for training purposes.

Table 2
Toy example of the proposed approach based on fuzzy distances.

(a) Confidence scores obtained from base classifiers

Class	Confidence scores generated by base classifiers		
	Inception V3	MobileNet V2	Inception ResNet V2
Superficial-Intermediate	0.321	0.242	**0.410**
Parabasal	0.165	**0.365**	0.252
Koilocytes	**0.425**	0.271	0.172
Dyskeratotic	0.085	0.043	0.029
Metaplastic	0.014	0.079	0.137

(b) Distance Measures

Class	Distance Measures			Product
	Euclidean	Manhattan	Cosine	
Superficial-Intermediate	**1.176**	**2.027**	**0.022**	**0.052**
Parabasal	1.288	2.218	0.046	0.131
Koilocytes	1.244	2.132	0.059	0.157
Dyskeratotic	1.642	2.843	0.090	0.419
Metaplastic	1.602	2.770	0.164	0.726

$j = 1, 2, \ldots, c$, the ensemble can be represented as follows: $P(\boldsymbol{x}_i) = (P_1^*(\boldsymbol{x}_i), P_2^*(\boldsymbol{x}_i), \ldots, P_M^*(\boldsymbol{x}_i))$.

To combine information from the three distance measures, we use the product rule for each class label, i.e., $\Lambda_j(\boldsymbol{x}_i) = P_j^E(\boldsymbol{x}_i) \times P_j^M(\boldsymbol{x}_i) \times P_j^C(\boldsymbol{x}_i)$, and assign the class label that minimizes equation below:

$$\hat{y}_i = \arg\min_j \{\Lambda_j(\boldsymbol{x}_i)\}, \qquad (1)$$

where $\hat{y}_i$ is the class label assigned to sample $\boldsymbol{x}_i$. The product rule works as a fuzzy measure, and the minimum of such defuzzified values is termed the final prediction. The product rule is used to defuzzify over majority vote or sum rule by considering it can normalize the range of values into a single one, while the sum rule or majority voting can not.

Table 2 presents a toy example of the formulation described above. Considering a particular sample $\boldsymbol{x} \in \mathcal{X}$, Table 2a bestows the confidence scores generated by the base classifiers corresponding to each class, i.e., $S_z^j(\boldsymbol{x}_i)$ (we used $z = 3$ in this work). Table 2b presents the distance measures for each class, i.e., $P_j(\boldsymbol{x}_i)$, calculated from the ideal solution **1**, which stands for the highest possible confidence score generated by a base classifier. Lastly, the product of these distance measures is calculated for each class, i.e., $\Lambda_j(\boldsymbol{x}_i)$, and the class with the smallest score is defined as the final prediction $\hat{y}_i$.

It is worth noting that distances differ in the order of their magnitude. Hence, it becomes essential to preserve their originality and normalize them into a single value to make the final prediction. The sum rule may not be appropriate since it is biased towards the distance value regarding the highest order of magnitude (e.g., Manhattan distance). Therefore, the product rule is a reasonable choice for such a task.

4. Results

4.1. Evaluation techniques

To validate the effectiveness of the proposed approach, we have used k-fold cross-validation scheme. Initially, the dataset is divided into k different parts. Each time while experimenting, we use one fold for the testing purpose and the remaining $(k-1)$ parts for the

Table 3
Results for a 5-fold cross validation methodology. The best results (absolute values) are highlighted in bold.

Metric	Inception V3	MobileNet V2	Inception ResNet V2	Proposed method
Accuracy	**96.63** ± **0.74**	94.19± 0.83	96.13 ± 0.46	96.47 ± 0.48
Precision	94.02 ± 0.79	94.20 ± 0.88	94.19 ± 0.40	**96.51** ± **0.54**
Recall	93.72 ± 0.68	94.23 ± 0.92	96.21 ± 0.39	**96.53** ± **0.54**
F-1	93.67 ± 0.74	94.17 ± 0.89	96.17 ± 0.42	**96.45** ± **0.54**

Table 4
Results obtained by the proposed method using the 5-fold cross validation technique on Herlev dataset. The best results (absolute values) are highlighted in bold.

Metric	Inception V3	MobileNet V2	Inception ResNet V2	Proposed method
Accuracy	97.67 ± 0.70	96.95 ± 0.90	97.21 ± 0.50	**98.58** ± **0.40**
Precision	97.80 ± 0.50	97.10 ± 0.80	97.38 ± 0.40	**98.65** ± **0.40**
Recall	97.58 ± 0.80	96.85 ± 1.00	97.12 ± 0.60	**98.53** ± **0.40**
F-1	97.66 ± 0.70	96.93 ± 0.90	97.20 ± 0.51	**98.58** ± **0.40**

training purpose. This process is repeated for k times. Fig. 3 illustrates such a procedure.

4.2. Evaluation metrics

The evaluation metrics used to determine the effectiveness of the proposed method can be found underneath:

- True Positive (TP): an outcome where the model correctly predicts the positive class.
- True Negative (TN): an outcome where the model correctly predicts the negative class.
- False Positive (FP): an outcome where the model incorrectly predicts the positive class.
- False Negative (FN): an outcome where the model incorrectly predicts the negative class.
- Accuracy is the ratio between correctly classified samples and the dataset size. We calculate accuracy using the following equation:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{2}$$

- Precision is defined as the ratio between the correctly predicted samples for a particular class and the total number of samples, and it is calculated as follows:

$$Precision = \frac{TP}{TP + FP}. \tag{3}$$

- Recall is defined as the ratio between the true positive samples and the total number of elements that belong to that positive class, and it is calculated as follows:

$$Recall = \frac{TP}{TP + FN}. \tag{4}$$

- F1 score is the harmonic mean between precision and recall score, and it is calculated as follows:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}$$

Table 5
State-of-the-art comparison on Herlev dataset.

Method	Accuracy
[17]	98.30
[24]	94.50
[7]	98.91
[27]	98.20
[41]	98.27
Proposed	**98.58**

4.3. Experimental results

Table 3 presents an initial comparison of accuracy, precision, recall, and F-1 scores (%) concerning the proposed approach and each model used to compose the ensemble individually. One should keep in mind that the proposed approach is neither parametric nor tunable, which makes it equally biased towards all base models. Since the proposed method is entirely based on the distance from the best possible solutions, a better confidence score obtained by the base classifier for a particular class ensures proper attention

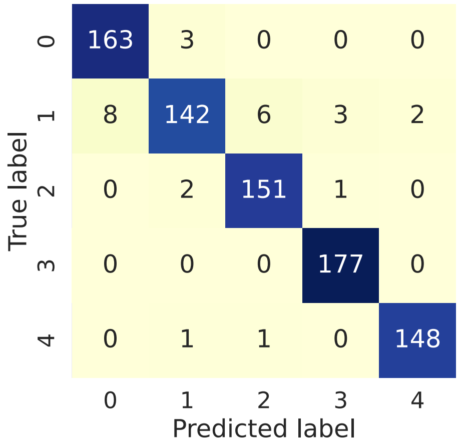

Fig. 4. Confusion Matrix for the fifth fold, the index can be referred to the classes in Table 1.

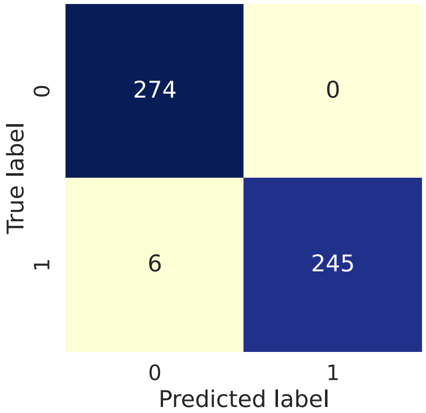

Fig. 5. Confusion matrix for the fifth fold on Herlev dataset. More details about the indices can be found in Section 4.4.

Table 6
Results obtained by the proposed method using the 5-fold cross validation on Mendeley LBC dataset. The best results (absolute values) are highlighted in bold.

Metric	Inception V3	MobileNet V2	Inception ResNet V2	Proposed method
Accuracy	99.16 ± 0.28	98.53 ± 0.44	99.16 ± 0.50	**99.68** ± **0.29**
Precision	98.14± 0.70	96.98 ± 0.90	98.14 ± 0.70	**99.34** ± **0.59**
Recall	99.31 ± 0.44	97.95± 0.95	99.31 ± 0.44	**99.87** ± **0.12**
F-1	98.70 ± 0.59	97.41 ± 0.87	98.70 ± 0.58	**99.60** ± **0.36**

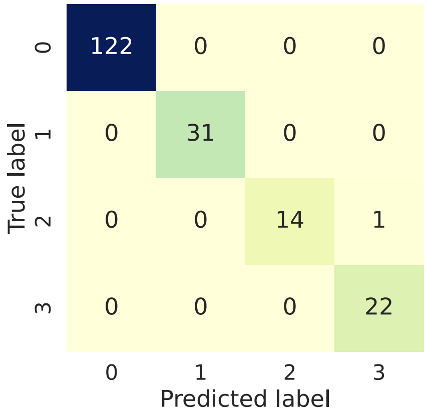

Fig. 6. Confusion matrix for the fifth fold on Mendeley LBC dataset. More details about the indices can be found in Section 4.4.

towards that particular class. One can perceive the proposed approach outperformed all single classifiers in all measures, but for the accuracy. Since the dataset is slightly imbalanced, standard accuracy may not be a healthy measure to be solely considered. The obtained confusion matrix is presented in Fig. 4.

Additionally, to evaluate the effectiveness of the proposed method under different scenarios, we also performed a 20-fold cross-validation scheme. Table 9 presents the results, where one can observe similar behavior regarding the 5-fold experiment (Table 3), but with the proposed approach outperforming the individual architectures in all measures. We can observe in Table 9 that the reported metric are higher than the previous experiments. The observed difference can be explained on the basis that, with much more training data: the deep learners are able to learn more local features and thereby producing much more confident outcomes and thus the the ensemble process is able to learn the aggregation thus proving the robust nature of the proposed ensemble method.

4.4. Additional tests

To validate the effectiveness of the proposed method on other datasets, we further evaluate our model on two more publicly available datasets namely: Herlev[2] (2-class) and Mendeley LBC[3] (4-class) using the 5-fold cross validation. The Herlev dataset consists of two classes namely abnormal (index-0) and normal (index-1). The Mendeley LBC dataset consists of four classes: negative for intra-epithelial malignancy (index-0), high squamous intra-epithelial lesion (index-1), squamous cell carcinoma (index-2), low squamous intra-epithelial lesion (index-3). For Herlev dataset, the obtained results are provided in Table 4 and the corresponding confusion matrix is presented in Fig. 5. Table 6 shows the results for Mendeley LBC dataset and the confusion matrix is presented in Fig. 6. The state-of-the-art comparisons for Herlev and Mendeley LBC datasets are given in Tables 5 and 7, respectively.

Table 7
State-of-the-art comparison on Mendeley LBC dataset.

Method	Accuracy
[23]	99.23
[40]	98.44
[42]	99.47
Proposed	**99.68**

Table 8
AUC-ROC scores concerning all five classes of the SIPaKMeD dataset. The reported scores are for the fifth fold.

Class	AUC-ROC Score
Dyskeratotic	0.98
Koilocytes	0.94
Metaplastic	0.98
Parabasal	1.00
Superficial-Intermediate	0.99

Table 8 presents the area under the ROC curve for the fifth fold of the experiment. Fig. 7 reflects loss variations during the convergence process using the 5-fold cross validation protocol.

5. Discussion

This section discusses some essential hyper-parameters used in this current experiment and their relevant analysis. Finally, we compare our method with some state-of-the-art approaches.

5.1. Hyper-parameter tuning

Choosing proper hyper-parameters is crucial in deep learning to control learning capabilities. Since the batch size and learning rate greatly influence convergence, we resorted to a grid search approach to seek suitable values in a validation set created by randomly sampling 20% of the training samples. Fig. 8 depicts an ablation study concerning different learning rates and batch sizes. We randomly selected 20% of the input images for testing and the remaining for training purposes for this experiment. We considered a learning rate range that is commonly adopted in the literature, i.e., $\{1e-3, 1e-4, 1e-5, 1e-6\}$ [36]. Concerning the batch size, we also used standard values, i.e., $\{8, 16, 32\}$. From Fig. 8, one can perceive that with a batch size of 16 and a learning rate of $1e-4$, the model performs best. However, when the learning rate goes down to $1e-6$, the model's performance drops significantly once the base learning rate used in this case cannot capture relevant information from the images. We use the Cross-Entropy [37] as the loss function and the Adam optimizer [38]. The hyper-parameters used for experimentation can be found in Table 10.

5.2. Analysis

Fig. 7 depicts the loss variation during training, which allows us to conclude that none of the CNN models has suffered from any significant overfitting. Also, it can be observed that in the initial

[2] http://mde-lab.aegean.gr/index.php/downloads.

[3] https://data.mendeley.com/datasets/zddtpgzv63/4.

Table 9
Fold-wise results of the 20-fold cross-validation technique. Acc, Prec, Rec and F1 refer to accuracy, precision, recall and F1 score, respectively. All metrics are reported in (%).

	Inception V3				MobileNet V2				Inception-ResNet V2				Ensemble			
Fold	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
Fold 1	96.55	96.51	96.33	96.40	96.55	96.52	96.39	96.38	96.55	96.48	96.36	96.41	97.04	97.00	96.88	96.91
Fold 2	94.58	94.54	95.13	94.77	94.58	94.33	94.73	94.41	96.05	95.99	96.48	96.17	95.56	95.44	96.00	95.64
Fold 3	95.56	95.33	95.62	95.44	93.59	93.36	93.65	93.40	94.08	93.95	94.61	94.21	95.07	94.80	95.18	94.97
Fold 4	96.55	96.78	96.46	96.57	94.08	94.33	94.02	94.10	97.04	97.29	96.95	97.09	97.53	97.68	97.45	97.54
Fold 5	97.53	97.45	97.64	97.53	94.08	94.50	94.43	94.20	97.53	97.54	97.58	97.51	97.53	97.61	97.78	97.58
Fold 6	97.53	97.48	97.55	97.51	96.05	96.08	96.10	96.03	98.02	98.01	98.01	98.01	98.52	98.49	98.55	98.51
Fold 7	96.05	96.06	96.20	96.05	96.05	96.23	96.19	95.94	97.53	97.56	97.64	97.58	98.02	98.09	98.13	98.06
Fold 8	94.58	94.30	93.94	94.07	94.08	94.06	93.85	93.83	96.05	96.69	95.60	96.00	97.04	97.33	96.53	96.78
Fold 9	97.04	97.01	96.95	96.97	95.56	95.73	95.19	95.37	95.07	94.95	95.17	95.01	97.53	97.44	97.49	97.43
Fold 10	93.06	93.09	93.25	92.90	92.07	92.94	92.60	92.45	95.04	95.19	95.23	95.20	96.03	96.39	96.13	96.21
Fold 11	94.05	93.42	95.48	94.23	94.05	93.65	95.66	94.40	96.03	96.75	95.42	95.98	98.01	97.82	98.42	98.10
Fold 12	94.05	94.12	94.38	94.19	92.07	92.43	92.50	92.43	99.00	99.00	99.04	99.01	97.52	97.54	97.67	97.59
Fold 13	95.04	94.26	94.78	94.42	93.56	92.87	93.16	92.89	96.03	95.71	95.83	95.72	97.52	96.96	97.25	97.09
Fold 14	95.04	95.07	94.85	94.85	95.54	95.23	95.38	95.14	97.02	96.99	97.04	96.99	97.02	96.99	97.04	96.99
Fold 15	95.04	95.08	95.05	95.04	91.08	90.79	90.99	90.68	96.03	96.00	96.01	95.96	97.52	97.46	97.64	97.51
Fold 16	95.04	95.02	94.89	94.92	88.11	88.68	87.86	88.03	96.03	96.22	95.94	96.03	95.54	95.81	95.53	95.61
Fold 17	96.03	96.02	96.11	96.05	94.55	94.52	94.49	94.38	97.02	97.19	96.83	96.93	96.53	96.51	96.61	96.53
Fold 18	93.56	93.57	93.52	93.48	92.57	92.74	92.45	92.46	95.04	94.93	94.93	94.90	95.54	95.45	95.46	95.43
Fold 19	95.04	95.06	94.94	94.97	95.54	95.47	95.46	95.45	97.52	97.48	97.32	97.37	98.01	97.91	97.86	97.88
Fold 20	94.05	93.79	93.64	93.59	94.55	94.21	94.29	94.21	96.03	95.92	95.71	95.79	96.03	95.77	95.74	95.74
Average	**95.30**	**95.20**	**95.34**	**95.20**	**93.92**	**93.93**	**93.97**	**93.81**	**96.44**	**96.49**	**96.39**	**96.39**	**96.96**	**96.92**	**96.97**	**96.91**
Std Dev	1.28	1.35	1.28	1.34	2.00	1.90	2.03	1.98	1.17	1.21	1.15	1.17	1.00	1.02	1.03	1.01

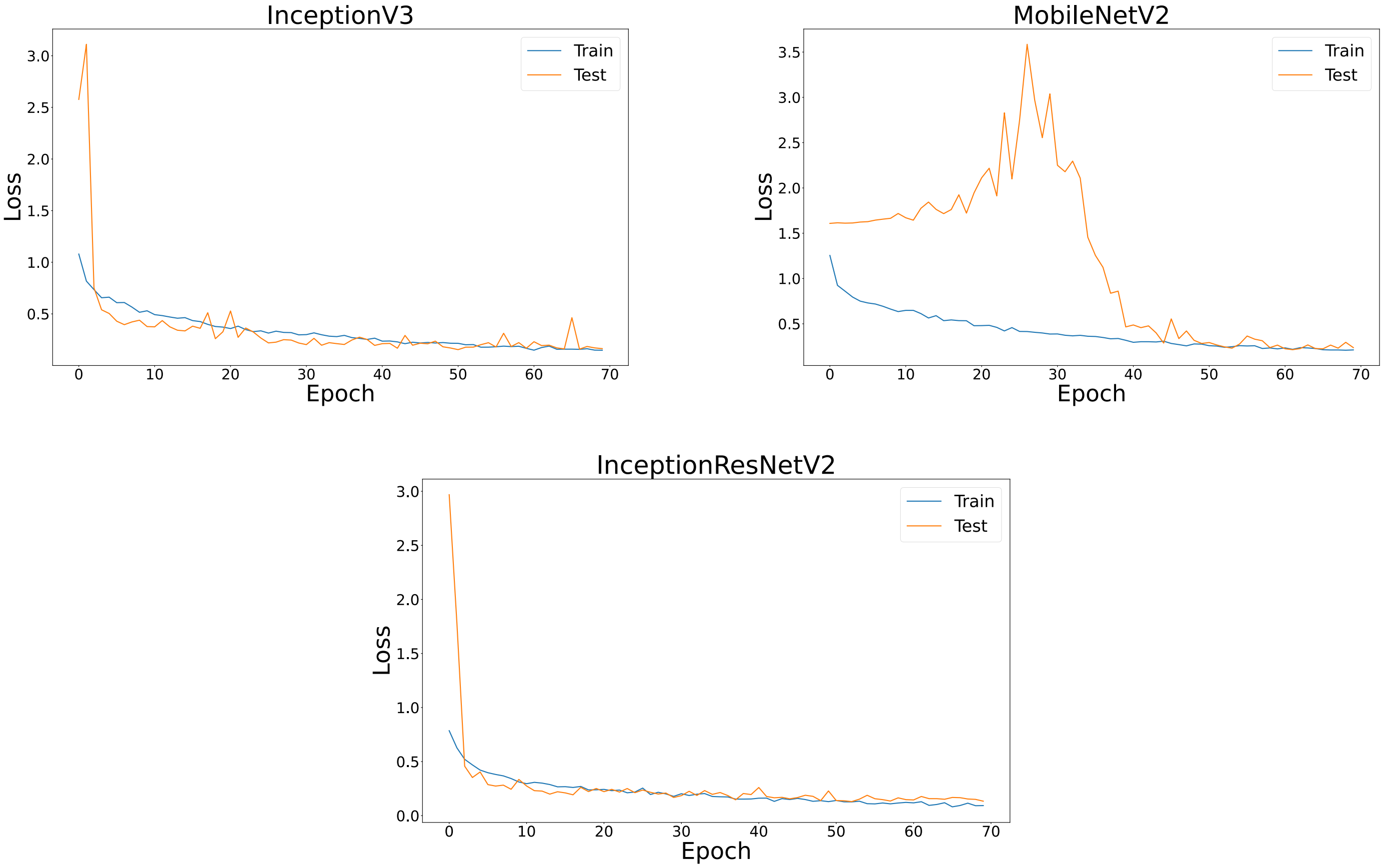

Fig. 7. Loss variation during the convergence process using the 5-fold cross validation protocol. These results correspond to the fifth fold.

epochs, the MobileNet V2 model is not able to converge. This behavior may be explained by the fact that in the initial epochs, the model is stuck at any local optimum. With progression in epoch and subsequently changing the learning rate, the model can finally converge to its optimum without overfitting. Whereas in the case of other deep learning models, in the initial few epochs, the models behave uncertainly, however, with progression in learning over the epochs, the models appear to give more stable results. It is interesting to note in Fig. 7 that in the final epochs, the value of test loss is very similar to the value of training loss. From this observation, we can safely comment that the deep learning models do not tend to overfit at any instance.

Table 10
The hyperparameters used for experimentation purposes.

Hyper-Parameter	Value/Method
Learning Rate	0.0001
Batch Size	16
Epochs	70
Optimizer	Adam
Loss	Cross-Entropy

Table 11
T-test comparison.

Method	*p*-value
[14]	0.00216
[23]	0.00501
[40]	0.04657

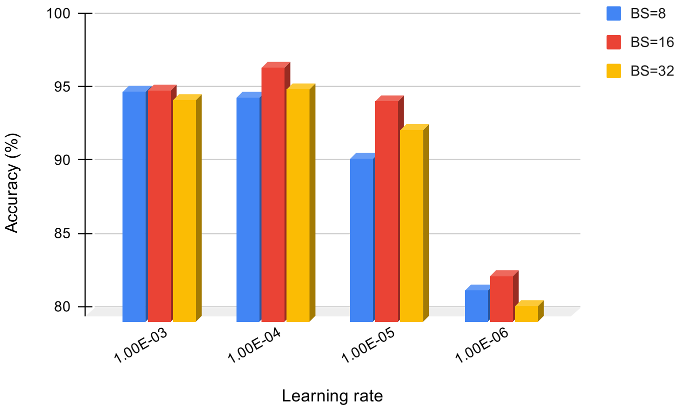

Fig. 8. Ablation study concerning the proposed approach with varied learning rates and batch sizes.

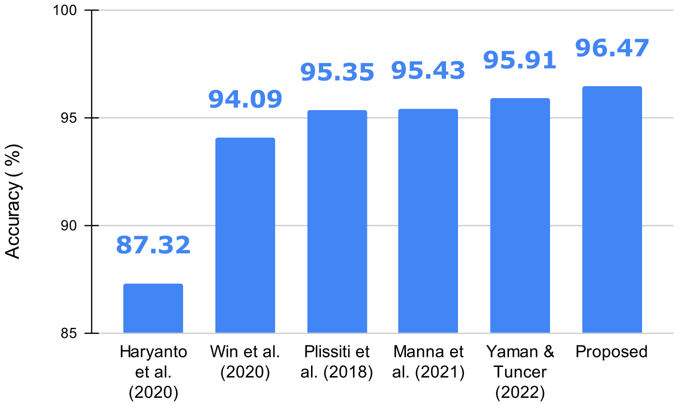

Fig. 9. Comparison with some state-of-the-art approaches.

Table 8 presents the area under the Receiver Operating Characteristics (ROC) curve concerning the fifth experimental fold. One can observe that the proposed method can satisfactorily handle the class imbalance issue, i.e., we can safely claim that the proposed method does not behave in a sub-optimal way for any of the classes. Hence, outstanding results can be observed in all five categories. The “Parabasal” category, for instance, comprises 787 samples, and it is the smaller one in the dataset. However, we can observe that the proposed approach achieves an AUC = 1.0, i.e., the highest possible one. Such an outcome corroborates one of the main motivations of the proposed approach, i.e., to make the learner less prone to imbalanced datasets and poor decisions from individual models.

5.3. Comparison with state-of-the-art approaches

We considered a fair comparison of the proposed approach with some works conducted recently [23,14]. We used a standard 5-fold cross-validation methodology to evaluate the effectiveness of all techniques on the SiPakMed dataset. Fig. 9 illustrates that the proposed approach outperformed the other techniques considered in our experiments. Out of these methods, the ones proposed by Haryanto et al. [39], Manna et al. [23], Plissiti et al. [14], Yaman and Tuncer [40] are based on deep transfer learning methods, and the one suggested by Win et al. [41] consists of two stages, i.e., segmentation and classification. Manna et al. [23] have used three traditional deep learning models and, in addition, they compared their achievements with several other transfer learning models. In contrast, it is worth noting that the proposed approach consists of some extra layers to learn more data-specific features. In addition, the proposed ensemble method can give a better aggregation of confidence scores aggregating information from different distances.

To evaluate the robustness of the proposed approach, we performed a statistical test to decide whether it performs better or not in comparison to the state-of-the-art methods considered in the manuscript. Our hypothesis is: *The proposed ensemble-based approach with fuzzy distances does not outperform the approaches compared in the experiments.* To check whether the hypothesis is true or false, we performed the well-known t-test using the 5-fold experiment. Table 11 presents the *p*-values obtained during the experiments[4].

Since we used a significance of 5%, we can claim the proposed approach is statistically superior to the other techniques compared in the statistical test. Such results corroborate our assumption that combining different distances using the fuzzy framework with an ensemble-based approach is promising and could lead to more accurate results than individual classifiers. Improving the accuracy of individual classifiers using a committee is not a novel idea, but it is expected that the individual members of the ensemble be complementary to each other, which does not seem to be the case here, for the individual techniques achieved similar results. Therefore, the proposed approach using the fuzzy framework can highlight the potential of individual techniques by considering different distance measures.

5.4. Advantages and disadvantages of the proposed method

5.4.1. Advantages

The advantages of the proposed method are listed below:

- The proposed ensemble technique is non-trainable and non-tunable, hence no extra effort or training data is required to get the optimal solution.
- Diversified distance metrics are considered to design the fuzzy distance based ensemble technique, thereby producing a robust decision making process.
- Deep learning architectures are often considered discriminative [8]. Thus the feature representation of three diverse architectures are considered as base learners. Such diversity allows the ensemble model to learn differently from the inputs, thereby helping to enhance the performance of the overall model.

5.4.2. Disadvantages

To have an unbiased view, we present the disadvantages of the proposed method below.

[4] We did not consider the work by Haryanto et al. [39], Win et al. [41] for they did not employ the 5-fold cross validation protocol.

- The proposed ensemble method is static in nature, therefore it may not be effective if the feature representation is biased towards any specific class.
- For an equal distribution (all classes with same probabilities), the proposed method would be erroneous. Since cosine distance measures an angle between two vectors, if they are linear concerning the origin, we have an angle with a null aperture. Therefore, the cosine distance will be zero in such cases.
- Deep learners are required to be very carefully trained using the right set of hyper-parameters, failing which the model is likely to overfit. Hence, just like any other ensemble technique the proposed ensemble method may not give desired results.

6. Conclusions and future work

Cervical cancer side-effects are very concerning. Medical professionals have made some significant efforts to deal with such a disease. However, with the overgrowing population, it is essential to explore CAD techniques to reduce the chances of human errors. We propose an ensemble approach for cervical cancer on PaP smear images. We consider three deep learning models, i.e., Inception V3, MobileNet V2, and Inception ResNet V2, with added extra layers to learn data-specific features. To aggregate the confidence scores obtained by the base learners, we present a novel non-tunable and non-trainable ensemble method able to learn better aggregation of confidence scores generated by the base learners.

The prime reason for using the fuzzy ensemble approach is to measure the importance of the confidence scores produced by the classification models into multiple distance spaces. The final prediction aims to minimize the distance values (i.e., the loss values) calculated in various distance spaces, thereby making the decision-making process more robust and reliable.

The proposed approach provides an end-to-end classification methodology for cervical cancer detection from PaP smear images, which may be employed for clinical use as per requirement. The proposed approach can be used to aggregate the outcomes of other deep learning models in varied domains. Also, optimization algorithms can be used for weighted defuzzification that would eventually incur additional training costs but can still be explored for better performance.

Since raw images are fed to deep learning models, further works can be focused on introducing attention mechanisms to highlight sensitive areas. We may also try to include other deep models to compose the ensemble, as well as to use other fuzzy-based functions to combine their outcomes.

Statements of ethical approval

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

Declaration of Competing Interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome

Acknowledgement

We are thankful to the Center for Microprocessor Applications for Training Education and Research (CMATER) research laboratory of the Department of Computer Science and Engineering, Jadavpur University, Kolkata, India, for providing infrastructural support to this project. The authors are also grateful to FAPESP grants #2013/07375-0, #2014/12236-1, and #2019/07665-4, and CNPq grants #307066/2017-7 and #427968/2018-6.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 68 (6) (2018) 394–424.
[2] M. Arbyn, E. Weiderpass, L. Bruni, S. de Sanjosé, M. Saraiya, J. Ferlay, F. Bray, Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis, Lancet Global Health 8 (2) (2020) e191–e203.
[3] T.A. Kessler, Cervical cancer: prevention and early detection, Semin. Oncol. Nurs. 33 (2) (2017) 172–183.
[4] R. Lozano, Comparison of computer-assisted and manual screening of cervical cytology, Gynecol. Oncol. 104 (1) (2007) 134–138.
[5] M.M. Ali, K. Ahmed, F.M. Bui, B.K. Paul, S.M. Ibrahim, J.M. Quinn, M.A. Moni, Machine learning-based statistical analysis for early stage detection of cervical cancer, Comput. Biol. Med. 139 (2021) 104985.
[6] M. Kaushik, R.C. Joshi, A.S. Kushwah, M.K. Gupta, M. Banerjee, R. Burget, M.K. Dutta, Cytokine gene variants and socio-demographic characteristics as predictors of cervical cancer: a machine learning approach, Comput. Biol. Med. 134 (2021) 104559.
[7] M.M. Rahaman, C. Li, Y. Yao, F. Kulwa, X. Wu, X. Li, Q. Wang, DeepCervix: a deep learning-based framework for the classification of cervical cells using hybrid deep feature fusion techniques, Comput. Biol. Med. 136 (2021) 104649.
[8] I. Arel, D.C. Rose, T.P. Karnowski, Deep machine learning - a new frontier in artificial intelligence research [research frontier], IEEE Comput. Intell. Mag. 5 (4) (2010) 13–18.
[9] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[10] A. Paul, R. Pramanik, S. Malakar, R. Sarkar, An ensemble of deep transfer learning models for handwritten music symbol recognition, Neural Comput. Appl. (2021), doi:10.1007/s00521-021-06629-9.
[11] A. Banerjee, P.K. Singh, R. Sarkar, Fuzzy integral-based CNN classifier fusion for 3D skeleton action recognition, IEEE Trans. Circuits Syst. Video Technol. 31 (6) (2021) 2206–2216, doi:10.1109/TCSVT.2020.3019293.
[12] N. Chakraborty, S. Kundu, S. Paul, A.F. Mollah, S. Basu, R. Sarkar, Language identification from multi-lingual scene text images: a CNN based classifier ensemble approach, J. Ambient Intell. Humanized Comput. 12 (7) (2021) 7997–8008.
[13] S. Tulyakov, S. Jaeger, V. Govindaraju, D. Doermann, Review of classifier combination methods, Mach. Learn. Doc. Anal. Recognit. (2008) 361–386.
[14] M.E. Plissiti, P. Dimitrakopoulos, G. Sfikas, C. Nikou, O. Krikoni, A. Charchanti, Sipakmed: a new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images, in: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 3144–3148.
[15] Y.R. Park, Y.J. Kim, W. Ju, K. Nam, S. Kim, K.G. Kim, Comparison of machine and deep learning for the classification of cervical cancer based on cervicography images, Sci. Rep. 11 (1) (2021) 1–11.
[16] K. Bora, M. Chowdhury, L.B. Mahanta, M.K. Kundu, A.K. Das, Pap smear image classification using convolutional neural network, in: Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing, 2016, pp. 1–8.
[17] L. Zhang, L. Lu, I. Nogues, R.M. Summers, S. Liu, J. Yao, DeepPap: deep convolutional networks for cervical cell classification, IEEE J. Biomed. Health Inf. 21 (6) (2017) 1633–1643.
[18] L. Nanni, S. Ghidoni, S. Brahnam, Handcrafted vs. non-handcrafted features for computer vision classification, Pattern Recognit. 71 (2017) 158–172.
[19] A.D. Jia, B.Z. Li, C.C. Zhang, Detection of cervical cancer cells based on strong feature CNN-SVM network, Neurocomputing 411 (2020) 112–127.
[20] A. Ghoneim, G. Muhammad, M.S. Hossain, Cervical cancer classification using convolutional neural networks and extreme learning machines, Future Gener. Comput. Syst. 102 (2020) 643–649.
[21] J. Shi, R. Wang, Y. Zheng, Z. Jiang, H. Zhang, L. Yu, Cervical cell classification with graph convolutional network, Comput. Methods Programs Biomed. 198 (2021) 105807.
[22] Q. Li, Z. Han, X.-M. Wu, Deeper insights into graph convolutional networks for semi-supervised learning, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[23] A. Manna, R. Kundu, D. Kaplun, A. Sinitca, R. Sarkar, A fuzzy rank-based ensemble of CNN models for classification of cervical cytology, Sci. Rep. 11 (1) (2021) 1–18.
[24] H. Lin, Y. Hu, S. Chen, J. Yao, L. Zhang, Fine-grained classification of cervical cells using morphological and appearance based convolutional neural networks, IEEE Access 7 (2019) 71541–71549.
[25] R. Elakkiya, K.S.S. Teja, L.J. Deborah, C. Bisogni, C. Medaglia, Imaging based cervical cancer diagnostics using small object detection-generative adversarial networks, Multimedia Tools Appl. (2021) 1–17.
[26] P. Wang, J. Wang, Y. Li, L. Li, H. Zhang, Adaptive pruning of transfer learned deep convolutional neural network for classification of cervical pap smear images, IEEE Access 8 (2020) 50674–50683.
[27] N. Dong, L. Zhao, C.-H. Wu, J.-F. Chang, Inception V3 based cervical cell classification combined with artificially extracted features, Appl. Soft Comput. 93 (2020) 106311.

[28] C. Zhang, D. Jia, N. Wu, Z. Guo, H. Ge, Quantitative detection of cervical cancer based on time series information from smear images, Appl. Soft Comput. 112 (2021) 107791.

[29] L. Cao, J. Yang, Z. Rong, L. Li, B. Xia, C. You, G. Lou, L. Jiang, C. Du, H. Meng, et al., A novel attention-guided convolutional network for the detection of abnormal cervical cells in cervical cancer screening, Med. Image Anal. 73 (2021) 102197.

[30] H. Guo, K. Zheng, X. Fan, H. Yu, S. Wang, Visual attention consistency under image transforms for multi-label image classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 729–739.

[31] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, Proc. IEEE 109 (1) (2021) 43–76, doi:10.1109/JPROC.2020.3004555.

[32] 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 (CVPR), 2016, pp. 2818–2826.

[33] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520.

[34] 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 (CVPR), 2016, pp. 770–778.

[35] 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.

[36] Y. Wu, L. Liu, J. Bae, K.-H. Chow, A. Iyengar, C. Pu, W. Wei, L. Yu, Q. Zhang, Demystifying learning rate policies for high accuracy training of deep neural networks, in: 2019 IEEE International Conference on Big Data (Big Data), IEEE, 2019, pp. 1971–1980.

[37] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, H. Li, Learning to rank: from pairwise approach to listwise approach, in: Proceedings of the 24th International Conference on Machine Learning, 2007, pp. 129–136.

[38] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[39] T. Haryanto, I.S. Sitanggang, M.A. Agmalaro, R. Rulaningtyas, The utilization of padding scheme on convolutional neural network for cervical cell images classification, in: 2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM), IEEE, 2020, pp. 34–38.

[40] O. Yaman, T. Tuncer, Exemplar pyramid deep feature extraction based cervical cancer image classification model using pap-smear images, Biomed. Signal Process. Control 73 (2022) 103428.

[41] K.P. Win, Y. Kitjaidure, K. Hamamoto, T. Myo Aung, Computer-assisted screening for cervical cancer using digital image processing of pap smear images, Appl. Sci. 10 (5) (2020) 1800.

[42] H. Basak, R. Kundu, S. Chakraborty, N. Das, Cervical cytology classification using PCA and GWO enhanced deep features selection, SN Comput. Sci. 2 (5) (2021) 1–17.