Medical image classification using a combination of features from convolutional neural networks

Rocha, Marina M. M.; Landini, Gabriel; Florindo, Joao B.

DOI:
10.1007/s11042-022-14206-y

Medical image classification using a combination of features from convolutional neural networks

Marina M. M. Rocha[a], Gabriel Landini[b], Joao B. Florindo[a,*]

[a] *Institute of Mathematics, Statistics and Scientific Computing - University of Campinas Rua Sérgio Buarque de Holanda, 651, Cidade Universitária "Zeferino Vaz" - Distr. Barão Geraldo, CEP 13083-859, Campinas, SP, Brasil*
[b] *Oral Pathology Unit, School of Dentistry, University of Birmingham, U.K.*

Abstract

Medical image classification is an important and challenging problem, since images are usually complex, variable and the amount of data is relatively constrained. Selecting optimal sets of features and classifiers is a crucial problem in this area. In this paper it is proposed an image classification method, named Hybrid CNN Ensemble (HCNNE), based on the combination of image features extracted by convolutional neural networks (CNN) and local binary patterns (LBP). The features are subsequently used to build an ensemble of multiple classifiers. More specifically, the Euclidean distance between LBP feature vectors of each training class and the confidence of CNN features classified by support vector machines are employed to compose the input of a multilayer perceptron classifier. Finally, these features are also used as input to other classifiers to compose the final voting ensemble. This approach achieved an accuracy similar to those of other state-of-the-art methods in texture classification and showed an improvement of 10% over the previously reported identification of a group of odontogenic oral cyst histological images, at a low computational cost. Three major contributions are presented here: 1) the combination of low and high level features assigning weights based on the confidence of the features for texture recognition; 2) the combination of automatically learned deep features with LBP by a multilayer perceptron based on the feature confidences; 3) state-of-the-art results are obtained in the odontogenic cyst categorization problem.

*Corresponding author.
Email addresses: m183983@dac.unicamp.br (Marina M. M. Rocha), G.Landini@bham.ac.uk (Gabriel Landini), florindo@unicamp.br (Joao B. Florindo)

Keywords: Classifiers ensemble, Texture recognition, Deep convolutional networks, Odontogenic cysts.

1. Introduction

Neural networks are powerful computational tools for performing tasks that would otherwise require human discernment and reasoning. Such algorithmic approaches are presently used in many aspects of modern society, from social media content recommendation [1], medical diagnosis [2] and DNA-RNA sequences predictions [3] to name a few. A common tasks entrusted to neural networks is that of image classification. Various new approaches and variations emerge every year, aiming at the design of methods with better accuracy and generalization, or less computational cost.

In this context, deep convolutional networks have shown outstanding performance in certain problems of image classification. However, dealing with complex, heterogeneous and small datasets (common in the medical domain) still poses an important challenge. An additional relevant problem is choosing the optimal set of approaches for a particular application, among the vast diversity of existing classification methods.

There are several possible ways of extracting features from images and those have an important impact on the overall performance of the classifier algorithms. Some features can be obtained from simple rules, for example the Histograms of Oriented Gradients (HOG) [4] (often called 'low-level' features) while convolutional neural networks (CNN) can be considered as 'high-level' feature extraction methods because of the computational complexity involved.

Choosing the optimal features can be decisive in the success rate of a classification method. One promising strategy is to combine different types of features, as recently explored by Forcén *et al.* 2019 [5]. The research question in this study is, in this way, whether such combination of low and high level features could be beneficial in texture image recognition, especially when small amount of data is available for training, as usual in medicine. It is also investigated how different classifiers can be combined by an ensemble to improve the performance of the individual classifiers. The Hybrid CNN Ensemble (HCNNE) model proposed in this paper is a new approach to feature combination that uses deep convolutional networks to extract features and simple multi-layer perceptron network to find the best feature combination. This work also approaches the challenges of classification tasks mentioned earlier, by implementing an ensemble of classifiers, which has showed positives results [6].

The texture datasets UIUC [7] and UMD [8] were used here to test the performance of the computational method proposed here. Since those datasets are widely used in classification research, the results obtained can be easily compared with those found elsewhere in the literature (the method achieved a classification performance comparable with other state-of-the-art methods). The methodology was also applied to the practical problem of supervised histological image classification of two types of odontogenic cysts of the jaws, a diagnostic task which is routinely done by specialists.

Contributions and novelties of this work may be summarized as follows:

1. This is the first time that such ensemble scheme based on a multi-layer perceptron is associated with convolutional neural features;
2. state-of-the-art results were obtained on the cysts database and the results were competitive with the state-of-the-art on benchmark databases;
3. the combination of low and high level features, differing from usual protocols in deep learning by the use of feature confidence to feed a multilayer perceptron classifier as part of the ensemble.

This work presents in the following section an overview of materials and papers related. The following section contains some theoretical foundation and more information about the methods applied in this study. The “Experiments” section approaches the details of the HCNNE model formulation, implementation and performance tests used, including a brief introduction to the cyst classification problem. The last two sections are dedicated to results analysis and conclusions.

2. Related works

Deep learning techniques have played fundamental role in image recognition in recent years. This subject has been thoroughly approached, for example, in Goodfellow *et al.* 2016 [9]. A further advance was the introduction of residual networks (ResNet) in Kaiming He *et al.* 2016 [10], with the use of transfer learning for pre-training networks in larger databases, such as ImageNet [11]. In Kumar *et al.* 2016 [12] those methods were applied to medical images, with a new approach that included fine-tuning. Recent works on this research field have also shown outstanding results by combining machine learning methods with other approaches, such as in Bacanin *et al.* 2022 [13] and Malakar *et al.* 2020 [14].

Another related strategy that has attracted interest in image recognition is the use of features extracted by convolutional neural networks (CNN) [15] as input to multiple classifiers. The use of CNNs and feature combinations have been exploited in recent studies, with promising results, such as in Ragab *et al.* 2020 [16], Attallah *et al.* 2020 [17, 18] and Anwar *et al.* 2020 [19]. Examples of classifiers that have been used are the Random Forest (RF) [20], Support Vector Machine (SVM) [21], Linear Discriminant Analysis (LDA) [22] and k-Nearest Neighbor (KNN) clustering [23].

Each of those classifiers has its own benefits and this naturally prompts the question of which classifier would be best to use. The informally called "No Free Lunch Theorem" [24] is an optimization statement often used in machine learning. In this context, it hypothesizes that it is not possible to have one single classifier that outperforms every other approach, no matter the task. The theorem and its implications in machine learning were discussed in Ho *et al.* 2002 [25].

To address this issue, a promising approach is to combine every classification through an ensemble of classifiers. This strategy has been used in previous works by Rokach *et al.* 2010 [26] and Ye Ren *et al.* 2016 [6]. Recent studies in medical applications also attested the potential of ensembling approaches, as in Ragab *et al.* 2019 [27], Attallah *et al.* 2022 [28] and Fouad *et al.* 2017 [29].

As mentioned earlier, the HCNNE model proposes a combination of image features extracted by different approaches. The 'low-level' feature extraction method chosen here was the local binary patterns (LBP) [30]. This has been extensively studied in previous texture classification works, such as Ojala *et al.* 2002 [30], Zhenhua Guo *et al.* 2010 [31] and Li Liu *et al.* 2017 [32].

The combination of image features was inspired by the work of Forcén *et al.* 2019 [5], applied to classification problems. This work differs from that of Forcén in using convolutional neural networks (CNN) as the 'high-level' feature extractor and adds a classifier ensemble to the procedure. This strategy of combining 'high' and 'low-level' features is also investigated in Attallah *et al.* 2020 [33].

The method presented is applied to a supervised classification problem of two types of jaw cysts from histological images. This task poses a significant challenge to computational algorithms, and otherwise would require the careful analysis of expert histopathologists. The use of deep learning in medical images and its challenges has been discussed in Litjens *et al.* 2017 [2]. The specific problem of odontogenic cysts classification was previously investigated by Landini 2006 [34] and Florindo *et al.* 2017 [35]. The results achieved in those papers are used here as benchmarks for comparison with the methodology presented.

3. Materials and methods

3.1. Feature extraction

Convolutional neural networks (CNN) are composed of convolutional layers, activation layers and pooling layers. The first layer is based on the convolution operation in the discrete and two dimensional domain [9]:

$$S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n),$$

where K is the network kernel, I is the input image, S is the obtained feature map and $*$ is the convolutional operator.

A way of visualizing this is sliding a matrix of small size (the kernel) along a matrix of larger size (input image) and an operation which computes a weighted sum in every possible position of the matrix, resulting in a new matrix (the feature map).

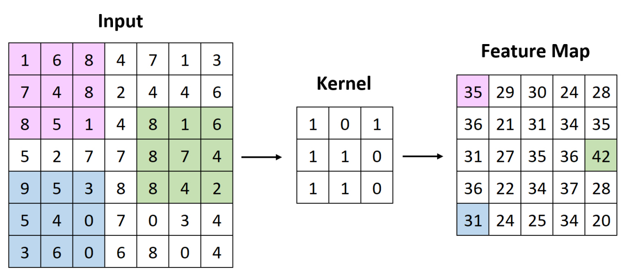


Figure 1: Scheme of convolutional application.

This operation carries important properties, one of those is *parameter sharing*, meaning that the same kernel is used along the entire image, significantly decreasing the number of parameters that need to be tracked and optimized. That does not only reduce computational burden, but also means that every learned feature (such as lines or edges) can be found anywhere in the picture.

After the convolutional layer, an activation function is applied. The most common one (and used here) is the Rectified Linear Unit (ReLU) that assumes the value zero for negative inputs, and the input value itself for positive ones.

A pooling layer then reduces the dimension of the resulting matrix by merging a pixel with its neighbors according to a given function. One of such functions is the max pooling operation, which returns the highest value in a pixel neighborhood.

Convolutional neural networks (CNN) have the property of detecting features with increasing complexity along its layers. For example, in a network

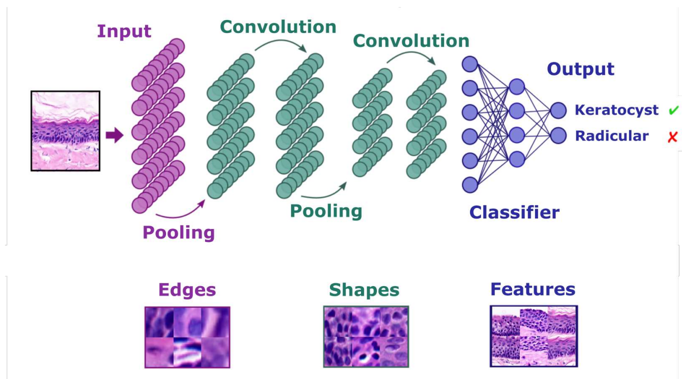


Figure 2: Scheme of convolutional neural network.

trained for identification of human faces, the first layers might detect lines and edges, while the last ones might be able to detect more complex structures, such as eyes. This property is what justifies the extraction of feature vectors from the last layers [36], since the aim is to work with features that carry the highest level information about the images. Those features were used here as inputs for the classifiers. Figure 3 shows examples of feature maps extracted from cysts samples by some of ResNet's convolutional layers.

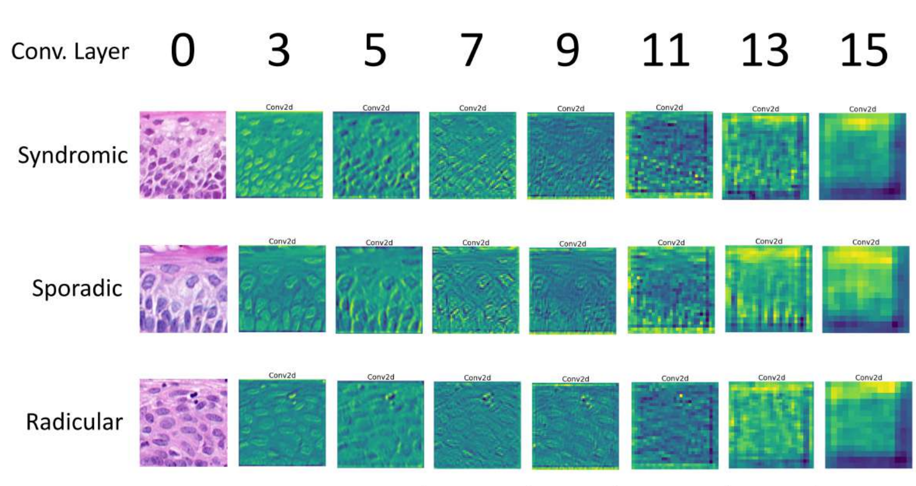


Figure 3: Examples of feature maps from some of the convolutional layers.

3.2. Classification

Classifiers are used in order to predict the class of given data point. In this work, the following methods were used:

- **Fully Connected Layer (FCL):** simple artificial neural network with one hidden layer of size $4n$, where n is the number of classes.

- **Support Vector Machine (SVM):** the algorithm tries to find a hyperplane that best separates the given data points [37]. Suppose a binary classification problem with dataset (x_i, y_i), where $i = 1, \dots, n$. In this case, the SVM model consists in solving the following optimization problem:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j K(x_i, x_j) \alpha_i \alpha_j, \\ \text{subject to} \quad & 0 \leq \alpha_i \leq \beta, \quad i = 1, \dots, n, \\ & \sum_{i=1}^{n} y_i \alpha_i = 0, \end{aligned}$$

  where β is a hyperparameter, K is the kernel function and α_i are Lagrange multipliers.

- **k-Nearest Neighbors (KNN):** consists in verifying the class of all data points within the neighborhood of a given data point, and then classifying this data point according to the most frequent class among its neighbors. [23].

- **Random Forest (RF):** combination of multiple decision-trees with subsampling strategies [20].

- **Linear Discriminant Analysis (LDA):** based on searching for a linear combination of parameters that best separates the classes. This is a statistic method that uses the concepts of expectation and covariance, given by the following equations:

$$\overrightarrow{w} \cdot \overrightarrow{x} > c$$

$$\overrightarrow{w} = \Sigma^{-1}(\overrightarrow{\mu_1} - \overrightarrow{\mu_0})$$

$$c = \frac{1}{2}(T - \overrightarrow{\mu_0}^T \Sigma^{-1} \overrightarrow{\mu_0} + \overrightarrow{\mu_1}^T \Sigma^{-1} \overrightarrow{\mu_1}),$$

where $\overrightarrow{x}$ is the vector representing the data point, $\overrightarrow{\mu}$ is the expectation vector, Σ is the covariance matrix and T is a predefined threshold [22].

3.3. Feature combination

Forcen *et al.* 2019 [5] exploited a different strategy, the premise being that low level feature vectors combined with high level features can increase the classification accuracy. [1]

Local Binary Patterns (LBP) is a simple yet efficient method for texture recognition. It consists of thresholding the image pixels in order to label each pixel as a binary value. This method was chosen as the low level feature extractor because of its simplicity. The features extracted by LBP are more likely to carry simple traits and patterns, which are important in texture recognition tasks. On the other hand, CNNs are known to generate highly complex features extracted from the last network layers and they were used as the high level method.

After applying LBP, every image is represented as a feature vector and those vectors were used to calculate the Euclidean distances between each testing and training images. Given a test image, the distance to every class was defined as the average distance between the five nearest images from each class, i.e. for every test image, there is a distance vector associated

[1] Here 'level' means the complexity of the feature, i.e. how much information about the image it holds.

$$\overrightarrow{d} = [d_1, d_2, \ldots, d_m],$$

where m is the number of classes.

The feature vectors extracted with the CNN were submitted to a SVM classifier, from which a confidence vector was obtained. This vector contains the confidence of the classifier that a particular image belongs to each class:

$$\overrightarrow{r} = [r_1, r_2, \ldots, r_m]$$

This confidence vector $\overrightarrow{r}$ is obtained by using Platt scaling [38], optimizing parameters A and B such that:

$$P(y_i|x_j) = r_i = \frac{1}{1 + \exp(A * \Phi(x_j) + B)},$$

where $\Phi(x_j)$ returns the distance of the sample x_j to the hyperplane optimized by the SVM method [39].

In order to combine both features into the same classifier, the low level features are used to calculate Euclidean distances between a test image and the training images from each class in some neighborhood. This procedure is similar to the KNN algorithm and it generates a distance vector $\overrightarrow{d}$ associated to each test image.

On the other hand, the high level features were used as input to an SVM classifier, resulting in score vectors $\overrightarrow{r}$ associated to each test image. Note that the score vector represents how confident the classifier is that each test image belongs to some class.

In summary, a vector $\overrightarrow{d}$ obtained from low level features and a vector $\overrightarrow{r}$ associated with high level features are used. To combine those two vectors, $\overrightarrow{d}$ and $\overrightarrow{r}$, three different approaches were employed, resulting in three new classifiers, that were named "Feature Combination" (FC):

- **SVM + LBP (FC1):** this combination was accomplished using the distances to calculate weights (w_i) for the score vectors ($\overrightarrow{r}$):

$$w_i = \left(\frac{1}{m} \sum_{\substack{j=1\ldots m \\ j \neq i}} d_j \right) / d_i$$

$$\overrightarrow{r_w} = [w_1 r_1, w_2 r_2, \ldots, w_m r_m]$$

- **SVM + LBP + FCL (FC2):** the score vector ($\overrightarrow{r}$) is concatenated with the distance vector ($\overrightarrow{d}$) and used as input for training a fully connected neural network with one hidden layer. This network is responsible for determining the best combination of scores and distances. The output is the classification of the image class. The class i in this case is assigned by solving the optimization problem

$$\text{argmin}_i C(W_i \sigma(\sum V_{jk} rd_k)),$$

where C is the cross-entropy loss, W_i are weights connecting the i^{th} output to the hidden layer, σ is the sigmoid function, V is th matrix of weights connecting the input and hidden layer, and rd is the vector resulting from the concatenation of r and d.

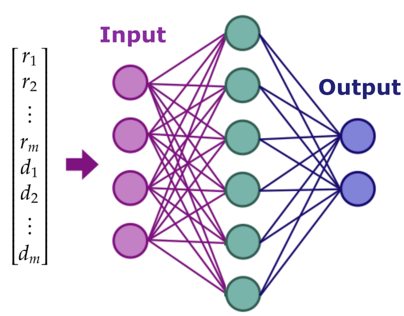


Figure 4: Visualization of SVM + LBP + FCL (FC2) method.

- **SVM + LDA + LBP + FCL (FC3):** similar to the previous classifier, but concatenating the SVM score vector with the LDA score vector and the distance vector.

4. Proposed Method

The Hybrid CNN Ensemble proposes classifying texture images using combinations of features extracted from convolutional neural networks (CNN) and local binary patterns (LBP), then applying those features to multiple classifiers and using an ensemble strategy to make the final classifications. The rationale behind the choice of LBP to provide the low level features here is twofold. First it is a straightforward and easy to interpret descriptor successfully used in texture recognition for several decades. Second it is well known for its high computational efficiency, being the faster algorithm among the most popular texture local descriptors. This model can be summarized into three steps:

- **Feature extraction:** the CNN extracts feature vectors from the image samples;
- **Classification:** each individual classifier is trained to predict the test samples classes (the classifiers are divided into standard ones and those that use the feature combination strategy);
- **Ensembling:** combines the decision of every individual classifier through simple voting.

These steps are represented in Figure 5.

4.1. Ensemble

Ensemble learning is a strategy widely used in machine learning. It consists of combining multiple classification methods in order to achieve better performance. The generalization and averaging obtained by ensembles can also be a workaround to problems like choosing the best classifier or getting stuck in local optimal minima. Using multiple different strategies simultaneously might also compensate for weaknesses and individual anomalies of each classifier.

In this study, each of the eight classifiers implemented (FCL, SVM, KNN, RF, LDA, FC1, FC2 and FC3) results in predictions for the classes of the test

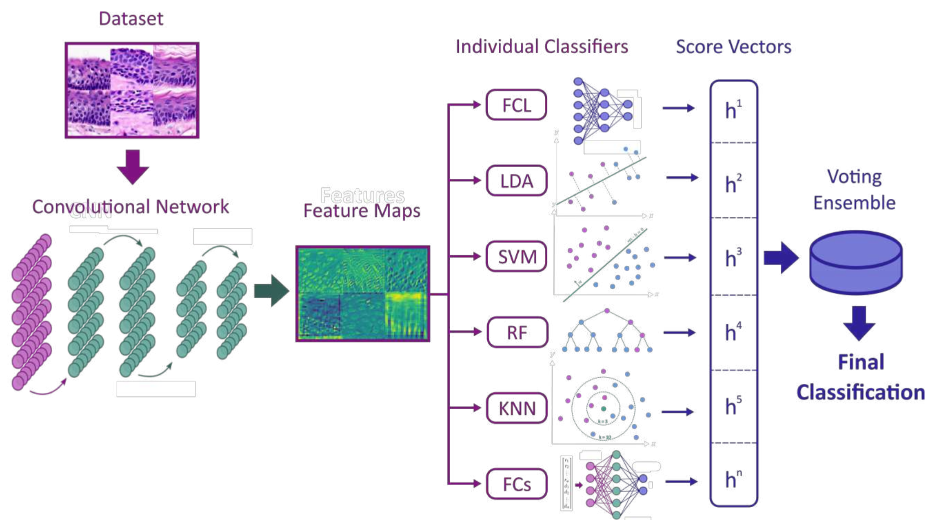


Figure 5: Scheme representing the entirety of the model.

images. To decide the final classification, an ensemble [26, 6] was implemented, where each classifier prediction represents a vote.

To increase accuracy, the worst two classifiers were removed from the ensemble. This is done by using a validation set to compare the accuracy h_i of each individual classifier $i = 1, \ldots, m$ after training is over, where m is the total number of classifiers. The remaining trained classifiers are then applied to the test dataset, returning a predicted class $y_i^j \in \{1, \ldots, n\}$ for each test image sample j, where n is the number of classes.

The ensembling strategy proposed here consists in choosing the most frequent class among the predictions from the individual classifiers, y_i^j for $i = 1, \ldots, (m-2)$. In other words, the class with the most "votes" corresponds to the final classification. When it comes to a draw, the classifiers with the next worst performance is removed, one by one, until there is no longer a tie.

This approach has many attractive features, such as not requiring a large database for training, nor high performance computing. Also, combining classifiers to improve accuracy is convenient because avoids the need of inserting new training data to the network.

4.2. Parameters settings

The databases used in this work were divided equally and randomly between test and train subsets. Regarding the architecture of the network, the chosen approaches were the pre-trained models ResNet [10] and AlexNet. The strategies used for computing the network parameters were as follows:

- **Fine-tuning:** after the network is pre-trained on ImageNet [11], all the parameters are optimized in the current database.

- **Fixed feature extraction:** the network is also pre-trained on ImageNet [11], but only parameters of the final layer are optimized in the current database.

Regarding the other parameters of the network, such as kernel size and strides, were used the standard values for ResNet and AlexNet, pre-trained on ImageNet. The results shown in this section were obtained by using 15 epochs, step size 7 and learning rate 0.0001.

5. Experiments

The accuracy is defined as the percentage of images classified correctly. Every configuration of the network was ran ten times, then the average accuracy and the standard deviation were calculated.

In medical applications, it is also important to measure the performance per case, or case-wise accuracy. This means that after the images have been labeled by the HCNNE, a simple ensemble by voting was used to decide the classification for each case. Therefore, case-wise accuracy is defined as the percentage of cases classified correctly.

The precision-recall curve was also used in order to calculate the average precision-recall score. This method gives a more reliable way of measuring the network performance. Precision, recall and F1-score values were also computed for the same purpose, according to the respective formulations.

$$\text{Precision} = \frac{TP}{TP + FP},$$

$$\text{Recall} = \frac{TP}{TP + FN},$$

$$\text{F1-score} = \frac{2(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}},$$

where TP, FP and FN represent, respectively, the total amount of samples classified as true positives, false positives and false negatives.

Other parameters observed were the average computing time for each run of the code and the average confusion matrix. The algorithm was implemented in Python 3.7 with the libraries Pytorch and Scikitlearn, the code is available on GitHub [2].

5.1. Cysts dataset

The dataset consisted of histological images of oral cysts, stained with haematoxylin and eosin). Cysts are pathological fluid-filled (sac-like) lesions, lined by epithelium. There are several types of cyst; here the interest was to investigate the potential for supervised classification of odontogenic keratocysts (OKC) *versus* the more common radicular cysts. OKCs can occur on their own (called 'sporadic OKCs') or as part of the Gorlin-Golz or Basal Cell Naevus syndrome (here referred to as 'syndromic' OKCs) [34]. The morphological differences between the two sub-types of OKC have been questioned, mostly because they are difficult to assess visually, and therefore pose an interesting diagnostic problem. While radicular cysts are inflammatory in origin (associated with the roots of non-vital teeth, mostly as consequences of untreated dental caries), OKCs show active growth and higher recurrence rates and this has raised long-standing arguments of whether OKCs should be considered benign cystic neoplasms rather than cysts. The database contained 65 images of

[2]https://github.com/MarinaRocha29/Hybrid-CNN-Ensemble

sporadic OKCs (denoted by **k**) from 13 cases, 40 images of syndromic from 8 cases (denoted by **s**) and 45 images of radicular cysts (denoted by **r**) from 9 cases (i.e. five images from each case, originally captured with a $\times 40$ objective and resized to 227×227 pixels).

Given the medical application of the problem, there are some classifications that can be of diagnostic relevance [40], [35]:

- $\mathbf{k} \times \mathbf{s} \times \mathbf{r}$: distinguishing between all of the classes.
- $\mathbf{ks} \times \mathbf{r}$: distinguishing between OKCs and radicular cysts.
- $\mathbf{k} \times \mathbf{s}$: classifying the sub-types of OKCs, sporadic and syndromic.

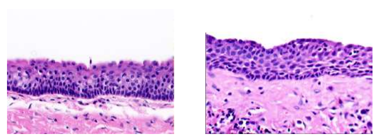


Figure 6: A case of sporadic odontogenic keratocyst (left) and a radicular cyst (right).

5.2. UMD and UIUC datasets

Given the popularity of benchmark databases of texture images, the performance of HCNNE on UIUC [7] and UMD [8] were obtained for comparison purposes. Those are datasets widely used in multiple state-of-the-art works on texture classification.

The UIUC and UMD datasets both contain 25 classes with 40 images each, resulting in 1000 images for each dataset. Every class has photographs of textures such as bark, marble and fur (for the UIUC dataset), wood floor, tile floor and flowers (for the UMD dataset) as shown in Fig. 7 and 8. The images are sized 1280×960 (UMD) and 640×480 (UIUC) pixels, with scale, rotation and viewpoint changes within each class.

6. Results and discussion

Figure 9 shows the average accuracy and standard deviation for each network configuration. The bars represent the accuracy for each individual classifier and

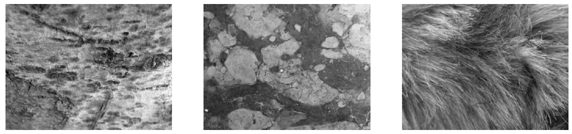


Figure 7: Examples of images from the UIUC texture dataset.

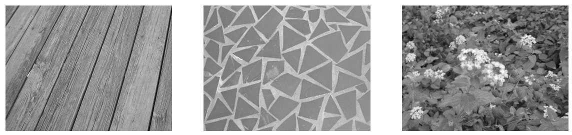


Figure 8: Examples of images from the UMD texture dataset.

for the ensemble approach. The best accuracy was achieved, in both cases, by the fine-tuned ResNet with the ensemble approach.

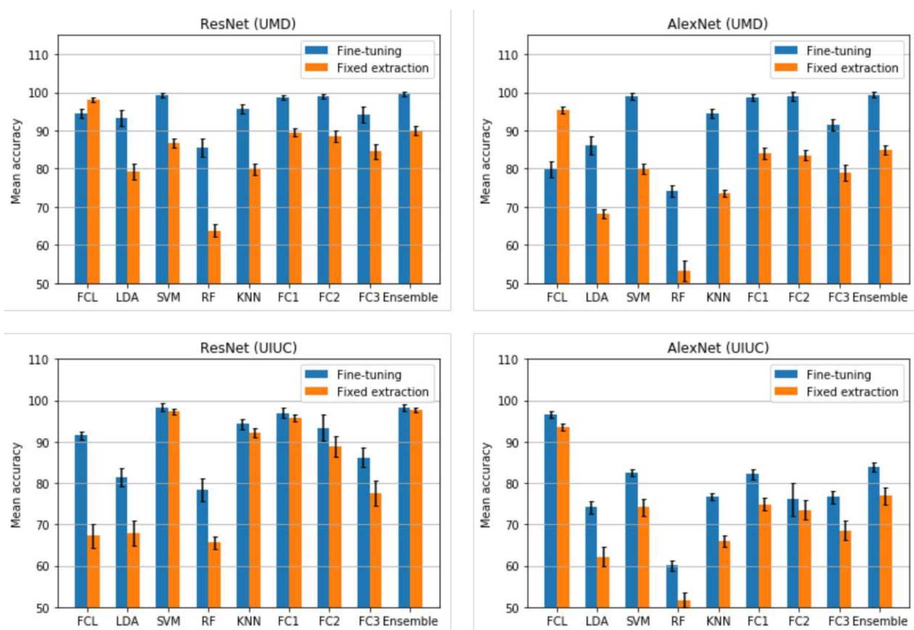


Figure 9: Average accuracy on UIUC and UMD texture databases.

Tables 1 and 2 show the elapsed time of the feature extraction and each individual classifier training procedures, followed by the total amount of time spent running the model. These results were obtained with an Intel(R) Core(TM) i7-8550U CPU 1.80GHz, 1992 Mhz, 4 Cores, 8 Logical Processors, 16GB RAM, GPU NVIDIA GeForce MX150 on Windows 10 version 1909, with Python 3.7.

Table 1: Average computing time in UIUC database.

hh:mm:ss	ResNet		AlexNet	
Structure	Fine-tuning	Fixed	Fine-tuning	Fixed
CNN (feature extraction)	00:28:55	00:21:34	00:27:37	00:19:10
FCL	00:01:08	00:00:51	00:01:21	00:00:47
LDA	00:00:01	00:00:01	00:00:01	00:00:01
SVM	00:00:12	00:00:09	00:00:14	00:00:07
RF	00:00:01	00:00:01	00:00:01	00:00:01
KNN	00:00:01	00:00:01	00:00:01	00:00:01
FC1	00:01:58	00:01:33	00:01:26	00:00:58
FC2	00:01:09	00:01:05	00:01:01	00:00:36
FC3	00:01:10	00:00:51	00:00:59	00:00:33
Total	00:34:52	00:26:09	00:32:42	00:23:04

Table 2: Average computing time in UMD database.

hh:mm:ss	ResNet		AlexNet	
Structure	Fine-tuning	Fixed	Fine-tuning	Fixed
CNN (feature extraction)	00:56:42	00:47:39	00:49:19	00:39:56
FCL	00:03:08	00:02:47	00:03:00	00:02:36
LDA	00:00:01	00:00:01	00:00:01	00:00:01
SVM	00:00:28	00:00:22	00:00:28	00:00:18
RF	00:00:01	00:00:01	00:00:01	00:00:01
KNN	00:00:01	00:00:01	00:00:01	00:00:01
FC1	00:03:42	00:03:09	00:03:12	00:02:50
FC2	00:02:50	00:02:32	00:02:09	00:02:02
FC3	00:02:39	00:01:55	00:01:46	00:01:27
Total	01:08:16	00:58:10	01:01:23	00:50:13

Another way of measuring the performance of the network is using reliability measures such as precision-recall curves and confusion matrices. To illustrate the HCNNE performance, Figure 10 shows examples of confusion matrices using AlexNet with fine-tuning. In Tables 3 and 4 is shown the average precision, recall and F1-score values for each network architecture. The values obtained also attest the efficiency of the strategy proposed in this work.

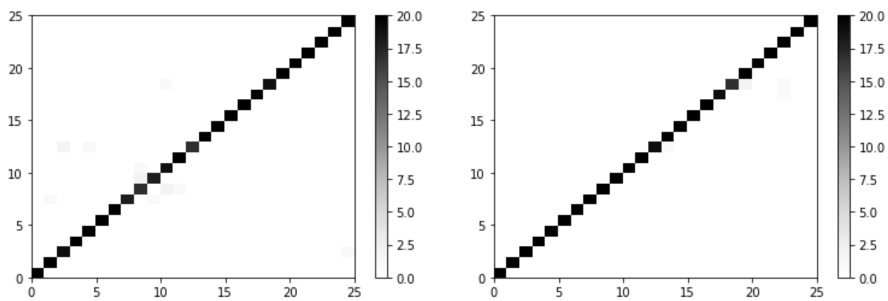


Figure 10: UIUC (left) and UMD (right) confusion matrices.

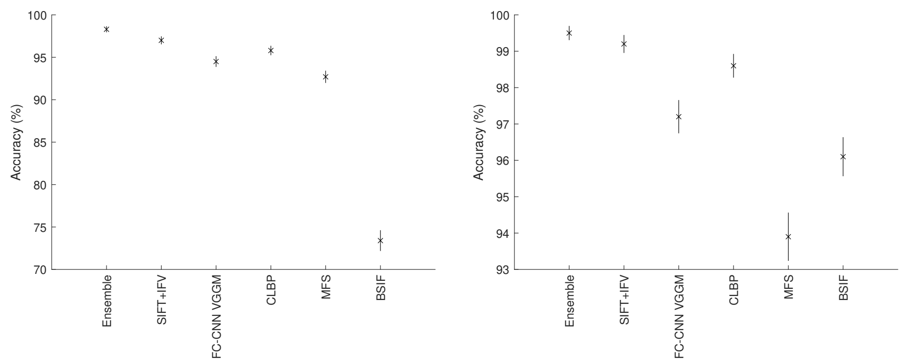
Table 3: UIUC dataset HCNNE performance scores.

	ResNet		AlexNet	
Structure	Fine-tuning	Fixed	Fine-tuning	Fixed
Precision	0.99	0.96	0.85	0.76
Recall	0.98	0.98	0.84	0.77
F1-score	0.99	0.98	0.84	0.76

Table 4: UMD dataset HCNNE performance metrics.

	ResNet		AlexNet	
Structure	Fine-tuning	Fixed	Fine-tuning	Fixed
Precision	1.00	1.00	0.92	0.86
Recall	1.00	1.00	0.91	0.84
F1-score	1.00	1.00	0.91	0.83

When it comes to classification problems, it is important to compare results



UIUC UMD

Figure 11: Confidence intervals for different methods on UIUC and UMD databases.

with those of previous works applied to the same databases. Table 5 shows such comparisons. The ensemble accuracy is similar to the best results obtained by several state-of-the-art methods. Figure 11 shows the confidence interval for some illustrative approaches presented in Table 5. The approach of confidence interval for a proportion is used, where the event "correct classification" is modeled as a binomial distribution. That figure confirms the competitiveness of the proposed method in such benchmark applications.

Figure 12 shows the average accuracy and standard deviation for each configuration of the network in the classification of the oral cysts investigated. Clearly, the problem of identifying OKCs from radicular cysts (ks $\times$ r) is the easiest one, while distinguishing between the two OKCs sub-types (k $\times$ s) is the most difficult one.

Similar results were found in the case-wise approach, as shown in Figure 13. The ensemble has outperformed every individual classifier in all the experiments. Classifiers using feature combination consistently reached higher accuracy values.

Tables 6 and 7 report the time spent on one run of the code. Considering the challenge of the task and the relatively modest hardware used in this experiment, these times are competitive and suggest the suitability of using the proposed algorithm in real world situations. Tables 8 to 11 list the average precision, recall

Table 5: Comparison with state of the art methods.

Method	UIUC (%)	UMD (%)
MFS [41]	92.7	93.9
LBP [30]	88.4	96.2
BSIF [42]	73.4	96.1
CLBP [31]	95.8	98.6
PLS [43]	96.6	99.0
SIFT + LLC [44]	96.3	98.4
SIFT + VLAD [44]	96.5	99.3
ScatNet [45]	88.6	93.4
SIFT + IFV [44]	97.0	99.2
FV-CNN AlexNet [46]	99.2	99.7
FC-CNN VGGM [46]	94.5	97.2
OTF [47]	98.1	98.8
WMFS [41]	98.6	98.7
BIFs SRC [48]	99.0	99.5
FC-CNN + FV-CNN AlexNet [46]	99.3	99.7
FV-CNN VGGM [46]	99.6	99.9
FC-CNN + FV-CNN VGGM [46]	99.6	99.8
FV-CNN VGGVD [46]	99.9	99.9
Ensemble	98.3	99.5

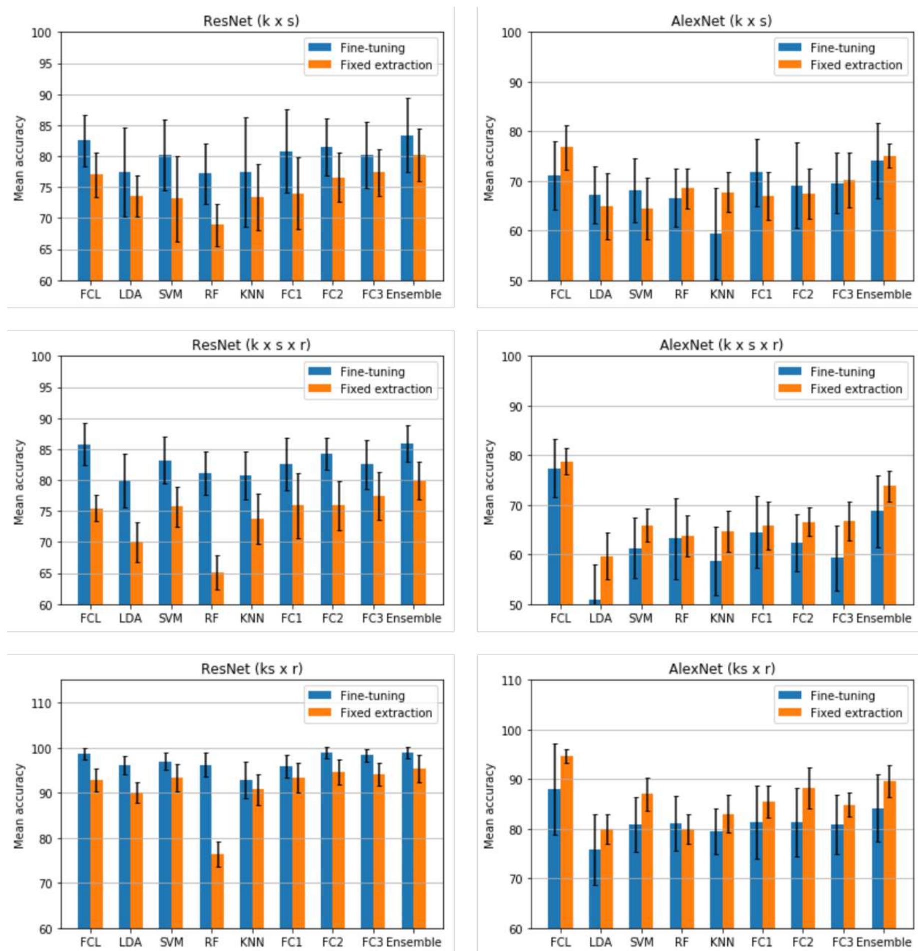


Figure 12: Average accuracy on the cysts database.

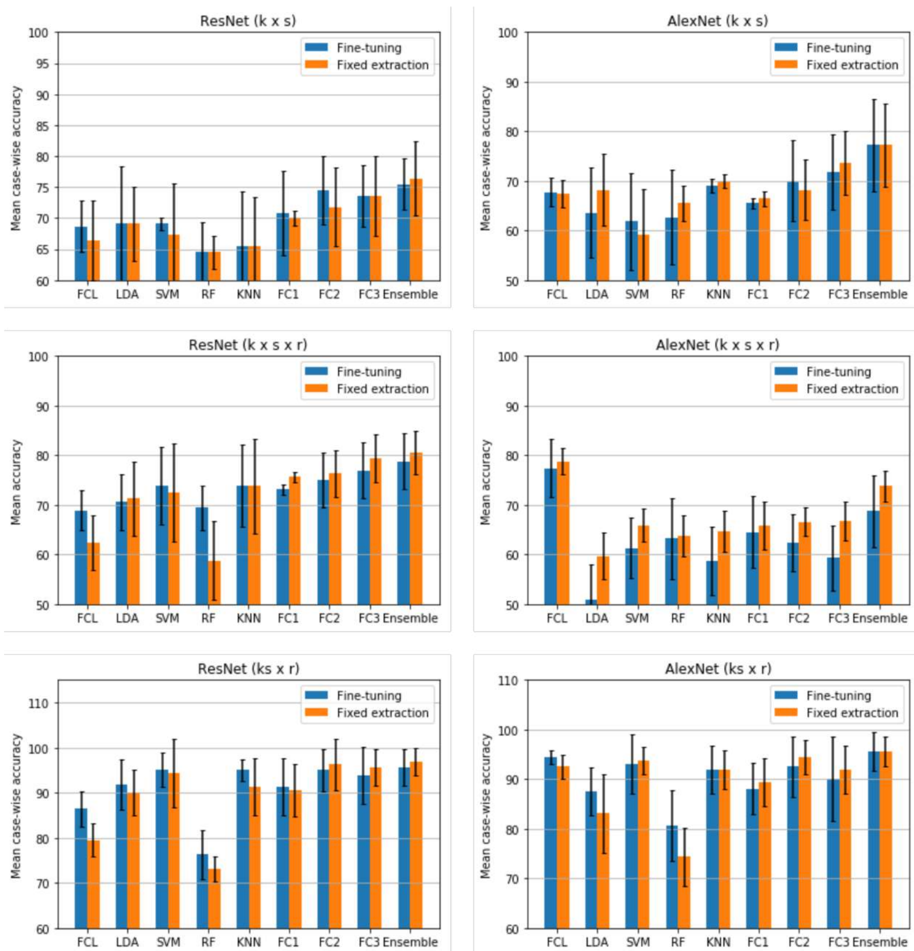


Figure 13: Average case-wise accuracy on the cysts database.

and F1-score values. The values obtained were compatible with the respective accuracy seen in Figures 12 and 13.

Table 6: Average computing time.

hh:mm:ss	k × s		k × s × r		ks × r	
Structure	Fine-tuning	Fixed	Fine-tuning	Fixed	Fine-tuning	Fixed
ResNet	00:03:17	00:03:12	00:07:41	00:05:36	00:07:02	00:04:15
AlexNet	00:03:42	00:02:42	00:06:20	00:05:06	00:06:11	00:04:37

Table 7: Case-wise average computing time.

hh:mm:ss	k × s		k × s × r		ks × r	
Structure	Fine-tuning	Fixed	Fine-tuning	Fixed	Fine-tuning	Fixed
ResNet	00:04:53	00:02:31	00:07:54	00:03:56	00:06:35	00:03:14
AlexNet	00:03:43	00:01:53	00:05:45	00:02:42	00:05:50	00:02:39

Table 8: Cysts dataset HCNNE performance metrics with ResNet.

	k × s		k × s × r		ks × r	
Structure	Fine-tuning	Fixed	Fine-tuning	Fixed	Fine-tuning	Fixed
Precision	0.81	0.83	0.86	0.88	0.97	0.99
Recall	0.78	0.77	0.86	0.83	0.97	0.98
F1-score	0.79	0.78	0.86	0.84	0.97	0.98

Table 9: Cysts dataset HCNNE performance metrics with AlexNet.

	k × s		k × s × r		ks × r	
Structure	Fine-tuning	Fixed	Fine-tuning	Fixed	Fine-tuning	Fixed
Precision	0.74	0.78	0.74	0.73	0.93	0.87
Recall	0.72	0.73	0.73	0.69	0.88	0.81
F1-score	0.73	0.74	0.73	0.70	0.90	0.83

Tables 12 to 17 are the average confusion matrices. The columns contain the predictions and the rows, the expected classifications. The values represent

Table 10: Cysts case-wise dataset HCNNE performance metrics with ResNet.

	k × s		k × s × r		ks × r	
Structure	Fine-tuning	Fixed	Fine-tuning	Fixed	Fine-tuning	Fixed
Precision	0.60	0.62	0.65	0.72	0.94	0.92
Recall	0.60	0.62	0.61	0.65	0.90	0.86
F1-score	0.60	0.62	0.62	0.65	0.92	0.88

Table 11: Cysts case-wise dataset HCNNE performance metrics with AlexNet.

	k × s		k × s × r		ks × r	
Structure	Fine-tuning	Fixed	Fine-tuning	Fixed	Fine-tuning	Fixed
Precision	0.61	0.58	0.69	0.56	0.77	0.87
Recall	0.60	0.58	0.69	0.55	0.78	0.87
F1-score	0.61	0.58	0.69	0.55	0.77	0.86

the percentage of samples classified over the total of samples in the expected class. Note that since the values are averaged, some values do not result in an integer number of samples.

Table 12: Average confusion matrix for k × s case.

Structure		Fine-tuning		Fixed	
		k	s	k	s
ResNet	k	0.92	0.08	0.88	0.12
	s	0.33	0.68	0.41	0.60
AlexNet	k	0.92	0.08	0.84	0.16
	s	0.64	0.36	0.35	0.65

Figure 14 shows two examples of precision-recall graphs, the measure of the network performance is associated with the area under the curve. Figure 15 contains two examples of graphs representing the average cost function along the epochs. In both figures, the examples are from the case of identifying OKCs and radicular cysts (ks × r) using ResNet and AlexNet, respectively.

The best accuracy achieved in the case where the three classes of cysts

Table 13: Average confusion matrix for ks × r case.

Structure		Fine-tuning		Fixed	
		ks	r	ks	r
ResNet	ks	0.99	0.01	0.97	0.03
	r	0.03	0.97	0.17	0.83
AlexNet	ks	0.97	0.03	0.96	0.04
	r	0.33	0.67	0.09	0.91

Table 14: Average confusion matrix for k × s × r case.

Structure		Fine-tuning			Fixed		
		k	s	r	k	s	r
ResNet	k	0.88	0.03	0.08	0.79	0.03	0.18
	s	0.02	0.96	0.02	0.10	0.86	0.04
	r	0.29	0.01	0.70	0.41	0.02	0.58
AlexNet	k	0.92	0.02	0.06	0.86	0.02	0.12
	s	0.04	0.87	0.09	0.06	0.90	0.04
	r	0.49	0.09	0.43	0.39	0.07	0.54

Table 15: Case-wise average confusion matrix for k × s case.

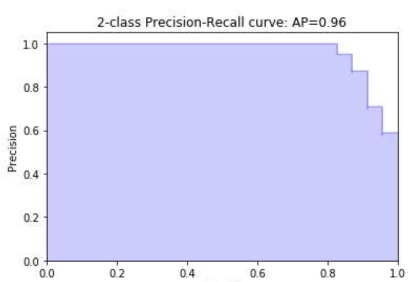
Structure		Fine-tuning		Fixed	
		k	s	k	s
ResNet	k	0.85	0.09	0.86	0.08
	s	0.67	0.34	0.75	0.25
AlexNet	k	0.83	0.11	0.73	0.22
	s	0.66	0.35	0.48	0.52

Table 16: Case-wise average confusion matrix for ks × r case.

Structure		Fine-tuning		Fixed	
		ks	r	ks	r
ResNet	ks	0.93	0.03	0.95	0.01
	r	0.34	0.58	0.60	0.32
AlexNet	ks	0.93	0.03	0.94	0.02
	r	0.10	0.82	0.18	0.74

Table 17: Case-wise average confusion matrix for k × s × r case.

Structure		Fine-tuning			Fixed		
		k	s	r	k	s	r
ResNet	k	0.85	0.03	0.12	0.85	0.06	0.08
	s	0.16	0.78	0.06	0.28	0.63	0.09
	r	0.63	0.08	0.29	0.67	0.13	0.21
AlexNet	k	0.87	0.03	0.11	0.79	0.04	0.17
	s	0.05	0.87	0.08	0.06	0.88	0.05
	r	0.57	0.07	0.37	0.53	0.08	0.40



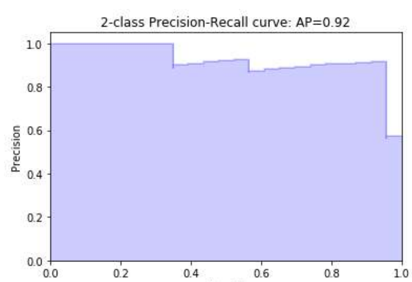


Figure 14: Precision-recall graphs for the ks × r case using ResNet and AlexNet, respectively.

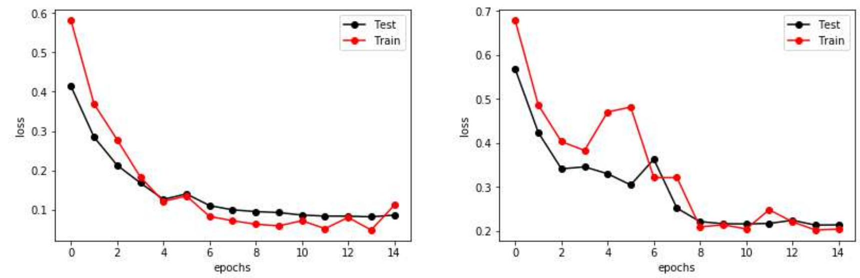
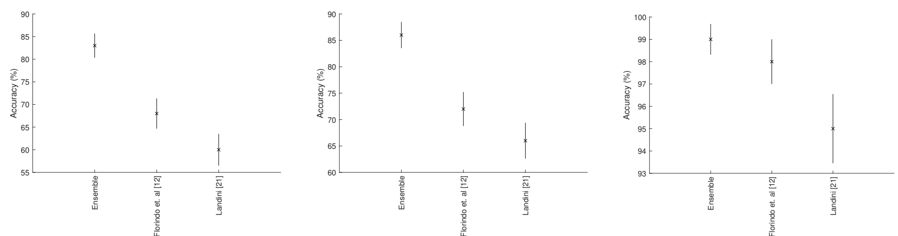


Figure 15: Graphs of the average cost function for the ks × r case using ResNet and AlexNet, respectively.



k × s   k × s × r   ks × r

Figure 16: Confidence intervals for different methods on cyst database.

are compared (k × s × r) was 86%. When distinguishing between OKCs and radicular cysts (ks × r) the best accuracy was 99%. And finally, the most difficult case, where the two types of OKCs are analysed, the best accuracy was 83%. All of those were achieved by the ensemble, using ResNet with fine-tuning.

Table 18 presents a comparison of these results with the ones found in Florindo et al. 2017 [35]. The ks × r problem is easily solved by all the methods, so it is not possible to make any comparison. Nonetheless, in the other two configurations it is possible to see that the proposed method has achieved higher accuracy, by a margin of 10% or more, and by 5% in the case-wise approach. Figure 16 shows the confidence interval for some illustrative approaches presented in Table 18. The presented proposal outperforms previous results published in the literature with significant margin both in k × s and k × s k × r problems.

Table 18: Comparison of results with the literature.

Source	k × s	k × s × r	ks × r
Landini [40]	60%	66%	95%
Florindo et. al [35]	68%	72%	98%
Case-wise [35]	71%	76%	100%
Ensemble	83%	86%	99%
Case-wise	77%	81%	97%

6.1. Discussion

The presented results confirm the expectations on the proposed method as being a competitive approach both in general texture recognition problems used for benchmark and especially on the medical task investigated here, namely, the identification of odontogenic oral cysts. An interesting observation in the medical task is provided by the confusion matrices and precision/recall metrics. In the k × s problem, a significant ratio of sporadic cysts are recognized as syndromic (corresponding to lower recall). This is motivated by the intra-group variability associated to the syndromic group, which in many situations behave like sporadic samples. Similar discrepancy is noticed when the algorithm attempts to categorize all three groups at once. In this case, radicular cysts are misclassified as syndromic. Again, the high variability of pixel patterns in the syndromic nuclei is a huge challenge for a precise recognition. Nevertheless, despite the existence of room for improvement in these tasks, the improvement over results previously reported is substantial. This is even more encouraging if one takes into account that the proposed methodology is relatively straightforward and achieves such promising results at low computational cost.

In practical terms, the achieved results represent an important improvement over the results previously published in the literature for the cyst problem, in particular, on the k × s and k × s × r tasks. These are, in fact, challenging problems, as the differences in pixel patterns discriminating sporadic and syndromic keratocysts are quite subtle, as previously reported in the literature

[40, 35]. In a more general viewpoint, the presented results suggest that the combination of low level features with features learned by a CNN and ensemble algorithms can be a powerful strategy in computer aided diagnostic when the amount of data for training and image resolution is limited. This is the case of the medical problem tackled here and also a common situation in many medical areas.

7. Conclusions

This work tackled a difficult medical image classification problem, using CNN and local binary patterns to extract features from odontogenic cysts images an then combining those features together to form multiple classifiers that would be finally combined through an ensemble.

This strategy applied to benchmark texture classification problems led to an accuracy of 98.3% and 99.5%, in UIUC [7] and UMD [8] databases, respectively. This is a competitive result, when compared with recent state-of-the-art approaches.

In the jaw cysts problem, HCNNE achieved 83.4% (k $\times$ s) and 85.9% (k $\times$ s $\times$ r) average accuracy, which represents an improvement of around 10% or more, compared to previous works.

The method developed in this paper aimed at solving difficult image classification problems, when the database is small. The strategy presented does not require new training data or high computational power, making it attractive for certain types of applications.

A limitation of the current study is its focus on the particular problem of oral cyst recognition. In this regard, future work plans include investigating the performance of the proposed methodology in other medical image problems and general image recognition. There is also room for further improvement in terms of the computational model, e.g. experimenting with more complex ensemble methods such as boosting and bagging, exploring different feature combinations and testing other network architectures and classifiers. Finally, there are plans to introduce modern strategies for feature pooling like those presented in [46], for example, into the ensemble pipeline.

Acknowledgements

Marina Rocha gratefully acknowledges the support of the National Council for Scientific and Technological Development, CNPq (Grant #121791/2019-0). Joao Florindo gratefully acknowledges the financial support of São Paulo Research Foundation (FAPESP) (Grant #2020/01984-8) and from National Council for Scientific and Technological Development, Brazil (CNPq) (Grants #306030/2019-5 and #423292/2018-8).

Conflict of interest statement

The authors certify that they have no known affiliations with or involvement in any organization or entity with any interest in the subject matter discussed in this paper.

Data availability statement

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

References

[1] Wei, J.; He, J.; Chen, K.; Zhou, Y.; Tang, Z. Collaborative filtering and deep learning based recommendation system for cold start items. *Expert Systems with Applications* **2017**, *69*, 29–39.

[2] Litjens, G.; Kooi, T.; Bejnordi, B. E.; Setio, A. A. A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J. A.; Van Ginneken, B.; Sánchez, C. I. A survey on deep learning in medical image analysis. *Medical image analysis* **2017**, *42*, 60–88.

[3] Alipanahi, B.; Delong, A.; Weirauch, M. T.; Frey, B. J. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology* **2015**, *33*, 831–838.

[4] Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). 2005; pp 886–893.

[5] Forcén, J.; Pagola, M.; Barrenechea, E.; Bustince, H. Combination of features through weighted ensembles for image classification. *Applied Soft Computing* **2019**, *84*, 105698.

[6] Ren, Y.; Zhang, L.; Suganthan, P. N. Ensemble classification and regression-recent developments, applications and future directions. *IEEE Computational intelligence magazine* **2016**, *11*, 41–53.

[7] UIUC Dataset. `http://www-cvr.ai.uiuc.edu/poncegrp/data/index.html`, Last accessed 06-2021.

[8] UMD Dataset. `http://cfar.umd.edu/\_fer/High-resolution-database/hrdatabase.htm`, Last accessed 06-2021.

[9] Goodfellow, I.; Bengio, Y.; Courville, A. *Deep learning*; MIT press, 2016.

[10] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; pp 770–778.

[11] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition. 2009; pp 248–255.

[12] Kumar, A.; Kim, J.; Lyndon, D.; Fulham, M.; Feng, D. An ensemble of fine-tuned convolutional neural networks for medical image classification. *IEEE journal of biomedical and health informatics* **2016**, *21*, 31–40.

[13] Bacanin, N.; Zivkovic, M.; Al-Turjman, F.; Venkatachalam, K.; Trojovskỳ, P.; Strumberger, I.; Bezdan, T. Hybridized sine cosine algorithm with convolutional neural networks dropout regularization application. *Scientific Reports* **2022**, *12*, 1–20.

[14] Malakar, S.; Ghosh, M.; Bhowmik, S.; Sarkar, R.; Nasipuri, M. A GA based hierarchical feature selection approach for handwritten word recognition. *Neural Computing and Applications* **2020**, *32*, 2533–2552.

[15] LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *nature* **2015**, *521*, 436–444.

[16] Ragab, D. A.; Attallah, O. FUSI-CAD: coronavirus (COVID-19) diagnosis based on the fusion of CNNs and handcrafted features. *PeerJ Computer Science* **2020**, *6*, e306.

[17] Attallah, O.; Anwar, F.; Ghanem, N. M.; Ismail, M. A. Histo-CADx: duo cascaded fusion stages for breast cancer diagnosis from histopathological images. *PeerJ Computer Science* **2021**, *7*, e493.

[18] Attallah, O.; Sharkas, M. GASTRO-CADx: a three stages framework for diagnosing gastrointestinal diseases. *PeerJ Computer Science* **2021**, *7*, e423.

[19] Anwar, F.; Attallah, O.; Ghanem, N.; Ismail, M. A. Automatic breast cancer classification from histopathological images. 2019 International Conference on Advances in the Emerging Computing Technologies (AECT). 2020; pp 1–6.

[20] Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32.

[21] Cortes, C.; Vapnik, V. Support-vector networks. *Machine learning* **1995**, *20*, 273–297.

[22] Rao, C. R. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)* **1948**, *10*, 159–203.

[23] Fix, E.; Hodges, J. L. Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique* **1989**, *57*, 238–247.

[24] Wolpert, D. H. The lack of a priori distinctions between learning algorithms. *Neural computation* **1996**, *8*, 1341–1390.

[25] Ho, Y.-C.; Pepyne, D. L. Simple explanation of the no-free-lunch theorem and its implications. *Journal of optimization theory and applications* **2002**, *115*, 549–570.

[26] Rokach, L. Ensemble-based classifiers. *Artificial Intelligence Review* **2010**, *33*, 1–39.

[27] Ragab, D. A.; Sharkas, M.; Attallah, O. Breast cancer diagnosis using an efficient CAD system based on multiple classifiers. *Diagnostics* **2019**, *9*, 165.

[28] Attallah, O. ECG-BiCoNet: An ECG-based pipeline for COVID-19 diagnosis using Bi-Layers of deep features integration. *Computers in biology and medicine* **2022**, *142*, 105210.

[29] Fouad, S.; Randell, D.; Galton, A.; Mehanna, H.; Landini, G. Unsupervised morphological segmentation of tissue compartments in histopathological images. *PloS one* **2017**, *12*, e0188717.

[30] Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence* **2002**, *24*, 971–987.

[31] Guo, Z.; Zhang, L.; Zhang, D. A completed modeling of local binary pattern operator for texture classification. *IEEE transactions on image processing* **2010**, *19*, 1657–1663.

[32] Liu, L.; Fieguth, P.; Guo, Y.; Wang, X.; Pietikäinen, M. Local binary features for texture classification: Taxonomy and experimental study. *Pattern Recognition* **2017**, *62*, 135–160.

[33] Attallah, O.; Sharkas, M. Intelligent dermatologist tool for classifying multiple skin cancer subtypes by incorporating manifold radiomics features categories. *Contrast media & molecular imaging* **2021**, *2021*.

[34] Landini, G. Quantitative analysis of the epithelial lining architecture in radicular cysts and odontogenic keratocysts. *Head & Face Medicine* **2006**, *2*, 4.

[35] Florindo, J. B.; Bruno, O. M.; Landini, G. Morphological classification of odontogenic keratocysts using Bouligand–Minkowski fractal descriptors. *Computers in biology and medicine* **2017**, *81*, 1–10.

[36] Bousetouane, F.; Morris, B. Off-the-shelf CNN features for fine-grained classification of vessels in a maritime environment. International Symposium on Visual Computing. 2015; pp 379–388.

[37] Burges, C. J. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery* **1998**, *2*, 121–167.

[38] Platt, J., et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* **1999**, *10*, 61–74.

[39] Lin, H.-T.; Lin, C.-J.; Weng, R. C. A note on Platt's probabilistic outputs for support vector machines. *Machine learning* **2007**, *68*, 267–276.

[40] Landini, G.; Othman, I. E. Architectural analysis of oral cancer, dysplastic, and normal epithelia. *Cytometry Part A: The Journal of the International Society for Analytical Cytology* **2004**, *61*, 45–55.

[41] Xu, Y.; Yang, X.; Ling, H.; Ji, H. A new texture descriptor using multifractal analysis in multi-orientation wavelet pyramid. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2010; pp 161–168.

[42] Kannala, J.; Rahtu, E. Bsif: Binarized statistical image features. Proceedings of the 21st international conference on pattern recognition (ICPR2012). 2012; pp 1363–1366.

[43] Quan, Y.; Xu, Y.; Sun, Y.; Luo, Y. Lacunarity analysis on image patterns for texture classification. Proceedings of the IEEE conference on computer vision and pattern recognition. 2014; pp 160–167.

[44] Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; Vedaldi, A. Describing textures in the wild. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014; pp 3606–3613.

[45] Bruna, J.; Mallat, S. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence* **2013**, *35*, 1872–1886.

[46] Cimpoi, M.; Maji, S.; Kokkinos, I.; Vedaldi, A. Deep filter banks for texture recognition, description, and segmentation. *International Journal of Computer Vision* **2016**, *118*, 65–94.

[47] Xu, Y.; Huang, S.; Ji, H.; Fermüller, C. Scale-space texture description on sift-like textons. *Computer Vision and Image Understanding* **2012**, *116*, 999–1013.

[48] Timofte, R.; Van Gool, L. A Training-free Classification Framework for Textures, Writers, and Materials. BMVC. 2012; p 14.

[49] Attallah, O.; Karthikesalingam, A.; Holt, P. J.; Thompson, M. M.; Sayers, R.; Bown, M. J.; Choke, E. C.; Ma, X. Using multiple classifiers for predicting the risk of endovascular aortic aneurysm repair re-intervention through hybrid feature selection. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine* **2017**, *231*, 1048–1063.