Majority voting of Heterogeneous Classifiers for finding abnormalities in the Gastro-Intestinal Tract

Zeshan Khan, Muhammad Atif Tahir
School of Computer Science, National University of Computer and Emerging Sciences, Karachi Campus, Pakistan
{zeshan.khan,atif.tahir}@nu.edu.pk

ABSTRACT

An endoscopy is a procedure in which a doctor uses specialized instruments to view and operate on the internal organs and vessels of the body. This paper aims to detect the diseases and abnormalities in the Gastro-Intestinal Tract using multimedia data. It differs from other projects in the medical domain because it does not use medical imaging like X-rays, CT scan etc. The dataset, which comprises of 5293 images, is provided by MediaEval Benchmarking Initiative for Multimedia Evaluation. The data is collected during traditional colonoscopy procedures. Techniques from the fields of multimedia content analysis (to extract information from the visual data) and machine learning (for classification) have been used. On testing data, 98% accuracy, 0.76 F_1 and an MCC of 0.75 is achieved using majority voting of logistic regression, random forest, and extra trees classifiers.

1 INTRODUCTION

Medical image diagnosis is one of the most challenging tasks pertinent to the industry of computer vision. Most of the work in the recent times has been done on CT-Scans, X-Rays, and MRI etc. The Medico Task of 2018 [5] [1] challenged their participants to predict the abnormalities in the Gastro-Intestinal tract through endoscopic examination [4]. This implies the presence of multimedia images instead of traditional medical images for the challenge [4]. Deep analysis on GI tract images can help to predict abnormalities and diseases in its initial stages. 5293 images were used for training purpose and the 8740 were reserved for testing data. Different pre-processing techniques were applied and machine learning models were deployed for accurate systems.

2 APPROACH

Feature Engineering is one of the most challenging and key part of any Machine Learning problem. Figure 1 shows the proposed model. Discriminating features are the requirement for the function approximation. The task organizers provided 6 pre-computed visual features for every image. These include JCD, Tamura, Color Layout, Edge Histogram, Auto Color Correlogram and PHOG. Alongside these pre-computed visual features, deep learning features are also used to extract meaningful information for classification. There are some visual features those can be extracted by using deep Networks. As the training dataset is of 5293 images is very low for the training of a deep learning model, a pre-trained model VGG19 is used [6]. VGG 19 is a very deep convolutional networks of up to 19 weight layers (16 convolutional layers and 3 fully-connected layers) for large scale image classification. With the help of pre-train process using large dataset from the ImageNet challenge and retraining of the last 2 layers with these medical images, the VGG 19 model is used to extract plentiful visual concepts.

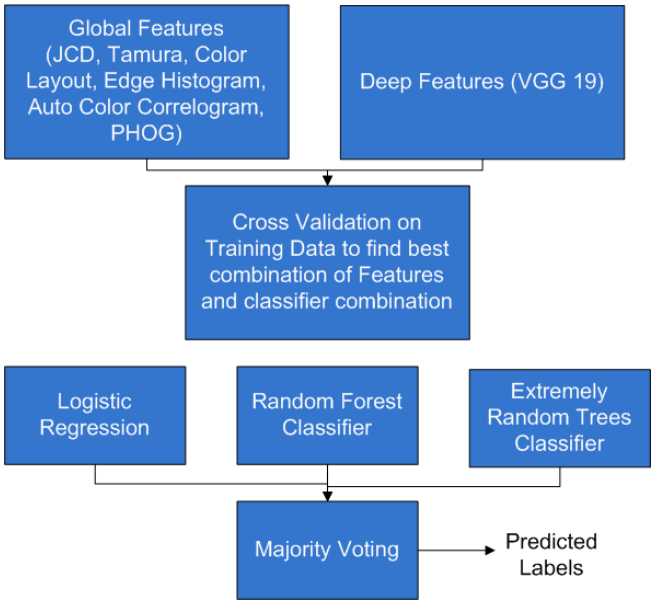

Figure 1: Proposed Model.

Classifiers are trained on the logistic regression [1], random forest [2] and extremely random trees classifier [3] for the features that have been extracted. There were two categories of the features including pre-computed texture features and VGG features, the features extracted by using VGG19 pre-trained model. Ensemble implies the fact that the final model makes use of weighted majority voting among all the independent models trained on all features. The weights of the ensemble are the percentage of accuracy measure of the independent classifier. It should be noted that various advanced machine learning techniques have been investigated but the best results were obtained using logistic regression, random forest and extremely random trees classifiers and thus reported in this paper.

The interesting characteristics of this competition included the limited data to train the models and the class imbalance. The technique of resampling is used to generate more data for each class. The resampling generated some more features of each class and resulted in the same number of instance for each of the available 16 classes. The resampling also increased the training dataset and the increased dataset is used to train and validate different models.

[1] http://www.multimediaeval.org/mediaeval2018/medico/index.html

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

3 RESULTS AND ANALYSIS

The linear regression, extremely randomized trees and random forest models have been implemented using Python's scikit-learn package. We trained logistic regression, random forest, and extremely random trees on both deep and global features. The results are first evaluated on training data using 10 Fold cross validation. By applying the proposed model, we obtained the accuracy of 97%, F_1 score of 0.90 and MCC of 0.81 on the 10 fold cross validation of the training data. Based on this initial investigation, the following runs are submitted to evaluate the performance of classifiers independently. The runs are submitted with the focus of 3 runs for speed results generation and 3 runs for accuracy.

- **Run1** Ensemble of 7 features [JCD, Tamura, Color Layout, Edge Histogram, Auto Color Correlogram, PHOG and VGG features] trained on 60 images, using voting of the logistic regression, random forest and extremely random trees classification algorithms.
- **Run2** Same as Run1 but trained on 300 images.
- **Run3** Same as Run1 but trained on all 5293 images.
- **Run4** Ensemble of 6 features [JCD, Tamura, Color Layout, Edge Histogram, Auto Color Correlogram, and PHOG] trained on 60 images, using voting of the logistic regression, random forest and extremely random trees classification algorithms.
- **Run5** Same as Run4 but trained on 300 images.
- **Run6** Same as Run4 but trained on all 5293 images.

Table 1: Accuracy, F_1, and MCC on different runs of testing data.

	Accuracy	F_1	MCC
Run1	0.956	0.625	0.614
Run2	0.957	0.587	0.603
Run3	0.954	0.549	0.572
Run4	0.961	0.611	0.597
Run5	0.976	0.745	0.741
Run6	**0.979**	**0.752**	**0.756**

Table 1 shows the summary of some evaluation criterions on best run. Accuracy of 97.9% is observed with F-score of 0.75 and MCC of 0.76. It is interesting to see that the best run is obtained by using just global features without using any deep learning features. We will investigate in future why deep features perform poorly. Initial investigation has indicated that a lot of samples that should belong to class "ulcerative-colitis" are misclassified as class "esophagitis" by using deep features. The best run is obtained using Run6 in which all 5293 images are used and this approach is basically ensemble of 6 features (JCD, Tamara, Edge Histograms, Color Layout, Auto Color Correlogram and PHOG). Logistic regression, random forest and extremely random trees is being used as a classifier with weighted majority voting. Table 2 is the confusion matrix of various classes. It is observed that total of around 1469 samples are misclassified. Two categories are mainly responsible for the misclassification which are "dyed-lifted-polyps" and "dyed-resection-margins". Around 500 samples are misclassified in these 2 categories (Tables 3 and 4). Table 5 shows the confusion matrix for polyps versus non-polyps class. Overall, performance is satisfactory but still there is a need to investigate state of the art texture and local features to further improve the performance.

Table 2: Confusion matrix of all classes. There are total 16 classes and summary of all classes is shown.

Predicted Actual	ALL	non-ALL
ALL	7271	1469
non-ALL	1469	129631

Table 3: Confusion matrix for class dyed-lifted-polyps versus non dyed-lifted-polyps. df = dyed-lifted.

Predicted Actual	df-polyps	non df-polyps
dyed-lifted-polyps	339	236
non-dyed-lifted-polyps	217	7948

Table 4: Confusion matrix for class dyed-resection-polyps versus non dyed-resection-polyps. df = dyed-lifted.

Predicted class Actual class	dr-margins	non-dr-margins
dyed-resection-margins	387	232
non-dyed-resection-margins	177	7944

Table 5: Confusion matrix for class polyps versus non non-polyps.

Predicted Actual	polyps	non-polyps
polyps	241	281
non-polyps	133	8085

4 CHALLENGES AND FUTURE WORK

It has been observed that results produced for many classes are quite accurate. However, there are some classes that are confusing the system. Future work aims to target these classes hierarchically and improve the performance using local features.

5 CONCLUSION

A model to classify gastro-intestinal abnormalities using endoscopic images is presented. Training (5293 samples) and Testing (8740 samples) data was provided by MediaEval Benchmarking Initiative for Multimedia Evaluation. As mentioned earlier in the introduction, the study used multimedia content analysis, machine learning and ensemble learning techniques for classification. The best of the results were found on majority voting of three models including logistic regression, random forest and extremely random trees classifier on 6 different features (including JCD, Tamura, Color Layout, Edge Histogram, Auto Color Correlogram and PHOG) which resulted in an accuracy of 97% with F1-score of 0.75 and MCC of 0.76 on testing data.

REFERENCES

[1] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer.

[2] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (01 Oct 2001), 5–32.

[3] Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine Learning* 63, 1 (2006), 3–42.

[4] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. 2017. KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSys'17)*. ACM, 164–169.

[5] Konstantin Pogorelov, Michael Riegler, Pal Halvorsen, Thomas de Lange, Kristin Ranheim Randel, Duc-Tien Dang-Nguyen, Mathias Lux, and Olga Ostroukhova. 2018. Medico Multimedia Task at MediaEval 2018. In *MediaEval18, 29-31 October 2018, Sophia Antipolis, France.*

[6] K. Simonyan and A. Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* (2014).