CHAPTER-5 HIGH VARIANCE AND LOW CORRELATION BASED DATA REDUCTION AND FEATURE SELECTION APPROACH

CAD system for CE is a data-intensive system. CE procedure of one patient yields approximately 60000 images and a feature vector for each image gets too large. In such a scenario a data reduction technique is a must. This chapter presents a high variance and low correlation (HVLC) based feature selection and data reduction approach with reference to ulcer detection using a CAD system. This approach can be used independently for data reduction and feature selection.

5.1 Introduction

A CAD system is needed to assist a physician with a fast and accurate analysis. CAD systems play an important role in searching for similar medical records, training inexperienced clinicians and assisting the medical experts to improve the accuracy of medical diagnosis [113].

One of the most common lesions of the GI tract is an ulcer. The mortality rate for bleeding ulcers is nearly 10% [50]. Two of the important causes of GI tract ulcers are non-steroidal anti-inflammatory drugs (NSAIDs) and bacteria called Helicobacter pylori (H. pylori) [65]. An untreated ulcer may lead to ulcerative colitis and Crohn’s disease. Thus a timely detection of ulcer is a must. Table 5.1 presents a summary of various ulcer detection systems.

Table 5.1: Summary of prior art on ulcer detection

Work	Features used	Method / classifier used	Limitations	Performance	Dataset size
[65]	The saliency of color and texture	SVM	Too few data samples	Accuracy=92.65% Sensitivity=94.12%	340 images
[44]	The chromatic moment for each patch	Neural network	The texture feature is neglected. Too few training samples	Specificity=84.68± 1.80 Sensitivity=92.97± 1.05	100 images
[8]	LBP histogram	Curvelet transformation, Multi-layer perceptron (MLP) and SVM	Too few samples for training	Accuracy=92.37% Specificity=91.46 % Sensitivity=93.28%	100 images
[53]	Dif lac analysis	De-noising using Bi-dimensional ensemble empirical mode decomposition (BEEMD)	Too few samples for training	Mean accuracy > 95%	176 images
[58]	Texture and colour	Vector supported convex hull method	Specificity is less. Skewed data	Recall = 0.932 Precision =0.696 Jaccard index= 0.235	36 images
[60]	Leung and Malik (LM) and LBP histograms	k-nearest neighbor (k-NN) classifier is used	Computationally intense. Skewed data	Recall =92% Specificity =91.8%	1750 images

By studying prior art, it is learned that automated ulcer detection systems are subject to following limitations: low detection rate, small-sized data leading to poor learning, skewed data leading to over-fitting and computationally intense techniques. To address the problems above, this study proposes an efficient design of a CAD system for ulcer detection in CE using optimized and clinically significant features. The subsequent section discusses the details regarding the proposed system. Major contributions to this work are:

- Identifying the most significant feature set for ulcers.
- HVLC based feature selection technique.
- Automated ulcer detection using an optimized feature set and SVM.
- A thorough comparison of the proposed feature selection technique with five other feature selection techniques.
- A thorough comparison of the proposed system two other systems.

5.2 Methods and Models

The proposed CAD system consists of four different stages namely, pre-processing, feature extraction, feature selection, and classification. This section explains the detailed significance of each stage. Figure 5.1 shows an overview of the proposed system and following is the procedure of the entire system:

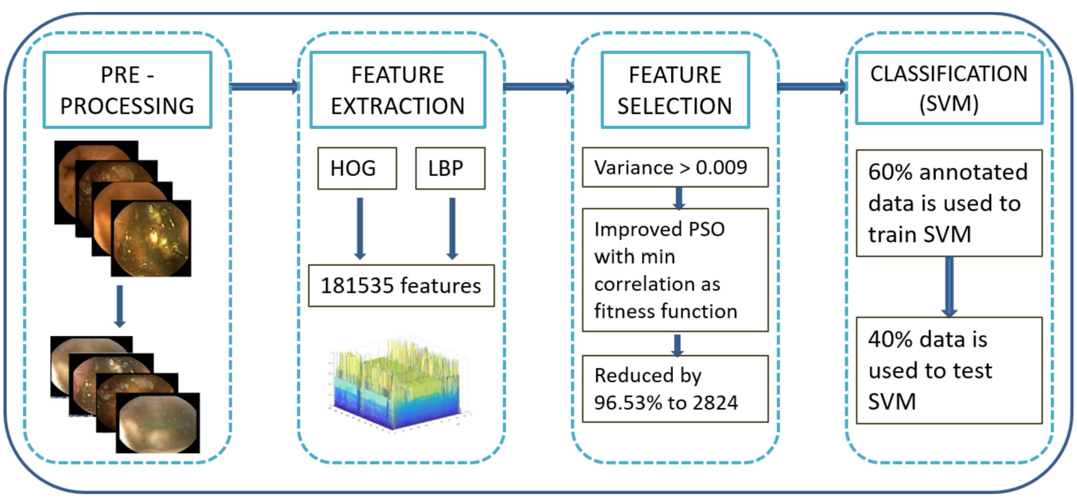


Figure 5.1: Overview of proposed CAD system for ulcer detection in CE images

Steps involved through the system are as follows:

- Load CE images.
- Perform contrast stretching for image enhancement.
- Extract HOG features.
- Extract LBP features.
- Reduce feature vector by selecting features with variance greater than a defined threshold value.
- Select desired number features using particle swarm optimization (PSO) with min correlation as the fitness function.
- Partition data into training and testing set.
- Train the classifier model on training data.
- Perform classification on the test data using the trained classifier model.

5.2.1 Pre-processing of Input Image

The image enhancement technique is used to make full use of all the information present in the image. Image enhancement provides a better perception of human visualization and better input for CAD systems [114]. Here, contrast stretching saturates top and bottom 1% of pixel intensity and maps the intensity of the given image to the desired image. Figure 5.2 shows a result of image enhancement using a contrast stretching technique.

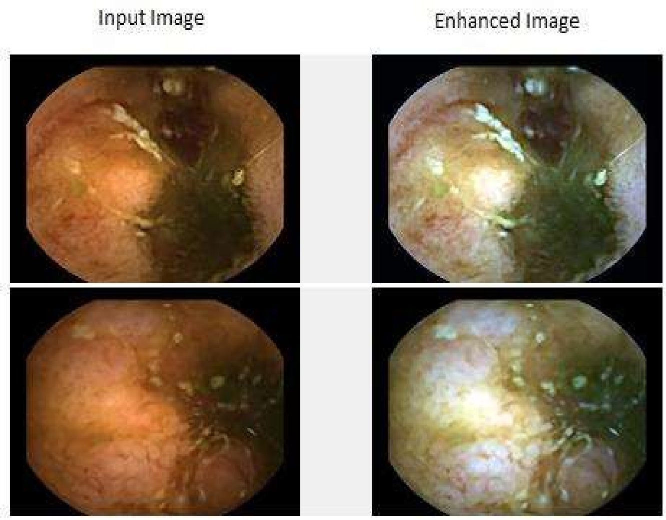


Figure 5.2: Input image and corresponding enhanced image

5.2.2 HVLC Based Data Reduction and Feature Selection Approach

The definition of feature selection problem is as follows: given a set of features, select a subset of features that perform best under some classification system. The optimal size of the feature set reduces the cost of recognition as well as lead to a better classification accuracy due to finite sample size [115]. HOG feature extraction process yields 181476 features, and LBP feature extraction leads to 59 features. Total of 181535 features is extracted from each of CE image in the dataset. Needless to say that this feature set produces extraordinary results for ulcer classification but, we must limit the number of features to a considerable number. To reduce the number of features and maintain the performance of classification, we propose a novel feature selection technique. Features with high variance can easily discriminate between two classes but, variance alone is not a good measure of information. Any two features can exhibit high variance but they may be correlated. Therefore, the proposed feature selection technique is designed on a dual criterion: high variance and low correlation. The proposed feature selection technique is termed as high variance low correlation (HVLC) technique. HVLC technique reduces the obtained feature set by 96.53% to 2824 number of features. The proposed technique encompasses a minimum correlation fitness function based particle swarm optimization (PSO). In PSO, a collection of particles moves around in the search space to find the optimal solution. It is influenced by a local best value called pbest and a global best value called gbest. pbest is the particle's own best value and gbest is the best value of the whole swarm. At each iteration, i^{th} particle updates position P as per (5.1) and velocity V as per (5.2).

$$P_i\,(t+1) = P_i\,(t) + V_i\,(t) \tag{5.1}$$

$$\begin{aligned} V_i\,(t+1) &= wV_i\,(t) \\ &+c1 * rand([0,1]) * \left(pbest - P_i(t)\right) \\ &+ c2 * rand([0,1]) * \left(gbest - P_i(t)\right) \end{aligned} \tag{5.2}$$

Where t and (t+1) are two successive iterations, inertia weight w, cognitive coefficient c1, and social coefficient c2 are constants. Also, c1 and c2 control the magnitude of steps taken by particle towards pbest (personal) and gbest (global) respectively. Correlation C between the given input data is computed by (5.3).

$$C(X,Y) = \frac{\sum_{i=1}^{n}(\mathrm{X_i} - \overline{\mathrm{X}})(\mathrm{Y_i} - \overline{\mathrm{Y}})}{\sqrt{\sum_{i=1}^{n}(\mathrm{X_i} - \overline{\mathrm{X}})^2 \sum_{j=1}^{n}(\mathrm{Y_j} - \overline{\mathrm{Y}})^2}} \tag{5.3}$$

Where X and Y are input data and $\overline{X}$ and $\overline{\mathrm{Y}}$ are mean of X and Y respectively. Algorithm 5.1 presents a detailed procedure to obtain HVLC reduced feature set.

Algorithm 5.1: HVLC based data reduction and feature selection

```
Initial feature set S=[1,2,...,n]
Final feature set Fs= Ø
Choose features with variance > 0.009
Surviving feature set Ss=[1,2,...,m] where  m < n
Set k = size of Fs from within Ss
Set values of PSO control parameters: w=0.2, c1=c2=2 and max iteration
Create and initialize particles with their P and V; initialize gbest of the
population as infinity
Repeat:
For i=1 to population
Compute C as per (5.3) as a ranking criteria
f = argmin(C)
EndFor
Update pbest = min(f)
Update gbest = min (gbest,pbest)
Update P using (5.1)
Update V using (5.2)
Until the termination criterion of max iterations is met
Return PSO selected values with min correlation
Final set of HVLC futures Fs=[1,2,...,k] where k < m
```

The value of the variance threshold is experimentally chosen to fit the application. These features are then fed to the classifier to classify between ulcers and normal images.

5.3 Classification using Different Classifiers

Ulcer detection in CE is a binary classification problem having exactly two classes namely ulcer and normal. An SVM finds the best possible hyperplane that can explicitly separate data points of two different classes. A hyperplane with the maximum margin between the two classes is the best hyperplane. The support vectors are the data points falling on the boundary of the slab parallel to the hyperplane. The other classifiers kNN (k nearest neighbor) and ensemble classifier are also experimented in this study. The subsequent section presents detailed result analysis and performance of the proposed system.

5.4 Results and Discussion

A total of 1200 images from CE videos [9] is extracted out of which 201 images are of ulcers, and 999 images are normal. The dimension of each image is 576x576 pixels. All the images were manually diagnosed and annotated by physicians providing the ground truth. To avoid imbalanced data and over fitting 100 ulcer images and 100 normal images are carefully chosen from the annotated dataset.

The performance of the automated ulcer detection system largely depends on the performance of the proposed feature selection technique. Therefore the performance of proposed technique is thoroughly compared with five other feature selection techniques namely correlation-based feature selection (CFS) technique [116], relief F [117], local learning-based clustering (LLCFS) [118], SVM-RFE [119] and maximum relevance minimum redundancy (mRMR) [120]. For each of the technique under the comparison, we have selected same number of features i.e 2824. Table 5.2 shows a comparative analysis of the proposed feature selection technique with five other techniques. As presented by Figure 5.3, feature set obtained by the proposed HVLC technique outperforms in the entire six performance criterion as compared to five other techniques belonging to statistical analysis, supervised and unsupervised learning. Accuracy, F-measure and MCC of the HVLC technique are 95%, 95.12%, and 90.11% respectively. The execution time in seconds for the pre-processing & feature extraction phase is 39.3 sec, that of feature selection is 132.5 sec and finally the classification phase took 25.8 seconds for execution.

Table 5.2: A comparative analysis of proposed feature selection with other methods

	Accuracy	Sensitivity	Specificity	Precision	F-measure	MCC
CFS	92.5	95	90	90.4	92.6	85.1
Relief F	92.5	95	90	90.4	92.6	85.1
LLCFS	90	87.5	92.5	92.1	89.7	80.1
SVM-RFE	75	75	75	75	75	50
mRMR	83.7	85	82.5	82.9	83.9	67.5
Proposed (HLVC)	**95**	**97.5**	**92.5**	**92.8**	**95.1**	**90.1**

Further, the ulcer detection system is compared with two other systems and two other classifiers. Suman et. al [3], extracts features from relevant color bands and ulcer images are classified using SVM. Koshy and Gopi [121], extracts contourlet transform and log Gabor based texture features and the classification task is performed using SVM. We have implemented both these prior art on our hardware and using our dataset. The optimized feature set is also experimented with kNN and ensemble classifiers in addition to SVM. The comparative results presented in Table 5.3 shows that the proposed system outperforms the prior art and the best system performance is achieved with the SVM classifier. As seen in Figure 5.4, the proposed CAD system for ulcer detection in CE outperforms the other two systems in terms of accuracy, sensitivity, F-measure, and MCC.

Table 5.3: A comparative analysis of the proposed system with other systems and classifiers

	Accuracy	Sensitivity	Specificity	Precision	F-measure	MCC
Suman et al. [3]	80	60	100	100	75	65.5
Koshy et al. [121]	88.7	77.5	100	100	87.3	79.5
HVLC+KNN	88.7	77.5	100	100	87.3	79.5
HVLC+ Ensemble	93.7	87.5	100	100	93.3	88.1
HVLC+SVM	**95**	**97.5**	**92.5**	**92.8**	**95.1**	**90.1**

However, it fails to outperform other systems in terms of specificity and precision. The reason for this is that the proposed systems have more false positives as compared to other systems. Approximately 8% of normal cases are misclassified as ulcer cases by the proposed system as compared to the other two systems. Illustration of misclassification is shown in Figure 5.5. If we further tune the model to overcome this then we are likely to misclassify ulcers as normal. So the trade-off is maintained.

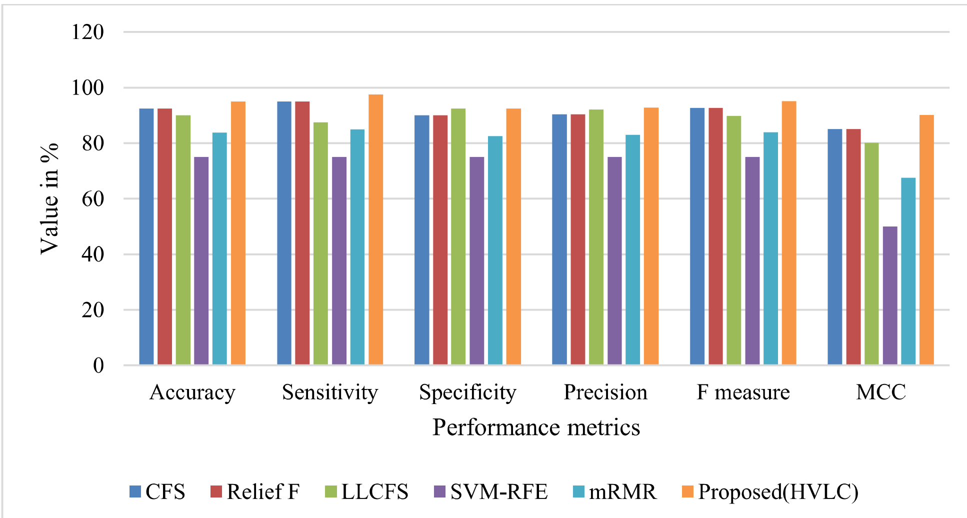


Figure 5.3: Performance comparison of feature selection techniques

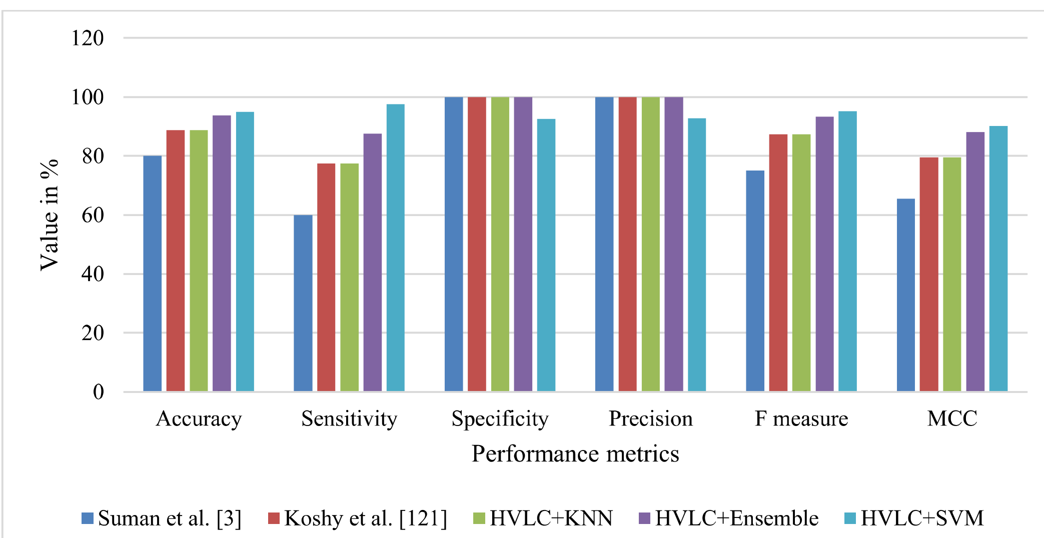
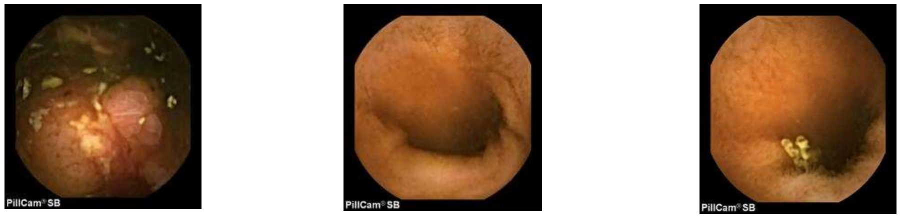


Figure 5.4: Performance comparison of CAD systems



(a) Ulcer (b) Normal (c) Normal detected as ulcer

Figure 5.5: Illustration of misclassifications

5.5 Conclusion

A CAD system can certainly assist the medical expert in providing a fast and accurate diagnosis. Also, it can widely be used as a remote diagnostic tool for geographically distant locations where medical experts may not be available. The proposed system significantly reduces the feature set and yet outperforms on all parameters. The performance of the proposed system performs with accuracy, F-measure, and sensitivity of 95%, 95.12%, and 97.5% respectively.