Neural Computing and Applications
https://doi.org/10.1007/s00521-020-05212-y

S.I. : WORLDCIST'20

Bag of feature and support vector machine based early diagnosis of skin cancer

Ginni Arora[1] · Ashwani Kumar Dubey[2] · Zainul Abdin Jaffery[3] · Alvaro Rocha[4]

Received: 8 June 2020 / Accepted: 13 July 2020

Abstract
Skin cancer is one of the diseases which lead to death if not detected at an early stage. Computer-aided detection and diagnosis systems are designed for its early diagnosis which may prevent biopsy and use of dermoscopic tools. Numerous researches have considered this problem and achieved good results. In automatic diagnosis of skin cancer through computer-aided system, feature extraction and reduction plays an important role. The purpose of this research is to develop computer-aided detection and diagnosis systems for classifying a lesion into cancer or non-cancer owing to the usage of precise feature extraction technique. This paper proposed the fusion of bag-of-feature method with speeded up robust features for feature extraction and quadratic support vector machine for classification. The proposed method shows the accuracy of 85.7%, sensitivity of 100%, specificity of 60% and training time of 0.8507 s in classifying the lesion. The result and analysis of experiments are done on the PH^2 dataset of skin cancer. Our method improves performance accuracy with an increase of 3% than other state-of-the-art methods.

Keywords Skin cancer · Computer-aided detection and diagnosis · Bag of feature · Support vector machine · Classification · SURF

1 Introduction

Skin cancer is a death-causing disease. According to the report of the Center for Disease Control and Prevention [22], somewhere in the years of 2010 and 2020, the extent of more malignancy cases in the USA went around 24% in men to over 1 million cases, and nearly 21% in women to over 900,000 cases for each year. Skin cancer is caused due to an increase in the number of cells in the body [1]. It is also caused by long-term or intense short-term exposure to the sun and due to the genetic issues.

The major categories of skin cancer are [17]:

(i) *Melanoma:* It is considered a vital type of skin cancer that appears as a dark spot on any skin surface. It is usually found over the neck, head, between the shoulders, on the soles of feet and palms or under the fingernails. Fair skin people are commonly affected by this type.
(ii) *Basal cell:* It is the least dangerous type of skin cancer that occurs in places that has more sun exposure. This type starts on the skin cell of the basal layer. It is also common in fair people.
(iii) *Squamous cell:* It is mostly seen in places of the body that is protected from ultraviolet rays like legs. It is common in dark people.

✉ Alvaro Rocha
amrrocha@gmail.com

Ginni Arora
garora@amity.edu

Ashwani Kumar Dubey
dubey1ak@gmail.com

Zainul Abdin Jaffery
zjaffery@jmi.ac.in

[1] Amity Institute of Information Technology, Amity University Uttar Pradesh, Noida, UP 201313, India

[2] Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, UP 201313, India

[3] Department of Electrical Engineering, Jamia Millia Islamia, New Delhi 110025, India

[4] ISEG, University of Lisbon, Rua do Quelhas, N° 6, 1200-781 Lisbon, Portugal

Published online: 21 July 2020

Traditionally, a direct biopsy method was used for the diagnosis of skin cancer. In a biopsy, the affected part is removed and given to pathological laboratories for further process, which is time-consuming, tender and invasive. Thus, computer-aided detection and diagnosis system was introduced to ignore the stated problem. This requires only a skin image for the detection of disease without any physical contact with a body. This process is fast, non-painful and noninvasive. There are majorly four parts for this system.

- Preprocessing
- Segmentation
- Feature extraction
- Classification

Initially, skin image is preprocessed in which processing of the image is done as smoothing, making edges sharp, removing noise, unidentified disrupted pixels and the elimination of long lines or hairs. After the removal of all types of noises, the processed image goes through the phase of image segmentation which segments the affected region of skin into two categories as effected and non-effected region. In the next phase, from the segmented image, unique features are extracted with feature extraction techniques. This phase is most important as if sometimes wrong selection or extraction of features may lead to inaccurate diagnosis. Once features are extracted, classification is applied to classify skin image into a normal skin or cancerous skin [11]. The complete process of computer-aided detection and diagnosis system is shown in Fig. 1.

Feature extraction is a procedure to identify features from an image. These features can be in any number from hundreds to thousands to millions. There are various parameters on which these features are extracted that can be in terms of shape, edge, color, texture, local, global, statistical and many more. But, these feature extractions are completely application dependent. Further, the role of feature vector comes in for identification of lesion through selection of only major features. Currently, many researches have been done in this field that resulted in variety of methods from their combination [13].

Thus, a study [8] based on the segmentation of dermoscopic images with the application of mathematical morphology was carried out. The image is segmented by isolating skin lesions from which shape, texture and color features are extracted. The extracted features are used to identify patients with melanoma and distinguish from non-melanoma cases. The classification is carried out by binary support vector machine (SVM) on images that are acquired from the database of the International Dermoscopy Society for the identification of melanoma.

Besides, a system for correctly classifying the lesion properties such as "Asymmetry" of an image was presented by Chakravorty et al. [7]. Respectively, the segmented image is preprocessed to show various combinations of structure, shape and color of the lesion across its axes. Along with shape-based features, the proposed method detects the lesion various properties based on structure and color which results in robustness in terms of performance.

Bag of features has also used with spatial information in which the distribution of features is considered [15]. For melanoma skin cancer, [4] represents computer-aided diagnosis (CAD) system framework. This system was developed using SVM with a histogram of oriented gradient optimized set (HOG)-based descriptors of lesions. The study also proposes the gray-level co-occurrence matrix (GLCM) and local binary pattern (LBP) utilized for surface highlights done by [6, 10]. The color moments and color histograms are the descriptors that are most in use for feature extraction [18].

Based on the decomposition of dermoscopic images using spatial and color features, a singular approach for feature extraction was presented by Ayadi et al. [2]. Even though the extracted features are prelim, the results obtained in terms of performance were competitive. Also, one of the most advantageous properties of this method is that without applying any preprocessing steps such as color enhancement, hair and noise removal, the performance is still satisfactory. For additional improvement of this method, more features should be incorporated and further decomposition of the image into deeper level results in good pixel value.

Several features were evaluated based on ABCD study [12, 13, 21] in which disease diagnosis is defined by the features extracted from the images. These features are based on area, border, color and diameter of the lesion. It diagnoses the skin disease images as melanoma, suspicious and benign categories. Also, the value of Total Dermatoscopic Value (TDV) supports the decision of diagnosis.

Recent researchers used a bag of feature with scale-invariant feature transform to develop code book learning method [9] achieved an accuracy of 82%. While there are widespread techniques for melanoma cancer, however,

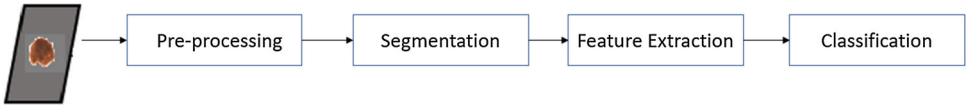

Fig. 1 Computer-aided detection and diagnosis system

other skin cancers recognition is still an unapproachable area for a CAD system. These problems may be dealt with well by way of feature descriptor-based methods for classification of skin lesion [5].

The contribution of this paper is:

1. To develop the computer-aided detection and diagnosis system through the fusion of bag of feature and speeded up robust features as a feature extraction technique.
2. To classify skin lesion into skin cancer or non-cancer class with quadratic support vector machine classifier.
3. To analyze the performance of the proposed method which is compared with other state-of-art methods.
4. To analyze the performance of the proposed method which is compared with various types of SVM classifiers.

The dataset for dermoscopic images is taken from the PH^2 database. This database provides three varieties of images as dermoscopic image, segmented image and region of interest (ROI) as shown in Fig. 2.

The remaining of the paper is systematized as follows: Sect. 2 focuses on the proposed methodology with its framework. Section 3 talks about the implementation, experiments and results. Section 4 is regarding the analysis of the proposed study with its limitations and implications. Finally, the conclusion of the paper is presented in Sect. 5.

2 Methods

This paper proposed the fusion of bag of feature (BoF) and speeded up robust features (SURFs) as a local feature descriptor with quadratic support vector machine as a classifier. BoF is a simple technique in terms of use and gives high performance. In this technique, the local features are grouped without following any sequence. The visual vocabulary is designed by extracting local features and vector quantized from the individual input image. These image features are the local features only, and the new features are added to the codebook. Histogram is used for the representation of these images.

Fig. 2 Skin lesion on left, segmented skin lesion in center and ROI on right [19]

SURF is majorly composed of two steps: feature extraction and feature description. For feature extraction, the Hessian matrix is used, and for feature description, wavelet is used as an orientation for the interest points. It is preferred over other feature descriptors such as SIFT due to its fast computation.

The flowchart for the proposed method is shown in Fig. 3, which mainly consists of two phases as follows:

1. Training (i.e., image set, feature extraction and classification)
2. Testing (i.e., input image, feature extraction and classification).

2.1 Training

The training phase includes three phases: input image, feature extraction and classification.

2.1.1 Input image

In this, setoff segmented training images are taken as label input images. The training images are divided into two classes as cancerous and non-cancerous images. A total of 100 images are used for training purposes.

2.1.2 Feature extraction

In feature extraction, fusion of the BoF method with SURF is applied. SURF algorithm is divided into three phases [14].

1. It selects the point of interest of an image using Hessian matrix: $H = \begin{bmatrix} L_{xx} & L_{xy} \\ L_{xy} & L_{yy} \end{bmatrix}$ where L_{xx} is a second derivative convolution of a Gaussian with the image at the point.
2. It uses the SURF neighborhood descriptor that describes pixel intensity distribution with a neighbor point of interest. The descriptor should be distinctive. The process [23] is shown in Fig. 4.
3. Comparison between features was obtained from images and matched the pairs.

BoF uses SURF features extraction that extracts the interest point of an image. It represents an image as an order less collection of local features. The extracted point of interest is stored in a feature vector. For quantification of the features, K-means clustering is used and by default, the size of the cluster used by BoF is 500 which can be increased or decreased. K-means clustering generates codebooks from k-means models and compares them with centroids in a codebook for vector quantization.

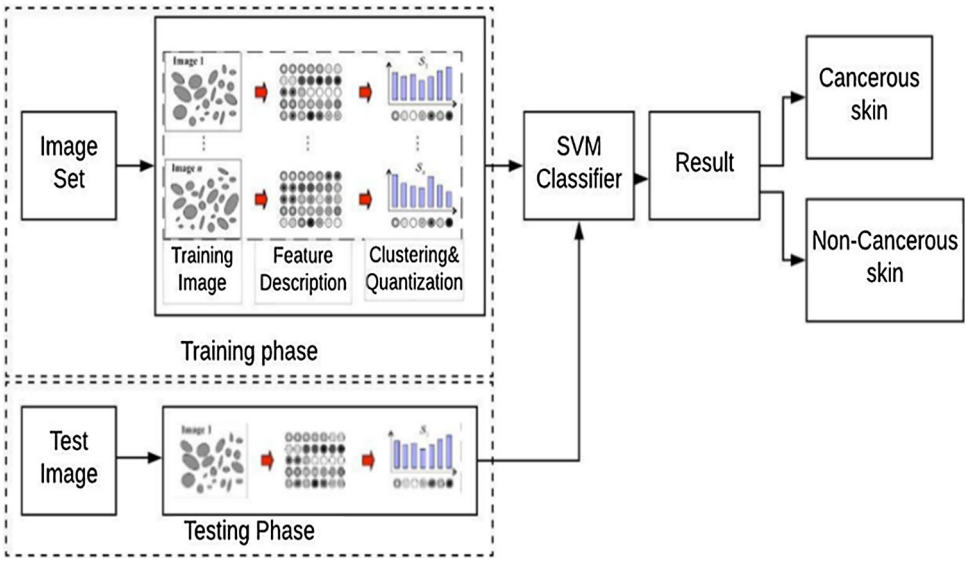

Fig. 3 Flowchart of the proposed method

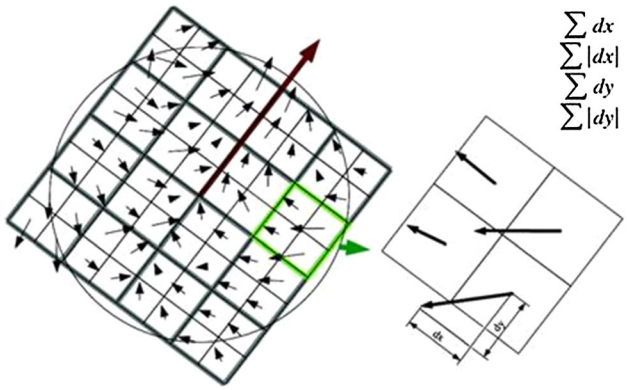

Fig. 4 SURF descriptors

2.1.3 Classification

Support vector machine (SVM) is a widespread technique for pattern recognition as it supports data with large dimensions and provides a better way to generalize the classes. Initially, SVM classifies images on a binary basis that is in two classes, but later, it is extended to support multi-class classification. In our work, the binary SVM classification is implemented for classification into cancerous or non-cancerous. The logic behind SVMs is mapping of original facts factors from the entry space to a high-dimension, or limitless-dimension, characteristic space such that the detection problem turns into easier inside the function area. The mapping is carried out through a satisfactory preference of mathematical functions known as the kernel [3].

The data set for training is assumed as $\{x_i, y_i\}^N i = 1$, with $x_i \in R^d$, where x_i = input vectors and $y_i \in \{-1, +1\}$ is the class label of x_i. SVMs map the d-dimension enter vector x from the input area to the d_h-dimension characteristic space with the use of a (non)linear feature. It computes the optimum separation hyperplane which classifies complete data keeping in view of the outermost data points. The linear separation hyperplane is computed as in Eq. 1:

$$f(x) = w^T x + b \tag{1}$$

where w and b are the weight vectors and bias, respectively. The generalization ability needs to be increased for SVM to compute the optimum separation hyperplane. Sometimes, if the training data are nonlinear, then classifier generalization ability is not so high, but then also hyperplanes can be optimally determined. To increase the linear ability, the feature space can be mapped to the original input space. The feature space for optimum separation hyperplane is given by Eq. 2:

$$f(x) = w^T \varphi(x) + b \tag{2}$$

where $\varphi(x)$ is the nonlinear vector function.

For any test data x, the decision function is given by Eq. 3 [3]:

$$f(x) = \text{sign}\left(w^T \varphi(x) + b\right) \tag{3}$$

2.2 Testing

The testing of the trained model is done for performance assessment of the proposed approach. Once the model is trained, cancerous and non-cancerous random images are taken as an input. Again, features are extracted as done in the training process. With the help of yfit = trainedModel.predictFcn (T), the model predicts whether it is cancerous or non-cancerous.

Algorithm of the proposed method:

Step 1: Train the model using the segmented image data set.
Step 2: Detect and extract features using the fusion of bag of features with SURF as a local feature descriptor.
Step 3: Compute feature vector and clustering using K-Means.
Step 4: Represent each trained image through a histogram.
Step 5: Train SVM classifier with a computed training image histogram.
Step 6: Test the model using test image.
Step 7: Repeat the previous steps from step 2 to step 4 for the test image.
Step 8: Predict the samples and determine their classification result according to their performance score of the model.

3 Results

A total of 100 images are acquired for the evaluation of performance for the proposed system, among which 50 are the cancerous and 50 are non-cancerous images. These sets of images are the segmented images that mean the lesion part is separated using PH^2 database. The proposed system is implemented through MATLAB.

3.1 Performances of fused bag of features

Table 1 illustrates a performance matrics that prompts improvement in the classification of skin cancer. The outcomes are shown with other types of classification models like a fine tree, linear SVM and linear discriminant performance against quadratic SVM. The performance is measured based on accuracy, specificity, sensitivity, speed and time. The performance matrics of the proposed approach with other techniques shows that proposed quadratic SVM gives 85.7% accuracy which is better than the other classifiers. Its prediction speed and training time are also better than the existing techniques. It has 100% sensitivity and 60% specificity. The proposed system gives satisfactory results in comparison with other classifiers.

Figure 5 shows the confusion matrix of 4 models as model1 for fine tree classifier, model2 for linear SVM classifier, model3 for linear discriminant classifier and model4 for quadratic SVM classifier. Model4 is giving 100% true positive and only 40% as a false negative prediction in comparison with other classifiers.

Receiver operating characteristics (ROCs) and area under curve (AUC) represents a false positive rate against true positive rate. These are used as performance measures. The higher the AUC means, the better the prediction model is for classification. ROC is the probability curve that means the curves for two classes should not overlap for better classification. In Fig. 6, it shows the ROC–AUC as model1 for fine tree classifier, model2 for linear SVM classifier, model3 for linear discriminant classifier and model4 for quadratic SVM classifier. The model4 is having high AUC and better ROC in comparison with other classifier models.

According to the sensitivity and specificity, overall classifier performance is evaluated. Sensitivity shows the proportion of the genuine fine projection to the sum of a wide variety of the doubtful lump area within the dataset. The capability to properly discover disease is sensitivity and is given by way of [16]:

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) \times 100\% \quad (3)$$

The specificity depicts the proportion of the incorrect tremendous projections to the whole range of the doubtful lump areas inside the image dataset. It is the capability to keep away from wrongly classifying normal tissue as ailment and is given by way of:

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}) \times 100\% \quad (4)$$

A ratio of the whole range of efficiently labeled regions and the entire range of regions defines the basic accuracy. It is represented as:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN}) \quad (5)$$

If the model correctly predicts the class and the actual result is also positive, then the result is TP. In the same way, if both prediction and the actual result are negative, then the classification result is TN. FP is vice versa of FN.

Table 1 Performance metrics

Measures	Fine tree	Linear SVM	Linear discriminant	Quadratic SVM
Accuracy %	42.9	71.4	78.6	**85.7**
Prediction speed obs/s	~200	~200	~200	**~200**
Training time s	0.92381	0.91318	0.88807	**0.85075**
Sensitivity %	44	100	99	**100**
Specificity %	40	20	60	**60**

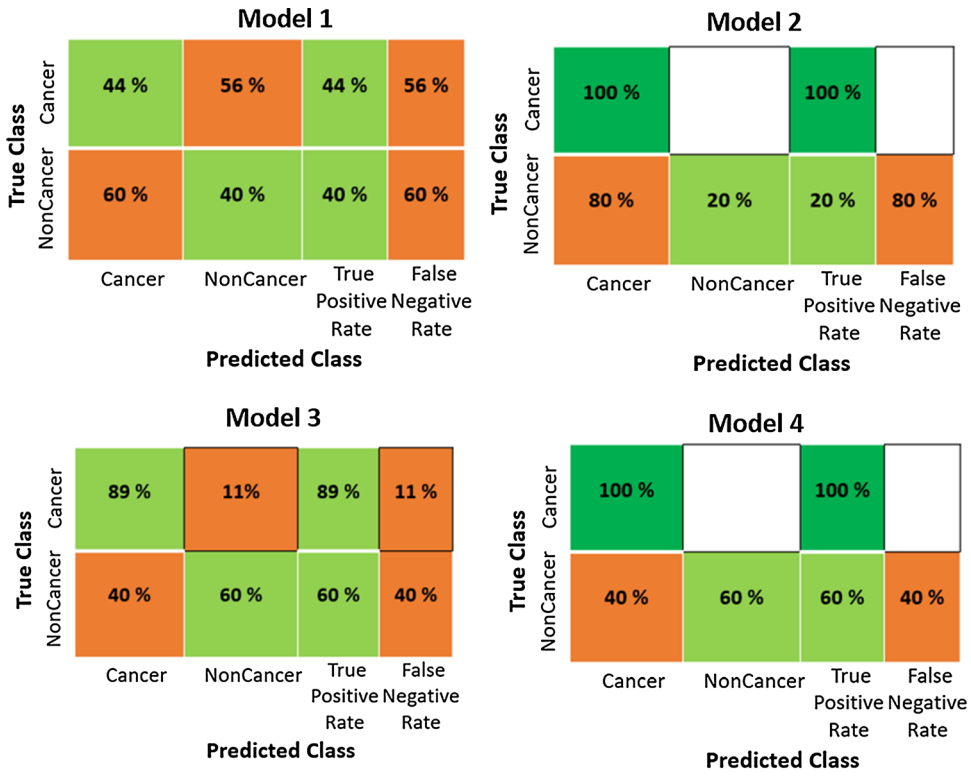

Fig. 5 Confusion matrix of model1 (fine tree), model2 (linear SVM), model3 (linear discriminant) and model4 (quadratic SVM)

3.2 Comparison with current state-of-the-art methods

The proposed method is compared with the work of various researchers inside the field of computer-aided detection and diagnosis system for skin cancer. The work is not able to be directly comparable with researchers due to variety in dataset and measures for validation. The results are shown in Table 2.

In the work, both local and spatial features are identified with the fusion of BoF with SURF and achieve 85.7% accuracy in identifying a lesion as cancer or non-cancer classification, highlighted in bold in Table 2. In the table, all methods using a similar procedure are compared with the proposed skin disease. These strategies were dependent on the investigation of the dermoscopic picture used to depict the examples of skin diseases or its hues.

4 Discussion

Skin cancer is one of the major concerns of almost every human being varying from type melanoma to non-melanoma. Every patient with type melanoma is at high risk in comparison with other skin cancers. Here, we aimed at identifying the type of skin lesion whether a lesion belongs to cancer or non-cancer class. If a patient comes to know that the lesion is of cancer class, then there is a need to visit the doctor; otherwise, normal medication can work. On the other hand, if a doctor comes to examine at preliminary stage that a lesion is of cancer class, then proper diagnosis can be proceed, otherwise not a serious case.

The major role is played by feature extraction during the complete process of skin lesion detection and diagnosis. There are various studies [5, 12, 18, 20, 21] in the literature that used color, texture, filters, wavelets or HSV for extracting relevant features. But the fusion of compatible and efficient techniques are of utmost importance. Recent studies have shown the role of SURF feature descriptor having good performance for classification.

This study evaluated the efficiency of computer-aided detection and diagnosis system when used for classification of skin lesion with the fusion of BoF and SURF using quadratic support vector machine classifier. The results confirm that this approach adds value in health care sector for accurate diagnosis of lesion as cancer or non-cancer. The advantage of using BoF with SURF is that computation is fast. Also, it reduces training time when used with quadratic support vector machine.

A direct comparative study between the different types of classifier is shown in this study to evaluate the

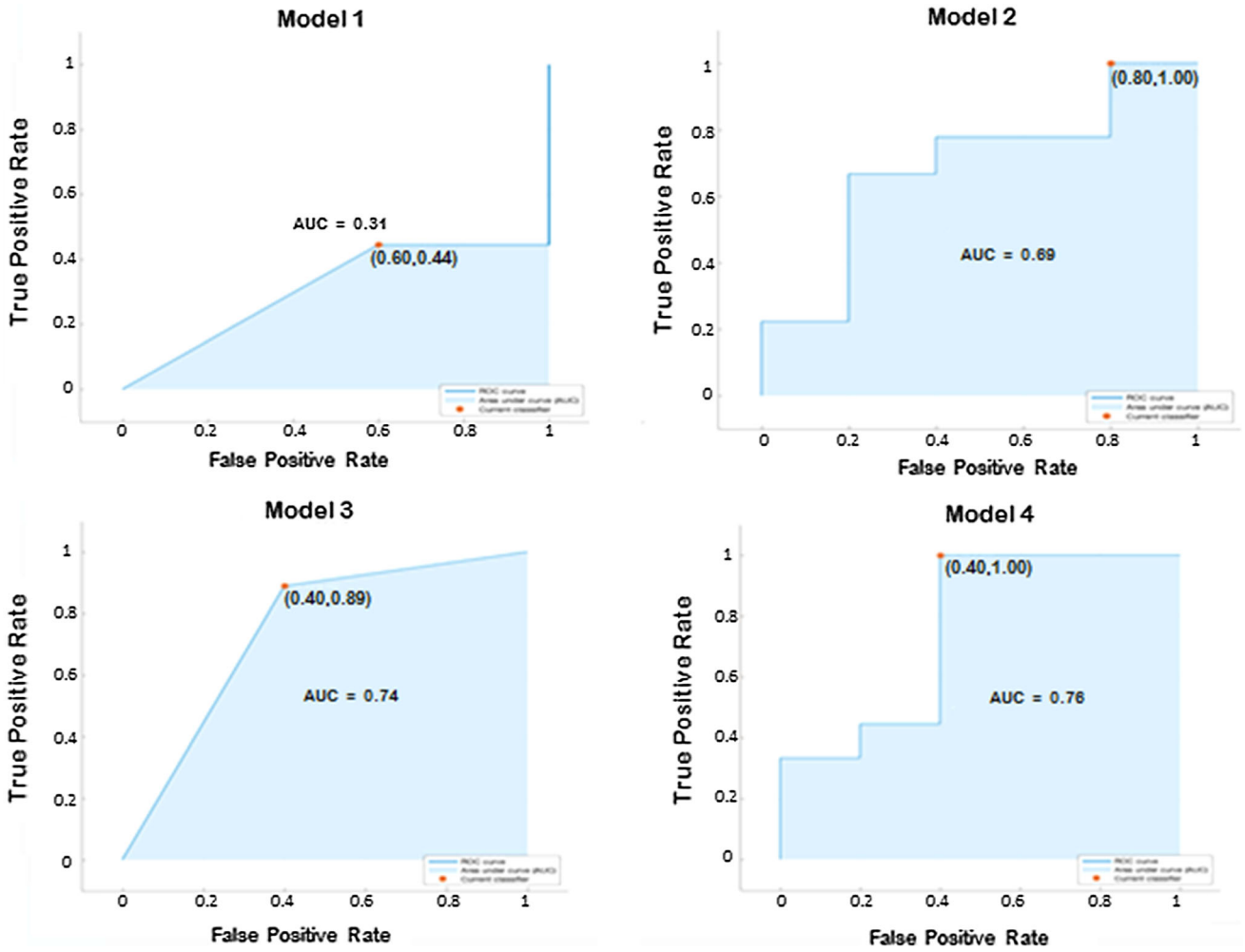

Fig. 6 ROC–AUC curve of model1 (fine tree), model2 (linear SVM), model3 (linear discriminant) and model4 (quadratic SVM)

Table 2 Correlation of the proposed method with the current advancement

Authors	Method	No. of classes	Accuracy (%)
Ballerini et al. [5]	Color and texture, K-NN	5	74
Situ et al. [18]	Color histogram, Gabor filter, BoVW, K-NN	2	82.2
Mikos et al. [12]	GLCM, PNN neural network	2	69.5
Wadhawan et al. [21]	Color histogram, Haar wavelet, SVM	2	76.4
Upadhyay and Chandra [20]	GLOH, HSVcolor, BoVW, SVM	6	78
Proposed method	BoF, SURF, SVM	2	**85.7**

performance of fused features with sub-variety of support vector machine classifier, where the quadratic classifier supports the features with less training time and high accuracy. The limitation of this study is the use of less number of classes of classification. The more number of classes or types of different skin diseases can improve accuracy and efficiency of the system.

5 Conclusion

This work has shown the fusion of bag of feature with speeded up robust features as feature extraction technique, and the classification is performed by using a quadratic support vector machine. This has given the growth of computer-aided detection and diagnosis system for

classifying a skin lesion as cancer or non-cancer. The proposed approach performed investigations on two diverse class tests of skin lesions. Besides, it is quick, precise and cost-effective which simply defines a lesion with an accuracy of 85.7%. The outcomes acquired in the proposed approach are very encouraging when contrasted with the current state-of-the-art and achieved 3% increase in performance measure. In the future, the proposed approach might be applied to various modalities of medical pictures identified with numerous other fundamental organs. Besides, a few alterations will be required in the proposed strategy with the goal that the level of mutilation might be considered, and it will assist with perceiving the more prominent number of class tests for a skin lesion.

References

1. Arora G, Dubey AK, Jaffery ZA (2018) Performance measure based segmentation techniques for skin cancer detection. In: Panda B, Sharma S, Roy N (eds) Data science and analytics. Singapore, Springer, pp 226–233
2. Ayadi W, Elhamzi W, Charfia L, Atri M (2019) A hybrid feature extraction approach for brain MRI classification based on bag-of-words. J Biomed Sig Process Control 48:144–152
3. Azad R, Babak A, Iman TK (2014) Optimized method for iranian road signs detection and recognition system. arXiv preprint arXiv:1407.5324
4. Bakheet S, An SVM (2017) Framework for malignant melanoma detection based on optimized HOG features. Open Access J Comput Sci 5:1–13
5. Ballerini L, Fisher RB, Aldridge B, Rees J (2013) A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In: Celebi M, Schaefer G (eds) Color medical image analysis. Dordrecht, Springer, pp 63–86
6. Bhuiyan A, Azad I, Uddin K (2013) Image processing for skin cancer features extraction. Int J Sci Eng Res 4(2):1–6
7. Chakravorty R, Liang S, Abedin M, Garnavi R (2016) Dermatologist-like feature extraction from skin lesion for improved asymmetry classification in PH2 database. In: Conf proc IEEE eng med-biol soc, pp 3855–3858
8. Chatterjee S, Dey D, Munshi S (2015) Mathematical morphology aided shape, texture and colour feature extraction from skin lesion for identification of malignant melanoma. In: Conf proc IEEE cond assess techniques in electrical systems, pp 200–203
9. Kai H, Xiaorui N, Si L, Yuan Z, Chunhong C, Fen X, Wanchun Y, Xieping G (2019) Classification of melanoma based on feature similarity measurement for codebook learning in the bag-of-features model. J Biomed Sig Process Control 51:200–209
10. Khakabi S, Wightona P, Leea T, Atkins M (2012) Multi-level feature extraction for skin lesion segmentation in dermoscopic images. In: Medical imaging 2012: computer aided diagnosis. https://doi.org/10.1117/12.911664
11. Mane S, Shinde S (2018) Skin cancer detection using dermoscopy images. Int J Comput Eng Appl 12:1–7
12. Mikos E, Sioulas I, Sidiropoulos K, Cavouras D (2012) An android-based pattern recognition application for the characterization of epidermal melanoma. J Sci Technol 7:67–72
13. Murumkar OS, Gumaste PP (2015) Feature extraction for skin cancer lesion detection. Int J Sci Eng Tech Res 4:1645–1650
14. Nasr S, Bouallegue K, Shoaib M, Mekki H (2017) Face recognition system using bag of features and multi-class SVM for robot applications. In: Conf proc on control, automation and diagnosis, pp 263–268
15. Passalis N, Raitoharju J, Tefas A, Gabbouj M (2019) Adaptive inference using hierarchical convolutional bag-of-features for low-power embedded platforms. In: Conf proc IEEE image processing, pp 3048–3052
16. Singh L, Jaffery Z (2018) Computer-aided diagnosis of breast cancer in digital mammograms. J Biomed Eng Technol 27:233–246
17. Singhal E, Tiwari S (2018) Skin cancer detection using artificial neural network. Int J Adv Res Comput Sci 6:149–157
18. Situ N, Yuan X, Chen J, Zouridakis G (2008) Malignant melanoma detection by bag-of-features classification. In: Conf proc IEEE eng in med-bio sci, pp 3110–3113
19. Teresa M, Pedro MF, Jorge M, Andre RSM, Jorge R (2013) PH^2—a dermoscopic image database for research and benchmarking. In: Conf proc IEEE eng med-biol soc, pp 5437–5440
20. Upadhyay PK, Chandra S (2019) An improved bag of dense features for skin lesion recognition. J King Saud Univ Comput Inf Sci. https://doi.org/10.1016/j.jksuci.2019.02.007
21. Wadhawan T, Situ N, Rui H, Lancaster K, Yuan X, Zouridakis G (2011) Implementation of the 7-point checklist for melanoma detection on smart hand held devices. In: Conf proc IEEE eng in medicine and biological sci, pp 3180–3183
22. Weir HK, Thompson TD, Soman A, Moller B, Leadbetter S (2015) The past, present, and future of cancer incidence in the United States: 1975 through 2020. Cancer 121:1827–1837
23. Woodruff C (2013) Computer vision: feature detection and matching. https://courses.cs.washington.edu/courses/cse576/13sp/projects/project1/artifacts/woodrc/index.htm. Accessed 23 Jan 2020