Mathematical Biosciences and Engineering

MBE, 19(6): 5850–5866.
DOI: 10.3934/mbe.2022274
Received: 27 December 2021
Revised: 23 March 2022
Accepted: 29 March 2022
Published: 07 April 2022

http://www.aimspress.com/journal/MBE

Research article

Recognition of bovine milk somatic cells based on multi-feature extraction and a GBDT-AdaBoost fusion model

Jie Bai[1,2], Heru Xue[1,2,*], Xinhua Jiang[1,2] and Yanqing Zhou[1,2]

[1] College of Computer and Information Engineering, Inner Mongolia Agricultural University, Hohhot 010018, China
[2] Inner Mongolia Autonomous Region Key Laboratory of Big Data Research and Application of Agriculture and Animal Husbandry, Hohhot 010018, China

* **Correspondence:** Email: xuehr@126.com.

Abstract: Traditional laboratory microscopy for identifying bovine milk somatic cells is subjective, time-consuming, and labor-intensive. The accuracy of the recognition directly through a single classifier is low. In this paper, a novel algorithm that combined the feature extraction algorithm and fusion classification model was proposed to identify the somatic cells. First, 392 cell images from four types of bovine milk somatic cells dataset were trained and tested. Secondly, filtering and the K-means method were used to preprocess and segment the images. Thirdly, the color, morphological, and texture features of the four types of cells were extracted, totaling 100 features. Finally, the gradient boosting decision tree (GBDT)-AdaBoost fusion model was proposed. For the GBDT classifier, the light gradient boosting machine (LightGBM) was used as the weak classifier. The decision tree (DT) was used as the weak classifier of the AdaBoost classifier. The results showed that the average recognition accuracy of the GBDT-AdaBoost reached 98.0%. At the same time, that of random forest (RF), extremely randomized tree (ET), DT, and LightGBM was 79.9, 71.1, 67.3 and 77.2%, respectively. The recall rate of the GBDT-AdaBoost model was the best performance on all types of cells. The F1-Score of the GBDT-AdaBoost model was also better than the results of any single classifiers. The proposed algorithm can effectively recognize the image of bovine milk somatic cells. Moreover, it may provide a reference for recognizing bovine milk somatic cells with similar shape size characteristics and is difficult to distinguish.

Keywords: bovine milk somatic cell; feature extraction; GBDT-AdaBoost fusion model; classifier

1. Introduction

Milk somatic cell number is an important index that reflects milk quality and cow health. An excessive somatic cell count will destroy the nutritional components in milk and indicate the occurrence of mastitis in dairy cows [1]. The somatic cells in milk are mainly white blood cells (like lymphocytes, macrophages, and neutrophils), accounting for about 99% of all somatic cells, with a small number of epithelial cells shed from mammary tissues, accounting for about 1% [2]. Mastitis can lead to a decrease in milk yield and economic loss and lead to changes in milk composition and nutritional composition. The number of various cells in milk will change depending on the infection degree of mastitis [3].

The commonly used detection methods for milk somatic cells are mainly divided into direct and indirect. The direct methods mainly include microscopy and fluorescence photoelectric counting instruments [2]. The indirect methods such as the California cell assay and Wisconsin mastitis test are accurate. However, the degree of automation is not high, and the workload and the cost of the measuring equipment are high [4,5].

In order to overcome the defects of the above methods, machine vision technology was introduced into cell recognition, mainly using dyed cells after a digital microscope color image recognition analysis. In terms of cell feature extraction, shape features [6], texture features [7], and color features [8] are usually used as the recognition features of cell images. In order to fully express cell information and further improve the accuracy of cell recognition, feature fusion is widely used in cell image recognition [9,10]. The construction of appropriate classifiers is another key problem in recognizing different cell image categories based on the cell extracted features. It is of great significance to study automatic recognition algorithms of milk cell microscopic images to monitor dairy cows' health status and ensure the quality of dairy products. Still, there is little research on milk somatic cell image classification and recognition. Gao et al. [11] used bi-directional two-dimensional principal component analysis to propose a rapid and accurate method to detect bovine mastitis. Gao et al. [12,13] also suggested a Relief F algorithm that could extract the features of milk somatic cells for classification. Zhang et al. [14] developed an algorithm based on the random forest method to achieve a recognition of 96%.

The machine learning methods commonly used in the field of cell recognition include support vector machine [15–17], K neighbor [18,19], random forest [20,21], naïve Bayes [5,22], logistic regression [23], extreme learning machine [24], and neural network [25,26]. Those methods can be applied to identify and classify various types of cells [27–29]. Still, those recognition methods each have perks and limitations. Because the image of milk cells contains a large amount of milk fat, milk protein, and cell debris, the image itself is complicated to interpret. The mentioned recognition methods have harsh requirements on the data set. When these classifiers are used for direct classification and recognition, the problem of weak generalization will appear. Therefore, in order to overcome the above problems, this study combined with the actual situation of milk somatic cell sample data and used for reference the ensemble learning method of recent popular research by scholars, that is, to complete high-precision classification tasks by integrating multiple weak classifiers into strong classifiers [30]. Common ensemble learning methods include parallel ensemble bagging [31], stacking [32], and serialization ensemble boosting [33]. AdaBoost is an adaptive boosting algorithm; compared with stacking and bagging integration, AdaBoost trains an optimal set of weak classifiers. This is done by adjusting the weight of samples and weak classifiers, improving generalization ability, obtaining higher

prediction accuracy, and reducing model overfitting. At present, this method has been widely applied in agricultural image processing, remote sensing image water information extraction, and fire smoke detection [34–36] but rarely in cell image classification and recognition [37].

Therefore, this paper proposes an algorithm based on multi-feature extraction and gradient boosted decision trees (GBDT)-Adaboost fusion model to recognize different types of milk cells. Firstly, according to the characteristics of milk cells, the color morphology and texture features were extracted and fused. Secondly, based on the extracted features, they were input into the fusion model designed in this paper for recognition. Finally, the effectiveness of the proposed method was verified by comparing algorithms. The results of this study could provide an efficient method for the identification and classification of milk somatic cells, which could help improve the automation of bovine mastitis detection.

2. Materials and methods

2.1. Sample image acquisition

The samples used in this paper were from the basic veterinary Laboratory of Veterinary College, Inner Mongolia Agricultural University. The milk somatic cell TIF images were 158 color images at 400× magnification under the microscope and with a rate of 2048*1536 pixels. From the 158 large color images, features from single-cell images were extracted from the large color images. Through veterinary pathology expert appraisal, the individual cells were classified into the four kinds of milk somatic cells, for a total of 392: epithelial cells (EPI), n = 65; lymphoid cells (LYP), n = 112; macrophage (MΦ), n = 81; neutrophils (NG), n = 134. Representative images are shown in Figure 1, where "1" represents MΦ, "2" represents EPI, "3" represents NG, and "4" represents LYP. EPI cells have a large size and a round or oval nucleus. MΦ cells are spherical with a diameter of 10-20 μm, and their nuclei are oval, kidney- and horseshoe-shaped, with abundant cytoplasm. LYP cells are spherical and can be divided into three types (large, medium, and small) according to their volume. Large LYPs are uncommon. Medium LYPs have a diameter of 9–12 μm, with rich cytoplasm and an oval- or kidney-shaped nucleus. Small LYPs represent the largest LYP number, accounting for about 90% of the total number of LYPs, with a diameter of 5–8 μm and a round nucleus, often with small depressions on one side little cytoplasm. NG cells are spherical and with a diameter of 9–12 μm. The nuclei are of various shapes. The nuclei are mostly trilobal. Some are sausage-shaped (called rod-shaped nuclei). At the same time, some are lobulated, with filaments connecting between the leaves (called leaf nucleus).

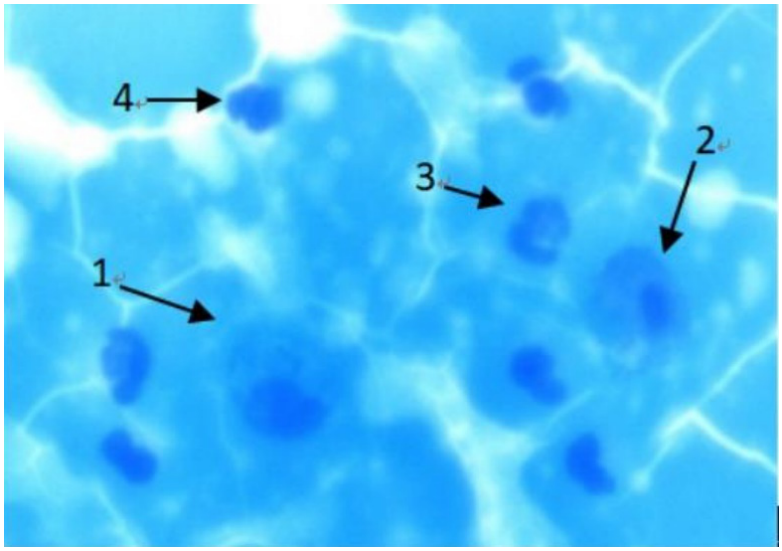

Figure 1. Image of cow milk somatic cells with a 400× magnification.

2.2. Image preprocessing and segmentation

In order to avoid the interference factors such as shadow brightness color saturation, cell fragments, and impurities, the original images were preprocessed, as shown in Figure 2. First, a calculation was used to contrast the milk somatic cell images in different color spaces. Then, the RGB color space was selected to deal with the cell image gray level of the space, and 3 × 3 median filter template and Gaussian filter were used for noise reduction processing. The most between-cluster variance method (OTSU) [38] was used for image binarization processing, using mathematical morphology closed operation noise to remove unnecessary holes in the cells and the open operation to remove slight noise in the image to optimize the boundary of the cell image. Finally, the k-means algorithm segmented the cell region from the image.

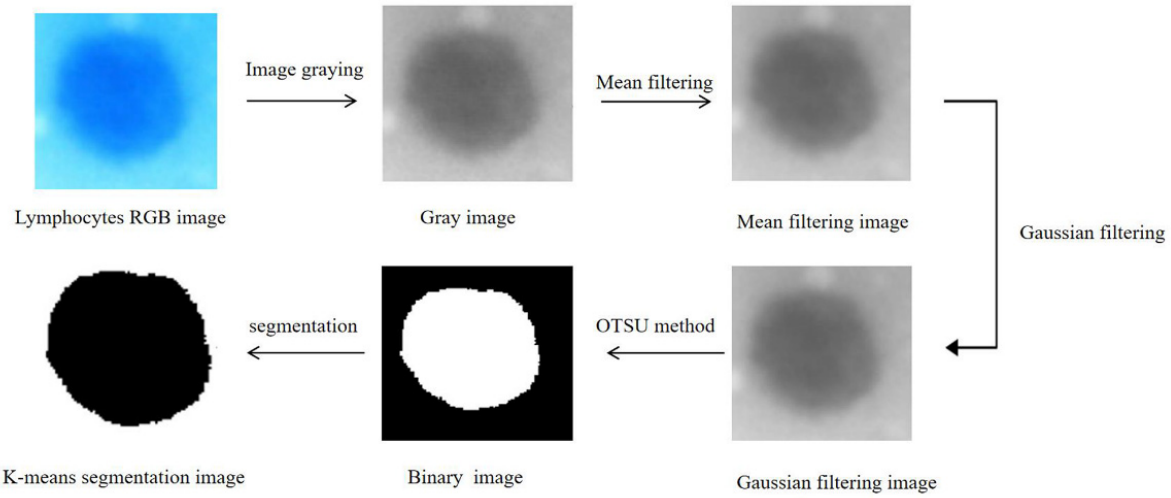

Figure 2. Flowchart of image preprocessing and segmentation.

2.3. Image preprocessing and segmentation

2.3.1. Color characteristics

After the cells are stained, the nucleus and cytoplasm will turn into different colors. In this paper, four features of cell images in gray space, including mean value, variance, energy, and contrast, were extracted as the characteristic color parameters.

2.3.2. Morphological characteristics

The morphology of the milk cells was observed under a microscope. Each type of cell image contains characteristic geometric information such as area, shape, number of lobes, and concave rate and proportion. The four types of cells have certain morphological differences degrees. In this paper, six geometric features and seven invariable moment features were extracted as 13 morphological features of somatic cells.

The parameters of geometric features contain much important information. In this paper, the area, circumference, and roundness of cells and nuclei were calculated as the main features [39]. The cell area was obtained by calculating the sum of the lengths of all horizontal line segments in the cell area [3]. The specific formula is shown in Eq (1).

$$A = \sum_{i=1}^{n}(y_{i2} - y_{i1}) \tag{1}$$

The cell perimeter was obtained by calculating the perimeter of the cell region boundary outline [3], and the formula was:

$$P = M_1 + \sqrt{2M_2} \tag{2}$$

Studies have found that roundness represents the complexity of the nucleus [3]. It is usually obtained by calculating the ratio of the roundness of the nucleus to the roundness of the whole cell, as shown in Eqs (3) and (4).

$$C = \frac{P^2}{4\pi A} \tag{3}$$

$$Y = \frac{C_n}{C_c} \tag{4}$$

The invariant moments were calculated through statistical moments [40]. For cell images, the experimental effect of using the nucleus is better than using the whole cell [41]. Therefore, this paper calculated the invariant moment characteristics of the nucleus region in the image of milk somatic cells to analyze it. The (p + q) order statistic of the invariant moment was defined as:

$$m_{pq} = \sum_x \sum_y x^p y^q Im(x, y) \tag{5}$$

$$\mu_{pq} = \sum_x \sum_y (x - x_c)^p (y - y_c)^q Im(x, y) \tag{6}$$

In the above formula, the center of mass is the coordinates of the gray center of the region, as shown in Eq (7). The gray image represents the sum of gray values and represents two first-order moments:

$$x_c = \frac{m_{10}}{m_{00}}, y_c = \frac{m_{01}}{m_{00}} \tag{7}$$

In order to ensure scale invariance, the normalized central moment was calculated. Seven invariant moment features were constructed by a linear combination of second and third-order central moments:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{r}}\ ,\ \gamma = \frac{p+q}{2} + 1, p + q = 2,3,4,\ldots \tag{8}$$

2.3.3. Texture features

In this paper, a gray-level co-occurrence matrix (GLCM) [42] and local binary pattern (LBP) [43] were used to extract texture features of somatic cell images. GLCM is a second-order statistic of image brightness change, which reflects texture feature information by calculating the joint probability density of two types of positions. In this paper, the gray level of the image was set as 16. In order to ensure the rotation invariance of feature parameters, the matrices at (0, 45, 90, 135) four different angles were calculated respectively, using contrast CON, otherness DISL, HOMO, ENT, ASM, COR. There were six statistics and 24 characteristic values to extract the texture information of milk somatic cell images. The following is the calculation formula of the six statistics. Row I and column J represent the normalized gray co-occurrence matrix, and L is the gray level progression.

$$CON = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1}(i-j)^2\hat{P}_\delta(i,j) \tag{9}$$

$$DISL = \sum_{i=0}\sum_{j=0}\hat{p}_\delta(i,j)|i-j| \tag{10}$$

$$HOMO = \sum_{i,j}\frac{\hat{p}_\delta(i,j)}{1+|i-j|} \tag{11}$$

$$ENT = -\sum_{i=0}^{L-1}\sum_{j=0}^{L-1}\hat{P}_\delta(i,j)\log\hat{P}_\delta(i,j) \tag{12}$$

$$ASM = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1}\hat{P}_\delta^2(i,j) \tag{13}$$

$$COR = \frac{\sum_{i=o}^{L-1}\sum_{j=0}^{L-1}ij\hat{P}_\delta(i,j)-\mu_x\mu_y}{\sigma_x^2\sigma_y^2} \tag{14}$$

$$\mu_x = \sum_{i=0}^{L-1} i \sum_{j=0}^{L-1}\hat{P}_\delta(i,j) \tag{15}$$

$$\mu_y = \sum_{i=0}^{L-1} i \sum_{j=0}^{L-1}\hat{P}_\delta(i,j) \tag{16}$$

$$\sigma_x^2 = \sum_{i=0}^{L-1}(i-\mu_x)^2 \sum_{j=0}^{L-1}\hat{P}_\delta(i,j) \tag{17}$$

$$\sigma_y^2 = \sum_{i=0}^{L-1}(i-\mu_y)^2 \sum_{j=0}^{L-1}\hat{P}_\delta(i,j) \tag{18}$$

LBP is a feature algorithm used to describe the local texture feature information of images [43]. The extraction method of the original LBP operator is simple. First, the center pixel of the 3 × 3 window is taken as the threshold value and compared with the adjacent pixels. The gray value with a large value was set to 1; otherwise, it was set to 0. It gives eight binary numbers made up of ones or zeros. The formula of the LBP feature extraction operator is shown in Eq (19)

$$LBP(x_c, y_c) = \sum_{i=0}^{i=8} s(p_i - p_c) \times 2^p \tag{19}$$

where *LBP* (x_c, y_c) is the central pixel, p_i is the gray value of adjacent pixels, p_c is the gray value of the central pixel, and p is the number of neighboring points. The function expression is shown in Eq (20):

$$s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{20}$$

An LBP operator can produce several different binary modes with the increase of sampling points, the number of binary modes, and a relatively sparse histogram. Therefore, the traditional LBP operator for dimension reduction makes as much as possible in fewer data to describe the image information. This paper adopted the equivalent of the LBP operator for dimension reduction [44]. Hence, 256 histogram statistics were obtained by LBP calculation, and 59 dimensional LBP features were finally obtained after dimensionality reduction.

2.4. Identification of milk somatic cells based on GBDT-Adaboost fusion model

GBDT is an integrated learning algorithm based on a series of ideas [45]. The core idea was to train the new weak classifier through the residual of the current model. Each training got a negative gradient value of the loss function. This value was taken as the approximate value of the residual. Finally, the result of each weak classifier was weighted and summed to get the final classifier. In this paper, the Light Gradient Boosting Machine (LightGBM) was selected as the weak classifier of the GBDT model by comparing many experiments. The grid search method was used to optimize the parameters of the LightGBM model. At the same time, the optimal LightGBM model was obtained by using the method of 10 folds cross-verification to calculate its hyperparameters. The specific algorithm flow of GBDT is as follows [46]:

Step 1: Initialize the weak classifier, assuming that the training set is: $\{(X_1, y_1), (X_2, y_2), \cdots, (X_n, y_{n1})\}$, number of iterations and loss function, $L(y_i, \gamma)$, $y_i = \{-1, 1\}$, Initializing weak classifiers:

$$f_0(x) = argmin_r \sum_{i=1}^{N} L(y_i, \gamma) \tag{21}$$

Step 2: m = 1, 2, …, M, Perform the following steps:

1) To: i = 1, 2, …, n, Calculate approximate residuals:

$$\gamma_{im} = -\left[\frac{\partial L(y_i, f(X_i))}{\partial f(X_i)}\right] f(x) = f_{m-1}(x) \tag{22}$$

2) The approximate residual value is used as training data to fit into a regression tree, which gives the leaf node domain R_{jm}, $j = 1, 2, \ldots, J_m$
3) For each node $j = 1, 2, \ldots, J_m$, calculate the best residual fitting value:

$$\gamma_{jm} = argmin_r \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(X_i) + \gamma) \tag{23}$$

4) Update classifiers:

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm}) \tag{24}$$

5) The final classifier is obtained:

$$\hat{f}(x) = f_M(x) \tag{25}$$

6) Calculation of prediction classification probability:

$$p_i = \frac{1}{1+e^{-\hat{f}(x_i)}} \tag{26}$$

2.5. AdaBoost algorithm

The AdaBoost algorithm obtains different test sample sets by changing the distribution weight of the samples [47]. The principle is to find the weights of the samples incorrectly classified by the weak classifier in training after each training. To increase their values and reduce the weights of the samples correctly classified, find a way to combine the weak classifier to minimize its weight coefficient and form the final strong classifier. In this paper, Decision Tree (DT) was selected as the weak classifier, and the specific algorithm flow is as follows [48]:

Step 1: Build sample data

$$T = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$$

Step 2: Initialize weights,

$$D_1 = (W_{11}, \dots, W_{1i}, \dots, W_{1N}), W_{1i} = \frac{1}{N}, i = 1,2, \dots, N$$

$$G_m(X): X \to \{-1, +1\}$$

$$\varepsilon_m = P[G_m(X_i) \neq Y_i] = \sum_{i=1}^{N} W_{mi} I[G_m(X_i) \neq Y_i]$$

$$\alpha_m = \frac{1}{2}\log\frac{1-\varepsilon_m}{\varepsilon_m}$$

$$D_{m+1} = (W_{m+1,1}, \dots, W_{m+1,i}, \dots, W_{m+1,N})$$

$$W_{m+1,i} = \frac{W_{mi}}{Z_m} exp[-\alpha_m Y_i G_m(X_i)], i = 1,2, \dots, N$$

Step 3: Build a strong classifier:

$$G(X) = sign\left[\sum_{i=i}^{M} \alpha_m G_m(X)\right]$$

2.6. GBDT-AdaBoost fusion algorithm

He et al. [49] applied the method of generating new features through the GBDT model to evaluate advertisement click-through rate. In this study, the GBDT model was used to generate the features of a new tree for each iteration. The feature selection and combination were automatically carried out. The new discrete feature vectors with a distinguishing degree were mined and input into the AdaBoost

model for training to achieve the final classification recognition. The training process of the GBDT-Adaboost fusion model is shown in Figure 3. The specific steps are as follows.

Step 1: Use the method to build the GBDT model and generate many decision trees.

Step 2: Input the original data into the decision tree generated in the previous step for prediction. At this time, the predicted value of each tree in the model was regarded as the new feature data and the new sample data.

Step 3: The new sample data was marked in the way of independent thermal coding, and its middle node was denoted as the node position of the sample output as 1, otherwise as 0, to obtain the position marker vector of each sample. The output of all samples formed a sparse matrix marking the leaf node position of the output of each decision tree.

Step 4: Take the new sample data as the input feature of weak classifier DT in AdaBoost model, build and train the GBDT-Adaboost model. The grid search method was used to obtain the optimal hyperparameter values of the GBDT-Adaboost model. Based on the sample set, the prediction model based on GBDT-Adaboost was trained, and the final prediction results of the model were output. The flow chart of the somatic cell recognition algorithm of cow milk-based on the GBDT-AdaBoost fusion model is shown in Figure 4.

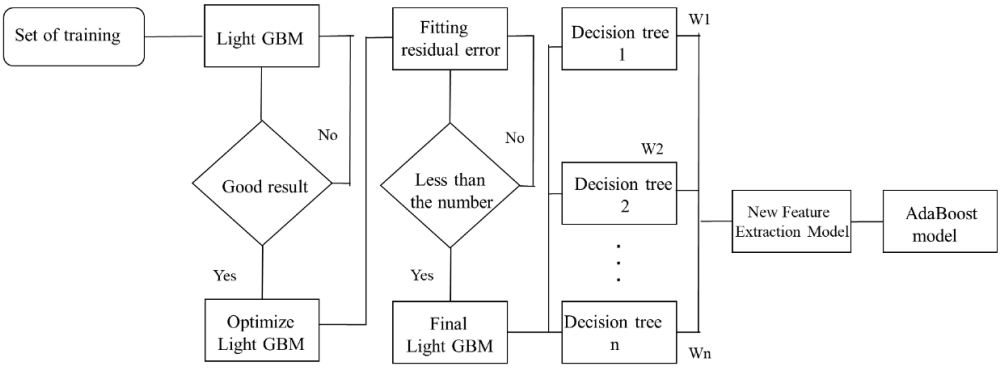

Figure 3. GBDT-AdaBoost fusion model training design.

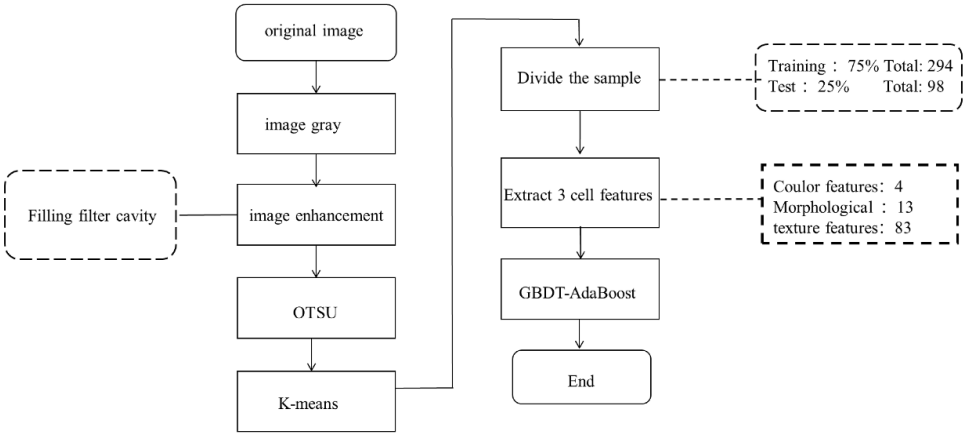

Figure 4. Algorithm flowchart.

3. Results and discussion

3.1. GBDT-AdaBoost analysis of fusion model recognition results

In this study, a total of 100 features, including four-color features, 13 morphological features (six geometric features and seven invariant moment features), and 83 texture features (24 GLCM features and 59 LBP features), were extracted from four types of preprocessed milk somatic cell images. Each feature was separately taken as the input feature of the GBDT-Adaboost fusion model, and the results are shown in Table 1. As can be seen from Table 1, the total accuracy of the separate classification of the three different features fluctuated greatly. Due to the inhomogeneity of nucleus and cytoplasm among somatic cells, the range of color difference between them was small, resulting in the lowest total accuracy of color features being 76.1%. Still, the sensitivity of texture features to chromatic aberration and illumination was weak, so the total accuracy based on texture features was the highest, reaching 97.3%. Because there were different differences among different types of somatic cells, and since some morphological characteristics have small differences, the total accuracy rate based on the morphological characteristics was 88.0%. Therefore, the accuracy and stability of single feature recognition were poor. In addition, according to the characteristics of various cells and the differences between each feature, the three types of extracted features were first fused in this experiment. Then they were used as input features of the GBDT-Adaboost model to achieve the final classification.

Table 1. The recognition results of a single feature.

Feature type	Accuracy/%				
	LYP	NG	MΦ	EPI	Overall Accuracy
Color features	77.6	70.2	76.6	79.7	76.1
Morphological features	87.5	87.5	86.5	90.6	88
Texture features	97.9	96.9	97.9	96.9	97.3

3.2. The GBDT-AdaBoost confusion matrix of the fusion model

Table 2. The method of the confusion matrix. TP: true positive; FN: false negative; FP: false positive; TN: true negative.

Actual category	Positive predict	Negative predict
Positive actual	TP	FN
Negative actual	FP	TN

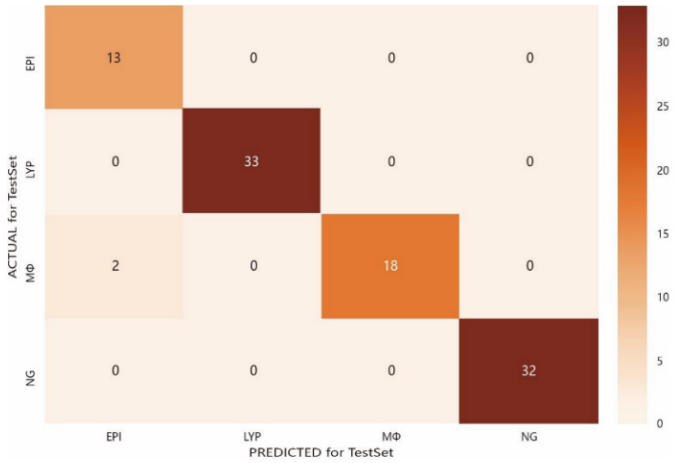

Figure 5. Confusion matrix of GBDT-AdaBoost fusion model.

A confusion matrix was used to display the classification results in this paper. The representation method of the confusion matrix is shown in Table 2, and the confusion matrix obtained in this experiment is shown in Figure 5. All three types of LYP and NG were correctly identified. At the same time, two of the MΦ were incorrectly classified as EPI, indicating that MΦ were similar to EPI in terms of color morphology and texture characteristics.

In this study, accuracy (A), precision (P), recall(R), and comprehensive evaluation index F (F1-Score) were used to evaluate the classification model [50], and the specific formulas are as follows:

$$A = \frac{TP+TN}{TP+TN+FP+FN} \tag{27}$$

$$P = \frac{TP}{TP+FN} \tag{28}$$

$$R = \frac{TP}{TP+FN} \tag{29}$$

$$F = \frac{2PR}{P+R} \tag{30}$$

As shown in Figure 5, for EPI and MΦ, the values of true positive (TP), true negative (TN), false positive (FP), and false-negative (FN) were 13, 83, 2, and 0 and 18, 78, 0, and 2, respectively. According to Eq (27), the accuracy of milk cell identification based on the GBDT-Adaboost fusion model was 98.0%. In contrast, the accuracy of RF, ET, DT and LightGBM was 79.9, 71.1, 67.3 and 77.2%, respectively. Therefore, the fusion model proposed in this paper improves recognition accuracy to a certain extent.

3.3. Comparison of classification results of different models

In order to more fully verify the effectiveness of the GBDT-Adaboost fusion model proposed in this paper, according to the above calculation Eqs (28)–(30), the fusion model and single classification model in this paper were compared and evaluated from three aspects of P, R, and comprehensive evaluation index F for each type of milk somatic cells. Tables 3–6 show the comparison results. Furthermore, receiver operating characteristic (ROC) was used to evaluate the classification performance, the ROC curves of the various classifications for all five models are

presented in Figure 6, And ROC curves achieved more excellent AUCs in GBDT-Aaboost model than in RF model, ET model, DT model and LightGBM model.

Table 3. Precision comparisons of different classification models.

Classification model	Precision rate of each class/%			
	LYP	NG	MΦ	EPI
RF	71.4	88.2	66.7	66.7
ET	75	76.9	56	71.4
DT	65.4	60	65.4	32.1
LightGBM	72.2	82.1	72.2	81.2
GBDT-AdaBoost	100	100	100	86.67

Table 4. Recall comparisons of different classification models.

Classification model	LYP	NG	MΦ	EPI
RF	90.9	88.2	66.7	66.7
ET	68.2	85.7	58.3	58.8
DT	54.8	65.4	34.6	60
LightGBM	96.3	71.9	61.9	72.2
GBDT-AdaBoost	100	100	100	90

Table 5. F1-Score comparisons of different classification models.

Classification model	LYP	NG	MΦ	EPI
RF	80	87	62.2	62.5
ET	71.4	81.1	57.1	64.5
DT	59.6	61.8	43.9	41.9
LightGBM	82.5	76.7	66.7	76.5
GBDT-AdaBoost	100	100	95	93

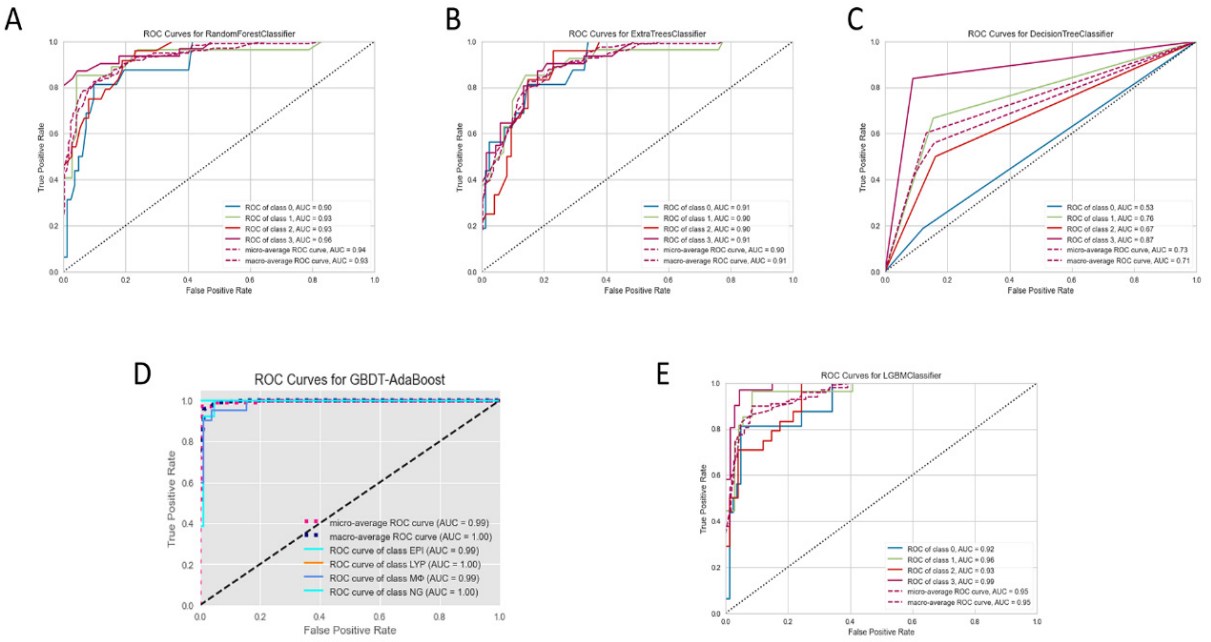

Figure 6. The ROC curves of the various classifications for all five models.

4. Conclusions

This paper proposes a milk somatic cell recognition algorithm based on the GBDT-Adaboost fusion model. Firstly, the original milk somatic cell images were processed by grayscale filtering, denoising, OTSU threshold segmentation, and other preprocessing to obtain the binary images of the cells. Then, the cell region was extracted by the k-means algorithm, and its mean value was extracted. The morphological features of the cells were extracted by calculating the area circumference and roundness. The LBP compared with GLCM was used to extract the texture features of the cell images and fused these features as the recognition features of the milk cells. Then, the color features such as mean-variance were extracted. The morphological features were extracted by calculating the area circumference and roundness. In order to thoroughly learn the data features, the extracted cell features were input into the GBDT model for optimization. Finally, these optimized features were input into the AdaBoost classifier for recognition. The model achieved 98.0, 96.8, 97.5 and 97.0% in classification accuracy, accuracy, recall rate, and F value of comprehensive evaluation index, respectively, which is better than the RF, ET, DT and LightGBM models. In future studies, we will further consider improving learning efficiency while reducing the calculation time. We plan to expand the image data set and extract the depth features of the cell image in combination with the deep learning method, to improve the cell recognition effect and improve the shortcomings of the proposed algorithm.

Acknowledgments

This study was funded by the National Natural Science Foundation of China (#61461041 and 31960494), the Inner Mongolia Autonomous Region Science and Technology Project (#2020GG0169), the Inner Mongolia Autonomous Region Higher Education Scientific Research

Project (#NJZY21486), and the Inner Mongolia Agricultural University Basic Subject Scientific Research Funding Project (#JC2018001).

Conflict of interest

All authors declare that they have no competing interests.

References

1. J. Y. Yang, C. Y. Niu, Y. Y. Liu, B. Q. Fu, J. Wang, Study on the necessity of somatic cell detection and measurement calibration of fresh milk, *Biotechnol. Bull.*, **334** (2020), 21–26. https://doi.org/10.13560/j.cnki.biotech.bull.1985.2019-1121
2. Y. C. Su, N. Zheng, S. L. Li, X. Y. Qu, X. W. Zhou, Research progress on the effect of somatic cell count in raw milk on milk quality and safety, *Food Sci.*, **39** (2018), 299–305. https://doi.org/10.7506/spkx1002-6630-201823043
3. J. X. Gao, Classification and recognition of polymorphic milk somatic cells based on feature fusion, *J. Inn. Mong. Agric. Univ.*, 2018.
4. J. J. Yan, Y. Gao, F. Gao, Research progress of milk somatic cell count detection, *Comput. Meas. Control.*, **2** (2016), 5–10. https://doi.org/0.16526/j.cnki.11-4762/tp.2016.02.002
5. J. C. Zhao, X. C. He, H. W. Gao, Research progress of milk somatic cell count detection methods, *China Cattle*, **13** (2014), 39–43. https://doi.org/10.3969/j.issn.1004-4264.2014.13.012
6. R. Nayar, D. Wilbur, D. Solomon, The bethesda system for reporting cervical cytology, in *Acta Cytologica*, (2008), 77–90. https://doi.org/10.1016/B978-141604208-2.10006-5
7. M. Wei, Y. Du, X. Wu, Q. Su, J. Zhu, L. Zheng, et al., A benign and malignant breast tumor classification method via efficiently combining texture and morphological features on ultrasound images, *Comput. Math. Methods Med.*, **2020** (2020), 5894010. https://doi.org/10.1155/2020/5894010
8. M. Habibzadeh, A. Krzyzak, T. Fevens, Comparative study of feature selection for white blood cell differential counts in low resolution images, *Artif. Neural Networks Pattern Recognit.*, 2014.
9. A. Behura, The cluster analysis and feature selection: perspective of machine learning and image processing, *Wiley*, 2021. https://doi.org/10.1002/9781119785620.ch10
10. A. Bodzas, P. Kodytek, J. Zidek, Automated detection of acute lymphoblastic leukemia from microscopic images based on human visual perception, *Front. Bioeng. Biotechnol.*, **8** (2020), 1005. https://doi.org/10.3389/fbioe.2020.01005
11. X. Gao, H. Xue, X. Pan, X. Jiang, Y. Zhou, X. Luo, Somatic cells recognition by application of gabor feature-based (2D)2PCA, *Int. J. Pattern Recog. Artif. Intel.*, **31** (2017), 1757009. https://doi.org/10.1142/S0218001417570099
12. X. Gao, H. Xue, X. Pan, X. Luo, Polymorphous bovine somatic cell recognition based on feature fusion, *Int. J. Pattern Recog. Artif. Intel.*, **34** (2020), 2050032. https://doi.org/10.1142/S0218001420500329
13. X. Gao, H. Xue, X. Jiang, Y. Zhou, Recognition of somatic cells in bovine milk using fusion feature, *Int. J. Pattern Recog. Artif. Intel.*, **32** (2018), 1850021. https://doi.org/10.1142/S0218001418500210

14. X. Zhang, H. Xue, X. Gao, Y. Zhou, Milk somatic cells recognition based on multi-feature fusion and random forest, *J. Inn. Mong. Agric. Univ.*, *Nat. Sci. Ed*., 2018.
15. S. U. Khan, N. Islam, Z. Jan, K. Haseeb, S. Shah, M. Hanif, A machine learning-based approach for the segmentation and classification of malignant cells in breast cytology images using gray level co-occurrence matrix (GLCM) and support vector machine (SVM), *Neural Comput. Appl.*, **2021** (2021), 1–8. https://doi.org/10.1007/s00521-021-05697-1
16. H. Gai, Y. Wang, L. Chan, B. Chiu, Identification of retinal ganglion cells from β-III stained fluorescent microscopic images, *J. Digit. Imaging*, **2** (2020), 1–12. https://doi.org/10.1007/s10278-020-00365-7
17. J. Rawat, A. Singh, H. S. Bhadauria, J. Virmani, J. S. Devgun, Computer assisted classification framework for prediction of acute lymphoblastic and acute myeloblastic leukemia, *Biocybern. Biomed. Eng.*, **37** (2017), 637–654.
18. V. Acharya, P. Kumar, Detection of acute lymphoblastic leukemia using image segmentation and data mining algorithms, *Med. Biol. Eng. Comput.*, **57** (2019). https://doi.org/10.1007/s11517-019-01984-1
19. H. B. Kmen, A. Guvenis, H. Uysal, Predicting the polybromo-1 (PBRM1) mutation of a clear cell renal cell carcinoma using computed tomography images and KNN classification with random subspace, *JVE J.*, **26** (2019), 30–34. https://doi.org/10.21595/vp.2019.20931
20. P. Mirmohammadi, M. Ameri, A. Shalbaf, Recognition of acute lymphoblastic leukemia and lymphocytes cell subtypes in microscopic images using random forest classifier, *Phys. Eng. Sci. Med.*, **44** (2021), 433–441. https://doi.org/10.1007/s13246-021-00993-5
21. S. Mishra, B. Majhi, P. K. Sa, L. Sharma, Gray level co-occurrence matrix and random forest-based acute lymphoblastic leukemia detection, *Biomed. Signal Process Control*, **33** (2017), 272–280. https://doi.org/10.1016/j.bspc.2016.11.021
22. N. Theera-Umpon, White blood cell segmentation and classification in microscopic bone marrow images, in *Fuzzy Systems and Knowledge Discovery* (eds. L. Wang, Y. Jin), Springer, (2005), 787–796. https://doi.org/10.1007/11540007_98
23. W. D. Lopes, D. Monte, C. Leon, J. Moura, C. Oliveira, Logistic regression model reveals major factors associated with total bacteria and somatic cell counts in goat bulk milk, *Small Rumin. Res.*, **198** (2021), 106360. https://doi.org/ 10.1016/j.smallrumres.2021.106360
24. L. W. Chen, X. P. Wu, C. Pan, Q. C. Hou, Application of extreme learning machine integration in bone marrow cell classification, *Comput. Eng. Appl.*, **51** (2015), 136–139. https://doi.org/10.3778/j.issn.1002-8331.1303-0219
25. A. X. He, B. Y. Wei, B. H. Zhang, B. T. Zhang, B. F. Yuan, B. Z. Huang, Grading of clear cell renal cell carcinomas by using machine learning based on artificial neural networks and radiomic signatures extracted from multidetector computed tomography images, *Acad. Radiol.*, **27** (2020), 157–168.
26. B. S. Divya, S. Kamalraj, H. R. Nanjundaswamy, Human epithelial type-2 cell image classification using an artificial neural network with hybrid descriptors, *IETE J. Res.*, **2018** (2018), 1–12. https://doi.org/10.1080/03772063.2018.1474810
27. F. Lavitt, D. J. Rijlaarsdam, D. Linden, E. Weglarz-Tomczak, J. M. Tomczak, Deep learning and transfer learning for automatic cell counting in microscope images of human cancer cell lines, *Appl. Sci.*, **11** (2021), 4912. https://doi.org/10.3390/app11114912

28. A. Kan, Machine learning applications in cell image analysis, *Immunol. Cell Biol.*, **95** (2017), 525–530. https://doi.org/10.1038/icb.2017.16
29. D. Kusumoto, S. Yuasa, The application of convolutional neural network to stem cell biology, *Inflammat. Regen.*, **39** (2019), 14. https://doi.org/10.1186/s41232-019-0103-3
30. X. Dong, Z. Yu, W. Cao, A survey on ensemble learning, *Front. Comput. Sci.*, **14** (2020), 241–258. https://doi.org/10.1007/s11704-019-8208-z
31. A. Andiojaya, H. Demirhan, A bagging algorithm for the imputation of missing values in time series, *Expert Syst. Appl.*, **129** (2019), 10–26.
32. Y. Hui, X. Mei, G. Jiang, T. Tao, Z. Ma, Milling tool wear state recognition by vibration signal using a stacked generalization ensemble model, *Shock*, **2019** (2019), 1–16. https://doi.org/10.1155/2019/7386523
33. B. Wang, J. Pineau, Online bagging and boosting for imbalanced data streams, *IEEE Trans. Knowl. Data Eng.*, **28** (2016), 3353–3366.
34. W. Zhan, D. He, S. Shi, Recognition of kiwifruit in field based on Adaboost algorithm, *Trans. Chin. Soc. Agric. Eng.*, **29** (2013), 140–146. https://doi.org/10.3969/j.issn.1002-6819.2013.23.019
35. J. Cao, L. Chen, M. Wang, H. Shi, Y. Tian, A parallel adaboost-backpropagation neural network for massive image dataset classification, *Sci. Rep.*, **6** (2016), 38201. https://doi.org/10.1038/srep38201
36. X. Wu, X. Lu, H. Leung, A video-based fire smoke detection using robust adaBoost, *Sensors*, **18** (2018), 3780. https://doi.org/10.3390/s18113780
37. Y. Wang, B. Zheng, M. Xu, S. Cai, J. Younseo, C. Zhang, et al., Prediction and analysis of hub genes in renal cell carcinoma based on CFS gene selection method combined with adaboost algorithm, *Med. Chem.*, **16** (2020), 654–663. https://doi.org/10.2174/1573406415666191004100744
38. J. Wang, Q. Zhou, A. Yin, Self-adaptive segmentation method of cotton in natural scene by combining improved Otsu with ELM algorithm, *Trans. Chin. Soc. Agric. Eng.*, **341** (2018), 181–188. https://doi.org/10.11975/j.issn.1002-6819.2018.14.022
39. S. H. Shirazi, A. I. Umar, S. Naz, M. I. Razzak, Efficient leukocyte segmentation and recognition in peripheral blood image, *Technol. Health Care*, **24** (2016), 335–347. https://doi.org/10.3233/THC-161133
40. X. F. Wang, D. S. Huang, J. X. Du, H. Xu., L. Heutte, Classification of plant leaf images with complicated background, *Appl. Math. Comput.*, **205** (2008), 916–926.
41. Y. K. Zhuang, P. Zhou, Automatic classification of blood leukocytes based on multiple evidence, *J. Zhejiang Sci. Tech. Univ.*, **30** (2013), 367–371.
42. Q. Wu, Y. Gan, B. Lin, Q. Zhang, H. Chang, An active contour model based on fused texture features for image segmentation, *Neurocomputing*, **151** (2015), 133–1141. https://doi.org/10.1016/j.neucom.2014.04.085
43. T. Ojala, M. Pietikainen, D. Harwood, A comparative study of texture measures with classification based on feature distributions, *Pattern Recognit.*, **29** (1996), 51–59. https://doi.org/10.1016/0031-3203(95)00067-4
44. H. Yang, J. Yin, M. Jiang, Perceptual image hashing using latent low-rank representation and uniform LBP, *Appl. Sci.*, **8** (2018), 317. https://doi.org/10.3390/app8020317

45. S. Lv, G. Liu, X. Bai, Multifeature pool importance fusion based GBDT (MPIF-GBDT) for short-term electricity load prediction, *IOP Conf. Series EES*, **702** (2021).
46. Y. X. Wang, Research on big data risk control model based on GBDT algorithm, *J. Zhengzhou Inst. Aeronaut. Ind. Manag.*, **167** (2020), 110–114.
47. J. Techo, C. Nattee, T. Theeramunkong, Boosting-based ensemble learning with penalty profiles for automatic Thai unknown word recognition, *Comput. Math. Appl.*, **63** (2012), 1117–1134.
48. D. Q. Han, T. X. Zhang, W. Shen, Lithology identification based on gradient lifting decision tree (GBDT) algorithm, *Bull. Mineral. Petrol. Geochem.*, **37** (2018), 1173–1180.
49. X. He, J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, et al., Practical lessons from predicting clicks on ads at facebook, *ACM*, **2014** (2014). https://doi.org/10.1145/2648584.2648589
50. W. Xie, Q. Chai, Y. Gan, S. Chen, X. Zhang, W. Wang, Strains classification of anoectochilus roxburghii using multi-feature extraction and stacking ensemble learning, *Trans. Chin. Soc. Agric. Eng.*, **36** (2020), 203–210.