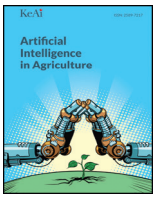
Artificial Intelligence in Agriculture 6 (2022) 100–110

Contents lists available at ScienceDirect

Artificial Intelligence in Agriculture

journal homepage: http://www.keaipublishing.com/en/journals/artificial-intelligence-in-agriculture/



Non-destructive silkworm pupa gender classification with X-ray images using ensemble learning

Sania Thomas *, Jyothi Thomas

Department of Computer Science and Engineering, Christ (Deemed to be University), Bangalore, India

ARTICLE INFO

Article history:
Received 2 March 2022
Received in revised form 4 August 2022
Accepted 5 August 2022
Available online 10 August 2022

Keywords:
Sericulture
Gender classification
Stratified k-fold cross-validation
Machine learning
AdaBoost

ABSTRACT

Sericulture is the process of cultivating silkworms for the production of silk. High-quality production of silk without mixing with low quality is a great challenge faced in the silk production centers. One of the possibilities to overcome this issue is by separating male and female cocoons before extracting silk fibers from the cocoons as male cocoon silk fibers are finer than females. This study proposes a method for the classification of male and female cocoons with the help of X-ray images without destructing the cocoon. The study used popular single hybrid varieties FC1 and FC2 mulberry silkworm cocoons. The shape features of the pupa are considered for the classification process and were obtained without cutting the cocoon. A novel point interpolation method is used for the computation of the width and height of the cocoon. Different dimensionality reduction methods are employed to enhance the performance of the model. The preprocessed features are fed to the powerful ensemble learning method AdaBoost and used logistic regression as the base learner. This model attained a mean accuracy of 96.3% for FC1 and FC2 in cross-validation and 95.3% in FC1 and 95.1% in FC2 for external validation.

1. Introduction

Sericulture is a widely used term for the cultivation of silkworms for the production of silk which comprises multiple fields together such as mulberry cultivation, rearing of cocoons, reeling of silk fibers, twisting, dyeing, and weaving (Ganga, 2019). Sericulture is mainly the combination of the cultivation of cocoons and the production of silk. The silkworm breeds play an important role in high cocoon yield and silk quality. The hybrid breed CSR2 x CSR4 was introduced in India in 1997 that has revolutionized the Indian sericulture industry. The CSR hybrids (single hybrids FC1 (CSR6 x CSR26), and FC2 (CSR2 x CSR27)) are robust and more productive, that can be easily reared by the farmers. However, more care is required in handling these pure races and small ignorance may lead to non-availability of parental cocoons. To overcome this, bivoltine double hybrid breeds were developed. The double hybrids can withstand adverse climatic conditions and hence resulting in crop stability than single hybrids. The double hybrid developed in India is (CSR2 x CSR27) x (CSR6 x CSR26) which can yield around 68.00 kg/100 dfls. Male silkworm cocoon has distinct finer silk filament than female (Zhang et al., 2010). Separating male and female cocoons based on gender improves the quality of the silk (Yu et al., 2005). The methods employed for the gender classification of cocoons include the Near-infrared spectroscopy method, Hyper Spectral Imaging technology, Optical penetration method, Camera imaging technology, X-ray method, etc. (Thomas and Thomas, 2020).

The near-infrared method for the silkworm gender classification (Tao et al., 2018a, 2018b; Zhu et al., 2018; Lin et al., 2019; Qiu et al., 2021) uses the spectral characteristics of the pupa for the classification. In terms of accuracy, the system provides good accuracy but the instrument and calibration costs are very high (Ozaki et al., 2018). The accuracy of the iPLS-CARS-PLSDA model by Qiu et al. is as high as 98.41% (Qiu et al., 2021). They have used near-infrared spectroscopy for the classification of male and female silkworm pupa. This method considers the pupa for gender classification which requires cutting of cocoon, which is a demerit. Any method which can find the volume of pupa accurately without cutting the cocoon may improve the result. Hyperspectral imaging method (Tao et al., 2018a, 2018b; Tao et al., 2019a, 2019b) uses the spectral and spatial information of the pupa for classification. The limitations include the high equipment cost and computational complexity (Schneider and Feussner, 2017). The optical penetration method (Sumriddetchkajorn and Kamtongdee, 2012; Sumriddetchkajorn et al., 2013; Kamtongdee et al., 2015; Sumriddetchkajorn et al., 2015) used the wavelength of light to discriminate the gender of silkworm pupa by locating the chitin gland inside the tale part of the female pupa. The chitin gland is not present in the male pupa. Slight displacement of pupa cannot obtain the region of interest which may cause misclassification in this case. The above-mentioned methods require the cutting of cocoons to find the features of the pupa. The camera imaging method applied for gender

* Corresponding author.
E-mail address: sania.thomas@res.christuniversity.in (S. Thomas).

https://doi.org/10.1016/j.aiia.2022.08.001

classification (Mahesh et al., 2017; Joseph Raj et al., 2019) is a low-cost method but in this method, cocoon characteristics are considered, and not possible to analyze the pupa characteristics. The X-ray method is a promising method that can be used to get the shape characteristics of the pupa without cutting the cocoon.

In imaging methods like the X-ray method and camera imaging method, researchers focus more on feature selection and classification algorithms based on machine learning methods. Agricultural researchers are working to develop fast, non-destructive, and robust methods to apply in the field of agricultural engineering. Przybyło et al. used Convolutional Neural Networks (CNN) for the acorn classification based on the color and intensity of the image of sections of the seeds (Przybyło and Jabłoński, 2019). Researchers have developed novel computer vision systems for the analysis and classification of seed images (Gulzar et al., 2020; Loddo et al., 2021; Loddo et al., 2021; Loddo et al., 2022). Loddo et al. have proposed a novel CNN architecture, SeedNet for seed image classification and retrieval (Loddo et al., 2021). Two very different datasets were tested using SeedNet and attained an accuracy of above 95% in both cases. Gulzar et al. proposed a seed image classification system using CNN and transfer learning. Fourteen commonly known seeds were classified using advanced deep learning techniques (Gulzar et al., 2020). In the validation set using 234 images, the proposed model attained an accuracy of 99%. In recent work, Loddo et al. proposed two plugins for seed image analysis, one able to extract morphological, texture, and color features from seed images, and the second one for the classification of the seeds based on the extracted features (Loddo et al., 2022).

X-rays are a type of electromagnetic radiation, the wavelength ranges from 0.1 to 10 nm which penetrates through the object (Moulet et al., 2017). Agricultural research such as quality determination of fruits and vegetables uses soft x-rays (Du et al., 2019; Xia et al., 2019; Van De Looverbosch et al., 2020). The soft X-ray wavelength ranges from 1 to 10 nm.

In the X-ray method, X-ray images of the cocoon were acquired using soft X-ray and then pre-processed. The shape features of the pupa were extracted and then classification was performed. Cai et al. used the X-ray method for the gender classification of the Pupa. In this, the author obtained an accuracy of 93.31% with the linear discriminant analysis used as a classifier (Cai et al., 2014). Various researchers have done gender classification of silkworm pupa using the methods discussed earlier. All the methods have attained an accuracy of above 90% and the minimum accuracy for gender classification obtained was 91.3% by Mahesh et.al for the CSR2 variant by fusing Zernike moment features with shape features of camera images (Mahesh et al., 2017).

In the proposed method Soft X-rays were used to acquire the X-ray images of the silkworm cocoon. An X-ray of the cocoon is used to extract the shape features of the pupa. 1156 FC1, and 1226 FC2 samples were used in this study. The shape features of the pupa such as width, height, area, perimeter, rectangularity, circularity, height to width ratio, solidity, convexity, volume, and weight were extracted. The novelty of this study includes the number of samples used each variety is high compared to other studies also tried a novel feature extraction method for width and height computation and a proposed classification model using AdaBoost integrated with Linear discriminant analysis. The major advantage of this study is that the cutting of cocoons is not required for gender discrimination. Cutting of cocoons leads to economic loss in the sericulture industry and also it can be done only by skilled laborers.

The remainder of the paper is organized as follows. Section 2 discusses the materials and methods which include sample collection, image pre-processing techniques, feature extraction, and classification method. The classification results are presented in Section 3 and it is discussed in detail in Section 4. The major conclusions are given in Section 5.

2. Materials and methods

The flow of the research work includes the sample collection of FC1 and FC2 cocoons then an X-ray image was acquired. Pre-processing was performed on the images for segmentation and noise reduction. The prominent features were extracted and the dataset was prepared. The entire dataset was divided into training and testing set on the ratio of 80:20. The training dataset was used for the classification model development with 10-fold cross-validation and the remaining 20% testing dataset was used for external validation. The entire process is depicted in Fig. 1.

2.1. Sample collection

Two single hybrid varieties of silkworm cocoon FC1(CSR6 x CSR26) and FC2(CSR2 x CSR27) were used for this research. Samples were collected from the state sericulture department authorized FC1 and FC2 silkworm rearer. A total of 2382 samples were used in this study. The samples include 1156 FC1 cocoons with 589 males and 567 females, 1222 FC2 cocoons with 623 males, and 599 females. Fig. 2 shows some of the silkworm cocoons used in the study. It is hard to discriminate between male and female cocoons with the human eye. In seed production centers cocoons are cut and the pupae are taken out for the gender classification. Fig. 3 shows some of the pupae of the single hybrid variety used for our work.

A single X-ray image of a single silkworm cocoon is used for the study. X-rays of the cocoon were acquired on the 10th day of cocooning. The pictorial representation of the x-ray imaging system is shown in Fig. 4 and some of the sample x-ray images of cocoons are shown in Fig. 5. Gender classification of the pupa was done with the help of experts in the silkworm seed production center, Palakkad, Kerala, India. Soft x-rays are employed for this study as soft x-rays are better for agricultural studies. The X-ray wavelength ranges from 1 to 10 nm, the X-ray tube voltage was 40 kV and the current was 0.6 mA.

2.2. Image pre-processing

2.2.1. Segmentation of the region of interest

Segmentation is a very important task in image classification and object detection problems. Region of interest (ROI) provides the most important information about an object in the image. Separating the ROI from the noise was performed by segmentation. Thresholding is one of the segmentation methods where it considers the pixel value. Simple thresholding can be done in different ways in which if the pixel value is greater than the threshold then the pixel value is assigned to a maximum specified value otherwise set to zero or this can be done in reverse. Another simple thresholding is that if the pixel value is greater than the threshold then that pixel value is truncated to the threshold value otherwise zero. If the problem does not require changing the source pixel value of the ROI then it can be done by making the pixels values to zero which are less than the threshold and keeping the pixel values greater than the threshold intact this can be done in reverse

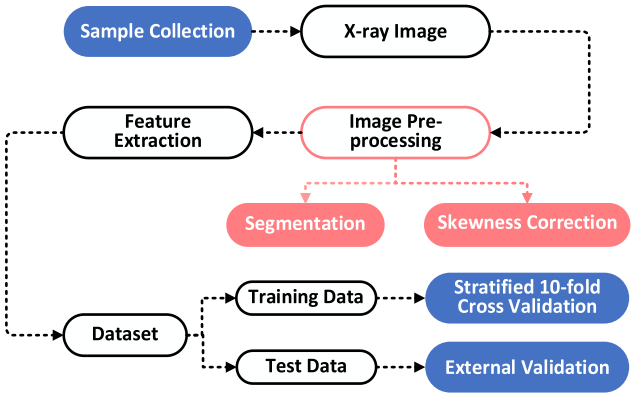


Fig. 1. Flowchart of the work.

Fig. 2. Silkworm cocoons of single hybrid variety.

also (Al-Amri and Kalyankar, 2010). The major disadvantage of this method is finding the threshold value of an image with different lighting conditions in different areas of the image.

In this study, the focus was given to the shape feature extraction of the image. Binary thresholding was an apt choice of segmentation, but identifying the optimum threshold value is a tedious process in simple binary thresholding. To overcome this Otsu's Binarization approach was used in this study. Otsu's is a global image thresholding method. In this method, it obtains the histogram of the image then computes the threshold value and replaces the image pixels with white if the pixel value is greater than the threshold and black otherwise. This best works with bimodal images. Otsu's binarization provides automatic thresholding while ordinary binary thresholding requires manual input of threshold value. Otsu's binarization best performs with different lighting conditions and provides the optimum threshold value. The x-ray images of the pupa are bimodal so Otsu's binarization can provide a better segmentation (Sund and Eilertsen, 2003).

$$\text{Img\_output} = \begin{Bmatrix} 255 & src(x,y) > t \\ 0 & otherwise \end{Bmatrix} \tag{1}$$

Otsu's binarization reduces the within-class variance and maximizes the between-class variance. The general representation of the weighted within-class variance and between-class variance of the two classes are expressed as

$$\sigma_w^2(t) = q_1(t)\sigma_1^2(t) + q_2(t)\sigma_2^2(t) \tag{2}$$

Fig. 3. Silkworm pupae of single hybrid variety.

$$\sigma_b^2(t) = q_1(t)q_2(t)[\mu_1 - \mu_2]^2 \tag{3}$$

where $q_1(t)$ and $q_2(t)$ are the probabilities of the two classes.

$$q_1(t) = \sum_{i=1}^{t} P(i) \tag{4}$$

$$q_2(t) = \sum_{i=t+1}^{I} P(i) \tag{5}$$

$P(i) = \frac{n_i}{n}$; n is the total number of pixels in an image and n_i is the number of pixels with an i intensity value.

The next step is to find the foreground and background means which are denoted by $\mu_1(t)$ and $\mu_2(t)$.

$$\mu_1(t) = \sum_{i=1}^{t} \frac{iP(i)}{q_1(t)} \tag{6}$$

and

$$\mu_2(t) = \sum_{i=1}^{t} \frac{iP(i)}{q_2(t)} \tag{7}$$

Now $\sigma_1^2(t)$ and $\sigma_2^2(t)$ are computed

$$\sigma_1^2(t) = \sum_{i=1}^{t} [i - \mu_1(t)]^2 \frac{P(i)}{q_1(t)} \tag{8}$$

$$\sigma_2^2(t) = \sum_{i=t+1}^{I} [i - \mu_2(t)]^2 \frac{P(i)}{q_2(t)} \tag{9}$$

Total variance can be denoted as the sum of within-class variance and between-class variance.

$$\sigma_T^2(t) = \sigma_w^2(t) + \sigma_b^2(t) \tag{10}$$

The optimum threshold value is computed by minimizing the within-class variance and maximizing between-class variance.

Before performing Otsu's binarization a gaussian blur of kernel size of 7×7 is performed to reduce the noise and improve the result. Fig. 6 shows the result of thresholding without performing Gaussian blur and the noise is visible in the result. Fig. 7 shows the result of thresholding after applying gaussian blur to the image. The result showed a noise-free segmented image.

2.2.2. Skewness correction

While taking the x-ray of the cocoon, the position of the pupa inside the cocoon is unpredictable and it can be in any direction. Therefore, skewness correction was performed on the image. The operation involves the identification of the angle of the pupa segmented from the x-ray image, and rotating the segmented image to correct the skew. Initially, a minimum area rectangle is drawn around the pupa and the rotation angle θ is measured by the angle between the horizontal x-axis and the first edge of the identified minimum area rectangle. The expected angle will be between −90 degrees to 0 degrees. To compute the actual angle, if the angle of rotation obtained is −45 degrees, then it is required to add 90 degrees to the angle and inverse it otherwise just inverse the angle. Based on this obtained angle the skewness correction was performed. The rotation matrix was identified for this purpose. A rotated image was obtained by simple matrix multiplication of the original image with the rotation matrix.

$$R_{\text{rotated}} = I_{\text{original}} \times M_{\text{rotation matrix}} \tag{11}$$

where R_{rotated} is the rotated image, I_{original} is the original image and $M_{\text{rotation matrix}}$ is the rotation matrix. The rotation matrix of angle θ is defined as

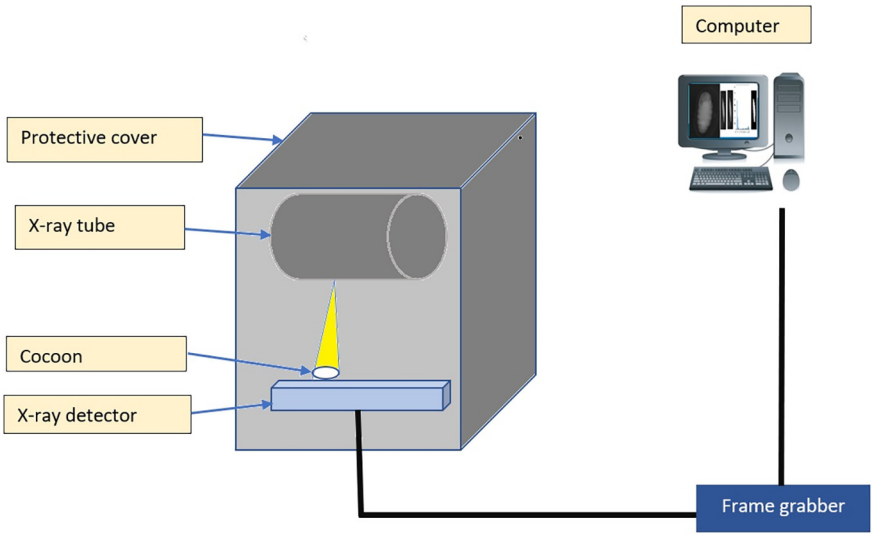


Fig. 4. X-ray imaging system.

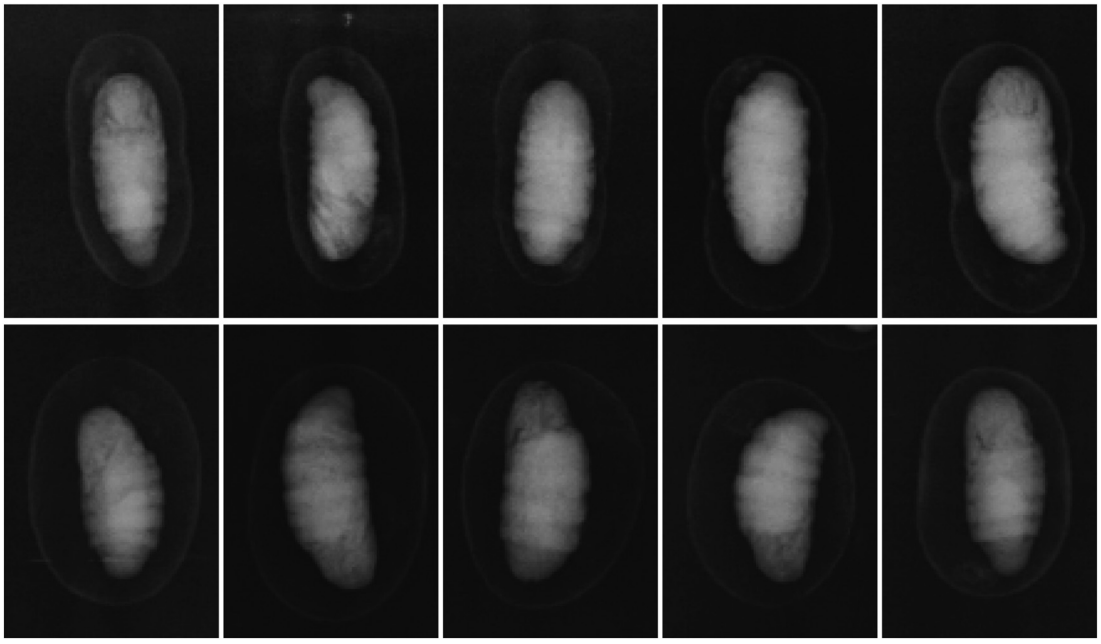


Fig. 5. X-ray images of the silkworm cocoon.

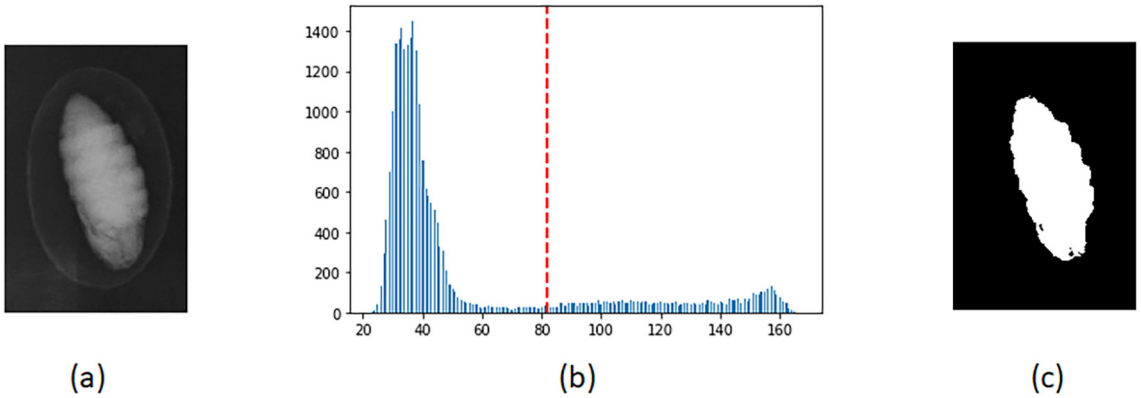


Fig. 6. Otsu's binarization without applying gaussian blur a) Original Image b) Histogram c) Segmented image.

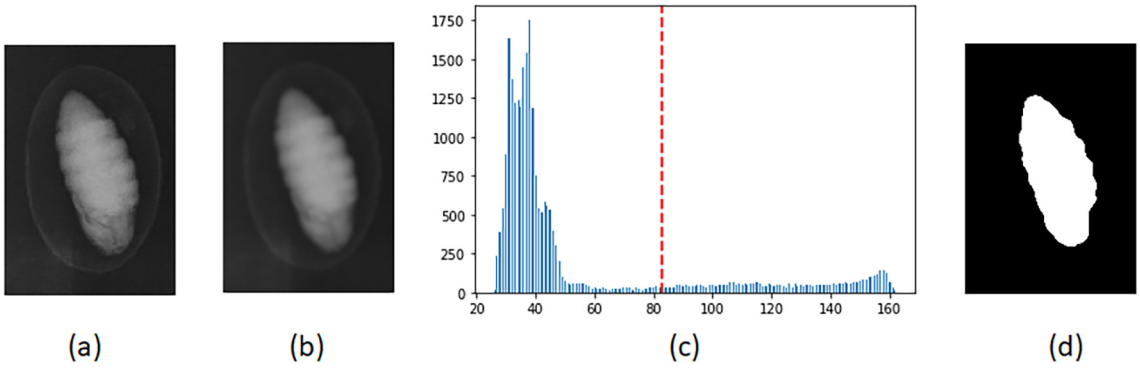


Fig. 7. Otsu's binarization after applying gaussian blur a) Original Image b) Applying gaussian blur c) Histogram d) Segmented image.

$$M_{\text{rotation matrix}} = \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix} \tag{12}$$

This matrix rotates the image about the origin and in this case, the origin is the center of the image. But if required to rotate the image in an arbitrary coordinate, a modified rotation matrix is used which includes translation, rotation, and translation. The modified rotation matrix is shown below.

$$M = \begin{bmatrix} \alpha & \beta & (1-\alpha) \times c_x - \beta \times c_y \\ -\beta & \alpha & \beta \times c_x + (1-\alpha) \times c_y \end{bmatrix} \tag{13}$$

where α is the scale value multiplied with the $cos\theta$ and β is the scale value multiplied with $sin\theta$. The scale value used for this problem is 1. c_x and c_y are the arbitrary center about which the rotation needs to be performed. In this problem, the center for rotation is used as the center of the contour identified around the pupa. This can be computed by identifying the moments of the image.

$$c_x = \frac{m_{10}}{m_{00}} \tag{14}$$

$$c_y = \frac{m_{01}}{m_{00}} \tag{15}$$

The skewness correction of the pupa image will help to identify the width of the pupa more accurately which is shown in Fig. 8. Subfigure 8. a, 4.b, 8.c shows the stages of skewness correction. Fig. 8.d shows the original image after skewness correction for better understanding and Fig. 8.e shows the contour drawn around the pupa after thresholding and skewness correction.

2.3. Feature extraction

2.3.1. Measuring width and length of the pupa

Methods that are used for the dimension measurement methods include ellipse fitting and minimum rectangular fit method which helps in the width and height estimation of regularly shaped objects. In the ellipse fit method, an ellipse is inscribed within the object contour and used a minor axis for width estimation and a major axis for height estimation. In the minimum rectangle fit method, a minimum area rectangle is fitted around the object, and then use the width of the rectangle for the width of the object and the height of the rectangle for the height of the object. These two methods fail in some cases. Hence a new method was proposed in this study.

In the proposed method, a point interpolation method was used. The steps are given below.

Step 1: Identify the center of the contour by finding the moment features of the contour. Centre of the contour was measured using eq. (14) and eq. (15).

Step 2: Identify the width and height of the image by considering the dimension of the image. Then calculate the extreme left, and right using the Eq. (16) and Eq. (17).

$$\text{Left end} = c_x - (w/2) \tag{16}$$

$$\text{Right end} = c_x + (w/2) \tag{17}$$

Step 3: Perform Interpolation from center to endpoints by keeping c_y as constant. The point which intersects with the contour is considered as the width point.

Step 4: Calculate the height by computing the extreme top and bottom points of the contour.

Step 5: Compute the distance of the points by Euclidean distance measure.

Fig. 9 shows the width and height calculated with the ellipse method (Fig. 9.a), minimum area rectangle method (Fig. 9.b), and proposed width calculation method (Fig. 9.c). From the figures, it is clear that the ellipse method and rectangular method are not providing the expected and accurate width of the pupa. Better accuracy of width and height was obtained in the proposed method.

2.3.2. Other predominant features

Other predominant features considered for this study were area, perimeter, volume, circularity, rectangularity, solidity, convexity, width to height ratio, and weight apart from width and height. The area and perimeter of the pupa were calculated by considering the area and arc length of the contour respectively. Circularity was measured by fitting a minimum enclosing circle around the pupa and dividing the area of the circle with the area of the pupa. Rectangularity was computed by fitting a minimum area rectangle around the pupa and dividing the area with the area of the pupa. Solidity was computed by finding the convex closure of the pupa and then dividing the area of the convex closure by the area of the pupa. Convexity was identified by dividing the perimeter of the convex hull with the pupa perimeter. Calculating volume from a two-dimensional image is a difficult task as it is not providing depth information. To overcome this issue manually checked the width and depth of the pupa using vernier caliper and found out that there was not much difference in the width and depth of the pupa. Width and height information was computed from the image. With the available height and width information, the volume is computed using the volume of the ellipsoid formula as the shape of the pupa resembles an ellipsoid. Table 1 shows the summary of predominant features and their calculations.

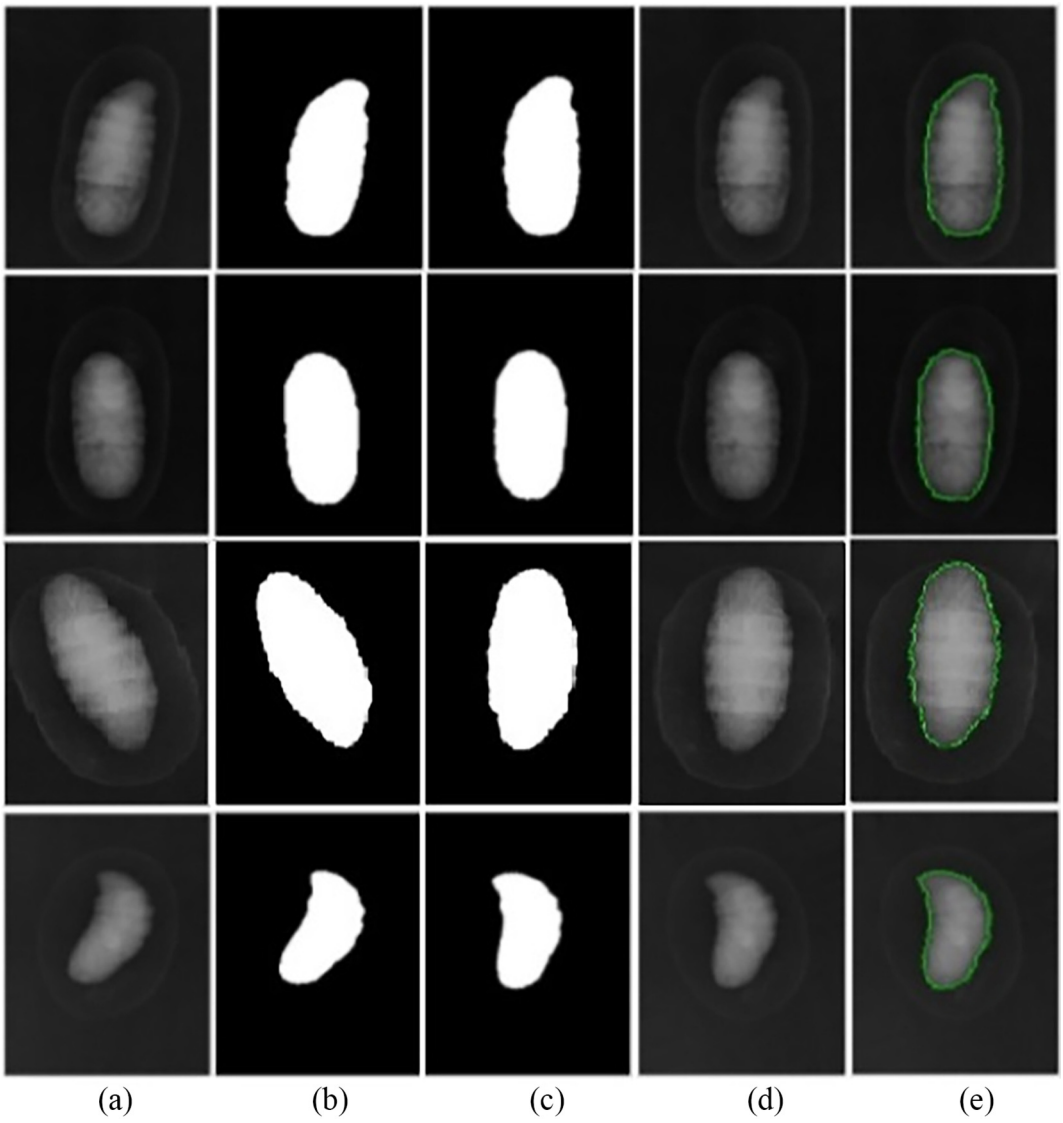


Fig. 8. Skewness correction a) Original image b) Segmented image c) Skewness corrected image d) Skewness corrected original image e) Original image with contour.

2.3.3. Dimensionality reduction

Dimensionality reduction techniques help in reducing redundancy, noise, and the complexity of the algorithm which in turn improves the accuracy of the model (Huang et al., 2019). Dimensionality reduction methods were applied in various fields such as data mining, pattern recognition, machine learning, etc., (Xu et al., 2019). Different dimensionality reduction methods such as Principal Component Analysis, Linear Discriminant Analysis, Truncated Singular Value Decomposition, t-distributed Stochastic Neighbor Embedding (t-SNE), and Multidimensional Scaling (MDS) were used and the results were compared.

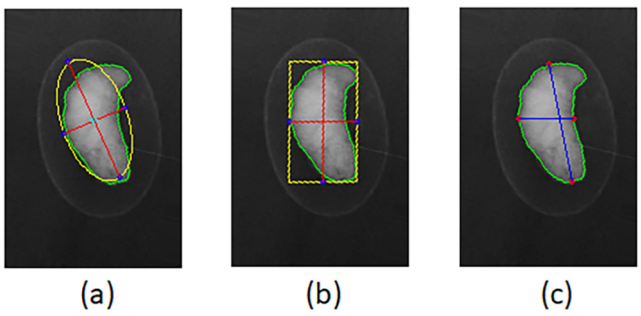


Fig. 9. Width and length estimation using various methods a) Ellipse method b) Rectangular method c) Proposed method.

Table 1
Predominant features.

Features	Calculation
Area	Area of the contour
Perimeter	Arc length of the contour
Circularity	Area/Area of the minimum enclosing circle
Rectangularity	Area/Area of the minimum area rectangle
Solidity	Area/Area of the convex closure
Convexity	Arc length of the convex closure/P
Width to height ratio	Width/height
Volume	$(4/3)^*(22/7)^*(\text{height}/2)^*(\text{width}/2)^2$
Weight	Weight of the cocoon with pupa
Width	Proposed method
Height	Extreme points of the top and bottom contour

2.4. Classification

In ensemble learning, a single problem is solved based on the training of multiple learners. A set of hypotheses are constructed and they are combined to get a solution for the problem. This consists of base learners or weak learners. In ensemble methods, the performance of the weak learners (base learners) is boosted. Base learners are developed from the training data by base learning algorithms such as decision-tree, logistic regression, etc. (Binson et al., 2021a, 2021b, 2021c; Schapire, 2013; Wang et al., 2021).

AdaBoost is used for the classification of the data. AdaBoost is an ensemble learning method that converts weak learners into strong (Schapire, 2013). It is employed for this research as it is the best model for a binary classification problem (Wang et al., 2021). AdaBoost is used for both classification and regression and it is a predictive joint learning algorithm. AdaBoost uses multiple iterations to generate a single composite learner strong and hence it is adaptive. "It creates the strong learner (a classifier that is well correlated to the true classifier) iteratively adding the weak learners (a classifier that is only slightly correlated with the true classifier). During each round of training, a new weak model is added and is trained with the weighted training data." This step is repeated until the development of a preset number of weak learners or no new enhancements could be made to the training data (Binson et al., 2021a, 2021b, 2021c). AdaBoost is used in different applications such as disease diagnosis (Sevinç, 2022; Binson et al., 2021a, 2021b, 2021c), quality prediction (Bai et al., 2021), gender classification (Wang, 2021), banking (Lahmiri et al., 2020), etc., In this study, two weak learners such as decision tree and logistic regression were used.

2.5. Computing environment

For this research python, 3.8.8 version was used along with OpenCV 4.5.2 and scikit learn 0.24.1 libraries. The operating system used was Microsoft windows 10, RAM 16GB, processor Intel core i7.

3. Results

3.1. Data analysis

3.1.1. 10-fold cross-validation

In this study, the classifier employed was AdaBoost and tried decision tree and logistic regression as weak learners with different dimensionality reduction techniques. To obtain the best performing model 10-fold cross-validation was employed (Bey et al., 2020). The performance matrices such as accuracy, precision, recall, f1 score, and AUC score were used for analyzing the performance of the classification model which is shown in Table 2. The accuracy score provides information about how effectively the model classified female cocoons and male cocoons. Precision gives information about how effectively the classifier performed the classification based on misclassification also known as false negatives. Recall gives the performance efficiency of the model based on false positives. F1 score provides an efficiency score of the model by using precision and recall which is the weighted average of precision and recall. Area under the ROC curve determines the effectiveness of the model in distinguishing the classes.

Table 2
Performance evaluation parameters.

Evaluation Metric	
Accuracy Score	$\left\{\frac{(True\_Positive+True\_Negative)}{(True\_Positive+False\_Positive+True\_Negative+False\_Negative)}\right\}$
Precision Score	$\left\{\frac{True\_Positive}{(True\_Positive+False\_Positive)}\right\}$
Recall Score	$\left\{\frac{True\_Positive}{(True\_Positive+False\_Negative)}\right\}$
F1 Score	$\left\{\frac{2\times(Recall\times Precision)}{(Recall+Precision)}\right\}$
AUC	Area under the ROC Curve

The classification of the dataset is done using Adaptive Boosting. To obtain performance improvements various dimensionality reduction technologies such as Principal Component Analysis, Linear Discriminant Analysis, Truncated Singular Value Decomposition, t-distributed Stochastic Neighbor Embedding (t-SNE), and Multidimensional Scaling (MDS) were used. Two weak learners such as decision tree and logistic regression were tried and it is found that logistic regression used as the weak learner for the AdaBoost performed better than decision tree as the weak learner in our binary classification problem. Table 3 shows the mean performance metrics score of FC1 cocoon classification of 10-fold cross-validation with different dimensionality reduction algorithms and logistic regression and decision tree as the weak learners which are graphically represented using Figs. 10 and 11. Fig. 10 visualizes the performance evaluation of AdaBoost with Logistic regression as a weak learner integrated with various dimensionality reduction methods of FC1 variety. Fig. 11 shows the evaluation metrics of AdaBoost with the decision tree as the weak learner and different dimensionality reduction methods for FC1. Hereby analyzing the performance metrics it is visible that the dimensionality reduction method LDA applied with the classifier AdaBoost and Logistic regression as the weak learner performed better with less time of 0.698 s for FC1.

Table 4 shows the mean score of the performance metrics such as accuracy,f1, precision, recall, AUC, and time of the FC2 cocoon gender classification with different dimensionality reduction techniques along with the boosting ensemble learning method AdaBoost with Logistic regression and decision tree as the classifier. Figs. 12 and 13 represents the graphical representation of the performance of the classifier with Logistic regression and decision tree used as the weak learners along with different dimensionality reduction methods. By analyzing the performance metrics, it shows that LDA along with AdaBoost and Logistic regression as the weak learner attained better performance in 0.732 s.

The classifier model is also tried without using dimensionality reduction techniques and the performance is depicted in Table 5. Analyzing the data in the table shows that the performance of the classifier is enhanced by the use of the Linear Discriminant Analysis dimensionality reduction method.

3.1.2. Proposed model

The proposed model for the classification of the pupa based on gender was designed using LDA + AdaBoost with Logistic regression as the weak learner. Hyperparameter tuning is done using the grid search method. The final model contains 100 estimators at which boosting is terminated. The weight applied to the boosting iterations is 0.0005. Stagewise Additive Modeling using a Multi-class Exponential loss function (SAMME) is used as the boosting algorithm. The proposed model is tested with the remaining 20% of data kept for the external validation purpose. External validation is performed to view the performance of the model. The data used for the external validation is obtained with the same acquisition condition as the main dataset. The number of images used for external validation is 232 images of FC1(111 males and 121 females) and 245 images of FC2(137 males and 108 females).

The results showed that the proposed model attained an accuracy of 95.3% in FC1 classification and 95.1% in FC2 classification. Fig. 14 shows the confusion matrix of the external validation of FC1 and FC2 using the proposed classifier. The performance matrix derived from this confusion matrix is depicted in Fig. 15.

4. Discussion

Gender classification of the pupa is an important task in the mulberry silkworm sericulture industry. This can help in improving the quality production of silk filament as well as export. The classification can be done before cutting the cocoon by using the X-ray imaging method which can penetrate through the cocoon so that we can extract

Table 3
Classifier performance with Different dimensionality reduction methods for FC1.

Classifier	Accuracy	F1	Precision	Recall	AUC	Time
PCA + AdaBoost(Logistic Regression)	0.957	0.953	0.984	0.926	0.96	3.639 s
PCA + AdaBoost(Decision Tree)	0.949	0.948	0.942	0.955	0.957	1.219 s
LDA + AdaBoost(Logistic Regression)	0.971	0.97	0.974	0.967	0.971	0.698 s
LDA + AdaBoost(Decision Tree)	0.968	0.967	0.967	0.966	0.969	0.778 s
MDS + AdaBoost(Logistic Regression)	0.955	0.952	0.959	0.946	0.954	2.219 s
MDS + AdaBoost(Decision Tree)	0.962	0.961	0.958	0.964	0.963	0.844 s
t-SNE + AdaBoost(Logistic Regression)	0.946	0.943	0.951	0.937	0.946	1.573 s
t-SNE + AdaBoost (Decision Tree)	0.827	0.834	0.777	0.91	0.876	0.806 s
SVD + AdaBoost(Logistic Regression)	0.955	0.951	0.981	0.924	0.953	1.389 s
SVD + AdaBoost(Decision Tree)	0.949	0.948	0.942	0.955	0.957	1.254 s

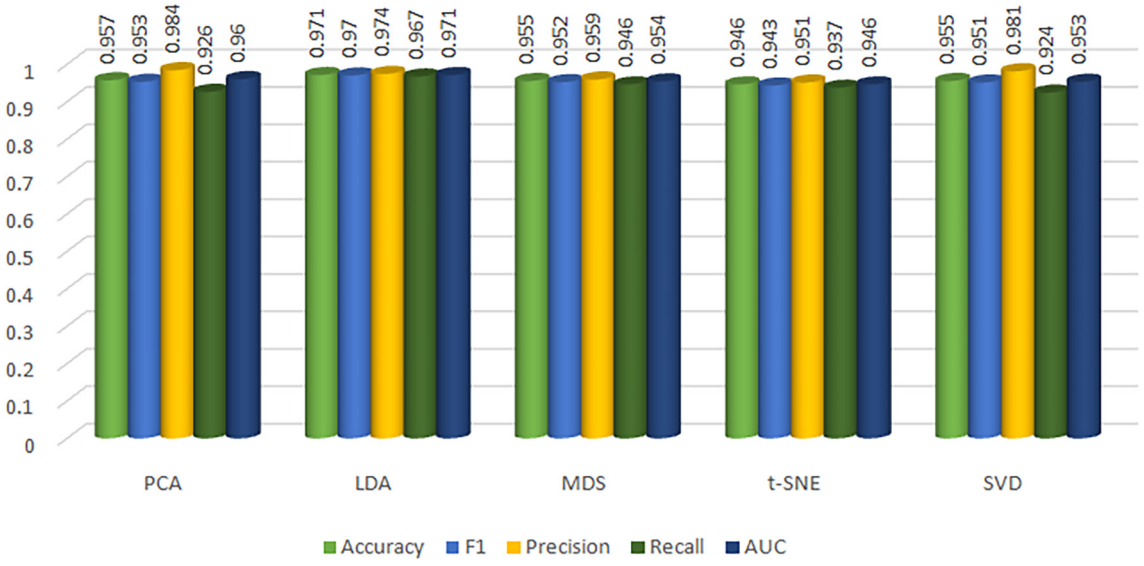


Fig. 10. FC1 data classification using AdaBoost (Logistic Regression) with different dimensionality reduction methods.

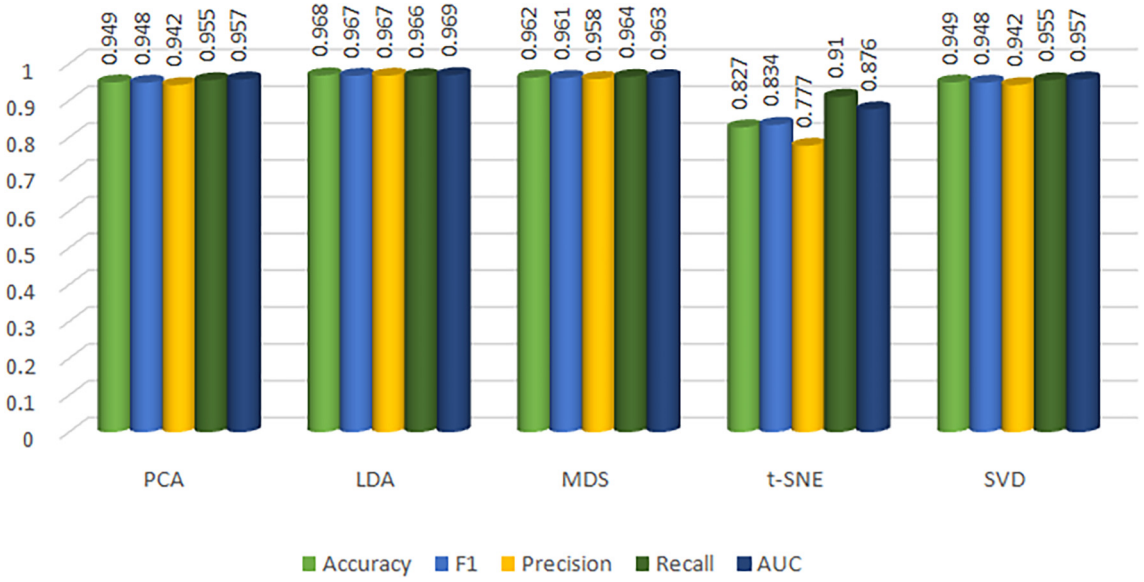


Fig. 11. FC1 data classification using AdaBoost (Decision Tree) with different dimensionality reduction methods.

Table 4
Classifier performance with Different dimensionality reduction methods for FC2.

Classifier	Accuracy	F1	Precision	Recall	AUC	Time
PCA + AdaBoost(Logistic Regression)	0.938	0.936	0.954	0.921	0.94	2.131 s
PCA + AdaBoost(Decision Tree)	0.867	0.871	0.847	0.898	0.872	9.732 s
LDA + AdaBoost(Logistic Regression)	0.963	0.963	0.962	0.965	0.963	0.732 s
LDA + AdaBoost(Decision Tree)	0.953	0.954	0.959	0.959	0.955	0.498 s
MDS + AdaBoost(Logistic Regression)	0.936	0.936	0.936	0.937	0.951	2.586
MDS + AdaBoost(Decision Tree)	0.938	0.936	0.959	0.916	0.938	0.669 s
t-SNE + AdaBoost(Logistic Regression)	0.9	0.9	0.898	0.904	0.899	1.513 s
t-SNE + AdaBoost (Decision Tree)	0.812	0.798	0.855	0.754	0.83	0.762 s
SVD + AdaBoost(Logistic Regression)	0.927	0.927	0.938	0.917	0.927	0.889 s
SVD + AdaBoost(Decision Tree)	0.867	0.871	0.847	0.898	0.872	0.974 s

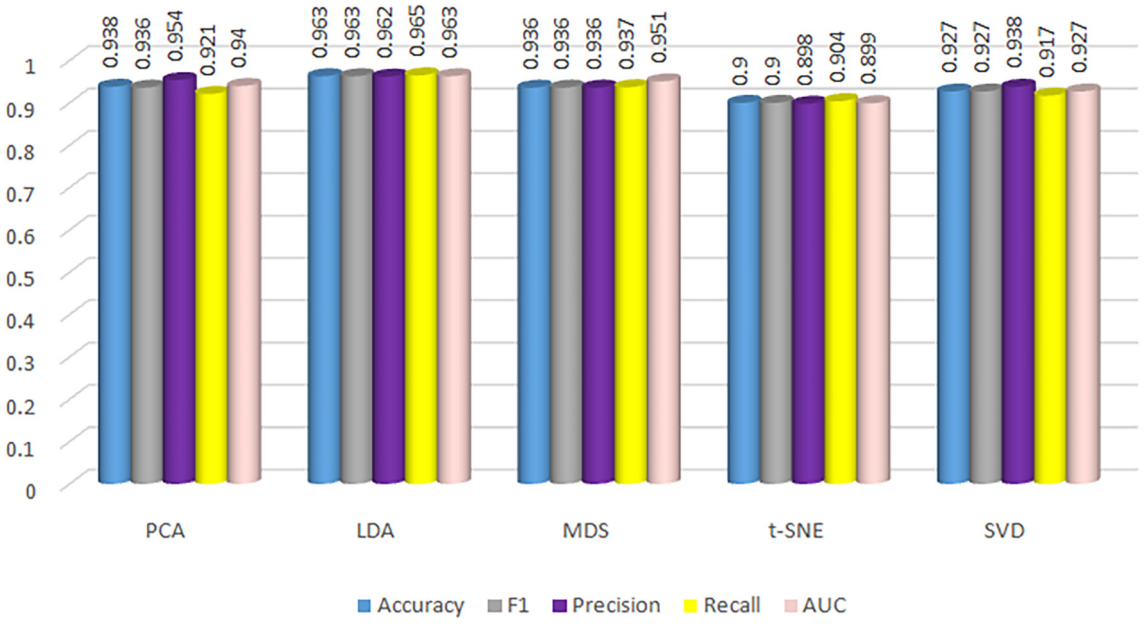


Fig. 12. FC2 data classification using AdaBoost (Logistic Regression) with different dimensionality reduction methods.

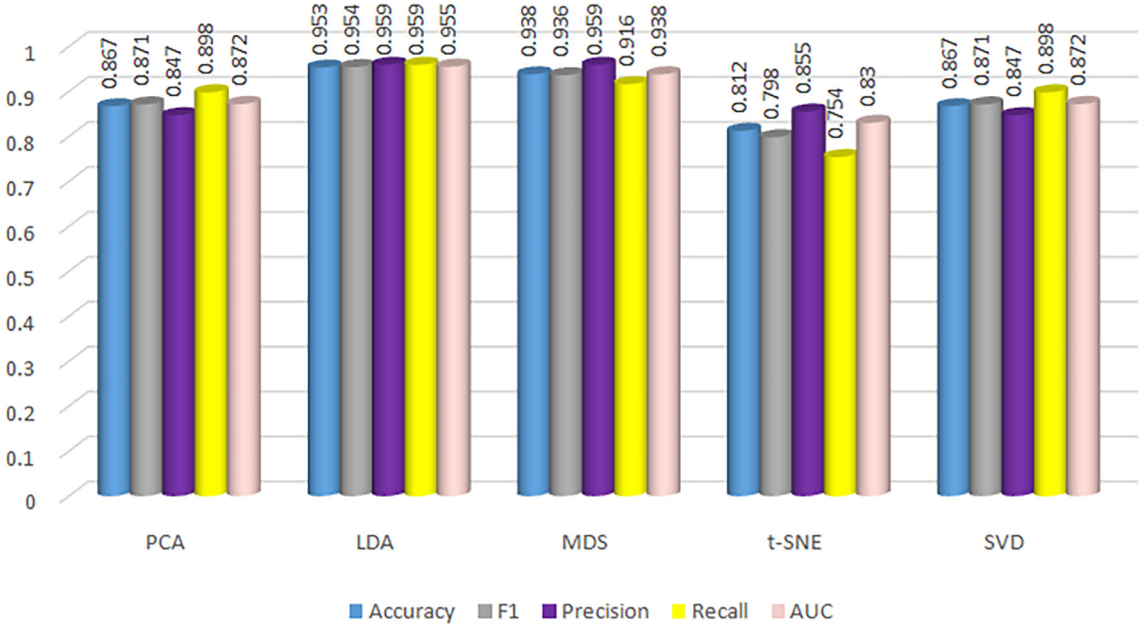


Fig. 13. FC2 data classification using AdaBoost (Decision Tree) with different dimensionality reduction methods.

Table 5
Classifier performance without dimensionality reduction methods.

classifier	Variety	Accuracy	Precision	Recall	F1-score	AUC
AdaBoost (Weak learner: Logistic regression)	FC1	0.921	0.914	0.924	0.918	0.935
	FC2	0.934	0.939	0.933	0.935	0.934
AdaBoost (Weak learner: Decision Tree)	FC1	0.903	0.913	0.936	0.924	0.934
	FC2	0.883	0.884	0.887	0.885	0.887

the pupa features without destructing the cocoon. The proposed width calculation method provides a better width compared to the minimum area rectangular fitting method and ellipse fitting method. The height computation was also better by considering the extreme points of the contour than the ellipse fitting and rectangular fitting method. Feature extraction also plays a major role in classifier performance.

The results session depicted the data analysis of the AdaBoost model with different weak classifiers such as decision tree and logistic regression and performed different dimensionality reduction methods. The classifier is analyzed with the help of the performance metrics such as accuracy, precision, recall, f1-score, and AUC. Based on the analysis the proposed model is developed with AdaBoost using Logistic regression as a weak learner with Linear Discriminant Analysis as the dimensionality reduction method. The AdaBoost model without using the dimensionality reduction method is also tried and the results are depicted in the Table 5. Comparing this table with Tables 3 and 4, it is clear that applying the dimensionality reduction method LDA enhances the performance of the classifier. This proposed model was tested with the validation dataset and obtained 95.3% of accuracy for FC1 and 95.1% for FC2 gender classification.

X-ray imaging for mulberry silkworm pupa classification used in the research by Cai et al. reported accuracy of 93.31% with the LDA classification model (Cai et al., 2014). In this model, the author used shapes describing features such as ellipse major axis, ellipse minor axis, the ratio of the major axis to minor axis, eccentricity, concave-convex characteristics, roundness, rectangularity, and complexity. The sample size used for this work was a total of 1071 of four hybrid varieties with around 531 female and 540 male cocoons.

The present research was carried out with 1156 FC1 and 1226 FC2 samples which helped in a better understanding of the features of the pupa. Other methods used for pupa gender classification were NIR spectroscopy (Tao et al., 2018a, 2018b; Zhu et al., 2018; Lin et al., 2019; Qiu et al., 2021), Hyper Spectral Imaging (HSI) (Tao et al., 2018a, 2018b; Tao et al., 2019a, 2019b), and the optical penetration method (Sumriddetchkajorn and Kamtongdee, 2012; Sumriddetchkajorn et al., 2013; Kamtongdee et al., 2015; Sumriddetchkajorn et al., 2015). The main concerns include that these methods required the destruction of cocoons for the pupa gender classification, the sample size used for each variety was less, and the cost of the machinery. Present research work showed a better accuracy compared with the existing methods and the sample size used was higher compared with the existing studies.

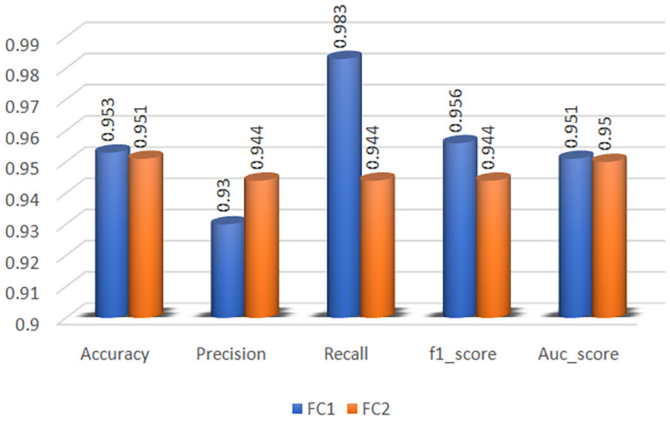


Fig. 15. Performance evaluation on external validation of FC1 and FC2.

If any method that accurately computes the volume irrespective of assuming it to be a particular shape will contribute more to the research. The researchers have not proved the information that the exposure of a cocoon to an X-ray can cause any damage to the pupa. Therefore, the proposed method can be adopted in seed production centers for the quality production of seeds. This requires more studies on how X-rays affect the pupa when cocoons are exposed to x-rays.

5. Conclusion

The Sericulture industry is an inevitable field that contributes to the cultural and economical development of a country. India is the second-largest exporter of silk. The quality of silk production can be improved by separately reeling silk fibers of male and female cocoons. This can avoid the mixing of silk. The proposed model uses the x-ray images of the cocoons through which we can identify the shape features of the pupa without destructing the cocoon. The proposed width extraction and height extraction provides more accurate feature extraction.

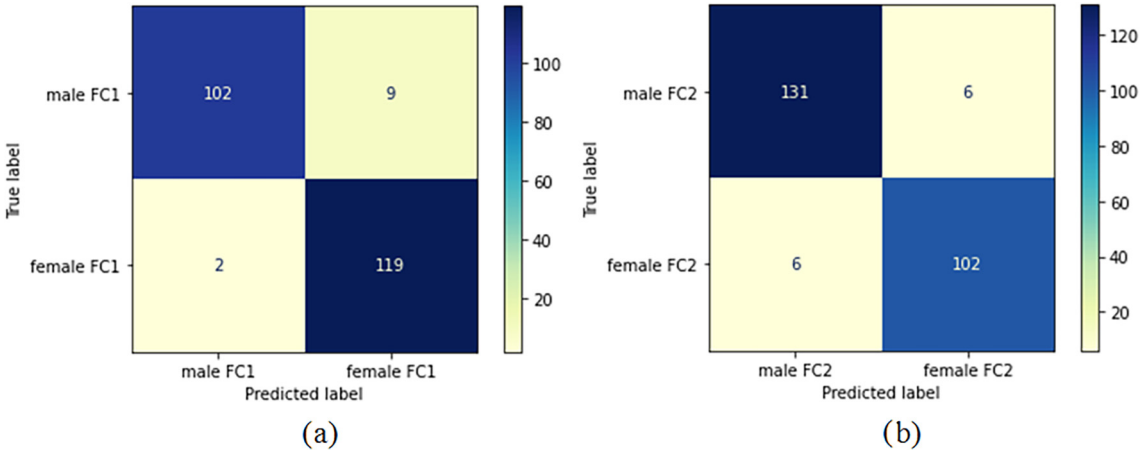


Fig. 14. Confusion matrix of External Validation a) Confusion Matrix of FC1 b) Confusion Matrix FC2.

Application LDA as the dimensionality reduction method enhanced the efficiency and performance of the classifier. The proposed model with 10-fold cross-validation in the training data provided a mean accuracy of 96.3% for both FC1 and FC2 variety. In external validation, the proposed model obtained an accuracy of 95.3% for FC1 and 95.1% for FC2.

Funding

This research was funded by the Department of Science & Technology (DST) with grant reference number: SEED/WS/2019/135.

CRediT authorship contribution statement

Sania Thomas: Conceptualization, Data curation, Investigation, Methodology, Software, Visualization, Writing – original draft. **Jyothi Thomas:** Conceptualization, Data curation, Funding acquisition, Methodology, Supervision, Validation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to acknowledge the support of the seed production center, Palakkad, Kerala, India for the assistance in sample collection.

References

Al-Amri, S.S., Kalyankar, N.V., 2010. Image segmentation by using threshold techniques. arXiv preprint 2 (5), 83–86 (arXiv:1005.4020).
Bai, Y., Xie, J., Wang, D., Zhang, W., Li, C., 2021. A manufacturing quality prediction model based on AdaBoost-LSTM with rough knowledge. Comput. Ind. Eng. 155, 107227.
Bey, R., Goussault, R., Grolleau, F., Benchoufi, M., Porcher, R., 2020. Fold-stratified cross-validation for unbiased and privacy-preserving federated learning. J. Am. Med. Inform. Assoc. 27 (8), 1244–1251.
Binson, V.A., Subramoniam, M., Mathew, L., 2021a. Detection of COPD and lung Cancer with electronic nose using ensemble learning methods. Clin. Chim. Acta 523, 231–238.
Binson, V.A., Subramoniam, M., Ragesh, G.K., Kumar, A., 2021b. Early detection of lung cancer through breath analysis using adaboost ensemble learning method. 2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS). IEEE, pp. 183–187.
Binson, V.A., Subramoniam, M., Sunny, Y., Mathew, L., 2021c. Prediction of pulmonary diseases with electronic nose using SVM and XGBoost. IEEE Sensors J. 21 (18), 20886–20895.
Cai, J.R., Yuan, L.M., Liu, B., Sun, L., 2014. Nondestructive gender identification of silkworm cocoons using X-ray imaging with multivariate data analysis. Anal. Methods 6 (18), 7224–7233.
Du, Z., Hu, Y., Ali Buttar, N., Mahmood, A., 2019. X-ray computed tomography for quality inspection of agricultural products: a review. Food Sci. Nutr. 7 (10), 3146–3160.
Ganga, G., 2019. An Introduction to Sericulture. Oxford and IBH Publishing.
Gulzar, Y., Hamid, Y., Soomro, A.B., Alwan, A.A., Journaux, L., 2020. A convolution neural network-based seed classification system. Symmetry 12 (12), 2018.
Huang, X., Wu, L., Ye, Y., 2019. A review on dimensionality reduction techniques. Int. J. Pattern Recognit. Artif. Intell. 33 (10), 1950017.
Joseph Raj, A.N., Sundaram, R., Mahesh, V.G., Zhuang, Z., Simeone, A., 2019. A multi-sensor system for silkworm cocoon gender classification via image processing and support vector machine. Sensors 19 (12), 2656.
Kamtongdee, C., Sumriddetchkajorn, S., Chanhorm, S., Kaewhom, W., 2015. Noise reduction and accuracy improvement in optical-penetration-based silkworm gender identification. Appl. Opt. 54 (7), 1844–1851.
Lahmiri, S., Bekiros, S., Giakoumelou, A., Bezzina, F., 2020. Performance assessment of ensemble learning systems in financial data classification. Intell. Syst. Account. Finance Manag. 27 (1), 3–9.
Lin, X., Zhuang, Y., Dan, T., Guanglin, L., Xiaodong, Y., Jie, S., Xuwen, L., 2019. The model updating based on near infrared spectroscopy for the sex identification of silkworm pupae from different varieties by a semi-supervised learning with pre-labeling method. Spectrosc. Lett. 52 (10), 642–652.
Loddo, A., Loddo, M., Di Ruberto, C., 2021. A novel deep learning based approach for seed image classification and retrieval. Comput. Electron. Agric. 187, 106269.
Loddo, A., Di Ruberto, C., Vale, A.M.P.G., Ucchesu, M., Soares, J.M., Bacchetta, G., 2022. An effective and friendly tool for seed image analysis. Vis. Comput. 1–18.
Mahesh, V.G., Raj, A.N.J., Celik, T., 2017. Silkworm cocoon classification using fusion of zernike moments-based shape descriptors and physical parameters for quality egg production. Int. J. Intell. Syst. Technol. Appl. 16 (3), 246–268.
Moulet, A., Bertrand, J.B., Klostermann, T., Guggenmos, A., Karpowicz, N., Goulielmakis, E., 2017. Soft x-ray excitonics. Science 357 (6356), 1134–1138.
Ozaki, Y., Huck, C.W., Beć, K.B., 2018. Near-IR Spectroscopy and its Applications. Molecular and Laser Spectroscopy. Advances and Applications, pp. 11–38.
Przybyło, J., Jabłoński, M., 2019. Using deep convolutional neural network for oak acorn viability recognition based on color images of their sections. Comput. Electron. Agric. 156, 490–499.
Qiu, G., Tao, D., Xiao, Q., Li, G., 2021. Simultaneous sex and species classification of silkworm pupae by NIR spectroscopy combined with chemometric analysis. J. Sci. Food Agric. 101 (4), 1323–1330.
Schapire, R.E., 2013. Explaining adaboost. Empirical inference. Springer, Berlin, Heidelberg, pp. 37–52.
Schneider, A., Feussner, H., 2017. Biomedical Engineering in Gastrointestinal Surgery. Academic Press.
Sevinç, E., 2022. An empowered AdaBoost algorithm implementation: a COVID-19 dataset study. Comput. Ind. Eng. 165, 107912 107912.
Sumriddetchkajorn, S., Kamtongdee, C., 2012. Optical penetration-based silkworm pupa gender sensor structure. Appl. Opt. 51 (4), 408–412.
Sumriddetchkajorn, S., Kamtongdee, C., Sa-Ngiamsak, C., 2013. May. Spectral imaging analysis for silkworm gender classification. Sensing Technologies for Biomaterial, Food, and Agriculture 2013. vol. 8881. SPIE, pp. 21–25.
Sumriddetchkajorn, S., Kamtongdee, C., Chanhorm, S., 2015. Fault-tolerant optical-penetration-based silkworm gender identification. Comput. Electron. Agric. 119, 201–208.
Sund, T., Eilertsen, K., 2003. An algorithm for fast adaptive image binarization with applications in radiotherapy imaging. IEEE Trans. Med. Imaging 22 (1), 22–28.
Tao, D., Wang, Z., Li, G., Qiu, G., 2018a. Accurate identification of the sex and species of silkworm pupae using near infrared spectroscopy. J. Appl. Spectrosc. 85 (5), 949–952.
Tao, D., Wang, Z., Li, G., Xie, L., 2018b. Simultaneous species and sex identification of silkworm pupae using hyperspectral imaging technology. Spectrosc. Lett. 51 (8), 446–452.
Tao, D., Qiu, G., Li, G., 2019a. A novel model for sex discrimination of silkworm pupae from different species. IEEE Access 7, 165328–165335.
Tao, D., Wang, Z., Li, G., Xie, L., 2019b. Sex determination of silkworm pupae using VIS-NIR hyperspectral imaging combined with chemometrics. Spectrochim. Acta - A: Mol. Biomol. Spectrosc. 208, 7–12.
Thomas, S., Thomas, J., 2020. A review on existing methods and classification algorithms used for sex determination of silkworm in sericulture. International Conference on Intelligent Systems Design and Applications. Springer, Cham, pp. 567–579.
Van De Looverbosch, T., Bhuiyan, M.H.R., Verboven, P., Dierick, M., Van Loo, D., De Beenbouwer, J., Sijbers, J., Nicolaï, B., 2020. Nondestructive internal quality inspection of pear fruit by X-ray CT using machine learning. Food Control 113, 107170.
Wang, J., 2021. Research on facial feature-based gender intelligent recognition based on the Adaboost algorithm. Int. J. Biom. 13 (1), 40–50.
Wang, X., Ma, Y., Hsieh, M.H., Yung, M.H., 2021. Quantum speedup in adaptive boosting of binary classification. Sci. China Phys. Mech. Astron. 64 (2), 1–10.
Xia, Y., Xu, Y., Li, J., Zhang, C., Fan, S., 2019. Recent advances in emerging techniques for non-destructive detection of seed viability: a review. Artif Intell Agr 1, 35–47.
Xu, X., Liang, T., Zhu, J., Zheng, D., Sun, T., 2019. Review of classical dimensionality reduction and sample selection methods for large-scale data processing. Neurocomputing 328, 5–15.
Yu, X.H., Wang, C.L., Jiang, J.C., Jia, Z.W., 2005. Effect of cocoon dry and cooking conditions and silkworm gender on silk quality. Silk Monthly 3, 12–15.
Zhang, Y., Yu, X., Shen, W., Ma, Y., Zhou, L., Xu, N., Yi, S., 2010. Mechanism of fluorescent cocoon sex identification for silkworms Bombyx mori. Sci. China.Life Sci. 53 (11), 1330–1339.
Zhu, Z., Yuan, H., Song, C., Li, X., Fang, D., Guo, Z., Zhu, X., Liu, W., Yan, G., 2018. High-speed sex identification and sorting of living silkworm pupae using near-infrared spectroscopy combined with chemometrics. Sens. Actuators B Chem. 268, 299–309.