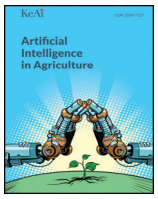
Artificial Intelligence in Agriculture 8 (2023) 1–19

Contents lists available at ScienceDirect

Artificial Intelligence in Agriculture

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



Lightweight convolutional neural network models for semantic segmentation of in-field cotton bolls

Naseeb Singh [a,*], V.K. Tewari [a], P.K. Biswas [b], L.K. Dhruw [a]

[a] Department of Agricultural and Food Engineering, IIT Kharagpur, Kharagpur 721 302, India
[b] Department of Electronics and Electrical Communication Engineering, IIT Kharagpur, Kharagpur 721 302, India

ARTICLE INFO

Article history:
Received 30 December 2022
Received in revised form 1 March 2023
Accepted 3 March 2023
Available online 7 March 2023

Keywords:
Cotton
Semantic segmentation
Convolutional neural network
Robotic harvesting
Image segmentation
Deep learning

ABSTRACT

Robotic harvesting of cotton bolls will incorporate the benefits of manual picking as well as mechanical harvesting. For robotic harvesting, in-field cotton segmentation with minimal errors is desirable which is a challenging task. In the present study, three lightweight fully convolutional neural network models were developed for the semantic segmentation of in-field cotton bolls. Model 1 does not include any residual or skip connections, while model 2 consists of residual connections to tackle the vanishing gradient problem and skip connections for feature concatenation. Model 3 along with residual and skip connections, consists of filters of multiple sizes. The effects of filter size and the dropout rate were studied. All proposed models segment the cotton bolls successfully with the cotton-IoU (intersection-over-union) value of above 88.0%. The highest cotton-IoU of 91.03% was achieved by model 2. The proposed models achieved F1-score and pixel accuracy values greater than 95.0% and 98.0%, respectively. The developed models were compared with existing state-of-the-art networks namely VGG19, ResNet18, EfficientNet-B1, and InceptionV3. Despite having a limited number of trainable parameters, the proposed models achieved mean-IoU (mean intersection-over-union) of 93.84%, 94.15%, and 94.65% against the mean-IoU values of 95.39%, 96.54%, 96.40%, and 96.37% obtained using state-of-the-art networks. The segmentation time for the developed models was reduced up to 52.0% compared to state-of-the-art networks. The developed lightweight models segmented the in-field cotton bolls comparatively faster and with greater accuracy. Hence, developed models can be deployed to cotton harvesting robots for real-time recognition of in-field cotton bolls for harvesting.

1. Introduction

Cotton (*Gossypium hirsutum* L.) is an important cash crop for the farmers of developed as well as developing countries and is presently harvested using machines like stripper or spindle pickers in developed countries and manually in developing countries (Bakhsh et al., 2017). Hand-picking of cotton bolls has certain advantages like better preservation of fiber characteristics, less trash content, and high harvesting efficiency but causes health hazards due to pesticide residues (Bakhsh et al., 2016, 2017; Memon et al., 2019), as well as it is a time-consuming and labor-intensive job. In the wake of a reduction in agricultural workers as well as an increase in agricultural labor costs in near future (Mehta et al., 2019), mechanical harvesting of cotton is necessary. But currently, cotton harvesting machines are the non-selective type which has many drawbacks as compared to manual hand picking. For instance, three times increase in trash content (Shukla et al., 2017) which decreases the quality of fiber (Tian et al., 2018), necessarily use of defoliation chemicals (Snipes and Baskin, 1994) which increases the harvesting cost, soil compaction due to movement of heavy machines on cotton fields (Braunack and Johnston, 2014) which decreases water and nutrients use efficiency (Colombi et al., 2018), having less picking efficiency (85.0–90.0% for spindle pickers) (Williford et al., 1994), and higher field losses during harvesting (Hughs et al., 2008), etc. Thus so, harvesting cotton manually and using presently available machines has its advantages and disadvantages. Using harvesting robots for cotton harvesting can be a better alternative as cotton harvesting robots can incorporate the advantages of hand picking as well as machines by replacing humans with machines for performing selective picking of cotton as performed by humans. The prospect of using cotton-picking robots has increased because of recent studies on agricultural autonomous platforms (Raikwar et al., 2022; Roshanianfard et al., 2020; Zaidner and Shapiro, 2016), which are well-suited to autonomous agricultural robots.

In the development of a cotton harvesting robot, the first major task is the in-field detection of cotton bolls which is a challenging task

* Corresponding author.
E-mail address: naseeb501@gmail.com (N. Singh).

https://doi.org/10.1016/j.aiia.2023.03.001

Table 1
Features extracted by researchers for cotton recognition.

Features extracted	Accuracy	Reference
Color features (YCbCr color space)	90.44%	Liu et al., 2011
Color features (RGB color space)	88.09%	Wang et al., 2008
Spatial features	88.0%	Yeom et al., 2018
Color features (RGB and YCbCr color space)	91.05%	Singh et al., 2021
Color and spatial features	84.60%	Sun et al., 2019

because of the non-uniformity in visual characteristics of cotton bolls, varying illumination conditions during harvesting operation, and complex background around the harvesting scene, due to which errors arises in cotton recognition. For the segmentation of cotton bolls, in past, numerous researchers used various image processing techniques for which color, shape, and texture features were extracted as shown in Table 1, and used as an individual basis or as a combination of multiple features. But, features extracted using hand-engineered-based methods are often low-dimensional (Wang et al., 2021). For the segmentation of cotton bolls with minimal errors in natural illumination with a complex background, high dimensional features need to be extracted which is a difficult task using present conventional image segmentation methods. Thus, a new technique barring conventional image segmentation methods needs to be implemented for cotton images which can extract high dimensional features automatically. A convolutional neural network is a technique that can extract high-dimensional features automatically through self-learned features. Using this convolutional neural network technique, numerous researchers achieved promising results for classification (Dyrmann et al., 2016; Jiang et al., 2019; Khanramaki et al., 2021; Waheed et al., 2020; Xie et al., 2021; Zhang et al., 2019) as well as for semantic segmentation (Azizi et al., 2020; Chen et al., 2020; Kang et al., 2021; Kestur et al., 2019; Tassis et al., 2021; Zabawa et al., 2020; Zou et al., 2021) tasks. For classification tasks, at the end of convolutional layers, dense layers are added while for semantic segmentation, fully convolutional neural networks are used in which a specific class is assigned to each pixel based on high features learned during the training of convolutional neural network models. For semantic segmentation of agricultural images, Tassis et al. (2021), Zou et al. (2021), Chen et al. (2020), Azizi et al. (2020), and Majeed et al. (2018) used fully convolutional neural networks and achieved a mean intersection over union (mean-IoU) value of 94.25%, 92.91%, 89%, 80.50%, and 59.0%, respectively. Regarding the use of convolutional neural networks for cotton segmentation, in the past, Li et al. (2017) and Li et al. (2016) employed fully convolutional neural networks in their studies to recognize in-field cotton bolls and achieved an intersection over union (IoU) score of 70.4% and 73.5%, respectively. Using convolutional neural networks, Tedesco-Oliveira et al. (2020) predicted cotton yield with an accuracy of 80.0%, while Singh et al. (2022) used existing state-of-the-art deep learning models to discriminate the cotton pixels from the sky, and achieved an IoU score of above 80.0%. Further applications of CNN in cotton include the prediction of cotton yarn intensity (Zhenlong et al., 2018) and the identification of infected cotton leaves (Yan et al., 2021).

For the working of cotton harvesting robots in real-time with higher productivity and lesser field losses, a high recognition rate with minimal errors as well as lower harvesting time is imperative. At present, the accuracy of cotton segmentation using a convolutional neural network is below 80.0% (Li et al., 2016, 2017), therefore, this study aimed to develop light-weight convolutional neural network (CNN) models which will be able to segment the cotton bolls with higher accuracy, minimal errors and comparatively a lower inference time. The performance of developed models were evaluated in terms of intersection over union, F1-score, accuracy, precision, and recall values. For a comprehensive evaluation, the performance of developed models was also compared with the cotton segmentation results obtained using existing state-of-the-art networks.

The major contributions of the present study are as follows:

1. Low-level and high-level features were combined with various filter sizes to create a neural network model for in-field cotton boll detection that can be deployed on small single-board computers like Raspberry Pi, etc.
2. The impact of the dropout rate in the convolutional layer as well as filter size on the performance of the convolutional neural network is discussed.
3. The ablation study of various proposed models provides insight into the implications of model pruning on their performance.

The rest of the paper is organized as follows. In Section 2, image dataset preparation, the methodology followed for CNN model development, their training, and quality criteria used for performance evaluation of CNN models is presented. The results obtained by the trained models to segment cotton bolls are presented in Section 3. In the end, Section 4 concludes the findings of this study.

2. Materials and methods

This section describes the image acquisition, labeling, preprocessing, and augmentation process. This section further describes the process followed for the development of three new architectures of CNNs. The training procedure of developed architectures is discussed along with the explanation of criteria matrices used in this study for the performance evaluation of developed CNN models.

2.1. Cotton image acquisition

Haryana is a major cotton-growing state in Northern India (Seidu, 2018) and the cotton images used in the present study were captured from a cotton field of *Kharanti* village, Rohtak district, India (29°01′30″N, 76°27′12″E). A total of one hundred RGB color images (having a resolution of 480 × 640 pixels) of cotton plants from the field with cotton bolls due for harvesting in a week were acquired under natural illumination conditions using a 0.9-megapixel digital camera (Logitech Webcam C270) having a fixed focal length of 4.6 mm and diagonal field of view of 55°. As the distance and angle of the captured scene of the cotton field will vary based on the position of vision sensors mounted on the robotic arm, therefore, for this study, with a purpose to make the image dataset heterogeneous and practical for in-field application of developed models, images were captured at random distances and angles to cotton plants.

2.2. Dataset construction

2.2.1. Image labeling

A labeled image is one in which each pixel is assigned a specific class manually and this process is called image labeling which is a fundamental pre-requisite to train the CNNs if the developed CNNs are for semantic segmentation tasks. In this study, image labeling was done for two-class binary classification for which pixels of cotton bolls were assigned a value of '*1*', while pixels of the region other than cotton bolls were assigned a value of '*0*' using an annotation app '*apeer*' provided by *ZEISS microscopy* (apeer [WWW Document], 2021).

2.2.2. Data augmentation

The training of CNNs required a sizeable number of labeled images (Zou et al., 2021), moreover, to achieve better semantic predictions using CNNs, image labeling should be carried out with maximal accuracy, and hence, labeling of images consumes a lot of time, human efforts, etc. Therefore, due to the limitation of resources in the labeling of images, an image augmentation technique in which synthetic images and their corresponding labeled images were produced artificially, was successfully implemented by various researchers (Azizi et al., 2020;

Chen et al., 2020; Kang et al., 2021; Zou et al., 2021) and achieved impressive accuracy in semantic segmentation tasks. Previous studies (Kang et al., 2021; Sun et al., 2021; Tassis et al., 2021; Zabawa et al., 2020; Zou et al., 2021) have shown that with the right data augmentation techniques, researchers may get good results with a limited number of images instead of a much larger one. Therefore, to increase the number of image datasets for the training and validation of CNNs, the image augmentation technique was implemented in this study. Out of hundred captured images, ninety images were taken randomly and the image augmentation technique was applied to each image. To accomplish this, images with random values were rotated from 10° to 60°, moved horizontally by 100 pixels, moved vertically by 100 pixels, flipped horizontally as well as vertically, upon which five new synthetic images and their corresponding labeled images were obtained. Fig. 1 shows the sample of augmented cotton images and corresponding masked images. Thus, the cotton dataset that was used for the training of CNNs consists of a total of 540 images out of which 405 images (75%) were used for the training of CNNs while the remaining 135 images (25%) were used for validation purposes. The remaining ten raw images were used for the testing of the proposed models.

2.3. Building convolutional structure of proposed models

Inspired by various state-of-the-art convolutional networks, three different custom architectures were designed in this study to recognize the cotton bolls using end-to-end fully convolutional networks. All three designed architectures have the same baseline which is as follows: (a) unless specified, all convolutions were performed using a 3×3 filter size with a stride of 1, (b) all max-pooling operations were performed using a 2×2 filter size with a stride of 2, (c) all transposed convolutions operations were performed using 2×2 with a stride of 2, (d) for all convolutional, transposed convolutional and max-pooling layers, padding was applied such that the output dimensions of feature map result in the same dimensions as of input feature map.

A non-linear activation function was applied after each convolutional layer, with the purpose to extract non-linear features from computed feature maps as well as increasing the accuracy of proposed convolutional networks (Jarrett et al., 2009). In the present study, the ReLU activation function proposed by Nair and Hinton (2010) was used as it leads to faster convergence (Glorot et al., 2011; Krizhevsky et al., 2012) and do not suffer from vanishing gradient problems.

Fig. 1. Sample of augmented raw images and corresponding masked images.

The ReLU activation function retains only the positive values of the computed feature maps while making the negative values equal to zero and mathematically, can be defined by Eq. (1).

$$f(x) = \max(x, 0) \tag{1}$$

In the last layer of each model, the soft-max activation function was used. During the training process of convolutional neural networks, at first, initial weights need to be provided to each layer after which these weights will be updated iteratively during the backpropagation process. This process of generation of initial weights according to the dimensions of layers is called *weight initialization* which is a vital operation for the network convergence (Kumar, 2017) as too-large or too-small weights initialization may lead to exploding and vanishing gradients issues, respectively (Bengio et al., 1994) during the training of the convolutional network. As in the present study, the *ReLU* activation function was used in neural layers, hence, as recommended by He et al. (2015), weights were initialized using *He* initialization for the custom proposed models. All models proposed in this study, are essentially comprised of encoder and decoder parts.

2.3.1. Building convolutional structure of model 1

For model 1, each convolutional block in the encoder part consists of one convolutional layer followed by batch normalization and a ReLU activation layer. After the successive inclusions of two such convolution blocks in architecture, a max-pooling layer was added to reduce the spatial dimensionality. This arrangement is repeated three times. At its end portion, the encoder consists of a max-pooling layer followed by a single convolutional block. Each block in the decoder part consists of a single number of transposed convolution layers, convolutional layer, batch normalization layer, and activation layer, sequentially as mentioned. This decoder block in the decoder up-sampled the input feature maps using the 2×2 transposed convolution. After the inclusion of four such blocks in architecture, at the end of the decoder, an additional 1×1 convolution is applied to reduce the depth of feature maps to the required number of classes which is two in the present study i.e., background and cotton classes, and produce the pixel-wise segmented image. At each convolutional block in the encoder, the number of feature channels increases progressively, while in the decoder at each decoder block, the number of feature channels decreased progressively. The architecture of proposed model 1 is illustrated in Fig. 2 along with the dimensions of feature maps and the number of filters.

2.3.2. Building convolutional structure of model 2

The pooling operation used in the encoder not only reduces the spatial dimensionality of feature maps but also results in the elimination of spatial information (Ronneberger et al., 2015) as well as generates checkerboard artifacts during up-sampling (Sugawara et al., 2019). Inspired by Ronneberger et al. (2015), in proposed model 2 and model 3, skip connections were used to pass features from the encoder to the decoder with the purpose to regain spatial information. For this, features maps in encoders were concatenated channel-wise to the features maps of identical dimensions in the decoder. As compared to model 1, layers in model 2 were increased to enhance the segmentation accuracy. But, the inclusion of additional layers in architecture may lead to an increase in training error due to degradation and vanishing gradient problems (He et al., 2015). Therefore, in the present study, the probable problem of increased training error has been alleviated by using the residual connection. Using the residual connection, two layers were directly connected while skipping one or more layers in between. Fig. 3 shows the residual block used in the architecture of model 2 to alleviated the training error issues. In this, the residual connection simply convolutes the input feature maps with kernel size of 1×1, output of which will be added to the output of the other stacked layer as shown in Fig. 3.

The encoder part of model 2 comprises three convolutional blocks, two residual blocks, and four max-pooling layers. Two residual blocks were successively added after the first convolutional block as shown in Fig. 4. Max-pooling layer was placed after each convolutional block and residual block except the last convolutional block. The decoder part of model 2 consists of four transposed convolution layers, with each of which feature maps from the encoder concatenated using skip connections. The concatenation of transposed convolutional and skip features were followed by a residual block and this arrangement was repeated four times in the decoder. The complete architecture of proposed model 2 is illustrated in Fig. 4 along with the dimensions of feature maps and the number of filters.

2.3.3. Building convolutional structure of model 3

Practically, during the operation of robotic harvesting in a cotton field, a visual sensor will capture cotton bolls at different viewing distances and angles due to which spatial layout, as well as the scale of cotton bolls, will be random. As mentioned by Li et al. (2017), single-instance cotton bolls (present globally in an image) are easier to segment as compared to multi-instance cotton bolls (present locally in an image). Hence, choosing a kernel size in such cases is always a challenging task as a larger kernel size is preferred for information that is present

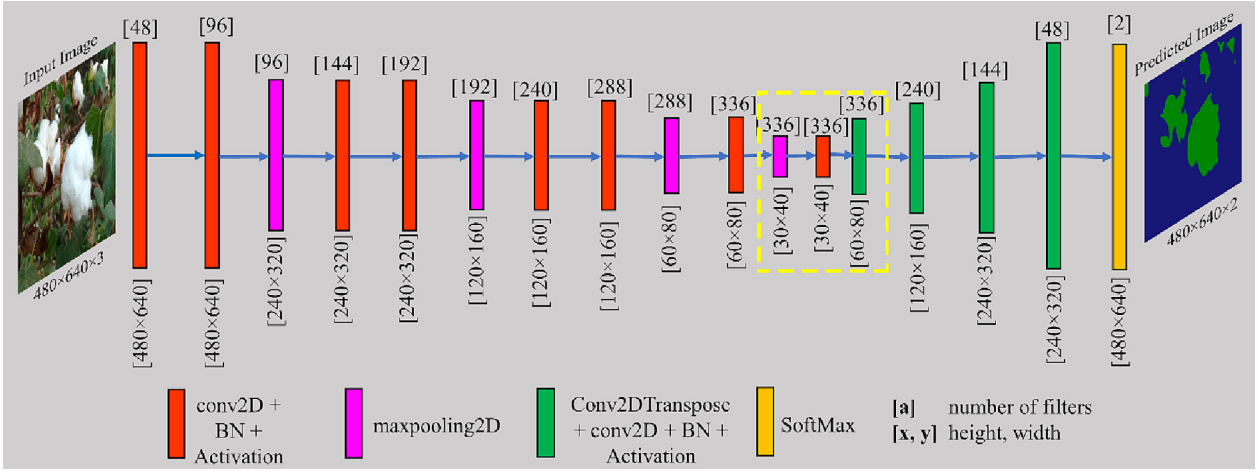


Fig. 2. Conceptual diagram of CNN model 1 (For yellow box refer to ablation experiments section).

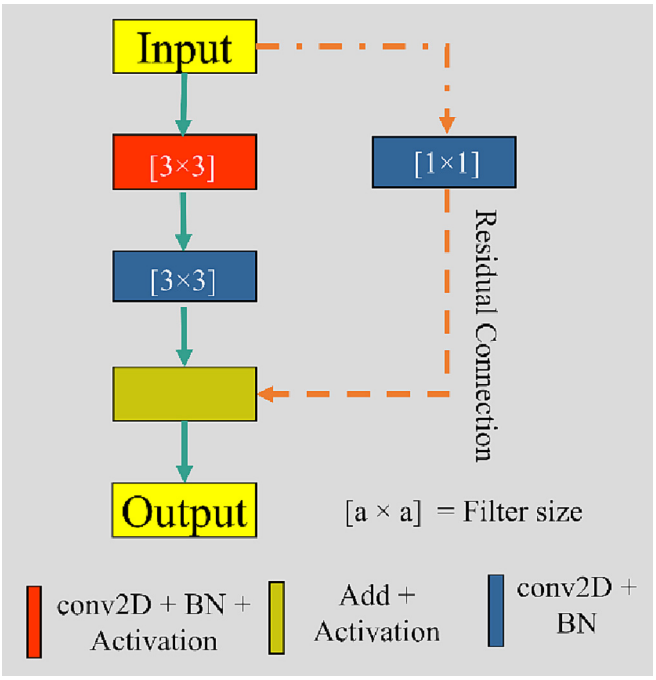


Fig. 3. Conceptual residual block for CNN model 2.

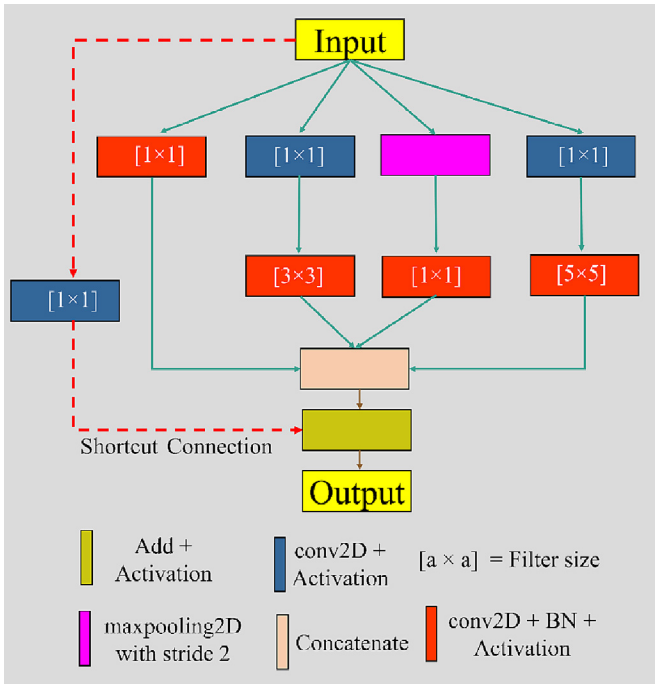


Fig. 5. Conceptual Inception-Residual block for CNN model 3.

globally while a smaller kernel size is preferred for the information which is present locally in an image. Therefore, inspired by Szegedy et al. (2014), an Inception-Residual block (IRB) as shown in Fig. 5 was built in the present study in which convolution on input with three different sizes of filters i.e., 1 × 1, 3 × 3, 5 × 5 was performed. A max-pooling operation was also performed with a kernel size of 3 × 3 with stride 2. Noteworthy, the 1 × 1 convolution proposed by Lin et al. (2014) was added before the 3 × 3 and 5 × 5 convolutions to limit the number of input channels, and so the computational expensiveness of Inception-Residual block (IRB). Additionally, a shortcut/residual connection is implemented in IRB to exploit the advantages of residual connections. The outputs from all these operations were concatenated and that was the output of the IRB. The output of IRB acts as input to the next layer.

The encoder part of model 3 consists of three convolutional blocks which comprised convolutional, batch normalization, and activation layers serially, three Inception-Residual blocks, and four max-pooling layers which were arranged as shown in Fig. 6. The decoder part consists of four transposed convolutions blocks which comprised of transposed convolutional, convolutional, batch normalization, and activation layers, serially to which feature maps from the encoder were concatenated using skip connections. The complete architecture along with the dimensions of feature maps and the number of filters is illustrated in Fig. 6.

2.4. Training of CNN models

Out of the total captured in-field cotton images, 10% of images were kept separate for the testing of developed CNN models, while the remaining 90% were used to create synthetic images which later utilize for training and validation purposes in 75:25 proportion, respectively. The optimization algorithm, learning rate, number of epochs, batch

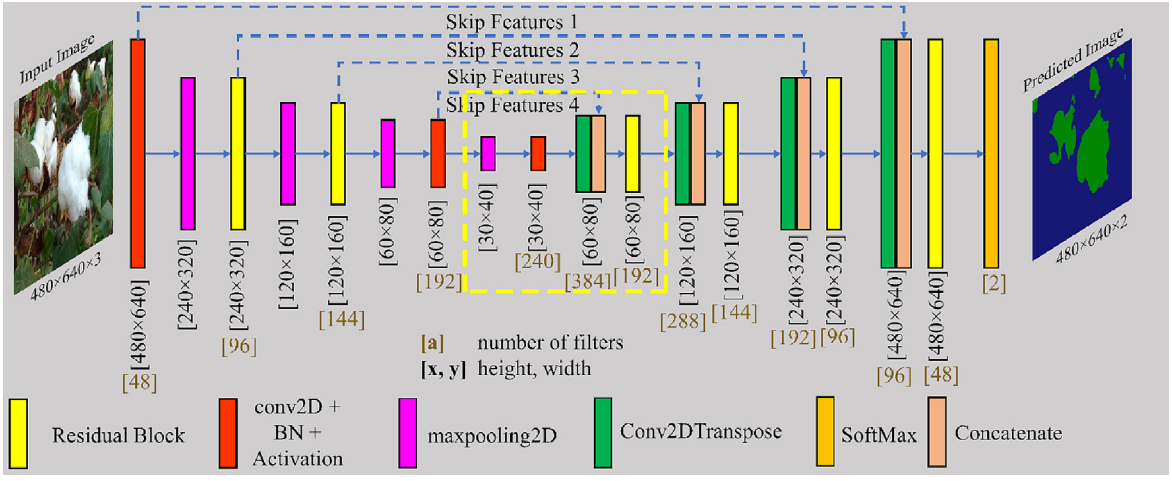


Fig. 4. Conceptual diagram of CNN model 2 (For yellow box refer to ablation experiments section).

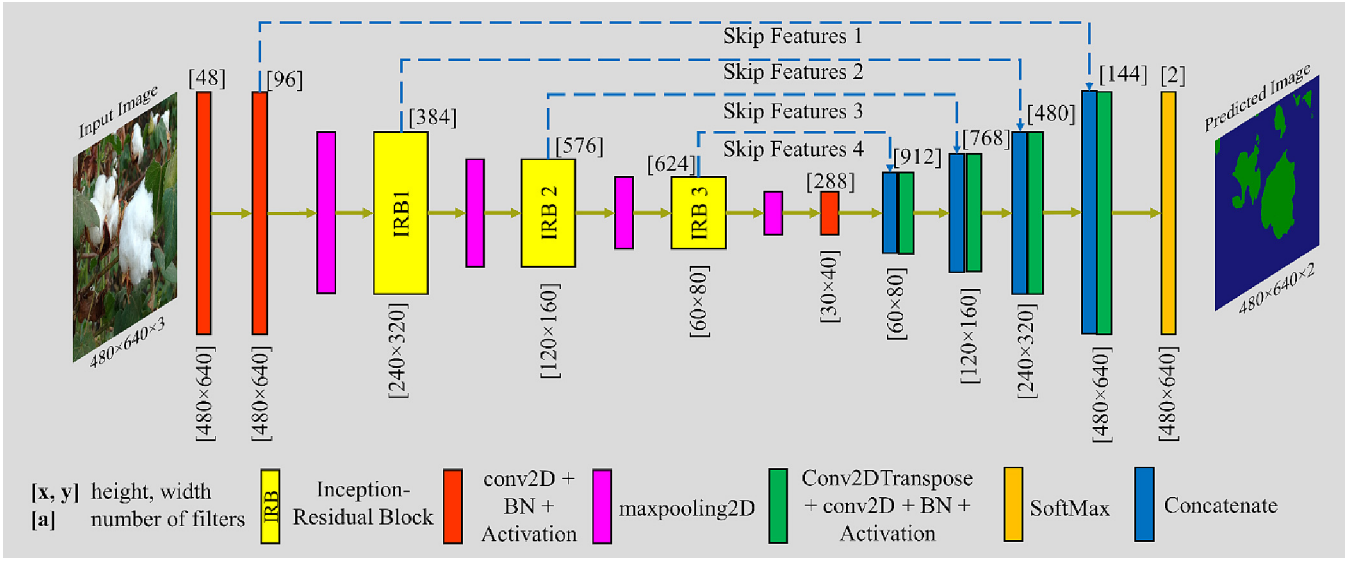


Fig. 6. Conceptual diagram of CNN model 3.

size, loss function, and weight initialization are some crucial hyperparameters that need to be tuned before training of neural network to achieve better results. The improper selection of these hyperparameters may result in poor generalization performance (Xing et al., 2018), overfitting, excess training time, over-consumption of computational resources (Xie et al., 2021), and may effects regularization significantly (Wilson and Martinez, 2003), etc. In the past, numerous researchers implemented the Adam optimizer (Bao et al., 2021; Gonzalez-Huitron et al., 2021; Waheed et al., 2020), used a learning rate of 0.001 (Gonzalez-Huitron et al., 2021; Kolhar and Jagtap, 2021; Zabawa et al., 2020), trained the models for one hundred epochs (Abdalla et al., 2019; Kolhar and Jagtap, 2021; Tassis et al., 2021; Zou et al., 2021) and achieved impressive results for agricultural images. Therefore, in the present study, models were trained for one hundred epochs having Adam optimizer (Kingma and Ba, 2017) as an optimization algorithm with a learning rate of 0.001 and a small batch size of five images as recommended by Kandel and Castelli (Kandel and Castelli, 2020).

Because of the existence of a higher class imbalance in the training dataset due to the substantial pixel ratio gap between cotton and background pixels, the accuracy of the trained model may not be satisfactory (Fujii et al., 2021). To deal with the class imbalance issue, multiple researchers (Badrinarayanan et al., 2017; Ngugi et al., 2020; Tang et al., 2020) successfully used the dice coefficient as a loss function in semantic segmentation tasks for agricultural images. Therefore, in the present study, the dice coefficient defined by Eq. (2) was used as a loss function in the training of proposed neural networks.

$$\mathrm{Dice}_{\mathrm{Loss}} = 1 - \frac{2 \times p_c^i \times q_c^i + \epsilon}{p_c^i + q_c^i + \epsilon} \tag{2}$$

where, $i = i^{th}$ sample in training data; $c = c^{th}$ class; p_c^i = one hot encoder of ground truth; q_c^i = probability of class c for i^{th} sample; $\epsilon = 1 \times 10^{-6}$.

A python-based open-source deep learning application programming interface named Keras v2.4.0 (Chollet, 2015) with an open-source python library called TensorFlow v2.6.1 (TensorFlow Developers, 2021) as its backend was used in this study to build the proposed models. Google Colaboratory (Google Colaboratory, 2021), an online cloud-based Jupyter notebook environment, was used to train proposed convolutional neural networks using available powerful hardware options as described in Table 2.

2.5. Selection of dropout rate

Dropout is a regularization technique, used to avoid overfitting in the neural networks by dropping out activation units randomly during training. This technique of regularization was successfully implemented in the fully connected layer of convolutional networks by various researchers for agricultural images (Khanramaki et al., 2021; Rahman et al., 2020; Waheed et al., 2020; Xie et al., 2021). You et al. (2020) used *DropBlock* (Ghiasi et al., 2018) and Li et al. (2017) added the dropout layer after the convolutional layer to avoid the overfitting of neural networks which build for weed detection and cotton segmentation, respectively. For analyzing the effects of the dropout layer on the performance of fully convolutional neural networks as well as to get the optimum value of dropout rate, proposed models having a fixed filter size of 3×3 in convolutional layers were trained with different dropout rates of 0%, 10%, 20%, 30%, and 40%. Except for the dropout rate, all other hyperparameters were kept the same for each case during training.

2.6. Selection of filter size

The size of the convolution filters determines the size of the receptive field from where information is extracted in convolutional layers. If the size of convolutional filters is small, then the extracted features will be highly local, while extracted features will be more expressive in case of a larger filter size (Szegedy et al., 2015). The in-field cotton bolls in captured images are present globally as well as locally for which larger filter size and smaller filter size should be preferred, respectively. To achieve the scale invariance for better performance from trained CNNs, the optimal filter size for convolutional layers was selected experimentally. For this, proposed models were trained with a filter size of 3, 5, and 7, while keeping other hyperparameters the

Table 2
Software and hardware specifications.

Name	Specifications
GPU	1xTesla K80, having 2496 CUDA cores, compute 3.7, 12 GB GDDR5 VRAM
CPU	1xsingle core hyperthreaded Xeon Processors @2.3Ghz
RAM	25 GB Available
Programming Language	Python

same for each case during training. The filter size which resulted in the best performance of the CNNs was selected to build the models.

2.7. Ablation study of architectures

A typical convolutional neural network consists of millions of trainable parameters (He et al., 2015; Ronneberger et al., 2015; Szegedy et al., 2015) out of which possibly some parameters may contribute meager or negligible to the output of the network and hence, can be removed without affecting the efficiency of the network (Molchanov et al., 2017). To remove the non-performing parameters of a CNN, ablation experiments were conducted by multiple researchers (Adhikari et al., 2019; Bao et al., 2021; Xie et al., 2021; You et al., 2020) in which certain components/layers of the network were pruned. The goal of the ablation study is to reduce the network size, computational expensiveness, training time, and inference time, providing the actual efficiency of the network is retained or degraded insignificantly.

In the present study, ablation experiments were conducted on proposed networks to reduce the size of networks and inference time while retaining the performance of networks. The ablation of neural networks in the present study was conducted in two ways: (i) by removing the number of layers, and (ii) by changing the number of filters in models. In model 1, from the encoder, the last max-pooling layer and the convolutional block was removed, while from the decoder, the first deconvolutional block was removed. The removed portion from model 1 is highlighted by yellow dotted lines in Fig. 2. This ablated model is referred to as model v1.1 in this study. In model 2, from the encoder, the last max-pooling layer and the convolutional block was removed, while from the decoder, the first deconvolutional layer, concatenate layer and residual block were removed. The removed portion from model 2 is highlighted by yellow dotted lines in Fig. 4. This ablated model is referred to as model v2.1. In model 3, Inception-Residual block number 3 (IRB3) was replaced with one convolutional block layer and referred to as model v3.1. For another ablation using model 3, IRB2 and IRB3 were replaced with a single convolutional block for each IRB and referred to as model v3.2. Regarding ablation experiments through the number of filters, for each model and its variants, the filters were decreased to 80%, 60%, and 40% from the numbers of filters stated in Figs. 2, 4, and 6. The performance of models and their variants were evaluated using evaluation matrices to select the best models among these in terms of segmentation accuracy, network size, and inference time.

2.8. Evaluation indexes

For the performance evaluation of proposed neural networks and comparisons among these, segmentation results obtained using networks were quantified using well-known evaluation metrics for agricultural images: F1-score (Chen et al., 2021; Kestur et al., 2019; Kolhar and Jagtap, 2021), intersection-over-union (IoU) (Sun et al., 2021; Zabawa et al., 2020; Zou et al., 2021), mean intersection-over-union (mean-IoU) (Long et al., 2015; Tassis et al., 2021; Xu et al., 2020), pixel accuracy (Kestur et al., 2019; Kolhar and Jagtap, 2021; Tassis et al., 2021), precision (Kestur et al., 2019; Zabawa et al., 2020) and recall (Chen et al., 2020; Tassis et al., 2021; Zou et al., 2021) metrics. IoU metric as defined by Eq. (3), measures the ratio of the number of pixels common between the predicted, and ground truth image and the total number of pixels present across both images. Mean-IoU is the average of IoU scores over semantic classes and is defined by Eq. (4). Pixel accuracy as defined by Eq. (5), measures the percent of pixels that were classified correctly. For a given class i, a precision metric can be defined as the fraction of correctly classified pixels of class i among the total pixels classified as class i pixels, whereas recall metric measures the fraction of correctly classified pixels of class i among the actual number of pixels of class i present in ground truth image. Precision and recall metrics are defined by Eqs. (6) and (7), respectively. The F1-score combine the precision and recall metrices into a single parameter for better analyzing and defined by Eq. (8).

$$IoU^i = \frac{TP_i}{TP_i + FN_i + FP_i} \tag{3}$$

$$mean - IoU = \frac{1}{k}\sum_{i=1}^{k} \frac{p_{ii}}{\sum_{j=1}^{k} p_{ij} + \sum_{j=1}^{k} p_{ji} - p_{ii}} \tag{4}$$

where p_{ij} represents the number of pixels of the class *i* that were predicted to be class j and k represents the number of semantic classes.

$$Pixel\ Accuracy^i = \frac{TP_i + TN_i}{TP_i + FN_i + FP_i + TN_i} \tag{5}$$

$$Precision^i = \frac{TP_i}{TP_i + FP_i} \tag{6}$$

$$Recall^i = \frac{TP_i}{TP_i + FN_i} \tag{7}$$

$$F1 - Score^i = \frac{2 \times Precision^i \times Recall^i}{Precision^i + Recall^i} \tag{8}$$

where true positive/true negative (TP_i/TN_i) are the pixels belonging to class *i* that is predicted correctly, false positive (FP_i) are pixels that are predicted as class *i* pixels but do not belong to class *i*, false negative (FN_i) are pixels that belong to class *i* but incorrectly classified as pixels of another class.

2.9. Comparison with the state-of-the-art networks

On considering the success of convolutional neural networks for classification tasks, Long et al. (2015) developed the first end-to-end trainable fully conventional network for image segmentation tasks. Later, Ronneberger et al. (2015) developed a novel CNN model for biomedical image segmentation. The architecture of their proposed model mainly consists of encoder and decoder portions in which the encoder learned the low-level features from the input image, while the decoder semantically maps those learned features onto the image pixels of the same size as of input image. Subsequently, numerous researchers (Adhikari et al., 2019; Badrinarayanan et al., 2017; Kolhar and Jagtap, 2021; Milioto et al., 2018; Peng et al., 2019) used convolutional encoder-decoder networks for semantic segmentation. The Visual Geometry Group (VGG) (Simonyan and Zisserman, 2015), ResNet (He et al., 2015), and InceptionV3 (Szegedy et al., 2015) networks achieve top-5 accuracy of 92.7%, 93.3%, and 93.9%, respectively on ImageNet dataset (Deng et al., 2009), which proves that these networks have good features extraction ability and often (Gao et al., 2020; Hecht et al., 2020; Majeed et al., 2018; Ou et al., 2019; Panda et al., 2022; Shah et al., 2022; Zou et al., 2021) used as a backbone in various CNN architectures developed for semantic segmentation. This study compared the proposed models to the aforementioned models while using the aforementioned networks as the encoder backbone. For this purpose, fully connected dense layers were pruned from these CNN models and an architecture similar to the U-Net model (Ronneberger et al., 2015) was constructed using up-sampling layers. Additionally, rather than starting the training process by randomly initializing the weights, weights of respective networks, trained on the ImageNet dataset (Deng et al., 2009) were used, while keeping the other hyperparameters unchanged from the training of the proposed models in the study. Tan and Le (2019) proposed a new model named EfficientNet with which they achieved better performance by balancing network depth, width, and resolution. For agricultural images too, multiple researchers (Atila et al., 2021; Duong et al., 2020; Yin et al., 2020; Zhang et al., 2020)

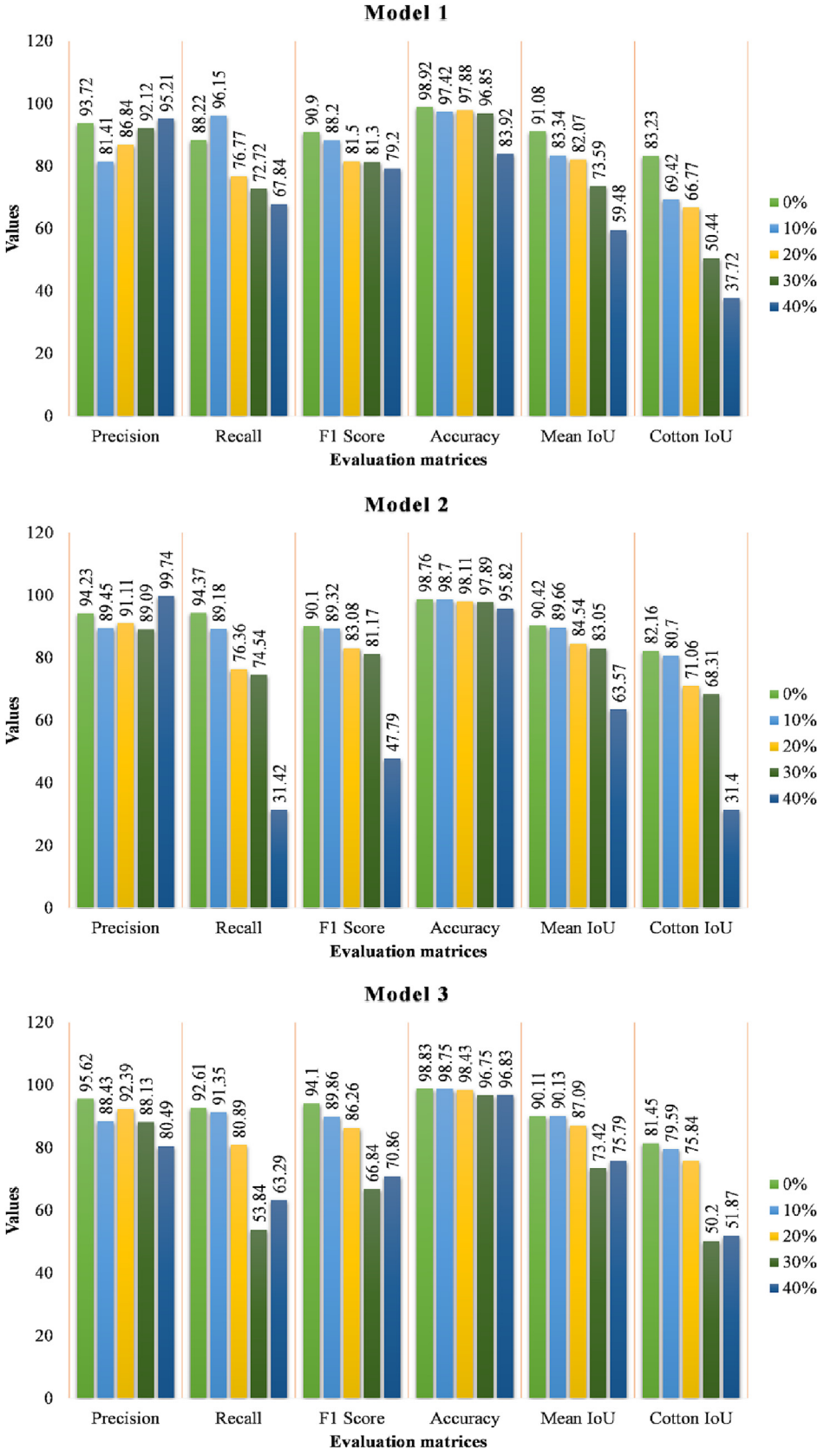


Fig. 7. Effects of dropout rate on the performance of proposed models.

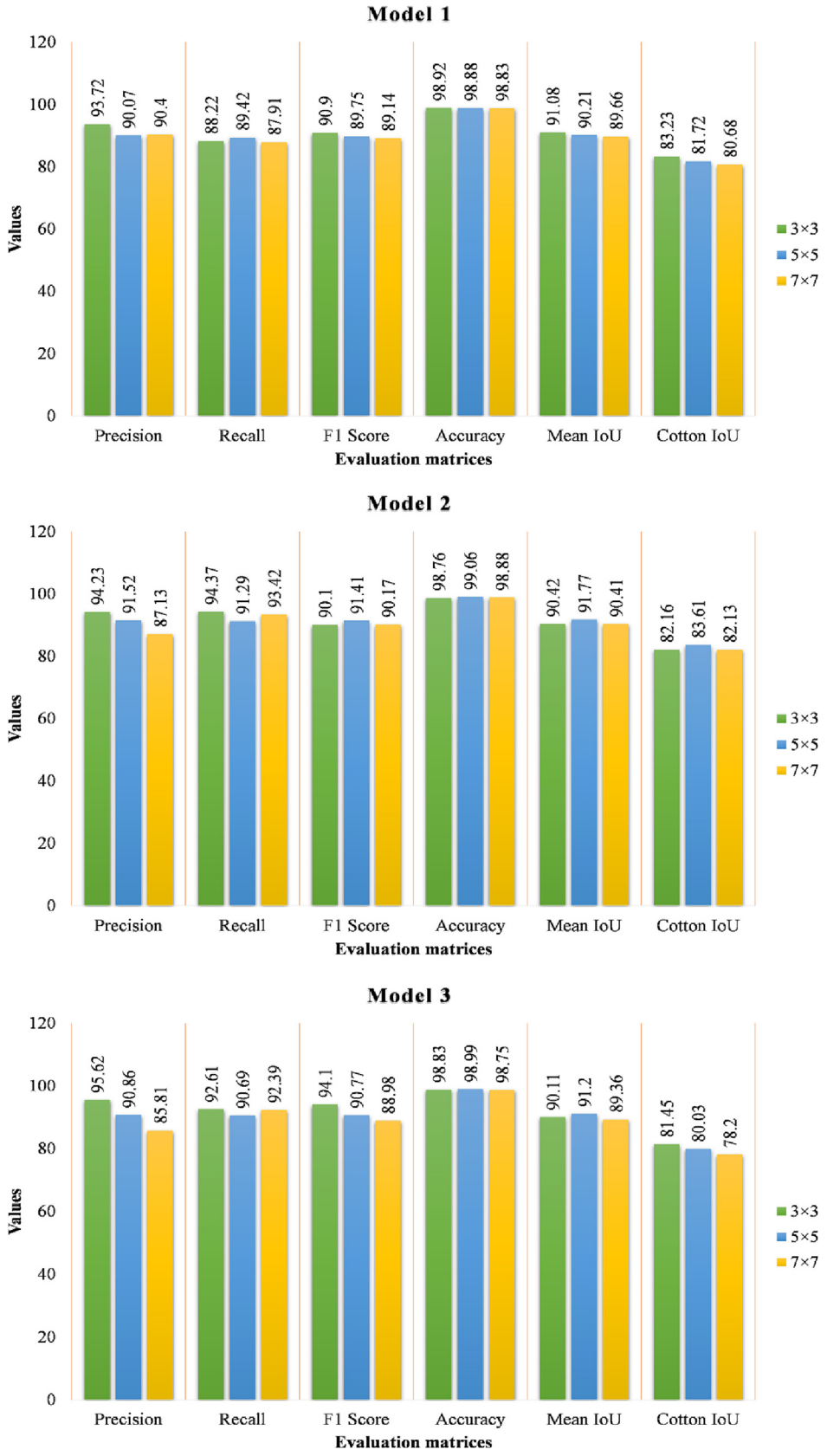


Fig. 8. Effects of filter size on the performance of proposed models.

achieved desirable results by implementing the various models of EfficientNet family in their studies. The EfficientNet models are smaller and faster as compared to other existing state-of-the-art models (Tan and Le, 2019), therefore, for comprehensiveness in evaluation, light-weight models proposed in the present study were also compared with the EfficientNet-B1 model.

3. Results and discussion

3.1. Effects of dropout rate on the performance of convolutional networks

The segmented results from models that were trained using different dropout rates were compared. From Fig. 7, it can be observed that with the increase in the dropout rate, the performance of the proposed models significantly decreased. The mean IoU is an important and commonly used parameter to evaluate the performance of models developed for semantic segmentation. It was observed that when the dropout rate increased from 0% to 40%, the mean- IoU value decreased by 31.60%, 14.32%, and 26.85% for model 1, model 2, and model 3, respectively. In addition, to mean-IoU values, cotton-IoU (Intersection-over-Union), F1-score, and pixel accuracy values also decreased significantly when dropout rates increased. Our analysis in this study shows that the dropout layer present in a fully convolutional neural network is ineffective and may decrease its overall performance.

The reason for the failure of the dropout technique is that the proposed models are fully conventional neural networks and cotton field images have strong spatial correlation and hence the extracted feature maps too. As feature maps activations are spatially correlated, therefore, despite the dropping-off of some random activations' units from feature maps, information still flows through the convolutional layers and the fully convolutional neural network shows no positive effect of dropout on its performance. Similar observations were also mentioned by Ghiasi et al. (2018) and Tompson et al. (2015) in their respective studies.

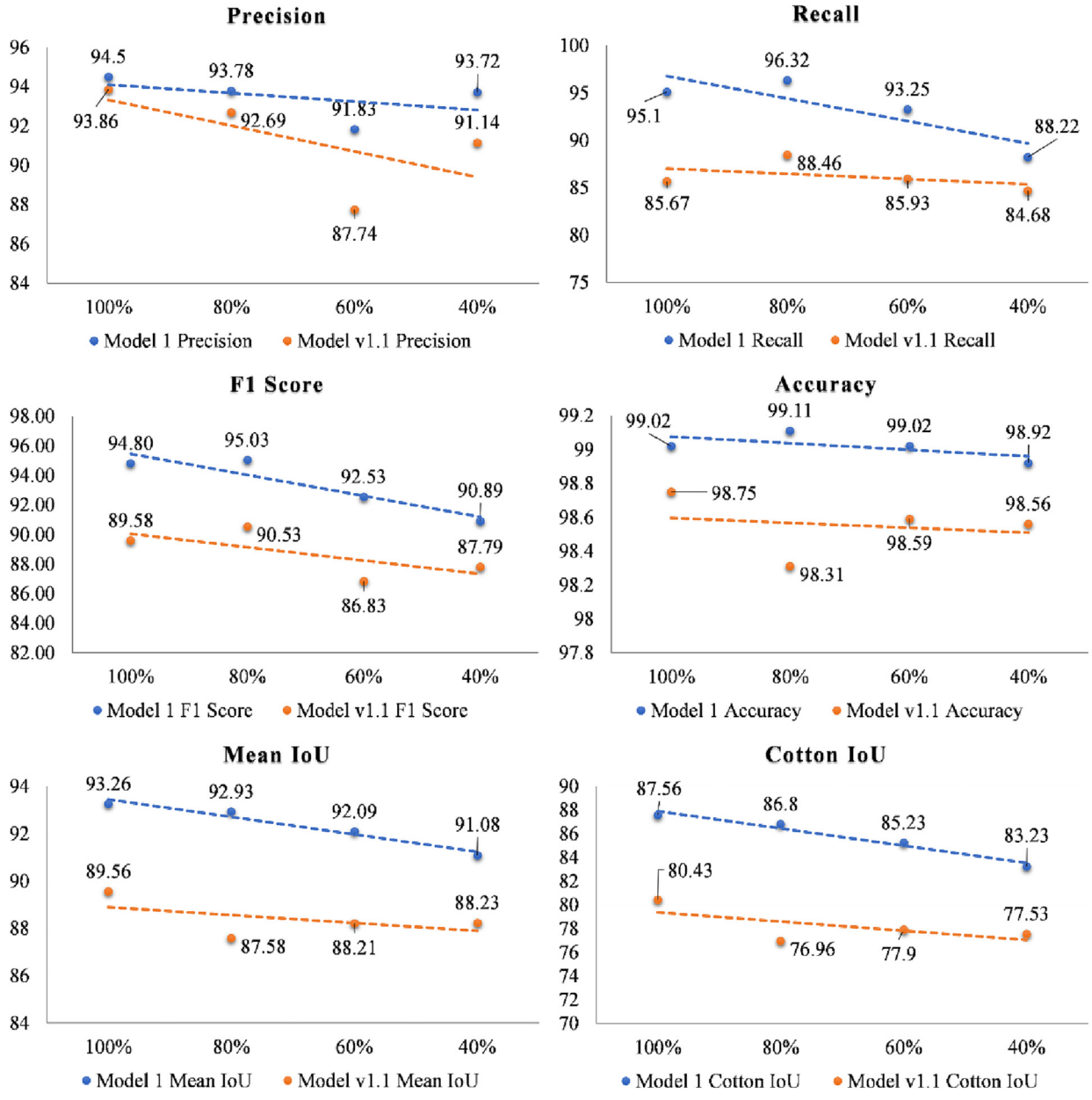


Fig. 9. Ablation study results in terms of evaluation matrices for model 1 and its variant.

3.2. Effects of filter size on the performance of convolutional networks

The segmented results from the models that were trained using filter sizes of 3, 5, and 7 in convolutional layers were compared to select the best filter size. From Fig. 8, it can be observed that there is an insignificant difference between the performance of proposed models while trained with different filter sizes. For the filter size of 3, 5, and 7, the mean IoU value for model 1, model 2, and model 3 was changed marginally while the computational expensiveness increased by 2.78 times and 5.45 times (Szegedy et al., 2015) for 5×5 and 7×7 filter size. Moreover, in terms of F1-score, accuracy, and cotton-IoU values too, no significant advantage was observed in favor of a larger filter size for the proposed models. Therefore, a 3×3 filter size was used in the convolutional neural networks of this study.

3.3. Results of ablation experiments

To reduce the size of models and their inference time, ablation experiments on three proposed models were conducted for which layer/components were pruned from the networks in the first approach while in the second approach, the number of filters was altered in models as well as in their derived variants. Figs. 9, 10, and 11 show the performance comparison of model 1, model 2, and model 3, respectively which were trained with 100%, 80%, 60%, and 40% of the total number of filters used in the proposed models as mentioned in Figs. 2, 4, and 6. The value of 100% denotes no change in the number of filters, while 80% denotes the 20% lesser number of filters as compared to the number of filters stated in Figs. 2, 4, and 6. A comparison was also made between the proposed models and their respective variants having a different number of filters. Zou et al. (2021) developed a convolutional neural network that outperform the U-net model by 1.78%, the DeepLabv3 network in Chen et al. (2021) outperform the U-net by 0.7%, and the AM-DeepLabv3 network in Kang et al. (2021) outperform the Sub-pixel-DeepLabv3+ network by 0.77% in terms of IoU score. Therefore, a threshold value of 0.80% was used in this study to categorize the change in the IoU score into significant or insignificant. If a change in the IoU score is greater than 1.0%, then only the change was considered significant.

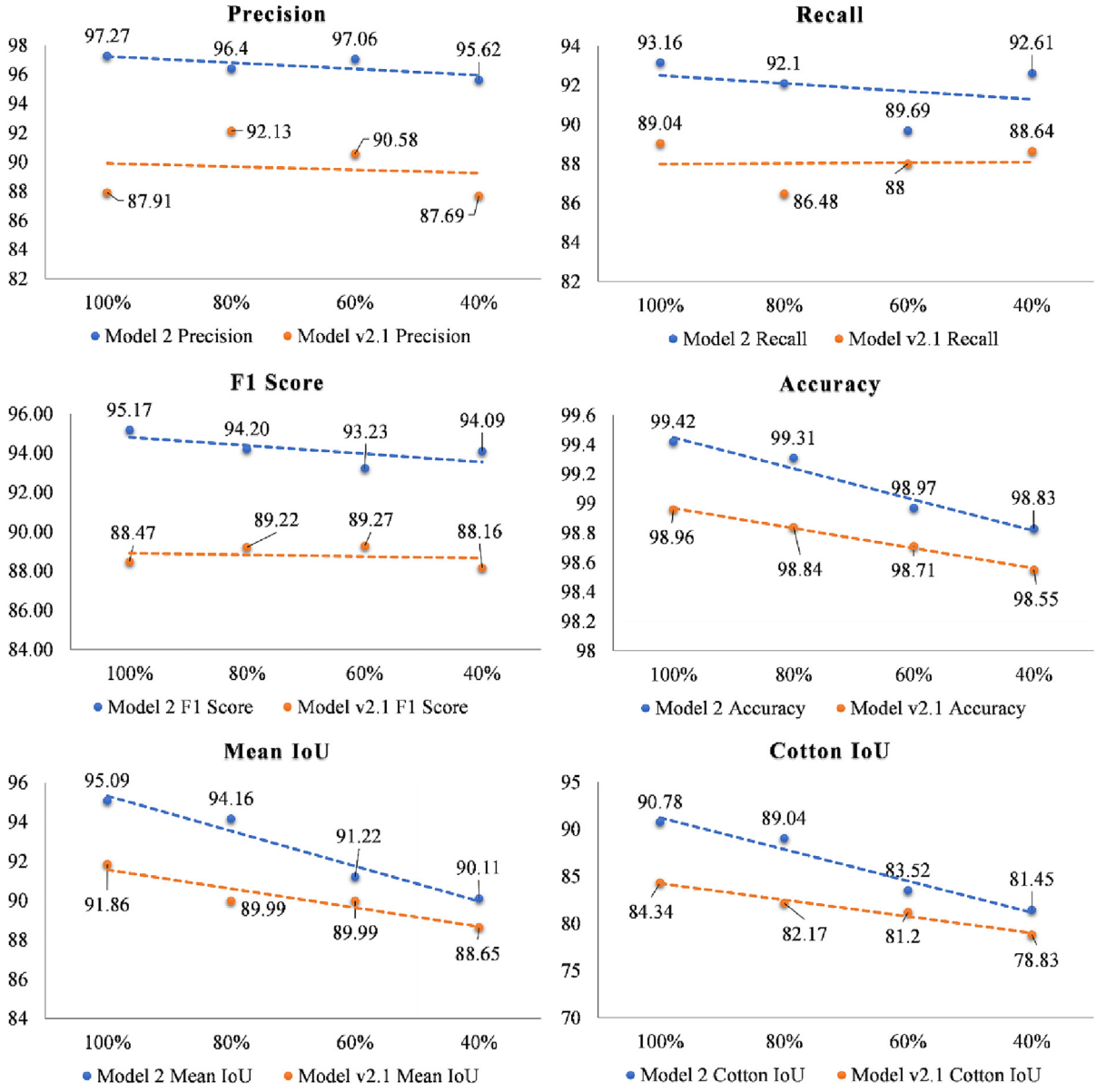


Fig. 10. Ablation study results in terms of evaluation matrices for model 2 and its variant.

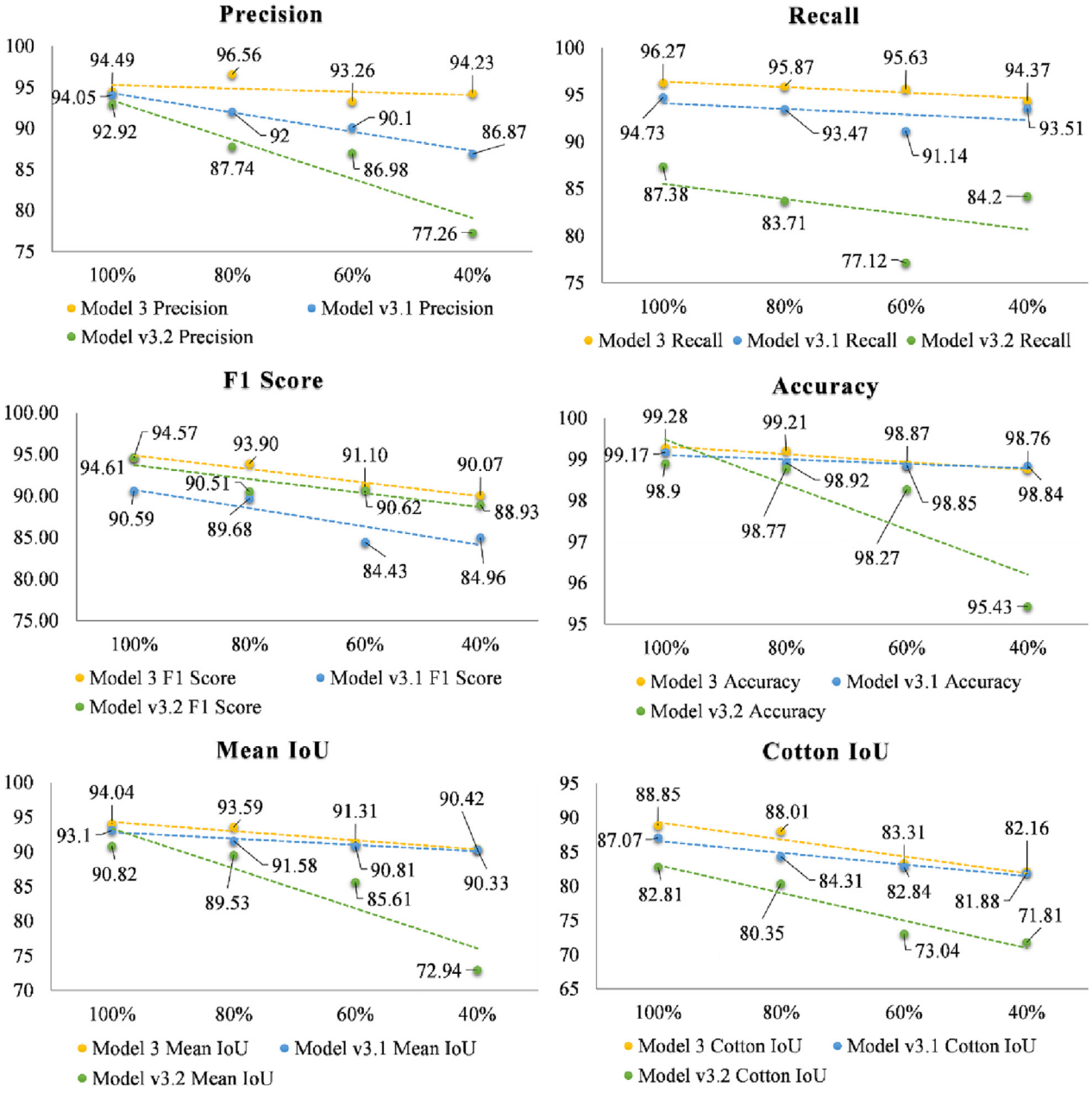


Fig. 11. Ablation study results in terms of evaluation matrices for model 3 and its variant.

The mean-IoU is the most important parameter to evaluate the performance of semantic segmentation. The value of the mean-IoU score, as observed in Figs. 9, 10, and 11, was decreased by 2.18%, 4.98%, and 3.62% on reducing the number of filters by 60% in model 1, model 2, and model 3, respectively. For 100% of the number of filters, a difference of 3.7% between model 1 and model v1.1, 3.23% between model 2 and model v2.1, and 3.22% between model 3 and model v3.2 was observed in the mean-IoU score. Similar patterns were also observed for precision, recall, F1-score, pixel accuracy, and cotton-IoU score of models and their respective variants. So, from Figs. 9, 10, and 11 we can conclude that either the number of layers/components gets pruned from the network or the number of filters gets reduced, and the performance of models and their respective variants gets decreased.

In addition to segmentation accuracy, inference time is also an important parameter that should be considered for the evaluation of models. Therefore, along with the cotton-IoU score, the inference time of models was compared as shown in Fig. 12. It can be observed from Fig. 12 that for 100% of the number of filters, there is a 7.13% reduction in cotton-IoU score for model v1.1 as compared to model 1 against a 15.34% reduction in inference time. A comparison between model 1 with no reduction in the filter and a 20% reduction in the number of filters shows that the cotton-IoU score decreased by 0.76% against a 20.45% reduction in inference time. The inference time will be further reduced to 298 ms by reducing the number of filters to 60%, but the cotton-IoU score also dropped by 2.0%. Hence, model 1 with a 20% fewer number of filters was selected for further fine-tuning. In the case of model v2.1 with 100% filters, there is a 6.44% reduction in cotton-IoU score as compared to model 2 against a 9.97% reduction in inference time. Comparing model 2 with no reduction in filters and a 20% reduction in the number of filters shows that the cotton-IoU score decreased by 1.74% which is considered to be a significant change based on the threshold value of 0.80% as described earlier. Hence, model 2 with a 100% number of filters was selected for further fine-tuning. A comparison of model 3 having no reduction in the number of filters with model v3.1 and model v3.2 shows

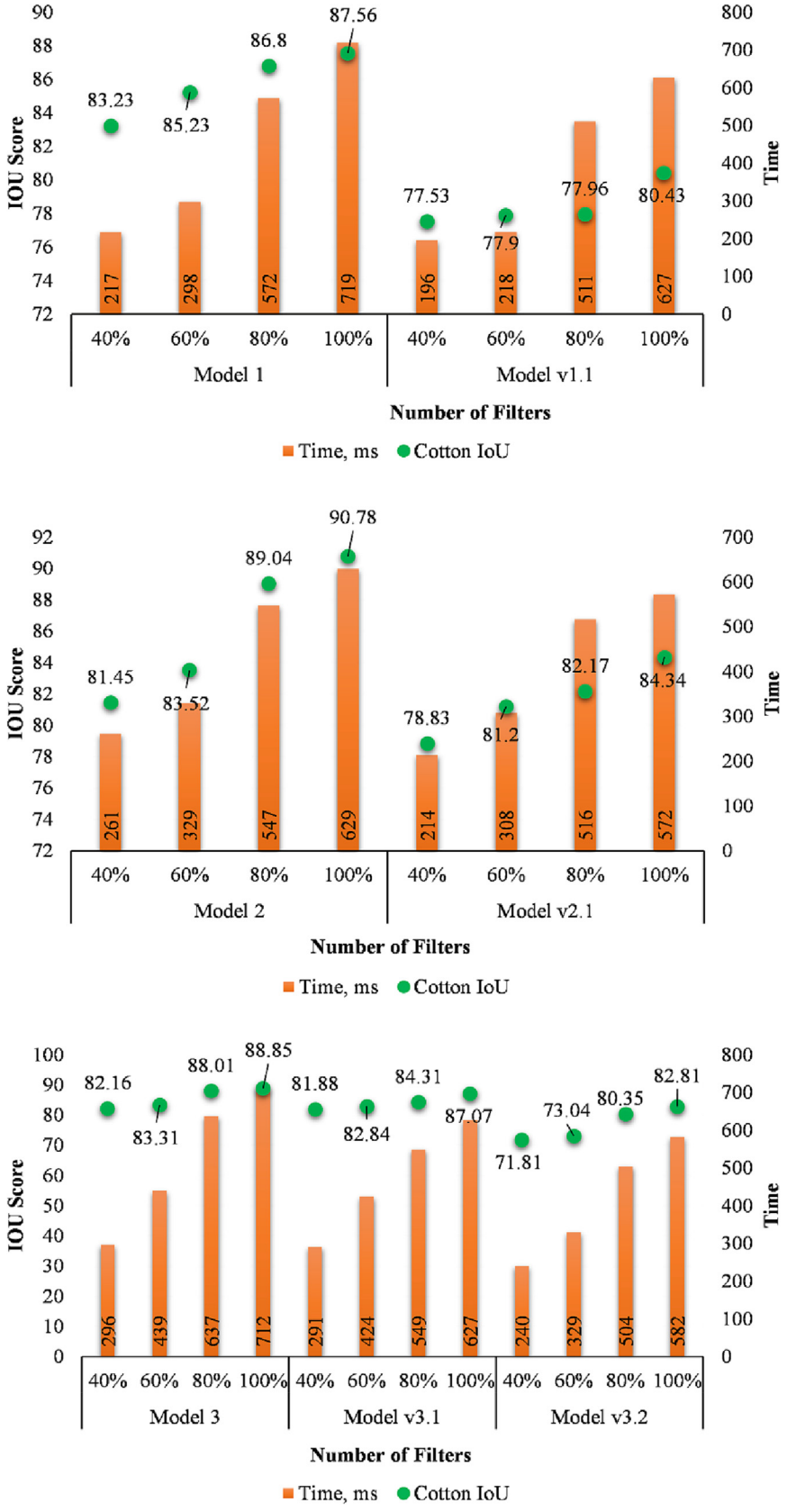


Fig. 12. Comparison of models and their variants in terms of cotton-IoU score and inference time.

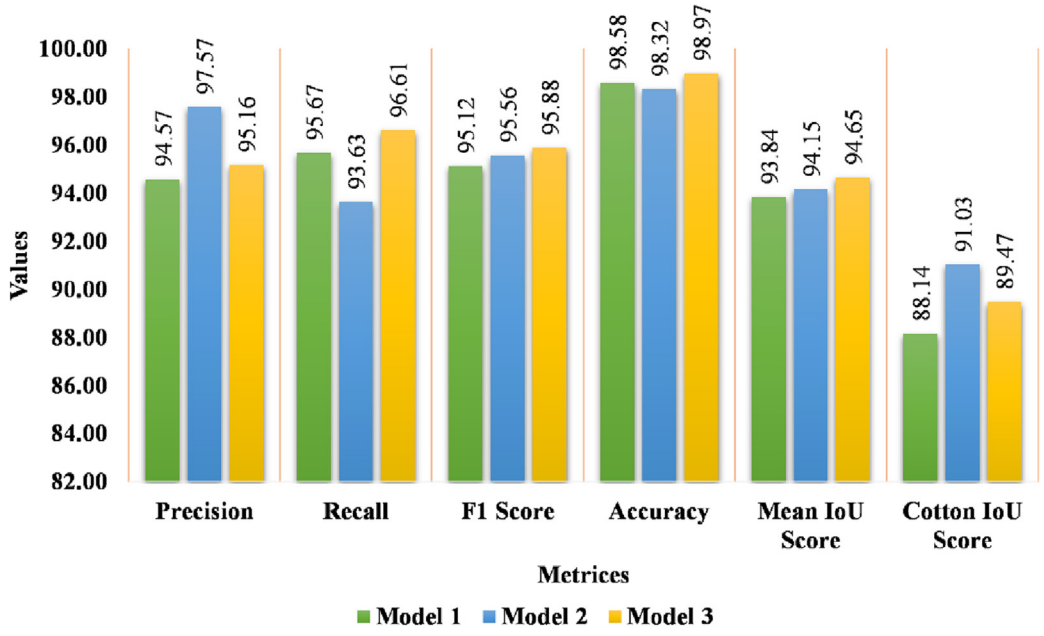


Fig. 13. Comparison of finally selected, and fine-tuned models in terms of evaluation matrices.

the 1.78% and 6.04% reduction in cotton-IoU score against 11.9% and 13.2% reduction in inference time. A comparison between model 3 with no reduction in filters and a 20% reduction in the number of filters shows that the cotton-IoU score decreased by 0.84% against a 10.53% reduction in inference time. In this case, the reduction in inference time is marginal plus the reduction in cotton-IoU score crossed the threshold value, and therefore, model 3 with a 100% number of filters was selected for further fine-tuning.

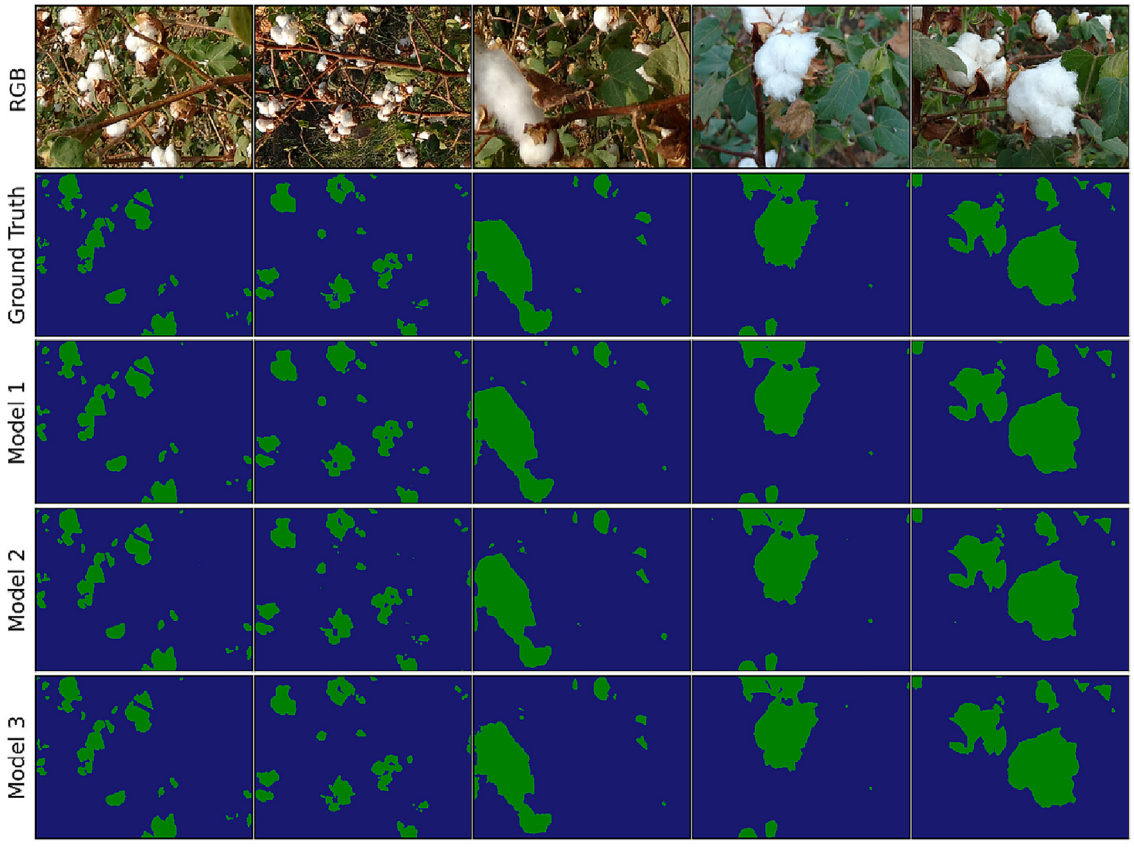


Fig. 14. Qualitative visualization of semantic segmentation results of cotton bolls obtained using developed models.

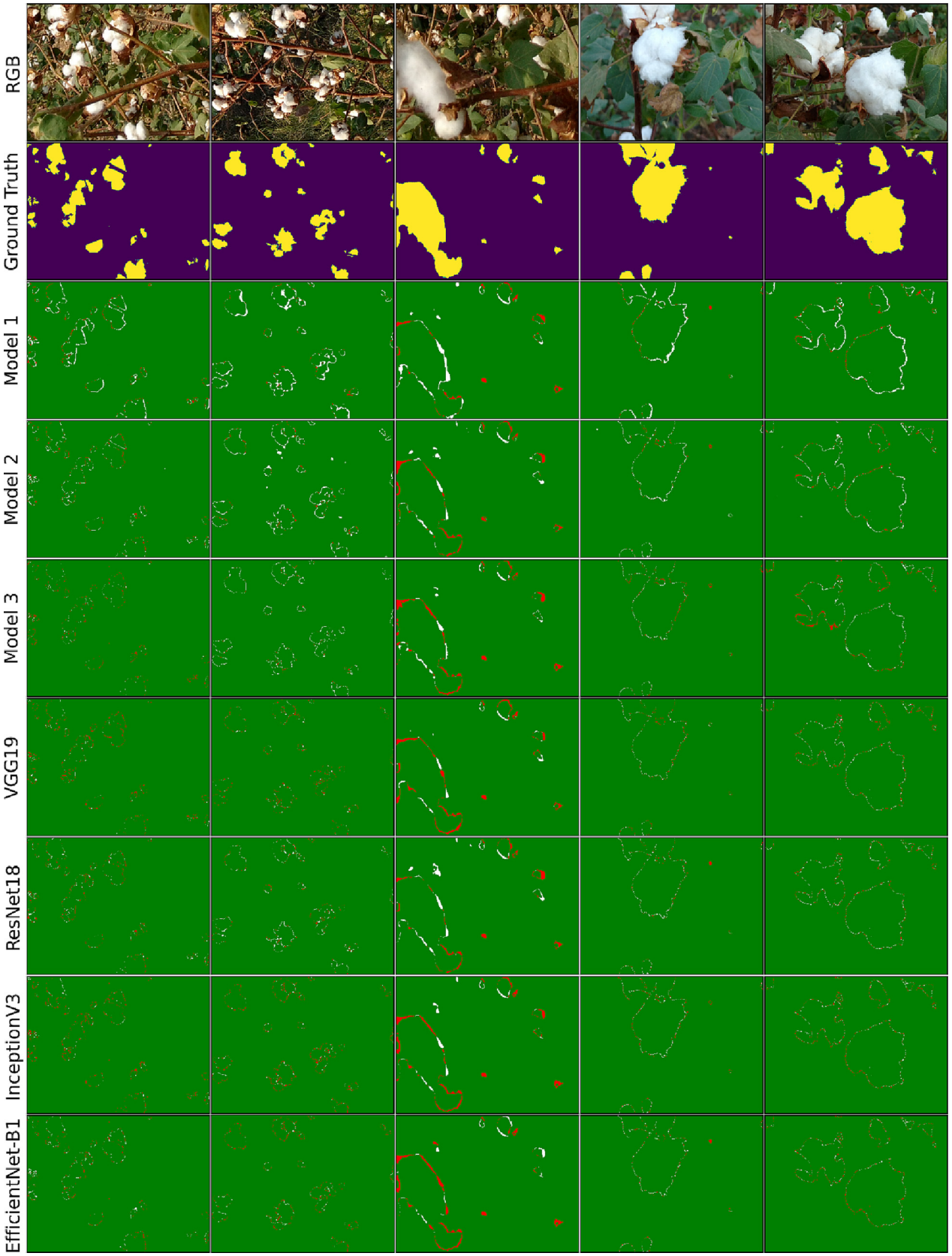


Fig. 15. Qualitative performance comparison of developed models with state-of-the-art convolutional networks for semantic segmentation of cotton bolls (The green, white, and red colors indicate the correct classification, false positive and false negative, respectively).

Table 3
Quantitative evaluation of segmentation results obtained using proposed models and state-of-the-art convolutional networks along with trainable parameters.

Model	Trainable parameters, M	Mean-IoU score, %	Precision, %	Recall, %	F-1 score, %
VGG19	29.05	95.39	93.47	96.13	94.79
ResNet18	14.33	96.54	96.10	95.84	95.96
InceptionV3	29.89	96.37	96.68	95.78	96.22
EfficientNet-B1	12.58	96.40	96.56	96.42	96.48
Model 1	3.48	93.84	94.57	95.67	95.12
Model 2	3.43	94.15	97.57	93.63	95.56
Model 3	8.17	94.65	95.16	96.61	95.88

3.4. Performance evaluation of neural networks

After analyzing the models in terms of segmentation accuracy and inference time in Section 3.3, it was observed that with a 20% reduction of filters in model 1, inference time was reduced by 20.45% against marginal (<0.77%) decrease in cotton-IoU score, while in case of model 2 and model 3, reduction in the number of filters resulted into a substantial reduction in cotton-IoU score. Therefore, considering the segmentation accuracy as well as inference time of proposed models and their variants, three models were selected for further fine-tuning. For ease of reference, these selected models will be designated as: *Model 1*: model 1 with 80% of the number of filters; *Model 2*: model 2 with 100% of the number of filters; *Model 3*: model 3 with 100% of the number of filters. These three selected models were trained for one-hundred epochs having Adam optimizer as an optimization algorithm with a learning rate of 0.001. The performance of these three trained models was evaluated using evaluation indexes as discussed earlier. From Fig. 13, it can be observed that the mean-IoU and cotton-IoU scores are above 93% and 88%, respectively for all three models. Model 3 outperforms the other two models with a pixel accuracy value of 98.97% and a mean-IoU score of 94.65%. In terms of the cotton-IoU score, the highest value of 91.03% was achieved by model 2. Li et al. (2016) employed a region-based semantic segmentation method to detect the in-field cotton bolls and achieved accuracy and IoU value of 97.0% and 73.5% respectively, while in another study, Li et al. (2017) achieved accuracy and cotton-IoU value of 97.14% and 70.5%, respectively using fully convolutional neural network for cotton segmentation. An open boll detection algorithm was proposed by Yeom et al. (2018) in which they achieved the highest values of 92.5%, 96.6%, 95.6%, and 96.1% for IoU, precision, recall, and F1-score, respectively. Singh et al. (2021) achieved precision and recall values of 97.53% and 88.69%, respectively using their proposed algorithms which were developed using color features. After analyzing the performance of the proposed models quantitively, it can be concluded that all three developed models can segment the in-field cotton bolls with higher accuracy. Fig. 14 shows the semantic segmentation results of cotton bolls obtained using model 1, model 2, and model 3. It can be visually observed that the proposed models segmented the single as well as multiple instances of in-field cotton bolls successfully.

3.5. Performance comparison with the state-of-the-art CNN models

For a comprehensive evaluation of the proposed models based on segmentation accuracy and inference time, four state-of-the-art convolutional networks namely InceptionV3, ResNet18, VGG19, and EffecientNet-B1 were used for the semantic segmentation of cotton bolls and segmentation results were compared with the proposed models. Fig. 15 shows the qualitative performance comparison of developed models with state-of-the-art convolutional networks. For a better comparison of the segmentation performance of networks, false positive and false negative errors for each model are shown in Fig. 15. The false-positive and false-negative errors were obtained using ground truth images. The green, white, and red colors in Fig. 15 indicate the correct classification, false positive and false negative, respectively. It can be observed visually from Fig. 15 that VGG19, ResNet18, EffecientNet-B1, and InceptionV3 perform slightly better than the proposed models as the number of white color pixels (false positive) and red color pixels (false negative) are comparably high in segmented results obtained using the proposed models. The slightly superior performance of state-of-the-art networks as compared to the proposed models is justifiable as the number of trainable parameters in state-of-art networks is multiple times higher than the proposed models. Table 3 shows the quantified comparison between state-of-the-art networks and proposed networks along with trainable parameters. Model 2 achieved the mean-IoU value of 94.15% while ResNet18 achieved the highest value of 96.54% among the state-of-the-art models. Therefore, despite having 76.06% fewer trainable parameters compared to ResNet18, the performance of model 2 in terms of mean-IoU and accuracy declined by 1.39% and 1.17%, respectively which is a marginal decline as compared to the decline in the number of trainable parameters. Model 3

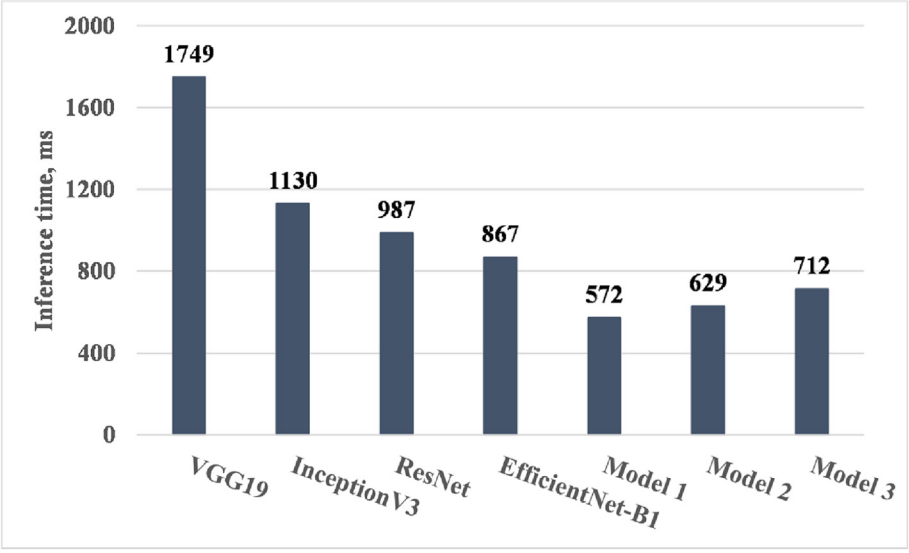


Fig. 16. Inference time of developed models and state-of-the-art convolutional networks for semantic segmentation of cotton bolls.

Table 4
Comparison of inference time among the CNN models.*

Model	InceptionV3	VGG19	ResNet	EfficientNet-B1	Model 1	Model 2	Model 3
InceptionV3	1.00						
VGG19	1.55	1.00					
ResNet18	0.87	0.56	1.00				
EfficientNet-B1	0.77	0.50	0.88	1.00			
Model 1	0.51	0.33	0.58	0.66	1.00		
Model 2	0.56	0.36	0.64	0.73	1.10	1.00	
Model 3	0.63	0.41	0.72	0.82	1.24	1.13	1.00

* Read as: Computationally, based on inference time, the VGG19 model is 1.55 times more expensive compared to the Inception model.

which has the highest number of trainable parameters among proposed models, achieved a mean-IoU of 94.65% which is just 1.89%, 1.72%, 1.75%, and 0.74% lesser than the values achieved using ResNet18, InceptionV3, EfficientNet-B1, and VGG19 networks, respectively, while the number of trainable parameters in these networks is higher by many folds.

For the neural networks developed for real-time application, inference time should be taken into account before its deployment. The segmentation algorithm with a higher inference time is not suitable for harvesting robots. Because, the higher inference time restricted its real-time application to some extent (De-An et al., 2011) as the theoretical field capacity of harvesting robots (Sakai et al., 2008) decreased proportionally with an increase in inference time. Therefore, the inference time required by networks for semantic segmentation of cotton bolls per image was compared in the present study. The combination of average time needed for model loading, image loading, pre-processing, and inference time is termed inference time in this study. To measure the inference time, in the present study, 15 numbers of images were used for predictions and the average time taken for their prediction was recorded. This procedure was repeated five times and the average of five rounds was considered as inference time. For this, a laptop having specifications of 8 GB RAM, Intel Core i5 CPU (without GPU), and Windows 10 operating system was used. Proposed models will be deployed to hardware having similar specifications as specified above, therefore, inference time was calculated using the same hardware. Fig. 16 shows the inference time of each network. The inference time for all proposed models is at least 22.0% less as compared to state-of-the-art networks. Among all state-of-the-art networks, EfficientNet-B1 consumed the lowest inference time of 867 ms, followed by the ResNet18 model with 987 ms. The proposed model 1 consumed the lowest inference time of 572 ms which is 52.0% less as compared to the EfficientNet-B1 network. Table 4 shows the comparison of inference time between the proposed models as well as state-of-the-art models. In comparing the EfficientNet-B1 model (smallest among the state-of-the-art models used in the present study) with the proposed Model 3 (biggest among the proposed models), it can be observed from Table 4 that the inference time for state-of-the-art models is at least 1.22 times higher than the proposed models. The inference time for VGG19 is 3.06 times higher when compared to model 1 against the marginal drop (1.55%) in the mean-IoU score for model 1. We can conclude that all proposed models are significantly faster than the state-of-the-art networks and hence, deployment of these developed models will increase the field capacity of harvesting robots with marginal losses in segmentation accuracy.

4. Conclusions

The detection of in-field cotton bolls is a challenging task due to various factors associated such as environment lighting, complex background, and non-uniformity in visual characteristics of cotton bolls. This study proposed three fully convolutional neural network models for the semantic segmentation of in-field cotton bolls. In these models, residual connections, skip connections, and multiple filter sizes were incorporated to enhance the segmentation performance. Effects of dropout rate in the convolutional layer and filter size were evaluated and found that in a convolutional layer, dropout is ineffective, while an increase in filter size reduced the segmentation accuracy for the cotton dataset. The cotton-IoU for all proposed models was found to be above 88.0%, in model 2 achieved the highest cotton-IoU of 91.03%. The proposed models achieved an F1-score and pixel accuracy values greater than 95.0% and 98.0%, respectively. Developed models were also compared with state-of-the-art networks such as VGG19, ResNet18, EfficientNet-B1, and InceptionV3. Despite having a limited number of trainable parameters as compared to state-of-the-art networks, the proposed models performed well with mean-IoU of 93.84%, 94.15%, and 94.65% against the mean-IoU value of 95.39%, 96.54%, 96.40%, and 96.37% obtained using VGG19, ResNet18, EfficientNet-B1, and InceptionV3 models, respectively. The inference time for the proposed models, which is an important factor for the real-time application of neural networks, was reduced up to 52.0% compared to state-of-the-art networks. Hence, it can be concluded that the three lightweight fully convolutional neural network models proposed in this study segment the in-field cotton bolls faster with greater accuracy and can be deployed to the cotton harvesting robots for real-time recognition of cotton bolls for harvesting.

CRediT authorship contribution statement

Naseeb Singh: Software, Conceptualization, Methodology, Investigation, Writing – original draft, Data curation. **V.K. Tewari:** Supervision, Resources, Writing – review & editing. **P.K. Biswas:** Supervision, Writing – review & editing. **L.K. Dhruw:** Software, Validation, Visualization, Data curation.

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.

Acknowledgments

The financial support was received from the Indian Institute of Technology Kharagpur, Kharagpur, West Bengal, India.

References

Abdalla, A., Cen, H., Wan, L., Rashid, R., Weng, H., Zhou, W., He, Y., 2019. Fine-tuning convolutional neural network with transfer learning for semantic segmentation of ground-level oilseed rape images in a field with high weed pressure. Comput. Electron. Agric. 167, 105091. https://doi.org/10.1016/j.compag.2019.105091.

Adhikari, S.P., Yang, H., Kim, H., 2019. Learning semantic graphics using convolutional encoder–decoder network for autonomous weeding in paddy. Front. Plant Sci. 10, 1404. https://doi.org/10.3389/fpls.2019.01404.

apeer [WWW Document], 2021. Apeer by ZEISS. https://www.apeer.com/home (accessed 11.1.21).

Atila, Ü., Uçar, M., Akyol, K., Uçar, E., 2021. Plant leaf disease classification using EfficientNet deep learning model. Ecol. Inform. 61, 101182. https://doi.org/10.1016/j.ecoinf.2020.101182.

Azizi, A., Abbaspour-Gilandeh, Y., Vannier, E., Dusséaux, R., Mseri-Gundoshmian, T., Moghaddam, H.A., 2020. Semantic segmentation: a modern approach for identifying

soil clods in precision farming. Biosyst. Eng. 196, 172–182. https://doi.org/10.1016/j.biosystemseng.2020.05.022.
Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39, 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615.
Bakhsh, K., Ahmad, N., Kamran, M.A., Hassan, S., Abbas, Q., Saeed, R., Hashmi, M.S., 2016. Occupational hazards and health cost of women cotton pickers in Pakistani Punjab. BMC Public Health 16, 961. https://doi.org/10.1186/s12889-016-3635-3.
Bakhsh, K., Ahmad, N., Tabasum, S., Hassan, S., Hassan, I., 2017. Health hazards and adoption of personal protective equipment during cotton harvesting in Pakistan. Sci. Total Environ. 598, 1058–1064. https://doi.org/10.1016/j.scitotenv.2017.04.043.
Bao, W., Yang, Xinghua, Liang, D., Hu, G., Yang, Xianjun, 2021. Lightweight convolutional neural network model for field wheat ear disease identification. Comput. Electron. Agric. 189, 106367. https://doi.org/10.1016/j.compag.2021.106367.
Bengio, Y., Simard, P., Frasconi, P., 1994. Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. 5, 157–166. https://doi.org/10.1109/72.279181.
Braunack, M.V., Johnston, D.B., 2014. Changes in soil cone resistance due to cotton picker traffic during harvest on Australian cotton soils. Soil Tillage Res. 140, 29–39. https://doi.org/10.1016/j.still.2014.02.007.
Chen, C., Li, B., Liu, J., Bao, T., Ren, N., 2020. Monocular positioning of sweet peppers: an instance segmentation approach for harvest robots. Biosyst. Eng. 196, 15–28. https://doi.org/10.1016/j.biosystemseng.2020.05.005.
Chen, Z., Ting, D., Newbury, R., Chen, C., 2021. Semantic segmentation for partially occluded apple trees based on deep learning. Comput. Electron. Agric. 181, 105952. https://doi.org/10.1016/j.compag.2020.105952.
Chollet, F., 2015. Keras.
Colombi, T., Torres, L.C., Walter, A., Keller, T., 2018. Feedbacks between soil penetration resistance, root architecture and water uptake limit water accessibility and crop growth – a vicious circle. Sci. Total Environ. 626, 1026–1035. https://doi.org/10.1016/j.scitotenv.2018.01.129.
De-An, Z., Jidong, L., Wei, J., Ying, Z., Yu, C., 2011. Design and control of an apple harvesting robot. Biosyst. Eng. 110, 112–122. https://doi.org/10.1016/j.biosystemseng.2011.07.005.
Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: a large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. Presented at the 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 https://doi.org/10.1109/CVPR.2009.5206848.
Duong, L.T., Nguyen, P.T., Di Sipio, C., Di Ruscio, D., 2020. Automated fruit recognition using EfficientNet and MixNet. Comput. Electron. Agric. 171, 105326. https://doi.org/10.1016/j.compag.2020.105326.
Dyrmann, M., Karstoft, H., Midtiby, H.S., 2016. Plant species classification using deep convolutional neural network. Biosyst. Eng. 151, 72–80. https://doi.org/10.1016/j.biosystemseng.2016.08.024.
Fujii, H., Tanaka, H., Ikeuchi, M., Hotta, K., 2021. X-net with different loss functions for cell image segmentation. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Presented at the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 3788–3795 https://doi.org/10.1109/CVPRW53098.2021.00420.
Gao, F., Wu, T., Chu, X., Yoon, H., Xu, Y., Patel, B., 2020. Deep residual inception encoder–decoder network for medical imaging synthesis. IEEE J. Biomed. Health Inform. 24, 39–49. https://doi.org/10.1109/JBHI.2019.2912659.
Ghiasi, G., Lin, T.-Y., Le, Q.V., 2018. DropBlock: A Regularization Method for Convolutional Networks. arXiv:1810.12890 [cs].
Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier neural networks. Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. Presented at the Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, pp. 315–323.
Gonzalez-Huitron, V., León-Borges, J.A., Rodriguez-Mata, A.E., Amabilis-Sosa, L.E., Ramírez-Pereda, B., Rodriguez, H., 2021. Disease detection in tomato leaves via CNN with lightweight architectures implemented in Raspberry Pi 4. Comput. Electron. Agric. 181, 105951. https://doi.org/10.1016/j.compag.2020.105951.
Google Colaboratory, 2021. Google Colaboratory [WWW Document]. https://colab.research.google.com/notebooks/basic_features_overview.ipynb (accessed 11.14.21).
He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs].
Hecht, H., Sarhan, M.H., Popovici, V., 2020. Disentangled autoencoder for cross-stain feature extraction in pathology image analysis. Appl. Sci. 10, 6427. https://doi.org/10.3390/app10186427.
Hughs, S.E., Valco, T.D., Williford, J.R., 2008. 100 years of cotton production, harvesting, and ginning systems engineering: 1907-2007. Trans. ASABE 51, 1187–1198. https://doi.org/10.13031/2013.25234.
Jarrett, K., Kavukcuoglu, K., Ranzato, M., LeCun, Y., 2009. What is the best multi-stage architecture for object recognition? 2009 IEEE 12th International Conference on Computer Vision. Presented at the 2009 IEEE 12th International Conference on Computer Vision, pp. 2146–2153 https://doi.org/10.1109/ICCV.2009.5459469
Jiang, B., He, J., Yang, S., Fu, H., Li, T., Song, H., He, D., 2019. Fusion of machine vision technology and AlexNet-CNNs deep learning network for the detection of postharvest apple pesticide residues. Artif. Intell. Agric. 1, 1–8. https://doi.org/10.1016/j.aiia.2019.02.001.
Kandel, I., Castelli, M., 2020. The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. ICT Express 6, 312–315. https://doi.org/10.1016/j.icte.2020.04.010.
Kang, J., Liu, L., Zhang, F., Shen, C., Wang, N., Shao, L., 2021. Semantic segmentation model of cotton roots in-situ image based on attention mechanism. Comput. Electron. Agric. 189, 106370. https://doi.org/10.1016/j.compag.2021.106370.
Kestur, R., Meduri, A., Narasipura, O., 2019. MangoNet: a deep semantic segmentation architecture for a method to detect and count mangoes in an open orchard. Eng. Appl. Artif. Intell. 77, 59–69. https://doi.org/10.1016/j.engappai.2018.09.011.
Khanramaki, M., Askari Asli-Ardeh, E., Kozegar, E., 2021. Citrus pests classification using an ensemble of deep learning models. Comput. Electron. Agric. 186, 106192. https://doi.org/10.1016/j.compag.2021.106192.
Kingma, D.P., Ba, J., 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs].
Kolhar, S., Jagtap, J., 2021. Convolutional neural network based encoder-decoder architectures for semantic segmentation of plants. Ecol. Inform. 64, 101373. https://doi.org/10.1016/j.ecoinf.2021.101373.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25, 1097–1105.
Kumar, S.K., 2017. On Weight Initialization in Deep Neural Networks. arXiv:1704.08863 [cs].
Li, Y., Cao, Z., Lu, H., Xiao, Y., Zhu, Y., Cremers, A.B., 2016. In-field cotton detection via region-based semantic image segmentation. Comput. Electron. Agric. 127, 475–486. https://doi.org/10.1016/j.compag.2016.07.006.
Li, Y., Cao, Z., Xiao, Y., Cremers, A.B., 2017. DeepCotton: in-field cotton segmentation using deep fully convolutional network. J. Electron. Imaging 26, 1. https://doi.org/10.1117/1.JEI.26.5.053028.
Lin, M., Chen, Q., Yan, S., 2014. Network in Network. arXiv:1312.4400 [cs].
Liu, J.-S., Lia, H.-C., Jia, Z.-H., 2011. Image segmentation of cotton based on YCbCr color space and fisher discrimination analysis. Acta Agron. Sin. 37, 1274–1279.
Long, J., Shelhamer, E., Darrell, T., 2015. Fully Convolutional Networks for Semantic Segmentation. arXiv:1411.4038 [cs].
Majeed, Y., Zhang, J., Zhang, X., Fu, L., Karkee, M., Zhang, Q., Whiting, M.D., 2018. Apple tree trunk and branch segmentation for automatic trellis training using convolutional neural network based semantic segmentation. IFAC-PapersOnLine, 6th IFAC Conference on Bio-Robotics BIOROBOTICS 2018. 51, pp. 75–80. https://doi.org/10.1016/j.ifacol.2018.08.064.
Mehta, C., Chandel, N., Jena, P., Jha, A., 2019. Indian agriculture counting on farm mechanization. Agric. Mech. Asia Africa Latin Am. 50, 84–89.
Memon, Q.U.A., Wagan, S.A., Chunyu, D., Shuangxi, X., Jingdong, L., Damalas, C.A., 2019. Health problems from pesticide exposure and personal protective measures among women cotton workers in southern Pakistan. Sci. Total Environ. 685, 659–666. https://doi.org/10.1016/j.scitotenv.2019.05.173.
Milioto, A., Lottes, P., Stachniss, C., 2018. Real-Time Semantic Segmentation of Crop and Weed for Precision Agriculture Robots Leveraging Background Knowledge in CNNs. https://doi.org/10.48550/arXiv.1709.06764.
Molchanov, P., Tyree, S., Karras, T., Aila, T., Kautz, J., 2017. Pruning Convolutional Neural Networks for Resource Efficient Inference. arXiv:1611.06440 [cs, stat].
Nair, V., Hinton, G.E., 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. Undefined.
Ngugi, L.C., Abdelwahab, M., Abo-Zahhad, M., 2020. Tomato leaf segmentation algorithms for mobile phone applications using deep learning. Comput. Electron. Agric. 178, 105788. https://doi.org/10.1016/j.compag.2020.105788.
Ou, X., Yan, P., Zhang, Y., Tu, B., Zhang, G., Wu, J., Li, W., 2019. Moving object detection method via ResNet-18 with encoder–decoder structure in complex scenes. IEEE Access 7, 108152–108160. https://doi.org/10.1109/ACCESS.2019.2931922.
Panda, M.K., Sharma, A., Bajpai, V., Subudhi, B.N., Thangaraj, V., Jakhetiya, V., 2022. Encoder and decoder network with ResNet-50 and global average feature pooling for local change detection. Comput. Vis. Image Underst. 222, 103501. https://doi.org/10.1016/j.cviu.2022.103501.
Peng, D., Zhang, Y., Guan, H., 2019. End-to-end change detection for high resolution satellite images using improved UNet++. Remote Sens. 11, 1382. https://doi.org/10.3390/rs11111382.
Rahman, C.R., Arko, P.S., Ali, M.E., Iqbal Khan, M.A., Apon, S.H., Nowrin, F., Wasif, A., 2020. Identification and recognition of rice diseases and pests using convolutional neural networks. Biosyst. Eng. 194, 112–120. https://doi.org/10.1016/j.biosystemseng.2020.03.020.
Raikwar, S., Fehrmann, J., Herlitzius, T., 2022. Navigation and control development for a four-wheel-steered mobile orchard robot using model-based design. Comput. Electron. Agric. 202, 107410. https://doi.org/10.1016/j.compag.2022.107410.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597 [cs].
Roshanianfard, A., Noguchi, N., Okamoto, H., Ishii, K., 2020. A review of autonomous agricultural vehicles (the experience of Hokkaido University). J. Terrramech. 91, 155–183. https://doi.org/10.1016/j.jterra.2020.06.006.
Sakai, S., Iida, M., Osuka, K., Umeda, M., 2008. Design and control of a heavy material handling manipulator for agricultural robots. Auton. Robot. 25, 189–204. https://doi.org/10.1007/s10514-008-9090-y.
Seidu, S.M., 2018. Growth and Instability in Cotton Cultivation in Northern India. EA 63. https://doi.org/10.30954/0424-2513.2.2018.20.
Shah, J., Gao, F., Li, B., Ghisays, V., Luo, J., Chen, Y., Lee, W., Zhou, Y., Benzinger, T.L.S., Reiman, E.M., Chen, K., Su, Y., Wu, T., 2022. Deep residual inception encoder-decoder network for amyloid PET harmonization. Alzheimers Dement. https://doi.org/10.1002/alz.12564 alz.12564.
Shukla, S., Arude, V., Deshmukh, S., Patil, P., Mageshwaran, V., Sundaramoorthy, C., 2017. Mechanical harvesting of cotton: a global research scenario and Indian case studies. Cotton Res. J. 8, 46–57.
Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs].
Singh, N., Tewari, V.K., Biswas, P.K., Pareek, C.M., Dhruw, L.K., 2021. Image processing algorithms for in-field cotton boll detection in natural lighting conditions. Artif. Intell. Agric. 5, 142–156. https://doi.org/10.1016/j.aiia.2021.07.002.

Singh, N., Tewari, V.K., Biswas, P.K., Dhruw, L.K., Pareek, C.M., Singh, H.D., 2022. Semantic segmentation of in-field cotton bolls from the sky using deep convolutional neural networks. Smart Agric. Technol. 2, 100045. https://doi.org/10.1016/j.atech.2022.100045.

Snipes, C.E., Baskin, C.C., 1994. Influence of early defoliation on cotton yield, seed quality, and fiber properties. Field Crop Res. 37, 137–143. https://doi.org/10.1016/0378-4290(94)90042-6.

Sugawara, Y., Shiota, S., Kiya, H., 2019. Checkerboard artifacts free convolutional neural networks. APSIPA Trans. Signal Inf. Process. 8, e9. https://doi.org/10.1017/ATSIP.2019.2.

Sun, S., Li, C., Paterson, A.H., Chee, P.W., Robertson, J.S., 2019. Image processing algorithms for infield single cotton boll counting and yield prediction. Comput. Electron. Agric. 166, 104976. https://doi.org/10.1016/j.compag.2019.104976.

Sun, K., Wang, X., Liu, S., Liu, C., 2021. Apple, peach, and pear flower detection using semantic segmentation network and shape constraint level set. Comput. Electron. Agric. 185, 106150. https://doi.org/10.1016/j.compag.2021.106150.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2014. Going Deeper with Convolutions. arXiv:1409.4842 [cs].

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2015. Rethinking the Inception Architecture for Computer Vision. arXiv:1512.00567 [cs].

Tan, M., Le, Q.V., 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. https://doi.org/10.48550/ARXIV.1905.11946.

Tang, H., Wang, B., Chen, X., 2020. Deep learning techniques for automatic butterfly segmentation in ecological images. Comput. Electron. Agric. 178, 105739. https://doi.org/10.1016/j.compag.2020.105739.

Tassis, L.M., Tozzi de Souza, J.E., Krohling, R.A., 2021. A deep learning approach combining instance and semantic segmentation to identify diseases and pests of coffee leaves from in-field images. Comput. Electron. Agric. 186, 106191. https://doi.org/10.1016/j.compag.2021.106191.

Tedesco-Oliveira, D., Pereira da Silva, R., Maldonado, W., Zerbato, C., 2020. Convolutional neural networks in predicting cotton yield from images of commercial fields. Comput. Electron. Agric. 171, 105307. https://doi.org/10.1016/j.compag.2020.105307.

TensorFlow Developers, 2021. TensorFlow. https://doi.org/10.5281/ZENODO.4724125.

Tian, J., Zhang, X., Zhang, W., Li, J., Yang, Y., Dong, H., Jiu, X., Yu, Y., Zhao, Z., Xu, S., Zuo, W., 2018. Fiber damage of machine-harvested cotton before ginning and after lint cleaning. J. Integr. Agric. 17, 1120–1127. https://doi.org/10.1016/S2095-3119(17)61730-1.

Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C., 2015. Efficient Object Localization using Convolutional Networks. arXiv:1411.4280 [cs].

Waheed, A., Goyal, M., Gupta, D., Khanna, A., Hassanien, A.E., Pandey, H.M., 2020. An optimized dense convolutional neural network model for disease recognition and classification in corn leaf. Comput. Electron. Agric. 175, 105456. https://doi.org/10.1016/j.compag.2020.105456.

Wang, Y., Zhu, X., Ji, C., 2008. Machine vision based cotton recognition for cotton harvesting robot. In: Li, D. (Ed.), Computer and Computing Technologies in Agriculture, Volume II, The International Federation for Information Processing. Springer US, Boston, MA, pp. 1421–1425 https://doi.org/10.1007/978-0-387-77253-0_92.

Wang, S., Li, Y., Yuan, J., Song, L., Liu, Xinghua, Liu, Xuemei, 2021. Recognition of cotton growth period for precise spraying based on convolution neural network. Inf. Process. Agric. 8, 219–231. https://doi.org/10.1016/j.inpa.2020.05.001.

Williford, J., Brashears, A., Barker, G., 1994. Harvesting. Cotton Ginners Handbook. USDA Agricultural Research Service, Washington, D.C, pp. 11–16.

Wilson, D.R., Martinez, T.R., 2003. The general inefficiency of batch training for gradient descent learning. Neural Netw. 16, 1429–1451. https://doi.org/10.1016/S0893-6080(03)00138-2.

Xie, W., Wei, S., Zheng, Z., Yang, D., 2021. A CNN-based lightweight ensemble model for detecting defective carrots. Biosyst. Eng. 208, 287–299. https://doi.org/10.1016/j.biosystemseng.2021.06.008.

Xing, C., Arpit, D., Tsirigotis, C., Bengio, Y., 2018. A Walk with SGD. arXiv:1802.08770 [cs, stat].

Xu, L., Li, Y., Xu, J., Guo, L., 2020. Two-level attention and score consistency network for plant segmentation. Comput. Electron. Agric. 170, 105281. https://doi.org/10.1016/j.compag.2020.105281.

Yan, T., Xu, W., Lin, J., Duan, L., Gao, P., Zhang, C., Lv, X., 2021. Combining multi-dimensional convolutional neural network (CNN) with visualization method for detection of *Aphis gossypii* glover infection in cotton leaves using hyperspectral imaging. Front. Plant Sci. 12, 604510. https://doi.org/10.3389/fpls.2021.604510.

Yeom, J., Jung, J., Chang, A., Maeda, M., Landivar, J., 2018. Automated open cotton boll detection for yield estimation using unmanned aircraft vehicle (UAV) data. Remote Sens. 10, 1895. https://doi.org/10.3390/rs10121895.

Yin, X., Wu, D., Shang, Y., Jiang, B., Song, H., 2020. Using an EfficientNet-LSTM for the recognition of single cow's motion behaviours in a complicated environment. Comput. Electron. Agric. 177, 105707. https://doi.org/10.1016/j.compag.2020.105707.

You, J., Liu, W., Lee, J., 2020. A DNN-based semantic segmentation for detecting weed and crop. Comput. Electron. Agric. 178, 105750. https://doi.org/10.1016/j.compag.2020.105750.

Zabawa, L., Kicherer, A., Klingbeil, L., Töpfer, R., Kuhlmann, H., Roscher, R., 2020. Counting of grapevine berries in images via semantic segmentation using convolutional neural networks. ISPRS J. Photogramm. Remote Sens. 164, 73–83. https://doi.org/10.1016/j.isprsjprs.2020.04.002.

Zaidner, G., Shapiro, A., 2016. A novel data fusion algorithm for low-cost localisation and navigation of autonomous vineyard sprayer robots. Biosyst. Eng. 146, 133–148. https://doi.org/10.1016/j.biosystemseng.2016.05.002.

Zhang, Shanwen, Zhang, Subing, Zhang, C., Wang, X., Shi, Y., 2019. Cucumber leaf disease identification with global pooling dilated convolutional neural network. Comput. Electron. Agric. 162, 422–430. https://doi.org/10.1016/j.compag.2019.03.012.

Zhang, P., Yang, L., Li, D., 2020. EfficientNet-B4-ranger: a novel method for greenhouse cucumber disease recognition under natural complex environment. Comput. Electron. Agric. 176, 105652. https://doi.org/10.1016/j.compag.2020.105652.

Zhenlong, H., Qiang, Z., Jun, W., 2018. The prediction model of cotton yarn intensity based on the CNN-BP neural network. Wirel. Pers. Commun. 102, 1905–1916. https://doi.org/10.1007/s11277-018-5245-0.

Zou, K., Chen, X., Wang, Y., Zhang, C., Zhang, F., 2021. A modified U-net with a specific data argumentation method for semantic segmentation of weed images in the field. Comput. Electron. Agric. 187, 106242. https://doi.org/10.1016/j.compag.2021.106242.