*ICIC Express Letters*
*Part B: Applications*
Volume **13**, Number **3**, March **2022**

ICIC International ©2022 ISSN 2185-2766
pp. **279–286**

# DAIRY CATTLE DETECTION IN LOOSE HOUSING CALVING PEN BY USING SEMANTIC SEGMENTATION NETWORKS

Swe Zar Maw[1,*], Thi Thi Zin[2] and Pyke Tin[2]

[1]Interdisciplinary Graduate School of Agriculture and Engineering
[2]Graduate School of Engineering
University of Miyazaki
1-1 Gakuen Kibanadai Nishi, Miyazaki 889-2192, Japan
[*]Corresponding author: swezarmaw@gmail.com
thithi@cc.miyazaki-u.ac.jp; pyketin11@gmail.com

Received July 2021; accepted September 2021

Abstract. *When it comes to controlling a cattle farm, being able to accurately forecast when calving will happen can be quite beneficial because it allows employees to assess whether or not assistance is required. If such help is not provided when it is required, the calving process may be prolonged, severely impacting both the mother cow and the calf's health. Multiple diseases may result from such a delay. During the production cycle, one of the most crucial events for cows is calving. An accurate video-monitoring technique for cows can spot abnormalities or health issues early, allowing for prompt and effective human interference. To make this surveillance automated, a crucial task is to detect the dairy cattle. For this purpose, in this research, we have proposed an effective semantic segmentation network for segmenting the cow from the 360-degree surveillance camera. The proposed network is a modified version of the U-Net architecture. An additional module is added in the U-Net architecture which is named as Convolutional Long Short-Term Memory (ConvLSTM) block. The ConvLSTM block allows for effective feature sharing between the less dense layers and denser layers. Experiments with our suggested method were carried out at a big dairy farm in Japan's Oita Prefecture. The suggested method's experimental findings demonstrate that it holds promise in real-world applications.*
**Keywords:** Neural networks, Calving, Segmentation, U-Net, Convolution, LSTM

1. **Introduction.** The livestock sector across the world, owing to the high demand for livestock, must make use of the bounded resources, which include farms and building. They also have to tackle the challenge of the shortage of labour [1]. To attain this goal, precision livestock farming comes as a handy option, as it is particularly efficient and economical. Gathering the information about each cattle's welfare, its well-being and behaviour have a significant effect on livestock regulation and choices [2,3]. The use of a camera is the popular and cost friendly solution, in today's world, to monitor livestock without being in direct contact. Image segmentation is required beforehand in procedures based on vision [4].

The segmented images are used to extract every visible characteristic which includes but are not limited to width, height, span, and posture. It is self-explanatory that image segmentation accuracy and competence has a say in following image analysis and optic-based, concurrent monitoring of animals and their health evaluation [5,6]. In contrast, the difference between frames and techniques builds on optical flow which have a hard time getting the desired accuracy when outdoor conditions are in play [7,8]. The dynamic conditions, unsteady background (employees walking on-farm, animals moving, etc.), low or variable light conditions pose a challenge that has a degrading effect on image segmentation [9].

DOI: 10.24507/icicelb.13.03.279

In recent years, Convolutional Neural Network (CNN) has established a great place in the field of bioinformatics [10], medical imaging [11] and much more [12]. CNNs study the features and then convey the spatial details and perceptual data that can be exploited in picture segmentation, owing to their leaning on huge, labelled datasets. Every part of this modern research demonstrates the potential for cow segmentation based on CNN in complex feedlot scenarios. Various deep learning-based networks have been used to segregate objects from their respective backgrounds [13].

One of the most painful, exhausting, and hard events a cow must go through in its production cycle is the calving stage [14]. Such problems arise with long terms spontaneous calving or long term and severe aided extraction. To help mitigate these losses and have better control over farm management, professionals on the farm need accurate and reliable tools for anticipating when the calving will take places and when they need to intervene [15]. Due to the changes going on and the difficulties faced by the cows during the calving stage, they become naturally vulnerable. Monitoring continuously at this time will aid in predicting the next time calving could occur and the calving behaviour [16]. Image processing tools and other cow monitoring techniques can provide visual data for analyzing and coping with other challenges [17].

Cow region segmentation plays an important role when it comes to cow activity detection. When it comes to detection of cow's activity, it is important to identify whole cow and not to skip any of its body part. By missing some regions of the body part, the whole activity of the cow can be wrongly recognized. To tackle the identified cow, we can employ a semantic segmentation algorithm to track the cows so that they can be provided with any necessary help being required by them.

In this paper, we mainly proposed a modified version of the U-Net architecture for segmentation of cow regions by adding an additional module, ConvLSTM block in the U-Net architecture. The U-Net architecture and ConvLSTM are mostly used in medical imaging and in this research, we apply these architectures for using cow region segmentation. The advantages of this proposed method are that it needs only very few annotation images, can perform on touching and overlapping objects for the same class and can localize and distinguish border by doing classification on every pixel compared with other state-of-the-art studies. The performance of the proposed architecture is better than that of original U-Net architecture. The cow region segmentation results of the proposed method increase 0.043 and 0.034 in dice score and IoU respectively than the original U-Net architecture.

The rest of the paper is organized as follows. In Section 2, we explained about the detailed information of the dataset that we used. In Section 3, we described the proposed architecture of this paper and in Section 4, we discussed about the experimental results. Finally, we concluded and illustrated the future work of our research in Section 5.

2. **Dataset.** When the calving was due in 72 hours, videos were recorded of each heifer and its respective actions. This paper details the findings of research conducted on a farm in the Oita Prefecture, Japan. In the center of the barn, the GV-FER5700 GeoVision 360-degree camera was installed. This camera would provide the best possible view of the pregnant heifers. The resolution of the camera was $2560 \times 2048$ pixels and it recorded with 30 frames per second. It was mounted 3 meters above the ground in each loose housing calving pen. The cows had a $7 \times 7$ square feet area, on a sawdust-covered surface, to roam around without any hindrance. The pregnant cows have different activities, so capturing the cow in different poses was an important task which was made sure. As for segmentation, it is necessary to have all possible positions while training the model. Figure 1 shows the sample of the dataset.

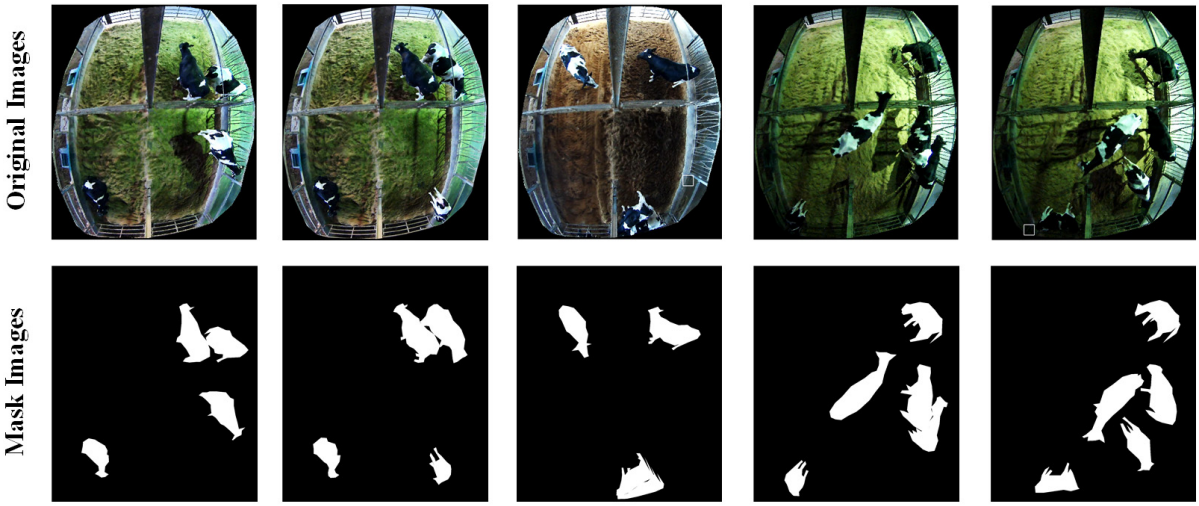


FIGURE 1. Sample images from dataset

3. **Proposed Architecture.** Figure 2 shows the visual representation of the proposed architecture. The proposed architecture is inspired from U-Net [18] and ConvLSTM [19]. The ConvLSTM is introduced in the original U-Net architecture by having some further modifications. In the following subsection, we go through several aspects of the network.

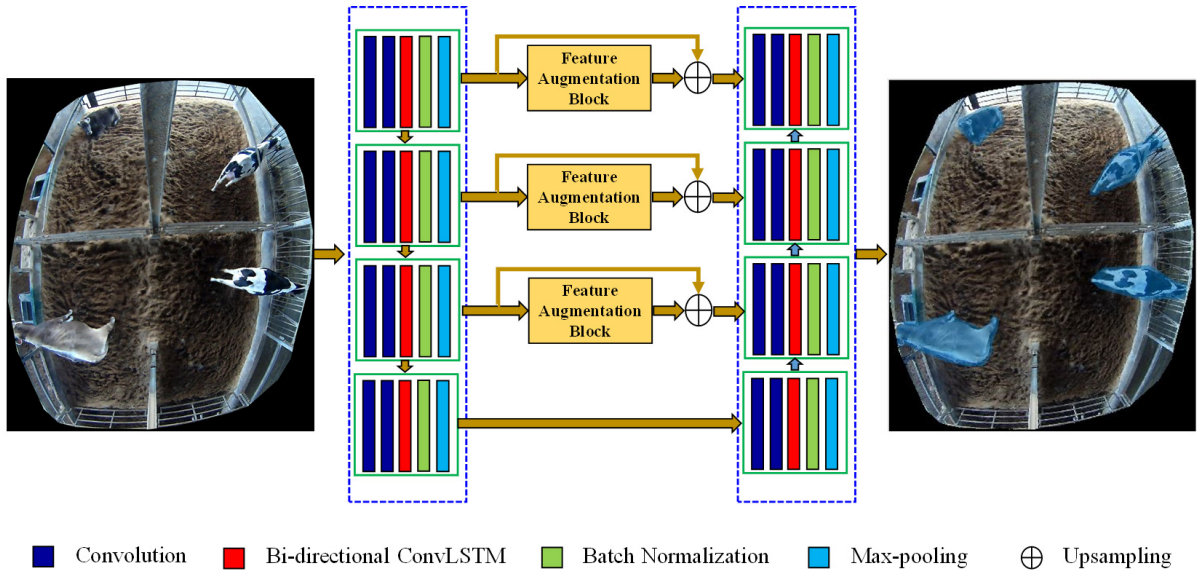


FIGURE 2. (color online) The proposed neural network architecture

3.1. **Encoder.** The encoder is the contracting process which consists of four phases. On every phase two convolutional layers are proceeded by a max-pooling layer and a ReLU activation layer. At each stage, the number of feature maps is doubled. Every convolutional layer uses filters of size $3 * 3$ and max-pooling uses a window of size $2 * 2$. The contracting route extracts features on every phase and increases the feature vector size, layer by layer. Finally, the encoding path's last layer provides a high-dimensional feature map with a lot of semantic information. In the last phase of the encoding route, the original U-Net has a sequence of convolutional layers. The technique learns multiple types of features by using a succession of convolutional layers in a network. Nonetheless, in subsequent convolutions, the network may learn duplicate characteristics. Therefore, in

observance to this problem and after having multiple experiments, the last block of the U-Net encoder path is removed. The concept of collective knowledge, in which feature vectors are utilized throughout the network, assists the network in improving its performance. It implies that feature vectors from all previous convolutional layer are fused with the feature vector from the last layer, and then sent to be used by the next convolutional layer without any modification.

3.2. **Decoder.** Each stage in the decode route begins with an up-sampling function applied to the preceding layer's output. The matching feature maps in the contracting path are clipped and transferred to the decoding path in the typical U-Net. The results of the up-sampling algorithms are then concentrated with these feature maps. However, in the proposed architecture, a ConvLSTM module is introduced in the U-Net architecture. The feature maps from contracting path are used for further feature extraction using ConvLSTM and then those extracted feature maps are concatenated with the up-sampled feature maps. However, to preserve the original information, a skip connection is used which bypasses the ConvLSTM block. On every decoding stage, a convolutional layer uses filters of size $2 * 2$ making half the number of feature channels and doubling the size of each feature vector when compared with encoding stages. It can be interpreted that the expanding route grows the dimension of the feature vector layer by layer until the final layer reaches the original dimensionality of the input image.

In decoder, after up sampling the feature map it undergoes the process of batch normalization. The distribution of activations fluctuates in the intermediate layers throughout the training stage, which is a concern. Because every layer in every training phase must learn to adapt to a new distribution, this issue slows down the training process. Batch normalization, which standardizes the inputs to a layer in the network by removing the batch mean and dividing by the batch standard deviation, is used to improve the stability of a neural network. The speed of a neural network's training process is impacted by batch normalization. Furthermore, the small regularization impact improves the model's performance in some situations.

3.3. **ConvLSTM.** Bi-directional ConvLSTM block is the major addition made to the original U-Net architecture. This block allows to get more significant feature maps from the different stages of the encoder side of the architecture. Recurrent Neural Networks (RNNs) are a type of network that was created particularly for processing sequence of samples. They output each layer to the next layer, as well as a hidden state for the current layer to utilize while processing the next input. RNNs excel at extracting correlations between samples.

Long Short-Term Memory (LSTM) is an evolved version of RNN that addresses the issue of gradient disappearance and explosion in RNN. In this article, LSTM is utilized to investigate the relationships between the characteristics and to forecast them one by one. LSTM can relate to the hidden state incorporating previous information while predicting upcoming characteristics. Although LSTM excels in time-series modeling, it is not without flaws, but the primary drawback of conventional LSTM networks is that they do not account for spatial correlation since complete connections are used in input to state, and state to state transitions. The generic LSTM models sequence information across the whole connection layer and flattens the input into a single dimensional vector, where image spatial information is lost.

ConvLSTM was suggested to overcome the discussed problem by including convolutional layer into input to state, and state to state transitions. ConvLSTM helps to retain the spatial information of the input image and in case of cow detection, this can be a great help. A convolution operation is utilized in ConvLSTM to convert input or state to the state. It can better mine the connection among pedestrian characteristics since it can capture the spatial information of attributes better than ordinary LSTM. ConvLSTM has

the following formulation:

$$i_t = sigmoid\left(W_{ix} * x_i + W_{ih} * h_{t-1} + W_{ic} \odot c_{t-1} + b_i\right) \tag{1}$$
$$f_t = sigmoid\left(W_{fx} * x_t + W_{fh} * h_{t-1} + W_{fc} \odot c_{t-1} + b_f\right) \tag{2}$$
$$o_t = sigmoid\left(W_{ox} * x_t + W_{oh} * h_{t-1} + W_{oc} \odot c_t + b_o\right) \tag{3}$$
$$g_t = \tanh\left(W_{gx} * x_t + W_{gh} * h_{t-1} + b_g\right) \tag{4}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{5}$$
$$h_t = o_t \odot \tanh\left(c_t\right) \tag{6}$$

The symbol '$*$' represents the convolution operator and '$\odot$' represents the element-wise multiplications. Further, *sigmoid*( ) symbolizes the logistic sigmoid component and tanh( ) symbolizes the hyperbolic tangent component, the subscript $t$ interpreted the $t$th step of ConvLSTM, '$i$' is the input gate, '$g$' is the input modulation gate, '$o$' is the output gate, '$f$' is the forget gate, '$h$' is the hidden state, '$c$' is the cell state, and '$x$' is the input data. The $x$, $c$, $h$, $f$, and $o$ are 3 dimensional tensors in which the 1st dimension contains time information, and the 2nd and 3rd dimensions provide spatial information in rows and columns. To save the spatial features of original image characteristics, the convolution process is utilized. The fundamental of ConvLSTM is similar to that of regular LSTM; the preceding layer's output is used as the input for the next layer. ConvLSTM may gain not only temporal but also spatial connections. This is analogous to how CNN's convolutional layer extracts spatial characteristics so that spatiotemporal features may be retrieved. Convolutional layer also has a spatial attention impact, since the region that corresponds to the targeted image characteristic has a higher activation response. The framework of ConvLSTM is shown in Figure 3.

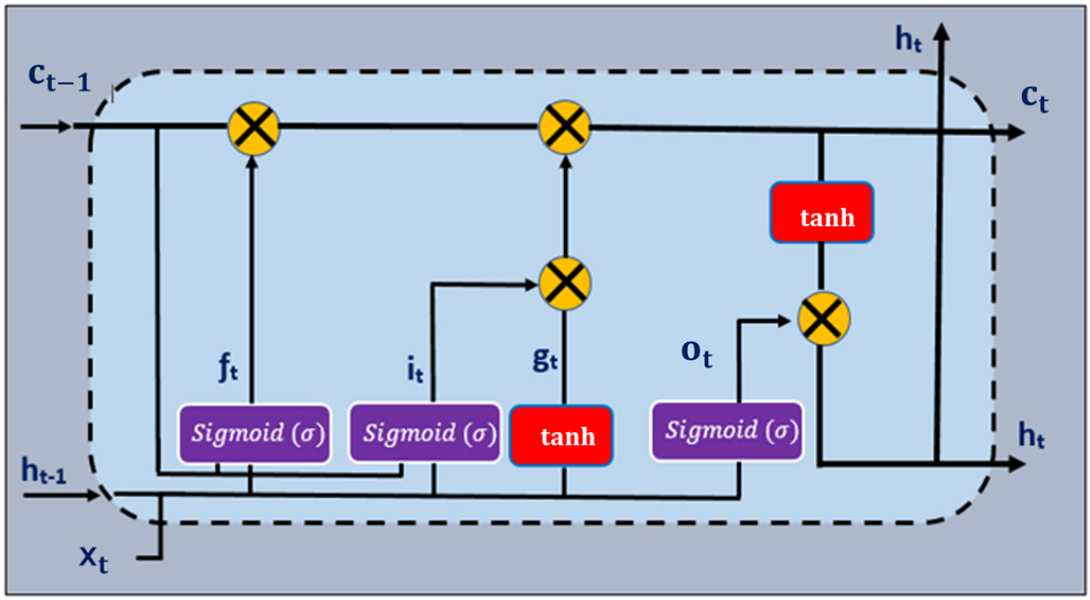


FIGURE 3. ConvLSTM framework

4. **Results and Discussion.** The total dataset was divided into three parts which are training, validation, and testing. The testing dataset was kept as unknown data to the architecture while the training and validation processes are working. For the purpose of performance measurement of the architecture, we have used dice score and Intersection over Union (IoU) as figure of merits. The formulation of both figure of merits is as follows:

$$\text{Dice}(x, y) = \frac{2|X \cap Y|}{|Y| + |X|} \tag{7}$$

$$\text{IoU}(X, Y) = \frac{X \cap Y}{X \cup Y} \tag{8}$$

where $X$ and $Y$ are mask image (ground truth) and predicted image, respectively. To compare the proposed model, we implemented original U-Net architecture and computed its results on the proposed as well as on the U-Net architecture. Table 1 shows the comparison between the two architectures. As shown in Table 1, the performance illustrated by the proposed architecture is better than that of original U-Net architecture. The architecture has shown an improvement of 0.043 and 0.034 in dice score and IoU, respectively. Figure 4 shows visual comparison where the first row of images is the input image while the second row of images is the ground truth. The third row and the fourth

TABLE 1. Comparison between U-Net and the proposed architectures on the testing dataset

| Architecture | Training dataset (# semantic labelled image) | Validation dataset (# semantic labelled image) | Testing dataset (# semantic labelled image) | Performance accuracy | |
|---|---|---|---|---|---|
| | | | | Dice score | IoU |
| U-Net | 3,252 | 407 | 407 | 0.833 | 0.831 |
| Proposed | 3,252 | 407 | 407 | 0.876 | 0.865 |

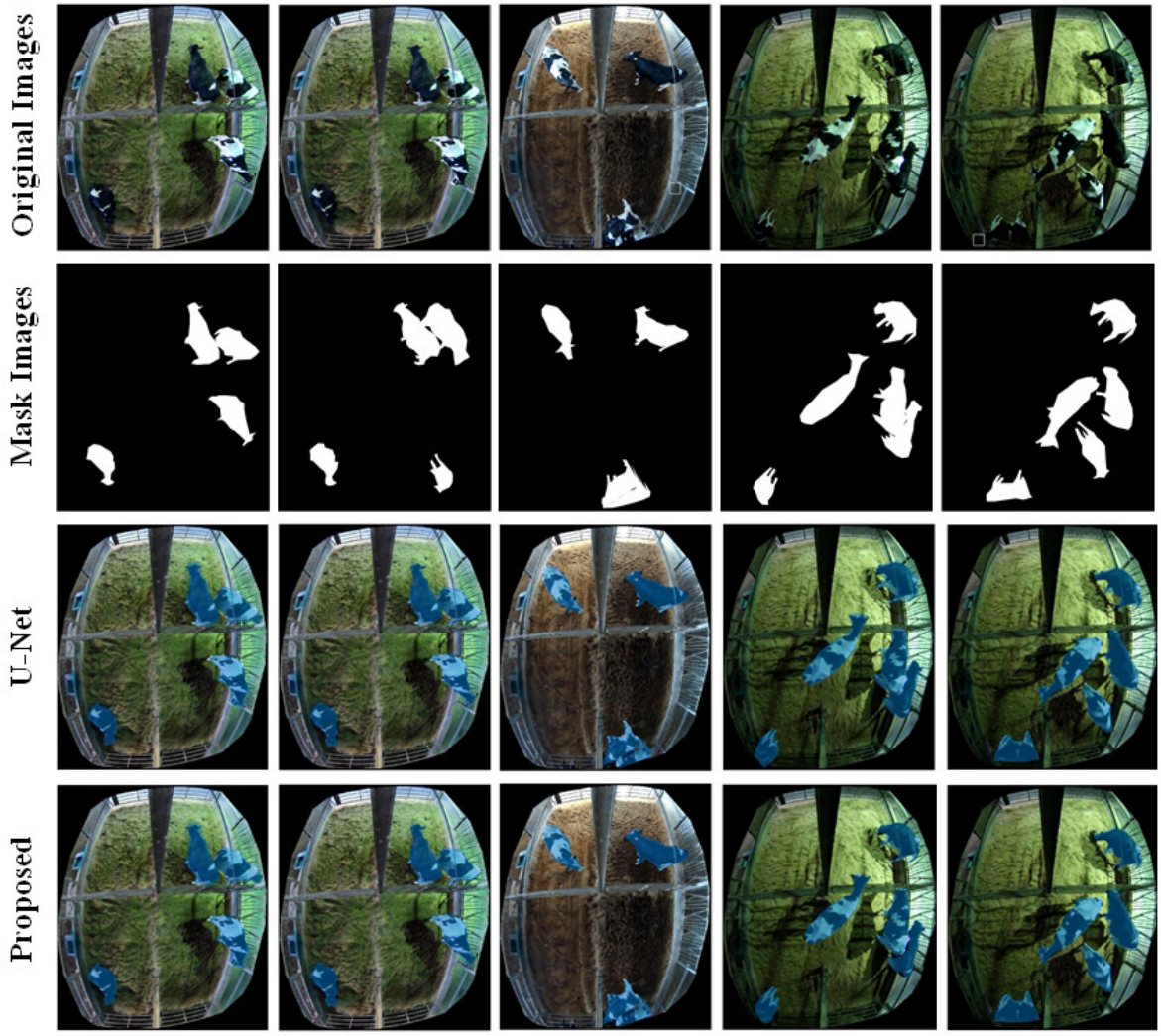


FIGURE 4. Visual comparison between U-Net and the proposed architectures

rows of the images present U-Net and the proposed architecture prediction, respectively. According to the experimental results, the U-Net architecture lacks few regions to get the segmentation of the whole part of the cow body. If the cow region is too small or one of the cows blocks another cow, the original U-Net architecture could not detect or segment that region. However, the proposed method can detect all cow regions and it does not miss any part of the cow body. And perhaps that is necessary for better activity recognition system.

5. **Conclusion and Future Work.** In this work, we have proposed a semantic segmentation framework for effective segmentation of cows. The effective segmentation system is necessary for further accessing the activity recognition of the cow. In the production cycle, one of the most crucial events for cows is calving and human interference is sometimes necessary in that. With the advancement in the field of computer vision and artificial intelligence, humans are equipped with the technology that can help to track the cow's activity and have decision on making upon that. The proposed architecture is the modified version of U-Net architecture. The proposed framework has introduced the ConvLSTM module in the real U-Net architecture. The proposed architecture has attained good results when compared with real U-Net architecture. The suggested method's experimental findings demonstrate that it holds promise in real-world applications.

This work is the first step towards the automation process. When the segmentation process was completed, the classification process will be performed. The classification will be carried out to recognize the activity of the cow. For this purpose, we intend the following future work to be done. Prepare the dataset for activity classification. Experts would be labeling the dataset of the images categorizing them in different activities. The current dataset is labelled for segmentation and now the segmented regions need to be labelled into different activities. With the production of new dataset, we will need to prepare a cascaded network where the first architecture would be the one proposed in this work while the second network will be of classification.

## REFERENCES

[1] G. Tara, Food sustainability: Problems, perspectives and solutions, *Proceedings of the Nutrition Society*, vol.72, no.1, pp.29-39, 2013.

[2] T. M. Banhazi, H. Lehr, J. L. Black, H. Crabtree, P. Schofield, M. Tscharke and D. Berckmans, Precision livestock farming: An international review of scientific and commercial aspects, *International Journal of Agricultural and Biological Engineering*, vol.5, no.3, pp.1-9, 2012.

[3] T. Emanuela, A. Finzi and M. Guarino, Environmental impact of livestock farming and precision livestock farming as a mitigation strategy, *Science of the Total Environment*, vol.650, pp.2751-2760, 2019.

[4] M. Tomás, P. P. P. Olea and J. V. L. Bao, Time to monitor livestock carcasses for biodiversity conservation and public health, *Journal of Applied Ecology*, vol.56, no.7, pp.1850-1855, 2019.

[5] P. Andrea, M. Guarino, L. Sartori and F. Marinello, A feasibility study on the use of a structured light depth-camera for three-dimensional body measurements of dairy cows in free-stall barns, *Sensors*, vol.18, no.2, DOI: 10.3390/s18020673, 2018.

[6] S. G. Matthews, A. L. Miller, T. PlÖtz and I. Kyriazakis, Automated tracking to measure behavioural changes in pigs for health and welfare monitoring, *Scientific Reports*, vol.7, no.1, pp.1-12, 2017.

[7] M. Garcia, L. F. Greco, M. G. Favoreto, R. S. Marsola, D. Wang, J. H. Shin, E. Block, W. W. Thatcher, J. E. P. Santos and C. R. Staples, Effect of supplementing essential fatty acids to pregnant nonlactating Holstein cows and their preweaned calves on calf performance, immune response, and health, *Journal of Dairy Science*, vol.97, no.8, pp.5045-5064, 2014.

[8] G. H. Mount, B. Rumburg, J. Havig, B. Lamb, H. Westberg, D. Yonge, K. Johnson and R. Kincaid, Measurement of atmospheric ammonia at a dairy using differential optical absorption spectroscopy in the mid-ultraviolet, *Atmospheric Environment*, vol.36, no.11, pp.1799-1810, 2002.

[9] T. A. Burnett, A. M. L. Madureira, F. S. Bruna, A. C. C. Fernandes and R. L. A. Cerri, Integrating an automated activity monitor into an artificial insemination program and the associated risk factors affecting reproductive performance of dairy cows, *Journal of Dairy Science*, vol.100, no.6, pp.5005-5018, 2017.

[10] R. U. Mobeen, K. J. Hong, H. Tayara and K. T. Chong, m6A-NeuralTool: Convolution neural tool for RNA N6-Methyladenosine site identification in different species, *IEEE Access*, vol.9, pp.17779-17786, 2021.

[11] R. U. Mobeen, S. B. Cho, K. J. Hong and K. T. Chong, BrainSeg-Net: Brain tumor MR image segmentation via enhanced encoder-decoder network, *Diagnostics*, vol.11, no.2, DOI: 10.3390/diagnostics11020169, 2021.

[12] A. S. Agoes, Z. Hu and N. Matsunaga, LICODS: A CNN based, lightweight RGB-D semantic segmentation for outdoor scenes, *International Journal of Innovative Computing, Information and Control*, vol.15, no.5, pp.1935-1946, 2019.

[13] A. A. Rafique, A. Jalal and K. Kim, Statistical multi-objects segmentation for indoor/outdoor scene detection and classification via depth images, *2020 17th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, pp.271-276, 2020.

[14] J. F. Mee, Managing the dairy cow at calving time, *Veterinary Clinics: Food Animal Practice*, vol.20, no.3, pp.521-546, 2004.

[15] C. E. Story, R. J. Rasby, R. T. Clark and C. T. Milton, Age of calf at weaning of spring-calving beef cows and the effect on cow and calf performance and production economics, *Journal of Animal Science*, vol.78, no.6, pp.1403-1413, 2000.

[16] M. B. Jensen, The early behaviour of cow and calf in an individual calving pen, *Applied Animal Behaviour Science*, vol.134, nos.3-4, pp.92-99, 2011.

[17] K. Sumi, T. T. Zin, I. Kobayashi and Y. Horii, A study on cow monitoring system for calving process, *2017 IEEE 6th Global Conference on Consumer Electronics (GCCE)*, pp.1-2, 2017.

[18] R. Olaf, P. Fischer and T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Sciencen*, N. Navab, J. Hornegger, W. Wells and A. Frangi (eds.), Cham, Springer, 2015.

[19] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong and W.-C. Woo, Convolutional LSTM network: A machine learning approach for precipitation nowcasting, *Advances in Neural Information Processing Systems*, pp.802-810, 2015.