A Deep Learning Approach for Cattle Classification to Enhance Livestock Monitoring

Farah Kareem[1], Syed Umaid Ahmed[1], Muhammad Farrukh Shahid[1], M. Hassan Tanveer[2]
[1]Department of AI and Data Science, FAST National University of Computer and Emerging Sciences (FAST-NUCES), Karachi, Pakistan.
[2]Department of Robotics and Mechatronics Engineering, Kennesaw State University, Marietta, GA, USA
*Correspondence: k238045@nu.edu.pk, mtanveer@Kennesaw.edu, k226020@nu.edu.pk , farahkareem395@gmail.com

Citation | Kareem. F, Ahmed. S. U, Shahid. M. F, Tanveer. M. H, "A Deep Learning Approach for Cattle Classification to Enhance Livestock Monitoring", IJIST, Special Issue pp 562-573, May 2026
Received | March 30, 2026 **Revised |** May 8, 2026 **Accepted |** May 13, 2026 **Published |** May 16, 2026.

The classification of cattle from frontal face images is an important step toward automated livestock management, disease tracking, and optimization of farm productivity. In this study, we compare the performance of three deep learning architectures, Vision Transformer (ViT), Swin Transformer, and ResNet18, for large-scale, multiclass cattle classification. The proposed system is trained and evaluated on a cow dataset from Karachi, Pakistan, which includes approximately 459 distinct classes, making it one of the largest publicly available cow image datasets. All the models were trained and tested in similar experimental conditions, which ensures a fair comparison. The highest classification accuracy was achieved by the Vision Transformer with 96.27%, compared to 95.86% and 94.48% in both ResNet18 and Swin Transformer, respectively. In addition, the ViT model attained a macro precision of 0.94, recall of 0.95, and F1-score of 0.94, while ResNet18 achieved 0.94, 0.95, and 0.94, and Swin Transformer achieved 0.92, 0.93, and 0.92, respectively. The training process converged within 100 epochs, with final training and validation losses of 0.0176 and 0.2420 for ViT, 0.0228 and 0.2172 for ResNet18, and 0.0527 and 0.3159 for Swin Transformer, indicating stable learning behavior across models. The obtained results indicate that transformer-based architectures effectively capture fine-grained facial features in cattle compared to traditional CNNs. In addition, the Top-5 accuracies of all models were more than 99%, which highlights the appropriateness of all models to large-scale, multiclass cattle identification. Hence, the proposed work illustrates that these models can improve classification accuracy, which will aid in accurate livestock tracking, traceability, and help to identify diseases in their early stages, which will eventually increase productivity and the development of sustainable agriculture. The current research also fits in with the United Nations Sustainable Development Goals, SDG 2 (Zero Hunger) and SDG 12 (Responsible Consumption and Production), as it contributes to effective livestock management and sustainable agricultural practices.

Keywords: Cattle Identification, Deep Learning, Convolutional Neural Network, Vision Transformer, Sustainable Agriculture.

Introduction:

Livestock management plays a crucial role in global agriculture by providing essential goods such as dairy, meat, and leather, and forms a significant part of the rural livelihoods [1]. In countries like Pakistan, where cattle farming constitutes a major component of the agricultural economy, accurate identification and classification of individual animals is required in order to manage livestock efficiently, monitor diseases and breeding programs, as well as monitor productivity. Traditional methods of identifying animals, including branding and ear tagging, routinely cause physical harm, increase stress levels in the animals, and can indirectly affect health and production [2].

Recent advances in deep learning have significantly improved image-based animal identification systems. Convolutional neural network (CNN) models, such as ResNet18 has demonstrated effectiveness in image classification objectives [3]. Furthermore, with the development of Vision Transformers (ViTs) [4] and their derivatives, such as the Swin Transformer [5], a more attractive option has emerged, which has shown superior capability in capturing long-range dependencies and fine-resolution image details, making it potentially useful in livestock biometrics and precision agriculture applications [6].

The Cow Frontal Face Dataset [7], which contains about 459 classes of cows in Karachi, Pakistan, is one of the largest datasets available in cattle-classification research. Individual cattle are very difficult to identify due to the high intra-class variation caused by poses, illumination, and viewpoint, along with visual resemblance between different individuals. Therefore, robust deep-learning models with strong feature-extraction capabilities are required.

Recent studies have explored deep learning and transformer-based approaches for livestock monitoring and precision agriculture applications [8] and [9]. However, limited work has been conducted on large-scale cattle classification using transformer-based architectures, particularly in the context of South Asian livestock systems. This research paper compares the performance of three modern architectures, ViT [4], Swin Transformer [5], and ResNet18 [3], with large-scale multi-class cattle classification. The deep learning models were trained and evaluated under identical experimental conditions, with classification accuracy used as the primary performance metric. The objective of the comparative analysis is to determine which model proves to be the most productive in terms of cattle classification, and the possible outcomes include livestock monitoring, traceability infrastructures, and AI-based precision agriculture.

Research Objectives:

The main objectives of this study are:

To compare transformer-based and CNN-based architectures on one of the largest available datasets for cattle classification.

Comparative performance of ViT, Swin Transformer, and ResNet18 with similar training conditions.

Analyze the suitability of deep learning methods to precision livestock farming and automated monitoring in the Pakistani environment.

Novelty and Contributions:

The novelty of this work includes:

A large-scale experimental evaluation on a 459-class cattle facial dataset, addressing a highly challenging fine-grained classification problem.

Experimental results demonstrate that Vision Transformer models outperform traditional CNN architectures in cattle identification tasks.

A comprehensive comparative analysis under similar training conditions implies a fair benchmarking of CNN and transformer-based models.

Insights into the applicability of transformer-based models for automated livestock tracking and precision agriculture in developing areas.

The work contributes to the United Nations Sustainable Development Goals, SDG 2 (Zero Hunger) and SDG 12 (Responsible Consumption and Production), as it enables the effective management of livestock and sustainable agricultural practices by facilitating the development of AI-based automation [10] and [11].

Literature Review:

The literature review aims to evaluate the existing literature about livestock management and the use and implementation of artificial intelligence (AI) to identify cattle and classify their breeds, thus contributing to the evolution and development of sustainable agriculture. Modern innovations in deep learning (DL), machine learning, and computer vision have attracted new possibilities in precision livestock farming, especially automating cattle-monitoring systems that are inherently productive and improve animal welfare without increasing environmental harm.

Several studies have evaluated and developed machine learning and deep learning techniques for cattle identification. In this perspective, ear tagging, branding, and radio-frequency identification (RFID), which are considered traditional ways of identification, are still common but cause physiological stress to the animals [12]. Traditionally, machine learning (ML) models, such as support-vector machines (SVM), k-nearest neighbors (KNN), and artificial neural networks (ANN), along with hand-engineered feature extractors such as local binary patterns (LBP), speeded-up robust features (SURF), and scale-invariant feature transform (SIFT), have been used to identify salient features such as coat patterns and muzzle prints [2] However, deep learning models, specifically convolutional neural networks (CNNs) and modern detection architectures, have proven to be more accurate and more robust in livestock detection and identification tasks [12]. Research has also investigated the performance of the machine-learning and deep-learning systems on popular image and video databases using both tag-based, deoxyribonucleic acid (DNA) based, and visual-based biometric systems [13]. Multi-feature decision-layer fusion has also been investigated, with better results regarding identification performance. The muzzle-pattern analysis, mutually used alongside facial recognition, has shown an increased accuracy, reliability, and robustness particularly in dynamic farm environments [14]. Kumar et al. propose a local feature-guided neural architecture to identify cows, where discrimination in local areas is emphasized to enhance strong performance against pose and illumination variations [15]. Out of 58 publicly available livestock-farming datasets under study, about 50 were of cattle. These results highlight the need to have more specialized datasets to optimize computer-vision systems in livestock management [16]. A study investigates accurate and non-invasive methods of identifying cattle with deep learning models and utilizes facial features to recognize individual livestock using image data [8]. Besides identification, computer vision technology has been utilized to support other livestock management activities. An artificial-intelligence-based system was proposed by Inam et al., which estimates cattle age based on dentition images, using YOLOv8 to detect and segment images and automatically predict the age group, which demonstrates the potential of deep learning in non-invasive livestock measurement [17].

Convolutional Neural Networks (CNNs) have paved the way for these image-based approaches by effectively learning discriminative visual features without manual feature engineering [18]. Akarsu et al. suggested employing deep features, which were obtained using ResNet-18, and Linear Discriminant Analysis (LDA) to improve the accuracy of animal classification [19]. According to their findings, the hybrid deep-classical approaches can significantly enhance the discriminative learning results in livestock recognition problems [19]. Similarly, individual identification accuracy of lightweight CNN models fine-tuned on dairy cow side-view images achieved 96.65% [20]. A refined breed detection model, RTDETR-Refa

(based on a ResNet18 backbone with RepConv and attention modules), succeeded in surpassing traditional ResNet18 and other comparisons, along with an accuracy increase of almost 1% but without decreasing detection speeds [21]. In related efforts, metric learning techniques combined with CNNs have been applied to Holstein-Friesian cattle using coat pattern recognition. Such models achieved an impressive 93.8% accuracy in identifying unseen individuals in test data, indicating strong generalization using SoftMax-based triplet loss mechanism [22].

The Vision Transformer (ViT) marked a paradigm shift in computer vision (CV) by applying transformer-based architecture, which was previously developed for natural language processing (NLP), directly to image patches [4]. Vision transformers ViTs learn long-range visual dependencies, which means that they are well-suited to fine-grained classification tasks by using the mechanisms of self-attention [4]. Transformer-based methods of cattle identification have been reported in recent literature. Specifically, Kumar et al. presented a multi-directional shifted patch encoding transformer models of non-invasive cattle detection and proved that patch-based detection can be used to successfully extract fine-grained biometric characteristics of cattle in image data [23]. The effectiveness of deep learning and machine-learning methods in predicting crop diseases has been validated by recent academic research studies in the agricultural sector. As an example, an interpretable machine-learning framework based on the Internet-of-Things has been applied to predict disease in pearl millet and provides high accuracy in real-world agricultural conditions [24]. Likewise, a hierarchical ViT (HViT) network, in a poultry disease analysis, reached an accuracy of 90.90% on validation, which is higher than all CNN baselines such as ResNet50 and VGG16 [25]. All these findings indicate a great applicability of ViTs to various fields of agriculture, especially in areas where discriminative features, including the pattern of coats, the shape of the muzzle, or facial features, play a critical role in the successful classification.

The Swin Transformer, which proposes a hierarchical structure with shifted-window self-attention, alleviates the weaknesses of ViTs in terms of data inefficiency and scale variance, improving efficiency and accuracy in general computer-vision problems [5]. The adoption of this in livestock management is not yet complete, but recent studies have shown its potential. In one of these studies, a Swin Transformer using a triplet network has been shown to recognize cattle nose-print with high accuracy (98.61% on a small dataset), demonstrating the feasibility of fine-grained biometric recognition in precision livestock farming despite the existing lack of data [26]. More recent transformer-based approaches, such as CattleDiT, a distillation-based transformer architecture, have been able to achieve higher accuracy in identifying cattle and also reduce the computational cost [27].

Despite recent studies demonstrating significant improvements using deep learning and transformer-based approaches, several limitations remain. Most of the existing works are evaluated on small-scale datasets with limited class diversity, restricting their applicability to real-world livestock environments. More recent works have explored computer-vision-based livestock analysis beyond identification, such as age estimation via dentition analysis using AI, to demonstrate how deep learning can be used to support a wide range of livestock management-related tasks, such as health monitoring and breeding analysis. However, the majority of the existing studies focus on particular tasks and do not provide a full-scale reference of modern deep-learning networks in full recognition of cattle on a large scale. Therefore, there has been no formal comparison of Vision Transformer (ViT) and Swin Transformer systems with long-established convolutional neural systems like ResNet -18 on large-scale cattle-face classification, especially in region-specific applications like Pakistani cattle databases. This gap is addressed by the current research, which compares these architectures on a standardized training procedure on a 459-class cattle-face dataset, therefore, offering practical implications of AI-based livestock control in real-life settings.

Materials and Methods:

Dataset Description: This study utilized the Cows Frontal Face Dataset [7], comprising approximately 459 classes of cows in Karachi, Pakistan, and is one of the largest datasets available for cattle classification studies. Each class corresponds to an individual animal identity, making the task suitable for individual cattle identification. The dataset employs facial and muzzle biometric features and provides a challenging benchmark for evaluating the performance of convolutional neural network (CNN) and transformer-based models in precision livestock farming.

Figure 1. Dataset Samples

Image Preprocessing: All images were scaled to 224×224 pixels, compatible with the input requirements of ResNet-18, Swin Transformer, and Vision Transformer models. To match pretrained weights, the ImageNet mean and standard deviation were used to normalize the images. This preprocessing step ensures consistent input dimensions, reduced computational complexity, and enhanced convergence during training by standardizing pixel intensity distributions.

Proposed Pipeline:

The overall pipeline of the proposed system consists of the following steps:

Dataset acquisition.

Image preprocessing includes resizing and normalization.

Data splitting into training and testing subsets.

Model training using ResNet18, Vision Transformer, and Swin Transformer architectures.

Model evaluation based on classification metrics.

The trained models are then used for inference to identify individual cattle from facial images.

Model Architectures:

ResNet18: ResNet-18 is a deep learning-based 18-layer convolutional neural network (CNN) that uses residual learning to mitigate the vanishing gradient problem, thus allowing training of deeper models to be effectively trained [3]. The main idea of ResNet is the residual connection, which was developed as [3].

$$y = F(x, \{Wi\}) + x \quad (1)$$

Here, x is an image input into the residual block, F (x, {Wi}) represents the residual mapping (stacked convolutional layers), and y is the output. In this research, ResNet18 was used as a baseline model. Standard residual blocks of ResNet18 consisting of convolution, batch normalization, and ReLU activation layers were used. The last fully connected layer was replaced to match the 459 classes of the cattle face dataset, while the other layers retained pretrained weights from ImageNet.

Vision Transformer (ViT): The model implements a transformer-based architecture, which was initially used in natural language processing, to image patches [4]. Every image is divided into fixed-size patches, embedded linearly and treated using multi-head self-attention layers, which can be expressed as [5]:

$$\text{Attention}\ (Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d}})\ V \quad (2)$$

where Q, K, V represent the query matrix, key matrix, and value matrix based on patch embeddings, and d is the key dimension. This mechanism enables the model to capture long-range dependencies between image patches. For this research, the ViT-Base Patch16-224 model was applied, in which the images are divided into 16×16 patches, which creates a chain of embedded tokens. The model has 768 embedding dimensions and 12 attention heads, enabling effective modeling of global contextual relationships.

Swin Transformer: This model uses a shifted-window attention mechanism, which makes features represented hierarchically with a minimal complexity of computation. The self-attention is calculated initially in non-overlapping windows and then moved to enable cross-window interaction [26].

$$y\_W = W - MSA(x\_W) + x\_W \quad (3)$$

$$y = SW - MSA(y\_W) + y\_W \quad (4)$$

where W-MSA and SW-MSA indicate window-based and shifted window multi-head self-attention, respectively. The Swin Transformer Base model is a hierarchical model with a 4×4 patch size and a 7×7 shifted window mechanism. The model uses multi-head self-attention on local windows and shifted windows, which enables the efficient learning of local and global features.

Training: The dataset and model training parameters are given in Table 1. The models were trained under the same experimental conditions to ensure a fair comparison of performance metrics.

Table 1. Dataset and Training Parameters Summary

Parameters	Description
Dataset	Cows Frontal Face Dataset
Number of Classes	459
Image Size	224 × 224 pixels
Training/Test Split	Balanced subsets
Batch Size No of Epoch	32 100
Preprocessing	Resize, To Tensor, Normalize
Loss Function	CrossEntropyLoss
Optimizer (ResNet18)	Adam, lr = 0.001
Optimizer (Swin Transformer)	AdamW, lr = 1e-4
Optimizer (ViT)	AdamW, lr = 1e-4
Pretrained Weights	ImageNet
Data Augmentation	None
Training Environment	GPU-enabled PyTorch

A learning rate scheduler was employed to reduce the learning rate during training for improved convergence. Early stopping was also taken into consideration to avoid overfitting by tracking validation loss. All models were trained for 100 epochs, and the best-performing model weights were selected based on validation performance.

To promote real-time inference in edge or GPU-enabled devices, The proposed models are designed for practical deployment in the systems of livestock monitoring. Transformer-based models prove to be much more efficient, and ResNet18 is lightweight,

thus suitable to implement it in farms to conduct automated cattle identification. These systems can be used to monitor continuously, identify diseases, and analyze their productivity in precision livestock farming.

Results and Discussion:

The section outlines the experimental findings, conducts detailed performance analysis, and provides a critical discussion on the suggested classification models based on the three deep-learning architectures.

Training Dynamics: The training loss and validation loss convergence curves of the three classification models provide useful information about the stability of learning, convergence rate, and the ability to generalize the information to the dataset.

Vision Transformer ViT (Training, Validation Loss Convergence Curve):

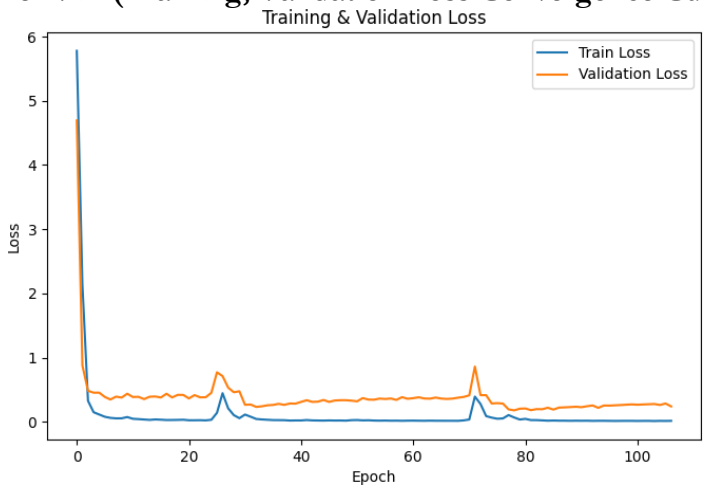

Figure 2. Training Validation Loss Convergence Curve ViT

In figure 2 the ViT model shows a rapid decrease in both training and validation loss at the initial epoch, indicating effective feature learning of global representations. After the convergence, the training loss decreases, whereas the validation loss has minor fluctuations, which is indicative of slight overfitting. Periodic spikes of the loss in later epochs demonstrate that there is sensitivity to the optimization dynamics; however, the stability is generally high, and the final validation loss is not excessive, which demonstrates strong generalization in the multi-class dataset.

ResNet-18 (Training, Validation Loss Convergence Curve):

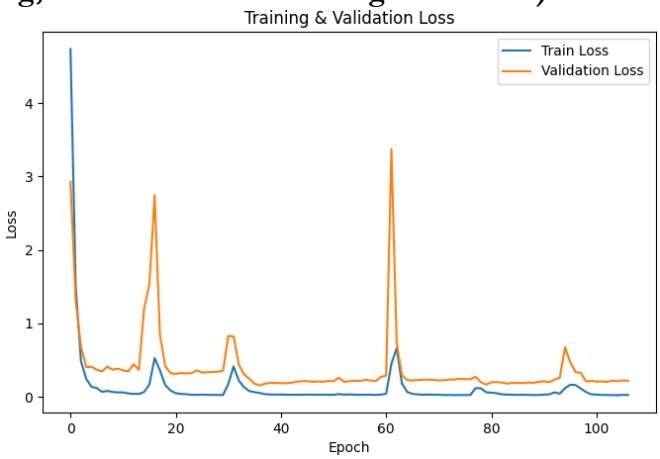

Figure 3. Training Validation Loss Convergence Curve ResNet-18

In figure 3 ResNet-18 exhibits smooth convergence, and the training loss and validation loss values decline continuously with the number of epochs. The close gap between the training loss and the validation loss emphasizes good generalization and strength. The

validation loss spikes vary temporarily, and the trends quickly stabilize, which is indicative of noise-resistance and the effective learning of features. Generally, ResNet-18 has consistent and stable training dynamics and little overfitting.

Swin-ViT (Training, Validation Loss Convergence Curve):

Figure 4. Training Validation Loss Convergence Curve Swin ViT

The Swin Transformer has a rapid initial convergence, which is a steady decrease in loss in the initial epochs as shown in figure 4. Training losses continue to drop significantly as the training continues, but validation loss levels off and even increases in each successive epoch, suggesting over-fitting due to large model capacity. However, Swin-ViT maintains competitive performance with respect to validation performance, which highlights its competence to represent hierarchical and local-global features interaction in complex visual data.

In all models, ResNet-18 exhibits the most stable training dynamics and better generalization, but ViT and Swin-ViT have faster convergence and will reduce training losses. In later epochs, they show evidence of overfitting. Transformer-based architectures are adept at encoding complex relationships among visual features, whereas ResNet-18 offers a better compromise between stability and generalization. These results highlight a trade-off between model complexity and training robustness in large-scale multi-class classification tasks.

Evaluation Metrics: This research utilized the Cows Frontal Face Dataset [7], comprising approximately 459 classes of cows in Karachi, Pakistan, which is one of the largest datasets available for cattle classification research.

Top-1 Accuracy (Overall Accuracy): Top-1 accuracy measures the fraction of test samples that the top-1 prediction of the model exactly matches the ground-truth class label [28].

$$Top-1\ Accuracy\ =\ \frac{1}{N}\sum_{i-1}^{N}1(yi\ =\ yi) \qquad (5)$$

Precision: Precision is used to determine the percentage of observed positive cases that are predicted to be positive [29].

$$Precision\ =\ \frac{TP}{TP+FP} \qquad (6)$$

Recall: Recall calculates how many of the actual positive cases were predicted correctly by the model [29].

$$Recall\ =\ \frac{TP}{TP+FN} \qquad (7)$$

F1-score: Precision and Recall are the harmonic mean to generate the F1-score. It strikes a trade-off between the two, particularly in the case of imbalanced class distribution [29].

$$F1\ =\ 2\times\frac{Precision\times Recall}{Precision+Recall} \qquad (8)$$

Top-5 Accuracy: Top-5 Accuracy calculates the proportion of test samples for which the ground-truth class appears among the model's top five predicted classes ranked by confidence [28].

$$Top - 5\ Accuracy\ = \frac{1}{N}\sum_{i-1}^{N} 1(yi\ \epsilon\ Top\ 5\ (pi)) \quad (9)$$

Mean Class Accuracy (MCA): Mean Class Accuracy is a measure of the mean classification accuracy across all classes, with the same weight given to each of the classes regardless of its size [30].

$$MCA\ = \frac{1}{C}\sum_{c=1}^{c} \frac{TPc}{Nc} \quad (10)$$

Results: In this research, we evaluated three deep learning architectures, Swin Transformer, Vision Transformer (ViT), and ResNet-18, on our dataset. Table 2 summarizes the classification metrics, including top-1 accuracy, precision, recall, and F1-score, top-5 accuracy, and mean class accuracy.

Table 2. Comparative Performance Evaluation of Models

Model	ViT	ResNet-18	Swin Transformer
Top 1 Accuracy	96.27%	95.86%	94.48%
Precision (weighted)	95.63%	95.46%	93.83%
Recall (weighted)	96.27%	95.86%	94.48%
F1- Score (weighted)	95.64%	95.31%	93.71%
Top 5 Accuracy	99.58%	99.58%	99.17%
Mean Class Accuracy	95.02%	94.65%	93.29%

Observations:

Top-1 Accuracy: The Vision Transformer has achieved the best classification accuracy of 96.27%, as compared to ResNet-18 (95.86%) and Swin Transformer (94.48%).

Precision & Recall: The models showed similar weighted precision (93-96%) and recall (94-96%) values, with a good performance with minimal false negatives and false positives.

F1-score: ResNet-18 and Swin Transformer achieved an F1-score of 95.31% and 93.71%, respectively, and ViT slightly outperformed them with 95.64%, showing a better precision-recall balance.

Top-5 Accuracy: Models all had high Top-5 accuracy, which highlights the strong identification performance under circumstances where the correct class is not at the top. ViT and ResNet-18 both performed with 99.58%, and Swin Transformer scored 99.17%.

Mean Class Accuracy: ViT achieved the most accurate mean class (95.02%), followed by ResNet-18 (94.65%) and Swin Transformer (93.29%), indicating the robustness of these models on both common and rare classes.

Misclassification Analysis: Due to the large number of classes (459), a complete confusion matrix is difficult to visualize; therefore, analysis is focused on misclassification trends among visually similar cattle classes, where most errors are based on high inter-class similarity.

Mapping Research Objectives:

The experimental results directly address the defined research objectives. Objective 1 is met by the comparative analysis of CNN and transformer-based models, in which ViT demonstrated the best classification accuracy. Objective 2 is affirmed through the fact that transformer-based architecture has been better than conventional CNN architectures in fine-grained features representation, as it has been supported by a higher degree of precision and recall. The high Top-5 accuracy (>99%) in all models supports objective 3 and shows that it can be used in practice to monitor animals.

Discussion:

The experimental evaluation of the three considered models, Swin Transformer, Vision Transformer (ViT), and ResNet-18, showed that ViT demonstrated the best

classification accuracy of 96.27% and, a precision of 95.63%, and a recall of 96.27%. ResNet-18 achieved an accuracy of 95.86%, and Swin Transformer reached a score of 94.48%. Transformer architecture outperformed the baseline convolutional neural network significantly with respect to all performance metrics.

The high performance of the Swin Transformer can be explained by the hierarchical representation learning and shifted-window self-attention, which allows simultaneously using both local and global contextual patterns in bovine images. These abilities are particularly relevant to livestock identification problems where they are necessary to use fine-grained discrimination of coat colors, ear shape, and facial features. However, in the current dataset, the ViT slightly outperformed the Swin Transformer on the Top-1 accuracy, which is indicative of the possibility that a pure global attention mechanism is more consistent with the given features.

Comparing transformers and CNNs, the results demonstrate that while CNNs (ResNet-18) remain competitive, they heavily utilize local receptive fields, which may limit their ability to capture global relationships in high-resolution livestock imagery. Transformers, however, are more effective at learning long-range dependencies, facilitating more successful discrimination in cases where animals have similar local textures, yet different overall shape or pattern configurations.

Practically, these results indicate that transformer-based architectures are better adapted to real-world livestock monitoring systems, where animals can be partially occluded or appear at different distances, or be captured under varying illumination conditions. This robustness is significant towards uncontrolled farm deployment.

Conclusion and Future Work: This research experimented on livestock recognition using Swin Transformer, Vision Transformer, and ResNet-18. Thus, the Vision Transformer became the most effective, achieving the highest classification accuracy of 96.27% and demonstrating competitive performance across all evaluation metrics. The results verify that transformer-based architecture offers a competitive advantage in the performance of traditional CNNs in identifying fine-grained animals. The findings further confirm the effectiveness of transformer-based architectures for fine-grained cattle identification. The results accommodate the specified research objectives since they prove the comparative performance of CNN and transformer-driven models as well as indicate their relevance to large-scale cattle classification. Furthermore, the high accuracy in classifications shows that these models can be used in real-world livestock monitoring systems.

This method in livestock surveillance systems can enable automated recognition and tracking, hence facilitating the application of precision agriculture, such as individual health monitoring, feeding-behavior identification, and breeding control. The proposed models will be beneficial in reducing the dependence on manual tagging, thereby facilitating efficient farm operations and animal welfare. Overall, the results indicate that transformer-based architectures can significantly enhance AI-powered livestock monitoring systems and thus allow scalable and sustainable agricultural solutions. The suggested framework also supports the United Nations Sustainable Development Goals, especially SDG 2 (Zero Hunger) and SDG 9 (Industry, Innovation and Infrastructure), as it allows for managing livestock efficiently, increasing productivity, and encouraging innovation in precision agriculture.

Future research must enhance this framework with the additional use of multimodal data combining visual recognition with supplementary data, video-based behavioral analytics, RFID data, and biometric data. It is expected that such a multimodal approach will increase the robustness and reliability of the system in more complex farm conditions. Furthermore, it will be useful to expand the dataset to more heterogeneous and diverse cohorts, i.e., include various breeds, different environmental factors, and different occlusion scenarios, as it will make the model more generalized and applicable to the real-world agriculture context.

Research on real-time inference by the implementation of optimized transformer models on edge devices would allow real-time monitoring of devices on-site without the need to have high-bandwidth connectivity. Lastly, the use of Explainable AI (XAI) frameworks is likely to be critical towards promoting model transparency and trust. XAI can encourage farmers and veterinarians to use it by providing interpretable insights into model decisions, which will facilitate informed decision-making. Comprehensively, this study shows that transformer-based architectures have significant flexibility in enhancing AI-based livestock monitoring, which can be used to create more scalable, intelligent, and sustainable agricultural solutions.

References:

[1] Munir Ahmad, Sagheer Abbas, "AI-Driven livestock identification and insurance management system," *Egypt. Informatics J.*, vol. 24, no. 3, p. 100390, 2023, doi: https://doi.org/10.1016/j.eij.2023.100390.

[2] Guoming Li, Galen E. Erickson, "Individual Beef Cattle Identification Using Muzzle Images and Deep Learning Techniques," *Animals*, vol. 12, no. 11, 2022, doi: 10.3390/ani12111453.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385*, 2015, [Online]. Available: https://arxiv.org/abs/1512.03385

[4] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, Oct. 2020, Accessed: May 16, 2024. [Online]. Available: https://arxiv.org/abs/2010.11929v2

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *arXiv:2103.14030*, 2021, [Online]. Available: https://arxiv.org/abs/2103.14030

[6] Christian Lamping, Gert Kootstra, "Transformer-based similarity learning for re-identification of chickens," *Smart Agric. Technol.*, vol. 11, p. 100945, 2025, doi: https://doi.org/10.1016/j.atech.2025.100945.

[7] Syed Umaid Ahmed, Jaroslav Frnda, "Dataset of cattle biometrics through muzzle images," *Data Br.*, vol. 53, p. 110125, 2024, doi: https://doi.org/10.1016/j.dib.2024.110125.

[8] Hua Meng, Lina Zhang, "Livestock Biometrics Identification Using Computer Vision Approaches: A Review," *Agriculture*, vol. 15, no. 1, p. 102, 2025, doi: https://doi.org/10.3390/agriculture15010102.

[9] Ishana Attri, Lalit Kumar Awasthi, "A review of deep learning techniques used in agriculture," *Ecol. Inform.*, vol. 77, p. 102217, 2023, doi: https://doi.org/10.1016/j.ecoinf.2023.102217.

[10] "Goal 2: Zero Hunger - United Nations Sustainable Development." Accessed: Mar. 17, 2026. [Online]. Available: https://www.un.org/sustainabledevelopment/hunger/

[11] "Goal 3 | Department of Economic and Social Affairs." Accessed: Sep. 02, 2025. [Online]. Available: https://sdgs.un.org/goals/goal3#targets_and_indicators

[12] Md Ekramul Hossain, Muhammad Ashad Kabir, "A systematic review of machine learning techniques for cattle identification: Datasets, methods and future directions," *Artif. Intell. Agric.*, vol. 6, pp. 138–155, 2022, doi: https://doi.org/10.1016/j.aiia.2022.09.002.

[13] Meghna Luthra, Meghna Sharma, Poonam Chaudhary, "Comprehensive Survey on Cattle Identification Approaches: From Traditional to Deep Learning Aspects," *J. Basic Sci. Eng.*, vol. 21, no. 1, 2024, [Online]. Available: https://www.yjgkx.org/uploads/archives/f3376599-35f1-46e5-8ad6-29c2269fdb27.pdf

[14] Dongxu Li, Baoshan Li, Qi Li, Yueming Wang, Mei Yang & Mingshuo Han, "Cattle identification based on multiple feature decision layer fusion," *Sci. Rep.*, 2024, [Online]. Available: https://www.nature.com/articles/s41598-024-76718-x

[15] Y. C. Binghao Ye, "Cow individual identification method based on local feature guided

neural network," *CFIMA 2024 - Proc. 2024 2nd Int. Conf. Front. Intell. Manuf. Autom.*, p. 109718, 2025, [Online]. Available: https://dl.acm.org/doi/10.1145/3704558.3704597

[16] A. Bhujel, Y. Wang, Y. Lu, D. Morris, and M. Dangol, "A systematic survey of public computer vision datasets for precision livestock farming," *Comput. Electron. Agric.*, vol. 229, p. 109718, Feb. 2025, doi: 10.1016/j.compag.2024.109718.

[17] A. Inam, H. Rehman, M. F. Shahid, and S. U. Ahmed, "Leveraging Artificial Intelligence for Age Prediction in Cattle based on Dentition," *Proc. 2025 4th Int. Conf. Comput. Inf. Technol. ICCIT 2025*, pp. 496–501, 2025, doi: 10.1109/ICCIT63348.2025.10989431.

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998, doi: 10.1109/5.726791.

[19] Elif Akarsu, Tevhit Karacalı, "Combining Deep Features with Classical Discriminants: High-Accuracy Animal Classification Using ResNet-18 and LDA," *Int. J. Innov. Res. Rev.*, vol. 9, no. 2, 2025, [Online]. Available: https://www.injirr.com/article/view/253

[20] Shijun Li, Lili Fu, "Individual dairy cow identification based on lightweight convolutional neural network," *PLoS One*, 2021, [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0260510

[21] Bingxuan Li, Jiandong Fang & Yvdong Zhao, "RTDETR-Refa: a real-time detection method for multi-breed classification of cattle," *J. Real-Time Image Process.*, vol. 22, 2025, [Online]. Available: https://link.springer.com/article/10.1007/s11554-024-01613-7

[22] W. Andrew, J. Gao, S. Mullan, N. Campbell, A. W. Dowsey, and T. Burghardt, "Visual identification of individual Holstein-Friesian cattle via deep metric learning," *Comput. Electron. Agric.*, vol. 185, p. 106133, Jun. 2021, doi: 10.1016/j.compag.2021.106133.

[23] N. Kumar and S. K. Singh, "Multi-Directional Shifted Patch Encoding With Transformers for Non-Invasive Cattle Identification," *IEEE Trans. Biometrics, Behav. Identity Sci.*, vol. 7, no. 4, pp. 620–631, 2025, doi: 10.1109/TBIOM.2025.3599374.

[24] Nidhi Kundu, Geeta Rani, "IoT and Interpretable Machine Learning Based Framework for Disease Prediction in Pearl Millet," *Sensors*, vol. 21, no. 16, p. 5386, 2021, doi: https://doi.org/10.3390/s21165386.

[25] Michael Agbo Tettey Soli, Dacosta Agyei, Waliyyullah Umar Bandawu, Leonard Mensah Boante, Justice Kwame Appati, "A Modified Hierarchical Vision Transformer Model for Poultry Disease Detection," *IET Image Process.*, vol. 19, no. 1, 2025, [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.70115

[26] Minyue Zhong, Yao Tan, "A Method for Recognition of Cattle Noseprint based Fusing Swin Transformer and Triplet Network," *ACM Int. Conf. Proceeding Ser.*, 2024, [Online]. Available: https://dl.acm.org/doi/10.1145/3652628.3652716

[27] N. Kumar and S. K. Singh, "CattleDiT: A Distillation-Driven Transformer for Cattle Identification," *IEEE Trans. Biometrics, Behav. Identity Sci.*, vol. 7, no. 4, pp. 824–836, 2025, doi: 10.1109/TBIOM.2025.3565516.

[28] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis. 2015 1153*, vol. 115, no. 3, pp. 211–252, Apr. 2015, doi: 10.1007/s11263-015-0816-y.

[29] D. . Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv Prepr. arXiv2010.16061*, 2020, [Online]. Available: https://arxiv.org/pdf/2010.16061

[30] G. L. Marina Sokolova, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009, doi: https://doi.org/10.1016/j.ipm.2009.03.002.