Smart Agricultural Technology 10 (2025) 100774

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

Vision transformers for automated detection of pig interactions in groups

Gbadegesin Taiwo *, Sunil Vadera, Ali Alameer

University of Salford, United Kingdom

ARTICLE INFO

Keywords:
Precision agriculture
Livestock management
Machine vision
Artificial intelligence in agriculture
Animal Behavior Detection

ABSTRACT

The interactive behaviour of pigs is an important determinant of their social development and overall well-being. Manual observation and identification of contact behaviour can be time-consuming and potentially subjective. This study presents a new method for the dynamic detection of pig head to rear interaction using the Vision Transformer (ViT). The ViT model achieved a high accuracy in detecting and classifying specific interaction behaviour as trained on the pig contact datasets, capturing interaction behaviour. The model's ability to recognize contextual spatial data enables strong detection even in complex contexts, due to the use of Gaussian Error Linear Unit (GELU) an activation function responsible for introduction of non-linear data to the model and Multi Head Attention feature that ensures all relevant details contained in a data are captured in Vision Transformer. The method provides an efficient method for monitoring swine behaviour for instance, contact between pigs, facilitating better livestock management and livestock welfare. The ViT can represent a significant improvement on current automated behaviour detection, opening new possibilities for accurate animal design and animal behaviour assessment with an accuracy and F1 score of 82.8 % and 82.7 %, respectively, while we have an AUC of 85 %.

Introduction

Pig behaviour influences both pork and financial profit by reflecting the health and growth state of the animals. Pigs need to be given close attention in order to have their behaviour monitored and recognized so it can be precisely managed [1]. Advancements in big data and technology are revolutionising livestock farming by enabling the tracking of animal activity through various sensors. These innovations incorporate artificial intelligence (AI), machine learning (ML), information and communication technology (ICT), and video surveillance. By leveraging AI and ML, these technologies are driving sustainable rural development and transforming the future of agriculture [2]. Three-axis acceleration sensors, for instance, are employed to track sows' prenatal behaviour traits in real time [3]. In order to enable precision feeding, Radio Frequency Identification (RFID) a technology that uses a radio device and a tag to identify an object, has replaced traditional ear tags, and pressure sensors are employed to track a sow's movements during parturition [4,5]. The applications of these in the real-world, reveals sensors' drawbacks are becoming more apparent [6]. The stress of putting on wearable devices on animals is another factor for consideration. There is an observed decline in pigs' mobility and touch due to installation of some field sensors put on them to generate some metrics Maselyne et al. [7]. As a result, non-contact computer vision technology has increasingly been adopted in commercial pig farming to monitor daily activities such as feeding, fighting, and drinking [8,9].

Deep learning has gained popularity in the field of computer vision and has seen several successes with object detection and image classification [10–12]. Object detection is one of the hotspots in computer vision as an extension of the image classification job. It involves not only classifying objects in an image but also locating the object's location and defining a bounding box around it. There are two types of deep learning algorithms for object detection: one-stage and two-stage. Two-stage techniques are required to construct an object-containing anchor box first, followed by fine-grained object recognition. With algorithms such as R-CNN used for representative models [13], though R-CNN [14], and SPPNet [15] possessed a higher accuracy but are slow in speed. On the other hand, single-stage algorithms, which are exemplified by the YOLO [16] series, SSD [17], and CenterNet [18], directly extract features from the network to predict the position coordinate of the object's class probability. As a result, they have a better balance between detection speed and accuracy than two-stage models.

Several approaches, including automated surveillance, have been utilised to monitor pig behaviour, daily activities, and drinking and feeding patterns, with the aim of enabling early detection of welfare and

* Corresponding author at: 43 Crescent, Salford M5 4WT, United Kingdom.
E-mail address: G.A.Taiwo@edu.salford.ac.uk (G. Taiwo).

https://doi.org/10.1016/j.atech.2025.100774
Received 4 October 2024; Received in revised form 5 January 2025; Accepted 6 January 2025
Available online 7 January 2025

health issues [19,20].

Hence, this gives room for the possibility that pigs abnormal behaviour could be captured and monitored with computer vision systems. One of these approaches is the MOT (Multi-Object Tracking) used by some researchers to monitor animal behaviours. Gan et al. [21] created a Convolutional Neural Network (CNN), method to classify various social activities among preweaning piglets in the swine. Alameer et al. [22] employed integration of two deep-learning developed detectors with the ability to track procedures in detection of pig's stance and drinking habits. Though the approach might be expensive, Psota et al. [23] employed a probabilistic tracking detector to monitor individual pigs in a group with identification of each pig critical spots with a CNN detector. While, Bhujel et al. [24]. developed a deep-learning-based pig posture and tracking technique to measure behavioural changes in pig environment with different greenhouse gas (GHG) levels, although Tu et al. study reveals that pig tracking and identification is still a major problem, considering target occlusion, light changing, incorrect ID tracking, and overlapping. For pig behaviour tracking applications, advanced detectors and MOT methods are being created in order to enhance the performance of the detector and tracker. Because existing methods rely on local receptive fields, they face limitations in addressing the issue of occlusion, which presents some accuracy issues. Additionally, the current solutions rely on picture masking and the usage of bounding boxes, both of which are labour-intensive and less effective.

This study therefore employs the use of ViT in detection of contact behaviour in pig-pen, considering ViT advantages over existing models such as contextual capturing of image information including long-range, needs no requirement for object detection hence can assist in capturing pig behaviour instantly, its encoder-only architecture helps to reduce overfitting on images, and faster at making inference compare to existing models.

2. Literature review

This section of the study discusses relevant literature related to application of machine learning and deep learning technology in exploring and modelling pig aggressive behaviour.

Machine learning approach

Conventional approaches to animal behaviour monitoring mostly rely on human eye observation, which is labour-intensive and prone to subjectivity. Pig breeding businesses are using automated video recognition techniques more often as a result of advancements in image and video processing technologies. Gronskyte et al. [25] used a charge-coupled device (CCD) camera to monitor the movement of pigs using an optical flow vector and fitted ellipse features in consecutive frames Kashiha et al. [26]. Weixing and Jin [27] used videos collected from the pigs' sides to create a joint angle waveform for the purpose of detecting lameness behaviour. While Li et al. [28] studies on particle filtering are reliable enough to be used in pig tracking. Weixing and Zhilei [29] has introduced an automatic method to identify respiratory behaviour in pigs that is based on the Freeman Code algorithm. Pigs' appearance in their typical living area has been used to identify a number of behaviours, including eating, drinking, and excretory behaviour, with a high degree of general sensitivity and accuracy Zhu et al. [30]. Using automatic video processing approaches based on fitted ellipse and dense trajectories features, it is possible to observe even the act of lying Nasirahmadi et al. [31].

More in-depth dimension information has been mined for behaviour identification in the last few years as 3D technology has advanced quickly. 3D point clouds and 2D depth photos are obtained using Kinect2.0's time of flight capability. Mittek et al. [32] collected pigs' cloud point information from 3D point cloud photos and tracked it using ellipsoidal fitting, Lee et al. [33] had utilised 2D depth photographs to offer an automatic recognition technique for aggressive behaviour detection. Andriamandroso et al. [34] assessed the iPhone 5s's Inertia Measurement Units (IMU) when it was placed on a cow's neck and developed the Decision Tree (DT), which has a 91 % accuracy rate, 91.1 % sensitivity, 90.9 % specificity, and 93.5 % precision in detecting grazing activity. 96.5 % accuracy, 53.1 % sensitivity, 99.4 % specificity, and 84.5 % prediction are obtained for rumination, whereas 87.6 % accuracy, 87.6 % sensitivity, 87.5 % specificity, and 79.1 % prediction are obtained for other activities. The effect of accelerometer/magnetometer placement (ear tag, collar (under neck), and halter) on the accuracy of grazing, standing, and ruminating classifications has been investigated by [35] The 3D accelerometer on the collar measures 12 Hz, while the accelerometer/magnetometer sample for the halter and ear tag measures 30 Hz. Stratified Cross Validation (SCV) and Leave-Out-One-Animal (LOOA) techniques were used to evaluate and apply the Random Forest Algorithm (RFA). As a matter of fact, the findings indicate that halters with Stratified Cross Validation (SCV) F-Scores of 91.4 %, 89 %, and 93.2 %, respectively, are superior for grazing, standing, and ruminating behaviours.

Barker et al. [36] compare the performance of decision tree built using an accelerometer measured at 12.5 Hz and a location to categorise actions (such as eating, non-feeding, and milking) on the one hand, and on the other, (lame and non-lame). For the analysis, a window size of 2 s was used. For milking, they received the following results for behaviour classification: accuracy: 94.2 %, sensitivity: 95.6 %, specificity: 94.0 %, and precision: 59.9 %. The non-feeding behaviour performances have the following parameters: sensitivity of 74.9 %, specificity of 91.3 %, accuracy of 80.8 %, and precision of 93.9 %. Performance metrics for feeding behaviour classification include 83.2 % accuracy, 83.5 % precision, 93 % specificity, and 65.3 % sensitivity. Additionally, they demonstrate that lame cows eat for shorter periods of time during the afternoon and throughout the day. [37], applied GPS data recorded at 0.2 Hz in grazing, resting, and walking behaviours are classified using four machine learning models (Random Forest, Naïve Bayes, J48, and JRip) and data mining techniques to extract characteristics. With the help of 10-fold cross-validation, the evaluation was completed. In terms of average accuracy, JRip and Random Forest performed the best, with respective F-measures of 76 % and 77 % and average accuracy of 85 % and 83 %.

Deep learning application

[38] introduced a novel technique for cow detection using side view photos and lightweight convolutional neural networks. The final recognition rate of the system was 97.95 %.

[39] employed the YOLO model to identify cow targets in a set of side-view cow photos. They then classified each cow by optimising the convolutional neural network model, achieving 96.65 % accuracy in each individual cow recognition. [38] examined deep learning-based techniques for tracking, segmentation, position estimation, target recognition, and classification of various animal species, including pigs, chickens, goats, and cows.

[40] examined three cutting-edge automated multi-object tracking techniques on two different pig datasets; the experimental findings of assessment metrics show the efficacy and resilience of the three suggested approaches on multi-object tracking systems. For individual pigs on a real farm, FairMOT obtains the best tracking performance with the suggested weighted-association technique. Zhang et al. suggested an automated multi-target tracking and detection system that works both during the day and at night for individual pigs. The evaluation's overall results of 94.72 % precision, 94.74 % recall, and 89.58 % MOTA demonstrate that our technique is capable of reliably detecting and tracking many pigs in difficult situations. [41] developed an automated system for identifying social connections with recordings of the pig engagement, the time, and the nature of the social contact were all ascertained by utilising key point-based body part detection in

conjunction with an algorithm for pig tracking.

According to the aforementioned study, computer vision technology based on machine and deep learning has a wide range of interesting applications in agricultural science. However, deep learning for instance can increase the possibilities for animal production in agriculture. There are currently very few study publications on the detection and recognition of pigs' aggressive behaviours and movement features, as the field is still in its early stages of development. Pig aggression can persist for several seconds to minutes and is a complicated interaction activity, which is one of the primary causes. Despite the approaches that have been proposed for contact detection in livestock, challenges including occlusion still exist. Hence, there is a need for an optimal approach that will alleviate these challenges.

3. Method and material

3.1. Dataset description

The dataset used for this study is pig contact dataset [42]; We extracted 433 images, which includes a csv file that contains the labelling of contact and no contact between pigs. The images were normalised with division by 255, the data preparation process include the augmentation with parameter settings such as horizontal flip, vertical flip, width shift range, height shift range, and nearest fill mode, this helps to provide more data point considering that 433 is a small sample size in image analysis, also image data generator was applied to ensure the images are being passed in batch size of 32 in training the Vision Transformer model.

3.2. Overview of ViT model

As highlighted in the previous section, a number of studies have used deep learning for contact detection in livestock with mixed results. Image classification is an unavoidable task in computer vision, this involves assigning a label to an image based on its contents. Deep CNNs, such as YOLOv7 Mimura et al., [43], have long been used as the standard of practice for image classification.

Recent advances in transformer design, which was originally presented for natural language processing (NLP), have shown considerable promise in competitive picture classification problems. The Vision transformer model is a more recent innovation and builds upon the successes of the Transformer model. In this section a model using the ViT was developed.

The Vision Transformer (ViT) model architecture was revealed in a research paper titled "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale" by Dosovitskiy et al. [44].

Vision Transformer (ViT) outperforms CNNs while requiring significantly fewer computational resources for pre-training. The ViT is a visual model built on the architecture of a transformer initially intended for text-based operations. The ViT model represents an input image as a set of image patches, similar to how word embeddings are represented when using transformers to text, and predicts the image's class labels directly. When trained on enough data, ViT outperforms a similar SOTA CNN with four times fewer CPU resources.

3.2.2. Vision transformers pig contact detection

As shown in Fig. 1, the pig contact dataset was received as a size of 224 × 224 pixels and 3 colour channels (RGB) in A, then patch was created with the image divided into 16 by 16 patches in B. Given the image size, the result was in a 14 by 14 grid of patches, which resulted in a total of 196 patches, it was then passed into C for flattening into a 1D vector, which consists of 768 elements (16×16×3 = 768) and vectors then get linearly projected into a sequence of vectors, that's known as embeddings.

In D the positional embeddings were added to each patch embedding in order to maintain the structure and spatial image information, it was then passed into E a special classification token known as [CLS] added at the beginning of the sequence, this makes our vectors equal to 197 (196 patches + 1 [CLS] token), each of size 768. This played a crucial purpose in summarising the entire image information for classification purposes.

The resulting token is passed into F, the Transformer Encoder that makes possible the processing of the sequence, which aids analysing and understanding of our image data, then forwarded into G, [CLS] Token to extract features after processing through the transformer encoder. The

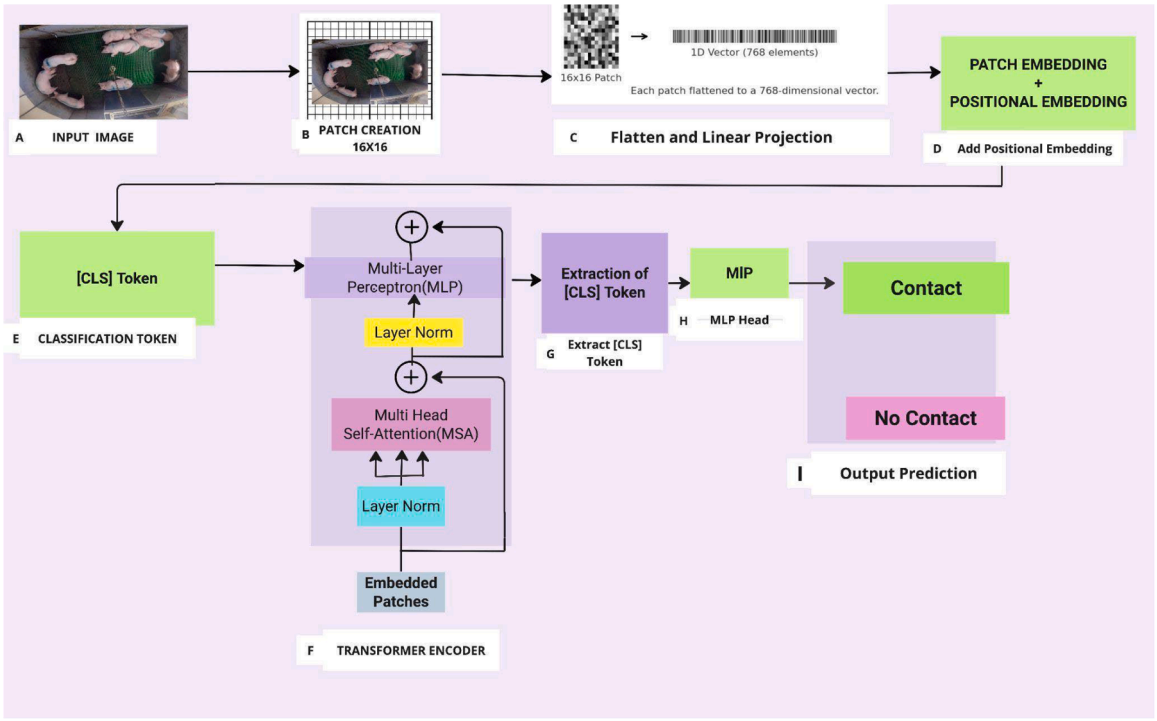

Fig. 1. Vision Transformer (ViT) Architecture for Detecting Pig Contact Behavior.

extracted features are passed into the MLP head (H) where the final application is decided and the output prediction (I) responsible for "Contact" or "No Contact".

3.2.3. Transformer encoder in vision transformer

Considering that Transformer encoder plays a significant role in Vision transformer, hence it becomes important to decentralise how it works. The following steps are involved in the architecture of Transformer encoder as shown in Fig. 2:

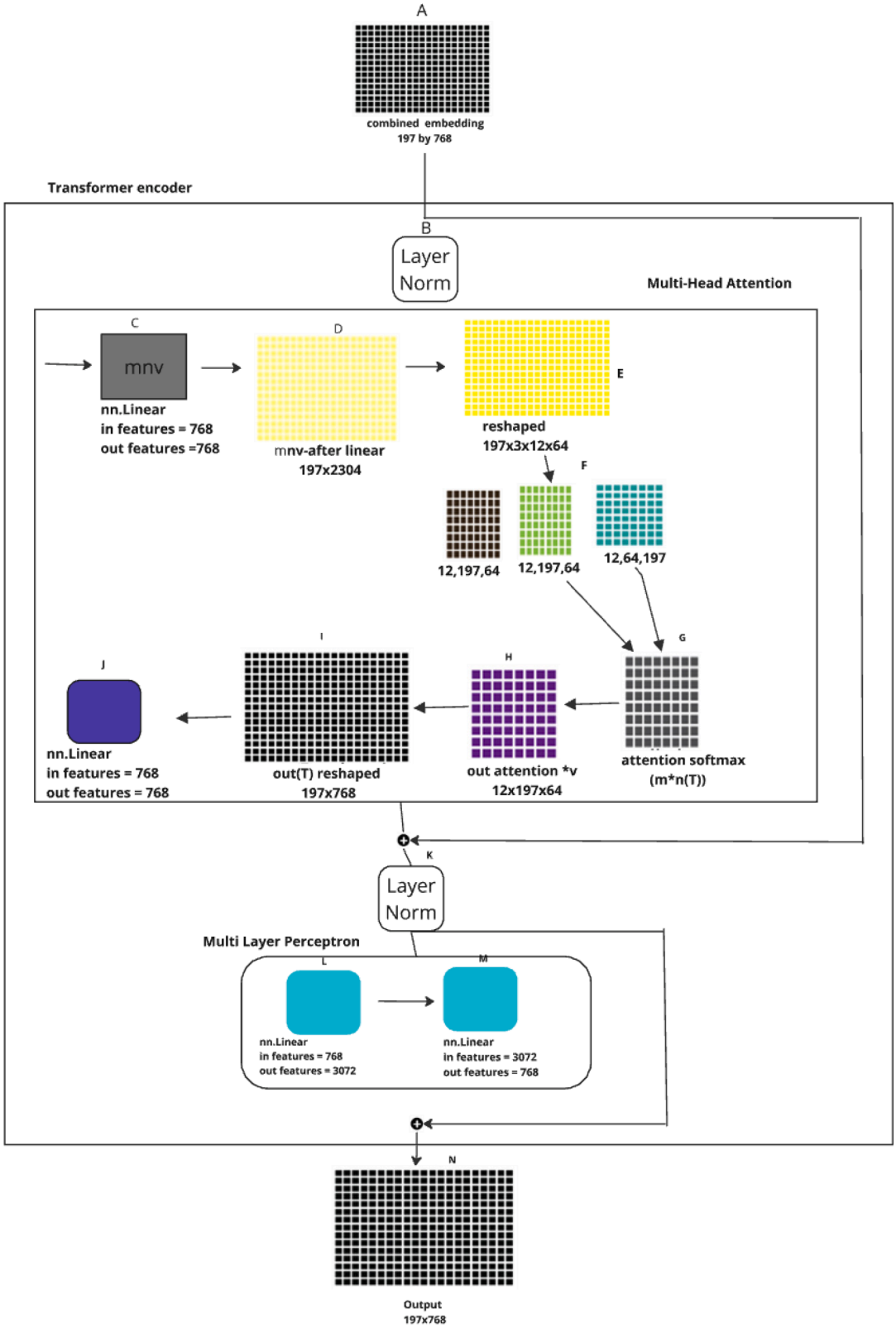

Fig. 2. Vision Transformer (ViT) encoder for Detecting Pig Contact Behaviour.

The A at the first and top layer of the transformer encoder receive the combined embeddings of shape 197 by 768. The input is then passed through the layer norm (B) then used to feed the Multi-Head Attention block. In the Multi Head Attention block, the input data are converted into 197 by 2304, with the use of a linear layer to get a mnv matrix (C). The mnv matrix (C) is then reshaped into 197 by 3 by 768 (D), and each of the matrices represent 197 by 768 (E).

These matrices (mnv) are thereafter reshaped into 12 by 197 by 64 to depict the 12 layers or attention heads (F). Once the matrices m, n, and v have been achieved, the attention operation is performed between the Multi Head Attention block. Then passed into (G) where attention softmax function is applied on the transpose n matrix by m matrix, and the obtained result is passed into (H) then multiplied by v matrix, then forwarded to (I) where the images are being restructured into 197 by 768, then passed into (J), a linear layer with an input and output of 768. This is the processes involved in the Multi-Head Attention block, the obtained result is then added to skip connection to achieve the final output that go through the layer norm (K) again before being passed into the Multi-Layer Perceptron block (MLP) (L and M), that contains two different linear layers with a gaussian error linear unit that serve as the activation function in a non-linear situation. The obtained results are then added to the skip connection to achieve the final result (N) from a single layer of the Transformer encoder.

3.2.3. Further Breakdown of visual transformer detection process

The pig images are received as fixed-size patches that are linearly embedded, then position embedding was added to maintain the structure and content of each image, and the obtained result is fed into a sequence of vectors to a transformer encoder. The classification token with a multilayer perceptron is then added for classification.

The images used in the study have 3 channels (RGB) input of pigs of size 224 by 224. From the figure above, patch size of 16 by 16 is created on the received image, hence a 14 by 14 of such patches are created. Hence, it becomes patching the image with a size of 16 by 16 by 3 that represents the number of channels.

The patch images are taken through the linear projection layer to a single 1 by 768 vector representation for each of the patch images; these patches are called patch embeddings. The patch embeddings of size 196 by 768, the position embeddings are added with a [cls] token for classification on the transformer encoder to the sequence hence both patch and position embeddings for maintaining positional image information of the patches becomes 197 by 768.

Thereafter, the positional and patch embeddings are passed into the transformer encoder and extract the learned representations of the class token. The output from the transformer encoder then becomes 1 by 768 that is then passed into the multi-layer perceptron head linear layer where the class predictions are obtained. The transformer encoder is further broken down from the combination of embeddings through to combined embeddings.

From Fig. 2 above, the combinations of shape 197 × 768 embeddings are accepted as input by the first layer of the Transformer Encoder. The inputs for every layer after that are the 197 × 768 output matrix from the Transformer Encoder's preceding layer. The Transformer Encoder of the ViT-Base architecture consists of a total of 12 of these layers.

The inputs enter the layer and are delivered to the Multi-Head Attention block after passing via a layer norm. To obtain the mnv matrix inside the Multi-Head Attention, the inputs are first transformed using a linear layer to a 197 × 2304 (768×3) shape. The m, n, and v matrices are represented by each of the three 197 × 768 matrices that was created after reshaping this mnv matrix. To represent the 12 attention heads, these m, n, and v matrices are further rearranged to 12 × 197 × 64. After obtaining the m, n, and v matrices, the attention operation is ultimately carried out within the Multi-Head Attention block, as indicated by the following equation:

$$Attention(mnv) = softmax\left(\frac{mn^T}{\sqrt{d_n}}\right)v$$

Where m, n, and v represents matrices of derived input embeddings, d represents the dimensions of the input embeddings, ˆT represents matrix transpose, while softmax is the activation function used for prediction of class probabilities of input pig images.

After the Multi-Head Attention block's outputs are obtained, they are combined to the inputs (skip connection) to produce the final outputs, which are then once more sent to Layer Norm before being supplied to the MLP Block. With two linear layers and a Gaussian Error Linear Unit (GELU) non-linearity, the MLP is a Multi-Layer Perceptron block. To obtain the final output from a single layer of the Transformer Encoder, the outputs from the MLP block are once more connected to the inputs (skip connection).

4. Results

To assess the result of this study that applies Vision Transformer in detection pig contact, at the time of this study there is no study that has applied Vision Transformer in Pig contact detection. This study used metrics such as accuracy and F1 score to evaluate the performance of the model, and the result showed an accuracy score of 82.8 % and F1 score of 82.7 %. Also, in the diagram Fig. 3 shows the classification measures with confusion metrics. Additionally, the model achieved an AUC of 85 %, showing our models strong ability to distinguish between contacts and no-contact instances.

Fig. 3 shows an accuracy result in the prediction of pig head-to-rear contacts or no contacts, the model correctly predicted 55 instances of no-contacts and 17 instances of contacts, but misclassified as no contact in 8 instances, and misclassified as contact in 7 instances. The obtained result is evaluated with AUC and ROC in Fig. 4

Further supporting this is an AUC score of 85 % regarding the effectiveness of the model in the prediction of pig contacts. This metric, therefore, shows the effectiveness of the model in ensuring a high true positive rate while keeping the false positive rate as low as possible, which is very important in practical applications for precision livestock farming.

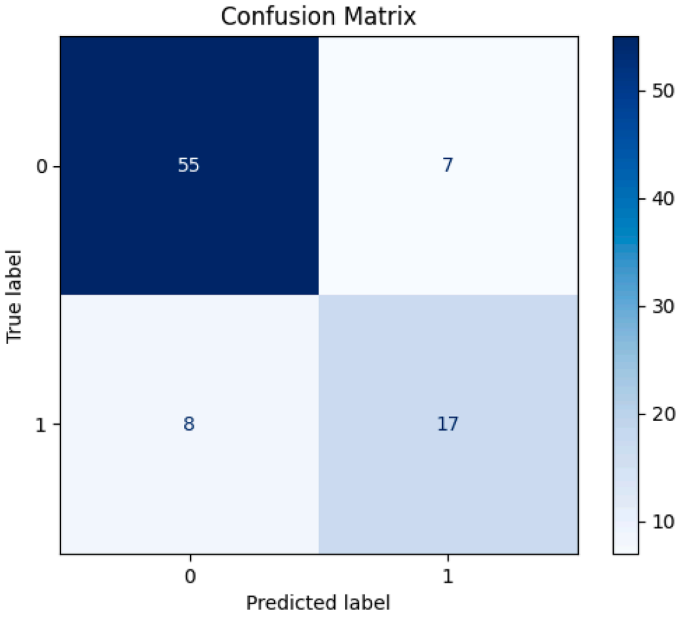

Fig. 3. Confusion Matrix for Vision Transformer (ViT) Model Predicting Pig Contact vs. No Contact.

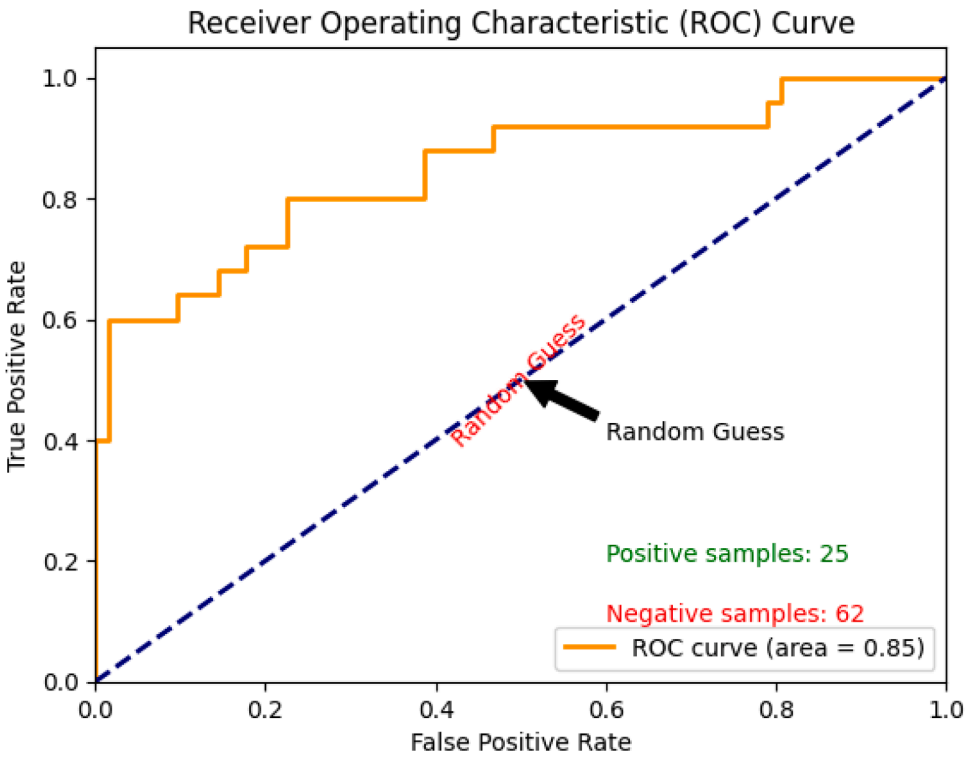

Fig. 4. ROC Curve for Vision Transformer-Based Pig Contact Detection Model (AUC = 0.85).

4.1. Contact detection

Fig. 5 shows how our model accurately detected a head-to-rear contact, while giving us the contact probability.

4.2. Comparison of our work with other architectures

To further prove the effectiveness of the proposed model, it was compared with heavy model (DenseNet121), lightweight (MobileNetv2)

Fig. 5. Sample Output of Vision Transformer-Based Pig Contact Detection with a 84 % Probability Score.

and residual network (ResNet50 using the exact dataset used in this research. Based on the results of the comparison as shown in Table 1 the model outperformed other models with an accuracy score of 82. 8 %, also with a balanced precision rate of 82. 6 % And recall rate of 82. 8 %. In the experimental findings ViT ranked as the best performer when compared to DenseNet121, MobileNetv2, and ResNet50.

5. Discussion

This study developed an advanced automated system capable of detecting contact behaviour in pigs using vision transformer model, it is interesting to contribute to existing knowledge in the application of AI technologies in pig contact behaviour and achieve an accuracy of 82.8 %,F1 score of 82.7 %, and AUC of 85 %with a total dataset of 433 pig images. Low standard deviations for accuracy and F1 score further give evidence on the model's reliability and consistency; these characteristics are very important in their application in precision livestock farming.

The high AUC score of 85 % indicates that the model we have created makes an excellent distinction between contact and no-contact instances, therefore providing accurate monitoring of the interaction among pigs. Provision of such a reliable detection capability would make possible early intervention and thus allow for improvement of animal welfare and farm management practices.

This study reveals the strength of Vision Transformer such that other models would either need a bounding box or masking to detect contact with the use of multi head attention. Vision Transformer can detect images hence it becomes less complex and faster compared to existing models.Unlike CNNs that possess only limited receptiveness, Vision transformers can attend to the whole images at once, which makes it more effective at understanding global context. Also, to improve the performance metrics of this model in detection of pig contact behaviour future study should consider using more dataset. Overall, this study offers different contributions to knowledge in the area of field social interaction:

1. Unlike Alameer et al. [45], who faced significant occlusion problems in their detection methods, we have created an approach that solves the challenge of occlusion, where parts of pigs may be partially blocked by a pig or object. This makes our work more reliable for a real-world farm environment
2. This study is the first study to apply Vision transformer to detection of pig head-to rear contact, this showcase a novel use of AI technology in livestock behaviour
3. High accuracy of 82.8 %, F1-Score of 82.7 % demonstrates the effectiveness and reliability of our model at detecting this behaviour.
4. We developed an efficient method with high accuracy that can detect pig-head contact behaviour without the need for bonding boxes, masking, this makes it easier to implement
5. Our approach enables understanding of global context at once, due to the ability of Vit to attend to all images at once, a key attribute that is missing in CNNs which only possess limited receptiveness.

6. Conclusion

Automated systems in pig farming could help considerably improve the level of early detection of behavioural changes, a critical indicator of health and welfare issues. In this paper, a new approach is proposed for the identification of head-to-rear contact behaviour of pigs with ViT, leaning away from CNN and other traditional machine learning techniques on which traditional methods have been very dependent. In this work,we developed a ViT-based model that has turned out to show high accuracy and consistency in contact behaviour detection in pigs. The model yielded an accuracy of 82.8 %, an F1 score of 82.7 %, and an area under the receiver operator characteristic curve of 85 %, hence showing us a good and reliable performance under different conditions and datasets. These results demonstrate that the Vision Transformer model reliably detects head-to-rear contacts in pigs, hence making a great leap forward in this area. The application of ViT in such a context is totally new, since very limited research has been carried out in this aspect using this sophisticated machine learning architecture for pig behaviour detection. This work has not only proved the power and robustness of ViTs in this area but also opened a way for their application in AgriTech. Also, the high performance and consistency of this model further proves its potential toward effective integration into automated systems for continuous monitoring of pig behaviour in aiding timely identification and management of aggression and other welfare issues. This development underlines the potential for ViTs to drive a revolution in precision livestock farming toward better welfare and farm productivity. This research work, hence, takes another key step toward the practical application of state-of-the-art machine learning techniques in farming practice and manages to demonstrate real benefits gained through the adoption of Vision Transformers for behavioural monitoring in livestock farming. It opens new paths of research and development into precision agriculture with improved capabilities in monitoring animal welfare.

Table 1
Comparison of the performance of our work with other architecture.

Architecture	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	AUC (%)
ViT(Our Work)	82.8	82.7	82.6	82.8	85.2
MobileNetV2	70.1	69.3	47.6	40.0	61.1
DenseNet 121	72.4	67.5	55.6	20.0	56.8
ResNet50	66.7	67.5	43.3	52.0	62.3

7. Future direction

In the future, we intend to apply our technique to varied livestock and larger datasets to test the Vision Transformer model's robustness and generalisation. In addition, we intend to refine the model so that it can precisely count the frequency of encounters, allowing for more detailed monitoring and better animal welfare management.

Ethics statement

Not applicable: This manuscript does not include human or animal research.

CRediT authorship contribution statement

Gbadegesin Taiwo: Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Sunil Vadera:** Writing – review & editing, Supervision. **Ali Alameer:** Writing – review & editing, Supervision, Conceptualization.

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.

Data availability

Data link was referenced in Manuscript

References

[1] Q. Yang, D. Xiao, A review of video-based pig behaviour recognition, Appl. Anim. Behav. Sci. 233 (2020) 105146.
[2] G.A. Taiwo, T.O. Akinwole, O.B. Ogundepo, Statistical Analysis of Stakeholders Perception on Adoption of AI/ML in Sustainable Agricultural Practices in Rural Development, in: XS. Yang, S. Sherratt, N. Dey, A. Joshi (Eds.), Proceedings of Ninth International Congress on Information and Communication Technology,

Springer, Singapore 1003, ICICT 2024 2024. Lecture Notes in Networks and Systems, 2024.
[3] N. Ringgenberg, R. Bergeron, N. Devillers, Validation of accelerometers to automatically record sow postures and stepping behaviour, Appl. Anim. Behav. Sci. 128 (1–4) (2010) 37–44.
[4] M. Martínez-Avilés, E. Fernández-Carrión, J.M. López García-Bones, J.M. Sánchez-Vizcaíno, Early detection of infection in pigs through an online monitoring system, Transbound Emerg Dis 64 (2) (2017) 364–373.
[5] R.J. Thompson, S. Matthews, T. Plötz, I. Kyriazakis, Freedom to lie: how farrowing environment affects sow lying behaviour assessment using inertial sensors, Comp. Electron. Agricul. 157 (2019) 549–557.
[6] R.M.A. Rechie, M. Kassim, N. Ya'acob, R. Mohamad, RFID monitoring system and management on deer husbandry, in: *IOP Conference Series: Earth and Environmental Science* (Vol. 540, No. 1,, IOP Publishing, 2020, July.
[7] J. Maselyne, W. Saeys, B. De Ketelaere, K. Mertens, J. Vangeyte, E.F. Hessel, S. Millet, A. Van Nuffel, Validation of a High Frequency Radio Frequency Identification (HF RFID) system for registering feeding patterns of growing-finishing pigs, Comp. Electron. Agricul. 102 (2014) 10–18.
[8] C. Chen, W. Zhu, T. Norton, Behaviour recognition of pigs and cattle: journey from computer vision to deep learning, Comp. Electron. Agricul. 187 (2021) 106255.
[9] D.A.B. Oliveira, L.G.R. Pereira, T. Bresolin, R.E.P. Ferreira, J.R.R. Dorea, A review of deep learning algorithms for computer vision systems in livestock, Livest Sci 253 (2021) 104700.
[10] A. Alameer, P. Degenaar, K. Nazarpour, Context-based object recognition: Indoor versus outdoor environments, *2*, Springer International Publishing, 2020, pp. 473–490.
[11] A. Alameer, P. Degenaar, K. Nazarpour, Objects and scenes classification with selective use of central and peripheral image content, J. Vis. Commun. Image Represent. 66 (2020) 102698.
[12] A. Alameer, P. Degenaar, K. Nazarpour, Biologically-inspired object recognition system for recognizing natural scene categories, in: *2016 international conference for students on applied engineering (ICSAE)*, IEEE, 2016, October, pp. 129–132.
[13] L.C. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang, H. Adam, Masklab: instance segmentation by refining object detection with semantic and direction features, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4013–4022.
[14] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell 39 (6) (2016) 1137–1149.
[15] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1904–1916.
[16] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A.C. Berg, Ssd: single shot multibox detector, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, Springer International Publishing, 2016, pp. 21–37. *October 11–14, 2016, Proceedings, Part I 14*.
[18] X. Zhou, D. Wang, P. Krähenbühl, Objects As Points, arXiv preprint arXiv: 1904.07850., 2019.
[19] M. Jiang, Y. Rao, J. Zhang, Y. Shen, Automatic behavior recognition of group-housed goats using deep learning, Comp. Electron. Agricul. 177 (2020) 105706.
[20] A. Yang, H. Huang, X. Yang, S. Li, C. Chen, H. Gan, Y. Xue, Automated video analysis of sow nursing behaviour based on fully convolutional network and oriented optical flow, Comp. Electron. Agricul. 167 (2019) 105048.
[21] H. Gan, M. Ou, E. Huang, C. Xu, S. Li, J. Li, K. Liu, Y. Xue, Automated detection and analysis of social behaviors among preweaning piglets using key point-based spatial and temporal features, Comp. Electron. Agricul. 188 (2021) 106357.
[22] A. Alameer, I. Kyriazakis, J. Bacardit, Automated recognition of postures and drinking behaviour for the detection of compromised health in pigs, Sci. Rep. 10 (1) (2020) 13665.
[23] B. T. Psota, E. Schmidt, T. Mote, C. Pérez, Long-term tracking of group-housed livestock using keypoint detection and map estimation for individual animal identification Sensors 20 (13) (2020) 3670.
[24] A. Bhujel, E. Arulmozhi, B.E. Moon, H.T. Kim, Deep-learning-based automatic monitoring of pigs' physico-temporal activities at different greenhouse gas concentrations, Animals 11 (11) (2021) 3089.
[25] R. Gronskyte, L.H. Clemmensen, M.S. Hviid, M. Kulahci, Pig herd monitoring and undesirable tripping and stepping prevention, Comp. Electron. Agricul. 119 (2015) 51–60.
[26] M.A. Kashiha, C. Bahr, S. Ott, C.P. Moons, T.A. Niewold, F. Tuyttens, D. Berckmans, Automatic monitoring of pig locomotion using image analysis, Livest. Sci. 159 (2014) 141–148.
[27] Z. Weixing, Z. Jin, Identification of abnormal gait of pigs based on video analysis, in: 2010 Third International Symposium on Knowledge Acquisition and Modeling, IEEE, 2010, pp. 394–397.
[28] Y. Li, L. Sun, X. Sun, Automatic tracking of pig feeding behavior based on particle filter with multi-feature fusion, Trans. Chinese Soc. Agricul. Eng. 33 (1) (2017) 246–252.
[29] Z. Weixing, W. Zhilei, Detection of porcine respiration based on machine vision, in: 2010 Third International Symposium on Knowledge Acquisition and Modeling, IEEE, 2010, pp. 398–401.
[30] W.X. Zhu, Y.Z. Guo, P.P. Jiao, C.H. Ma, C. Chen, Recognition and drinking behaviour analysis of individual pigs based on machine vision, Livest. Sci. 205 (2017) 129–136.
[31] A. Nasirahmadi, O. Hensel, S.A. Edwards, B. Sturm, A new approach for categorising pig lying behaviour based on a Delaunay triangulation method, Animal 11 (1) (2017) 131–139.
[32] M. Mittek, E.T. Psota, J.D. Carlson, L.C. Pérez, T. Schmidt, B. Mote, Tracking of group-housed pigs using multi-ellipsoid expectation maximisation, IET Comp. Vision 12 (2) (2018) 121–128.
[33] J. Lee, L. Jin, D. Park, Y. Chung, Automatic recognition of aggressive behavior in pigs using a kinect depth sensor, Sensors 16 (5) (2016) 631.
[34] A.L.H. Andriamandroso, F. Lebeau, Y. Beckers, E. Froidmont, I. Dufrasne, B. Heinesch, P. Dumortier, G. Blanchy, Y. Blaise, J. Bindelle, Development of an open-source algorithm based on inertial measurement units (IMU) of a smartphone to detect cattle grass intake and ruminating behaviors, Comp. Electron. Agricul. 139 (2017) 126–137.
[35] A. Rahman, D.V. Smith, B. Little, A.B. Ingham, P.L. Greenwood, G.J. Bishop-Hurley, Cattle behaviour classification from collar, halter, and ear tag sensors, Infor. Proc. Agricul. 5 (1) (2018) 124–133.
[36] Z.E. Barker, J.V. Diosdado, E.A. Codling, N.J. Bell, H.R. Hodges, D.P. Croft, J. R. Amory, Use of novel sensors combining local positioning and acceleration to measure feeding behavior differences associated with lameness in dairy cattle, J. Dairy Sci. 101 (7) (2018) 6310–6321.
[37] M.L. Williams, N. Mac Parthaláin, P. Brewer, W.P.J. James, M.T. Rose, A novel behavioural model of the pasture-based dairy cow from GPS data using data mining and machine learning techniques, J. Dairy Sci. 99 (3) (2016) 2063–2075.
[38] G. Li, Y. Huang, Z. Chen, G.D. Chesser Jr, J.L. Purswell, J. Linhoss, Y. Zhao, Practices and applications of convolutional neural network-based computer vision systems in animal farming: a review, Sensors 21 (4) (2021) 1492.
[39] W. Shen, H. Hu, B. Dai, X. Wei, J. Sun, L. Jiang, Y. Sun, Individual identification of dairy cows based on convolutional neural networks, Multimed. Tools Appl 79 (2020) 14711–14724.
[40] Q. Guo, Y. Sun, C. Orsini, J.E. Bolhuis, J. de Vlieg, P. Bijma, P.H. de With, Enhanced camera-based individual pig detection and tracking for smart pig farms, Comp. Electron. Agricul. 211 (2023) 108009.
[41] M. Wutke, F. Heinrich, P.P. Das, A. Lange, M. Gentz, I. Traulsen, F.K. Warns, A. O. Schmitt, M. Gültas, Detecting animal contacts—A deep learning-based pig detection and tracking approach for the quantification of social contacts, Sensors 21 (22) (2021) 7512.
[42] A. Alameer, S. Buijs, N. O'Connell, L. Dalton, M. Larsen, L. Pedersen, I. Kyriazakis, Automated detection and quantification of contact behaviour in pigs using deep learning, Biosyst. Eng (2022).
[43] K. Mimura, K. Nakamura, K. Yasukawa, E.C. Sibert, J. Ohta, T. Kitazawa, Y. Kato, Applicability of Object Detection to Microfossil research: Implications from Deep Learning Models to Detect Microfossil Fish Teeth and Denticles Using YOLO-v7, Earth and Space Science, 2024.
[44] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, 2021. An Image is Worth 16x16 Words: transformers for Image Recognition at Scale, conference paper at ICLR 2021.
[45] A. Alameer, S. Buijs, N. O'Connell, L. Dalton, M. Larsen, L. Pedersen, I. Kyriazakis, Automated detection and quantification of contact behaviour in pigs using deep learning, Biosyst. Eng. 224 (2022) 118–130.