2026 Volume: 02 Issue: 05 www.ijrpa.com ISSN 2456-9995 Research Article

International Journal Research Publication Analysis

Page: 01-20

MULTIMODAL ANIMAL BEHAVIOR MONITORING USING AUDIO AND VIDEO ANALYSIS

[*1]Ms. Apeksha R Kulkarni, [1]Ms. Deeksha S., [1]Ms. Harshitha M., [2]Mrs. Vindhya Ramachandran

[1]Jyothy Institute of Technology.

[2]Assistant Professor, Department of Computer Science and Engineering Jyothy Institute of Technology.

Article Received: 23 March 2026
Article Revised: 13 April 2026
Published on: 03 May 2026

***Corresponding Author: Ms. Apeksha R Kulkarni**
Jyothy Institute of Technology.

DOI: https://doi-doi.org/101555/ijrpa.2291

ABSTRACT

It is necessary to monitor the animals and livestock for health, welfare and productivity in the contemporary pre-cision farming domain. Traditional visual inspection techniques for monitoring are labor-intensive and inaccurate, thus raising the requirement for intelligent and automated monitoring meth-ods. With the breakthrough of artificial intelligence techniques, sensor-based, vision-based and multi-modal behavior recognition methods have been designed and implemented. Vision-based behavior recognition methods based on deep learning networks, such as convolutional neural networks (CNN) and transformer-based networks (e.g., ViT) have shown excellent performance for recognizing actions and postures of livestock. IoT based sensor systems achieve a real-time monitoring and anomalous detection on the animals with the aid of unsupervised learning methods. Additionally, multi-modal methods combining the audio and visual and sensor data are robust and effective to address some of the drawbacks of single modal based techniques, providing a higher recognition accuracy and a better generalization abil-ity. Meanwhile, emerging technology including vision-language model (VLM), edge computing framework, and lightweight de-tection architectures facilitate intelligent monitoring on resource constrained system for real-time deployment.However, challenges of small data set, varied environmental interference and com-putational intensity exist and further exploration is necessary to achieve robust and accurate intelligent livestock monitoring. In this paper, we will provide an in-depth survey on current animal behavior recognition techniques, focusing on methods, performance and research issues and discuss future opportunities to develop accurate and real-time

intelligent monitoring systems for the smart livestock.

***INDEX TERMS*:** Animal behavior recognition, livestock monitor-ing, deep learning, computer vision, Internet of Things (IoT), multimodal learning, anomaly detection, precision agriculture.

I. INTRODUCTION

The monitoring of livestock and animal behavior plays an indispensable role in animal welfare, early disease diagnosis, and production in current precision agriculture. The traditional ways of monitoring animals require extensive labor and time; they also often come with human errors. With the continuous increase in the number of livestock in modern farms, intelli-gent and automatic monitoring systems are required for real-time behavior detection.

In recent times, advanced AI techniques, such as deep learning have facilitated vision-based methods for the recogni-tion of behaviors in animals. Convolutional Neural Networks and transformer-based models, for example, has been known for the high performance in posture and activity detection. Visual Transformers which are transformer-based models, in particular, performs better than CNN models in animal posture detection [1]. Similarly, activity detection from sensors data can also achieve the state of the art results by utilizing deep learning and transfer learning models, thereby improving the efficiency and reducing training time [10].

Aside vision-based methods, the use of IoT and sensor-based monitoring systems for animals is also trending in current research. These wearable sensors measure multidimen-sional data of animals' activity such as movement, location and temperature to allow for anomaly detection in real-time. Methods such as unsupervised Autoencoders, K-Means and Isolation Forest have shown remarkable performance in anomaly detection, by identifying behavior anomalies without requiring labeled data [3]. The observed behavior abnormali-ties often indicate health problems or unusual environmental stimuli for the animal.

To further improve monitoring efficiency, hybrid sensor-based and vision-based methods has been developed by fusing data from various sensors. Computer vision methods that detect and classify livestock behavior in terms of standing, eating and resting has also been widely investigated using YOLO-based methods [4]. In particular, research focuses on developing lightweight models for on-line detection of animal behavior anomalies in limited computational resource environ-ments [12]. Further enhancement of monitoring by enabling

on-line processing on edge devices such as Raspberry Pi by eliminating the reliance on cloud computation has also been reported [7].

While some research utilizes a single modality data for be-havior analysis, multimodal methods by fusing various sensors data such as vision and audio can perform better in detecting behavior and emotion of animals [5], [6]. These models use various architectures such as CNN-LSTM and CNN-Vision methods to accurately detect behaviors and emotions in ani-mals. AnimalFormer utilizes detection, segmentation and pose estimation approaches for rich animal behavior feature extrac-tion from videos [2]. Vision-language models, more recently, have been explored for better understanding animal behavior, combining image and text features for better performance and interpretation of animal behaviors [11].

For non-livestock animal behavior, new advanced models, such as species-specific activity detection by leveraging deep learning architectures such as transformer models and multi-modal semantic query network models have been developed and the accuracy has been shown to be more efficient than previous models in recognizing different behaviors in animals [8]. These studies emphasize the necessity of creating special architectures tailored to different types of behaviors exhibited by different animals.

Although, state-of-the-art works have been achieved, some challenges remain, such as insufficient available of annotated data, environmental issues (lighting, occlusion) and the need of real-time processing. This survey is about the works that address the challenges above by classifying into sensor-based, vision-based and multimodal approach. We also will discuss about the challenge and the future research directions in the livestock monitoring field.

II. Components of Animal Behavior Monitoring Systems

Animal behavior monitoring systems are comprised of vari-ous interrelated components that work collectively to acquire, process and analyze behavior data. The following are a few data acquisition, feature extraction, behavior analysis and decision-making components.

A. Data Acquisition

The first step of animal behavior monitoring system involves data acquisition. Various sensors and devices are used to collect raw animal data. Image and video data is captured by vision-based systems to analyze animal activity and posture. IoT devices, such as accelerometers, GPS, and temperature sensors, are used by sensor-based systems to record physio-logical and movement data of animals. In an attempt to pro-vide a detailed understanding of animal

behavior, multimodal systems integrate data collected from different modalities like audio and wearable sensors along with video and text data. These systems continuously monitor animal behavior without any intrusion.

B. Feature Extraction

The extracted features are processed to obtain meaningful representation that may be used for analyzing animal behavior. In vision-based systems, images are analyzed by deep learning models like Convolutional Neural Networks (CNNs) and Vi-sion Transformers to extract spatial information. Other models like YOLO are employed to detect objects (animals) in the scene and track them for subsequent actions.

In sensor-based systems, time series data is analyzed using both temporal and statistical features. Movement patterns, activity levels and changes in the physical status of the animals that are crucial for the recognition of behaviors can be observed from these extracted features.

C. Behavior Analysis and Classification

Based on the extracted features, this component identifies and classifies the behavior exhibited by the animals. Both machine learning and deep learning based models like CNNs, Long Short-Term Memory (LSTM) networks, and Recurrent Neural Networks (RNNs) have been widely employed for behavior classification. Complex interactions among various data modalities are captured by Transformer-based models and multimodal architectures, which further improve the classi-fication performance. Animals' eating, resting, walking, and abnormally expressed behaviors are now capable of being distinguished from others using this technology.

D. Anomaly Detection

Animal behavior anomaly detection is the crucial task for raising early alarms about health-related issues of animals and unusual physiological changes. Unsupervised learning approaches, such as Autoencoders, K-Means clustering, and Isolation Forests, are predominantly applied for this purpose by detecting unusual patterns and changes with respect to standard animal behavior. Such systems will promote better animal welfare, decrease economic loss in livestock farming.

E. Decision Support and Visualization

The last component aims at translating the analyzed results to farmers or relevant stake

holders through a readable and accessible interface. Typical ways of information delivery for animal behavior analysis include a web interface and dash-boards for visualizing animal behaviors, trends, and anomaly alerts. They typically are based on the technologies of the edge and cloud platform for continuous real-time monitoring. This will help the farmers make reasonable decisions regarding animal health, food, and management.

III. Advantages and Disadvantages of Animal Behavior Monitoring Systems

Animal behavior monitoring systems are beneficial to Live-stock management, welfare and productivity, but the existing system face some disadvantages which should be address to be efficiently used in real-world scenarios.

A. Advantages

1. **Early disease identification:** The automated systems detect abnormalities in the animal behavior from its early stage so the animal is provided with early treatment, which helps to overcome from severe diseases.
2. **Less human interference:** In the above systems we don't have to watch them constantly thus saving from labors cost and human errors.
3. **Real-time monitoring:** The systems are enabled to monitor live feed with high speed, due to the usage of IOT devices and edge computing techniques, they alarm about every abnormality occurring during that live feed in the animal.
4. **Accuracy:** Deep learning models like CNN, transformer, and multimodal network present high accuracy for recognizing posture and behavior in animal which is one of the primary requirements.
5. **Non-invasive monitoring:** Vision and multimodal based models are non-invasive means no direct interaction with animals is required, which prevents stress among animals and allow observation of the behavior in their natural state.
6. **Scalability:** The systems are useful in larger farms and can monitor large number of animals at a time with the usage of a number of cameras.
7. **data driven decisions:** Integration of artificial intelli-gence with analytics give insight information of animal be-havior and help the farmer take informed decisions regarding feeding, treatment, etc.

B. Disadvantages

1. **high cost:** Deployment cost of sensors, cameras and computing infrastructure required are quite high, especially for smaller farms.
2. **Lack of labeled data:** To implement Deep learning models, we require large quantity of data; but for animal behavior there are limitations in getting it.
3. **Environmental factors:** Lighting conditions, occlusion, weather conditions and the complexity of backgrounds affect the vision based models for animal behavior identification.
4. **computational limitations:** models like transformer, multimodal networks need more computational powers which can hinder its real-time deployment and also be expensive.
5. **Sensor noise and unreliable data:** wearable sensors can lead to noises or fragmented data because of their design limitations or the interference from the external environment.
6. **generalization problem:** Model trained on one dataset will not perform well on another with a different species, breed or farming environments.
7. **privacy and ethical concerns:** continuous monitoring and storing of animals' data raises security and privacy issues and questions regarding ethical use of surveillance data.

IV. PHASES OF ANIMAL BEHAVIOR MONITORING SYSTEMS

Animal behavior monitoring systems are structured as a series of stages beginning from data collection through de-cision making. Each stage is an integral step in the process of transforming raw data into behavioral understanding.

A. Data acquisition Phase

The first stage is to collect raw data from various data sources including cameras, microphones and IoT based wear-able devices. In vision-based systems, cameras acquire image and video while sensors provide movement and environmental data. Multi-modal data acquisition schemes with two or more data source types may be used to enhance the accuracy of data.

B. Pre-processing Phase

In the second stage, raw data collected from various sources is refined and cleaned for subsequent processing. Image and video data will undergo noise removal, resizing and image normalization. Data from sensors may also need to be filtered for noise and errors, such as removal of invalid data entries.

C. Feature extraction Phase

The third stage of the system, Feature extraction, is the extraction of relevant patterns from the data pre-processed above. In vision-based system Deep Neural networks like CNNs, Vision Transformers can be utilized to extract features such as movement, posture and body pose. From sensor based systems, a combination of temporal features such as mean and variation of values in a time series, and statistical features may be extracted.

D. Behavior recognition Phase

The fourth stage, Behavior recognition, aims at classi-fying animals' activities such as feeding, resting, walking and aberrant actions, through Machine learning models or Deep learning based systems such as CNNs, LSTMs and transformer-based models. Similar to the above, multi-modal models with multiple data source types can contribute to the accuracy of behavior classification.

E. Anomaly detection Phase

In this stage of anomaly detection, behavior which might indicate illness or distress is identified. Anomalous behaviors such as absence of normal actions such as feeding or the presence of erratic behaviors are detected. This may be done through supervised or unsupervised methods such as autoen-coders, k-means and isolation forests, applied to deviation of animals from its learned behavior patterns. These are particularly important for smart livestock farming systems that can act as a precursor to early warnings.

F. Decision-making and Alert Phase

The last stage involves the extraction of actionable insights based on the animal's behavior pattern, as illustrated in the figure. Based on abnormal behavior, an alert may be generated and sent to the farmers and concerned persons, and to facilitate a prompt action plan. Data obtained are presented using visualization dashboards and are streamed to cloud based platforms in order to assist informed decision-making in smart agriculture.

V. LITERATURE SURVEY

Over the past few years, several animal behavior recognition methods including vision-based, sensor-based, and multimodal approaches have been designed and developed. These different methods are expected to increase the accuracy of livestock monitoring, and facilitate early anomaly detection, and smart decision-making in animal management systems.

A. Vision-Based Approaches

In vision-based systems, cameras and deep learning tech-niques are used to conduct an accurate analysis of animal posture and behavior without any contact. Comparative studies such as [1] conducted between the Transformer based architec-tures and the CNN based models show the better performance of the Vision Transformers over CNN based architectures for the posture detection task. Also, fine tuning of a ResNet based model using transfer learning drastically increased the activity recognition accuracy and at the same time reduced the training effort [10]. Also in the case of behavior recognition,object detection algorithms such as YOLO are extensively used. In [4], the application of a YOLOv8 based system in livestock management with IoT infrastructure is shown and high accu-racy for cattle behavior analysis such as feeding and resting are demonstrated. There also has been efforts in designing lightweight detection framework to perform real time abnor-mal behavior detection such as YOLO-PetX [12]. Vision based AI systems have also been adopted in precision agriculture systems for on-line monitoring of livestock behavior and their health as they provide both visual and biological data in an integrated way [9]. These systems are, however, susceptible to changes in illumination, the presence of obstructions and complex backgrounds.

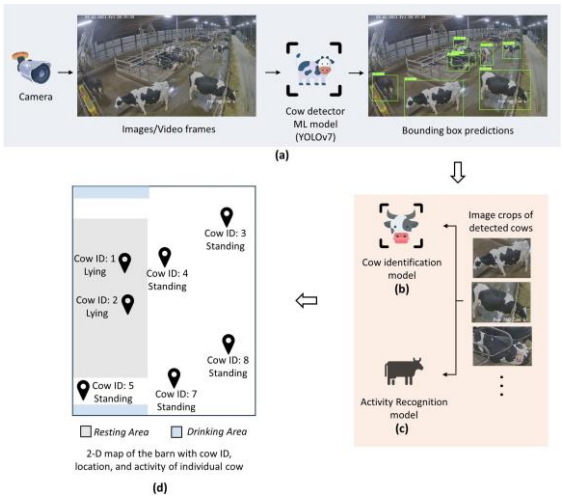


Fig. 1. Vision-based animal behavior recognition using deep learning and object detection.

B. Sensor-Based Approaches

In sensor-based approaches, portable sensing devices and IoT tools are used to extract data about physiology and motion.

In [3], anomaly detection approach has been designed using an IoT sensors and various unsupervised learning algorithms such as Isolation Forests and Autoencoders to automatically recognize abnormalities in the livestock behavior and behavior anomaly. Such a method can recognize disease and stress at an early stage as it does not require any labeled data. Deep learning-based system such as activity recognition system utilizing sensor data is also able to achieve a good level of accuracy using transfer learning so that the large training process on many labeled datasets is reduced [10]. Sensor-based methods can work well for a continuous monitor but they are known to contain noise, and limited contextualization power.

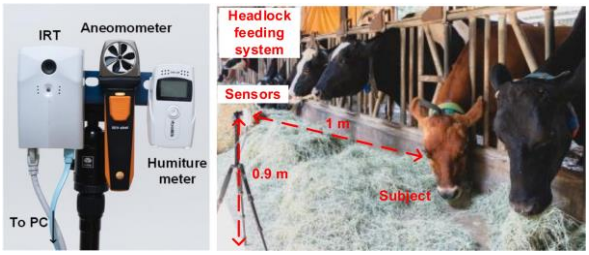


Fig. 2. Sensor-based livestock monitoring using IoT and wearable devices.

C. Multimodal and Hybrid Approaches

To deal with the limitations of single-modality systems, extensive researches on multimodal methods which utilize combined features from audio, video, and sensor data have been done in this area. Such approaches can incorporate addi-tional and complementary information from different sources and further improve the reliability and accuracy of the behavior recognition system. In [5], the framework adopted a combina-tion of CNN-LSTM to process the audio signal and ResNet for extracting the visual features to recognize dog emotions and achieve better performance. A fusion based model combined with camera images and wearable sensors data was proposed in [6] and performed better than single-modality systems for motion-context interaction captured from these sensors.

State-of-the-art multimodal frameworks could perform be-havior analysis with more sophisticated multiple stages pro-cessing. For example, the AnimalFormer [2] integrated with object detection, segmentation, and pose estimation models which enabled the extraction of fine-grained animal behav-ioral features and provides detailed information about pos-ture, movement, and interactions. On the other hand, the edge computing based systems [7] are designed for real-time monitoring by doing data processing at the edge devices (e.g. Raspberry Pi), minimizing the dependency on network transmission and enabling the use in

remote and resource-limited farm environments.

More recent transformer based multimodal architectures were also investigated, such as [8] proposed a dual-phase semantic query network for species-specific activity recogni-tion and demonstrated superior performance across a variety of animal behaviors. Moreover, vision-language models [11] can be utilized for behavior analysis through combining visual features with language representations, making the extracted behaviors more interpretable and semantically enriched. The developed models can not only contribute to improved classi-fication accuracy, but also generated rich descriptive informa-tion, which is crucial for an intelligent monitoring system.

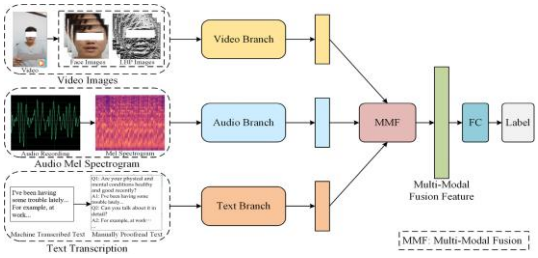


Fig. 3. Multimodal system integrating audio, video, and sensor data for behavior recognition.

D. Summary

It can be observed from the literature that the vision based approaches allow to conduct the monitor without any inter-action and sensor-based system helps in detecting anomaly in real-time whereas multimodal system provide accuracy. However, data scarcity, complexity of background and variety of condition remain some issues which must be addressed before wide scale implementation.

TABLE I COMPARISON OF EXISTING APPROACHES.

Ref	Model	Data	Type	Contribution	Acc
[1]	ViT vs CNN	Img	Vision	ViT better	92%
[2]	AnimalFormer	Vid	Multi	Pose+Detect	High
[3]	Autoencoder	Sensor	IoT	Anomaly detect	High
[4]	YOLOv8	Vid	Hybrid	Real-time	88%
[5]	CNN+LSTM	A+I	Multi	Emotion detect	High
[6]	YOLO+Sensor	V+S	Multi	Fusion model	High
[7]	Edge System	Vid	Vision	Low latency	Efficient
[8]	MSQNet	Vid	Multi	Activity recog	72%
[9]	Vision AI	V+S	Hybrid	Smart farming	-
[10]	ResNet TL	Sensor	IoT	High accuracy	98%
[11]	BLIP	Img+Text	Multi	Caption+Classify	92%
[12]	YOLO-PetX	Img	Vision	Lightweight	72%

VI. General Applications

Smart farming and intelligent monitoring solutions employ animal behavior monitoring systems with a blend of artificial intelligence, IoT, and deep learning to various fields.

A. Smart livestock farming

Automated monitoring systems are developed for animal behavior such as feeding behavior, motion behavior, resting behavior for production, optimum resource usage and effective farm management in precision farming.

B. Health monitoring and disease detection

Behavior analysis techniques is widely applied to monitor and early diagnose disease or abnormal animal status by changes of motion, posture or activity. The early diagnosis results in timely intervention of disease, reduction of the animal loss rate.

C. Animal welfare and stress monitoring

The welfare status of animals is checked by detecting stress, abnormal behavior or abnormal environmental conditions us-ing multimodal monitoring systems with audio analysis and video analysis. Emotional state and behavioral changes were detected by analysis.

D. Surveillance and farm security

The systems for continuous monitoring and surveillance of livestock and animal shelter facilities; computer vision technology can recognize unusual activities, trespassing and abnormal behavioral patterns to prevent damages or loss.

E. Wildlife monitoring and research

These systems are utilized for research on wildlife behavior, with behavior recognition used for the classification of wildlife motion, behavior and activity for conservation research.

F. Smart pet monitoring

Intelligent pet monitoring and care systems monitor pets' behavior, detect unusual behavior to judge pets' health state, and make alerts to the owners in time.

G. Real-time monitoring using edge computing

Edge computing for behavior analysis using devices such as Raspberry Pi can reduce the latency of analysis, thus suitable for the animal monitoring system in the isolated farm.

VII. System Architecture

In general, Animal behavior monitoring systems are multi-layered architecture, integrated data acquisition, processing, analysis, and decision making. As designs of the architec-tures vary by study, most of them comprise vision system, sensor, and intelligent model: The

data acquisition layer represents the system's layer that collect raw data, the main used sensors consist of cameras, microphones, and IoT-based wearable sensors. Vision-based systems rely on a continual stream video capture for the determination of animal poses and activities [1] [4], while sensor-based system focuses on the detection of some physiology or motion related data, such as temperature, acceleration and locations [3]. Besides traditional system designs, the novel systems integrated the sensors on an IoT-based platform, aiming at real time monitoring of an-imal behavior. :contentReference[oaicite:0]index=0 The **data transmission and storage layer** conveys the detected data using wireless transmission methods, and store them on cloud or edge-based server. Recent architectures use edge devices (Raspberry Pi, for example) to perform the data processing so as to decrease dependence on cloud servers, which can achieve low latency and high efficiency data management.

The system using the combined advantage of edge and cloud platforms can enable fast data transmission and real time analytics in the environment of smart agriculture. :contentRef-erence[oaicite:1]index=1 The **processing and feature extrac-tion layer** is used to convert the raw data into understandable features. Computer vision techniques are mainly applied on images for pose and object detection: it could be CNNs, YOLO or Vision Transformers. :contentReference[oaicite:2]index=2 For sensor-based systems, statistical and temporal features were extracted. However, some multimodal systems combined features from different sources to increase the robust and precision of the architecture [5] [6]. The **behavior analysis and classification layer** involves analyzing animal behaviors and classifying the behaviors of animal into different cate-gories using some machine learning and deep learning models. The prevalent algorithms are CNNs, LSTMs or transformer-based methods [1] [10]. New and complex models such as AnimalFormer have been developed which can implement detection, segmentation and pose estimation simultaneously.
:contentReference[oaicite:2]index=2 The final layer is the **de-cision making and alert system**, which output and visualized results to farmers and trigger alarms for abnormal activities, and also could offer feedback to users for better livestock care. This kind of system aims at scalability, real time processing capability and data fusion using the combination of sensors. It can be widely used in precision livestock farming.

VIII. METHODOLOGY

Different methods have been used to detect and classify animals behaviors based on different

data types and models. Based on literature review we could divide them into vision-based, sensor-based and multimodal based methods.

A. Vision-Based Methodologies

Vision-based methodologies utilize images and video as input for information extraction of the animals' behaviors. Common Deep Learning models CNN, YOLO and Vision Transformers are used for the extraction of the necessary features, classification and detection of the animals behaviors [1, 4, 12]. For instance, in [1], a comparison between two CNN models and a transformer-based model was carried out for animal behavior posture classification, and the transformer-based model presented better results than both CNN based models. YOLO-based detection system is often utilized in real-time detection of animal actions [4, 12]. In spite of being non-invasive approaches, vision-based methods are highly sensitive to lighting conditions and occlusions.

B. Sensor-Based Methodologies

Sensor-based methodologies incorporate a wide variety of sensors like IoT enabled devices, wearable devices, etc to detect information regarding animals behavior, physiologi-cal data etc. Unsupervised learning based methods such as Autoencoders, K-Means clustering and Isolation Forests are incorporated for anomaly detection, which does not rely on a labeled data set for classification of animals' behaviors [3].

Unsupervised models allow for non-invasive detection and tracking of animals behavior without a lot of information of the task. Deep neural networks based models can also be used for animals classification using transfer learning, especially when the number of animal species are high [10]. In contrast to vision-based approaches, sensor-based methods don't offer a wide range of context.

C. Multimodal and Hybrid Methodologies

Multi-modal approaches combines features extracted from various data sources such as audio, video, sensor data etc to improve the accuracy of the animal behavior detection and classification task [5, 6]. For example, a CNN-LSTM model is presented in [5] that utilizes both audio and video features to detect the emotions in dogs, and a hybrid system utilizing cameras and wearable sensors has been developed for higher accuracy animal behavior classification in [6]. More advanced multi-modal methods like AnimalFormer, that extracts pose, performs body segmentation, and detects the presence of surrounding objects related to animal behavior [2]. The per-formance of deep learning models can be enhanced for animal behavior classification using transformer-based networks such as MSQNet, which can also detect animals in a variety of fields with a very high accuracy [8]. Other than that, vision-

language models are also being used for the detection of ani-mal behaviors in order to achieve more intuitive interpretation of behavior by combining vision-language modalities [11]. An edge-based method has also been developed in order to achieve real-time detection of animal behavior in an embedded environment like Raspberry Pi [7].

D. Comparative Analysis

The following conclusions are drawn by comparing the literature presented:

- Vision-based methods can gather the most spatial and positional data of the animals' activities, however are not always very feasible to implement, and are influenced by the light conditions.
- Sensor-based methods are good for constant animal be-havior tracking at any location or time of day and are generally non-invasive but lack information regarding the contextual analysis of the situations.
- Multimodal methods are capable of generating better and more accurate results for detection and classification of animal behavior because of the combined analysis of more data sources.
- Both transformer-based and hybrid models result in more efficient results in comparison to traditional methods.

Thus we can consider multimodal and hybrid methods to be more prospective for the future.

IX. Results and Analysis

This section provides a comparison between the reported re-sults in the literature for different animal behavior monitoring systems. Specifically, it compares the performance, robustness and applicability of different methods.

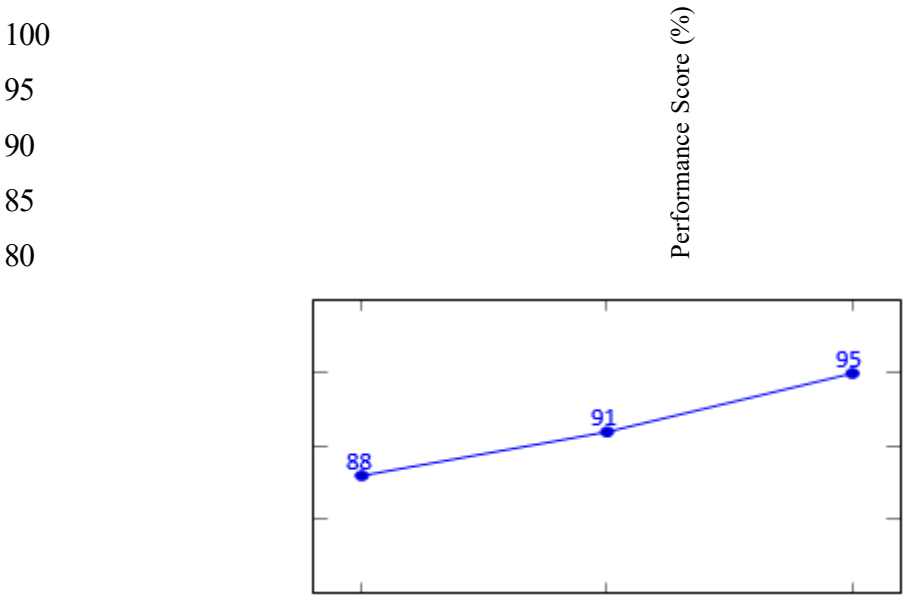
A. Performance of Vision-Based Models

Vision-based approaches demonstrate strong performance for tasks of posture and activity recognition. As mentioned in [1], transformer-based models such as Vision Transformers outperform CNN-based models like EfficientNet and Con-vNeXt for accurate prediction, reaching 92.31However, vision-based models with high accuracy are sensitive to lighting conditions, occlusion and background complexities and thus can limit the performance in real world scenarios.

B. Performance of Sensor-Based Systems

Sensor-based systems are proved to be effective for long-time monitoring at real time. In

[3], unsupervised learning



Sensor Vision Multimodal Approach

methods such as Autoencoders and Isolation Forests can detect anomalous behaviors with high accuracy and low false positive rate. Besides, the transfer learning applied deep neural networks on sensor data achieved detection accuracy as high as 98.40

C. Performance of Multimodal and Hybrid Systems

Multimodal systems combining information from sound, video and sensor data achieved even better performance than single-modality ones. For example, CNN-LSTM model for audio and ResNet model for video were used to provide complementary information, which significantly increased the accuracy of emotion recognition [5]. Similarly, fused camera and wearable sensor data reached a higher accuracy for behavior classification than either alone [6]. Besides that, more advanced model like AnimalFormer can provide detailed infor-mation on animal behavior by using detection, segmentation and pose estimation [2]. Vision-language model has also been used to analyze animal behavior which achieve 92.4

D. Comparative Analysis

Overall comparison shows that vision-based systems can achieve a high accuracy and obtain detailed information from spatial patterns but are subject to environments. Sensor-based systems can detect anomaly with high efficiency and stable performance but are limited in context. Multimodal system can increase the robustness and accuracy. And hybrid/transformer-based model is more capable in complex behavior recognition. In

summary, the results show that multimodal or hybrid system would be a better choice to develop animal behavior monitoring tools as they incorporate the benefit of both vision-based and sensor-based system and avoid their weaknesses.

TABLE II PERFORMANCE COMPARISON OF APPROACHES.

Approach	Accuracy Range	Key Strength
Vision-based	84% – 92%	High spatial detail
Sensor-based	Up to 98%	Real-time anomaly detection
Multimodal	90%+	Robust and accurate

Fig. 4. Performance comparison showing superiority of multimodal ap-proaches

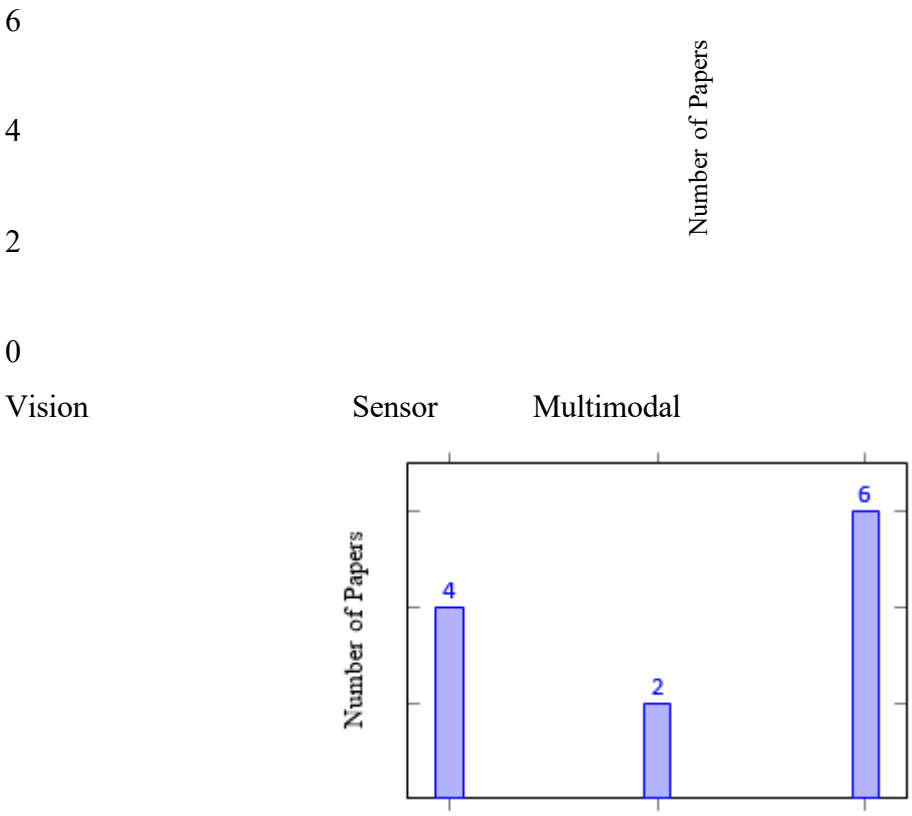


Fig. 5. Distribution of approaches used in surveyed papers.

X. LIMITATIONS OF EXISTING SYSTEMS AND PROPOSED IMPROVEMENTS

Even though a lot of effort has been put in to design a system that monitors the behavior of the animal, we have noticed some issues in most of the present systems which makes it difficult to implement such systems in practical use. The issues mentioned above can be addressed by,

A. Limitations of Existing Systems

1. **Single Modality based System:** In general, all present systems only consider a single modality such as visual data, or sensor data while analyzing the behavior of an animal. Visual based systems are affected by several environmental changes like illumination changes, occlusions, and surroundings vari-ations [1], [4] whereas sensor based systems cannot consider the context or visual information about an animal's behavior [3], [10].
2. **Insufficient availability of datasets:** Deep learning based models require huge amounts of labelled data to classify an animal's behavior, however the labelled data required for training the models in livestock behavior is not sufficiently available, so the models are not generalized well for different animals and conditions.
3. **Environmental and Practical challenges:** Environmen-tal conditions are an obvious concern during the practical im-plementation of the system. Variations in lighting, background noise, camera angle, weather conditions and so on affect the model performance.
4. **Computationally Expensive:** Most state-of-art models like transformers and multimodal networks require massive amounts of computational power for training and running these models in real-time [8], [11]. These models are generally not portable for the device with limited computational resources.
5. **Sensor noise and inconsistencies in data:** For IoT systems the sensors may contain noise and missing values which may result in failure to predict the behavior of the animals accurately [3].
6. **No real-time and scalable system:** Many of the previous systems were designed for experimental setup and cannot be generalized to large scale farms, as they were not developed with real-time in mind.
7. **Lack of Interpretability:** The deep learning based models are black-box models, and it is difficult for the farmers to explain why a particular behavior has been detected by the model.

B. Improvements in Our System

Our system aims at providing an efficient and an effective alternative to existing systems by considering the following features:

1. **Multi modal approach:** Our proposed system takes the audio and visual modality of the animal's behavior and fuses them. This ensures better behavior detection due to multiple factors. Not only does this model take into account the visual input of an animal, but it

also uses the sound input made by the animal to further classify its behavior.

2. **Real-time monitoring:** Our system monitors the animals real-time and provides immediate alert for abnormal behaviors enabling intervention to stop extreme harm to the animals.
3. **Better robustness:** By considering both audio and visual inputs the model is more robust to changes in lighting condi-tions and occlusion. The system is also robust to background noise in terms of accuracy.
4. **Lightweight and computational efficiency:** Optimized deep learning models are used for faster inference making it suitable for use in embedded systems with low computational resources. It can easily be deployed at edge devices.
5. **Advanced anomaly detection:** The system will fur-ther consider advanced algorithms to detect subtle behavioral changes as early indicators of stress, illness or other abnor-malities.
6. **Scalability to large farms:** The system should not only monitor a single animal, but the proposed system must monitor multiple animals at large scale farms effectively.
7. **User friendly interface:** A simple visual representation, using dashboards and alerts should be provided to help the farmers understand the different detected behaviors.
8. All in all, our system attempts to build an animal behavior monitoring system considering various issues from the previ- ous systems and thereby making the system a better choice in the long run.

XI. CONCLUSION

In this paper, a comprehensive review of the existing animal behavior monitoring systems that are based on vision, sensors and multimodal approaches were summarized. In addition, the development of intelligent animal behavior monitoring in precision farming driven by artificial intelligence, deep learning, and IoT were introduced.

It showed that vision-based method is an non-invasive approach that obtains comprehensive information about spa-tial characteristics of behaviors; sensor-based method obtains continuous information and detects abnormal behaviors in real-time. However, individual vision-based or sensor-based approach is susceptible to environmental changes and does not obtain sufficient contextual information. A multimodal system by fusion with multiple types of sensors, such as audio and video sensors combined with other types, can provide better accuracy and robustness. In the future, multimodal approach is more suitable for building robust animal behavior monitoring systems.

In the paper, various deep learning models such as CNN, transformer and combined approaches were investigated and summarized how they are utilized to conduct the task of behav-ior recognition. Compared with individual models, multimodal approaches with transformer achieve significantly higher per-formance in real and complex scenarios.

Despite the great development of the technology, there are still challenging issues such as the shortage of data set, com-plexity of calculation and limitation of real time deploying. These problems must be solved to develop the intelligent animal behavior monitoring system into real application.

As the next generation system, multimodal animal behavior monitoring system has greater prospects to improve the ac-curacy and reliability. The research directions of future work should focus on the efficiency, robustness and scalability of developing light-weight real time models that suitable for actual farming environment.

REFERENCES

1. Kwatra and A. Chug, "Analyzing the Performance of Transformer-Based and CNN-Based Models for Animal Posture Recognition," *IEEE*, 2025.
2. Q. Taha Razzaq and A. Iqbal, "AnimalFormer: Multimodal Vision Framework for Behavior-based Precision Livestock Farming," Tibbling Technologies, 2024.
3. S. T., N. Manaswini, M. Keerthana, P. Vedasri, N. Srija, and V. Supraja, "Anomaly Detection in Livestock Behavior via IoT Sensors and Unsu-pervised Learning," in *Proc. Int. Conf. Emerging Technologies and Fu-ture Innovations (ETFI)*, 2026, doi:10.1109/ETFI68128.2026.11484090.
4. M. Farhan, G. S. W. Thaha, and K. Mutijarsa, "Cattle Anomaly Behavior Detection System Based on IoT and Computer Vision in Precision Livestock Farming," Institut Teknologi Bandung, Indonesia, 2024.
5. W. N. B. Perera and S. P. K. Arachchi, "Deep Learning Multimodal Approach for Dog Emotion Detection in Images, Audios and Videos," in *Proc. Int. Conf. Advances in Technology and Computing (ICATC)*, 2024, doi:10.1109/ICATC64549.2024.11025337.
6. Kim and N. Moon, "Dog Behavior Recognition Based on Multimodal Data from a Camera and Wearable Device," Hoseo University, Korea, 2024.
7. Wu, L. Zhao, H. Ma, C. Feng, Y. Jin, and Z. Sun, "Dual-Model Livestock Visual Warning System," Hangzhou Dianzi University, China, 2024.
8. Yu, J. Varghese, F. Demirkiran, P. Buonaiuto, X. Li, and M.-Chang, "Dual-Phase MSQNet for Species-Specific Animal Activity Recognition," in *Proc. IEEE Int. Conf.*

Multimedia and Expo Workshops (ICMEW), 2024, doi:10.1109/ICMEW63481.2024.10645377.

9. Shukla, K. Gowri, and G. Sudhamsu, "Fostering Smart Agriculture: Using Vision-Based AI for Livestock Managing," Maharishi University of Information Technology and Presidency University, 2024.
10. S. Mekruksavanich, P. Jantawong, and A. Jitpattanakul, "ResNet-based Deep Neural Network using Transfer Learning for Animal Activity Recognition," in *Proc. 6th Int. Conf. Information Technology (InCIT)*, 2022, doi:10.1109/InCIT56086.2022.10067405.
11. M. Ahmad, W. Zhang, M. Smith, B. Brilot, and M. Bell, "Toward Multimodal AI for Livestock: Vision-Language Modelling for Non-Invasive Cattle Behaviour Analysis in Smart Barns," Hartpury University and University of the West of England, 2024.
12. Xu, X. Liao, X. Cheng, and B. Chao, "YOLO-PetX: Enhanced YOLO-Based Recognition of Abnormal Dog Behaviors in Intelligent Pet Care Applications," *IEEE*, 2025.