Avian Activity Classification Using Recurrent Networks to Fuse Videos with Metadata on Imbalanced Datasets

Xijun Wang
Dept. of Computer Science, Northwestern University,
Evanston, IL, USA

Adam Szymanski
Argonne National Laboratory,
Lemont, IL, USA

Yuki Hamada
Argonne National Laboratory,
Lemont, IL, USA

Aggelos K. Katsaggelos
Dept. of Electrical and Computer Engineering,
Northwestern University,
Evanston, IL, USA

ABSTRACT

Activity classification plays a crucial role in various real-life scenarios involving both humans and animals. There is an increasing need for precise activity classification focused on avian-solar interactions, as the usage of solar energy facilities, such as photovoltaic array power stations, has been observed to impact bird species richness, behavior, and activity. However, there has been no work to develop an automated system to monitor and classify these avian-solar interactions. All current methods rely on human observers, which is time and human resources costly and subject to errors related to searcher efficiency. With the recent success of Deep Learning models in activity classification problems, this paper develops a recurrent neural network-based model to automatically classify six avian activities around solar energy facilities. Our proposed model integrates critical feature engineering metadata with video frame data, enabling improved learning and more accurate activity classification. Furthermore, we address the challenge of data imbalance during training and demonstrate the efficacy of our model in detecting and classifying different activities within video tracks. Additionally, we analyze the saliency/backpropagation map of the trained proposed model and validate its decision-making rationale.

CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**.

KEYWORDS

Activity classification, recurrent networks, bidirectional LSTM fusion model, data imbalance, saliency map analysis

ACM Reference Format:
Xijun Wang, Adam Szymanski, Yuki Hamada, and Aggelos K. Katsaggelos. 2024. Avian Activity Classification Using Recurrent Networks to Fuse Videos with Metadata on Imbalanced Datasets. In *2024 10th International Conference on Computing and Artificial Intelligence (ICCAI 2024), April 26–29, 2024, Bali Island, Indonesia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3669754.3669792

ICCAI 2024, April 26–29, 2024, Bali Island, Indonesia
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1705-5/24/04
https://doi.org/10.1145/3669754.3669792

1 INTRODUCTION

With the success of deep learning (DL) methods for various tasks in the computer vision field, using them for video-based activity classification has been a promising application in recent years, both for human and animal activities [3, 4, 6, 9, 12, 15, 22, 24]. Among them, only a small portion is for animal activity classification. It is a more challenging task due to the unique challenges such as unpredictable behaviors exhibited by different animals and the scarcity of available datasets. Many state-of-the-art video-based animal activity classification DL models proposed in the literature use a dual-stream-based model, i.e., they use RGB raw frames and their corresponding optical flow as the two input sources. For example, [5] and [7] used a convolutional neural network (CNN) model to extract deep features from the RGB raw video frames and their corresponding optical flow, and then fuse these two streams in the later part of the model to predict cattle or mouse behaviors. Schindler et al. [16] followed a similar strategy to classify a broader range of wild animal activities. Considering the nature of video data, it is reasonable to adopt recurrent neural networks (RNNs), as in [10]. After fusing the two streams, the authors utilized a Long Short-Term Memory (LSTM) classification network to predict the salmon's feeding activity. In this paper, we focus on the video-based animal activity classification task, particularly avian/bird activity. We propose a DL-based method to solve this task, in particular, we design a Bidirectional LSTM (Bi-LSTM) -based model to classify avian/bird activities.

It is well known that solar energy is an important green energy source, and solar panels have been placed in numerous open areas. However, these solar energy facilities can impact bird communities from different aspects, e.g., bird species richness, migrations, and activities [11, 20, 23]. Therefore, it is crucial to gain a precise understanding of the interactions between birds and solar energy infrastructure to ensure the continued deployment of utility-scale solar energy facilities. Currently, monitoring avian-solar interactions and classifying avian activities rely on human observers. However, training and repeated deployment of human observers are costly and time-consuming, and subject to errors associated with search efficiency. It also generates potential safety concerns for observers working outdoors. Therefore, our objective in this study is to develop a DL model capable of automatically classifying the interaction activities between birds and solar panels.

Few works have used DL methods to study bird behavior via video data in the current literature. The existing DL-based bird

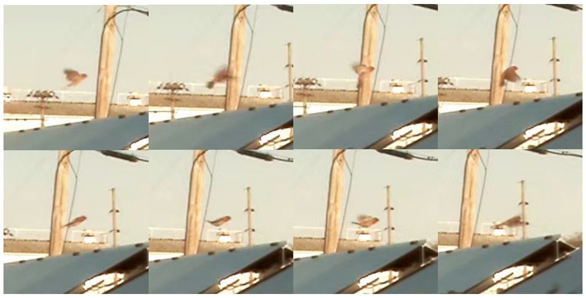


Figure 1: Video track example.

activity classification works mainly utilize acoustic signals [2, 8, 21]. Since March 2020, researchers at Argonne National Laboratory (Argonne) have collected over 6,000 hours of daytime video at five operational solar energy facilities using high-definition true-color video cameras. By processing these videos, Argonne has generated and labeled bird-centered video tracks of avian activities around the solar energy facilities, such as fly over above, fly-through, and perching. In this paper, we train our proposed model based on this avian activity video dataset. Several challenges need to be overcome along the way, such as the heavy imbalance between activities categories, which is also a common issue in many activity classification tasks [13, 14]. In addition, compared to the state-of-the-art video-based animal activity classification works, our avian activity classification is a more challenging task. The video-based animal activity data used in the previous works have no camera movement and a relatively still background [5, 7, 10, 16], but in this avian activity video dataset, video tracks are bird-centered and the birds usually fly at a fast speed, which causes a rapidly changing background in the video tracks. Furthermore, the bird objects can be tiny when they fly far away from the camera. To address these challenges, we incorporate not only the video track frames but also critical engineering features (metadata) provided by Argonne along with each video track into our proposed model, resulting in improved predictions.

Additionally, few DL video-based activity classification works analyze their model's sanity, which determines whether the trained models make reasonable decisions. When the model makes the decision, we expect it to mainly look at the areas where the activity is being observed, for example, the areas around the object whose activity is being classified throughout the video. These analyses are common and important in the state-of-the-art DL image-based classification approaches [1, 18, 19]. In summary, the main contributions of our work are: 1) We propose a novel DL method for video-based avian activity classification. Our proposed Video-Meta Fusion Bi-LSTM model has been designed to classify six bird activities leveraging both the video data and additional engineered features (meta-data). 2) We effectively address the challenge of an imbalanced dataset for training the model. 3) After our model is trained, we verify that our model looks at the right areas to classify the corresponding activities by computing and examining its saliency/backpropagation maps of the input video frames. 4) To the best of our knowledge, this study is the first to utilize RNNs for video-based avian activity classification while also analyzing the decision-making rationale of the proposed DL model in animal activity classification.

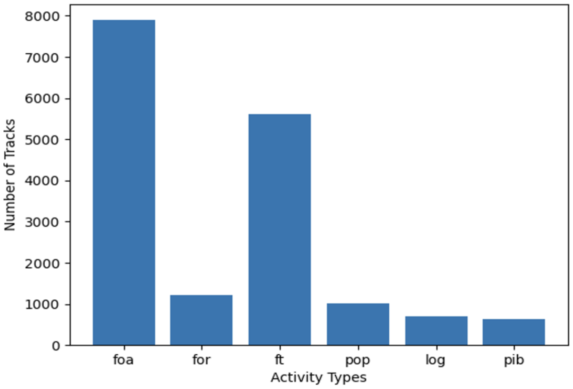


Figure 2: Number of samples in each activity.

2 METHODOLOGY

2.1 Avian activity dataset

Argonne has provided us with an avian activity dataset that consists of video "tracks" and their corresponding feature engineering features (metadata). Each track is a bird-centered sequence of frames for a single bird. An example can be found in Figure 1. The metadata, containing the bird's trajectory (x/y coordinate information) and its moving speed, is pre-computed by Argonne researchers and directly provided within the dataset for each track.

Based on the activities of birds around the solar energy facilities, the dataset is categorized into six activities:

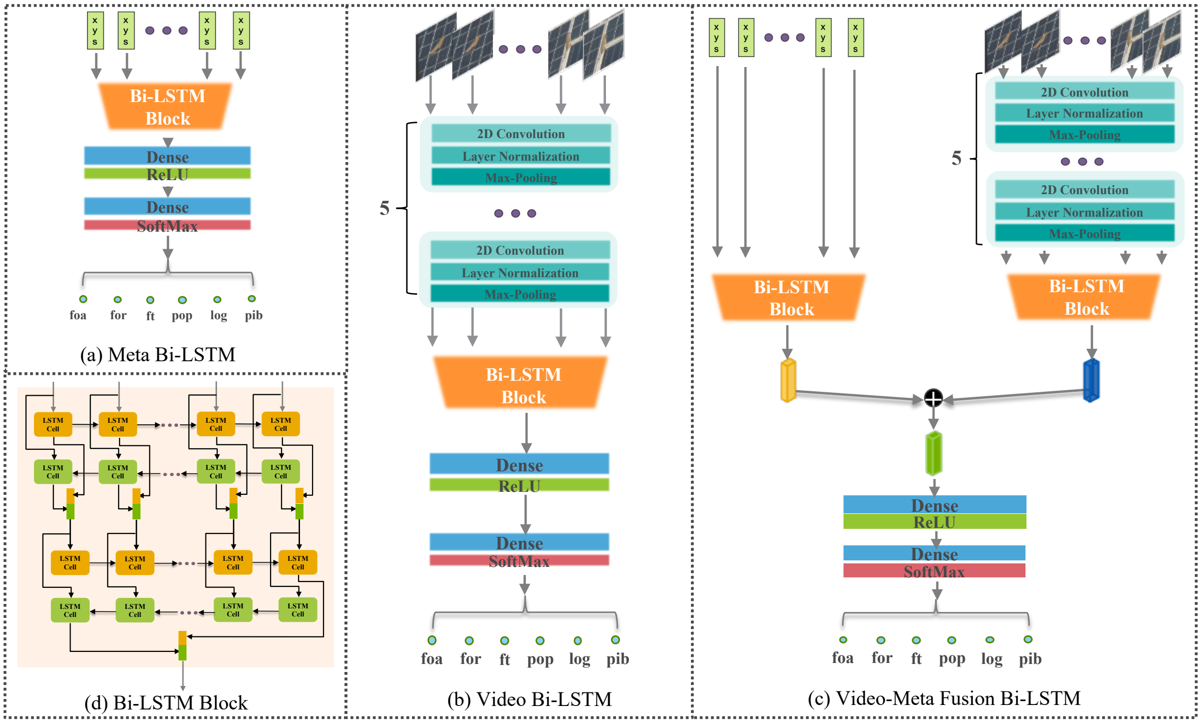
(1) Fly over above (foa): a bird flying high above solar panels.
(2) Fly over reflection (for): a bird's reflection flying over panels.
(3) Fly through (ft): a bird flying near solar panels.
(4) Perch on panel (pop): a bird flying into the frame lands on any part of the panel structure, or a bird on the panel flies away.
(5) Land on the ground (log): a bird flew in and landed on the ground, a bird on the ground flew away, or a bird is moving on the ground.
(6) Perch in background (pib): a bird flying into the frame lands on objects other than panels, or a bird on a non-panel object flies away.

Figure 2 shows the number of samples for each activity category. We can clearly see a heavy imbalance in the dataset. There are two major activities - the foa and ft, and four minor activities.

2.2 Model description

Before proposing the final fusion model, we first introduce two sub-models that utilize either the metadata sequence input (Meta Bi-LSTM) or the video sequence input (Video Bi-LSTM).

2.2.1 Video Bi-LSTM & Meta Bi-LSTM. The Video Bi-LSTM architecture takes the video frame sequence as input and predicts the probability for each activity category. As shown in Figure 3(b),



(a) Meta Bi-LSTM

(d) Bi-LSTM Block

(b) Video Bi-LSTM

(c) Video-Meta Fusion Bi-LSTM

Figure 3: Meta Bi-LSTM, Video Bi-LSTM, Video-Meta Fusion Bi-LSTM models' architecture.

to efficiently extract hierarchical features from input frames, after the video frame sequence is inputted into the model, we first use convolutional, normalization and pooling layers to extract the deep features from them. Then, to incorporate both forward and backward context into the model's understanding of the sequence, these features are then flattened into vectors and passed through a Bi-LSTM block (composed of two stacked bidirectional LSTM layers, as shown in Figure 3(d)), followed by fully-connected dense layers. Finally, the output layer utilizes a softmax function to provide the predicted probability for each activity category. The Video Bi-LSTM model can be described as:

$$p = f_{video}(\boldsymbol{v}), \tag{1}$$

where $\boldsymbol{v}$ denotes the input video frame sequence, p is the output vector of the predicted probabilities for C activity categories, and $f_{video}(\cdot)$ denotes the Video Bi-LSTM model.

In the dataset, some bird video tracks can be very long and last for numerous frames, and we would like our model to get information from both past and future frames during training. Therefore, we employ bidirectional LSTM layers rather than pure LSTM layers when building the models [17].

As explained in Section 1, incorporating engineered features can be beneficial for our task. The current video tracks are bird-centered, which loses the general bird trajectory information in the larger camera field of view. Since the x/y coordinate information represents the trajectory of the bird's movement, and the speed of the bird can differ for different activities along the trajectories, we choose to include the x and y coordinates and the bird's speed as the critical metadata. Figure 3(a) shows the Meta Bi-LSTM model, which solely uses the metadata sequence as input. That is, the input of this model is a sequence of 3-dimensional vectors (x-coordinate, y-coordinate, speed). Unlike the Video Bi-LSTM which takes the high-dimensional image sequence as input, we do not need to use the convolution layers in the beginning to extract the hierarchical features, cause the input now is very low-dimensional. We directly use the Bi-LSTM block followed by fully-connected dense layers. Similarly, the output layer utilizes a softmax function to provide the predicted probability for each activity category. The Meta Bi-LSTM model can be described as:

$$p = f_{meta}(\boldsymbol{m}), \tag{2}$$

where $\boldsymbol{m}$ denotes the input metadata sequence, p represents the output vector containing the predicted probabilities for C activity categories, and $f_{meta}(\cdot)$ denotes the Meta Bi-LSTM model.

2.2.2 *Video-Meta Fusion Bi-LSTM.*

The final proposed model, Video-Meta Fusion Bi-LSTM (VM Bi-LSTM), is shown in Figure 3(c). It incorporates two input streams: the video sequence and the metadata sequence. It fuses these two streams after the Bi-LSTM blocks, where two feature representations are captured and learned from the metadata and video frames along the temporal dimension. The addition of these two deep feature vectors is then input into the following dense layers. The Fusion Bi-LSTM model can be described as:

$$p = f_{fusion}(\boldsymbol{v}, \boldsymbol{m}), \tag{3}$$

where the VM Bi-LSTM model, denoted as $f_{fusion}(\cdot)$, takes both the video frame sequence $\boldsymbol{v}$ and its corresponding metadata sequence $\boldsymbol{m}$ as inputs. p is the output vector of the predicted probability for

C activity categories. The proposed VM Bi-LSTM model can then learn to classify the activity by combining the information from engineered features and raw video sources. In section 4, we show that the proposed fusion model can achieve better results than the two sub-models.

2.3 Training objective

We use the categorical cross-entropy loss as the loss function for training the Video Bi-LSTM, Meta Bi-LSTM, and VM Bi-LSTM models:

$$\mathcal{L} = -\sum_{c=1}^{C} y_c log(p_c), \tag{4}$$

where $[y_1, y_2, ...y_C]$ is the one-hot ground-truth activity label vector with C activity categories, and p_c is the predicted probability for the cth activity category.

2.4 Data augmentation in metadata and video data

Our models are trained with a heavily unbalanced dataset, and this causes severe issues if we directly use it to train our models. Therefore, we need to balance the dataset. In general, we use up-sampling methods for the four minor activities: we apply image augmentation on each frame in the video tracks, like changes in brightness, saturation, or contrast. Since we are augmenting the video tracks instead of the single images, we use the same augmentation process on all frames within each video track to maintain temporal consistency along these frames. The augmentation process can differ from track to track. Besides, we keep the same metadata values for those augmented tracks as the tracks they are upsampled from.

3 EXPERIMENTS AND ANALYSIS

This section presents the experimental results and comparisons among the Video Bi-LSTM, Meta Bi-LSTM, and VM Bi-LSTM models on the avian activity dataset. We use the saliency map to analyze our model and validate the rationality of its predictions. Lastly, we assess the performance of our model.

3.1 Experimental setup

Dataset. The avian activity dataset provided to us by Argonne is not publicly available. It consists of a total of 17,059 bird-centered video tracks, with the two major activities (foa and ft) accounting for over half of the dataset. The metadata information is pre-computed by Argonne researchers and provided to us directly within the dataset for each track. When utilizing this dataset, We separate the dataset into training (80%), validation (10%), and testing (10%) sets. As described in section 2.4, during training the Video Bi-LSTM and Meta Bi-LSTM models, we perform data augmentation to up-sample the four minor activities (for, pop, log, and pib) to 2000 tracks each and randomly down-sample the two major activities to 2000 tracks each. In the case of the VM Bi-LSTM model, which is larger than the two sub-models and has more trainable parameters, we slightly increase the up-sampling and down-sampling numbers to 2500 tracks each category. The video track frames have dimensions of 200 × 200 pixels. During training, validating, and testing, we center-crop the frames to 100 × 100 pixels.

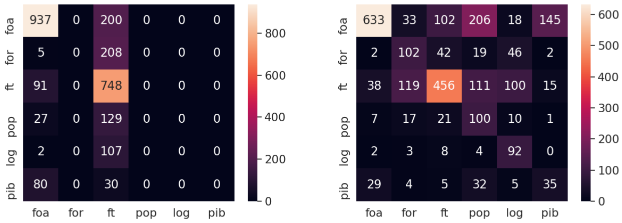


Figure 4: The confusion matrices of Meta Bi-LSTM model trained with unbalanced (left) or balanced (right) datasets.

Train & test process. During training, we set the total epochs to 40. We use the validation loss to monitor the training process and apply early stopping with patience 10. Adam is used as the optimizer with the learning rate equal to 0.001. The batch size is chosen to be one as the length of video tracks in the dataset varies a lot.

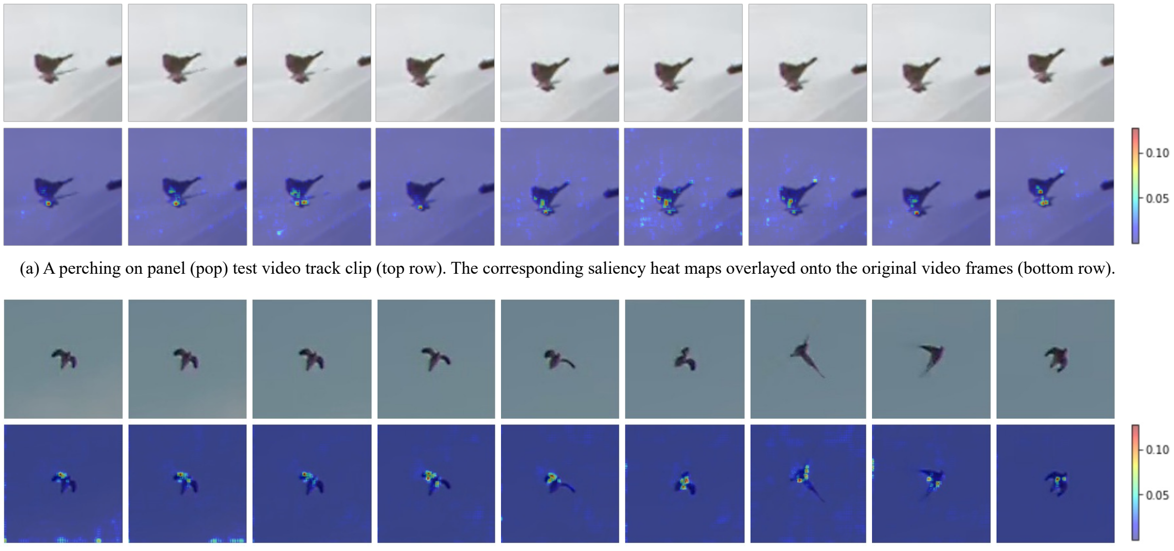
3.2 Results and comparisons

To demonstrate the limitations of training models with an unbalanced dataset, we present a comparison of confusion matrices in Figure 4 between the Meta Bi-LSTM models trained on the original unbalanced dataset and the balanced dataset. When trained with the unbalanced dataset, we can see that the model can only learn to classify the two major activities (foa, ft), while considering all the other minor activities as foa or ft, as shown in Figure 4. After using the balanced dataset, the model now makes predictions across six activities, i.e., it successfully recognizes that there are six classes rather than just two.

Using the balanced dataset, we train and compare the performance of Video Bi-LSTM, Metadata Bi-LSTM, and Video-Meta Fusion Bi-LSTM models. Table 1 presents the comparison of test accuracy. We can see that the fusion model performs the best, it can reach 76.9%, 89.9%, and 95.2% accuracy in Top1, Top2, and Top3 test accuracy, respectively. This comparison confirms our intuition that fusing the critical metadata with the raw video input enables DL models to leverage more information, enhance their learning process, and improve their ability to distinguish between different activities.

Table 1: Test accuracy of Meta Bi-LSTM, Video Bi-LSTM, and Video-Meta Fusion Bi-LSTM models. If the top 2 or top 3 predicted activities contain the given ground-truth activity, then we count it as a correct classification within the Top2 or Top3 test accuracy correspondingly.

	Top1	Top2	Top3
Meta Bi-LSTM	55.3%	78.2%	89.9%
Video Bi-LSTM	72.5%	86.4%	94.0%
Video-Meta Fusion Bi-LSTM	**76.9%**	**89.9%**	**95.2%**



(a) A perching on panel (pop) test video track clip (top row). The corresponding saliency heat maps overlayed onto the original video frames (bottom row).

(b) A fly over above (foa) test video track clip (top row). The corresponding saliency heat maps overlayed onto the original video frames (bottom row).

Figure 5: Saliency maps of the test video tracks.

3.3 Model analysis

We use the saliency/backpropagation maps [19] to validate that our trained model reasonably classifies each activity. The saliency maps indicate the areas of the frames that the model focuses on when making decisions. For an input video sequence with N contiguous frames, denoted as F_i ($i = 1, 2, ..., N$), its saliency map W_i for each frame is calculated according to

$$W_i = |\frac{\partial s_c}{\partial F_i}|, \tag{5}$$

where s_c is the scalar score of class c - the model's output before the final SoftMax layer for the cth class. The derivative in the equation calculates the gradient of s_c at (each pixel of) F_i, therefore, the computed saliency maps indicate the areas/pixels in the video frames which affect the c class score the most. Higher values signify a stronger impact of the corresponding areas/pixels in the frame on the classification decision made by our model regarding activity c.

In Figure 5, we show the saliency maps of a pop test video track and a foa test video track which are correctly classified by the trained VM Bi-LSTM model's top-1 predicted activity. We can see that the trained model primarily directs its attention towards the bird object in most frames. Moreover, for the pop video track, the model also focuses on the areas of the bird reflected on the solar panels, and this is undoubtedly reasonable. As in a lot of the pop tracks, the bird perches and stays on the panels, and the reflection on the panel is an important factor in determining if the bird is staying/perching on the panels.

4 CONCLUSIONS

In this paper, we proposed the Video-Meta Fusion Bi-LSTM model to solve the avian activity classification problem. We fuse the information from raw video RGB frames and the feature engineering metadata when building the model. This fusion strategy is able to gain improved test accuracy results over the Video Bi-LSTM and Meta Bi-LSTM models. Additionally, we employ a consistent data augmentation method to up-sample the video and metadata to solve the dataset's imbalance issue. By looking into the saliency heatmaps, we observe that the proposed model makes its decisions in a rational manner.

ACKNOWLEDGMENTS

This research was funded by the U.S. Department of Energy, Energy Efficiency and Renewable Energy, Solar Energy Technologies Office, grant number 36473.

REFERENCES

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems* 31 (2018).

[2] Mark Anderson, John Kennedy, and Naomi Harte. 2021. Low resource species agnostic bird activity detection. In *2021 IEEE Workshop on Signal Processing Systems (SiPS)*. IEEE, 34–39.

[3] Nutchanun Chinpanthana and Yunyu Liu. 2020. Human Activities of Daily Living Recognition with Graph Convolutional Network. In *Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence*. 305–310.

[4] Anand Dubey, Niall Lyons, Avik Santra, and Ashutosh Pandey. 2022. XAI-BayesHAR: A novel Framework for Human Activity Recognition with Integrated Uncertainty and Shapely Values. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 1281–1288.

[5] Alvaro Fuentes, Sook Yoon, Jongbin Park, and Dong Sun Park. 2020. Deep learning-based hierarchical cattle behavior recognition with spatio-temporal information. *Computers and Electronics in Agriculture* 177 (2020), 105627.

[6] Jen-Cheng Hou, Aileen McGonigal, Fabrice Bartolomei, and Monique Thonnat. 2022. A Self-Supervised Pre-Training Framework for Vision-Based Seizure Classification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1151–1155.

[7] Marcin Kopaczka, Daniel Tillmann, Lisa Ernst, Justus Schock, René Tolba, and Dorit Merhof. 2019. Assessment of Laboratory Mouse Activity in Video Recordings Using Deep Learning Methods. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 3673–3676.

[8] Mario Lasseck. 2018. Acoustic bird detection with deep convolutional neural networks. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*. 143–147.

[9] Yiying Li, Yulin Li, and Yanfei Gu. 2020. Channel-Wise Spatial Attention with Spatiotemporal Heterogeneous Framework for Action Recognition. In *Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence*. 334–338.

[10] Håkon Måløy, Agnar Aamodt, and Ekrem Misimi. 2019. A spatio-temporal recurrent network for salmon feeding action recognition from underwater videos in aquaculture. *Computers and Electronics in Agriculture* 167 (2019), 105087.

[11] Ramadan J Mustafa, Mohamed R Gomaa, Mujahed Al-Dhaifallah, and Hegazy Rezk. 2020. Environmental impacts on the performance of solar photovoltaic systems. *Sustainability* 12, 2 (2020), 608.

[12] B Natarajan, R Elakkiya, R Bhuvaneswari, Kashif Saleem, Dharminder Chaudhary, and Syed Husain Samsudeen. 2023. Creating Alert messages based on Wild Animal Activity Detection using Hybrid Deep Neural Networks. *IEEE Access* (2023).

[13] Samira Pouyanfar, Shu-Ching Chen, and Mei-Ling Shyu. 2018. Deep spatio-temporal representation learning for multi-class imbalanced data classification. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*. IEEE, 386–393.

[14] Samira Pouyanfar, Yudong Tao, Haiman Tian, Shu-Ching Chen, and Mei-Ling Shyu. 2019. Multimodal deep learning based on multiple correspondence analysis for disaster management. *World Wide Web* 22, 5 (2019), 1893–1911.

[15] Abhisek Ray, Maheshkumar H Kolekar, R Balasubramanian, and Adel Hafiane. 2023. Transfer learning enhanced vision-based human activity recognition: a decade-long analysis. *International Journal of Information Management Data Insights* 3, 1 (2023), 100142.

[16] Frank Schindler and Volker Steinhage. 2021. Identification of animals and recognition of their actions in wildlife videos using deep learning techniques. *Ecological Informatics* 61 (2021), 101215.

[17] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45, 11 (1997), 2673–2681.

[18] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[19] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*. Citeseer.

[20] Anil Kumar Sisodia et al. 2019. Impact of bird dropping deposition on solar photovoltaic module performance: a systematic study in Western Rajasthan. *Environmental Science and Pollution Research* 26, 30 (2019), 31119–31132.

[21] Anshul Thakur, Arjun Pankajakshan, and Padmanabhan Rajan. 2018. Learned aggregation in CNN: all-conv net for bird activity detection. In *Detection and Classification of Acoustic Scenes and Events 2018 Workshop*.

[22] Hadiqa Aman Ullah, Sukumar Letchmunan, M Sultan Zia, Umair Muneer Butt, and Fadratul Hafinaz Hassan. 2021. Analysis of Deep Neural Networks For Human Activity Recognition in Videos–A Systematic Literature Review. *IEEE Access* (2021).

[23] Elke Visser, Vonica Perold, Samantha Ralston-Paton, Alvaro C Cardenal, and Peter G Ryan. 2019. Assessing the impacts of a utility-scale photovoltaic solar energy facility on birds in the Northern Cape, South Africa. *Renewable energy* 133 (2019), 1285–1294.

[24] Zhen Xing, Qi Dai, Han Hu, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. 2023. Svformer: Semi-supervised video transformer for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18816–18826.