Engineering Applications of Artificial Intelligence 157 (2025) 111372

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Research paper

A multi-head deep fusion model for recognition of cattle foraging events using sound and movement signals

Mariano Ferrero [a], José O. Chelotti [a,b], Luciano S. Martinez-Rau [a,c,*], Leandro D. Vignolo [a], Martín Pires [d], Julio R. Galli [d,e], Leonardo L. Giovanini [a], H. Leonardo Rufiner [a,f]

[a] Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), FICH-UNL/CONICET, 3000 Santa Fe, Argentina
[b] TERRA Teaching and Research Center, University of Liège, Gembloux Agro-Bio Tech (ULiège-GxABT), 5030 Gembloux, Belgium
[c] Department of Computer and Electrical Engineering, Mid Sweden University, Sundsvall, Sweden
[d] Facultad de Ciencias Agrarias, Univ. Nacional de Rosario, S2125 Zavalla, Argentina
[e] Instituto de Investigaciones en Ciencias Agrarias de Rosario, IICAR, Facultad de Ciencias Agrarias, UNR-CONICET, S2125 Zavalla, Argentina
[f] Laboratorio de Cibernética, Facultad de Ingeniería, Univ. Nacional de Entre Ríos, Oro Verde 3100, Argentina

ARTICLE INFO

Keywords:
Deep learning
Information fusion
Convolutional neural networks
Recurrent neural networks
Precision livestock farming
Ruminant foraging behaviour

ABSTRACT

Monitoring feeding behaviour is a relevant task for efficient herd management and the effective use of available resources in grazing cattle. The ability to automatically recognise animals' feeding activities through the identification of specific jaw movements allows for the improvement of diet formulation, as well as early detection of metabolic problems and symptoms of animal discomfort, among other benefits. The use of sensors to obtain signals for such monitoring has become popular in the last two decades. The most frequently employed sensors include accelerometers, microphones, and cameras, each with its own set of advantages and drawbacks. An unexplored aspect is the simultaneous use of multiple sensors with the aim of combining signals in order to enhance the precision of the estimations. In this direction, this work introduces a deep neural network based on the fusion of acoustic and inertial signals, composed of convolutional, recurrent, and dense layers. The main advantage of this model is the combination of signals through the automatic extraction of features independently from each of them. The model has emerged from an exploration and comparison of different neural network architectures proposed in this work, which carry out information fusion at different levels. Feature-level fusion has outperformed data and decision-level fusion by at least a 0.14 based on the F1-score metric. Moreover, a comparison with state-of-the-art machine learning methods is presented, including traditional and deep learning approaches. The proposed model yielded an F1-score value of 0.802, representing a 14% increase compared to previous methods. Finally, results from an ablation study and post-training quantisation evaluation are also reported.

1. Introduction

The intensification of livestock production systems requires innovative tools to improve efficiency while mitigating environmental negative impacts. Traditional methods of livestock management, often based on herd-level observations, may overlook individual behavioural patterns, leading to suboptimal resource use and increased environmental footprints.

Individualised livestock monitoring offers significant economic benefits, including improved feed efficiency, reduced effluents, and enhanced animal health management (Laca, 2009). For instance, by accurately detecting foraging events, farmers can fine-tune feed distribution, ensuring that animals receive adequate nutrition without overfeeding. This precision not only reduces feed costs – a major expense in livestock systems – but also minimises competition for limited resources. Furthermore, early detection of irregular behaviours through individual monitoring can aid in identifying health issues (Morgan-Davies et al., 2024), reducing veterinary costs and potential production losses.

From an environmental perspective, individualised monitoring contributes to sustainability by promoting optimal grazing practices. Overgrazing, a common issue in unmanaged systems, can lead to soil

* Corresponding author at: Department of Computer and Electrical Engineering, Mid Sweden University, Sundsvall, Sweden.
E-mail address: luciano.martinezrau@miun.se (L.S. Martinez-Rau).

https://doi.org/10.1016/j.engappai.2025.111372
Received 10 September 2024; Received in revised form 22 May 2025; Accepted 1 June 2025
Available online 21 June 2025

degradation, loss of biodiversity, and decreased carbon sequestration. By accurately identifying and managing foraging behaviour, producers can implement rotational grazing strategies that enhance pasture resilience and soil health. Additionally, better feed management reduces greenhouse gas emissions per unit of production, aligning with global goals to mitigate climate change.

With regard to the traditional monitoring of feeding behaviour, two principal activities are considered: grazing and rumination. Despite this, a more fine-grained classification might be possible including drinking, chewing, foraging, and walking, among others (Kilgour, 2012; da Silva Santos et al., 2023). Including these activities might contribute to provide a comprehensive analysis of feeding patterns, nutritional intake, and overall well-being. By accounting for these behaviours, a more complete understanding of the animal's feeding dynamics can be achieved.

Each period of the key activities mentioned before (grazing and rumination) may last from minutes to hours and consists of sequences of specific jaw movement (JM) events that allow their accurate identification and tracking.

These events are classified as bite, chew, and chew-bite (a combination of the two previous events) (Laca and WallisDeVries, 2000; Ungar et al., 2006). Monitoring the occurrence of these events and activity periods allows for the estimation of dry matter intake (Chelotti et al., 2024), the detection of the presence of a disease or condition (Calamari et al., 2014; Paudyal et al., 2018), the prediction of states of stress (Herskin et al., 2004) or anxiety (Bristow and Holmes, 2007), and approximating the calving moment (Büchel and Sundrum, 2014; Clark et al., 2015), to name a few examples.

Continuous direct observation of cattle behaviours represents a challenge, especially when dealing with a significant number of animals distributed across extensive areas. This challenge has driven research into the use of sensors for monitoring relevant livestock behaviours. Various types of sensors have been proposed, allowing for differentiation between those which are positioned on the animal (commonly referred to as "wearables") and those situated externally. The former has been the predominant choice in the literature, with motion sensors being the preferred option, followed by acoustic sensors (Andriamandroso et al., 2016; Chelotti et al., 2024).

Acoustic sensors are able to capture signals with high discriminative power, although the disadvantage is the difficulty in processing them due to the volume of generated information. On the other hand, the processing of IMU signals is simpler due to the smaller number of samples per second. Although these signals record important information about position, turns and other head movements, the discrimination of different JM events might be challenging (da Silva Santos et al., 2023).

While the use of a single sensor has been the most extensively studied approach, the combination of signals from multiple sensors has yet to be fully explored. This represents an advantage in this problem due to the ability to have complementary information to reduce environmental noise, make the system more robust to failures, and improve detection capabilities, among others. This promising approach can be addressed through the use of data fusion strategies combining the most used signals in the state-of-the-art: motion and audio signals.

In the context of information fusion, three main levels of abstraction are frequently employed in situations where data comes from multiple sensors. These are data fusion, feature fusion, and decision fusion (Hall and Llinas, 1997; Qiu et al., 2022). Data fusion level refers to the premature combination of acquired signals from sensors to create a unique signal with several channels, regardless of whether pre-processing is performed or not. In this context, a common approach consists of the creation of multimodal signals by stacking raw signals. On the other hand, the feature-fusion level involves extracting representative values of each signal (usually using fixed-size windows) and then constructing a vector of fixed-dimension elements. The main idea is to combine information from all available signals in this single representation, generating some independence between specific properties of each signal (Spinsante et al., 2016). Feature generation can be manual (i.e. following a feature engineering approach) or automatic (i.e. self-learned features in a deep learning approach). Finally, the decision-level fusion builds a system that combines predictions from underlying systems, each of which analyses information from a single sensor (Garcia-Ceja et al., 2018). Consequently, the system endeavours to optimise the output by combining or selecting hypotheses generated by simpler systems, in accordance with a comparable methodology to ensemble methods (Dietterich, 2000). To create a final decision, traditional approaches could be employed (such as majority voting) in addition to machine learning models (for instance decision trees or logistic regression).

This paper presents a multi-head convolutional neural network (CNN) - recurrent neural network (RNN) approach for the recognition of JM events in grazing cattle. The approach fuses information from acoustic and inertial measurement units (IMU) signals at the feature-level without any prior preprocessing or feature extraction. The proposed model is capable of detecting and classifying JM events simultaneously, distinguishing between five different classes. An investigation into the efficacy of different information fusion architectures has been conducted to identify the optimal configuration for enhancing recognition results in this context. Furthermore, the proposed method has been subjected to empirical evaluation and benchmarked against a range of state-of-the-art alternatives. Experiments were performed to show the superiority of multimodal approaches over unimodal solutions and to illustrate the advantages of deep architectures over traditional machine learning approaches.

An in-depth exploration of the technical details and implications involved in implementing the proposed model is beyond the scope of this study.

The main contributions of this publication are the following:

(a) It presents a multi-head CNN-RNN model that performs information fusion at the feature-level.

(b) It proposes and evaluates different architectures of deep neural networks that perform data fusion at different levels.

(c) It examines the effectiveness and accuracy of the proposed solution by comparing the obtained results with those obtained by state-of-the-art methods.

(d) It presents an ablation study to analyse the benefits of each part of the proposed model.

Our proposed multi-head deep fusion model, leveraging sound and movement signals, provides a novel approach to detecting cattle foraging events with high precision. By integrating these modalities, our work addresses the gap in individualised livestock monitoring technologies and supports sustainable and economically viable livestock production systems.

The structure of the remaining parts of the article is as follows: Section 2 introduces a short overview of the state-of-the-art regarding automatic monitoring of ruminant feeding behaviour. Section 3 describes the proposed feature-level fusion model as well as other fusion level architectures proposed and analysed. Section 4 is dedicated to the experimentation including a description of the adopted methodology. Several comparisons are also presented in this section. Finally, conclusions, limitations, and future research lines are discussed in Section 5.

2. Related work

In the last few decades, ruminant feeding monitoring has attracted scientific attention due to the existing challenges and potential benefits from a practical point of view. Machine learning algorithms are proposed as a means of creating systems capable of working in this context. This section describes the recent developments in ruminant feeding monitoring analysing the most common sensing principles adopted.

2.1. Sensors

Motion sensors allow for the identification of specific ruminant behaviours based on changes in body posture. The principle of motion sensing and its location on the animal determines which movements can be monitored. Accelerometers have been the most studied sensor (Aquilani et al., 2022), due to their low cost, compact size, and low power consumption. Another advantage of the signals captured by this sensor is the low computational cost required for processing them, as they operate at sampling frequencies below 100 Hz. In the context of ruminant feeding monitoring, the use of motion sensors has been primarily focused on detecting activities such as rumination, grazing, and drinking (Aquilani et al., 2022). However, their use for specifically detecting JM events poses challenges due to the limited discriminatory power of the signals captured for this purpose (Chelotti et al., 2024). A variety of approaches have been explored, including the use of accelerometers (Tani et al., 2013; Oudshoorn et al., 2013; Bloch et al., 2023), accelerometers and gyroscopes (referred to as IMUs) (Andriamandroso et al., 2015; Li et al., 2022), and accelerometers, gyroscopes, and magnetometers (referred to as inertial and magnetic measurement units) (Liu et al., 2023).

In free-grazing conditions, acoustic sensors have been demonstrated to be a valuable tool for monitoring feeding behaviour (Ungar et al., 2006). Microphones positioned on the animal's forehead are able to capture sounds produced by the teeth, transmitted through the bones, cavities, and soft tissues of the head (Laca et al., 1992; Chelotti et al., 2024). The information captured in these signals allows for the precise recognition of JM events, as well as grazing and rumination activities (Navon et al., 2013; Chelotti et al., 2018). However, the challenge in exploiting these signals lies in the presence of environmental noise and the computational requirements to process them. Furthermore, the volume of information generated in a given time period is greater than that produced by motion sensors.

2.2. Machine learning approaches

With regard to the development of an automated system capable of classifying JM events and feeding activities, machine learning techniques have been extensively studied (Chelotti et al., 2024). The most commonly used approaches follow a classic pattern recognition pipeline: pre-processing, feature extraction, and classification (Bishop, 2006). Nevertheless, certain limitations have been observed in the classification of JM events (Martiskainen et al., 2009; Greenwood et al., 2017) and feeding activities (Giovanetti et al., 2017). One of the principal limitations of these approaches is the necessity to manually specify the input features of the machine learning models. This aspect introduces a challenge in this problem because there is no consensus on which features should be employed (Chelotti et al., 2024).

As an attempt to address this issue, within the field of deep learning, the use of CNNs has emerged. These architectures are capable of automatically learning features by adapting the filters or weights contained in the network. Li et al. (2021) evaluated the use of CNNs on time-frequency representations of acoustic signals to classify JM events in dairy cows. The reported results are comparable or superior to those obtained through traditional schemes. Wang et al. (2021) explored the use of different deep neural network architectures to classify JM events in sheep from audio files. The proposed approach detects JM events using a heuristic method and subsequently performs classification using deep neural networks. Specifically, the use of fully-connected neural networks (FNNs), CNN, and RNN is evaluated. The input to the CNN and RNN is obtained by calculating Mel-frequency cepstral coefficients. In the case of the FNN, the input data consists of the raw signal corresponding to the previously detected event. Ferrero et al. (2023) proposed a full end-to-end approach which combines FNN, CNN, and RNN to recognise JM events from acoustic signals. The model input constitutes signal chunks extracted using fixed-length time windows. The comparison with other state-of-the-art methods demonstrated a clear improvement over traditional approaches. Nunes et al. (2021) presented a similar approach using RNN to classify JM events in horses from acoustic signals with promising results. The use of deep neural networks has also been applied to inertial signals in the context of recognising feeding activities (Peng et al., 2019; Pavlovic et al., 2021; Wu et al., 2022; Bloch et al., 2023), with promising results.

Architectures that have yielded very good results in related problems such as attention mechanisms (Topaloglu et al., 2023; Aydogmus et al., 2023), have not been applied in this context. One explanation for this may be due to the scarcity of labelled data, which may be an impediment to train models with these characteristics.

2.3. Multimodal learning outside JM events recognition

The utilisation of independent sensing principles for the monitoring of feeding behaviour has been extensively addressed. However, the integration of diverse complementary information sources to achieve more robust and scalable performance in dynamic real-world environments is a promising and underexplored area of study (Chelotti et al., 2024). The use of multimodal systems has been demonstrated to be beneficial in other areas, including speech recognition (Mroueh et al., 2015), emotional state recognition (Tzirakis et al., 2017), and human activity recognition (Nweke et al., 2019).

Arablouei et al. (2023) proposed a method that combines an accelerometer with global navigation satellite system (GNSS) data to classify feeding activities in cows. The solution involves first extracting a set of features from inertial signals and another set from GNSS signals. Subsequently, information fusion is explored at the feature and decision level. A FNN was used to construct the classification model. The reported results demonstrate that information fusion leads to superior outcomes compared to unimodal systems.

The evidence presented in this section indicates the existence of an untapped potential for enhancing JM events recognition. This potential is based on the utilisation of multimodal signals, which allows the exploitation of the advantages offered by each sensing principle. Furthermore, another aspect that has not been studied thus far is the generation of deep learning architectures capable of merging these signals and autonomously learning features, subsequently enabling the recognition of the JM events present in them. Results reported in the literature Ferrero et al. (2023) suggest that the combination of convolutional and recurrent architectures emerges as a promising line of research on this problem.

3. Methodology

This section describes a multimodal deep learning architecture based on the combination of three types of neural networks: CNN (Lecun et al., 1998), RNN (Rumelhart et al., 1986) and FNN (Bishop, 2006). In the following, a brief introduction to these architectures is provided. Then, a detailed description of the proposed method is presented with other proposed architectures which perform fusion at different levels are also introduced. Lastly, the dataset used in the experimentation is described.

3.1. CNN, RNN and FNN

FNN refers to a traditional neural network architecture in which each node belonging to a layer is connected with all nodes of the previous layer. This architecture has been used in classification and regression problems (Bishop, 2006). There are usually three types of layers including input, hidden, and output layers. While the neurons of the input layer represent the features provided to the network (input data or outputs from other networks), each neuron of the hidden and output layers represents a processing element that combines the output of incoming connected neurons using a non-linear activation function.

The overall formal representation for a single hidden layer network is expressed in Eq. (1).

$$y_k(x,w) = \sigma\left(\sum_{j=1}^{M} w_{j_i}^{(2)} h\left(\sum_{i=1}^{D} w_{j_i}^{(1)} x_i + w_{j_0}^{(1)}\right) + w_{k_0}^{(2)}\right) \quad (1)$$

Herein, y_k denotes the output of the neuron k based on the input vector x of size D and a set of weights w, h denotes the activation function of M neurons in the hidden layer, whereas σ represents the activation function of the output neuron. The strength of the connections between neurons in FNNs (w in Eq. (1)) is controlled using weights, which are optimised during the training process to adapt the model outputs to a set of desired values (Bishop, 2006).

CNNs (Lecun et al., 1998) are one of the most widely used architectures in recent decades. These networks usually consist of several convolutional layers, and each layer contains one or more filters (a set of arbitrary decimal numbers) to produce an output feature map of its inputs. In the learning stage, the weights of the filters (used in traditional convolutional mathematical operations) are adjusted to approximate the outputs using optimisation strategies as described above for FNNs. By doing this, the layers are capable of learning different high- and low-level patterns without explicit domain knowledge. In the field of information fusion, several sub-models (usually referred to as heads) could be independently applied to input signals to extract relevant features from them. In the case of a one-dimensional (1D) CNN with n heads, the expression of the output value z at position i in feature map m at layer l of head c can be denoted by Eq. (2).

$$z_i^{clm} = h\left(\sum_{j=0}^{F-1} x \times w_j^{clm}\right) \quad (2)$$

Here, h indicates the activation function for the kernel of size F and weights w_j, and x represents the signal affected by the kernel.

In CNNs, convolutional layers are complemented by other types of layers, such as pooling, batch normalisation, and dense layers. Pooling layers perform simple mathematical operations on patches of the feature maps, such as extracting the maximum value, to reduce the dimensionality of the input. Batch normalisation layers, on the other hand, perform input standardisation to speed up the network training process. Dense layers are equivalent to hidden layers in FNNs and allow the network to adapt the intermediate representations learned by the convolutions to effectively influence the final output. The connection between convolutional and dense layers is established by a flattening operation to convert the output of the convolutional layers into a 1D vector.

Although FNNs can be used in problems with sequential or time series data, they present certain challenges that make them inappropriate in these scenarios. To address this limitation, RNNs emerged (Rumelhart et al., 1986). In this architecture, layer outputs are connected as inputs to the same layer. A variation of an RNN known as Bidirectional RNN (Schuster and Paliwal, 1997) adds a copy of the proposed network trained on the reverse data sequence. Both independently trained RNNs are then connected to the next layer of the network.

Early RNN architectures have certain drawbacks related to the ability to learn efficiently from long sequences and new alternatives have been proposed. Gated recurrent units (GRUs) are a type of RNN in which each neuron has two different gates: reset and update (Cho et al., 2014). These gates control how much information from previous and current states is used. A GRU architecture, in contrast with simple RNNs, effectively captures long-term dependencies in sequences by addressing the vanishing gradient problem. Additionally, GRUs are computationally more efficient and require fewer parameters than Long Short-Term Memory (Hochreiter and Schmidhuber, 1997), another RNN type which includes three gates, making them faster to train while still providing improved performance over simple RNNs, especially in tasks requiring memory of long-term dependencies.

A representation of a GRU cell is shown in Fig. 1, and the associated mathematical expression is given in Eqs. (3) to (6).

$$r_t = \sigma\left(W_r x_t + W_r h_{t-1} + b_r\right) \quad (3)$$

$$z_t = \sigma\left(W_z x_t + W_z h_{t-1} + b_z\right) \quad (4)$$

$$n_t = \phi\left(W_n + r_t \odot \left(W_n h_{(t-1)}\right) + b_n\right) \quad (5)$$

$$h_t = (1 - h_z) \odot n_t + z_t \odot h_{(t-1)} \quad (6)$$

Herein, x_t represents the input vector, h_t the output vector, and z_t, n_t and r_t are the update, new and reset gate vectors, respectively at time t. σ and ϕ represent the activation functions, whereas W and b are the parameters matrices and the bias vector of each gate, respectively. Bidirectional GRUs (BGRUs) have shown promising results in sound events detection (Yihan et al., 2021) and classification (Zhu et al., 2020).

Stochastic gradient descent and backpropagation (Rumelhart et al., 1986) are very common algorithms to perform parameter optimisation in neural networks. In this context, artificial neural networks tend to overfit training data. To reduce the possibility of this, a dropout operation is used. This regularisation technique introduces random cuts between layer connections during training (Hinton et al., 2012).

3.2. Proposed model architecture

Several deep neural network architectures could be proposed to merge the available acoustic and motion signals in this problem. Here, an architecture has been chosen that is capable of extracting features from each signal independently and combining them into a common feature space (feature-level fusion) by using CNNs. The rationale behind this choice lies in the fact that architectures performing feature fusion have proven beneficial in related problems where combining data from different types of sensors is required (Son and Kang, 2023; Islam et al., 2023; Tan et al., 2024). Furthermore, since each signal captures particular properties of the phenomenon of interest using a different sensing principle (sounds of the JM events, and displacement and rotation of the animal head), it is expected that extracting specific features from each of them will be advantageous compared to generating a single signal with multiple channels.

To solve the problem of JM events recognition (which implies detection and classification), a hybrid multimodal network architecture is presented, composed of multi-head 1D-CNN, RNN, and FNN. To the best of our knowledge, this study represents one of the first multimodal approaches to the problem of JM events recognition using acoustic and IMU signals. The input to the network is represented by frames, which are extracted from the raw signals using fixed sliding time windows without any prior preprocessing or feature extraction. The model classifies each window into one of five possible classes: bite, chew-bite, grazing-chew, rumination-chew, and no-event (to represent the absence of any particular JM event). Hence, the proposed method addresses the challenges of both detecting and classifying JM events simultaneously.

An overall graphical representation of the proposed model composed of three blocks is presented in Fig. 2. The model processes chunks of input signals computed using a time window duration of 300 ms, with a 50% overlap between consecutive windows. The first block introduces a multi-head CNN combining three independent 1D CNNs. This block extracts low- and high-level features from acoustic and movement signals independently and performs dimensionality reduction at the same time. Each head of the CNN is composed of a normalisation layer (or re-scaling in the audio head), a sequence of 1D convolutional layers, followed by a max pooling layer. A flatten operation is also used in each head, and those values are finally concatenated to create a unique 1D feature vector representation. The

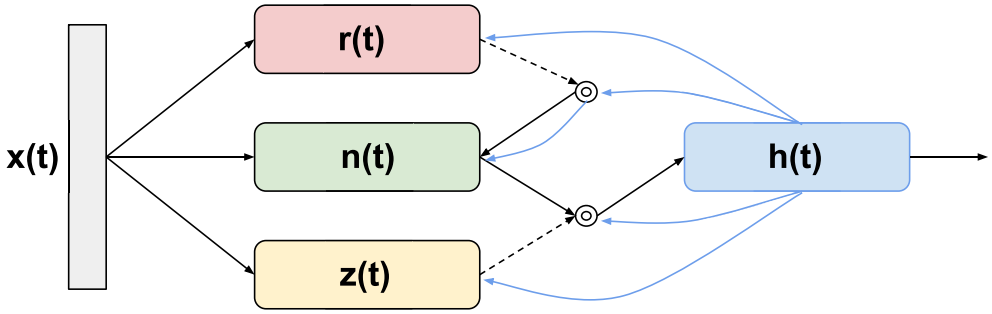

Fig. 1. GRU cell diagram including the different gates and their connections.

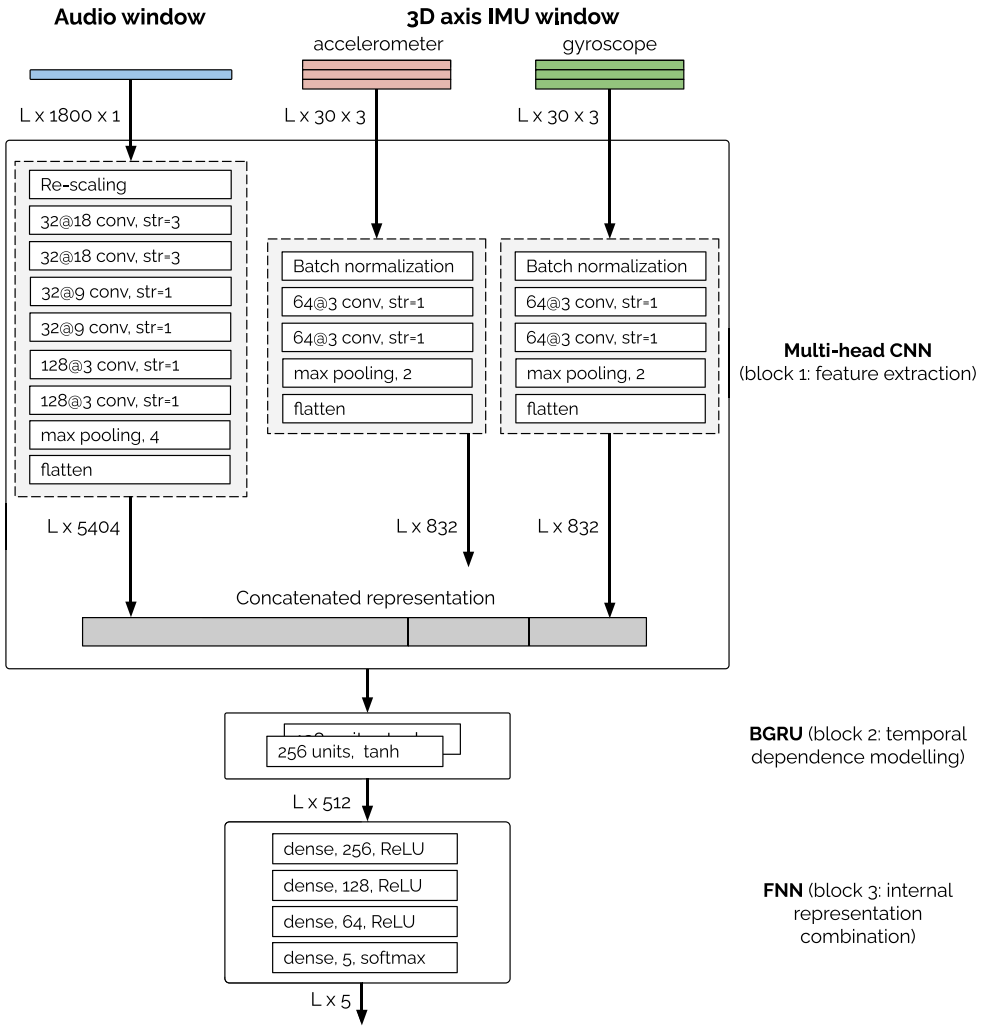

Fig. 2. Proposed method architecture: input signals correspond to audio and movement chunks extracted using fixed length time windows. Each convolution layer shows the number of kernels and kernel size (ReLU was used as activation function), whereas max pooling layers specify the filter size. Dense layers indicate the number of neurons and activation function. At each step the feature dimensions are given, L being the number of windows in the sequence.

second block introduces an RNN, consisting of a BGRU layer of 256 cells, giving the model the ability to capture temporal dependencies present in data. The last block of the model introduces an FNN, which combines information in dense layers and predicts class probabilities for each input window. The first and third blocks are enclosed within time-distributed wrappers so that the same layers and parameters are applied to each window of the input sequences. The rectified linear unit (ReLU) was used for all convolutional layers, whilst the cells of the BGRU use hyperbolic tangent and sigmoid. All dense layers of the FNN use ReLU as well, except for the last dense layer, which uses the softmax function for the final classification. The total number of parameters of the model is 11,704,478.

3.3. Different information fusion strategies

As mentioned in Section 1, there are three main levels at which data fusion can take place: data, features, and decisions. While the proposed model performs feature-level fusion using a multi-head CNN, other architectures that perform fusion at data and decision levels have been proposed and explored as well.

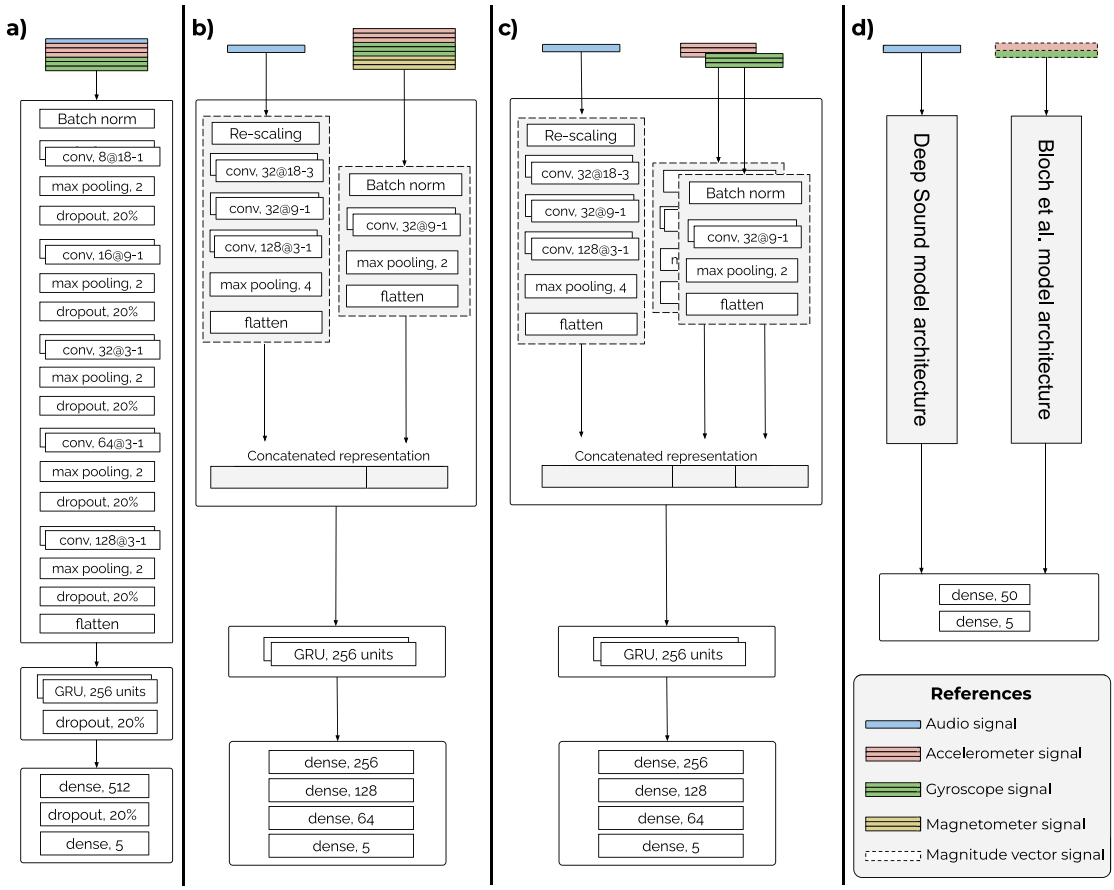

Fig. 3. Illustration of the architectures for different fusion levels, where each level represents the configuration that reached the best results. (a) data-level fusion; (b) feature fusion with two independent CNN and feature concatenation; (c) feature fusion with three independent CNN and feature concatenation (proposed model); (d) decision fusion using an FNN for the final decision model. In all cases, the best results were obtained with a window size of 0.3 s.

For comparison purposes, the best-performing model architectures for the different levels of signal fusion were determined in each case (Fig. 3). In particular, for the feature-fusion level, a variation of the proposed model with 2-heads CNN is included. Several models were evaluated for all fusion levels by varying the number of layers, the size and quantity of filters, and the inclusion of intermediate layers and operations, such as max pooling or dropout (for example, the use of dropout operations has been evaluated in all architectures but it only improves at data-level fusion). Different sizes of the window used to extract data from input signals were also studied. Based on previous studies, durations of 0.3, 0.5, and 1 s were selected for comparison (Alvarenga et al., 2020; Ferrero et al., 2023). Different combinations of input signals were also evaluated, using: (a) all available raw signals; (b) raw audio, accelerometer, and gyroscope signals; (c) raw audio signal, and accelerometer and gyroscope vector's magnitude calculated using Eq. (7)

$$s = \sqrt{s_x^2 + s_y^2 + s_z^2} \tag{7}$$

In the data-level fusion architecture (Fig. 3a), signals from sound, accelerometer, and gyroscope are concatenated at the initial stage creating a single input to the classifier. Due to differences in the number of samples in each signal, the data from the IMU has been resampled in order to match the sampling frequency of the audio signal.

Feature-level fusion has been evaluated using a multi-head CNN on two main approaches: (i) a 2-head CNN (Fig. 3b), which uses one CNN for all data from an IMU sensor; and (ii) a 3-head CNN (Fig. 3c), which represents the proposed model presented in Section 3.2. In both cases, an intermediate representation is constructed by doing a concatenation of automatically extracted features from convolutional layers. This combination approach was selected to deal with the difference of feature space size between heads' outputs, and to provide to the following layers all the available information. Other methods for IMU heads (which share the same input size) were tested – in particular average, maximun, and multiplication – with no improvements.

Decision-level fusion was explored by implementing two base models which process input signals from each sensor independently (Fig. 3d). Audio signals were processed using the architecture proposed by Ferrero et al. (2023), whereas the proposed architecture by Bloch et al. (2023) was used to process inertial signals. The output probabilities of these models are then introduced to a meta-classifier to make a final output decision. Combinations of different base models were also evaluated, including the former two base models, and those proposed by Chelotti et al. (2018) (called Chew-Bite Intelligent Algorithm (CBIA)) and by Alvarenga et al. (2020). Decision trees and multilayer perceptrons were explored as meta-classifiers, as well as traditional methods such as majority voting. In all cases, model weights have been initialised randomly.

3.4. Dataset

The fieldwork to collect the dataset occurred on 1st August 2022 at the Campo Experimental J.F. Villarino, Facultad de Ciencias Agrarias, Universidad Nacional de Rosario (UNR) located in the city of Zavalla, Argentina. The area of 450 hectares is made up of several research and productive subsystems, which are representative of the activities in the area of influence (pork, dairy, beef, and crops). In particular, the dairy subsystem can be characterised as a medium-sized, intensified pastoral-based dairy farm with 140–165 milking cows, with an individual daily

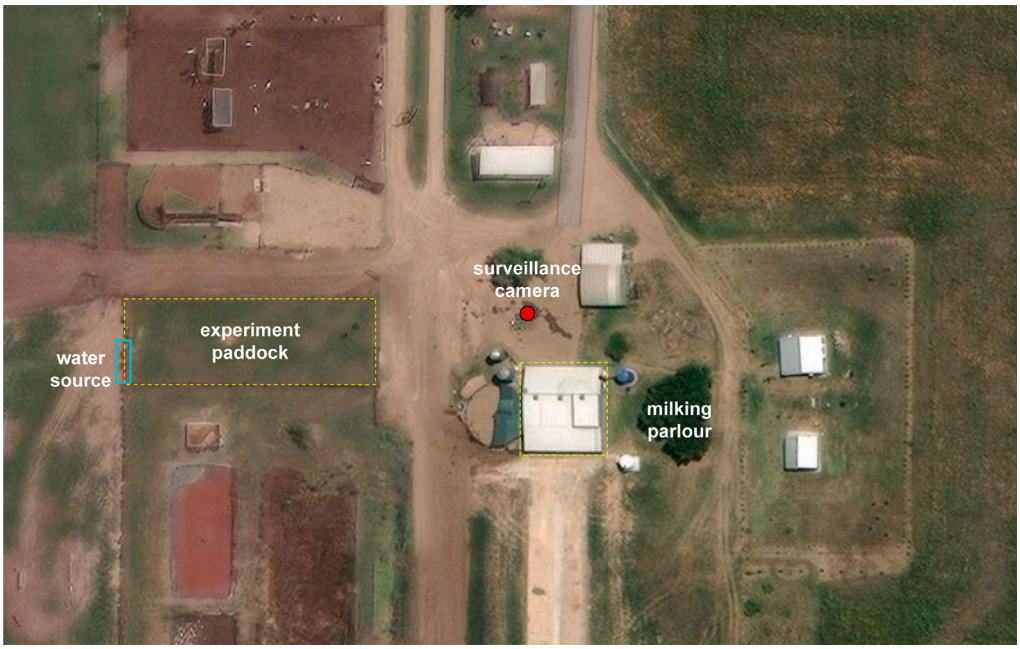

Fig. 4. Satellite image of the dairy facilities detailing experimental paddock area, water source, surveillance camera position, and milking parlour.

production of 24–27 l of milk. The protocol used to conduct the experiment has been evaluated and approved by the Committee on Ethical Use of Animals for Research of the UNR.

The paddock area was approximately 1.200 m^2 (20 × 60 m) and was fully enclosed with fences. This place was covered with naturalised perennial grasses (with dominance of *Lolium sp*, *Festuca sp* and *Cynodon sp*). The experimental cows were free to graze within the paddock, and they had permanent access to a watering trough.

This area was permanently monitored by an outdoor dome video camera positioned at a lateral distance of 30 m from the paddock to assist during the labelling process. Fig. 4 introduces a satellite view of the dairy facilities with references to the most important places for the experiment. In addition, two observers with knowledge of animal behaviour manually logged the main behaviours and significant activities on spreadsheets throughout the experiment. Data have been obtained from three 4-year-old lactating Holstein cows weighing 570–600 kg. All cows were tamed and trained in the experimental routine before the final recordings. Each animal was equipped with an acquisition data device consisting of an external microphone (IP57 100 mm, −42 ± 3 dB, SNR 57 dB) plugged via a 3.5 mm jack to a Moto G6 smartphone.[1] Each device was fixed inside a plastic box and secured to prevent internal movements. This same instrumentation has been used in another similar study (Andriamandroso et al., 2017). Microphones were located on the cow's forehead and covered with rubber foam to isolate them from wind-induced noise and protect them from other frictions. Boxes were mounted to the top side of a halter neck strap (Fig. 5).

Data signals were recorded and synchronised using a specifically developed and tested Android application running in the Moto G6 smartphones, using the internal IMU and the external microphones. Three-dimensional IMU signals were recorded using a sampling rate of 100 Hz. Audio recordings were stored using high-efficiency advanced audio coding (Bosi et al., 1997) with a sampling rate of 44.1 kHz and a bit rate of 128 kbps, single channel (mono). The experiment lasted approximately 6 h (from 09:11:22 to 15:10:20) thus a total of 18 h were generated in total. For this study, all audio signals were resampled to 6 kHz. Although the experiments were conducted in a confined area, animals were exposed to environmental noise conditions such as bird chirps, wind gusts, and movements that are not directly related to JM events.

From the collected signals, a total of 29 segments were carefully chosen for annotation with a duration of 9 min and 31 s on average and a standard deviation (SD) of 1 min and 57 s. Because of the high time demand for the labelling process, a representative subset of signals was selected. A total of 4 h, 36 min and 1.4 s have been annotated. The size of the generated dataset represents a significant increase compared to other datasets used in previous studies (Vanrell et al., 2020; Martinez-Rau et al., 2023). Each segment corresponds to a particular feeding activity (grazing or rumination) and is composed of a sequence of quasi-periodic JM events.

To create event labels, two experts in ruminant foraging behaviour independently delimited the JM events (including event label, start, and end time) by watching and listening to the acoustic signal. The agreement result was 97.63% on average. Both experts worked together to achieve a final decision in case of disagreement.

Based on previous studies (Martinez-Rau et al., 2022), four mutually exclusive labels were treated: bite, grazing-chew, rumination-chew, and chew-bite (a compound movement which is composed of a chew followed by a bite when the animal closes its jaw). Rumination-chew and grazing-chew are events that differ primarily in the feeding activity in which they occur. In the case of rumination, the animal is generally in a state of rest (standing or lying down) and only chew events are present. During grazing, the cow is typically foraging for food (walking, searching, tearing off plants) so the movement of its body and head is recurrent. Chews alternate with bites, or they are even combined (chew-bite). Another difference between rumination-chew and grazing-chew events is the energy of the signal recorded by the acoustic sensor, being higher in the case of grazing (Martinez-Rau et al., 2022). A visual representation of a typical waveform of each JM events from the acoustic signals is presented in Fig. 6. The number of labelled samples for each JM event in the dataset and duration statistical values are presented in Table 1.

4. Experiments, results and discussions

In this section, the methodology selected to drive the experimentation is explained, and the results and discussions of performed experiments are presented as well.

[1] Moto G6 smartphone specifications.

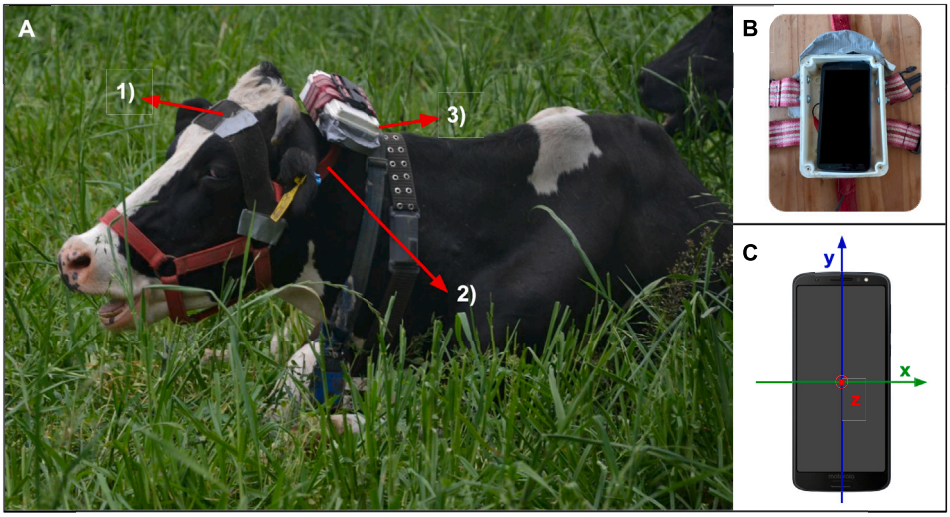

Fig. 5. Experimentation setup description. (A) Cow in the paddock during a rumination period with external microphone (1), halter (2), and plastic box (3). (B) Moto G6 placed in a plastic box; (C) axis from IMU sensors orientation: *x*-axis is aligned with a tail-to-head vector of the animal, *y*-axis describes sideway movements, whereas *z*-axis captures up and down movements.

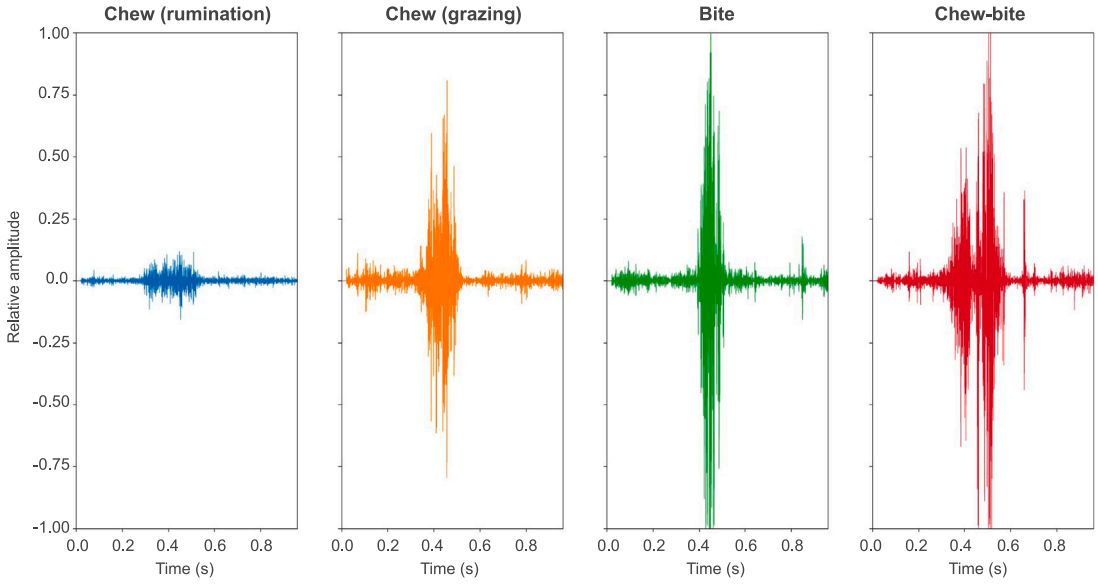

Fig. 6. Characteristic waveform of the 4 different JMs events classes considered in the study.
Source: Adapted from Martinez-Rau et al. (2023).

Table 1
Number and duration of annotated jaw movements (JM) events from acoustic signals before windows extraction.

JM	Number	Duration [s]		
		Mean	Min	Max
Bite	2234	0.33 ± 0.084	0.115	0.926
Chew-bite	6605	0.436 ± 0.087	0.187	0.961
Grazing-chew	6905	0.323 ± 0.066	0.144	0.665
Rumination-chew	2751	0.341 ± 0.051	0.167	0.806
Overall	18,495	0.362 ± 0.092	0.115	0.961

4.1. Experimental settings

From the total of 29 signal segments, 24 were used for model selection purposes. All models were trained and evaluated using a 5-fold cross-validation (CV) scheme with each fold containing 4 or 5 segments. Each fold contains 1 segment from a rumination period and the rest from grazing intervals. This relation between grazing and rumination was proposed to balance the number of JM events. While grazing includes grazing-chews, bites and chew-bites, rumination only contains ruminating-chews. The remaining 5 segments were separated for test purposes, meaning the evaluation of the generalisation capability of the model performing the best on validation sets. The separation of data into different sets was conducted before the experimentation stage, and these sets remained constant throughout this stage. In order to solve class imbalance, the weights of training samples were adapted according to Eq. (8).

$$W_{ic} = \frac{N_{max}}{N_c} \tag{8}$$

where W_{ic} is the weight of instance i associated with class c; N_{max} is the number of instances of the majority class and N_c is the number of instances of class c. Experiments using a data augmentation approach

Table 2
Information fusion architectures (Fig. 3) results based on F1-score, precision, recall, and error rate by class and overall results. In all cases, the average and the SD across validation sets during the 5-fold CV phase are reported.

	Data level	Feature level (2-heads CNN)	Feature level (3-heads CNN)	Decision level
	F1-score			
Bite	0.403 ± 0.066	0.581 ± 0.096	**0.662 ± 0.006**	0.469 ± 0.274
Chew-bite	0.624 ± 0.035	0.797 ± 0.005	**0.811 ± 0.027**	0.733 ± 0.145
Grazing-chew	0.389 ± 0.041	**0.809 ± 0.026**	0.805 ± 0.038	0.562 ± 0.334
Rumination-chew	0.013 ± 0.022	**0.870 ± 0.049**	0.827 ± 0.146	0.670 ± 0.195
Overall	0.450 ± 0.036	0.793 ± 0.040	**0.802 ± 0.033**	0.656 ± 0.207
	Precision			
Bite	0.357 ± 0.134	**0.758 ± 0.051**	0.717 ± 0.039	0.587 ± 0.147
Chew-bite	0.517 ± 0.033	0.717 ± 0.084	**0.747 ± 0.052**	0.663 ± 0.183
Grazing-chew	0.386 ± 0.052	**0.728 ± 0.025**	0.719 ± 0.050	0.656 ± 0.186
Rumination-chew	0.062 ± 0.085	0.856 ± 0.026	**0.866 ± 0.029**	0.676 ± 0.192
Overall	0.430 ± 0.045	0.742 ± 0.046	**0.749 ± 0.038**	0.660 ± 0.176
	Recall			
Bite	0.528 ± 0.090	0.488 ± 0.130	**0.618 ± 0.081**	0.445 ± 0.297
Chew-bite	0.788 ± 0.058	**0.908 ± 0.022**	0.890 ± 0.010	0.839 ± 0.074
Grazing-chew	0.397 ± 0.046	0.910 ± 0.028	**0.917 ± 0.018**	0.559 ± 0.377
Rumination-chew	0.007 ± 0.013	**0.887 ± 0.081**	0.822 ± 0.216	0.674 ± 0.202
Overall	0.474 ± 0.033	0.852 ± 0.031	**0.864 ± 0.029**	0.658 ± 0.238
	Error rate			
Bite	1.643 ± 0.504	0.674 ± 0.084	**0.624 ± 0.090**	0.786 ± 0.221
Chew-bite	0.950 ± 0.094	0.471 ± 0.149	**0.418 ± 0.077**	0.669 ± 0.437
Grazing-chew	1.248 ± 0.133	**0.431 ± 0.058**	0.447 ± 0.101	0.625 ± 0.340
Rumination-chew	1.053 ± 0.045	**0.262 ± 0.086**	0.302 ± 0.197	0.662 ± 0.401
Overall	1.015 ± 0.107	0.337 ± 0.064	**0.327 ± 0.050**	0.513 ± 0.31

were evaluated as an alternative to sample weighting to solve the class imbalance, with inferior results.

For unification process during training, all windows extracted from each signal were converted into smaller sequences of a fix number of windows. Based on this, each example provided to the model consists of a sequence of L windows. Different values have been evaluated for this parameter and L = 46 emerged as the one that obtained the best results in preliminary experiments. The length of the original signal included in each sequence varies according to the window size, being for example 6.9 s for a window size of 300 ms with 50% overlap. A padding operation was used to complete the missing windows in those shorter sequences if necessary.

All the necessary code was developed using Python version 3.10.12 and it is available in the project repository.[2] Several utilities from Python library scikit-learn 1.2.2 have been used, in particular label encoders, k-fold extraction, grid search, and the implementation of traditional machine learning algorithms (such as decision trees). Tensorflow 2.12.0 was used to define and train the neural network architectures. Experiments were performed using an Intel Core™ i7-8700 3.20 GHz CPU, 64 GB RAM and a dual NVIDIA GPU configuration composed of 24 GB GeForce RTX 3090 and 24 GB RTX A5000.

For training, the Adam optimiser (Kingma and Ba, 2014) was chosen, utilising a total of 1400 epochs with an early stopping tolerance of 50 epochs. The batch size was set to 5, and categorical cross-entropy was employed as the loss function. Default values were retained for the remaining parameters.

4.2. Evaluation metrics

The process of JM events recognition involves initially detecting the event, i.e., recognising the onset and offset, and subsequently, assigning a class to the event. In this scenario, detection errors directly impact the classification task. Based on this, the problem addressed in this work requires the use of an evaluation methodology that takes into account both aspects.

The *sed_eval* toolbox (Mesaros et al., 2016, 2021) has been selected to calculate the performance during experimentation. This tool has been used in numerous studies related to event recognition in sounds (Serizel et al., 2020; Venkatesh et al., 2022). Furthermore, it is a comprehensive open-source toolbox that implements a range of metrics suitable for the objectives set in this work.

Given a reference event, the criterion used by the tool to classify a prediction generated by a system as correct includes three conditions: **(a)** the onset of the predicted event must fall within the interval defined by the onset of the reference event ± tolerance value (300 ms); **(b)** the offset of the predicted event must fall within the interval defined by the offset of the reference event ± tolerance value (300 ms); **(c)** the class of both events must be equivalent. Fig. 7 introduces examples where different situations for conditions (a) and (b) can be observed.

Regarding classification results, the metrics expressed in Eq. (9) to (12) have been used:

$$precision = \frac{TP}{TP + FP} \tag{9}$$

$$recall = \frac{TP}{TP + FN} \tag{10}$$

$$F1 - score = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{11}$$

$$Error - rate = \frac{S + D + I}{N} \tag{12}$$

where TP denotes true positive, FP false positive, FN false negative, S substitutions (correct detected JM events in system output but incorrectly labelled), I insertions (detected JM events for the system output that do not exist in the ground truth) and D deletions (ground truth JM events that are not detected). Metrics were computed for each class individually as well as for the overall multi-class. The overall metrics handle the multi-class imbalanced condition by computing micro (class) averages (Sokolova and Lapalme, 2009). Micro average computation implies that TP, FP, and FN are obtained by summing up samples through all classes. For instance, the term TP is ultimately

[2] Project repository link: https://github.com/sinc-lab/chewbite-deep-fusion.

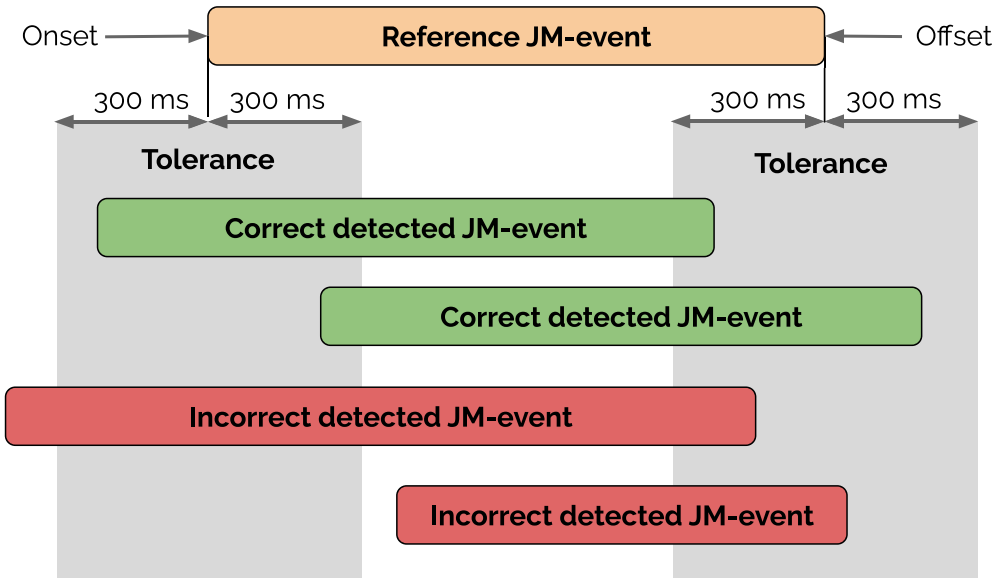

Fig. 7. Illustration of evaluation procedure implemented by the sed_eval toolbox used in this article. Two pairs of JM events (one pair correct and one pair incorrect) with respect to a reference JM event.
Source: Adapted from Mesaros et al. (2016, 2021).

Table 3
Performance of the proposed model with 0.3, 0.5, and 1 s time windows, each with a 50% overlap.

	F1-score	Precision	Recall	Error rate
0.3 s	**0.802 ± 0.033**	**0.749 ± 0.038**	**0.864 ± 0.029**	**0.327 ± 0.05**
0.5 s	0.507 ± 0.254	0.496 ± 0.238	0.524 ± 0.263	0.769 ± 0.245
1 s	0.297 ± 0.229	0.314 ± 0.207	0.295 ± 0.233	1.006 ± 0.119

represented as $TP_{gc} + TP_{rc} + TP_{cb} + TP_b$, denoting the number of true positives for grazing-chews, rumination-chews, chew-bites, and bites, respectively. With the exception of the error rate, for all other metrics between Eqs. (9) and (12) higher values are indicative of a better model.

4.3. Fusion level comparison

An evaluation of the classification performance of the considered level fusion architectures from Fig. 3 is presented in Table 2. The information fusion scheme that achieved the best results was the feature-level in all analysed metrics. In particular, the proposed model with 3-heads CNN scored the best based on the overall F1-score. In addition, the decision-level model outperformed the data-level architecture. For all metrics, data-level fusion presented the lowest performance. Particularly remarkable is the incapacity of this architecture to recognise associated rumination-chews events.

One possible interpretation of this comparison might be that feature-level fusion is more suitable for this task compared to data-level and decision-level fusion because it leverages the strengths of both sensor modalities (acoustic and inertial signals) by integrating informative features that are automatically extracted from each sensor domain. This might be interpreted as a specific characteristic of the presented approach, due to different conclusions reported in other studies with other model specifications (Nweke et al., 2019; Arablouei et al., 2023). In data-level fusion, raw sensor data from different modalities are directly combined, which can lead to issues due to the heterogeneity of the data types (e.g., sampling rates, signal formats), making it difficult for the model to effectively exploit the complementary nature of each sensor. On the other hand, decision-level fusion combines predictions from separate models trained on individual sensor types. While this approach might capture some modality-specific insights, it does not capitalise on the synergistic relationships between features from different sensors during the learning process.

Although the structures of both feature-fusion level architectures are similar, the 2-head CNN architecture obtained slightly inferior results. Apart from that, the use of magnetometer signals does not appear to provide benefits over using only the accelerometer and gyroscope signals. This is probably related to the fact that the execution of JM events does not have a relation with changes to any particular location, something that is measured by this sensor. In fact, the performance of the model drops when using this signal, due to the need to process data that apparently does not contain discriminative power in this context.

When comparing the recognition performance for individual JM event classes, worse results were obtained with the minority class (bite), even with the use of different weights per class to counteract the data imbalance. These results are consistent with previous studies (Martinez-Rau et al., 2022; Ferrero et al., 2023).

Overall, there is a reduced variability in the metrics (F1-score, precision, and recall) obtained across the first three fusion levels, indicating stability in the performance of the models (Fig. 8). The decision-level model is an exception because the SD between the folds is significant.

4.4. Effect of time window size and quantisation

The performance of the proposed model has been evaluated for different sliding input window sizes. Table 3 introduces the results using three different sizes of sliding windows: 0.3 s, 0.5 s, and 1 s. The overlap between two consecutive windows was 50%. The reported results include the average values per metric for the different validation folds, as well as the SD.

A window size of 0.3 s exhibited the best metrics, while the use of 1 s windows performed the worst. Based on these results, it would be useful to use short time windows, similar to the average duration of JM events (Table 1).

Conversely, the use of longer time windows seems to worsen the performance. There are two likely causes for this: firstly, when extracting 1 s fragments, two consecutive JM events could be partially included, generating chunks with valuable information that are categorised as the absence of JM events — "no-event" class (Fig. 9). Lastly, the detection of JM events represents a challenge for the tolerance value selected for evaluation purposes, since JM events generally have a duration shorter than the window size.

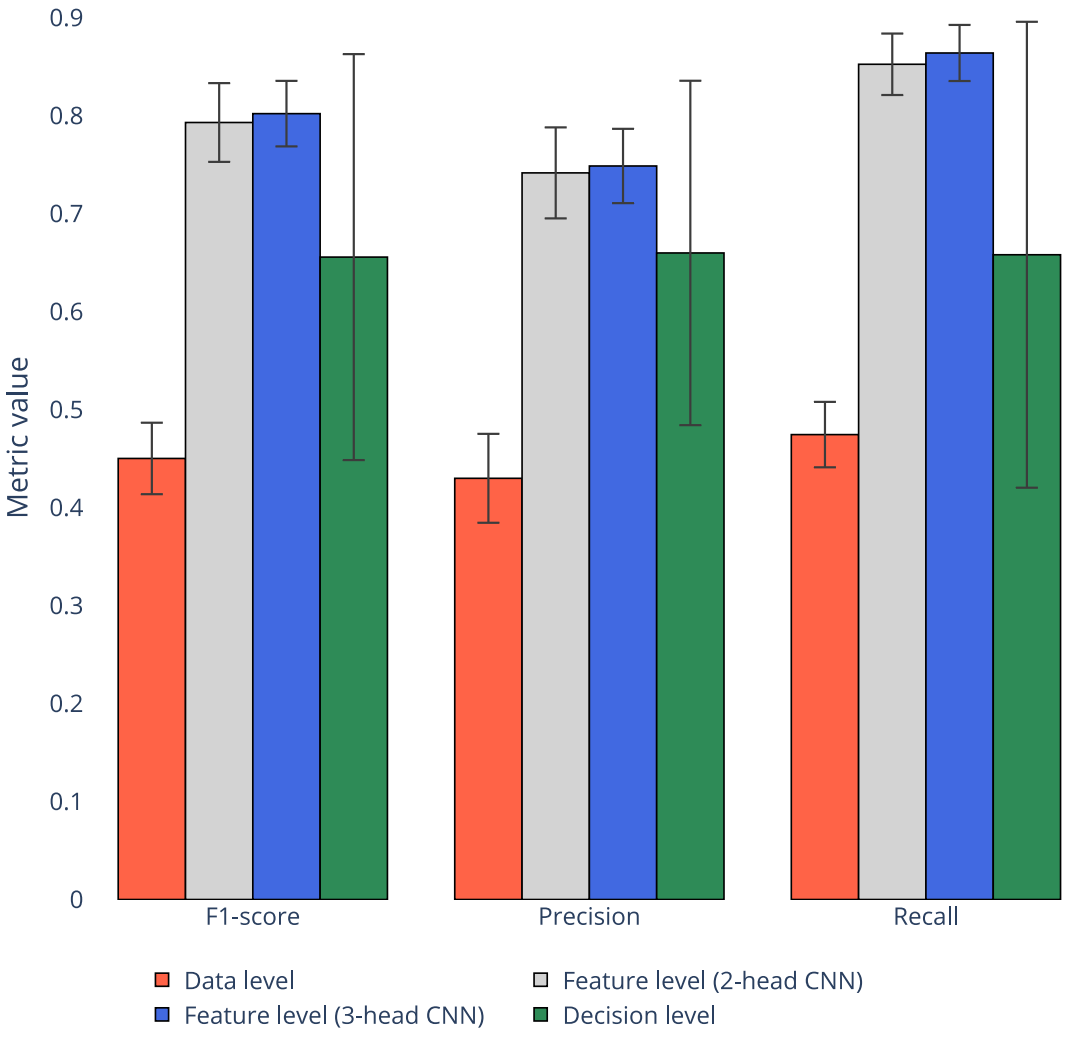

Fig. 8. Comparison of the results obtained by different fusion levels based on the overall F1-score, precision, and recall.

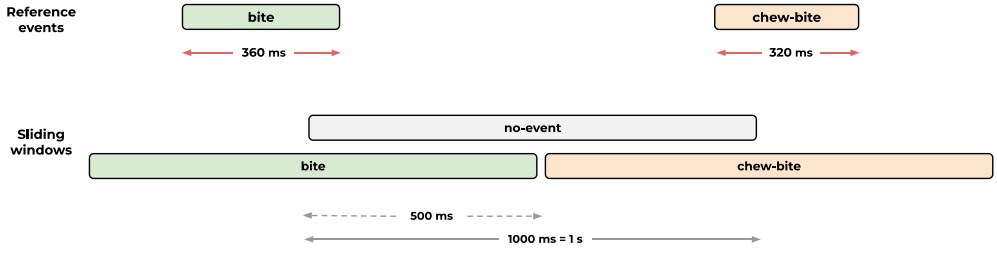

Fig. 9. An example representation for a time window length of 1000 ms = 1-s with two reference events and three extracted windows.

Table 4
Comparison of presented model results using quantisation.

Weights precision	F1-score	Precision	Recall	Error rate
float 32	**0.802 ± 0.033**	**0.749 ± 0.038**	**0.864 ± 0.029**	**0.327 ± 0.05**
float 16	0.791 ± 0.05	0.742 ± 0.054	0.848 ± 0.045	0.335 ± 0.069

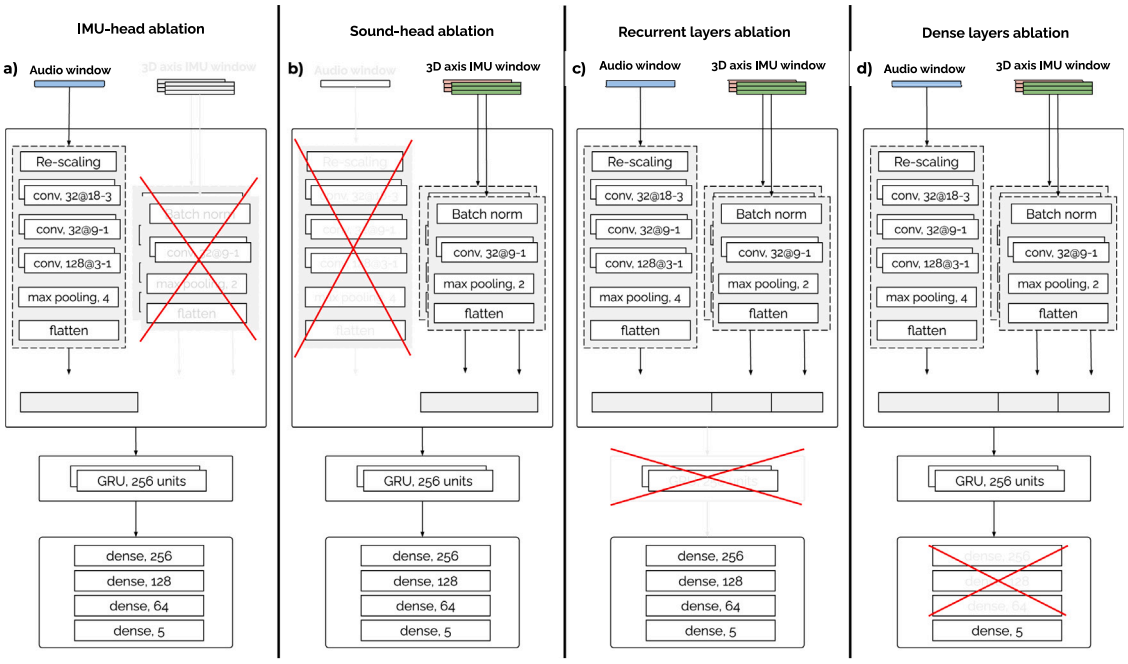

Fig. 10. Different architectures proposed in the ablation study. (a) proposed model with only sound head; (b) proposed model with only IMU head; (c) proposed model with no recurrent block; (d) proposed model with only one dense layer in the last block.

Quantisation is employed to optimise the model by lowering the precision of weights, which reduces memory usage and computational requirements, making the model more suitable for deployment on resource-constrained devices. Results from the exploration of post-training quantisation are presented in Table 4, which helps to analyse their effects on key metrics such as F1-score, precision, recall, and error rate. Those results demonstrate that quantisation using a float 16 precision (instead of the default float 32) achieves significant efficiency gains while maintaining acceptable performance levels.

4.5. Comparison between the proposed model and state-of-the-art methods

The performance of the proposed model (Fig. 3c) was compared against different state-of-the-art methods. Four unimodal models were selected to encompass four combinations, integrating audio and movement signals using both traditional and deep neural network methods:

1. The CBIA is a pattern recognition method that processes acoustic signals to perform event detection using thresholds, feature extraction over the detected event, and then classification using an FNN (Chelotti et al., 2018).
2. The Deep Sound architecture combines convolutional, recurrent, and fully connected layers to recognise (detect and classify) JM events using sliding windows (Ferrero et al., 2023).
3. The traditional approach proposed by Alvarenga et al. (2020) processes motion signals using sliding windows, and a specific feature engineering process is proposed for the classification of short-duration activities in ruminants for each window.
4. The deep architecture proposed by Bloch et al. (2023) consists of CNN and FNN to recognise feeding activities in ruminants using motion input signals.

All selected methods have been trained and validated using the same dataset partitions that were used for the exploration of the fusion level architectures.

The average and SD values for the different validation partitions during the 5-fold CV process are shown in Table 5. It can be seen that for all analysed metrics, the proposed model outperforms all unimodal methods, while the Deep Sound and CBIA models are in second and third performance rank, respectively. Regarding the unimodal approaches, there is a remarkable improvement in acoustic methods (CBIA and Deep Sound) compared to movement-based methods (Alvarenga and Bloch). This acknowledges the previous statement regarding the advantages of sound over inertial signals to recognise JM events. On the other hand, even though the use of deep architectures offers better results in sound processing, the opposite occurs in the case of signals extracted from the IMU.

Different input signal alternatives were evaluated for the model proposed by Bloch et al. (2023), including the use of raw signals, the calculation of magnitude vectors from each signal, and the use of band-pass filters (as described by the authors) as well as their omission. The results obtained in all cases were worse than those reported in Table 5 (where the Hamming filter proposed by the authors was included and all the raw signals were used as input).

As previously mentioned, for movement-signals-based options is it noteworthy that the deep learning models underperform the classic models. This suggests, in conjunction with the results reported by Alvarenga et al. (2020), that a more exhaustive exploration of deep architectures that allow automatically obtaining more representative variables from data could be beneficial.

On the other hand, the results from motion methods are observed to be significantly lower. This seems to indicate a clear difficulty in recognising short-duration events (such as JM events) using these signals, which is aligned with what was previously mentioned in the related work section. It can be established that this type of signal offers an advantage when used in conjunction with audio signals, but their independent use in this problem is insufficient.

4.6. Ablation study and test performance

In order to evaluate the capabilities of each component in the proposed model, four different ablation experiments have been conducted. The architectures explored in those experiments are introduced in Fig. 10 highlighting the differences with the proposed model. Two of them were focused on the input blocks: Fig. 10(a) the proposed model without IMU heads and Fig. 10(b) the proposed model without the sound head. These experiments are important from a practical point

Table 5
Overall results obtained for the multi-head CNN-RNN fusion proposed model and selected state-of-the-art algorithms.

	F1-score	Precision	Recall	Error rate
Alvarenga et al. (2020)	0.251 ± 0.015	0.188 ± 0.015	0.381 ± 0.008	1.977 ± 0.129
Bloch et al. (2023)	0.125 ± 0.009	0.123 ± 0.012	0.127 ± 0.007	1.615 ± 0.067
CBIA	0.606 ± 0.066	0.627 ± 0.063	0.587 ± 0.072	0.499 ± 0.074
Deep Sound	0.704 ± 0.025	0.650 ± 0.030	0.767 ± 0.020	0.453 ± 0.052
Proposed model	**0.802 ± 0.033**	**0.749 ± 0.038**	**0.864 ± 0.029**	**0.327 ± 0.05**

Table 6
Performance of the ablation study including four architectures and the proposed model for cross-validation folds and test set. Inference time refers to calculations to process 1 min of signal. V: Validation. T: Test. PM: Proposed model. b.1 and b.2 reflect the extraction of gyroscope and accelerometer head from the architecture proposed in (b).

		F1-score	Precision	Recall	Error rate	Parameters	FLOPs	Inference time (s)
a	V	0.576	0.542	0.615	0.536	11,678,470	9.3×10^{10}	0.215 ± 0.031
	T	0.686	0.660	0.713	0.388			
b	V	0.155	0.182	0.156	1.144	11,605,214	8.6×10^{9}	0.211 ± 0.014
	T	0.001	0.087	0.001	1.004			
b.1	V	0.011	0.068	0.006	1.044	11,605,214	8.5×10^{9}	0.208 ± 0.008
	T	0.001	0.026	0.001	1.016			
b.2	V	0.165	0.185	0.167	1.150	11,605,214	8.5×10^{9}	0.208 ± 0.008
	T	0.002	0.033	0.001	1.023			
c	V	0.607	0.473	0.851	1.008	298,142	3.7×10^{7}	0.133 ± 0.008
	T	0.574	0.437	0.838	1.146			
d	V	0.738	0.690	0.795	0.427	11,531,998	1.59×10^{11}	0.209 ± 0.009
	T	0.743	0.697	0.795	0.444			
PM	V	0.802	0.749	0.861	0.325	11,704,478	1.59×10^{11}	0.217 ± 0.034
	T	**0.813**	**0.771**	**0.859**	**0.306**			

of view. During the execution of a multimodal system, if one of the inputs is lost or has strong interference, the performance could be severely affected. In this situation, it is often convenient to discard one of the inputs. In the context of this application specifically, this is commonly seen in environments where animals are confined (barn). In these cases, the signal-to-noise ratio of the sound is low due to the noise and reverberations and it is often convenient to discard this data and use only motion data. The execution of experiments with controlled noise in future studies will allow an evaluation of which signal is most convenient to be used in these scenarios, noisy sound or IMU signals.

The remaining two experiments were focused on specific blocks of the original model: Fig. 10(c) the proposed model without recurrent layers (block 2) and Fig. 10(d) the proposed model with only the last dense layer in block 3. These experiments seek to simplify the structure of the proposed model without greatly affecting performance. Simplifying the model can reduce the risk of overfitting and the amount of data needed for the model to achieve good performance.

The results of the ablation study in terms of performance metrics, number of model parameters, floating point operations (FLOPs), and inference time are presented in Table 6, including the average performance on validation folds as well as on the test set. It can be observed that in all cases, the elimination of a specific part from the proposed model worsens the performance, pointing out that all parts play an important role and have an impact on the final architecture.

The worst results were exhibited by option (b), that is, using only motion data. These results were expected because, in the particular case of JM events recognition, sound signals offer more discriminative power than motion signals (Chelotti et al., 2024). This option also shows convergence issues when trying to predict the test set. Furthermore, the concept of option (a) (considering only the sound input) achieves similar results in the test set to those indicated in the CNN-RNN acoustic method (Ferrero et al., 2023). When using only motion data, the gyroscope head (b.2) performs better in general than accelerometer head (b.1) both on validation and test data.

The final dense layers of the FNN are responsible for generating the final output of the model by combining the features obtained in the previous layers. Removing this set of layers reduces the overall performance of the model, as can be seen in results achieved by option (d). This option achieved the best performance of the four ablated models (except for recall where option (c) reported the higher values), but still underperformed the proposed model. Moreover, given the large number of parameters of the proposed model, the removal of all recurrent layers simplifies the model and considerably reduces the risk of overfitting. When removing these layers – option (d) –, the model performance was also highly damaged, thus confirming the importance of the temporal component in this problem. To the best of our knowledge, no specific study confirms the temporal dependence of this phenomenon.

Regarding the difference between the values obtained in validation and test, except for options (b) and (c), some improvements are observed in the test performances. This may be caused because the amount of data with which the models are trained varies substantially, being 25% larger in the test set. It is important to highlight that the test data set includes signals extracted from the same fieldwork, where the animals, equipment and experimental conditions were the same. If any of these conditions vary, the performance of the models may not be the same. This aspect is of special interest and should be studied in the future.

With respect to the inference times in Table 6, ten executions per method were run on the same hardware, and the average and the SD of times to predict 1 min of signal are reported.

In addition, the sum of FLOPs required to process one chunk from input signal (300 ms) is presented in this table. The same methodology reported in Ferrero et al. (2023) was used to calculate these costs. From this comparison, it can be concluded that recurrent layers represent the biggest impact in terms of the processing time and operations of the proposed model, directly affected by the total number of parameters included in the block with RNNs layers. Inference times remained at similar levels without considerable differences between the proposed model and the rest of the options.

5. Conclusions

In this study, a multi-head CNN-RNN was introduced for JM event detection and recognition in grazing cattle. The model includes acoustic

and IMU signals as inputs. The proposed architecture was compared with several different proposals among the three main data-level fusion strategies: data-level, feature-level, and decision-level. Variations in the number of layers, kernels, CNN heads, and kernel sizes were evaluated during the exploration. Additionally, different combinations of input signals were tested.

The results suggest that the proposed model for feature-level fusion is the more appropriate strategy in this context, using an independent CNN head for each input signal, achieving an average micro F1-score of 0.802. The contribution of each part of the model was also assessed and presented in an ablation study. Additionally, the effect of different window sizes was analysed, showing a clear advantage when using a size close to the average duration of the JM events (or even smaller). The proposed model clearly outperformed the state-of-the-art methods by at least 10% (micro F1-score).

This study pioneers research into the effectiveness of information fusion strategies for the detection and recognition of JM events in grazing cattle. The results demonstrate that the use of both sound and motion signals provides a clear advantage over unimodal solutions.

Despite the set of model architectures explored during experimentation, there are other potential changes that could be beneficial. The use of different fusion methods other than those used during experimentation, such as attention layers, will be evaluated in future works.

It is important to note several limitations of the presented model. From a practical point of view, its ability to generalise should be tested across different scenarios, including variations in recording devices, environmental conditions, sensor placement, different herd management, and even its applicability to other ruminants. Additionally, external noises can independently affect both signals, which may affect the model's performance. Future studies should investigate the robustness of the model and explore potential improvements in this aspect. Ethical considerations related to sensor intrusiveness and animal welfare should be carefully evaluated before large-scale adoption. Future research should focus on overcoming these limitations to ensure the feasibility and scalability of AI-driven event detection in grazing cattle.

Regarding technical limitations, the difficulty of obtaining high-quality labelled data sets with a larger amount of data – which is a critical aspect in this context – represents a challenge. The exploration of techniques to help overcome this problem, such as transfer learning or the use of semi-supervised approaches, will be evaluated in future research.

Finally, technical analysis and likely implications of the presented model are not covered in this study. However, one might notice that the information fusion imposes the need for both signals synchronisation since mismatches can degrade the performance of the model. Another aspect related to the use of BGRU layers is the need of a buffer and, consequently, a potential problem for real-time applications. The model structure might increase resource demands in comparison with other approaches, which may limit implementation on edge devices. The use of knowledge distillation or model pruning while retaining performance might be an interesting line of research in this direction.

CRediT authorship contribution statement

Mariano Ferrero: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **José O. Chelotti:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Data curation, Conceptualization. **Luciano S. Martinez-Rau:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Data curation, Conceptualization. **Leandro D. Vignolo:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Martín Pires:** Resources, Investigation, Data curation. **Julio R. Galli:** Writing – review & editing, Resources, Funding acquisition, Data curation. **Leonardo L. Giovanini:** Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **H. Leonardo Rufiner:** Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Funding acquisition, 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.

Acknowledgements

This work has been funded by Universidad Nacional del Litoral, Argentina, CAID 50620190100080LI and 50620190100151LI, Universidad Nacional de Rosario, Argentina, projects 2013-AGR216, 2016-AGR266 and 80020180300053UR, Agencia Santafesina de Ciencia, Tecnología e Innovación (ASACTEI), project IO–2018—00082, Consejo Nacional de Investigaciones Científicas *y* Técnicas (CONICET), project 2017-PUE sinc(i). Authors would like to thank the dedication and perceptive help by Campo Experimental J. Villarino Dairy Farm staff for their assistance and support during the completion of this study. Authors also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan XP GPU used for this research. Finally, our thanks to Constanza Quaglia for her enormous contribution to this work through the development of the Android application used to capture signals during fieldwork.

Data availability

Data will be made available on request.

References

Alvarenga, F.A.P., Borges, I., Oddy, V.H., Dobos, R.C., 2020. Discrimination of biting and chewing behaviour in sheep using a tri-axial accelerometer. Comput. Electron. Agric. 168, 105051.

Andriamandroso, A.L.H., Bindelle, J., Mercatoris, B., Lebeau, F., 2016. A review on the use of sensors to monitor cattle jaw movements and behavior when grazing. Biotechnol. Agron. Soc. Env. Pages 27, 3–286.

Andriamandroso, A.L.H., Lebeau, F., Beckers, Y., Froidmont, E., Dufrasne, I., Heinesch, B., Dumortier, P., Blanchy, G., Blaise, Y., Bindelle, J., 2017. Development of an open-source algorithm based on inertial measurement units (IMU) of a smartphone to detect cattle grass intake and ruminating behaviors. Comput. Electron. Agric. 139, 126–137.

Andriamandroso, A., Lebeau, F., Bindelle, J., 2015. Changes in biting characteristics recorded using the inertial measurement unit of a smartphone reflect differences in sward attributes. In: 7th Conference on Precision Livestock Farming, vol. 28, pp. 3–289.

Aquilani, C., Confessore, A., Bozzi, R., Sirtori, F., Pugliese, C., 2022. Review: Precision livestock farming technologies in pasture-based livestock systems. Animal 16 (1), 100429.

Arablouei, R., Wang, Z., Bishop-Hurley, G.J., Liu, J., 2023. Multimodal sensor data fusion for in-situ classification of animal behavior using accelerometry and GNSS data. Smart Agric. Technol. 4 (100163), 100163.

Aydogmus, O., Bingol, M.C., Boztas, G., Tuncer, T., 2023. An automated voice command classification model based on an attention-deep convolutional neural network for industrial automation system. Eng. Appl. Artif. Intell. 126, 107120.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer New York.

Bloch, V., Frondelius, L., Arcidiacono, C., Mancino, M., Pastell, M., 2023. Development and analysis of a CNN- and Transfer-Learning-Based classification model for automated dairy cow feeding behavior recognition from accelerometer data. Sensors 23 (5).

Bosi, M., Brandenburg, K., Quackenbush, S., Fielder, L., Akagiri, K., Fuchs, H., Dietz, M., 1997. ISO/IEC MPEG-2 advanced audio coding. J. Audio Eng. Soc. 45, 789–814.

Bristow, D.J., Holmes, D.S., 2007. Cortisol levels and anxiety-related behaviors in cattle. Physiol. Behav. 90 (4), 626–628.

Büchel, S., Sundrum, A., 2014. Short communication: Decrease in rumination time as an indicator of the onset of calving. J. Dairy Sci. 97 (5), 3120–3127.

Calamari, L., Soriani, N., Panella, G., Petrera, F., Minuti, A., Trevisi, E., 2014. Rumination time around calving: An early signal to detect cows at greater risk of disease. J. Dairy Sci. 97 (6), 3635–3647.

Chelotti, J.O., Martinez-Rau, L.S., Ferrero, M., Vignolo, L.D., Galli, J.R., Planisich, A.M., Rufiner, H.L., Giovanini, L.L., 2024. Livestock feeding behaviour: A review on automated systems for ruminant monitoring. Biosyst. Eng. 246, 150–177.

Chelotti, J.O., Vanrell, S.R., Galli, J.R., Giovanini, L.L., Rufiner, H.L., 2018. A pattern recognition approach for detecting and classifying jaw movements in grazing cattle. Comput. Electron. Agric. 145, 83–91.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. EMNLP, Association for Computational Linguistics., Doha, Qatar, pp. 1724–1734.

Clark, C.E.F., Lyons, N.A., Millapan, L., Talukder, S., Cronin, G.M., Kerrisk, K.L., Garcia, S.C., 2015. Rumination and activity levels as predictors of calving for dairy cows. Animal 9 (4), 691–695.

Dietterich, T.G., 2000. Ensemble methods in machine learning. In: Multiple Classifier Systems. Springer Berlin Heidelberg., pp. 1–15.

Ferrero, M., Vignolo, L.D., Vanrell, S.R., Martinez-Rau, L.S., Chelotti, J.O., Galli, J.R., Giovanini, L.L., Rufiner, H.L., 2023. A full end-to-end deep approach for detecting and classifying jaw movements from acoustic signals in grazing cattle. Eng. Appl. Artif. Intell. 121, 106016.

Garcia-Ceja, E., Galván-Tejada, C.E., Brena, R., 2018. Multi-view stacking for activity recognition with sound and accelerometer data. Inf. Fusion 40, 45–56.

Giovanetti, V., Decandia, M., Molle, G., Acciaro, M., Mameli, M., Cabiddu, A., Cossu, R., Serra, M.G., Manca, C., Rassu, S.P.G., Dimauro, C., 2017. Automatic classification system for grazing, ruminating and resting behaviour of dairy sheep using a tri-axial accelerometer. Livest. Sci. 196, 42–48.

Greenwood, P.L., Paull, D.R., McNally, J., Kalinowski, T., Ebert, D., Little, B., Smith, D.V., Rahman, A., Valencia, P., Ingham, A.B., Bishop-Hurley, G.J., 2017. Use of sensor-determined behaviours to develop algorithms for pasture intake by individual grazing cattle. Crop. Pasture Sci. 68 (12), 1091.

Hall, D.L., Llinas, J., 1997. An introduction to multisensor data fusion. Proc. IEEE 85 (1), 6–23.

Herskin, M.S., Munksgaard, L., Ladewig, J., 2004. Effects of acute stressors on nociception, adrenocortical responses and behavior of dairy cows. Physiol. Behav. 83 (3), 411–420.

Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., Kingsbury, B., 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process. Mag. 29 (6), 82–97.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Islam, M.M., Nooruddin, S., Karray, F., Muhammad, G., 2023. Multi-level feature fusion for multimodal human activity recognition in internet of healthcare things. Inf. Fusion 94, 17–31.

Kilgour, R.J., 2012. In pursuit of normal: A review of the behaviour of cattle at pasture. Appl. Anim. Behav. Sci. 138 (1), 1–11.

Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization. Int. Conf. Learn. Represent..

Laca, E.A., 2009. Precision livestock production: tools and concepts. Rev. Bras. Zootec. 38, 123–132.

Laca, E.A., Ungar, E.D., Seligman, N.G., Ramey, M.R., Demment, M.W., 1992. An integrated methodology for studying short-term grazing behaviour of cattle. Grass Forage Sci. 47 (1), 81–90.

Laca, E.A., WallisDeVries, M.F., 2000. Acoustic measurement of intake and grazing behaviour of cattle. Grass Forage Sci. 55, 97–104.

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.

Li, Y., Shu, H., Bindelle, J., Xu, B., Zhang, W., Jin, Z., Guo, L., Wang, W., 2022. Classification and analysis of multiple cattle unitary behaviors and movements based on machine learning methods. Anim. (Basel) 12 (9).

Li, G., Xiong, Y., Du, Q., Shi, Z., Gates, R.S., 2021. Classifying ingestive behavior of dairy cows via automatic sound recognition. Sensors 21 (15).

Liu, M., Wu, Y., Li, G., Liu, M., Hu, R., Zou, H., Wang, Z., Peng, Y., 2023. Classification of cow behavior patterns using inertial measurement units and a fully convolutional network model. J. Dairy Sci. 106 (2), 1351–1359.

Martinez-Rau, L.S., Chelotti, J.O., Ferrero, M., Utsumi, S.A., Planisich, A.M., Vignolo, L.D., Giovanini, L.L., Rufiner, H.L., Galli, J.R., 2023. Daylong acoustic recordings of grazing and rumination activities in dairy cows. Sci Data 10 (1), 782.

Martinez-Rau, L.S., Chelotti, J.O., Vanrell, S.R., Galli, J.R., Utsumi, S.A., Planisich, A.M., Rufiner, L.L., 2022. A robust computational approach for jaw movement detection and classification in grazing cattle using acoustic signals. Comput. Electron. Agric. 192, 106569.

Martiskainen, P., Järvinen, M., Skön, J.-P., Tiirikainen, J., Kolehmainen, M., Mononen, J., 2009. Cow behaviour pattern recognition using a three-dimensional accelerometer and support vector machines. Appl. Anim. Behav. Sci. 119 (1–2), 32–38.

Mesaros, A., Heittola, T., Virtanen, T., 2016. Metrics for polyphonic sound event detection. NATO Adv. Sci. Inst. Ser. E Appl. Sci. 6 (6), 162.

Mesaros, A., Heittola, T., Virtanen, T., Plumbley, M.D., 2021. Sound event detection: A tutorial. IEEE Signal Process. Mag. 38 (5), 67–83.

Morgan-Davies, C., Tesnière, G., Gautier, J., Jørgensen, G., González-García, E., Patsios, S., Sossidou, E., Keady, T., McClearn, B., Kenyon, G., Grøva, L., Decandia, L., Halachmi, I., Dwyer, C., 2024. Review: Exploring the use of precision livestock farming for small ruminant welfare management. Animal 18, 101233, Selected keynote lectures of the 74th Annual Meeting of the European Federation of Animal Science (Lyon, France).

Mroueh, Y., Marcheret, E., Goel, V., 2015. Deep multimodal learning for audio-visual speech recognition. In: 2015 IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP, pp. 2130–2134.

Navon, S., Mizrach, A., Hetzroni, A ., Ungar, E.D., 2013. Automatic recognition of jaw movements in free-ranging cattle, goats and sheep, using acoustic monitoring. Biosyst. Eng. 114 (4), 474–483, (Special Issue: Sensing Technologies for Sustainable Agriculture).

Nunes, L., Ampatzidis, Y., Costa, L., Wallau, M., 2021. Horse foraging behavior detection using sound recognition techniques and artificial intelligence. Comput. Electron. Agric. 183, 106080.

Nweke, H.F., Teh, Y.W., Mujtaba, G., Al-garadi, M.A., 2019. Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. Inf. Fusion 46, 147–170.

Oudshoorn, F.W., Cornou, C., Hellwing, A.L.F., Hansen, H.H., Munksgaard, L., Lund, P., Kristensen, T., 2013. Estimation of grass intake on pasture for dairy cows using tightly and loosely mounted di- and tri-axial accelerometers combined with bite count. Comput. Electron. Agric. 99, 227–235.

Paudyal, S., Maunsell, F., Richeson, J., Risco, C., Donovan, D., Pinedo, P., 2018. Rumination time and monitoring of health disorders during early lactation. Animal 12 (7), 1484–1492.

Pavlovic, D., Davison, C., Hamilton, A., Marko, O., Atkinson, R., Michie, C., Crnojević, V., Andonovic, I., Bellekens, X., Tachtatzis, C., 2021. Classification of cattle behaviours using Neck-Mounted Accelerometer-Equipped collars and convolutional neural networks. Sensors 21 (12).

Peng, Y., Kondo, N., Fujiura, T., Suzuki, T., Wulandari, H., Itoyama, E., 2019. Classification of multiple cattle behavior patterns using a recurrent neural network with long short-term memory and inertial measurement units. Comput. Electron. Agric. 157, 247–253.

Qiu, S., Zhao, H., Jiang, N., Wang, Z., Liu, L., An, Y., Zhao, H., Miao, X., Liu, R., Fortino, G., 2022. Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges. Inf. Fusion 80, 241–265.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323 (6088), 533–536.

Schuster, M., Paliwal, K.K., 1997. Bidirectional recurrent neural networks. IEEE Trans. Signal Process. 45 (11), 2673–2681.

Serizel, R., Turpault, N., Shah, A., Salamon, J., 2020. Sound event detection in synthetic domestic environments. In: ICASSP 2020-2020 IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP, IEEE, pp. 86–90.

da Silva Santos, A., de Medeiros, V.W.C., Gonçalves, G.E., 2023. Monitoring and classification of cattle behavior: a survey. Smart Agric. Technol. 3, 100091.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. 45 (4), 427–437.

Son, C.-S., Kang, W.-S., 2023. Multivariate CNN model for human locomotion activity recognition with a wearable exoskeleton robot. Bioeng. (Basel) 10 (9).

Spinsante, S., Angelici, A., Lundström, J., Espinilla, M., Cleland, I., Nugent, C., 2016. A mobile application for easy design and testing of algorithms to monitor physical activity in the workplace. Mob. Inf. Syst..

Tan, T.-H., Chang, Y.-L., Wu, J.-R., Chen, Y.-F., Alkhaleefah, M., 2024. Convolutional neural network with multihead attention for human activity recognition. EEE Internet Things J. 11 11 (2), 3032–3043.

Tani, Y., Yokota, Y., Yayota, M., Ohtani, S., 2013. Automatic recognition and classification of cattle chewing activity by an acoustic monitoring method with a single-axis acceleration sensor. Comput. Electron. Agric. 92, 54–65.

Topaloglu, I., Barua, P.D., Yildiz, A.M., Keles, T., Dogan, S., Baygin, M., Gul, H.F., Tuncer, T., Tan, R.-S., Acharya, U.R., 2023. Explainable attention resnet18-based model for asthma detection using stethoscope lung sounds. Eng. Appl. Artif. Intell. 126, 106887.

Tzirakis, P., Trigeorgis, G., Nicolaou, M., Schuller, B., Zafeiriou, S., 2017. End-to-end multimodal emotion recognition using deep neural networks. IEEE J. Sel. Top. Signal Process..

Ungar, E., Ravid, N., Zada, T., Ben-Moshe, E., Yonatan, R., Baram, H., Genizi, A., 2006. The implications of compound chew–bite jaw movements for bite rate in grazing cattle. Appl. Anim. Behav. Sci. 98 (3–4), 183–195.

Vanrell, S.R., Chelotti, J.O., Bugnon, L.A., Rufiner, H.L., Milone, D.H., Laca, E.A., Galli, J.R., 2020. Audio recordings dataset of grazing jaw movements in dairy cattle. Data Brief 30, 105623.

Venkatesh, S., Moffat, D., Miranda, E.R., 2022. You only hear once: A YOLO-like algorithm for audio segmentation and sound event detection. NATO Adv. Sci. Inst. Ser. E Appl. Sci. 12 (7), 3293.

Wang, K., Wu, P., Cui, H., Xuan, C., Su, H., 2021. Identification and classification for sheep foraging behavior based on acoustic signal and deep learning. Comput. Electron. Agric. 187 (106275), 106275.

Wu, Y., Liu, M., Peng, Z., Liu, M., Wang, M., Peng, Y., 2022. Recognising cattle behaviour with deep residual bidirectional LSTM model using a wearable movement monitoring collar. Collect. FAO Agric. 12 (8), 1237.

Yihan, C., Min, G., Zhiqiang, L., 2021. Sound event detection based on bidirectional temporal convolutional network and gated recurrent unit. In: 2021 20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS). IEEE, pp. pages 445–450.

Zhu, Z., Dai, W., Hu, Y., Li, J., 2020. Speech emotion recognition model based on Bi-GRU and focal loss. Pattern Recognit. 140, 358–365.