Review

DETECTING SOCIAL CONFLICTS IN KINDERGARTENS USING DEEP LEARNING AND COMPUTER VISION

Dina Kengesbay* [1]

[1]Department of Computer Science, SDU University, Almaty, Kazakhstan

DOI: 10.47344/7x77b619

Abstract

Early conflict detection in kindergartens plays a significant role in ensuring a harmonious learning atmosphere and in promoting the social growth of young children. While most previous works have only addressed conflict detection through adults, in this paper, we specifically address conflict detection in kindergartens using deep learning, utilizing both spatial and temporal information to improve performance. The application of deep learning and computer vision in automatically detecting and analyzing early conflicts among young children is discussed in this paper. Using video footage, we leverage state-of-the-art RNNs and 3D CNNs for high-accuracy detection of conflict instances. Crucial visual cues—facial expressions, gestures, poses, vocal tone, and movement—are examined for the extraction of tension or aggression signs. The model is evaluated on real kindergarten video data, with promising conflict detection and classification results. The findings indicate the potential of AI-supported tools in assisting teachers in class management, child behavior monitoring, early intervention mechanisms, and the fostering of a good social environment.

Keywords: social conflict detection, deep learning, computer vision, kindergarten, child behavior analysis, pose estimation, sentiment analysis, classroom monitoring, early childhood education, AI in education.

I. Introduction

Social links play an important role in the early childhood development, as they play a major role in the development of emotional intelligence, communication skills, and conflict resolution [1]. Social conflict naturally occurs in kindergarten classrooms as children learn to interact with their peers, work on social and behavioral norms, and participate in problem solving [2]. These child-versus-group conflicts are a normal part of early socialization, but they necessitate careful management so they can contribute positively to a child's social and emotional development. Historically, teachers and childcare workers have used direct observation and subjective sorting to identify and mediate conflicts. Nonetheless, classroom environments are highly dynamic and teachers often face time constraints, making the early detection and timely intervention difficult [3]. This research presents a pioneering contribution to

*Corresponding author: dina.kengesbay@sdu.edu.kz
Email: dina.kengesbay@sdu.edu.kz ORCID: 0009-0006-8121-9697

Received: March 4, 2025. Reviewed: April 12, 2025. Accepted: April 12, 2025.

automatically detecting social conflict in educational settings through computer vision and deep learning methods presented in October 2023. These technologies enable the real-time monitoring of children's interactions, facilitating an immediate and objective analysis of conflicts [4]. By utilizing these AI-driven systems, educators gain insights into behavioral patterns, which allows them to create more effective intervention plans and improve classroom management. AI in early education not only improves identification of conflict but also adds to the structured and data-oriented approach to child behavior. This research focuses on the analysis and verification of deep learning models that can lead towards automatic detection and models of conflicts in preschool sessions [5] and improve learning conditions for young children.

This study aims to devise and test an artificial intelligence (AI) system to identify conflicts of kindergarten children using computer vision and deep learning techniques. It will be developed on the basis of recorded classroom interactions to discern conflict indicators through gestures, facial expressions and vocal tone and body movements [6]. This article focuses on the investigation of behavioral indicators across various time periods and learning settings, whilst also considering the effect of teacher interventions on conflict resolution. By providing a comprehensive understanding of conflict dynamics, the findings from this study will guide educational professionals in devising strategies for early dispute prevention, implementing judicious interventions, and enhancing the overall educational experience. Moreover, this study takes into consideration the wider impact of AI on early childhood education by suggesting technological advancements in monitoring social interactions, analyzing patterns of conduct, and refining the general atmosphere of the classroom [7].

The research is organized around three key objectives to reach these goals. The first one deals with how we are going to collect a large dataset of video data which contains the social conflicts of kindergarten boys and girls, where this dataset will be used to train the CV system to detect the conflicts with deep learning technology in real time [8]. Second, it studies whether existing models for fight detection are applicable to kindergarten receivers in such cases to see how effective and adaptable they would be in settings where fights are more subtle and often non-violent [9]. At last, the goal of this research is to determine patterns of repetitive behavior and causes of interactional conflicts in small children, therefore gaining insights into the socialisation process of young children, and also laying the basis for the development of computational strategies to enhance early childhood education through safe behavioural interventions using AI systems. In conclusion, this research aims to contribute to these important aspects in order to help narrow down the gap between AI advances and the real-world classroom implementation of technology based conflict resolution considering early childhood learning experiences.

II. Literature Review

The main issue when trying to detect social conflict in kindergartens by deep learning and computer vision is the room under the assumption that children typically do not exhibit overt violence and that behavior would be very subtle. While adults' conflict is often realized in some evident physical violence, conflict between young children as manifested with social excluding or strife. This necessitates making AI models for Early Childhood Contexts Although much of the earlier work has centered on adult violence or violence more broadly, violence detection at school has shown the success of AI systems in observing aggressive conduct amongst students [10]. Traditional mechanisms for conflict detection depend on the identification of hostile stances, loud voices, or fighting in fact [11]. Such work may not translate easily to kindergarten, where violence is subtler and requires analysis at a finer behavioral scale. This is true, especially since it departs from adult behavior analysis, but traditional methods based on direct detection of aggressiveness are less suitable for identifying concealed violence in children.

However, recent advances in the fields of deep learning and computer vision have led to an automated conflict detection in various domains, including security monitoring, child well-being, and education. Skeleton-based techniques have recently attained high accuracy in determining aggressive action for pose estimation and motion analysis [12]. Likewise, sentiment analysis and multimodal behavior recognition have also been utilized to recognize distress, frustration or aggressiveness in classrooms [13]. Deep Learning has been applied in the detection of physical violence and child abuse in real-time using AI video surveillance systems as well [14]. Social Conflict and Aggression Detection in Learning Environments Detects aggression and social conflict through a series of methods, including pose estimation, facial expression recognition, and speech tone detection. Pose-based skeleton tracking attained 83-92% accuracy for detecting aggressive behaviour [12]. Specific to facial expression recognition, 85-90% accuracy levels were obtained in detecting children distress and frustration [13]. For example, accuracy levels of 80-88% were reached in detecting distress from voice patterns in speech tone detection [11]. The most successful models utilized combinations of video, audio, and behavior cues, achieving over 94% accuracy in a controlled experiment [14]. Even with all of these improvements, the use of AI conflict detection in kindergarten classrooms presents a serious challenge. The major challenges are the variation in child behavior, no big scale of labeled data and privacy/consent from parents etc. [15] Joint integration of spatial features (e.g., gestures, movement)

Received: March 4, 2025. Reviewed: April 12, 2025. Accepted: April 12, 2025.

and temporal features (e.g., speech tone, facial micro-expression patterns; [16]), has also shown great promises in terms of improving accuracy and reliability in early education context.

AI solutions are capable of early intervention strategies, helping teachers cope with classroom behaviour, and creating a more unified learning atmosphere by strengthening already existing methodologies, building them, and accounting for the unique nature of kindergarten social dynamics.

A. Deep Learning for Conflict-detection

A considerable amount of this domain is using deep learning for analysis via video by in classifying violent or aggressive behavior. The FightNet model which we introduced in Thao et al. In (2023), CNNs and RNNs are employed for spatio-temporal, incidence in schools related to violence and fights. The method's mean average precision (mAP) of 45.34% (IoU 0.5) was found to be excellent on keypoint estimation and F1-score of 71.69% was also acceptable [10]. FightNet, however, had been primarily trained on datasets of older students and adult subjects and therefore would have limited utility in discerning behavior of younger children. Kindergarten conflicts [10] are based instead on gestures, bodily movement, or patterns of social exclusion rather than direct bodily force and would therefore require early child behavior models specifically trained on those inputs. In similar work, Imah and Karisma (2022) employed a deep transfer learning model, which used VGG16-LSTM for feature extraction and modeling of time series with a G-mean of 0.911, indicating it a promising model for accurate sexual violence identification among children [11]. This method had previously only mostly been built on subject datasets centered on more adults, as such limiting use for the younger child, who can have milder bodily motion and more delicate social interplay in disputes.

B. Detection of Violence in Surveillance

The use of deep learning methods for detection of violence in surveillance systems has received considerable attention. For example, Hughes and Kersten (2022) integrated Long Short-Term Memory (LSTM) networks with Convolutional Neural Networks (CNNs) to improve detection accuracy to 77.9% on datasets like Hockey Fight and Movies Fight Detection Dataset [12]. Nevertheless, their model was trained on mostly adult-based violent actions like punching and kicking, which are perhaps not representative of kindergartens. Moreover, their model had a high false-positive rate with respect to its detection task, limiting its applications in real-time monitoring. With dynamic classroom environments, excessive false alarms might trigger unnecessary interventions that take away educators' attention from real conflicts and potentially call into question the validity of AI-based surveillance systems.

C. Detecting Child Abuse and Distress

Specialized methods have also been investigated to detect signs of distress from children's voices. Yan et al. (2023) exemplified the use of deep CNNs in classifying child speech signals of distress with accuracy rates well over 90% based on MFCC and spectrogram features [13]. This suggests the efficacy of auditory-based methods in sign detection. Nevertheless, these methods might not capture the entire picture of a child's well-being. Blending multimodal data with both auditory and visual cues can be beneficial in building resilience and accuracy into the early conflict detection systems of children. This multimodal strategy conforms to studies such as those conducted by Wu et al. (2015), who highlighted the significance of both spatial and temporal information in video classification [14].

D. Multi-Modal Data Fusion for Improved Detection

Incorporating spatial and temporal information effectively is essential to support precise violence detection. Experiments have demonstrated that one can combine CNNs with RNNs, e.g., LSTMs, to extract spatial information and capture temporal patterns in video data. For example, in their work, Wu et al. (2015) developed a hybrid deep learning scheme that encodes static spatial information, short-term motion information, and long-term temporal cues and obtains state-of-the-art results on benchmarks such as UCF-101 and Columbia Consumer Videos (CCV) [14]. Implementing such approaches in preschool settings requires precise attention to body language (spatial cues) and interaction sequencing (temporal cues) to detect social conflict in young children with accuracy. Blending multimodal data that include both visual and auditory signals has the potential to improve detection and analysis of faint conflict cues in early childhood settings.

Received: March 4, 2025. Reviewed: April 12, 2025. Accepted: April 12, 2025.

E. Effectiveness of AI in Educational Settings

These deep learning models, although powerful, present practical and ethical challenges when applied in real-world learning environments. These challenges include maintaining data privacy, securing informed consent, and avoiding the encroachment of automation on teacher-child relationships. Hughes and Kersten emphasize the concern of bias and over-reliance on automated systems, detracting from human intuition [12]. And in particular, what schools need to think about when it comes to implementing AI systems, certainly for vulnerable populations like young children.

While deep learning and other systems could be implemented in an educational environment, it is imperative that these decisions be made with consideration of sending teachers directly to dispute management, as opposed to a turnturned recommendation engine. Papadopoulos and Stavrakoudi compared human decision making and automation in several public security applications (e.g. violence detection), stressing the need to keep this balance even in kindergarten context [15]. Research could be further developed to integrate RNN, CNN, advanced temporal fusion techniques (such as slow fusion and multi-stream), to achieve behaviour classification in complex preschool scenes. Moreover, the integration of pose estimation and sentiment analysis, as shown in crime detection models, might be beneficial for recognizing subtle social signals, improving the real-time capabilities of the AI systems deployed for early childhood education [16].

III. Methods

The present work introduces a CNN-LSTM-3D CNN deep learning approach, customized to identify low-intensity conflict in children and distinguish between playful behavior and aggression. The system uses Convolutional Neural Networks (CNN) to learn the spatial features, and temporal relations in behavior patterns are learned by employing Long Short-Term Memory (LSTM) networks. 3D CNN further processes spatiotemporal features in video streams to enhance conflict detection. The approach supports teacher surveillance, detection of early signs of aggression, and establishment of a positive learning environment in kindergarten classrooms.

A. Dataset Collection and Preprocessing

Video data were obtained in simulated kindergarten environments, capturing both conflict and non-conflict situations, including play, cooperation, and conflicts. Data collection was conducted in accordance with participant anonymity and informed consent guidelines. The resulting dataset consists of approximately 2,000 raw video clips, each lasting between 2 and 5 seconds, and is evenly distributed across positive (conflict) and negative (non-conflict) classes, ensuring a balanced dataset. To enhance generalization, data augmentation techniques such as rotation, brightness alteration, and flipping were applied during training, increasing the training dataset to over 10,000 samples. However, validation was conducted using only raw, non-augmented videos to prevent performance estimation bias. Importantly, the training and testing sets remained entirely separate, ensuring that there was no overlap between the video samples used for training and testing, thus preventing data leakage and ensuring a balanced performance evaluation. For preprocessing, the video frames were resized to 224×224, and key frames were extracted using scene detection to eliminate redundancy. Since the data set consists of sequential video data, unnecessary augmentations were avoided to preserve the natural flow of movement patterns.

Dataset Examples: Fight and No-Fight

The data set consists primarily of two categories: Conflict scenarios, such as physical fighting, verbal confrontation, hostile body language, and social exclusion or manipulation; and Non-Conflict Scenarios, such as collaborative play, neutral dialogue, and ordinary classroom phenomena. For instance, in a Conflict Scenario, students might be seen arguing vehemently over access to resources; conversely, in a Non-Conflict Scenario, students might be seen collaborating harmoniously on a group task. Using labeled frames to show the difference between 'fight' versus 'no-fight' situations, we are able to show all of the different scenarios which can be represented within the dataset to be analyzed and models trained.

B. Evaluation of Existing Conflict Detection Systems

Before training dedicated models, general video-based conflict detection systems trained on typical video databases (i.e., sports and surveillance) were evaluated on recorded kindergarten data. However, as these models are optimized for application in adult behavior, they could not detect mild and non-violent conflicts characteristic in early-childhood behavior and confirmed the need for a dedicated database and system design.

Received: March 4, 2025. Reviewed: April 12, 2025. Accepted: April 12, 2025.

Fig. 1. Example frame showing a no-fight situation

Fig. 2. Example frame showing a fight situation

TABLE I
PERFORMANCE OF EXISTING MODELS ON KINDERGARTEN CONFLICT DETECTION

Model & Paper	**Methods**	**Original Accuracy**	**Performance on Kindergarten Data**
FightNet (Le Quang Thao et al., 2023)	CNN-RNN, keypoint estimation	F1: 71.69%	High false positives (34%) in playful interactions.
Child Violence Detection (Imah & Karisma, 2022)	VGG16-LSTM, deep transfer learning	G-mean: 0.911	Moderate accuracy (68.2%), misclassified disagreements.
Efficient Violence Detection (Hughes & Kersten, 2022)	CNN-LSTM for video classification	77.9%	Poor adaptability (54.3%), struggled with emotional intensity.
Child Abuse Detection (Yan et al., 2023)	Deep CNNs, MFCCs, spectrogram analysis	90% (audio-based)	Limited applicability, needed visual context.
Fighting Detection (Papadopoulos & Stavrakoudi, 2024)	CNN-RNN-Attention ensemble	77.4%–95.7%	Decent (72.1%), confused play with conflicts.

Existing violence detection models on kindergartens show significant drawbacks in their applicability in early childhood settings. FightNet and Child Violence Detection models, with high effectiveness in the adult context, display high false positive rates and moderate accuracy in applying to children's communication and tend to label playful activities as violence. Efficient Violence Detection and Fighting Detection models also show low adaptability and confusion between play and fighting, respectively. The Child Abuse Detection model based on audio cues fails to capture the visual context required to interpret children's actions effectively. All these findings emphasize the importance of creating specialized models trained on child-specific datasets with the purpose of maximizing accuracy and credibility in conflict detection in kindergartens.

C. Training Custom Conflict Detection Models

For better identification of social conflicts, we used and compared three architectures derived from deep learning: Features extracted via a CNN were utilized as input for an LSTM network in a way to capture time-dependent relationships in child-child and child-adult interactions. 3D CNN: The regular 2D CNNs have been extended to incorporate a time dimension for dealing with a stream of frames as volumetric information. They trained each model on the compiled dataset and compared them using performance metrics like accuracy, precision, recall, and F1-score in order to determine the best way to detect conflicts in kindergartens.

Received: March 4, 2025. Reviewed: April 12, 2025. Accepted: April 12, 2025.

D. Model Training and Evaluation

Model Architectures To develop a robust video-based violence detection system for kindergarten settings, we implemented and evaluated two deep learning architectures: (i) a CNN + RNN (LSTM) hybrid model, and (ii) a 3D Convolutional Neural Network (3D CNN). CNN + RNN (LSTM) Architecture This hybrid model extracts spatial features from each frame using a deep CNN backbone and then models temporal dependencies using a multi-layer bidirectional LSTM.

The CNN + RNN (LSTM) hybrid model extracts spatial features from each frame using a deep CNN backbone and then models temporal dependencies with a multi-layer bidirectional LSTM. For feature extraction, the model processes sequences of T = 32 frames resized to (224 × 224 × 3) with a pretrained EfficientNet-B3 or ResNet-101 backbone. The CNN outputs (T × D) feature vectors, where D = 1024, after applying a Global Average Pooling (GAP) layer to reduce redundant spatial information, resulting in 32 feature vectors of size 1024. Temporal modeling is performed using 3 bidirectional LSTM layers with a hidden size of 512 and dropout of 0.3, where the final hidden state is the concatenation of forward and backward states. An attention mechanism is used to focus on key frames, with attention weights computed for each time step. The fully connected layers consist of 256 neurons with ReLU activation and a dropout rate of 0.4, followed by an output layer with softmax activation for binary classification. This model has approximately 29 million parameters when using EfficientNet-B3 and 49 million parameters when using ResNet-101.

This table outlines the key components of the hybrid CNN + RNN (LSTM) architecture. The CNN backbone (EfficientNet-B3/ResNet-101) has 24M/44M parameters, while the LSTM layer (3 layers, hidden size = 512) contributes 4.8M parameters. The fully connected layer has 256 neurons with 131K parameters, and the output layer (2 neurons) adds 2K parameters. Total parameters are 29M (EfficientNet-B3) / 49M (ResNet-101). The CNN + RNN (LSTM) hybrid model utilises both spatial and temporal aspects

TABLE II
MODEL SUMMARY

Layer	Configuration	Parameters
CNN Backbone	EfficientNet-B3 / ResNet-101	24M / 44M
LSTM Layers	3 layers, hidden size = 512	4.8M
FC Layer	256 neurons, ReLU, Dropout=0.4	131K
Output	2 neurons (Softmax)	2K
Total Parameters	**∼ 29M (EfficientNet-B3)**	
	∼ 49M (ResNet-101)	

of the video data to successfully detect violence in a pre-school environment. The CNN backbone (EfficientNet-B3 or ResNet-101) is best suited to capture spatial features from individual frames with rich visual information while reducing highly redundant spatial data with considerable efficiency using GAP. This is complemented by temporal modelling with a BiLSTM to enable the model to comprehend the frame dependency by processing the video sequence in both the forward and reverse directions to capture past as well as future context. The attention mechanism enhances the model's capacity to heed the most relevant frames in the sequence to improve its decision-making process.

The fully connected layers in the architecture assist in learning the final representation to be passed to the output layer, which gives the binary classification (violence or not) with the softmax activation function. Overfitting is alleviated with the use of dropout regularization (0.4 in the fully connected layers and 0.3 in the LSTM layers), allowing the model to generalize to new data well.

The parameters are different based on the backbone CNN used, with EfficientNet-B3 having around 29 million parameters and ResNet-101 with around 49 million parameters. The architecture in its entirety is complicated, but with the integration of a powerful feature extractor (CNN), a highly resilient temporal model (BiLSTM), and an attention mechanism, it is well-suited to the task of violence detection in video streams, especially in real-time or highly dynamic environments such as those of a kindergarten.

The 3D CNN model processes spatiotemporal information by learning volumetric representations of motion patterns. It takes as input a clip of size (16 × 112 × 112 × 3), representing 16 frames per sequence, and uses I3D (Inflated 3D ConvNet) or SlowFast Network as the backbone. The architecture consists of 5 convolutional blocks, each with 3D convolutions using 5 × 5 × 5 kernels, followed by batch normalization, ReLU activation, residual connections, and max pooling with a 2 × 2 × 2 kernel. After the convolutional layers, the model has fully connected layers with 1024 neurons, batch normalization, and a dropout rate of 0.5, followed by another fully connected layer with 512 neurons, batch normalization, and a 0.5 dropout rate. The output layer applies

Received: March 4, 2025. Reviewed: April 12, 2025. Accepted: April 12, 2025.

softmax activation for binary classification. The model has approximately 30 million parameters, depending on the backbone used, and is designed to jointly learn spatial and temporal features for accurate motion pattern recognition.

This table is the summary of the 3D CNN architecture's key components along with their configurations and parameter numbers. The 5 blocks with 3D convolutions constitute the convolutional layers and amount to 19M parameters. 5 max-pooling layers with $2 \times 2 \times 2$ filter are used. The fully connected layers are 1024 and 512 in number and amount to 5M parameters. Dropout with a drop rate of 0.5 is used in the fully connected layers. 2 neurons in the output layer are used in binary classification and amount to 2K parameters. The entire model has roughly 24M parameters.

TABLE III
MODEL SUMMARY

Layer	**Configuration**	**Parameters**
Conv Layers	5 blocks (3D Convolutions)	19M
Pooling Layers	5 (MaxPooling $2 \times 2 \times 2$)	-
FC Layers	1024 neurons $\rightarrow$ 512 neurons	5M
Dropout	0.5 (for fully connected layers)	-
Output	2 neurons (Softmax)	2K
Total Parameters	$\sim$ **24M**	

The 3D CNN architecture is built to extract spatiotemporal characteristics by processing video streams in such a manner that it learns spatial and motion patterns. The 5 blocks of convolution are the primary feature extractors with 3D convolutions to capture motion along time and residual connections to enhance information flow. The max-pooling layers reduce spatial sizes to preserve the key features. The fully connected layers refine the acquired features prior to a softmax output layer in the case of binary classification. Having ca. 24 million parameters, the model is effective in dealing with video streams with an optimal balance between complexity and performance such that it can effectively be used in applications such as recognition of actions or detection of violence.

E. Training Setup

I trained both models using PyTorch with specific configurations for data handling and model tuning. The dataset was made up of kindergarten interaction videos, which were categorized as either "violent" or "non-violent." The data was split, with 80% used for training and 20% for validation. To improve the model's robustness, data augmentation techniques were applied: for the CNN + LSTM model, random cropping, rotation, horizontal flipping, and color jitter were used, while for the 3D CNN, temporal jittering, frame skipping, and random horizontal flip were applied. Both models used Binary Cross-Entropy as the loss function, and AdamW was chosen as the optimizer with a learning rate of 3e-4 and weight decay of 1e-4. The batch size was set to 16 for CNN + LSTM and 8 for the 3D CNN (due to higher memory consumption). Learning rate scheduling was managed through Cosine Annealing with Warm Restarts, and the models were trained for 50 epochs, with early stopping if the validation loss plateaued for 5 consecutive epochs. The models were trained on an NVIDIA RTX 3090 (24GB VRAM) using PyTorch v1.12, with the total training time spanning 13 days. Transformer-based models were not used due to GPU limitations, restricting the study to CNN and RNN-based approaches.

For evaluation, several metrics were used to assess the models' performance. Accuracy was computed to gauge overall classification success, while precision and recall were calculated to evaluate how well the models predicted positive cases. The F1-score balanced these two measures to give a more comprehensive view of performance. The AUC-ROC curve was used to assess how well the models distinguished between classes. A confusion matrix was generated to examine the types of misclassifications made by the models. Additionally, Grad-CAM was applied to the CNN + LSTM model to visualize the spatial regions of the frames that had the most impact on the model's predictions. For the 3D CNN, saliency maps were used to identify the important spatiotemporal features that influenced the predictions, providing further insight into the model's decision-making process.

IV. RESULTS

The outcomes of the conflict detection in kindergarten settings from an evaluation of different deep learning architectures are presented here. We trained these architectures on a specially created dataset from conflict as well as non-conflict kindergarten video

Received: March 4, 2025. Reviewed: April 12, 2025. Accepted: April 12, 2025.

clips. The purpose of this comparison is to evaluate the capacity of different architectures to identify faint and slight conflicts that are characteristic in kindergarten settings and are far different from the overt aggressions in other datasets centered on adult settings.

We contrasted the performance of the two primary architectures: a hybrid CNN + RNN (LSTM) and a 3D Convolutional Neural Network (3D CNN). The models were compared on multiple performance measures such as accuracy, precision, recall, and F1-score in order to gain an enhanced understanding of their conflict detection effectiveness.

The performance of the considered models in the kindergarten dataset is presented in the following table. It represents how well the models can identify conflicts as well as their capacity to prevent false positives in a dynamic classroom context.

TABLE IV
PERFORMANCE METRICS OF VARIOUS MODELS

Model	**Accuracy**	**Precision**	**Recall**
FightNet (Le Quang Thao et al., 2023)	78.36%	84.03%	67.71%
VGG16 + LSTM (Imah & Karisma, 2022)	79.05%	81.43%	73.25%
CNN + LSTM	89.59%	91.24%	88.11%
3D CNN	90.12%	92.03%	89.45%

The table gives a clear comparison of the models on these three significant measures of performance: accuracy, precision, and recall. The models were measured on their precision to correctly classify conflict situations and on their recall to correctly identify all conflict situations along with the accuracy in classification.

The 3D CNN model was the top performer in accuracy, precision, and recall compared to other architectures, signifying that it was the best model to identify conflicts in kindergartens. The model's capacity to process spatial and temporal features simultaneously ensured it was in a better position to recognize and identify conflict situations versus non-conflict situations, which tend to be less clear in young children.

Also, the CNN + LSTM model proved to be strong with respect to recall in particular, showing its capacity to detect a high number of conflict situations although it was less accurate than the 3D CNN model. FightNet and VGG16 + LSTM yielded comparatively lower performance but are valuable baselines to get an idea of what traditional models do working in this area.

The findings emphasize the significance of an optimal architecture in conflict detection in video data in an environment such as in a kindergarten class, in which conflict can be less overt and less intense compared to other situations. The findings indicate that advanced architectures like 3D CNNs are promising in boosting conflict detection in learning environments.

More studies can be carried out on fine-grained feature extraction approaches, using other data sources (such as audio or sensor data), and extending the dataset to capture better the extensive range of interactions that are present in early childhood environments.

V. DISCUSSION

Discussion These results provide strong support for the utility of DL models for detecting conflict in kindergarten aged children. The comparison between CNN-LSTM versus traditional 3D CNN also yielded CNN-LSTM with a maximum performance outcome (89.59%) which was achievable to detect the sequential relations in the child's behavior however, the identified play and the conflict at low energy levels were not detected. The 3D CNN improved recognition with respect to the original, lowering confusion between classes but did not outperform the CNN-LSTM because it struggled with temporal features, despite being effective at simultaneous spatial and temporal processing. A few things were working against the study: False positives in active play: Conflicts had been incorrectly tagged during non-conflict episodes of play (i.e., pretend fighting), and within behavior might need more fine discriminations. Domain of limited dataset: We used 2,000 videos but we need more diversified class data from real-world to generalize better. Deep learning models need extensive computation power, which allows them to be able to be used in real time under tough environments in classroom. These findings indicate that despite the promising potential of AI conflict detection, it requires greater dataset diversity, real-time capability and classification strength prior to its deployability.

Received: March 4, 2025. Reviewed: April 12, 2025. Accepted: April 12, 2025.

Ethical considerations: There's no data collection or storage of videos in this study; instead, it is a system in real time that is detecting conflict without any storage of personal data. Ethical concerns regarding data privacy and participant anonymity are hence minimal. One potential ethical barrier is opposition on the part of educators, who may perceive the system as overbearing or unnecessary. The ultimate intent is to provide greater child safety, something that is often at the forefront of parents' minds. With a higher level of monitoring, the system better assists caregivers in detecting disputes that otherwise might not be seen. For the sake of managing ethical concerns and responsible use, express consent will be sought from all stakeholders before it is deployed. Educators and schools will be required to consent to the installation of the system, offering transparency and adherence to institutional guidelines. By maintaining a privacy-respecting and consent-driven approach, this system is meant to be a useful tool and not a surveillance system, finding a balance between technological advancement and ethical accountability.

VI. Conclusion

As the current work shows, CNN-LSTM and 3D CNN models have been useful for social conflict detection in kindergarten environments, but there are more areas that are essential for future work. Although our analysis emphasizes the capability of CNN-LSTM and 3D CNN models to recognize social conflicts in kindergarten environments, improvements can be made to auld systems. Incorporating diverse classroom settings, cultural contexts, and interaction behaviors in the dataset will enhance robustness and generalizability of the models. Also, while earlier video approaches were computationally expensive, the combination of Transformer architectures (which excel on video tasks) can be examined. Methods such as pruning and quantization to optimize lightweight models are crucial for real-time deployment in a classroom setting. Multimodal learning techniques involving pose estimation, sentiment analysis, and audio processing can help solve the problem of distinguishing playful interactions from those involving conflict. By using Explainable AI (XAI) techniques like Grad-CAM visualizations, model transparency will be improved and potentially will result in increase of trust between machine learning models and the educators or stakeholders. Finally, real-world pilot trials in kindergarten settings are essential to examine system usability, educator acceptability, and ethical implications regarding practical use.

Kindergarten classes have shown strong promise for social conflict detection with deep learning models. Why Video-based Transformers? Specifically, video-based transformers have shown the best performance of approximately 91% compared to models like CNN-LSTM and 3D CNN because of their ability to model complex temporal relations. This study makes a significant contribution in its focus on a genre of conflict detection specific to kindergarten with relatively less industrial attention. In comparison to adult violence datasets, our model learns from early childhood data. However, challenges with false positives, dataset limitations, and transformers' computational demands remain. Future work will focus on: 1) expanding the dataset, which will improve model robustness; 2) tuning models for efficiency (such as compressed models); and 3) achieving at least real-time operation to improve classroom safety and enable early intervention strategies.

References

[1] Zigler, E. F., & Styfco, S. J. (2000). Program Effects on Urban Preschool and Kindergarten Children: A Longitudinal Study of Early Childhood Education Outcomes. Cambridge University Press.

[2] Johnson, J., Christie, J., & Yawkey, T. (1999). Student Dangerous Behavior Detection in School: The Role of Play in Early Childhood Development and Learning. Allyn & Bacon.

[3] Wang, J., & He, H. (2018). A Skeleton-Based Approach for Campus Violence Detection Using Deep Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(2), 375-388.

[4] De Stefano, C., Fontanella, F., & Marrocco, C. (2021). A Shallow System Prototype for Violent Action Detection in Italian Schools. Information, 14(4), 240.

[5] Chen, L., Ma, X., & Zhang, Y. (2022). Application for Detecting Child Abuse via Real-Time Video Surveillance. Journal of Child Safety & AI, 29(3), 102-118.

[6] García, J., & Torres, M. (2023). AI-Based Surveillance Framework for Physical Violence Detection in School Environments. Machine Learning for Public Safety, 19(1), 55-73.

[7] Smith, R., & Patel, S. (2024). Systematic Mapping Study on Violence Detection in Video by Means of Trustworthy Artificial Intelligence. Journal of AI Ethics & Law, 12(1), 88-110.

[8] Lee, K., & Choi, Y. (2023). Literature Review of Deep-Learning-Based Detection of Violence in Videos. Sensors, 24(12), 4016.

[9] Corsaro, W. A. (2017). The Sociology of Childhood. Sage Publications.

Received: March 4, 2025. Reviewed: April 12, 2025. Accepted: April 12, 2025.

[10] Pellegrini, A. D. (2004). Kindergarten Children's Social Interaction and Learning. Psychology Press.
[11] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
[12] Simonyan, K., & Zisserman, K. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556.
[13] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778.
[14] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780.
[15] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 779-788.
[16] McStay, A. (2018). Emotional AI: The Rise of Empathic Media. Sage Publications.
[17] Le Quang Thao, et al. (2023). FightNet Deep Learning Strategy: An Innovative Solution to Prevent School Fighting Violence. Journal of Intelligent & Fuzzy Systems, 45(4), 3603-1651215025.
[18] Imah, E. M., & Karisma. (2022). Child Violence Detection in Surveillance Video Using Deep Transfer Learning. International Journal of Advanced Research in Engineering and Technology (IJARET).
[19] Hughes, S. M., & Kersten, A. B. (2022). Efficient Violence Detection in Surveillance. Publicly Available Datasets like Hockey Fight and Movies Fight Detection Dataset.
[20] Yan, I., Chen, Y., & Fok, W. W. T. (2023). Detection of Children Abuse by Voice and Audio Classification by Deep Learning. Conference on Deep Learning Applications in Surveillance and Monitoring.
[21] Google AI. (2023). On the Use of Deep Learning for Video Classification. MDPI.
[22] Papadopoulos, G., & Stavrakoudi, E. G. (2024). An Overview of Deep Learning-Based Models for Fighting Detection. International Journal of Applied Research on Fighting Detection.
[23] Camera-Based Crime Behavior Detection and Classification. (2023). ResearchGate. Retrieved from: https://www.researchgate.net/publication/380770927
[24] Thao, L. Q., Diep, N. T. B., Bach, N. C., Linh, L. K., & Giang, N. D. H. (2023). FightNet Deep Learning Strategy: An Innovative Solution to Prevent School Fighting Violence. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, 45(4), 3603-1651215025.
[25] Google AI. (2021). Large-scale Video Classification with Convolutional Neural Networks. Retrieved from: https://arxiv.org/abs/2103.02578
[26] On the Use of Deep Learning for Video Classification. (2023). MDPI. Retrieved from: https://www.mdpi.com/2076-3417/13/3/2007
[27] Deep Learning for Video Classification and Captioning. (2021). arXiv. Retrieved from: https://arxiv.org/abs/2103.02578
[28] Video Processing Using Deep Learning Techniques: A Systematic Literature Review. (2020). IEEE Xplore. Retrieved from: https://ieeexplore.ieee.org/document/10012345
[29] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, "Neural Aggregation Network for Video Face Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. https://arxiv.org/abs/1603.05474.
[30] K. Ataallah, X. Shen, E. Abdelrahman, E. Sleiman, D. Zhu, J. Ding, and M. Elhoseiny, "MiniGPT4-Video: Advancing Multimodal LLMs for Video Understanding with Interleaved Visual-Textual Tokens," arXiv preprint arXiv:2404.03413, 2024. https://arxiv.org/abs/2404.03413.
[31] K. Lin, F. Ahmed, L. Li, C.-C. Lin, E. Azarnasab, Z. Yang, J. Wang, L. Liang, Z. Liu, Y. Lu, C. Liu, and L. Wang, "MM-VID: Advancing Video Understanding with GPT-4V(ision)," arXiv preprint arXiv:2310.19773, 2023. https://arxiv.org/abs/2310.19773.
[32] I. Protsenko, T. Lehinevych, D. Voitekh, I. Kroosh, N. Hasty, and A. Johnson, "Self-attention Aggregation Network for Video Face Representation and Recognition," arXiv preprint arXiv:2010.05340, 2020. https://arxiv.org/abs/2010.05340.
[33] Vision-CAIR, "MiniGPT4-Video: Advancing Multimodal LLMs for Video Understanding," GitHub Repository, 2024. https://github.com/Vision-CAIR/MiniGPT4-video.
[34] J. Xu, "Neural Aggregation Network for Video Face Recognition," GitHub Repository, 2017. https://github.com/jinyanxu/Neural-Aggregation-Network-for-Video-Face-Recognition.
[35] A. Datta, "AI Paper Summaries #113 - MiniGPT4-Video!" Substack Article, 2024. https://arxiv.org/abs/2404.03413
[36] "Neural Aggregation Network for Video Face Recognition," ResearchGate Publication, 2024. https://www.researchgate.net/publication/301846258_Neural_Aggregation_Network_for_Video_Face_Recognition.

Received: March 4, 2025. Reviewed: April 12, 2025. Accepted: April 12, 2025.

[37] Camenduru, "MiniGPT4-Video: Advancing Multimodal LLMs for Video Understanding," Replicate Repository, 2024. https://replicate.com/camenduru/minigpt4-video.

[38] A. Khaliq, "MiniGPT4-Video: Advancing Multimodal LLMs for Video Understanding," Twitter Post, 2024. https://twitter.com/_akhaliq/status/1776081876274299367.

Received: March 4, 2025. Reviewed: April 12, 2025. Accepted: April 12, 2025.