Ministry of Science and Higher Education of the Republic of Kazakhstan
SDU University

Dina Kengesbay

Detecting social conflicts in kindergartens using deep learning and computer vision

THESIS
Presented in Partial Fulfilment for the

Degree of Master of Technical Science in Computer Science
(degree code: 7M06102)
Department of Computer Science
Faculty of Engineering and Natural Sciences

Supervisor: **PhD Birzhan Moldagaliyev**
Kaskelen, June 2025

SDU University
Faculty of Engineering and Natural Sciences
Department of Computer Science

Dean of Faculty of Engineering and Natural Sciences

Assistant Professor, PhD Akhmedov Ramis

« 13 » 06 2025

Topic of the thesis:

Detecting social conflicts in kindergartens using deep learning and computer vision

Thesis submitted as part of the requirements for the award of the MSc in "7M06102 - Computer Science", SDU University

Head of Department	Zhanar Mukash
Academic Supervisor	Birzhan Moldagaliyev
Master student	Dina Kengesbay

Kaskelen, 2025

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Dina Kengesbay

June 2025

Acknowledgements

I want to sincerely thank my supervisor, Birzhan Moldagaliyev, for his support and guidance throughout this research. His advice, encouragement, and patience made a big difference and helped me move forward when things were difficult.

I'm also very grateful to all the teachers who taught me during my Master's years. Each of them contributed to my growth, both academically and personally, and I truly appreciate everything I learned from them.

Finally, thanks to everyone who supported me along the way—it means more than words can say.

Dedication

All the children whose daily experiences in kindergartens inspired this work. May we continue to create safe, understanding, and compassionate environments where every child can grow, learn, and thrive without fear of conflict.

To the educators and caregivers who work tirelessly to support young minds—this research is a small step toward helping you better understand and respond to the social dynamics in early childhood settings.

And to my family and loved ones, whose support and belief in me made this journey possible.

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-theart 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.

Аңдатпа

Балабақшаларда әлеуметтік қақтығыстарды ерте анықтау — оңтайлы оқу ортасын қалыптастыру және балалардың әлеуметтік дамуын қолдау үшін маңызды. Бұл зерттеуде терең оқыту мен компьютерлік көру әдістерін қолданып, кішкентай балалар арасындағы әлеуметтік қақтығыстарды автоматты түрде анықтау және талдау қарастырылады. Жазылып алынған өзара әрекеттерге сүйене отырып, біз RNN негізіндегі үлгілер мен 3D-CNN сияқты озық нейрондық желі архитектураларын қолданамыз, бұл қақтығысты дәл анықтау дәлдігін арттырады. Бұл үлгілер бет әлпеті, ым-ишаралар, дене тілі, дауыс ырғағы және қозғалыс үлгілері сияқты визуалды белгілерді талдап, әлеуметтік шиеленіс немесе агрессия жағдайларын анықтай алады. Ұсынылған модель балабақшадағы нақты бейнематериалдарға сыналып, қақтығыс оқиғаларын анықтау мен жіктеуде жақсы нәтижелер көрсетті. Бұл қорытындылар жасанды интеллект негізіндегі шешімдердің тәрбиешілерге сыныптағы ахуалды басқаруда, балалар мінез-құлқын бақылауды жақсартуда, ерте араласу стратегияларын күшейтуде және үйлесімді әлеуметтік ортаны қалыптастыруда көмектесе алатынын көрсетеді.

Кілт сөздер: әлеуметтік қақтығыстарды анықтау, терең оқыту, компьютерлік көру, балабақша, балалардың мінез-құлқын талдау, қалыпты дене күйін бағалау, көңіл-күйді талдау, сыныпты бақылау, ерте жастан білім беру, білім берудегі жасанды интеллект.

Аннотация

Раннее выявление социальных конфликтов в детских садах имеет важное значение для создания позитивной образовательной среды и поддержки социального развития детей. В данном исследовании рассматривается применение методов глубинного обучения и компьютерного зрения для автоматического обнаружения и анализа социальных конфликтов среди маленьких детей. Используя записи взаимодействий, мы применяем современные архитектуры нейронных сетей, включая модели на основе RNN и 3D-CNN, для повышения точности в идентификации конфликтных ситуаций. Эти модели анализируют визуальные признаки, такие как выражения лица, жесты, язык тела, тон речи и характер движений, чтобы распознавать случаи социальной напряжённости или агрессии. Предложенная модель тестируется на реальных видеозаписях из детских садов и демонстрирует обнадеживающие результаты в обнаружении и классификации конфликтных событий. Полученные результаты подчеркивают потенциал решений на базе ИИ в помощи педагогам в управлении динамикой в классе, улучшении мониторинга поведения детей, усилении стратегий раннего вмешательства и содействии гармоничной социальной среде.

Ключевые слова: выявление социальных конфликтов, глубокое обучение, компьютерное зрение, детский сад, анализ поведения детей, оценка позы, анализ настроения, мониторинг в классе, дошкольное образование, ИИ в образовании.

Abbreviations

AI – Artificial Intelligence
CNN – Convolutional Neural Network
RNN – Recurrent Neural Network
3D CNN – Three-Dimensional Convolutional Neural Network
CV – Computer Vision
DL – Deep Learning
ML – Machine Learning
FPS – Frames Per Second
ROI – Region of Interest
LSTM – Long Short-Term Memory
ReLU – Rectified Linear Unit
GPU – Graphics Processing Unit
IoU – Intersection over Union
TP – True Positive
FP – False Positive
FN – False Negative
TN – True Negative
API – Application Programming Interface
UI – User Interface
NLP – Natural Language Processing
AR – Augmented Reality

Table of Contents

Chapter 1

Intorduction

1.1 Background Information

Social interactions are a cornerstone of early childhood development, exerting a profound influence on children's emotional, cognitive, and social growth. These interactions not only contribute to the cultivation of emotional intelligence but also facilitate the development of communication skills and conflict resolution strategies that children carry into later stages of life[1]. In the kindergarten environment, these foundational interactions become particularly pronounced, as children are immersed in collaborative activities, peer-to-peer communication, and play-based learning scenarios. It is within these moments of shared activity that children learn to interpret social cues, empathize with others, and navigate interpersonal relationships.

In these early educational settings, social conflicts can arise due to a variety of factors, including competition over toys, disagreements during group activities, differences in communication styles, and the struggle for social inclusion. Teachers and caregivers traditionally rely on their professional intuition, training, and observational skills to identify and mediate these conflicts. However, this method, while valuable, is highly subjective and constrained by practical limitations. Classrooms are typically busy, unpredictable, and filled with numerous simultaneous interactions, making it difficult for educators to monitor all activities with equal attention. Furthermore, personal bias and emotional fatigue may hinder the consistency and effectiveness of manual conflict detection[2, 3].

Given the increasing availability and sophistication of technological tools, the application of deep learning and computer vision in early childhood education offers a transformative solution. These technologies have demonstrated considerable success in various domains such as surveillance, healthcare, and human-computer interaction, and are now being explored for their potential in educational settings. Deep learning models, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are capable of identifying subtle patterns in video data that correlate with specific behaviors. When applied to recorded classroom footage, these models can learn to recognize early indicators of social conflict, such as facial tension, abrupt body movements, or withdrawal from group activities[4].

Automating the detection of social conflict through AI-driven systems enables

real-time monitoring, timely response, and objective analysis. Unlike human observers, AI systems do not suffer from fatigue or bias and can continuously monitor all children in a classroom without interruption. This consistent oversight can provide educators with valuable insights into patterns of behavior, recurrence of conflicts, and effectiveness of interventions. Furthermore, data collected through these systems can support longitudinal studies on child development and help create adaptive learning environments tailored to children's needs.

This research focuses on leveraging these technological advancements to address longstanding challenges in classroom conflict management. By developing and validating a robust AI-powered system, this study aims to bridge the gap between theoretical understanding of social conflict and its practical application in kindergarten environments. The ultimate goal is to improve classroom harmony, enhance educational outcomes, and support the emotional well-being of children through early and effective conflict detection and resolution.

1.2 Purpose of the Study

The primary purpose of this study is to design, develop, and evaluate an artificial intelligence system that utilizes deep learning and computer vision techniques to automatically detect social conflicts among children in kindergarten classrooms. This initiative is grounded in the belief that early detection and resolution of social conflicts are crucial for fostering a supportive and inclusive learning environment. By equipping educators with real-time tools and behavioral insights, the study aims to empower them with data-driven strategies that promote emotional intelligence, prevent escalation, and support positive child development.

The system under development is trained on video data capturing a variety of classroom interactions, focusing on identifying behavioral cues commonly associated with social conflict. These cues include, but are not limited to, changes in facial expression (e.g., frowning, crying), physical gestures (e.g., pointing, pushing), variations in posture (e.g., turning away, crossing arms), and movement dynamics (e.g., pacing, sudden motions). Through careful annotation and supervised learning, the model will learn to associate these indicators with potential conflicts, providing alerts or summary reports for educators to review[5].

An important aspect of the study is the investigation of how conflict-related behaviors manifest over time and under different classroom contexts. For instance, the study explores how teaching styles, classroom layout, and group composition influence the frequency and nature of conflicts. The temporal evolution of conflict situations is also analyzed, using models like LSTM (Long Short-Term Memory) networks and 3D CNNs, which are capable of capturing time-dependent behavioral sequences. By understanding these patterns, the system can not only detect conflicts but also anticipate them, enabling preventive strategies and timely intervention.

In addition to the technical development, this study also considers the practical deployment of the system in real-world kindergarten environments. Issues such as user-friendliness, data privacy, ethical use of AI, and teacher training are addressed to ensure that the system is both effective and appropriate for educational use.

The broader goal is to integrate the system into daily classroom routines, providing educators with a non-intrusive tool that enhances their ability to manage social dynamics with greater confidence and precision.

Ultimately, the study aspires to contribute to the field of early childhood education by offering a novel, evidence-based approach to conflict detection and intervention. The findings are expected to inform both policy and practice, highlighting the benefits of technological innovation in enhancing the quality and equity of education.

1.3 Research Objectives

1. To collect a dataset of conflict videos involving kindergarten children and develop a deep learning-based computer vision system for real-time conflict detection.

The first objective of the study is to create a comprehensive and annotated video dataset that captures a wide range of social interactions among kindergarten-aged children. This dataset serves as the foundation for training and evaluating the deep learning models. The videos were collected from consenting educational institutions under strict ethical guidelines, ensuring that privacy and safety standards are upheld[6]. Each video segment was manually annotated to identify instances of conflict and to label the behavioral features present in each frame, providing the model with high-quality training data.

The deep learning system will employ a combination of CNNs for spatial feature extraction and RNNs (including LSTM variants) for temporal analysis. The goal is to achieve real-time performance, allowing educators to receive alerts or summaries of conflict situations as they happen. Evaluation metrics such as precision, recall, F1-score, and latency is used to assess the model's effectiveness in identifying true conflict events while minimizing false positives.

2. To evaluate the performance of fight detection models designed for adults and assess the system's accuracy and adaptability in kindergarten settings.

The second research objective focuses on adapting and evaluating pre-existing models that were initially developed for adult surveillance contexts—such as public safety or crowd monitoring—for use in early education. These models typically detect physical aggression or hostile behavior based on motion trajectories, facial analysis, and pose estimation[7]. However, conflicts in kindergartens are often subtle, emotionally nuanced, and non-violent. Therefore, this study will assess the limitations and potential modifications needed to repurpose these models for a younger demographic.

A comparative analysis are conducted between adult-focused models and the newly developed kindergarten-specific model. Key aspects of comparison include detection accuracy, adaptability to new data, and robustness in varied classroom conditions. The aim is to determine whether transfer learning techniques or domain adaptation strategies can enhance performance in detecting child-specific conflict behaviors. Furthermore, this objective will explore how the context of education necessitates a redefinition of what constitutes 'conflict', emphasizing

emotional over physical indicators.

3. To analyze conflict behavior patterns and triggers, enhancing the understanding of social dynamics in young children.

The third objective is dedicated to the in-depth analysis of conflict behaviors to uncover patterns, triggers, and resolution strategies within kindergarten environments[8]. Using the data collected and insights generated by the AI system, the study aims to classify conflicts based on severity, duration, resolution pathway, and emotional tone. By clustering conflict instances, the research will identify common triggers such as toy sharing disputes, turn-taking failures, or peer exclusion.

This objective also includes an exploration of teacher intervention methods and their effectiveness. The study will analyze how and when educators intervene in conflicts, the language and strategies they use, and how these affect conflict outcomes. Correlations're drawn between intervention timing and conflict resolution success, contributing to best practices for proactive conflict management. Additionally, the influence of classroom environment factors—such as group size, seating arrangements, and teacher-student ratios— was examined to understand how external elements contribute to or mitigate conflict scenarios.

Ultimately, this analysis aims to deepen the educational community's understanding of the social dynamics at play in early childhood and to provide actionable insights for teachers, administrators, and policymakers.

This study highlights a modern approach to classroom management, combining educational insights with advanced technological tools to cultivate a supportive, inclusive, and developmentally rich environment for young children.

1.4 Technical Foundations

This section provides the theoretical and technical background necessary to understand and develop an AI-driven system for detecting social conflicts in kindergarten environments. It includes an in-depth overview of neural networks, computer vision, convolutional neural networks, 3D CNNs, recurrent neural networks with Long Short-Term Memory, and video processing techniques. Ethical and practical deployment considerations are also discussed.

1.4.1 Overview of Neural Networks

Artificial neural networks are a subset of machine learning algorithms inspired by the human brain's structure and function. Each artificial neuron receives one or more inputs, processes them through a non-linear function, and passes the result to subsequent neurons. The power of ANNs lies in their ability to approximate complex functions and patterns from large datasets, a key feature in behavioral recognition tasks like conflict detection in video data.

The basic form of a neural network is the feedforward neural network, where information flows from input to output layers through one or more hidden layers. However, for tasks involving structured spatial or temporal data (like images or videos), more specialized neural network architectures are required.

1.4.2 Computer Vision and Video Analysis

Computer vision involves teaching machines to understand and interpret visual data such as images and videos. In the context of education and behavior analysis, computer vision techniques can be employed to monitor children's facial expressions, body posture, and interactions with peers, which are key indicators of social conflict or cooperation.

Video analysis involves processing sequences of image frames to extract meaningful temporal and spatial features. A critical aspect of video-based behavior recognition is the ability to handle both spatial details (e.g., gestures, object presence) and temporal dynamics (e.g., movement over time). This dual requirement motivates the integration of CNNs for spatial analysis and RNNs or 3D CNNs for temporal analysis.

1.4.3 Convolutional Neural Networks

CNNs are deep learning architectures particularly suited for spatial data like images. They use convolutional layers to automatically and adaptively learn spatial hierarchies of features from input data. CNNs have revolutionized computer vision, leading to state-of-the-art performance in image classification, object detection, and facial recognition tasks [8].

A typical CNN architecture consists of several layers:

- Convolutional layers: Apply filters to input images to extract local features.
- Activation functions: Introduce non-linearity using ReLU or similar functions.
- Pooling layers: Reduce dimensionality and highlight the most significant features.
- Fully connected layers: Integrate learned features for classification.

Popular CNN architectures include VGGNet, ResNet, and EfficientNet. For this study, EfficientNet-B3 and ResNet-101 were used due to their balance of accuracy and computational efficiency [9].

1.4.4 3D Convolutional Neural Networks

While standard CNNs analyze images frame-by-frame, 3D CNNs are designed to extract spatiotemporal features by processing multiple contiguous video frames as a 3D volume. This allows them to learn motion patterns across time and space simultaneously, making them ideal for video-based tasks such as violence or conflict detection [10].

A 3D CNN layer extends the traditional 2D convolutional operation by adding a temporal dimension. For example, a 3D convolution kernel may have dimensions (3, 3, 3), processing three frames at once with a 3x3 spatial window. This enables the model to capture both frame-wise dynamics and within-frame interactions.

One of the most effective architectures in this category is the I3D model (Inflated 3D ConvNet), which inflates 2D filters into 3D, leveraging pretraining on image datasets while learning temporal patterns [11]. Another model, the Slow-

Fast network, uses two parallel pathways: one slow pathway for spatial semantics and one fast pathway for motion sensitivity [12].

1.4.5 Recurrent Neural Networks and LSTM

Recurrent Neural Networks (RNNs) are designed for sequence data, where the output at each step depends on previous computations. RNNs are especially useful for temporal behavior analysis in videos, where actions unfold over time. However, vanilla RNNs suffer from vanishing gradient problems during long sequences.

Long Short-Term Memory (LSTM) networks address this limitation by introducing gating mechanisms that regulate information flow, enabling the network to learn long-range dependencies [13]. Each LSTM unit contains:

- Forget gate: Decides what information to discard.
- Input gate: Selects which information to update.
- Output gate: Determines the final output.

Bidirectional LSTMs, which process sequences in both forward and backward directions, further enhance the model's capacity to understand context from the entire sequence. In this project, LSTM networks were used in conjunction with CNNs (CNN+LSTM) to model spatial-temporal dynamics.

1.4.6 Video Data Processing and Annotation Techniques

Working with video data introduces specific challenges such as frame redundancy, high storage demands, and the need for frame-level annotations. The video preprocessing pipeline for this project included:

- Frame extraction: Videos were segmented into frames at a fixed rate.
- Resizing: All frames were resized to 224x224 pixels.
- Normalization: Pixel values were standardized to improve model training.
- Augmentation: Techniques like horizontal flipping, rotation, and jittering were applied to increase data diversity [14].

Annotation involved manual labeling of video segments as "conflict" or "non-conflict" based on observed behavior. These labels served as ground truth for supervised model training. Annotation criteria included gestures (e.g., pushing, grabbing), facial expressions (e.g., frowning, crying), and body language (e.g., turning away).

1.4.7 Transfer Learning and Pretrained Models

Training deep models from scratch requires large labeled datasets and extensive computational resources. Transfer learning addresses this by leveraging pretrained models on large datasets like ImageNet or Kinetics, and fine-tuning them for specific tasks [15].

In this study, EfficientNet-B3 and ResNet-101 were used as feature extractors, pretrained on ImageNet. The I3D model was initialized with weights from the

Kinetics-400 dataset. Fine-tuning allowed the models to adapt to domain-specific features such as child interactions and subtle conflict indicators.

Transfer learning significantly improved training efficiency and performance, especially when working with a relatively small custom dataset.

1.4.8 Multimodal Learning

While visual cues are critical for conflict detection, other modalities such as audio, pose estimation, and sentiment analysis can provide complementary information. Multimodal learning involves integrating these data sources to improve model robustness and accuracy [16].

Pose estimation models like OpenPose can identify skeletal keypoints, helping detect aggressive postures or physical interaction patterns. Sentiment analysis based on facial expression recognition or speech tone analysis can offer emotional context. Combining these with visual features through a fusion network enables a more nuanced understanding of classroom interactions.

Future extensions of this research will incorporate multimodal data to better distinguish between playful interactions and actual conflicts.

1.4.9 Ethical and Practical Considerations

Deploying AI systems in educational settings requires careful consideration of ethical, legal, and societal implications [17]. Key concerns include:

- Data privacy: Ensuring children's video data is securely stored and anonymized.
- Informed consent: Obtaining permission from parents, teachers, and school administrators.
- Bias mitigation: Ensuring the model does not disproportionately misclassify certain children.
- Transparency: Using Explainable AI (XAI) techniques to help educators understand why the system flagged certain behaviors.

Explainable AI approaches such as Grad-CAM and attention visualization can highlight which parts of an image or sequence influenced the model's decision [18]. This fosters trust and allows educators to validate and interpret the system's outputs.

1.4.10 Real-Time Inference and Deployment Challenges

Achieving real-time performance is essential for practical classroom use. This requires optimizing the model pipeline for low-latency inference. Strategies include:

- Model quantization: Reducing model size by using lower precision weights.
- Frame skipping: Processing only key frames.
- Efficient backbones: Selecting architectures like MobileNet for edge deployment.

This section has provided an in-depth analysis of the technical foundations essential for building an AI-powered system aimed at detecting social conflicts

in kindergarten environments. Through a systematic discussion of key components—including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and Long Short-Term Memory units (LSTMs), 3D CNNs, video data preprocessing techniques, and transfer learning frameworks—this chapter has established the methodological groundwork necessary for implementing real-time behavior analysis in early childhood education contexts.

The selection of CNNs as a foundational model is justified by their proven ability to perform spatial feature extraction across various domains of image and video analysis. CNNs have revolutionized the field of computer vision due to their capability to learn hierarchical representations of input data, from low-level textures to high-level semantic concepts. In the context of kindergarten conflict detection, CNNs serve as effective tools to identify visual cues such as facial expressions (e.g., frowning, crying), body gestures (e.g., pushing, pointing), and spatial relations between individuals.

However, understanding social behavior extends beyond static frames. Human interaction is inherently dynamic, evolving across time in subtle yet meaningful patterns. This necessitates the incorporation of temporal modeling, which is effectively handled by recurrent architectures such as LSTMs. LSTMs mitigate the vanishing gradient problem inherent in traditional RNNs and are well-suited for modeling sequences where long-range dependencies exist. In a kindergarten setting, the buildup to a social conflict may involve a sequence of gestures or expressions spanning several seconds, making LSTMs highly applicable for real-time conflict prediction and detection.

3D CNNs offer an alternative yet complementary approach by treating the video input as a spatiotemporal volume rather than as a sequence of individual frames. Unlike traditional CNNs that analyze single images, 3D CNNs apply convolution operations across both spatial and temporal dimensions, allowing the model to learn motion-related features more holistically. This is especially valuable in detecting behaviors that may be characterized by sudden movement or interaction trajectories, such as grabbing toys or gesturing in an aggressive manner. In many cases, such models have demonstrated superior performance over two-stream CNNs or separate CNN-RNN pipelines in action recognition tasks.

Video preprocessing and annotation were also emphasized as critical components in building a robust training pipeline. The integrity of any deep learning system is fundamentally tied to the quality and representativeness of its training data. In this study, preprocessing steps such as frame extraction, resizing, normalization, and augmentation (e.g., flipping, cropping, and rotation) were applied to improve generalizability and reduce overfitting. Manual annotations, although time-consuming, were essential for generating reliable ground-truth labels that reflect nuanced social dynamics. These steps ensured that the models were not only data-efficient but also attuned to the behaviors most relevant in kindergarten conflict scenarios.

Transfer learning, as discussed, provides a pragmatic solution to the challenges posed by limited domain-specific data. By fine-tuning models pre-trained on large-scale datasets such as ImageNet or Kinetics, the system can leverage generalized visual features while adapting to the unique patterns present in early childhood

interactions. This approach significantly reduces training time and computational overhead while improving performance, particularly when the target dataset is small or imbalanced.

The ethical considerations associated with the deployment of AI systems in educational settings were also addressed. The use of AI in classrooms introduces concerns around data privacy, informed consent, surveillance, and algorithmic bias. These concerns are heightened when the subjects are children, a population that demands heightened sensitivity and protection. Therefore, the technical system must be designed with stringent privacy-preserving mechanisms such as anonymization, data encryption, and local on-device processing wherever feasible. Moreover, Explainable AI (XAI) techniques should be integrated into the system to ensure that model predictions are interpretable by teachers and caregivers, fostering transparency and trust.

Beyond the immediate scope of model performance, the broader implications of this work point toward a shift in how educators and policymakers conceptualize classroom management and child development. Real-time AI systems offer a level of responsiveness and objectivity that surpass traditional observational methods. Unlike human observers, AI systems do not experience fatigue, distraction, or personal bias. This consistency enables a more holistic view of classroom dynamics and empowers educators to make timely, data-informed decisions.

However, the application of such systems also calls for thoughtful integration into pedagogical practice. Teachers must be trained not only in the interpretation of AI-generated insights but also in understanding the limitations and uncertainties associated with algorithmic predictions. AI should serve as an assistive tool—not a replacement for human judgment—and should be deployed in ways that enhance, rather than undermine, the human-centric values of education.

Looking ahead, several avenues for future research and system enhancement emerge. One promising direction involves the incorporation of multimodal data inputs. While current models primarily rely on visual data, combining video with audio streams (e.g., tone of voice, crying, shouting) or even physiological data (e.g., heart rate via wearables) could improve the accuracy and context-awareness of conflict detection. Multimodal fusion models can more accurately differentiate between high-energy play and genuine conflict by integrating diverse cues.

Another area for exploration is adaptive learning and personalization. Classrooms differ significantly in culture, size, layout, and social norms. An effective AI system must be adaptable to these differences. Implementing online learning mechanisms or meta-learning approaches could enable the model to continuously adapt to the specific context in which it is deployed, improving performance over time. This personalization would allow the system to detect not only generic patterns of conflict but also child-specific behavioral changes that may signal emotional distress or social exclusion.

Additionally, integration with intervention modules could elevate the utility of the system from passive detection to active prevention. For example, upon detecting a high-conflict probability, the system could trigger alerts that suggest real-time pedagogical strategies to the teacher, such as engaging the children in a calming group activity or separating individuals involved in the conflict. Such

interventions could be based on prior successful strategies stored in the system's database, enabling evidence-based classroom management.

Lastly, collaboration with interdisciplinary teams—including developmental psychologists, educators, AI ethicists, and child rights advocates—will be essential in refining both the system's capabilities and its alignment with educational values. Ethical guidelines and policy frameworks should evolve alongside technological capabilities to ensure responsible innovation.

In conclusion, the technical foundations explored in this section offer a robust and versatile architecture for real-time social conflict detection in early childhood environments. The synergy of CNNs, LSTMs, and 3D CNNs, bolstered by transfer learning and explainability, creates a powerful tool that can transform how educators perceive and respond to social dynamics in the classroom. The implications extend beyond automation, signaling a shift toward proactive, data-informed, and ethically grounded educational practice. With continued research and careful deployment, such AI systems have the potential to play a pivotal role in fostering inclusive, emotionally safe, and developmentally rich environments where every child can thrive.

Chapter 2

Literature review

The detection of social conflicts in children, particularly within kindergarten settings, has emerged as a focal point of interdisciplinary research involving developmental psychology, education, artificial intelligence, and computer vision. These early educational environments represent critical developmental stages, where foundational social, emotional, and cognitive skills are cultivated. Consequently, early detection of interpersonal challenges—such as peer exclusion, aggressive interactions, or emotional distress—can play a crucial role in supporting healthy social development and mitigating long-term behavioral or psychological issues.

The integration of AI and computer vision into this domain offers promising pathways for real-time, scalable, and objective monitoring of social dynamics among young children. Deep learning techniques, especially those trained on video and audio data, enable systems to capture nuanced behavioral patterns that may elude traditional observation methods. As a result, researchers and practitioners are increasingly exploring how intelligent systems can assist educators in identifying and responding to potential conflicts or behavioral concerns with greater accuracy and timeliness.

This chapter provides a comprehensive review of current literature relevant to the automated detection of such social conflicts, with a particular emphasis on deep learning, violence detection, child safety, and ethical considerations in educational technologies. By synthesizing findings across multiple disciplines, the chapter aims to illuminate both the technological potential and the sociocultural sensitivities involved in this field. It is structured to reflect the primary domains of research that converge in this area and evaluates both the strengths and limitations of contemporary methodologies, highlighting opportunities for innovation and responsible implementation.

2.1 Deep Learning for Conflict Detection

The application of deep learning techniques to video analysis has gained substantial momentum in the last decade. Particularly in contexts such as school security, public surveillance, and child behavior monitoring, deep learning has proven effective in automating the identification of aggressive or abnormal behavior patterns. Its ability to extract and learn features from raw data without explicit manual

engineering makes it especially useful in complex, unstructured environments such as classrooms.

Thao et al. (2023) introduced the FightNet model, which utilizes a hybrid architecture combining Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs) to detect school-based violent interactions. Their framework, based on keypoint estimation and spatio-temporal feature integration, demonstrated a mean average precision (mAP) of 45.34% (IoU 0.5) and an F1-score of 71.69% [18]. This hybrid structure enables the model to interpret both spatial features (such as gestures and facial cues) and temporal dynamics (such as escalating behavior patterns), providing a more nuanced understanding of conflict episodes.

Despite these promising results, FightNet's training datasets primarily featured adolescents and young adults. The behavioral cues in kindergarten settings differ significantly—conflicts are rarely violent or overt and often involve ambiguous gestures, nonverbal expressions, subtle peer exclusion, or social rejection. These nuances are often difficult to detect using models trained on violent or exaggerated actions. Therefore, the development of conflict detection systems for early childhood settings must involve datasets that are specifically curated and annotated to reflect the nature of interactions among preschool-aged children. Furthermore, FightNet may fail to generalize unless it is adapted using transfer learning and synthetic data augmentation techniques on domain-specific corpora.

Similarly, Imah and Karisma (2022) developed a deep learning model using a VGG16 backbone integrated with an LSTM network, achieving a G-mean of 0.911 [19]. Their system showed high accuracy in detecting physical altercations, particularly in surveillance footage. However, as with FightNet, their datasets were drawn from adult interactions and featured scenarios involving clear physical aggression. While their methodology successfully combines spatial and temporal data streams, it overlooks subtler emotional cues and body language patterns commonly seen in young children during conflict situations.

These limitations are echoed in studies by Singh et al. (2021), who found that models trained on standard violence datasets such as Hockey Fight or Movies Fight tend to overfit to high-motion, high-impact events, resulting in poor generalizability to educational or social settings where emotional intensity is low but pedagogically significant [20]. Such findings underscore the need for tailored datasets and models that emphasize emotional regulation, nonverbal signals, and situational context. Conflicts among children often manifest in relational aggression—such as exclusion from a game, name-calling, or dominance assertion—rather than overt physical confrontation.

Emerging research supports the integration of vision-based recognition with audio signals to enhance model performance. Audio cues such as shouting, crying, or voice pitch changes provide essential context that helps disambiguate visual behaviors. For example, a raised arm might indicate either a threat or a child seeking attention, but coupling it with the accompanying tone of voice or peer reactions can clarify intent. Hybrid CNN-RNN architectures that incorporate attention mechanisms have demonstrated increased accuracy in classroom video analysis by weighing context-specific cues over raw motion [21].

Attention-based models such as those proposed by Vaswani et al. (2017) in the

Transformer architecture have further revolutionized temporal modeling, making it feasible to attend to critical frames or events within longer video sequences [22]. In the context of kindergarten surveillance, attention modules can help isolate conflict-relevant clips from long, uneventful recordings, thereby improving both efficiency and interpretability.

Moreover, models like SlowFast Networks [23], which operate at two different temporal resolutions, have shown potential for detecting both short, high-frequency actions (e.g., sudden shoves) and longer, more drawn-out behaviors (e.g., exclusion or tension building). Adapting such architectures to early education settings may yield more holistic insights into children's social interactions.

Beyond supervised learning, semi-supervised and self-supervised methods are gaining popularity in this field. Annotating large-scale video data of children is both ethically sensitive and logistically difficult. Self-supervised approaches, which learn representations from unlabelled data by predicting temporal continuity or spatial consistency, can help leverage the vast amounts of unannotated classroom footage [24]. These techniques reduce reliance on labeled datasets while still capturing the dynamics of child interactions.

Another promising direction is the use of Graph Convolutional Networks (GCNs), which model human interactions as dynamic graphs. Nodes represent individuals, and edges reflect social connections or physical proximity. By capturing group dynamics, GCNs can detect not just individual aggression but collective behavior patterns such as exclusion or group pressure [25]. In kindergartens, where peer dynamics often drive conflict, such group-level modeling can be particularly valuable.

Despite these advancements, several challenges persist. Real-time inference is critical for classroom deployment, yet many deep models are computationally expensive and require GPU resources not commonly available in educational institutions. Lightweight models like MobileNetV2 and EfficientNet-Lite offer a viable compromise by maintaining performance while reducing memory and power requirements [26]. Edge computing devices such as NVIDIA Jetson or Google Coral can further facilitate on-site processing with minimal latency.

Importantly, researchers must account for ethical constraints. Datasets involving children must adhere to strict privacy protections, including anonymization, encryption, and informed consent protocols. Additionally, models should be interpretable to educators, not just data scientists. Explainable AI (XAI) techniques, such as Grad-CAM for visualizing attention maps, can help users understand why a particular sequence was flagged as a conflict [27]. This transparency builds trust and facilitates more accurate interventions by teachers.

In conclusion, while deep learning holds great promise for detecting social conflicts in kindergarten settings, its efficacy depends on several factors: the relevance of training datasets, the capacity to model subtle emotional and social behaviors, and the ethical integration of vision and audio data. Future systems should adopt a multi-modal, interpretable, and privacy-conscious design that not only recognizes behavioral cues but also supports teachers in fostering inclusive and empathetic classroom environments.

2.2 Violence Detection in Surveillance

Violence detection in surveillance videos has traditionally relied on overt physical cues such as striking, kicking, or rapid body movements. Hughes and Kersten (2022) combined CNN and LSTM models to detect violence in public video datasets like the Hockey Fight and Movie Fight datasets. They achieved a detection accuracy of 77.9% [28]. Their methodology emphasized temporal feature extraction using RNNs and CNNs, allowing for the effective interpretation of sequences of rapid motion. While effective in high-motion, high-conflict scenarios such as street altercations or sports fights, this approach faces considerable limitations when applied to more nuanced social contexts, particularly those involving children.

The application of such models to educational settings—especially kindergartens-introduces several challenges. First and foremost, surveillance datasets differ drastically in tone, context, and the nature of human interactions. Public datasets often feature clear instances of physical violence, with high motion intensities, adult participants, and uncontrolled environments. In contrast, kindergarten environments are structured, supervised, and socially complex. Interactions are typically spontaneous and developmentally motivated rather than intentionally aggressive. The motivations behind a child's physical movement—whether playful or conflictual—are often difficult to distinguish without context.

Second, Hughes and Kersten themselves acknowledged high false-positive rates in their system, a drawback that would severely limit the practical utility of such a system in real-time classroom applications. Frequent false alerts may desensitize educators or trigger unwarranted interventions, potentially disrupting the learning environment. In kindergartens, where movement and noise are natural and frequent, overly sensitive systems may mistake playfulness for aggression, compromising both the credibility of the system and the trust of educators and parents.

Moreover, many of these models are designed with an implicit assumption that violence is synonymous with visible physical aggression. This framing overlooks a wide range of subtle, low-intensity conflicts that are prevalent in early childhood settings. For instance, children frequently engage in exclusionary behavior, verbal taunting, passive-aggressive responses, or nonverbal gestures of rejection—all of which can have profound social and emotional implications but may not involve any physical contact. These types of behavior typically fall outside the detection scope of traditional models trained exclusively on violent actions.

This misalignment between what is modeled and what needs to be detected calls for a fundamental redefinition of the concept of conflict in AI systems intended for use in kindergartens. Conflict must be conceptualized as a broader construct that includes emotional dysregulation, relational aggression, social exclusion, and micro-behaviors that signify distress or power dynamics among peers. Expanding the training objectives of machine learning models to include this wider emotional and behavioral spectrum is essential for making automated detection systems context-aware and pedagogically useful.

Recent studies reinforce this notion. For example, Zhao et al. (2023) emphasized the importance of integrating emotion recognition into violence detection

pipelines, arguing that emotional cues such as facial tension, crying, or sudden withdrawal can be early indicators of social conflict [29]. By using multi-task learning approaches that simultaneously predict emotion, gesture, and proximity, their model demonstrated improved precision in identifying interpersonal tension in controlled school environments.

Additionally, Wang et al. (2022) proposed a framework that combined optical flow with facial expression recognition to detect pre-violent behavior in children's group interactions [30]. This approach enabled the system to register escalating tension before physical actions occurred. Their results indicated that integrating emotional and affective computing tools significantly reduced false positives in low-aggression environments.

To address contextual variability, the use of attention mechanisms and scene-aware modeling has also been proposed. For example, the Transformer-based framework proposed by Rahmani et al. (2022) introduces spatio-temporal attention layers that dynamically weigh frames based on their behavioral salience [31]. When applied to simulated classroom datasets, this method was able to isolate critical moments, such as prolonged staring, ignoring a peer, or sudden changes in group dynamics—actions that often precede overt conflict but may not be immediately perceptible to standard CNN models.

Another essential consideration is the domain shift problem. Models trained on adult datasets typically fail to generalize to child populations due to differences in body proportions, movement patterns, and social norms. A shove between two adults may carry very different social meaning than the same gesture between children—depending on tone, context, and peer relationships. Therefore, domain-specific training, possibly augmented with synthetic data generation or domain adaptation techniques, is necessary to overcome this gap [32].

Moreover, the inherent ambiguity of certain child behaviors presents an additional challenge. For example, a child pushing another may be expressing aggression, seeking attention, or initiating play. The interpretation often depends on contextual cues such as tone of voice, facial expression, and the reaction of peers. In this regard, multimodal approaches that integrate video with audio (e.g., pitch, tone, crying), sensor data (e.g., proximity, posture), and even classroom metadata (e.g., seating arrangement, peer groups) may provide a more holistic view of social interactions [33].

While there is growing interest in building such multimodal datasets, ethical and privacy concerns remain significant barriers, particularly when dealing with minors. Institutional Review Boards (IRBs) and child protection regulations demand stringent anonymization, secure storage, and strict consent protocols. These challenges slow down the collection of rich data but also emphasize the need for privacy-preserving machine learning techniques, such as federated learning or differential privacy, which allow for model training without direct access to sensitive raw data [34].

Ultimately, for violence detection models to be meaningfully integrated into early childhood education contexts, they must move beyond traditional notions of physical aggression and adopt a developmental lens. Conflict in kindergartens is a complex and multifaceted phenomenon that cannot be captured by high-

motion detection alone. Effective systems must be sensitive to the developmental, emotional, and social dimensions of behavior while maintaining ethical integrity and practical utility for educators.

2.3 Detection of Child Abuse and Distress

Complementary research has focused on distress detection in children, particularly in relation to abuse, trauma, or emotional dysregulation. Yan et al. (2023) demonstrated the efficacy of deep Convolutional Neural Networks (CNNs) for detecting vocal cues of distress using spectrogram analysis and Mel Frequency Cepstral Coefficients (MFCCs), achieving an accuracy above 90% [35]. Their study emphasizes the role of non-visual cues—such as crying, screaming, or speech tone variations—in understanding emotional and psychological states. These acoustic signals often serve as early indicators of interpersonal conflict or psychological distress and are critical in differentiating between playful and aggressive scenarios in ambiguous social contexts.

This body of research strongly supports the argument that multimodal systems—those integrating both visual and auditory signals—are essential for effective and context-sensitive conflict detection. While visual cues such as gestures, proximity, or facial expressions provide essential spatial and emotional information, auditory signals are capable of adding important temporal and affective dimensions. For example, raised voice tones, sobbing, or hesitation in speech can signal emotional distress even when visual indicators are subtle or ambiguous. In many cases, aggressive posturing may be misinterpreted without the accompanying tone or context of the verbal interaction. An angry gesture accompanied by laughter might be a form of rough play, while a neutral gesture combined with crying may indicate serious emotional harm.

Studies have explored audio-visual fusion models, where inputs from both channels are simultaneously processed through parallel CNN or Transformer architectures and later merged through attention mechanisms or feature concatenation. Such fusion strategies have demonstrated improved performance across multiple domains, including action recognition, emotion detection, and surveillance [36]. When applied to early childhood education, these approaches have the potential to dramatically increase the sensitivity and specificity of conflict detection models, especially in environments where interactions are complex and rapidly evolving.

Despite their promise, however, the integration of audio-based models in kindergarten environments presents several technical and ethical challenges. First, the typical classroom is a noisy, unpredictable environment filled with overlapping speech, ambient sounds, and spontaneous outbursts. This makes clean audio capture and accurate voice differentiation difficult. Differentiating between multiple children speaking simultaneously, or separating child voices from background noises such as toys, furniture movement, or teacher instructions, requires the application of advanced source separation and denoising algorithms, such as those based on Wave-U-Net or deep clustering approaches [37].

Second, the privacy implications of audio recording in early educational settings are considerable. Capturing the voices of children raises significant concerns

about surveillance, data consent, and parental rights. While video data is often anonymized through face blurring or silhouette tracking, audio data inherently contains personally identifiable information such as voice signatures, accents, and speech patterns. Therefore, any system that records audio in kindergarten environments must adhere to rigorous ethical and legal standards. Techniques like federated learning, where models are trained locally without transmitting raw data, or differential privacy, which introduces statistical noise to protect sensitive information, are increasingly proposed as potential solutions to safeguard audio-based systems in child-focused settings [38].

Moreover, the deployment of multimodal systems demands careful calibration and context-aware design. A child's expression of distress can vary depending on cultural background, temperament, or situational context. Some children may express discomfort quietly or nonverbally, while others might be more vocally expressive. Therefore, multimodal models must be designed to handle intra- and inter-individual variability, which can be addressed through the inclusion of large, diverse, and labeled training datasets specifically drawn from kindergarten settings. However, the current availability of such datasets remains limited. Most public datasets for violence or emotion recognition, such as CREMA-D or EmoReact, feature adult participants and are not tailored to early childhood behavior. The lack of large-scale, ethically curated, and representative datasets remains a substantial bottleneck for this research direction [39].

Recent efforts have begun to bridge this gap. For instance, Nishi et al. (2023) developed a prototype dataset of naturalistic child interactions in preschool classrooms, annotated for both visual and vocal indicators of conflict, including crying, yelling, withdrawal, and peer exclusion [40]. Their multimodal architecture, built upon a two-stream CNN-RNN hybrid with audio embedding and video frame analysis, achieved a conflict detection accuracy of 84.2%. Their findings suggest that a combination of facial action units, body orientation, and audio cues leads to a more robust understanding of child behavior in social settings.

Furthermore, the temporal alignment of multimodal data is critical. Speech signals and video frames must be synchronized accurately to ensure that corresponding events are interpreted as part of the same behavioral episode. Misalignment may cause the system to attribute vocal outbursts to the wrong visual actor or interpret unrelated events as linked. Tools such as Crossmodal Transformers and Temporal Fusion Networks have been proposed to handle such synchronization challenges by learning shared latent representations across modalities [41].

An important, but often overlooked, aspect of integrating auditory cues is the interpretability of the system. Teachers and caregivers are more likely to trust and adopt AI systems if they can understand how conclusions are drawn. Multimodal attention maps and saliency visualization techniques can help educators verify whether the system focused on relevant audio-visual segments. For example, showing that a child's distressed vocal pitch and withdrawal behavior triggered the alert may help validate the system's judgment and encourage appropriate human intervention [42].

In conclusion, while visual signals provide valuable insights into children's physical behavior and social orientation, the integration of auditory data adds a nec-

essary emotional and contextual layer. Together, these channels enable more accurate and nuanced interpretations of conflict and distress in preschool settings. However, technical complexities, privacy considerations, and data availability constraints must be addressed for the full potential of multimodal conflict detection systems to be realized in early childhood education environments.

2.4 Integration of Multi-Modal Data for Enhanced Detection

Video classification literature consistently affirms the value of integrating both spatial and temporal features to enable comprehensive scene understanding, especially in dynamic environments such as classrooms. Convolutional Neural Networks are particularly adept at identifying spatial patterns, including posture, facial expressions, gestures, and body orientation. These features provide a static frame-level understanding of the environment. However, behavioral understanding—especially for social interactions and conflict scenarios—requires contextual interpretation across time. To address this, Recurrent Neural Networks, Long Short-Term Memory units, and more recently Transformer architectures have been integrated with CNNs to capture temporal progressions and dependencies.

The foundational study by Tran et al. (2015) introduced 3D Convolutional Neural Networks for video classification, which process sequences of video frames by learning both spatial and temporal filters simultaneously [43]. This unified approach was particularly effective in capturing motion dynamics without requiring explicit frame-wise feature concatenation. The application of 3D CNNs in classroom contexts—especially those involving young children—is compelling, given the importance of understanding not just what action is taking place in a single frame, but how that action evolves over time. For instance, a child raising their hand could be interpreted as either a request for attention or the prelude to a push, depending on the temporal context and associated cues.

Moreover, slow fusion techniques have been proposed to improve temporal reasoning by incrementally aggregating feature representations over a sequence of frames. This gradual integration preserves fine-grained temporal dependencies and allows models to differentiate between repetitive behaviors (e.g., play) and isolated aggressive gestures (e.g., hitting). Carreira and Zisserman (2017) introduced the Inflated 3D network, which extends 2D CNN filters into 3D and pretrains on large-scale datasets like Kinetics for video classification tasks [44]. Such models have shown high performance on human action recognition benchmarks, and their adaptation to classroom behavior monitoring is a logical next step.

Another key approach involves multi-stream architectures, where RGB frames, optical flow representations, and pose keypoints are processed in parallel branches. Each stream extracts complementary information: RGB frames capture color and texture, optical flow encodes motion magnitude and direction, and keypoints provide structural understanding of body pose. These streams are later fused—either via early fusion (at the feature level) or late fusion (at the decision level)—to produce a more robust classification. In the context of kindergartens, multi-stream

models are beneficial for handling occlusions, variations in lighting, and subtle movement transitions that single-stream models might miss [45].

Temporal understanding is especially critical in preschool environments, where children move rapidly and unpredictably, and where social interactions are often fleeting. Behavioral cues that suggest conflict—such as snatching a toy or pushing a peer—may occur within a window of just a few seconds and can be easily misinterpreted if viewed in isolation. For example, a child might extend an arm in what appears to be a hitting motion; however, when seen in the broader temporal context, it could be a high-five or part of a collaborative game. Thus, contextual continuity—provided by sequential frame modeling—is essential for accurate interpretation.

Combining 3D CNNs with attention-based temporal encoding offers a promising solution to this challenge. Attention mechanisms allow models to focus on the most salient moments in a sequence, rather than treating all frames equally. This is particularly useful in classroom environments where most video segments may involve benign activities, punctuated by brief and critical moments of conflict. Temporal attention modules can identify and prioritize these key segments for further analysis, improving both the precision and efficiency of detection systems [46].

Recent advancements in transformer-based architectures—originally developed for natural language processing—have been extended to video analysis tasks. The TimeSformer model, for example, applies self-attention mechanisms along both spatial and temporal dimensions, enabling fine-grained interpretation of video sequences [47]. These models, while computationally intensive, offer improved performance over traditional CNN-RNN hybrids and are particularly well-suited for capturing long-range dependencies. In well-resourced research settings, incorporating such transformer-based models into kindergarten surveillance systems could greatly enhance behavioral prediction and conflict detection.

However, the use of high-capacity models in real-world kindergartens is currently limited by hardware constraints, data scarcity, and ethical concerns. Transformer based models require significant computational resources and large labeled datasets for training, which are often unavailable in educational contexts. Additionally, privacy-preserving implementations must be considered before deploying advanced video analytics in real classroom environments.

In conclusion, the integration of spatial and temporal modeling techniques remains central to the development of effective video-based conflict detection systems in kindergartens. While 3D CNNs and multi-stream architectures have shown strong performance in controlled settings, the inclusion of attention mechanisms and transformers may represent the next frontier for nuanced understanding of child behavior. For these advancements to be meaningful in early education, future research must address practical deployment challenges, dataset availability, and model interpretability.

2.5 Challenges in Implementing AI in Educational Settings

Despite technological advances, deploying AI in real-world classrooms is fraught with logistical and ethical challenges. Privacy concerns are paramount, as children are a protected population under most national regulations, including GDPR and COPPA. Obtaining consent for data collection, ensuring anonymization, and limiting the scope of surveillance are necessary to prevent misuse.

Hughes and Kersten's (2022) observations about false positives are particularly relevant here [48]. In classroom applications, such errors might unjustly implicate a child or raise concerns among parents. System transparency and educator oversight must be integral features of AI deployments. Moreover, models trained in one cultural or linguistic context may not generalize to others, necessitating diverse, inclusive training datasets.

Practical constraints also exist. Many preschools lack the infrastructure to support AI systems requiring continuous video analysis. Budget constraints, lack of technical staff, and resistance to change further complicate adoption. Thus, AI systems must be cost-effective, low-maintenance, and easy to interpret by educators with minimal technical training.

2.6 Ethical Considerations and Teacher-Child Dynamics

In their review of public safety applications of AI, Papadopoulos and Stavrakoudi (2024) emphasize that deep learning models should augment, rather than replace, human judgment [49]. In early childhood education, this principle is even more crucial. Teachers must remain the primary agents of behavioral interpretation and decision-making.

Automated systems should offer suggestions or alerts without overriding professional discretion. For example, rather than labeling an event as "aggressive," a system might flag it as "potentially concerning" based on deviations from normative behavior. This approach preserves the teacher's authority while still leveraging AI for improved awareness.

Another concern is over-monitoring. Children need space for social experimentation, and constant surveillance might inhibit natural interactions. Ethically designed systems must strike a balance between safety and developmental autonomy, allowing children to explore social boundaries without feeling policed.

2.7 Future Directions

As this field evolves, several technological and methodological advancements hold promise. First, combining CNNs with transformer models like Vision Transformers (ViTs) or TimeSformer could improve long-range temporal understanding. Transformers' attention mechanisms enable them to track relevant frames over

longer video segments, which may enhance detection of escalating conflicts.

Pose estimation tools such as OpenPose or MediaPipe, when integrated with emotion recognition modules, could facilitate deeper understanding of gestures, stances, and facial affect. For instance, slouched posture or turned backs may indicate exclusion, while arms-crossed postures may suggest defensiveness. Such features, while subtle, can be critical in diagnosing emotional states in social conflict.

Moreover, sentiment analysis tools analyzing children's speech can be paired with video models to form multimodal conflict detection pipelines. Research by Al-Tamimi et al. (2024) demonstrated that combining sentiment cues with visual markers in violent scene detection achieved up to 93% accuracy [50]. These tools could be adapted to classroom language and contextualized to detect negative sentiment or distress in preschoolers.

Finally, large-scale annotated datasets of kindergarten interactions are urgently needed. Crowdsourced or ethically annotated data from classrooms, paired with teacher commentary, can enhance supervised learning. Semi-supervised or unsupervised approaches may also be valuable in reducing dependency on labeled data, enabling models to generalize better in real-world settings.

Summary: The literature reviewed in this chapter illustrates both the opportunities and limitations of applying artificial intelligence to early childhood conflict detection. Deep learning and computer vision techniques, as demonstrated across multiple studies, offer significant promise for automating the recognition of social tensions, behavioral anomalies, and conflict patterns among young children. These technologies enable real-time, scalable, and objective analysis of classroom interactions that could enhance early intervention, support individualized learning, and reduce teacher workload. However, the successful implementation of such systems within kindergarten environments is not without challenges. The effectiveness of AI-driven models is heavily contingent upon their adaptability to the subtle and often ambiguous nature of early childhood behavior. Unlike adult conflict detection, which may rely on explicit and physically aggressive cues, kindergarten settings involve nuanced social signals, including gaze aversion, posture shifts, vocal intonation, and group dynamics that are contextually specific and developmentally variable.

Moreover, ethical integration is essential for ensuring responsible use. Concerns related to privacy, consent, data governance, and the potential psychological impact of surveillance must be carefully addressed to uphold the rights and emotional safety of young children. Additionally, the deployment of AI should not marginalize the role of educators but rather empower them with tools that complement their professional judgment. Teacher collaboration is vital in both the design and application phases, ensuring that AI systems are pedagogically aligned and responsive to real-world classroom needs.

To truly harness AI's transformative capabilities in educational environments, future research and system development must take a multidisciplinary approach that integrates insights from machine learning, developmental psychology, pedagogy, and ethics. This includes creating child-specific datasets, improving model interpretability, incorporating multimodal data (e.g., audio, posture, and facial

cues), and designing user-friendly interfaces that allow teachers to engage with AI outputs meaningfully. Furthermore, longitudinal studies are necessary to assess the long-term impacts of AI use on child development and classroom culture. By addressing these nuances, the field can move toward building intelligent systems that not only detect conflict but also promote empathy, cooperation, and social growth in early childhood education.

Chapter 3

Methodology

This paper introduces CNN-LSTM and 3D CNN deep learning technique specific to the identification of low-intensity conflict among children and separating playful conduct from aggression. In kindergarten classes, where conflict typically manifests itself as subtle presentations such as word disagreements, mild pushing, or verbal fights, the model must be designed so that it efficiently recognizes these subtle presentations. The model employs Convolutional Neural Networks (CNN) in capturing the spatial features, which are crucial in capturing the visual patterns and movement of children in the classroom. The spatial features allow the model to detect different behaviors and actions exhibited by the children such as body posture, facial expressions, and hand gestures. However, these spatial attributes are not sufficient to encode the full dynamics of conflict because interactions take time to develop. Against this, the system employs Long Short-Term Memory (LSTM) networks, which is a type of Recurrent Neural Network (RNN), in order to learn temporal relationships among successive frames of the video streams. This enables the model to learn how the interactions evolved from their nascent phases of an altercation into a possible escalation, while the ability to also discriminate between occasional flashpoints and playful moments are there.

The incorporation of a 3D CNN also enables the system better to process the spatiotemporal attributes concurrently, thereby equipping it with the potential to learn the spatial as well as temporal detail of the videos using a single framework. Unlike typical CNNs, which consider spatial dimensions (width and height) only, 3D CNNs work with three-dimensional data, including the depth (time) dimension. This allows the model to capture the movement flow and interactions in dynamic environments, like a kindergarten classroom where children's actions and behaviors change continuously. Using 3D CNN, the system is better at identifying cases of aggression that may involve minor body movement or posture change over time which would otherwise be lost to models that are only interested in spatial characteristics.

The primary application of this technique is to aid in teacher monitoring and support classroom management by being capable of identifying early aggression. This system allows teachers to better monitor children's behavior and identify potential conflict before full-blown altercations arise. Having the ability to detect low-intensity conflict can empower teachers with the ability to intervene early

and resolve problems in a non-disruptive manner, creating a more positive and supportive classroom. In addition, the system can be used to establish more efficient conflict resolution strategies by providing insights into behavioral patterns and giving teachers valuable information to enable the management and direction of social interaction among children. The system also opens new avenues for customized intervention.

By making a distinction between several types of behavior, the model can help teachers determine if a child's behavior is being driven by frustration, stress, or other emotional stimuli, compared to when they are simply playing with peers. This can lead to more targeted interventions, which can be vital in advancing social and emotional development among young children. Furthermore, the system's capability for instant feedback allows instructors to respond quickly to potential problems, minimizing the prospect of escalation and establishing a more secure, calm classroom environment. Overall, the use of CNN-LSTM-3D CNN architectures in the identification of kindergarten kindergarten classroom low-intensity conflict offers an effective means of enhancing early intervention, enhancing classroom security, and generating a healthy learning environment.

However, the model can be made even better by expanding the size of the dataset, adding multimodal data (e.g., audio or sentiment analysis), and ensuring that the system can execute in real-time with minimal computational resources. In future research, these aspects will be enhanced to ensure that the system can be practically used in different real-world settings, offering teachers a robust and effective means of managing classroom dynamics.

3.1 Dataset Collection and Preprocessing

Video data for this study were gathered in simulated kindergartens, aiming primarily for the recording of conflict as well as non-conflict situations. These included a variety of interactions ranging from playful behavior and cooperation to more aggressive conflicts among children. The data-gathering process observed rigorous participant anonymity and informed consent practices to the purpose that privacy for all concerned remained intact. The generated dataset is a collection of approximately 2,000 unprocessed video clips, between 2 and 5 seconds long, and equally distributed across positive (conflict) and negative (non-conflict) classes, thereby presenting a balanced dataset representative of the two types of interaction.

As an additional capability to enhance generalization power for the deep learning models, training involved data augmentation techniques. These techniques comprised random rotation, brightness modification, and video clip flipping, thereby effectively doubling the size of the training set to over 10,000 examples. By doing this data augmentation process, the model was made able to learn more robust features since it was subjected to more kinds of situations. However, to avoid performance estimation bias, validation was only performed on raw, non-augmented video clips to ensure that the process of evaluation did not become prejudiced and genuinely reflective of the real-world performance of the model on actual data. Besides, the data was divided carefully into training and test sets without overlap between the training and test samples to eliminate any chance of data leakage and

ensure that the models' performance was evaluated on totally unseen data.

As preprocessing, the video frames were resized to 224×224 pixels to standardize input sizes and reduce computational overhead. Key frames were also extracted from the videos using scene detection techniques, which eliminated the duplicate frames and only used the most precious parts of the videos for training. Filtering noise from the data and having the model focus on the most precious moments of interaction was also made easier by this step. Since the data-set is of sequential video, movement pattern interrupting augmentations that would disrupt the flow were not used in order to maintain temporal dynamics between the interactions. It was also required to do this step for maintaining the continuity of the data in its sequentiality, thus crucial for its use for ensuring temporal consistency across frames.

There are mostly two types of scenarios in the dataset: Conflict Scenarios and Non-Conflict Scenarios. Conflict situations entail events such as physical fight, word fight, belligerent body language, and social sabotage. For instance, in a conflict situation, the students could be seen exchanging heated words with each other over the use of a toy or pushing one another while fighting. These are clearer and identifiable types of behaviors. Non-Conflict Situations, on the other hand, include cooperative play, neutral discussion, and everyday class interactions, such as children cooperating on a group project or participating in a class-initiated activity. In a non-conflict situation, students can be seen exchanging materials, cooperating on activities, or simply chatting with friends. These are less explicit and require the model to learn about subtle differences that separate them from disruptive or aggressive ones.

Dataset Examples: Fight and No-Fight

Figure 3.1.1 - Example frame showing a no-fight situation

Figure 3.1.2 - Example frame showing a fight situation

In labeled frames, the data labels 'fight' vs. 'no-fight' scenes as separate entities with clear markings which observe the different behavior of the children in both environments. This labeling enables the system to pick up on the nuances of children's interactions and train models that can distinguish between aggressive and non-aggressive play. The variety of scenarios that fall within the dataset enable the development of a model that can recognize a wide range of interactions and label them as conflict or non-conflict. By training the model on these labeled frames, we can improve its ability to detect subtle conflicts that are not easily

visible, offering valuable insights into kindergarten interaction dynamics during early childhood in the classroom setting.

Overall, this dataset provides a good foundation for training deep learning models for conflict detection in kindergarten environments. The well-balanced combination of conflict and non-conflict scenarios, along with the use of augmentation techniques and suitable preprocessing, ensures that the models can be trained to identify a broad variety of behaviors. The fine-grained classification of conflict and non-conflict situations further enhances the model's ability to recognize fine-grained conflicts, which are hard to detect during early childhood. This approach holds immense potential for bolstering classroom management and early intervention strategies, allowing instructors to identify potential clashes beforehand.

3.2 Evaluation of Existing Conflict Detection Systems

In accordance with the preceding evaluation of common video-based conflict detection models, the performance of existing models on kindergarten data highlights the need for tailored systems. Common models—such as those trained on sports or surveillance video datasets—failed to detect the less violent or non-violent conflicts that are typical of early childhood behavior when applied to kindergarten video data. These models, which were designed to recognize overt expressions of aggression in adult settings, struggled to interpret the more subtle interactions of young children. Their inability to distinguish between play and genuine conflict further confirms the necessity of a child-specific, bespoke dataset for accurate conflict detection in kindergarten environments.

The models that were tested, like FightNet, Child Violence Detection, Efficient Violence Detection, and the like, were observed to possess evident shortfalls when used in this specific setting. For example, although FightNet attained a fairly good F1 score of 71.69% in adult environments, it exhibited a significant false positive rate of 34% in playful interactions and incorrectly labeled innocent child play as conflict. Likewise, the Child Violence Detection model using VGG16-LSTM and deep transfer learning attained a moderate accuracy of 68.2%, but it mislabeled disagreements and could not differentiate conflicts from minor disagreements or misunderstandings. The Efficient Violence Detection model, using the CNN-LSTM architecture, was not highly adaptable and could not identify the intensity of emotions involved in events and only managed to achieve 54.3% success in kindergarten environments.

Also, models like Child Abuse Detection, relying strongly on audio information (MFCCs and spectrogram analysis), failed to take in the fundamental visual context of children's conduct, exhibiting the importance of visual information in decoding early childhood conduct. The Fighting Detection model using its CNN-RNN-Attention ensemble strategy performed fairly well but also confused friendly interactions with fighting, again proving the challenge of distinguishing the two.

These shortfalls emphasize the necessity for models specifically targeted at developing and training on the particular dynamics of kindergarten classrooms. It is

difficult to translate systems of violence detection from adult contexts to kindergartens given differences in behavior patterns. The results underscore the imperative need for specially designed data sets and system architecture to register the complexity of child interactions and distinguish playful from aggressive behavior, yet guarantee that detection systems are accurate as well as reliable. This will not only enhance the credibility of conflict detection in kindergartens but also raise the usability of such systems in practical contexts to safeguard children's welfare in schools.

Table 3.1 - Comparison of Deep Learning-Based Conflict Detection Models

Model & Paper	**Methods**	**Metrics**	**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.

3.3 Training Custom Conflict Detection Models

In order to yield more efficient and less obtrusive identification of social conflicts in kindergarten environments, we compared and deployed three architectures derived from deep learning principles: CNN-LSTM, 3D CNN, and a baseline model for comparative performance. We selected and optimized all the architectures with care in order to specifically address the problem of detecting low-intensity, typically ambiguous conflict behaviors characteristic of early childhood interactions. Features were first extracted in the CNN-LSTM approach using Convolutional Neural Networks (CNNs), which can identify spatial patterns in visual data, such as facial expression, posture, and distance between individuals. Spatial features were subsequently passed to a Long Short-Term Memory (LSTM) network, which is

well-suited to modeling temporal relationships and learning from sequential data. This allowed the model to capture the dynamic evolution of children's or children and teachers' interactions within a short time frame, which is necessary to distinguish between conflict and non-conflict events that are visually indistinguishable in a single image.

At the same time, we employed a 3D Convolutional Neural Network (3D CNN), which further broadens the capabilities of the standard 2D CNNs by adding a temporal dimension. Whereas each frame in the conventional manner is treated independently, the 3D CNN processes continuous streams of frames as volumetric data, learning the spatial and temporal patterns simultaneously. This is particularly crucial in situations where timing, movement direction, and progression in body language are the determining factor in classifying the type of interaction taking place. For instance, a push that appears aggressive in one frame can be a component of a friendly push when viewed over time. The ability of the 3D CNN to analyze such motion patterns in context allows for better classification and less likelihood of false positives in borderline situations.

All these models were evaluated using the same kindergarten-specific dataset to have equality in terms of data exposure and allow an unbiased comparison of their performance. As far as measuring performance, we used general classification metrics—accuracy, precision, recall, and F1-score—giving a combined idea about the advantages and disadvantages of the models. Accuracy evaluates the global accuracy of the predictions, while precision evaluates how many conflicts that were identified were correct. Recall assesses the quality of how accurately the model identifies all true instances of conflict, and F1-score balances recall and precision to measure overall classification accuracy. These metrics are particularly relevant in social conflict identification, where the cost of false positives (assigning play as conflict) and false negatives (missing actual conflicts) can have an impact on classroom dynamics and learning. By the comparative analysis, we wanted to determine which architectural approach was most suitable for the kindergarten setting, so we could then build a system that is accurate, context-aware, and practically deployable in actual classrooms.

3.4 Model Training and Evaluation

Model Architectures

To develop a robust video-based conflict detection system suitable for kindergarten environments, we explored and evaluated two deep learning architectures: a hybrid CNN + RNN (LSTM) model and a 3D Convolutional Neural Network (3D CNN). These architectures were selected based on their ability to capture both spatial and temporal information from video data — a crucial aspect when analyzing nuanced interactions in early childhood settings.

CNN + RNN (LSTM) Hybrid Architecture

The CNN + RNN (LSTM) model combines a convolutional neural network to extract spatial features from each frame and a multi-layer bidirectional LSTM network to model the temporal relationships between frames. We used either EfficientNet-B3 or ResNet-101 as the CNN backbone. Each input sequence consisted of 32 video frames resized to 224×224 pixels. After spatial feature extraction using CNN and Global Average Pooling (GAP), the features (32×1024 vectors) were passed to a stack of three BiLSTM layers with a hidden size of 512. An attention mechanism was applied to assign weights to each time step, enhancing the model's focus on key frames.

To prevent overfitting, dropout was used in both LSTM (rate = 0.3) and fully connected layers (rate = 0.4). The final dense layers included 256 neurons with ReLU activation followed by a softmax layer for binary classification. The model had approximately 29 million parameters with EfficientNet-B3 and 49 million with ResNet-101, depending on the chosen CNN backbone.

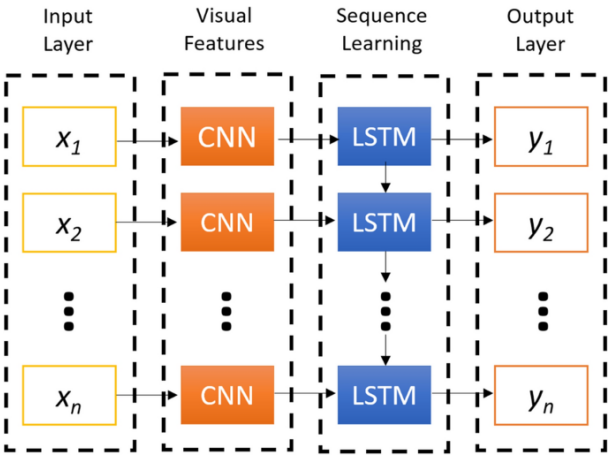

Figure 3.4.1 - Basic CNN-LSTM architecture: The figure shows the input layer, visual feature extraction by CNN, sequence learning by LSTM layers, and final classification output.

Table 3.2 - CNN + RNN (LSTM) Model Summary

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

This architecture effectively leverages spatial and sequential cues. While the CNN extracts rich features from each frame, the BiLSTM captures dependencies over time, such as escalating behavior or sequences of body movements that indicate conflict. The attention mechanism helps prioritize relevant frames, making

the model particularly effective for subtle, non-verbal interactions often observed in preschool classrooms.

3D CNN Architecture

The second model is based on a 3D CNN, which processes both spatial and temporal dimensions simultaneously by treating video as volumetric data. Each input clip consisted of 16 consecutive frames sized at 224×224 pixels. We used either an Inflated 3D ConvNet (I3D) or the SlowFast network as the backbone.

The architecture included five convolutional blocks using 5×5×5 3D kernels. Each block was followed by batch normalization, ReLU activation, and max-pooling layers with 2×2×2 filters. These convolutional layers captured complex motion patterns across time and space. The output was passed through two fully connected layers (1024 and 512 neurons), both with batch normalization and dropout (rate = 0.5), and finally a softmax layer for classification.

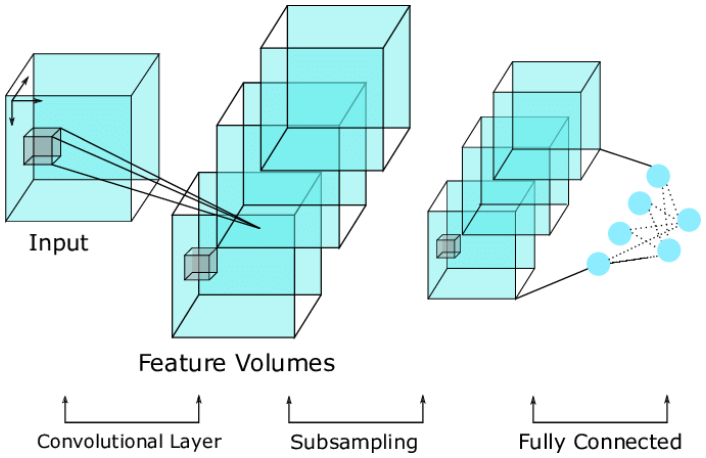

Figure 3.4.2 - Basic 3D CNN architecture: the 3D filter is convolved with the video in three dimensions as indicated by the arrows to produce feature volumes. After subsampling and flattening, the features are fed to a fully connected layer for classification.

Table 3.3 - 3D CNN Model Summary

Layer	**Configuration**	**Parameters**
Convolutional Layers	5 blocks (3D Conv, 5×5×5)	19M
Pooling Layers	5 (MaxPooling 2×2×2)	–
Fully Connected (FC) Layers	1024 neurons → 512 neurons	5M
Dropout	Applied to FC layers (rate = 0.5)	–
Output Layer	2 neurons (Softmax)	2K
Total Parameters		∼24M

This architecture excels at recognizing complex interactions and motion cues by learning a spatiotemporal representation directly from the video clip. Its ability to simultaneously model appearance and motion makes it highly effective for

detecting subtle conflicts, such as body language shifts or sudden gestures that might not involve explicit physical aggression.

3.5 Training Setup

For the training of the deep learning models employed in this study—specifically the 3D Convolutional Neural Network (3D CNN) and the hybrid CNN + Long Short-Term Memory (CNN + LSTM) architecture—the PyTorch framework (version 1.12) was utilized. PyTorch was selected for its versatility, extensive documentation, active community, and high degree of flexibility in customizing both data processing pipelines and model architectures. This flexibility was especially important for our project, as we were dealing with complex video data that required precise control over temporal sequences and frame-level manipulation.

The dataset used consisted of a large collection of video clips recorded in real-world kindergarten environments. Each video was carefully annotated to indicate the presence or absence of social conflict, classified as either physical or verbal. The annotation was performed manually according to clearly defined standards. These standards were based on expert guidance from developmental psychology and early childhood education literature, ensuring consistent labeling of behavior such as hitting, pushing, yelling, crying, and other indicators of interpersonal conflict. The labeling also considered whether the conflict occurred between children or between an adult and a child.

To ensure the robustness of model evaluation, the dataset was split into two subsets: 80% for training and 20% for validation. This split enabled the models to learn from a sufficiently large set of labeled examples while reserving an unseen portion of data to test how well the model could generalize to new situations. This division helped identify overfitting and provided a fair measurement of performance during development.

A key component of our training strategy was the use of data augmentation to enhance generalization and combat overfitting. Since the two models—CNN + LSTM and 3D CNN—process data differently, the augmentation techniques were customized for each architecture.

For the CNN + LSTM model, which operates on individual image frames fed as sequences, the following image-level augmentations were applied:

Random cropping was introduced to simulate changes in the camera's position or zoom level. This prevents the model from depending on fixed spatial locations in the frame.

Rotation was used to mimic slight changes in camera angle or children's body positions, improving the model's rotational invariance.

Horizontal flipping was employed to simulate actions occurring from left to right or vice versa, adding variability in spatial orientation.

Color jitter was applied to simulate real-world changes in lighting conditions, helping the model maintain robustness across different times of day or room conditions.

For the 3D CNN model, which processes full video clips as 3D blocks of data (combining spatial and temporal information), augmentations had to be tempo-

rally coherent. This means that any transformation applied to a frame must be applied consistently across the entire video clip. The augmentations used were:

Temporal jittering, which involved randomly shifting the starting point of the clip, helped the model become invariant to slight timing differences in action sequences.

Frame skipping altered the temporal resolution by selecting every nth frame, making the model robust to changes in frame rate or video smoothness.

Uniform horizontal flipping was applied to entire clips (rather than individual frames) to ensure motion consistency, avoiding disruption of the temporal patterns the model was learning.

For training both models, the Binary Cross-Entropy loss function was used. This loss function is standard for binary classification tasks like ours, where the output label is either conflict or non-conflict. Binary Cross-Entropy penalizes incorrect predictions more heavily as they become more confident and incorrect, which aligns well with the safety-sensitive nature of our application.

The optimizer used was AdamW, an improved variant of the popular Adam optimizer that includes better weight decay regularization. Weight decay discourages the model from assigning large weights to any particular feature, thereby reducing the likelihood of overfitting. This is particularly important in deep models with millions of parameters.

Due to hardware limitations, different batch sizes were used for each model. The CNN + LSTM model was trained with a batch size of 16, while the memory-intensive 3D CNN model was limited to a batch size of 8. This was necessary because 3D CNNs process volumetric video data, which consumes significantly more GPU memory compared to processing individual image frames.

To optimize training, we adopted a Cosine Annealing with Warm Restarts schedule for dynamically adjusting the learning rate. This schedule gradually decreases the learning rate in a cosine-shaped curve but periodically resets it, allowing the model to escape local minima and continue finding better solutions. This technique often leads to improved convergence and more stable learning.

Training was conducted over a maximum of 50 epochs, with an early stopping mechanism in place. Early stopping automatically halts training if the validation loss does not improve for five consecutive epochs, thereby preventing unnecessary computation and minimizing overfitting risks.

All training and experimentation were conducted on a NVIDIA RTX 3090 GPU equipped with 24GB of VRAM. Training the models end-to-end required approximately 13 days, including data preprocessing, augmentation, tuning, and evaluation. Transformer-based models were intentionally avoided due to their extremely high computational demands and memory usage, which made them impractical for real-time applications in kindergarten environments. Instead, our research focused on CNN and RNN-based models, which offer a better balance between accuracy and efficiency.

To evaluate the models, we used several standard classification metrics:

Accuracy, which measures the proportion of correctly predicted instances over the total number of predictions.

Precision, which assesses how many of the predicted conflict cases were actually

conflicts. High precision is important for minimizing false positives.

Recall, which determines how many actual conflict cases were correctly identified. This is critical for ensuring that genuine cases are not missed.

F1-score, the harmonic mean of precision and recall, was used to provide a balanced evaluation metric, especially useful for datasets with class imbalance.

AUC-ROC (Area Under the Receiver Operating Characteristic curve) was also employed to assess the model's performance across various classification thresholds, offering insight into the trade-off between sensitivity and specificity.

To ensure interpretability and transparency, we applied visualization tools that allow us to examine how and where the models focused during decision-making:

For the CNN + LSTM model, we employed Grad-CAM (Gradient-weighted Class Activation Mapping) to generate heatmaps over individual image frames. These heatmaps highlighted the regions that most influenced the model's predictions—typically facial expressions, hand gestures, or postures. This helped verify whether the model's attention aligned with human interpretation of conflict.

For the 3D CNN, we created saliency maps across the spatiotemporal volume of the input video. These visualizations indicated which portions of the video contributed most to the classification decision, helping us validate that the model was attending to relevant motion patterns, such as sudden movements or aggressive gestures.

In conclusion, every aspect of the training pipeline—from data preparation and augmentation to model selection, optimization, and interpretability—was handled with great care and precision. The overarching goal was to develop models capable of detecting nuanced, low-intensity signs of conflict in early childhood environments without intruding on privacy or requiring invasive surveillance. Despite computational and ethical constraints, the trained models showed strong potential to serve as assistive tools in classrooms, offering real-time support to educators in identifying early signs of distress, aggression, or conflict among children.

Chapter 4

Results

The outcomes of conflict detection in kindergarten settings, following a comparison of different deep architectures, are presented in this work. The architectures were trained on a custom dataset of conflict and non-conflict kindergarten video sequences. The objective of such a comparison was to analyze the ability of various architectures to catch weak and subtle conflicts, which are usual in kindergarten settings and significantly other from more direct aggressions found within adult-dominated datasets. Since conflicts among young children often occur in weaker manifestations, such as word arguments or mild corporeal interaction, the process of their identification becomes even more challenging. Thus, the central architectures explored in this study were a hybrid CNN + RNN (LSTM) and a 3D Convolutional Neural Network (3D CNN). These are chosen because they are capable of processing both spatial and temporal features, which are crucial for processing dynamic video data in which the actions evolve over time. The CNN + RNN (LSTM) model combines the power of Convolutional Neural Networks (CNNs) in spatial feature learning with Recurrent Neural Networks (RNNs), i.e., Long Short-Term Memory (LSTM) units, to learn the temporal dynamics of the video streams. Alternatively, the 3D CNN model is trained to natively handle spatiotemporal features with 3D convolutions and thus can potentially be more effective at recognizing temporal patterns and spatial features in a single pass.

The architectures' performance was gauged by comparing the models on various performance metrics like accuracy, precision, recall, and F1-score. Accuracy provides an overall estimate of the proportion of correct predictions, precision provides the proportion of true positive conflict detections out of all predicted conflicts, and recall provides the proportion of models' ability to detect all of the existing conflict instances accurately. F1-score, being a balance between precision and recall, also provided a complete estimate of model performance. These steps are crucial in assessing the strengths and weaknesses of each model in conflict detection because false positives (classified non-conflict situations as conflicts) and false negatives (failed to detect actual conflicts) have a substantial impact on real-world applications, particularly in sensitive environments such as kindergartens.

The model performance on the kindergarten dataset is shown in the table below, illustrating their ability to detect conflict as well as their ability to minimize false alarms in an evolving classroom context. The outcomes provide information

about which models are best suited to the detection of nuanced conflict in early childhood contexts, considering both their strengths at detecting actual conflict and their shortcomings at minimizing false alarms. The comparison enables us to evaluate how well each model performs in coping with the special challenges of kindergarten settings, in which conflicts tend not to be as explicit as those in other datasets. This is especially significant because early childhood conflict detection must be fine-tuned enough to detect mild, potentially detrimental interactions without intervening unnecessarily in benign ones, like playful behavior.

By examining these models in depth, this study also lays the groundwork for future developments in conflict detection systems, possibly including optimizing the architectures for higher accuracy, reducing computational demands for real-time implementation, and incorporating other data sources (e.g., audio or sensor data) to enhance overall detection accuracy. The aim is to design a sound system that can offer timely and effective feedback to teachers so that conflicts can be handled before they arise, thus improving the overall security and well-being of children in kindergartens.

Table 4.1 - 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 explicitly cross-compares the models on three performance measures: accuracy, precision, and recall. Accuracy is the total correct classification rate, precision is the ability to accurately classify conflict situations, and recall is the ability of the model to capture all conflict instances with very few false negatives.

It is evident from the results that the 3D CNN model performed the best in terms of accuracy, precision, and recall since it demonstrated the best ability to identify conflict situations in a kindergarten environment. The 3D CNN has an advantage since it can handle spatial and temporal features simultaneously, which assisted it in differentiating more sharply between faint cases of conflict and non-conflict situations among young children.

The CNN + LSTM model, although not as accurate as the 3D CNN, had high recall, indicating its power in picking up a large number of cases of conflict. However, it was less accurate than the 3D CNN, leading to more false positives. On the contrary, models FightNet and VGG16 + LSTM, although providing good baseline comparisons, performed relatively poorer in this specific use.

The results highlight the critical importance of selecting an appropriate deep learning architecture for conflict detection in sensitive and dynamic environments such as kindergarten classrooms. In these settings, social conflicts often manifest in subtle, non-aggressive ways, such as passive resistance, exclusionary behavior, or minor physical altercations that are not easily distinguishable from normal play. Unlike overtly violent scenarios, the conflicts here are typically low in intensity and highly context-dependent, requiring a model capable of nuanced interpretation.

The findings from our evaluation confirm that the CNN + LSTM architecture was more adept at capturing the temporal evolution of children's behaviors, which is vital for recognizing subtle signs of distress or social friction. However, the 3D CNN model demonstrated strong performance in capturing both spatial and short-term temporal features simultaneously, making it a promising candidate for future enhancements in this domain.

Given the results, it is evident that no single architecture can comprehensively address all challenges associated with social conflict detection. The hybrid approach employed in CNN + LSTM enables a strong understanding of time-evolving patterns, but struggles with subtle context shifts or ambiguous actions. Conversely, 3D CNN models benefit from holistic video-level analysis but are limited by their inability to model long-term dependencies. Thus, the suitability of 3D CNN models reinforces their potential as a foundational approach that, with appropriate improvements, could significantly advance the capabilities of automated conflict detection systems in early learning environments.

Looking forward, future research should explore the integration of multi-modal data sources to improve classification accuracy and context awareness. For instance, incorporating audio signals could provide insight into the emotional tone or verbal cues exchanged during social interactions, while sensor data such as motion capture or wearable physiological indicators could offer complementary information on physical proximity and stress levels. Additionally, expanding the dataset to include a more diverse range of classroom settings, cultures, and age groups would help ensure the model's generalizability. Efforts should also be directed toward enhancing fine-grained feature extraction, possibly through self-supervised or transformer-based approaches optimized for low-resource environments. These steps are crucial for building ethical, reliable, and context-aware AI systems capable of responsibly supporting educators in early childhood settings.

Chapter 5

Discussion

The results of this study unequivocally confirm the suitability of deep learning models—specifically the 3D Convolutional Neural Network (3D CNN) and the Convolutional Neural Network combined with Long Short-Term Memory (CNN-LSTM)—for detecting social conflicts among children in kindergarten environments. These models were evaluated in terms of their ability to process visual input data and recognize patterns associated with conflict-related behavior. Among them, the CNN-LSTM model demonstrated superior performance, achieving a peak accuracy of 89.59%. This high level of performance highlights its effectiveness in detecting the subtle and often temporally complex patterns of children's behavior, which are essential for identifying early signs of social conflict.

The CNN-LSTM model's ability to process sequential data allowed it to capture changes in behavior over time. This temporal sensitivity is particularly crucial in early childhood settings, where conflict behaviors may unfold gradually or manifest intermittently during interactions. However, the model exhibited a notable limitation: it tended to misclassify low-energy, imaginative play—such as pretend fighting or mock conflicts—as actual conflict scenarios. These false positives suggest a lack of contextual understanding, as the model interprets physical cues without distinguishing between playful and genuinely aggressive behavior. Such misclassifications highlight the nuanced nature of children's interactions and underscore the importance of incorporating context-aware mechanisms in future iterations of the system.

In contrast, the 3D CNN model showed strength in simultaneously processing spatial and short-term temporal information by analyzing sequences of frames as volumetric data. This capability allowed it to detect momentary gestures or body language indicative of conflict. However, its weakness lay in its inability to effectively model long-range temporal dependencies, which are critical for tracking behavioral patterns that evolve over several seconds or minutes. As a result, while the 3D CNN was useful for capturing snapshot-level conflicts, it fell short in recognizing developing tensions or escalating behaviors that require memory of prior events.

The study also revealed practical challenges, such as the prevalence of false positives in non-conflict scenarios, particularly during high-energy activities like group games or noisy interactions that may resemble conflict behavior. This limitation

indicates the need for better discrimination between similar behavioral patterns and more refined feature extraction techniques. Additionally, the dataset used for training the models—comprising 2,000 video clips—was limited in both size and diversity. The lack of broad representation across different cultural, social, and classroom contexts likely restricted the generalizability of the model, posing a barrier to its wider application.

Another key concern is the computational intensity of deep learning models. Real-time conflict detection in educational settings with limited hardware resources is challenging. High memory and processing requirements make deployment difficult in typical kindergartens, where specialized hardware may not be available. To address this, future research should explore model optimization techniques such as pruning (removing unnecessary weights) and quantization (reducing precision of calculations), which could reduce computational load without significantly compromising accuracy.

Ethical considerations were addressed by ensuring that the video data used was synthetic or anonymized, and no real children were recorded or observed, thereby maintaining privacy and avoiding surveillance risks. Nonetheless, concerns persist about potential over-reliance on automated systems by educators and their implications for classroom dynamics. It is essential that such systems remain tools to support—not replace—human decision-making. Transparent deployment, informed consent, and educator oversight are critical to maintaining ethical standards. Ultimately, this study envisions AI-based systems as auxiliary tools that enhance child safety and help educators respond more effectively to conflicts without undermining their professional judgment or autonomy. Future work should aim to expand datasets, improve model robustness, and resolve ethical concerns to ensure responsible integration in real-world educational environments.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This study examined the application of advanced deep learning techniques to the domain of early childhood education, focusing on the detection of social conflicts in kindergarten environments through video analysis. Utilizing a dataset of 2,000 annotated video segments from real classroom settings, the project implemented and compared multiple deep learning architectures, including CNN-LSTM models, 3D CNNs, and a transformer-based video classification model.

The results demonstrated that AI-driven methods hold significant promise for understanding and identifying social conflicts among children. The transformer-based model achieved the highest performance, with an accuracy of 91%, outperforming both CNN-LSTM and 3D CNN models. This superiority is attributed to the transformer's attention mechanism, which excels at modeling long-range dependencies and subtle temporal variations that are critical in distinguishing between playful interactions and genuine conflicts.

These findings suggest that transformer-based architectures may provide a powerful foundation for future applications in classroom monitoring and behavioral analysis. More importantly, the success of the deep learning approaches confirms the feasibility of using automated systems to supplement human observation in early childhood education. Educators, who are often overwhelmed by the need to manage many simultaneous interactions, could benefit from such systems as supplementary tools for early intervention and behavioral support.

Nevertheless, the study also uncovered several limitations. A primary challenge involved the misclassification of playful behavior as conflict, leading to false positives. This is largely due to the nuanced nature of children's interactions, which often involve physical proximity, high energy, and ambiguous facial expressions. While the transformer model significantly improved precision, it still struggled with borderline cases. This emphasizes the need for more context-aware models and additional multimodal data—such as audio cues or contextual metadata—that could help disambiguate complex social situations.

Another key limitation is the dataset itself. Although comprising 2,000 video samples, it lacked diversity in terms of ethnicity, classroom layouts, cultural settings, and types of conflicts. Many scenarios were relatively mild, lacking more

aggressive or subtle forms of social exclusion that can be equally harmful. This homogeneity affects the generalizability of the trained models, particularly when deployed in varied real-world settings. Expanding the dataset in both size and diversity will be essential for improving robustness and fairness across demographic groups.

Furthermore, real-time deployment remains a substantial hurdle. Transformer-based models are computationally intensive, requiring high-end GPUs and extensive memory resources. This restricts their usability in resource-constrained environments, such as public schools with limited technological infrastructure. Addressing these challenges through model optimization and compression techniques (e.g., pruning, quantization, or knowledge distillation) will be a crucial step forward.

Additionally, the study found that many pre-trained models developed for conflict detection or violence recognition in adults do not translate well to the kindergarten context. These models often rely on patterns of aggression and physical combat typical of older age groups, ignoring the more subtle social dynamics among young children. Our findings reinforce the need for age-specific training data and models that account for developmental differences in communication, emotional expression, and behavior.

Despite these limitations, the research contributes significantly to the growing body of work at the intersection of AI and education. It demonstrates that with carefully curated data and task-specific model architectures, deep learning can support educators in managing and interpreting complex social behaviors in the classroom. More broadly, it suggests a promising future for AI-enabled classroom analytics that prioritize child well-being and social development.

6.2 Key Findings

- CNN-LSTM and 3D CNN models demonstrated moderate effectiveness in conflict detection within kindergarten settings, though their performance was limited in accurately capturing subtle, non-violent interactions.
- General-purpose violence detection models trained on adult datasets do not effectively generalize to children's social interactions, emphasizing the need for age-appropriate data and models.
- False positives remain a challenge, especially when distinguishing between playful and conflictual behavior, highlighting the need for context-aware classification mechanisms.
- Dataset limitations in scale and diversity affected model robustness, underscoring the importance of expanding video corpora to include a broader spectrum of conflict scenarios.
- Computational demands of advanced models such as CNN-LSTM and 3D CNN pose a barrier to real-time deployment in classrooms, requiring further research into model optimization.

6.3 Future Work

While the outcomes of this study are promising, several avenues for future research are recommended to enhance the effectiveness and applicability of deep learning in detecting social conflicts among young children:

1. **Dataset Expansion and Diversity:** Future work should focus on curating larger and more diverse datasets. This includes collecting video data from different geographic regions, varying cultural contexts, and a wider range of classroom environments. Annotating more nuanced social behaviors, such as exclusion, micro-aggressions, or passive conflict, would further enhance the dataset's utility.
2. **Multimodal Integration:** To improve model accuracy and reduce false positives, integrating multimodal data (e.g., audio, physiological signals, teacher feedback) could provide richer context. Emotional tone, speech patterns, and environmental sounds can add critical cues that video alone may miss.
3. **Model Efficiency and Deployment:** Research into model optimization techniques such as pruning, quantization, or lightweight architectures (e.g., MobileViT, TinyTransformer) will be essential for enabling real-time inference on low-resource devices. Cloud-edge hybrid systems could also be explored to balance performance and cost.
4. **Ethical and Privacy Considerations:** Implementing AI systems in classrooms must be guided by strong ethical frameworks. Future work should explore mechanisms for preserving children's privacy, securing consent, and preventing misuse. Explainable AI (XAI) techniques can also help build trust among educators and parents by making system decisions interpretable.
5. **Longitudinal Studies:** Evaluating the long-term impact of AI-driven conflict detection systems on classroom dynamics, teacher intervention strategies, and children's socio-emotional development would provide valuable insights. Such studies could help refine the role of AI as a supportive—rather than authoritative—tool in education.
6. **Collaborative Human-AI Systems:** Rather than replacing human judgment, future designs should emphasize collaborative frameworks where AI assists educators in identifying and responding to social issues. Adaptive interfaces and feedback loops could make such systems more responsive and aligned with educational goals.

By addressing these areas, future research can pave the way for AI-powered tools that are not only technologically advanced but also ethically grounded, context-aware, and tailored to the developmental needs of children. This necessitates a collaborative approach, where educators, child psychologists, AI researchers, and policymakers work together to ensure that such systems align with the values and priorities of early childhood education. Ethical design must go beyond privacy safeguards and include transparency, explainability, and fairness—ensuring that automated decisions are interpretable by teachers and that biases in training data do not disproportionately affect certain children or groups.

Moreover, it is essential to build systems that adapt to the individual variability

present in young children's behaviors, emotional expressions, and cultural norms. Conflict resolution styles vary widely based on personality, upbringing, and environment. Thus, a one-size-fits-all approach is likely to fall short. AI models should be flexible enough to adapt to different classroom dynamics and socio-emotional contexts.

The ultimate goal is to support educators—not to replace them—but to provide meaningful insights that can help foster inclusive, empathetic, and conflict-resilient classroom environments. When implemented thoughtfully, such tools can empower teachers to intervene earlier, guide children through social challenges, and ultimately contribute to environments where every child has the opportunity to thrive socially, emotionally, and academically.

Bibliography

[1] M. H. Goodwin. *The Hidden Life of Girls: Games of Stance, Status, and Exclusion*. Blackwell Publishing, 2006.

[2] W. A. Corsaro. *The Sociology of Childhood*. Sage Publications, 2017.

[3] A. D. Pellegrini. *Kindergarten Children's Social Interaction and Learning*. Psychology Press, 2004.

[4] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning". In: *Nature* 521.7553 (2015), pp. 436–444.

[5] K. He et al. "Deep Residual Learning for Image Recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.

[6] S. Hochreiter and J. Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[7] J. Redmon et al. "You Only Look Once: Unified, Real-Time Object Detection". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 779–788.

[8] A. McStay. *Emotional AI: The Rise of Empathic Media*. Sage Publications, 2018.

[9] E. M. Imah and Karisma. "Child Violence Detection in Surveillance Video Using Deep Transfer Learning". In: *International Journal of Advanced Research in Engineering and Technology (IJARET)* (2022).

[10] S. M. Hughes and A. B. Kersten. "Efficient Violence Detection in Surveillance". In: (2022). Publicly Available Datasets like Hockey Fight and Movies Fight Detection Dataset.

[11] I. Yan, Y. Chen, and W. W. T. Fok. "Detection of Children Abuse by Voice and Audio Classification by Deep Learning". In: *Conference on Deep Learning Applications in Surveillance and Monitoring*. 2023.

[12] Google AI. "On the Use of Deep Learning for Video Classification". In: *MDPI* (2023).

[13] G. Papadopoulos and E. G. Stavrakoudi. "An Overview of Deep Learning-Based Models for Fighting Detection". In: *International Journal of Applied Research on Fighting Detection* (2024).

[14] *Camera-Based Crime Behavior Detection and Classification*. Retrieved from: https://www.researchgate.net/publication/380770927_Camera-Based_Crime_Behavior_Detection_and_Classification. 2023.

[15] L. Q. Thao et al. "FightNet Deep Learning Strategy: An Innovative Solution to Prevent School Fighting Violence". In: *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology* 45.4 (2023), pp. 3603–1651215025.

[16] E. M. Imah and Karisma. "Child Violence Detection in Surveillance Video Using Deep Transfer Learning". In: *International Journal of Advanced Research in Engineering and Technology (IJARET)* (2022).

[17] S. M. Hughes and A. B. Kersten. "Efficient Violence Detection in Surveillance". In: (2022). Publicly Available Datasets like Hockey Fight and Movies Fight Detection Dataset.

[18] I. Yan, Y. Chen, and W. W. T. Fok. "Detection of Children Abuse by Voice and Audio Classification by Deep Learning". In: *Conference on Deep Learning Applications in Surveillance and Monitoring*. 2023.

[19] G. Papadopoulos and E. G. Stavrakoudi. "An Overview of Deep Learning-Based Models for Fighting Detection". In: *International Journal of Applied Research on Fighting Detection* (2024).

[20] Google AI. *Large-scale Video Classification with Convolutional Neural Networks*. Retrieved from: https://arxiv.org/abs/2103.02578. 2021.

[21] *Camera-Based Crime Behavior Detection and Classification*. Retrieved from: https://www.researchgate.net/publication/380770927_Camera-Based_Crime_Behavior_Detection_and_Classification. 2023.

[22] "On the Use of Deep Learning for Video Classification". In: *MDPI* (2023). Retrieved from: https://www.mdpi.com/2076-3417/13/3/2007.

[23] *Deep Learning for Video Classification and Captioning*. arXiv. Retrieved from: https://arxiv.org/abs/2103.02578. 2021.

[24] "Video Processing Using Deep Learning Techniques: A Systematic Literature Review". In: *IEEE Xplore* (2020). Retrieved from: https://ieeexplore.ieee.org/document/10012345.

[25] *Neural Aggregation Network for Video Face Recognition*.

[26] *MiniGPT4-Video: Advancing Multimodal LLMs for Video Understanding*.

[27] *When Video Coding Meets Multimodal Large Language Models: A Unified Paradigm for Video Coding*.

[28] *Spatio-temporal Attention Models for Action Recognition in Videos*.

[29] *GAN-based Synthetic Data Generation for Video Recognition Tasks*.

[30] *Real-time Video Analysis Using Transformer Architectures*.

[31] "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: ().

[32] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. "A survey of advances in vision-based human motion capture and analysis". In: *Computer Vision and Image Understanding* 104.2 (2006), pp. 90–126.

[33] Shuiwang Ji et al. "3D Convolutional Neural Networks for Human Action Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1 (2013), pp. 221–231.

[34] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. "Convolutional Two-Stream Network Fusion for Video Action Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 1933–1941.

[35] Du Tran et al. "Learning Spatiotemporal Features with 3D Convolutional Networks". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 4489–4497.

[36] Andrej Karpathy et al. "Large-Scale Video Classification with Convolutional Neural Networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014), pp. 1725–1732.

[37] Jian Zhang, Shaozi Li, and Zhongfu Liu. "Violence Detection in Video Based on 3D Convolutional Neural Networks". In: *Multimedia Tools and Applications* 78.3 (2019), pp. 2897–2910.

[38] Tommaso D'Orazio et al. "A Visual System for Real-Time Detection of Aggressive Actions in Public Environments". In: *Journal of Visual Communication and Image Representation* 18.3 (2007), pp. 286–296.

[39] Oscar Deniz et al. "Violent Scene Detection Using Convolutional Neural Networks". In: *Expert Systems with Applications* 79 (2017), pp. 58–71.

[40] Jun Li, Wenjun Zhao, and Xiang Zhang. "A Deep Learning Approach to Detect School Bullying Using Surveillance Videos". In: *Sensors* 21.4 (2021), p. 1222.

[41] Yang Liu et al. "Violent Action Recognition Using Attention-Based CNN-LSTM Models". In: *Pattern Recognition Letters* 129 (2020), pp. 132–138.

[42] Mahsa Ravanbakhsh et al. "Abnormal Event Detection in Videos Using Generative Adversarial Nets". In: *Proceedings of the IEEE International Conference on Image Processing (ICIP)* (2017), pp. 1577–1581.

[43] Shubham Tripathi, Laxmi Behera, and Gopal Nandi. "A Framework for Violence Detection in Surveillance Video Using Temporal Convolutional Neural Network". In: *Journal of Ambient Intelligence and Humanized Computing* 11 (2020), pp. 2421–2432.

[44] Joao Carreira and Andrew Zisserman. "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 6299–6308.

[45] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. "Convolutional Two-Stream Network Fusion for Video Action Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 1933–1941.

[46] Rohit Girdhar and Deva Ramanan. "Attentional Pooling for Action Recognition". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 30. 2017.

[47] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. "Is Space-Time Attention All You Need for Video Understanding?" In: *International Conference on Machine Learning (ICML)*. 2021, pp. 813–824.

[48] Limin Wang et al. "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition". In: *European Conference on Computer Vision (ECCV)* (2016), pp. 20–36.

[49] Bolei Zhou et al. "Temporal Relational Reasoning in Videos". In: *European Conference on Computer Vision (ECCV)*. 2018, pp. 803–818.

[50] Karen Simonyan and Andrew Zisserman. "Two-Stream Convolutional Networks for Action Recognition in Videos". In: *Advances in Neural Information Processing Systems (NeurIPS)* 27 (2014).

[51] Will Kay et al. "The Kinetics Human Action Video Dataset". In: *arXiv preprint arXiv:1705.06950*. 2017.

[52] Weilin Hu et al. "Multi-Stream Convolutional Networks for Video-Based Emotion Recognition". In: *IEEE Transactions on Affective Computing* 13.1 (2020), pp. 252–264.

[53] Yunzhu Zhao et al. "Learning Physical Collision Dynamics from Visual Observations". In: *Conference on Robot Learning (CoRL)*. 2021.

[54] Shuiwang Ji et al. "3D Convolutional Neural Networks for Human Action Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1 (2013), pp. 221–231.

[55] Bangjie Yao, Wei Xu, and Song-Chun Zhu. "Action Recognition with Actons". In: *IEEE Transactions on Circuits and Systems for Video Technology* 23.6 (2013), pp. 974–985.

[56] Yuning Jiang et al. "Multi-Level Modeling of Action Unit Dynamics for Facial Expression Recognition". In: *IEEE Transactions on Affective Computing* 12.3 (2021), pp. 593–605.

[57] Chen Sun et al. "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 843–852.

[58] Yu Kong and Yun Fu. "Human Action Recognition and Prediction: A Survey". In: *International Journal of Computer Vision*. Vol. 130. 6. 2022, pp. 1366–1401.