Department of Electrical and Computer Engineering
North South University

Senior Design Project

Age Suitability Test Based on Video Analysis

NOURASH AZMINE CHOWDHURY	2012017042
DIPTA NEOGI	2013100042
MST MORIOM AKTER	2013049642

Faculty Advisor:

Dr. Mohammad Ashrafuzzaman Khan (AZK)

Assistant Professor

ECE Department

Spring, 2024

LETTER OF TRANSMITTAL

May, 2024

To

Dr. Rajesh Palit
Chairman,
Department of Electrical and Computer Engineering
North South University, Dhaka

Subject: **Submission of Capstone Project Report on "Age Suitability Test Based on Video Analysis"**

Dear Sir,

With due respect, we would like to submit our **Capstone Project Report** on **"Age Suitability Test Based on Video Analysis"** as a part of our BSc program. The report deals with classifying video-based age certification. This project was precious to us as it helped us gain experience in the practical field and apply it in real life. We tried to the maximum competence to meet all the dimensions required from this report.

We will be highly obliged if you kindly receive this report and provide your valuable judgment. It would be our immense pleasure if you find this report useful and informative and have an apparent perspective on the issue.

Sincerely Yours,

.....
Nourash Azmine Chowdhury
ECE Department
North South University, Bangladesh

.....
Dipta Neogi
ECE Department
North South University, Bangladesh

.....
Mst. Moriom Akter
ECE Department
North South University, Bangladesh

APPROVAL

Nourash Azmine Chowdhury (ID # 2012107042), Dipta Neogi (ID # 2013100042), and Mst. Moriom Akter (ID # 2013049642) from the Electrical and Computer Engineering Department of North South University has worked on the Senior Design Project titled "Age suitability test based on Video" under the supervision of Dr. Mohammad Ashrafuzzaman Khan, partial fulfillment of the requirement for the degree of Bachelor of Science in Engineering and has been accepted as satisfactory.

Supervisor's Signature

…….

Dr. Mohammad Ashrafuzzaman Khan

Assistant Professor

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

Chairman's Signature

…….

Dr. Rajesh Palit

Professor

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

DECLARATION

This is to declare that this project is our original work. No part of this work has been submitted elsewhere partially or fully for the award of any other degree or diploma. All project related information will remain confidential and shall not be disclosed without the formal consent of the project supervisor. Relevant previous works presented in this report have been properly acknowledged and cited. The plagiarism policy, as stated by the supervisor, has been maintained.

Students' names & Signatures

_____

1. Nourash Azmine Chowdhury

_____

2. Dipta Neogi

_____

3. Mst Moriom Akter

ACKNOWLEDGEMENTS

The authors would like to express their heartfelt gratitude towards their project and research supervisor, Dr. Mohammad Ashrafuzzaman Khan, Assistant Professor, Department of Electrical and Computer Engineering, North South University, Bangladesh, for his invaluable support, precise guidance, and advice about the experiments, research, and theoretical studies carried out during the current project and also in the preparation of the current report.

Furthermore, the authors thank the Department of Electrical and Computer Engineering, North South University, Bangladesh, for facilitating the research. We would also like to thank my friends for helping us with this project. The authors would also like to thank their loved ones for their countless sacrifices and continual support.

ABSTRACT

Age Suitability Test Based on Video Analysis

These days, video content has flooded social media. However, it may not be suitable for all viewers because it generally contains extreme violence, explicit language, sexual content, or other sensitive materials. In this paper, we introduce a deep neural network model, developed exclusively for the purpose of automatically categorizing raw video data. This way, our optimizer does this using the Adam Optimizer, and subsequently this makes it possible for our algorithm to then be able to classify the videos into the right classification and assign ratings in conformity to the existing movie rating system (G, PG, PG-13, R). Agiometry has a good classification accuracy, 77.65%.

TABLE OF CONTENTS

LETTER OF TRANSMITTAL 1

APPROVAL 3

DECLARATION 4

ACKNOWLEDGEMENTS 5

ABSTRACT 6

LIST OF FIGURES 9

LIST OF TABLES 10

Chapter 1 Introduction 11

1.1 Background and Motivation 11

1.2 Purpose and Goal of the Project 11

Chapter 2 Research Literature Review 14

2.1 Existing Research and Limitations 14

Chapter 3 Methodology 17

3.1 System Design 17

3.2 Hardware and/or Software Components 17

Chapter 4 Investigation/Experiment, Result, Implementation, Analysis and Discussion 25

Chapter 5 Impacts of the Project 32

5.1 Impact of this project on societal, health, safety, legal and cultural issues 32

5.2 Impact of this project on environment and sustainability 33

Chapter 6 Conclusions 34

6.1 Summary 34

6.2 Limitations 34

6.3 Future Improvement 35

References 36

LIST OF FIGURES

Figure 1: Flowchart Showing Model Training to Implementation Procedure 17

Figure 2: Block of LSTM at any timestamp {t} 20

Figure 3: LRCN architecture 22

Figure 4: Total Loss vs Total Validation loss for ConvLSTM model 26

Figure 5: Total Accuracy vs Total Validation Accuracy for ConvLSTM 26

Figure 6: Total Loss vs Total Validation loss for LRCN model 27

Figure 7: Total Accuracy vs Total Validation Accuracy for LRCN 27

Figure 8: Training Accuracy vs Validation accuracy for SimCLR 28

Figure 9: Training loss vs Validation Loss for SimCLR 28

Figure 10: Predicted classification results from video frame 29

Figure 11: Screeshot Website homepage 29

Figure 12: Screeshot of after video uploaded in the website in time of processing 30

Figure 13: Screeshot of showing the result of which category the video belong. 30

LIST OF TABLES

Table-1: Dataset Distribution 17
Table-2: Model Architecture 24
Table-3: Results of Precision, Recall, F1 Score of ConvLSTM model 26
Table-4: Results of Precision, Recall, F1 Score of LRCN model 27
Table-5: Results of Precision, Recall, F1 Score of SimCLR model 28

Chapter 1 Introduction

1.1 Background and Motivation

Rapid deployment of technology has put in place uncontrolled access for children to online video content, which could be either good or bad. In either way, one concerns the security of a child while surfing through a massive pool of online video outlets. This work aims to implement a smart
system that will detect contents considered safe for underage viewers with the help of an LRCN neural network, based on the video available. Using implementations of LRCN neural networks for content analysis and declaring whether the content is safe for underage viewers, this app will be able to keep children away from risky or unpalatable content, hence leading to a safer space on
the net.

The age-appropriateness testing of video would be one of the best tools for many applications, such as the analysis-based LRCN neural network in video, parental control, content filtering applications, age-restricted content delivery by many content providers to their customers, and so on. The implementation of this system ensures that people, especially children, are protected from
inappropriate content that can harm them.

Most of the online platform and other streaming contents contain a lot of information that might not be healthy for several age groups. Most age concern is displayed in a display of what goes on in the minds of children when they access such content. An age-appropriateness test can go a long
way to rid one's content of what is not suitable for the kids' age.

1.2 Purpose and Goal of the Project

Our project aimed at the development of a seniority test based on video analysis using LRCNs aims to achieve several critical objectives, including:

The main objectives of the project are listed below as project objectives that use LRCN techniques to create an age-appropriateness test based on video analysis:

Content safety: The most important goal here is that video content should be appealing or totally safe for audiences of various ages, especially children, from content that is not suitable or harmful.

Regulatory Compliance Help to regulate platforms and content providers with the different laws requesting age compliance, especially with the online and video game distribution of films and its media.

Better user experience: this would definitely make an enhancement in the user experience by creating content that goes with the age and maturity level of the viewers, ultimately satisfying them and making them engage more.

Parental Control: It offers a very potent tool for the parents to use to limit the content which they allow their children to browse on the Internet.

Although working with video platforms may mean great impacts to many different areas of the digital world in order to integrate this technology, it is a beneficial expansion of safety precautions.

The following are the objectives of the project "Age suitability test based on video analysis by LRCN method. " :

1. Child Safety Assurance: This is the core objective of the project in making sure that children are safe on internet video platforms. Safeguarding children involves protection from offensive, graphic, violent, or otherwise harmful content.

2. Tailor-made dataset development: Create rich and varied datasets that would describe exactly the type of data a child is most likely to encounter on the Internet.

3. Development of the LRCN model: We are developing a state-of-the-art ANN based on LRCN, the first capable one for real-time video input analysis. The developed model will support the inclusion of multiple modalities in predicting the content suitability for children.

4. Training and Validation: Train the LRCN model on the curated dataset to achieve high recognition accuracy of safe and unsafe video content. Metrics such as accuracy, precision, recall, and F1 show nicely performing models.

5. Real-time Analysis: Realize this model where a system can be developed and used for real-time video content analysis during an upload or streaming on every video platform to screen out the undesirable content for children.

6. Content Filtering and Alerts: The software must incorporate content filtering and alertness combined with blocking and warning controls that will display a warning notice to the user, parents, or guardians, and block any warning service when detecting any content that is potentially harmful. It should have possibly adjustable settings to meet the preferences of individual users.

7. Feedback from Users and Learning: In fact, have backward feedback through continuous improvement in the classification accuracy of the material. Proceed to fine-tune the model by obtaining feedback from users to identify new forms of unsafe content with a better approach.

8. Contribution to Research: Improved techniques on the basis of LRCN in the area of deep learning for video analysis and take the work of research that contributes to increasing the accuracy while detecting child safety content.

9. Compatibility and Scalability: Design the system in such a manner that it shall be adjustable with the increase in compatibility of global video platforms' variance and the devices of transmission to give complete value with penetration and coverage.

10. Continual improvement: Update and develop further the system to answer new needs of online installation and the needs of the users. Follow the development of newly emerging technologies and threats.

The project will have applied these objectives to make the Internet environment a little safer for those children and, thus, empower the parents and guardians with the means to secure their children against potentially harmful video content.

Chapter 2 Research Literature Review

2.1 Existing Research and Limitations

This research study [1] explores the fusion of text and images using deep learning and multi-modal approaches. That enhances the classification of film age appropriateness. It also aims to automate the tasks traditionally done by censorship bodies like the MPAA and BBFC. By creating a bimodal dataset combining film scripts with images from IMDB, the authors investigate the impact of integrating visual cues on age-appropriateness predictions, demonstrating the potential of deep learning in Film Studies and Digital Humanities.

The paper [2] uses a deep neural network with attention mechanisms to predict movie suitability for children and young adults based on scripts to determine MPAA ratings based on movie dialogues. It has achieved an 81% weighted F1 score. It excelled by seven percent over traditional methods and capably solved the problem of choosing the suitable content for young users.

In this study [3], they employed an innovative deep-learning methodology to detect violence in movies to automatically identify and remove violent scenes, aiming to create "violence-free" versions. They claimed that shot segmentation and selective extraction of mainframes could be employed along with entering a deep-learning violence classification model as a means to protect those who are incapable against violent occurrences, particularly kids.

This paper [4] introduces a deep-learning framework for accurately detecting and filtering adult content in video sequences to protect children and vulnerable individuals from inappropriate material. It uses both spatially self-learned features as well as temporal cues for ensuring accurate classification of porn materials by employing Long-Term Short Memory (LSTM) technology and its variations.

The paper [5] focuses on the affective characterization of movie scenes using content analysis and physiological changes. While watching a movie video clip, participants' physical feelings and self-appraised emotional reactions were recorded by the people who were doing the research. This study investigates how digital content properties, physiological signals, and self-appraisals can all be combined to interpret an emotional scene within a film. It underscores the

possibility of using extraneous peripheral signals for emotion measurement and video indexing based on feelings.

The paper [6] presents a content-based movie analysis and indexing approach that integrates audio and visual cues to extract high-level semantic information from movies. The goal is to detect significant movie occurrences and provide them with semantic annotations for their content indexation. Blending sound and picture indications results in more profound comprehension, generalization process, and indexation of video content.

This system achieved an average accuracy of 84% with low false acceptances and rejections. The adaptive silence detector achieved an average accuracy of 63.25%, which outdid the global silence model because of its low false acceptance rate.

This study [7] presented an automated system, DOVE, for detecting violence in movies, which utilizes color histograms, skin and blood component identification, and motion intensity analysis to determine violent scenes. In simplified terms, DOVE works wonderfully by employing four intricately interconnected modules to sieve through a broad spectrum of scenes. For instance, during the experimentation conducted on such movies as Gladiator and Passion of the Christ, it was observed that the rate of accurate positive responses was very high, showing that this software program can recognize even the minutest details of violence within moving images whose input is provided either through video or through files stored on hard disks repositories which should also be considered to improve its performances anyway because redundancy sometimes leads us astray though sometimes it helps overcome malfunctions like erring sensors for example. Nevertheless, there were problems in identifying scenes in films like Kill Bill and What A Girl Wants, which necessitated system enhancements to handle the same. In order to enhance the precision and efficiency, the thresholds have to be adjusted; also, inter-blob distances should be taken into account. Overall, DOVE's success in recognizing violent scenes demonstrates its potential for advancing content analysis in the film industry.

The study [8] discusses video classification models that recognize violent actions. The parameters for analysis, like precision, recall, precision-recall curve, and F1 score, are discussed. The 3DCNN model shows lower precision as recall increases, indicating inclusivity but more false positives.

In contrast, the ConvLSTM model maintains high precision with rising recall, making it suitable for the task. It provides insights into the performance of different models for detecting violent behavior in videos, highlighting the importance of precision and recall in model evaluation.

This paper [9] used the simCLR approach that falls in the family of frameworks for contrastive learning of visual representations. It significantly improves the baseline in performance on self-supervised, semi-supervised, and transfer learning tasks. Concerning guidelines on the constituent elements of the framework experiment with different design choices, noting the gain in performance over the existing methods, it shows a slight difference in the components. Most components of the SimCLR framework, such as Sobel filtering, extra color distortion, and combined augmentations with motion blur, are more relevant to performance in predictive tasks. Larger data augmentation compositions increase accuracy rates for ResNet-50 models trained at multiple augmentation levels. Also, selecting a non-linear head at the end of the network with some loss functions and optimal temperature settings for different batch sizes will make our framework work well in practice for learning visual representations. Results show the SimCLR framework outperforms the previously mentioned methods and, in some cases, is on par with supervised pre-trained models in different settings. It achieved an accuracy of 85.8%. In general, the simplicity and efficiency of the SimCLR framework bring forward the importance of carefully designed choices and data augmentation strategies for learning better visual representations with contrastive self-supervision methods.

Chapter 3 Methodology

3.1 System Design

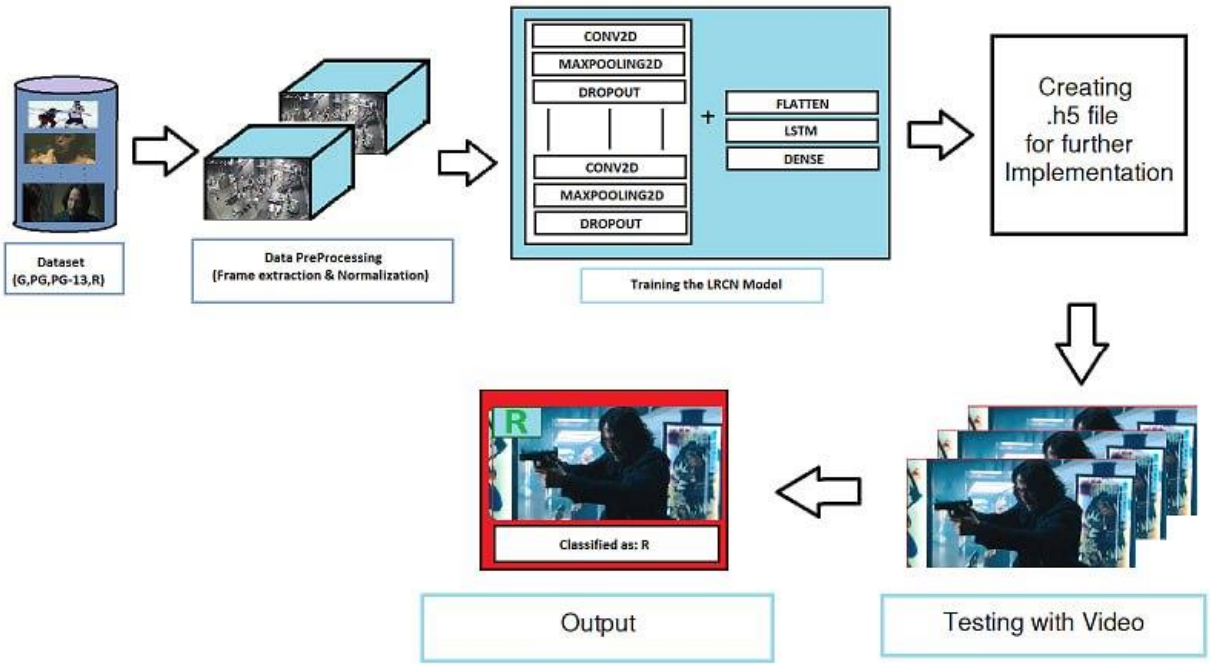

Figure 1: Flowchart Showing Model Training to Implementation Procedure

3.2 Hardware and/or Software Components

3.2.1 Dataset

We have gone through several movies and internet top content hubs to collect the dataset. We have customized our data according to the class types, including G, PG, PG-13, and R. Here is each class's detailed distribution of videos. We have used LSTM and LRCN for this project.

Classes	G	PG	PG-13	R
No of videos	81	66	90	62

Table 1: Dataset Distribution

3.2.1.1 Dataset Preprocessing

We have given customized image height and width to 64 and 64 and set the sequence length to 20. Now, we have introduced a new function that extracts the required frames from a video after resizing and normalizing them. The provided function initializes a list to store the processed frames and reads the video using OpenCV's `VideoCapture` object. To sample frames evenly across the video length, the total number of frames is determined, and an interval is calculated to ensure that a fixed number of frames (defined by `SEQUENCE_LENGTH`) are extracted. The function sets the video position for each frame, reads it, and resizes it to a predefined height and width (`IMAGE_HEIGHT` and `IMAGE_WIDTH`). It also normalizes the pixel values to a range of 0 to 1 by dividing by 255. The processed frames are later added to the list.

3.2.2 Model Description

LSTM: Long Short-Term Memory architecture: designed to handle sequential inputs and make an attempt to bypass some of the vulnerabilities in regular RNNs, the primary one being the vanishing gradient problem. Many sequence-related activities, such as time series, natural language processing, and speech recognition, find LSTMs particularly well-suited.

Each of the video frame data is stated 64X64 size and the sequence length is stated 20. For the ConvLSTM structure we have created a Convolutional Function. But for the baseline LSTM structure is defined.

Here are the key components and characteristics of the LSTM architecture:

1. Cell State and Hidden State

2. Gates: For information control, LSTMs implement three types of gates, which include input gates, forget gates, and output gates.

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i)$$
$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f)$$
$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o)$$

where,

$$
\begin{aligned}
i_t &\rightarrow represents\ input\ gate.\\
f_t &\rightarrow represents\ forget\ gate.\\
o_t &\rightarrow represents\ output\ gate.\\
\sigma &\rightarrow represents\ sigmoid\ function.\\
w_x &\rightarrow weight\ for\ the\ respective\ gate(x)\ neurons.\\
h_{t-1} &\rightarrow output\ of\ the\ previous\ lstm\ block(at\ timestamp\ t-1).\\
x_t &\rightarrow input\ at\ current\ timestamp.\\
b_x &\rightarrow biases\ for\ the\ respective\ gates(x).
\end{aligned}
$$

3.Long-Term Dependencies: LSTMs were designed in such a way that they could explicitly learn long-term dependencies. Therefore, through its gating mechanisms, LSTMs can decide when to reset a read, write, or cell state and hence still be able to learn information that is relevant to the task even after many time steps.

4.Training and Backpropagation: LSTMs are trained using backpropagation through time (BPTT), similar to traditional RNNs. All these depend on gradient-based optimization methods like stochastic gradient descent (SGD) to update the weights in the bid to minimize the loss function.
The equations for the cell state, candidate cell state and the final output:

$$
\begin{aligned}
\tilde{c_t} &= tanh(w_c[h_{t-1}, x_t] + b_c)\\
c_t &= f_t * c_{t-1} + i_t * \tilde{c_t}\\
h_t &= o_t * tanh(c^t)
\end{aligned}
$$

where,

$$
\begin{aligned}
c_t &\rightarrow cell\ state(memory)\ at\ timestamp(t).\\
\tilde{c_t} &\rightarrow represents\ candidate\ for\ cell\ state
\end{aligned}
$$

LSTMs are appropriate for a wide class of problems in which dependencies in sequences have a great meaning. They perform very well in modeling long-range dependencies and dealing with the vanishing gradient issue, making them the most preferable for a wide range of data-processing tasks in sequences; this goes from text and speech recognition to time series prediction.

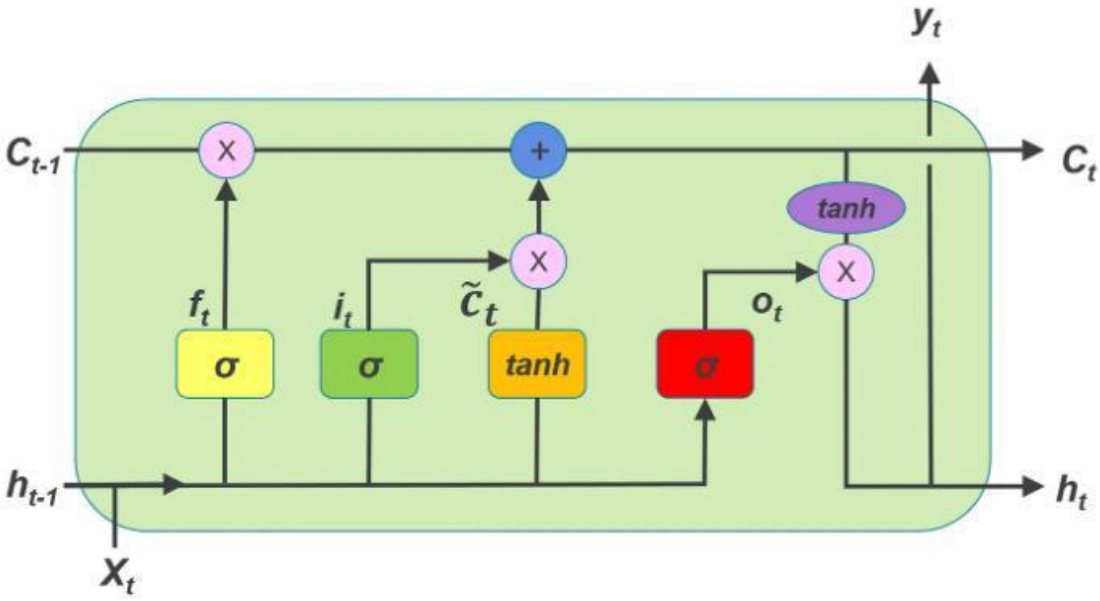

Figure 2: Block of LSTM at any timestamp {t}

LRCN:

An LRCN is a Long-Short Term Memory Recurrent Convolutional Network, which is a deep learning design that couples both the power of a Convolutional Neural Network and a Recurrent Neural Network for the processing of sequential data, further designed for processing where the data are either spatially or temporally organized. This architecture has been applied to a couple of domains, such as computer vision, natural language processing, and video analysis.

The major motivation behind LRCN is in handling the limitations posed by a conventional CNN and RNN architecture in sequences of data. Even though CNNs can capture spatial features really well from images or videos, they do not naturally handle temporal dependencies. On the other hand, RNNs, as good models for sequential data, often do not capture the spatial structure in an image or video.

This is enabled by a combination of both the CNN and the RNN, which attained spatial and temporal information in a joint architecture. Developing LRCN captures both spatial and temporal information in a single unified architecture. This is how the elements of the LRCN composition come into play:

1. **Convolutional Neural Network (CNN):** At the input stage, LRCN usually starts with a Convolutional Neural Network (CNN). The CNN extracts spatial features present in each frame of the input data, which can be in the form of pictures or video frames. These features are further propagated through the next layers, and the CNN part can contain architectures like VGG, ResNet, or Inception.
2. **Temporal Encoding:** After processing the feature of each frame by the CNN, it is reshaped into a 3D tensor containing the temporal dimension. That is, it represents a sequence of the feature maps over time.
3. **Recurrent Neural Network (RNN):** The output is passed into an RNN, more precisely a recurrent layer like LSTM or GRU. These RNNs will be useful in modeling temporal dependencies since they take into consideration the information in neighboring frames. This might be used for video analysis or action recognition with captioning.
4. **Task-Specific Output Layer:** The output from the RNN is further processed with fully connected layers toward the prediction of the specific task at hand. In the video captioning task, the model would generate an output of a textual description of the video, while in the action recognition task, it would output the classification of actions being undertaken.

The LRCN architecture has proven very good in applications where understanding data in its spatiotemporal nature has to be combined. Some of the common applications are subtitling of videos, summarization from videos, action recognition, and gesture recognition among others. It seems that LRCN suits many problems at the same time because of the synergy of CNNs for spatial feature extraction and RNNs for sequential modeling.

But it is pertinent to mention that LRCN is one of the architectures among many to deal with sequential data; task performance can be considered task-specific and dataset-dependent.

Research towards the variations and improvements in the design parameters is an area of constant focus and is usually made toward making these models efficient and effective in a wide array of tasks.

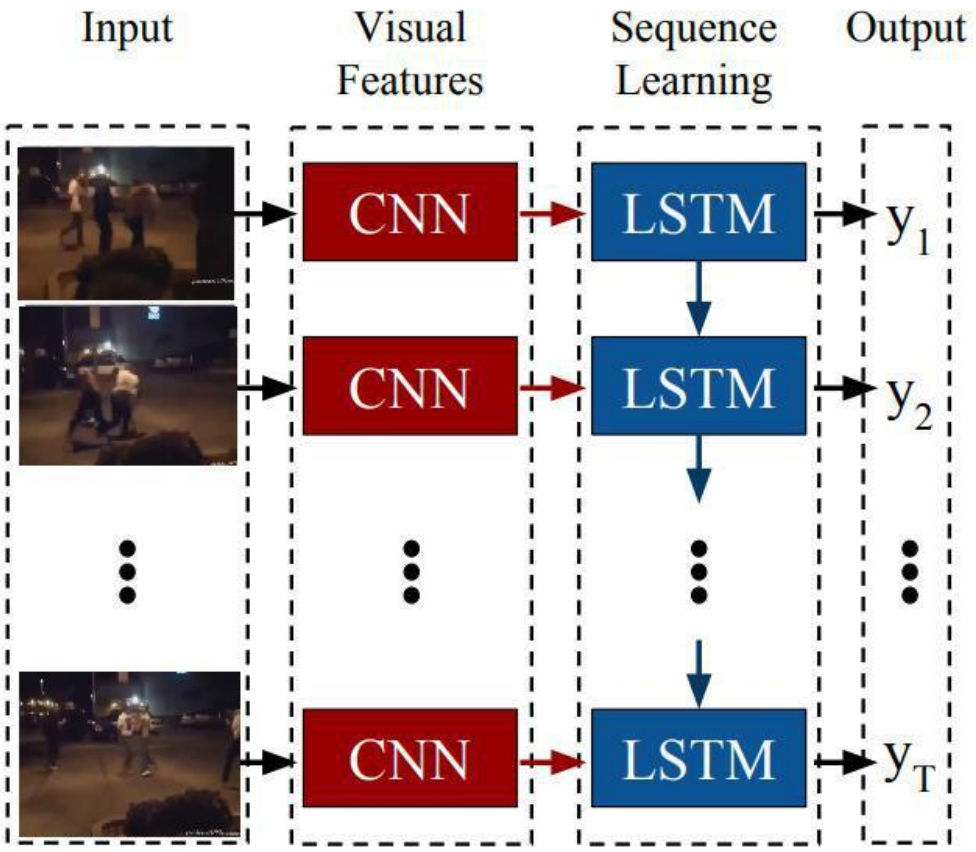

Figure 3: LRCN architecture

SimCLR:

Google Research researchers presented a self-supervised learning framework called SimCLR (A Simple Framework for Contrastive Learning of Visual Representations). By optimizing agreement amongst variously enhanced perspectives of the same data example, it seeks to learn visual representations from unlabeled data source. The main elements of SimCLR consist of:

Data Enrichment: Every data example receives a different set of augmentations to produce a variety of displays. These enhancements may consist of Gaussian blur, color distortion, and random cropping.

Base Encoder: A neural network, typically a variant of ResNet, is used to encode each augmented view into a feature vector. The encoder learns to produce similar representations for different views of the same image and dissimilar representations for views of different images.

Projection Head: On top of the encoded feature vectors, a small neural network is added to further augment it, leading to better quality of learnt representations.

The contrastive nature of loss functions pulls augmented versions of the same image close to each other (positive pairs) and pushes the representation of different images away from each other (negative pairs).

The learned representations could even be further fine-tuned with a labeled dataset to generate a supervised version for SimCLR, thus ensuring better performance on tasks at a distant horizon, including picture categorization.

3.3 Model Architecture

We used a maxpooling2D layer, a dropout layer, and a conv2D layer to generate the LRCN model. Following several iterations of this sequence orientation, we have added flatten, LSTM, and finally a thick layer. ReLu is utilized as the activation function in Conv2d, and softmax activation is employed for the dense layer.

Layer Type	**Output Shape**	**Details**
Input Layer	(20, 64, 64, 3)	Input shape: (SEQUENCE_LENGTH, IMAGE_HEIGHT, IMAGE_WIDTH, 3)
TimeDistributed Conv2D	(20, 64, 64, 16)	16 filters, kernel size (3, 3), padding='same', activation='relu'
TimeDistributed MaxPooling2D	(20, 16, 16, 16)	Pool size (4, 4)
TimeDistributed Dropout	(20, 16, 16, 16)	Dropout rate: 0.25
TimeDistributed Conv2D	(20, 16, 16, 32)	32 filters, kernel size (3, 3), padding='same', activation='relu'
TimeDistributed MaxPooling2D	(20, 4, 4, 32)	Pool size (4, 4)
TimeDistributed Dropout	(20, 4, 4, 32)	Dropout rate: 0.25
TimeDistributed Conv2D	(20, 4, 4, 64)	64 filters, kernel size (3, 3), padding='same', activation='relu'
TimeDistributed MaxPooling2D	(20, 2, 2, 64)	Pool size (2, 2)
TimeDistributed Dropout	(20, 2, 2, 64)	Dropout rate: 0.25
TimeDistributed Conv2D	(20, 2, 2, 64)	64 filters, kernel size (3, 3), padding='same', activation='relu'
TimeDistributed MaxPooling2D	(20, 1, 1, 64)	Pool size (2, 2)
TimeDistributed Flatten	(20, 64)	
LSTM	32	32 units
Dense	4	4 units

Table 2: Model Architecture

Chapter 4 Investigation/Experiment, Result, Implementation, Analysis and Discussion

Investigation/Experiment:

This work contributes to the field of video age certification through the adoption of state-of-the-art deep learning models, the ConvLSTM and LRCN models, to tackle accurately determining age as the primary requirement for using deep learning models.

This research was instigated by the construction of a dedicated dataset for the certification of Video Age. Since there was the difficulty of obtaining a proper good existing dataset, an old continuous process of data collection from various sources was started. To this respect, the gathered data was arranged diligently according to the MPA standard guidelines into four categories: 'G,' 'PG,' 'PG-13,' and 'R.' This is really a systematic collection process of data and classifying them into these categories, accomplished through regular collection and sorting efforts made from movies and other relevant sources.

A lot of preprocessing works out after obtaining the dataset. All video files are read frame by frame and resized uniformly to a standard resolution of 64 × 64 in order to keep the computation minimum. In addition, normalization was also adopted for making the data so that pixel values stick within the bounding of [0, 1] to follow faster convergence during the network training.

For model evaluation, the dataset has been cautiously split into the training dataset, which comprises 75%, and the testing dataset, which comprises the remaining 25%, so as to give an assurance of the dataset shuffling being carried out equitably, avoiding any kind of inherent bias.

Classification of age appropriateness in a video has been achieved using two robust models, ConvLSTM and LRCN. Since the ConvLSTM is useful due to the convolutional and LSTM layers, this will be oriented in order to critically scrutinize the video frames spatially and temporally to make a decision correctly relating to age classification. In the other direction, the LRCN model combines the strengths of the Conv2D and LSTM layers toward focusing on the recognition of complex patterns and interdependencies within video sequences.

Result:

Implementing ConvLSTM and LRCN models for video age certification yielded significant insights. Notably, the ConvLSTM model achieved an approximate test accuracy of 61.22%. In contrast, the LRCN model outperformed, exhibiting a higher test accuracy of around 77.65%. And in SimCLR we got 68% accuracy.

Here is the precision, recall, and F1 score for each class for the ConvLSTM model:

Class	Precision	Recall	F1-Score
G	0.37	0.50	0.43
PG	0.29	0.31	0.30
PG-13	0.67	0.25	0.40
R	0.54	1.00	0.70

Table 3: Results of Precision, Recall, F1 Score of ConvLSTM model

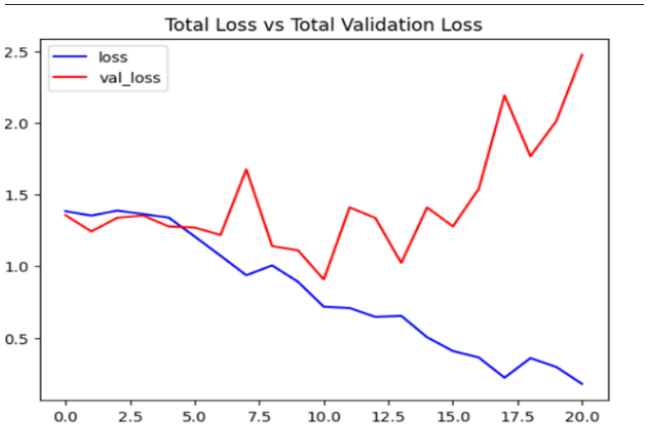

Figure 4: Total Loss vs Total Validation loss for ConvLSTM model

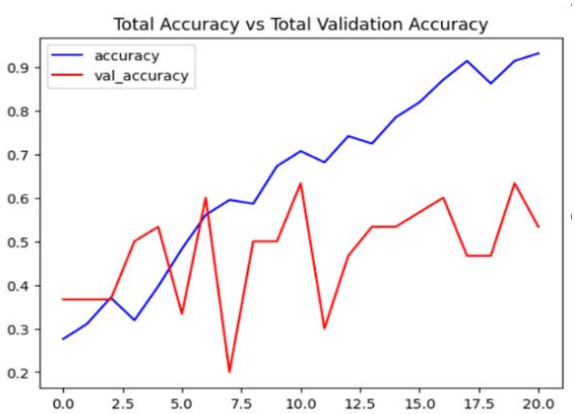

Figure 5: Total Accuracy vs Total Validation Accuracy for ConvLSTM Model

Here is the precision, recall, and F1 score for each class for the LRCN model:

Class	Precision	Recall	F1-score
G	0.71	0.60	0.65
PG	0.77	0.77	0.77
PG-13	0.88	0.85	0.87
R	0.68	0.87	0.76

Table 4: Results of Precision, Recall, F1 Score of LRCN model

From LRCN, we get a Training accuracy of 92.13% and a validation accuracy of 66.67%. After evaluating the test set, we get a test accuracy of 77.65%. Here are the training accuracy and validation accuracy graph.

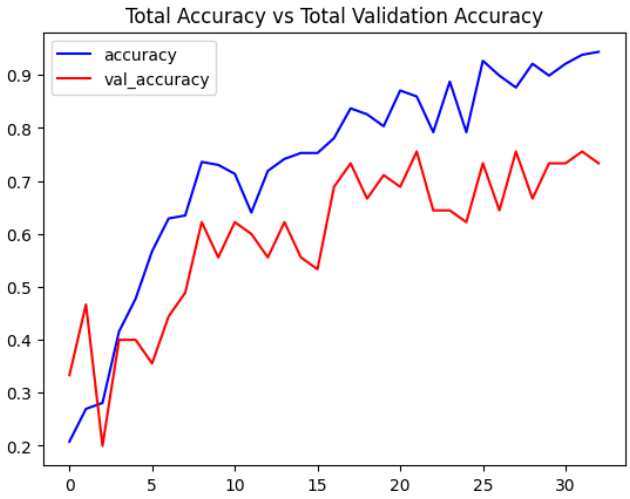

Figure 6: LRCN Training Accuracy vs Validation accuracy

The Training loss for LRCN is 18.17%, and the validation loss is 80.51%. For the test set evaluation, it is 83.58%.

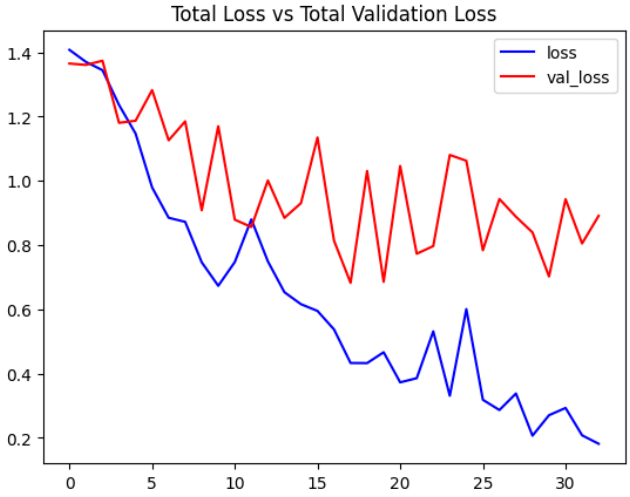

Figure 7: LRCN Training loss vs Validation Loss

Here is the precision, recall, and F1 score for each class for the SimCLR model:

Class	Precision	Recall	F1-Score
G	0.75	0.45	0.56
PG	0.90	0.69	0.78
PG-13	0.75	0.67	0.71
R	0.52	1,00	0.68

Table 5: Results of Precision, Recall, F1 Score of SimCLR model

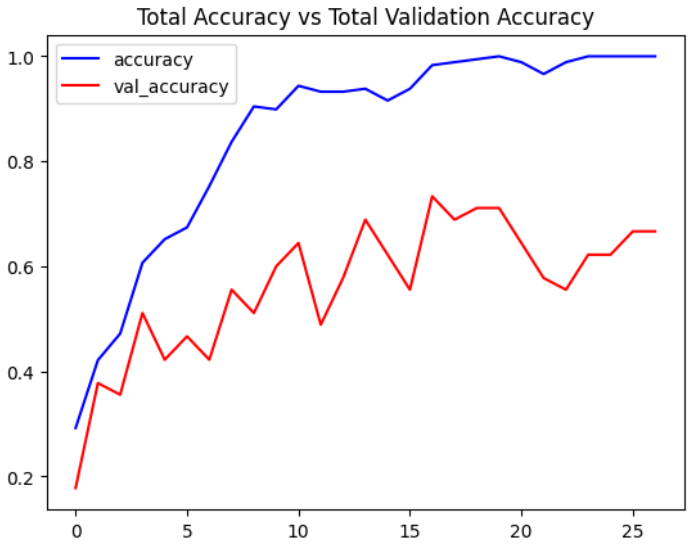

Figure 8: Training Accuracy vs Validation accuracy for SimCLR

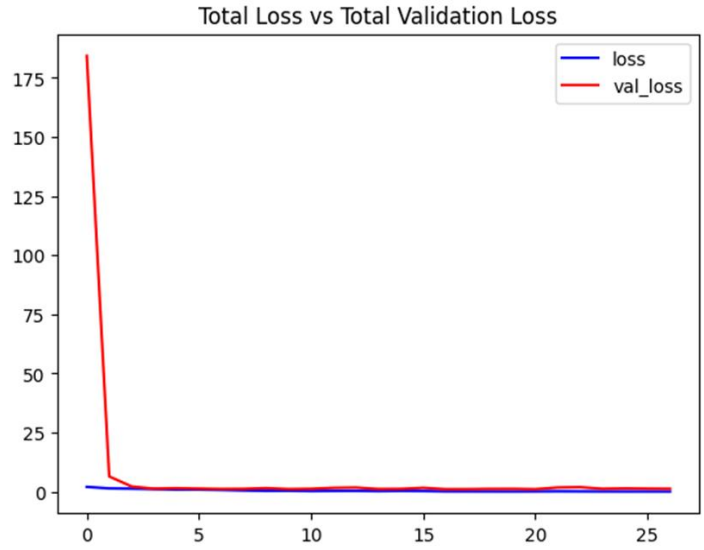

Figure 9: Training loss vs Validation Loss for SimCLR

For the output after the video analysis is done, our model prediction result on the top corner of the video is shown. Here are the detailed results.

Figure 10: Predicted classification results from video frame

Implementation:

For the implementation part, we have developed a website that uses the LRCN.h5 file for the classification of the uploaded video. We have used Flusk, HTML, CSS, javascript, python, and other libraries for this implementation task. However, there were some things that could be improved in this task. The processing time of the video was very long in the website according to the duration of the video.

Here is some screenshot of the website.

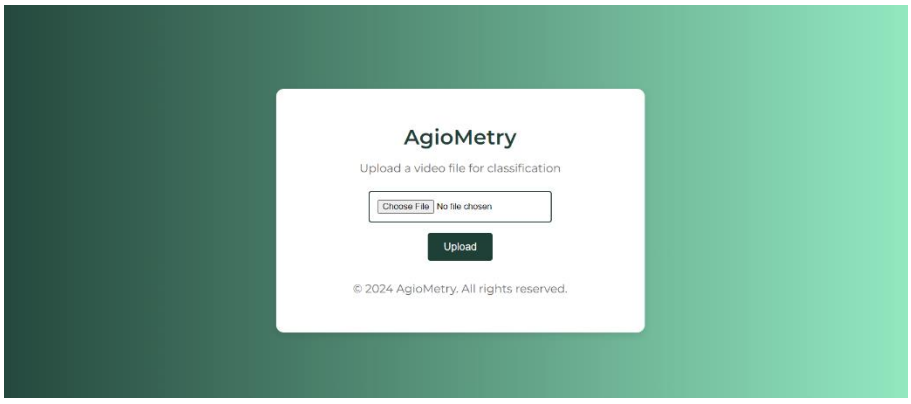

Figure 11: Screeshot Website homepage

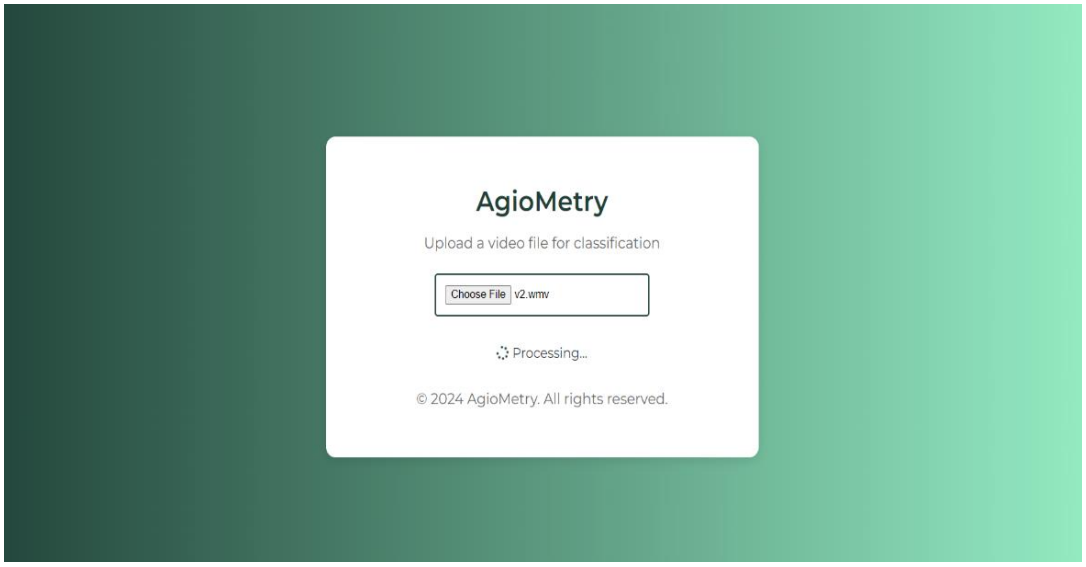

Figure 12: Screeshot of after video uploaded in the website in time of processing

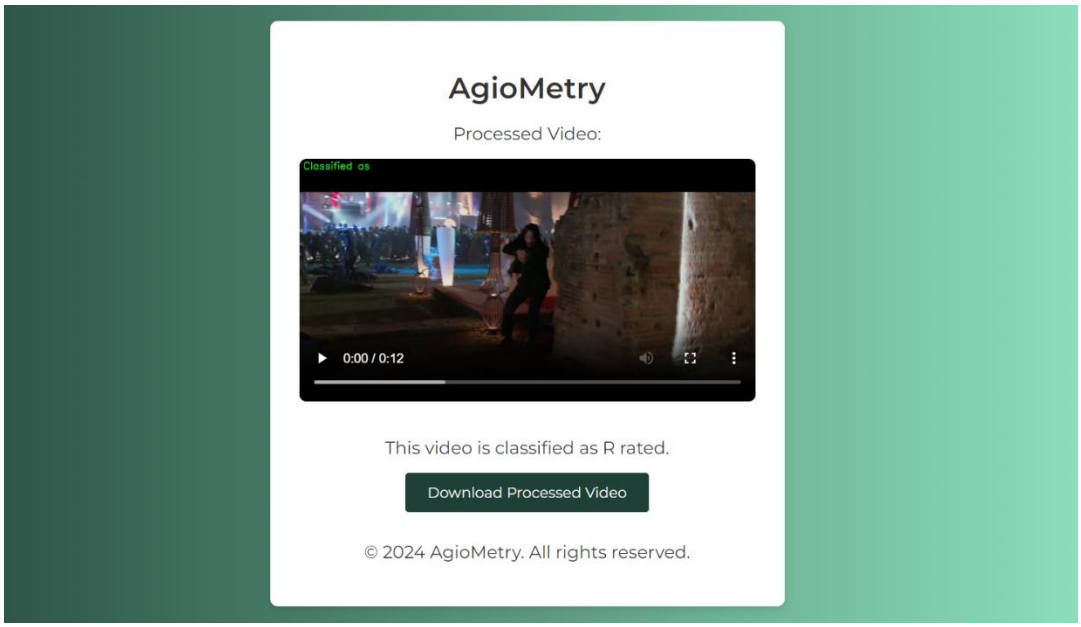

Figure 13: Screeshot of showing the result of which category the video belong. In this case the video is R rated.

Analysis:

The observations that follow shall make differences in performance metrics between ConvLSTM,simCLR and LRCN models apparent, differences which warrant an in-depth analysis of the underlying factors influencing their usefulness in video age certification.

The disparity in the test performance of both ConvLSTM and LRCN models, although being relatively moderate, is very important. The fact that this model achieved great accuracy of 77.65% when compared to the ConvLSTM model, which was 61.22%,for simCLR it was 68% which vocalizes the open debate about architectures and possible caveats with one model versus the other.

They all manages sequences lasting 20 frames, which are uniformly resized to 64x64 pixels, and share the same formatting of the input data. However, differences in their architecture are inherently different. The whole ConvLSTM model was built around those ConvLSTM2D layers, which actually have the capacity to capture spatial and temporal features in a video. On the other side, the LRCN model was working within a hybrid structure using Conv2D to facilitate feature extraction in the spatial domain, together with LSTM to recognize the dependent relations in the temporal sequence. Most probably, some architectural design discrepancies have been the key reasons for the different performances.

A difference in performances could be attributed to the size of the training dataset. Maybe the small size of the training dataset limited the training; thus the generalization ability of the ConvLSTM model was restricted to enable it infer more difficult patterns.

Discussion:

The LRCN model got just a wee bit more test accuracy than the ConvLSTM model and simCLR model, proving very slight differentiations in its performance. That is to say, the same set of data went into these three models.

The small difference in accuracy between the models indicates the LRCN model could have a slight edge in understanding complex video patterns compared to ConvLSTM and simCLR. The accuracy of the model is disappointing. All models need more samples in training.

Going forward, possible strategies will include enlarging the training dataset for the ConvLSTM model in such a manner that it acquires much better performance, and enhancing diagrams in order to capture intricate dependencies within video sequences. A comprehensive aggregation of data from movies, among other diverse sources, is a constant activity in the enhancement of the dataset for training. This assures better performance within the forthcoming iterations of the model.

Chapter 5 Impacts of the Project

5.1 Implications of the project in terms of social, health, safety, legal and cultural factors

The implications of this video age certification project are great and far-reaching: there are implications in the social, health, safety, legal, and cultural domains.

Societal Impact: The pervasiveness of digital media in society is shaping contemporary norms and behaviors. Precise video age certification will support controlled exposure of the people to content that is of a suitable age, especially the minors. The project will add to the making of a more responsible and knowledgeable digital society due to the operation of effective certification mechanisms. It gives room for parental guidance and a safer online atmosphere since the choices will be informed by what kind of content is appropriate for the age groups.

Health and Safety: The project has taken some precautions in preventing mentally conditioned situations that may be induced by early exposure to inappropriate or harmful content, especially for youth. Acts of violence, inappropriate language, or mature themes are some of the key things that mentally conditioned situations can be brought on by. Preventing circumstances of exposure to such content can really help with mental health by providing the online space with safe, healthy, and secure environments.

Legal Implications: It is important for content that is being consumed online, especially by minors, to not cause a breach of legal standards and regulations. Thus, a well-working video age certification system, according to the legal requirements, can help protect the system from possible legal effects of distribution of inappropriate content to minor audiences. This project could serve to meet the legal requirements and standards for a systematic certification and regulation of video content to related approaches that support a more legally compliant digital space.

Cultural Impact: Videos are strong cultural bearers and influence the perception that society has. Effective measures of verification show respect and protection of cultural values, sharing only well-graded content. This goes a long way in preserving cultural sensitivity and norms, thereby allowing the digital environment to be sensitive and very inclusive to respecting all cultural diversities. It also helps in the distribution of culturally fitting content, thereby adding to needs of the preservation and celebration of many cultural views within the digital platform.

It transcends the impact of the project at the very least toward a more educated and responsible digital culture, mental health, legal compliance, and diverse cultural values. It shows a step toward a more conscious and inclusive space for many users, making it a safer, more sensitive environment culturally, and on the right side of the law.

5.2 Impact of this project on environment and sustainability

According to him, content ratification towards a certain age group is done through the project and will go a long way to help the users be more responsible in content consumption, saving bandwidth and reducing unnecessary data flows. By promoting the same, he argues, it contributes indirectly to streamlining the digital infrastructure, possibly resulting in energy savings, both in data centers and transmission networks. The idea is that the user, with responsible access to content, aims at a more sustainable digital ecosystem.

Deep learning projects need much processing power, and the level of energy consumption is high; hence, indirect environmental damage. The project architecture and infrastructure should be optimized in such a way that it helps to decrease the environmental footprint. Not a huge initiative to require detailed hardware configurations but sustainable hardware practices need to be put in place to help reduce the electronic waste. This will be done with the use of energy-efficient components of hardware and data storage techniques that are friendly to the environment. Even with a small project, the introduction of more efficient elements and the introduction of sustainable procedures can considerably lower the carbon footprint of a project.

Responsible processing and content use will be the object of attention, and optimization for hardware will be considered because good sustainable hardware practice is key for the mitigation of the environmental impact that this project can have.

Chapter 6 Conclusions

6.1 Summary

Thus, this project developed a very strong system for video age certification using the models of ConvLSTM and LRCN. The primary objective was the assurance for users to watch their contents in a well-devised manner of content classified under different age groups. Thus, the creation would not only suffice for age-appropriate access but also be directed towards the efficient utilization of bandwidth, thereby helping digital infrastructure to be streamlined and further possibly saving energy in data centers and transmission networks. Age-appropriate allocation was to be performed, and thus, proper analysis of video sequences through the implementation of the models was to be done very carefully.

Developing a model on this project calls for a comprehensive dataset and a very extensive preprocessing procedure. In this regard, models were facing the problem of generalization and achieving the most desirable performance due to data constraints. This also impresses the need for sustainability in the project because deep learning models are very resource-intensive in the use of energy and environmental considerations.

6.2 Limitations

A lot of limitation points were identified in the whole project:

Constraints in Data Collection: The unavailability of a suitable existing dataset compelled the development of a data-set tailored for this project. Therefore the data was collected and categorised under the limitation of time and complexity.

Resource Intensiveness: The processing time for deep learning models that require an enormous amount of resources to train leads to concerns about energy consumption and increased hardware requirements.

Model Sensitivity: Both the ConvLSTM and LRCN models were sensitive to variations in the data, which reduced the generalization ability of such models to new data and, in fact, possibly limited the application of these models to any practical scenario.

Computational Complexity: The models of the ConvLSTM and LRCN classes are complex, and their computational requirements are high, which would make training more laborious and, in general, adversely affects scalability.

Subjectivity in Age Classification: A high level of subjectivity in the tagging of the age category revealed an intrinsic problem: age appropriateness may be considered by cultural, social, and individual factors that take the model's estimations far from the exact value.

6.3 Future Improvement

For future improvements:

Dataset Expansion: We plan to expand the dataset in the future by collecting a much broader spectrum of diverse video content. This should be in favor of generalization ability and accuracy in age classification.

Applying methods: As our model’s accuracy and other performance metrics is not good enough , we will try out different methods such as semi-supervised learning. We would like to try out different approaches like Contrastive video representation learning (CVRL) and semi supervised approach for simCLR method to improve the performance of the model.

More effective implementation: To apply on real life data, we plan to design a browser extension that will access the social media’s video content and warn the users before hand about the content type. Thus the main purpose of creating a heathier tech environment for all type users can be achieved.

References

1. Mohamed, E., 2021. Combining Text and Images for Film Age Appropriateness Classification. *Procedia Computer Science*, *189*, pp.242-249.
2. Shafaei, M., Smailis, C., Kakadiaris, I.A. and Solorio, T., 2021. A Case Study of Deep Learning Based Multi-Modal Methods for Predicting the Age-Suitability Rating of Movie Trailers. *arXiv preprint arXiv:2101.11704*..
3. Khan, S.U., Haq, I.U., Rho, S., Baik, S.W. and Lee, M.Y., 2019. Cover the violence: A novel Deep-Learning-Based approach towards violence-detection in movies. *Applied Sciences*, *9*(22), p.4963.
4. Kusrini, K.U.S.R.I.N.I., Setyanto, A.R.I.E.F., Agastya, I.M.A., Hartatik, H.A.R.T.A.T.I.K., Chandramouli, K.R.I.S.H.N.A. and Izquierdo, E.B.R.O.U.L., 2022. a Deep-learning framework for accurate and robust detection of adult content. *Journal of Engineering Science and Technology*.
5. [Soleymani, M., Chanel, G., Kierkels, J.J. and Pun, T., 2009. Affective characterization of movie scenes based on content analysis and physiological changes. International Journal of Semantic Computing, 3(02), pp.235-254.
6. Li, Y., Narayanan, S. and Kuo, C.C.J., 2004. Content-based movie analysis and indexing based on audiovisual cues. IEEE transactions on circuits and systems for video technology, 14(8), pp.1073-1085.
7. Clarin, C., Dionisio, J., Echavez, M. and Naval, P., 2005. DOVE: Detection of movie violence using motion intensity analysis on skin and blood. PCSC, 6, pp.150-156.
8. Patel, D., Shah, J., Pandey, P., Katre, N., & Tawde, P., 2023. Assessing the Performance of Video Classification Models in Identifying Violent Actions. International Journal for Research in Applied Science and Engineering Technology. https://doi.org/10.22214/ijraset.2023.55877.
9. Chen, T., Kornblith, S., Norouzi, M. and Hinton, G., 2020, November. A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PMLR.