National College of Ireland

Dance Video Classification into Relevant Street Dancing Styles using Deep Learning Techniques

MSc Research Project
Data Analytics

Dhanshree Bauskar
Student ID: x19230460

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Dhanshree Bauskar
Student ID:	x19230460
Programme:	MSc. in Data Analytics
Year:	2022
Module:	Research Project
Supervisor:	Dr. Catherine Mulwa
Submission Due Date:	16/12/2021
Project Title:	Dance Video Classification into Relevant Street Dancing Styles using deep learning Techniques
Word Count:	
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	16th December 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Dance Video Classification into Relevant Street Dancing Styles using Deep Learning Techniques

Dhanshree Bauskar
x19230460

Abstract

Computer vision is the emerging research area in technical field, and video classification is considered as one of the categories of computer vision. In past few years, deep learning approaches have elongated in video domain. Combination of deep learning approaches and computer vision have solved complex tasks. Consequently, deep learning approach is used to solve computer vision problem like video classification. As a increased people's interest in the field of dance, this research intends to classify 10 different dance styles using the database provided by AIST dance academy. Dance as a whole consists of Body movements, musical pieces, facial expressions and hand gestures which makes it complex problem to classify dance videos. Using VGG-16 model for pre-processing and training of data gave the accuracy result of 75.86% followed by the accuracy obtained by VGG-19 model that is 68.96% comparatively Convolutional neural network has the lowest accuracy. The model is evaluated with different evaluation criteria.

1 Introduction

Dance is a popular art among people of all age groups. Dance is a combination of different poses, actions, and body movements which forms different dance styles. Dance can be performed solo, in a group or as a couple. Many literatures have been performed on solo dance form recognition. There are 28 basic dance types around the world, and every dance type has different style to performance. Here, we focus on one of the dancing styles which is street dancing, it involves large number of social dance styles like hip-hop, popping, breakdance, locking, jazz etc. It can be performed in a group, as a competition or solo. The deep learning techniques like neural networks is been admired to solve many problems related to image processing, Speech Recognition and video processing.

Project Background and Motivation : Classification of videos can be done with different basis as video consists of audio-visual effects, they can be classified by audio or recognized by the actions performed in the video. The dataset is taken from AIST dance video database. It is large-scale database publicly available for use of academic purpose. The dataset contains a large number of dance videos of 10-20 seconds each. The dataset demonstrates 10 dance genres like ballet jazz, street jazz, krump, house, wack, lock, pop, break, LA- style hip-hop and middle hip-hop.

The motivation behind this research is the rising interest of people of different age groups towards the dance. As every research has some base study, similarly for this

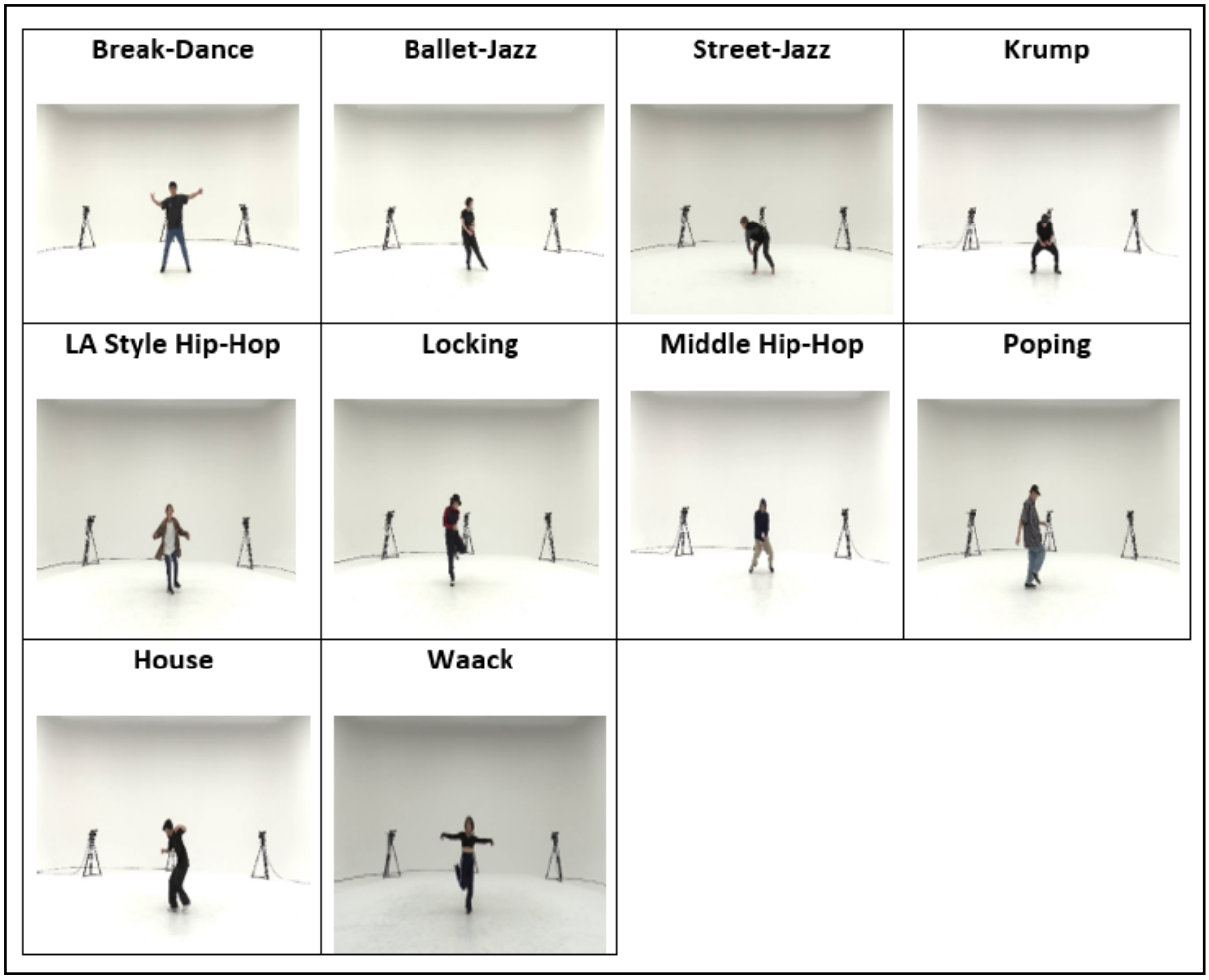

Figure 1: Street Dancing Style Categories

research, Human Action Recognition can be considered as the base study. As, dance is the combination of human actions whether it be hand gestures, facial expression or leg movements. The evaluation of applied model will be based on the accuracy,precision and recall. In figure 1, ten distinct classes which are to be classified in this research can be seen.

1.1 Research Question

How can deep learning techniques help to improve identification and classification of different Street Dancing Style?
Neural network models like VGG-16, VGG-19 and CNN are used to analyse whether the model can recognise various dance forms and is able to categorise under specific dance styles. Dance styles of Street dancing around the world will be classified using Neural Networks. Deep learning is a subset of Machine Learning where by using multiple neural layers higher level of features can be extracted from the raw input. The pre-trained models of neural networks have an ability to extract important features from a video. Each model has its own unique technique to process data and generate results.

Every research has some objectives to achieve and complete. The Research Objectives

Table 1: An overview of Project Objectives

ID	Objectives	Description
1	Literature Review	Critically examine peer-reviewed literature and studies performed in the area of video classification
2	Methodology	Methodology approach for dance video classification
2.1	Exploratory Data Analysis	Exploring and understanding the data
2.2	Data pre-processing and Feature extraction	Prepare data for analysis, identify features needed for the classification of video.
2.3	Design process flow	The overall process flow of the research.
3	Implementation, Evaluation and Results of VGG-16 ,VGG-19 and CNN	Building models, evaluating and generating optimal results
4	Discussion and Comparison	Comparison of implemented models with existing studies

for this paper are mentioned in table 1.1:

This document consists of different parts starting with the literature study in section 2 followed by the methodology section 3 where methods used and process flow of the project is discussed briefly and then implementation section of the research gives details of how the models were implemented and evaluated on the basis of some evaluation criteria and a review on the results obtained in the implementation. As implementation, evaluation and results are inter-dependent, for every model applied its implementation and results are discussed in the section 4. A discussion on the results obtained in the project as well as results already captured by people and comparison of the models are investigated in section 5 followed by concluding the project and future work is included in section 6.

2 Related Work

2.1 Introduction

In the last decade, several researches have been performed on videos that include understanding and exploration of the videos. Studies show that there are various process by which a video can be classified, some analysts have proposed their own techniques and some have improvised the techniques in order to give best possible results. This section is been divided into three subsections based on the literature study for this research like section 2.2 explains classifying Indian dance forms which includes the study conducted to categorise Indian classical dance forms, the second is 2.3 proposed techniques in this area where the new techniques introduced by some researches are discussed and third section 2.4 is identifying human action recognition which focuses on the studies conducted on action recognition.

2.2 Classifying Indian Dance Forms

An algorithm (Shubhangi and Tiwary; 2017) was proposed to recognise Indian classical dance forms like 'Bharatnatyam', 'Kathak', 'Odissi' with 15 different gestures. The database for this research consisted of 100 images which were then split for training and testing purpose. Another research conducted by (Naik and Supriya; 2020) includes 2-D dance image classification using ResNet34 model and fast-AI. Using this CNN architecture, the author was able to achieve the accuracy of approximately 79%. The proposed techniques can be extended by using 3-D dataset with geometric deep learning Approach. Convolutional Neural Networks is an powerful tool in the world of Artificial Intelligence and Computer Vision Kishore et al. (2018) have performed human action recognition on Indian Classical Dance. The dataset for the study was collected online and offline from YouTube. The offline dataset was created using 200 videos of dance mudras/poses. Around 60 frames were made and CNN was trained with 8 different sample sizes and 2 samples were used for testing. With different CNN architecture data was tested with the accuracy of 93.33%. Another study conducted by Biswas et al. (2021) used two pre-trained models VGG-16 and VGG-19 for classifying 8 classes of classical dance form. The dataset used was taken from Kaggle. Out of both the models VGG-19 outperforms VGG-16 with the accuracy of 91.7% whereas VGG-16 can give aximum accuracy of 89.6%. Performance of the models were compared on the basis of accuaracy, precision and recall. In order to increase performance, necessary changes were made in the dense layers of the model. A study named 'Nrityabodha' was conducted by Mohanty et al. (2016) on deep understanding of Indian classical dance using deep learning approaches.A dataset was generated for single hand gestures and double hand gestures in dance, using kinect sensors a dataset for the dance poses was created. Convolutional neural network was used to classify dance poses and hand gestures. As the experiment's denouement, transfer learning beats the traditional supervised learning methods. This study can be extended by applying the proposed method on lyrics and music. Several challenges like occlusions, variation in viewpoint, ambiguity in the movement of hand gesture were faced while implementation. Using a pre-trained model SVM or CNN classifier with MNIST, CIFAR-10, the network resulted with the accuracy of 89.5%. 'Rasabodha' an understanding Indian Classical Dance form by recognising emotions with the help of deep learning approach. Recognising emotions from a video or images is a challenging task as over barriers like costumes, make-up and clutter in enactment, Mohanty and Sahay (2018) had to classify emotions equivalent to 'Navarasas', it is one of the emotions from classical dance. Two datasets CVLND-RGB and CVLND-D were used in this study, and a real world dataset for 'navarasa' was used for the reference to evaluate on the model. Hu and Zhai (2010) has used SVM classifier to extract feature from the cultural relic video, with the experiment it can be seen that the method performs better than the others. The model was constructed with the eigen-vectors and the image frames were divided into 7 categories. An ancient classical dance 'Bharatnatyam' is considered to be having comples gestures and facial expressions. it uses Support Vector Machine as a claasifier with the acccracy of 71.36%. Venkatesh and Babu (2016) have claimed that the project gives insight of predictive analysis, expression classification and analyze dancers in the video.

.

2.3 Proposed Techniques in the area of Dance Video Classification

Maintaining traditional techniques of identifying the video content, new techniques are introduced one of them is proposed by Bakalos et al. (2018) , a framework to recognize the dance poses in the video. For this, kinect sensors were used for RGB capturing and real-time depth sensing. This method involves genuine postures, a sequence of postures will be drawn by the kinect sensor and an appropriate label will be assigned to each frame. A research study by Shuhei Tsuchida (2019) was performed on the same database as used in this study. They proposed four baseline methods to classify dance genre using LSTM and SVM classifier. The two methods were adaptive and L-fixed method. Adaptive method uses beat positions like one, two, three and four are treated as one unit corresponding to the video frames whereas in L-fixed method vectors accumulate within the fixed length units. The two models were applied and results were compared accordingly. The best result of accuracy as 91.4% was obtained by using L-fixed method with LSTM model. Maintaining traditional techniques of identifying the video content, a convolutional graph based video representation was proposed by (Mao et al.; 2019) . With this method, a complex relationship between the frames is addressed by applying graph network hierarchically within frame sequence. Based on the results obtained in this research, the proposed technique outperformed its alternatives such as GRU and LSTM. Support vector machine(SVM) is another classifier which when used with Shannon Entropy Algorithm gives challenging results, as discovered by Maale and Pushpanjali (2019). The proposed algorithm when used with SVM classifier to classify the video dataset resulted with 94.4% accuracy in compared to existing system which yeilded 74.8% accuracy. The existing system here refers to the neural networks. Eventually, this proposed algorithm can be used for real time dance video or it can be extended for classifying western dance as stated by the author.

Dance pose classification methods are still missing detailed analysis and feedback, The paper presented by Matsuyama et al. (2021) have used a different approach to the problem where the authors used three Dimensional joints of the body and wearable sensors, they used a LSTM (Long Short term memory) deep learning model with temporal and trajectory wise structure to classify ballroom dance forms. Thirteen different ballroom dance poses were classified, in order to curate the dataset, performers used a wearable sensor and video were also made. The implemented model was able to achieve 93 percent accuracy. The good approach which they followed is that they have made the dataset public in order to progress the research in the related field.

A study was conducted on how to enhance extraction of Dynamic information from Video Classification, and two solutions were put forward by Li et al. (2021). First, with the use of MIB (Motion Intensification Block) a specific set of channels are encoded explicitly to compare with the patterns extracted by other channels and calculating diversity of convolution neural network. Secondly, to increase the intensity of fused-features which reflects the importance of various channels a spatial-temporal squeeze and excitation block is used. With this improvisation, the dynamic information in 2D backbone network is retrieved without implementing complex architectural convolutions. The dataset used for this experiment was Diving48. When compared with state of art algorithms, the proposed technique gave competing results.

2.4 Identifying Human Actions in the video

Dance motion consists of human actions hand movements, leg movements and so studying human action identification can also be considered for this study. A technique involving CNN and DB-LSTM was introduced by Ullah et al. (2018) in which the video features help in building the convolutional neural network and then result obtained is fed into deep- bidirectional LSTM. In order to decrease the chances of redundancy and complexity, every sixth frame of the video is considered while processing with the video. This division of videos into chunks based on the time interval were examined whether the proposed method can recognise the complex sequence of videos or not. The normal human activities like sleeping, tilting, walking have been experimented on various models like SVM, RCNN and CNN as a base model, the accuracy results obtained about 98.90% for Sample data, 92.50% for SVM Model and 97.92% for RCNN model respectively. With the few modification in the model, it can be used for advance dataset as suggested by (Junagade and Kulkarni; 2020). Robust Non-Linear Knowledge Transfer Model (R-NKTM) proposed by Rahmani et al. (2018) in the area of action recognition. R-NKTM is trained with dummy variable and does not need re-training or fine-tuning the model. An unsupervised technique was used in the study to classify among Action 3D dataset. According to the study conducted by Md Faridee et al. (2019) Body Sensor Network based Convolutional Neural Networks proved efficient in recognizing small steps of dance activities as compared to the traditional techniques of feature engineering approaches. A solution was presented by the Schonfeld (2009) to classify and recognise the multiple interactive activities with the Markov models are known as multiple dimensions. Markov models are distributed multi-dimensional models, as the hidden layers of the model gives the solution for training and classification for general algorithms like forward-backward, viterbi and expectation-maximization. Activity recognition is monitoring human behaviour,

An analysis was performed by Nweke et al. (2018) on different deep learning techniques used for activity recognition which enables automatic feature extraction. deep learning methods like Convolutional Neural Network, recurrent neural network, Auto-encoder, , advantages and disadvantages. With the use of deep learninig method categorisation like generative, hybrid and discriminative. Feature learning can be restricted boltzmann machine were discusssed on the basis of their performance, characteristics enhanced by combining hybrid method with discriminative and generative model.

Scene classification is an another technique in the field of Computer Vision. It can be used in many areas like with images, videos, robotics and video surveillance as suggested by Ye et al. (2018). The main challenge comes as how to increase feature extraction accuracy for the videos having complex backgrounds. Using CNN networks with increased depth helped in improving accuracy for feature extraction. The structure of a network consists of 10 layers out of which 7 layers focuses on extracting features from the video with coal mining concept and the rest 3 layers are connected layers with 'Softmax' loss function. This mine scene classification has overcome the drawbacks of traditional techniques.

2.5 Critique and Comparison of Techniques and Models used in Video Classification

Many studies have been conducted in the area of video classification and computer vision. The novel techniques were proposed out of which some gave excellent results like the one

with 91.4% accuracy with the technique proposed by Shuhei Tsuchida (2019) whereas some studies compared the results of different models based on evaluation criteria. A study conducted by Saikia and Saharia (2021) was not able to extract features of FFBS class in the dataset, using the methods feature matrix feature extracion from videos were performed. A study used NetVLAD and NetFV models to classify videos and as a evaluation method huber loss functions were used. This experiment was performed on the large-scale dataset that is YouTube video data of 8 million videos. The author Shin et al. (2019) performed 2.5 epochs and found optimal performance improvement in the classification. With the help of dance poses indian classical dance can be categorised as stated by Samanta et al. (2012), they used SVM classifier on the video dataset collected from YouTube. Finally 86.6% of accuracy was achieved also it overcomes the problem of 'Bags-of-Words'.

Table 2: Comparison of Literature in Dance Video Classification

Technique Used	**Results Obtained**	**Authors**
SVM Classifier	86.6% accuracy	Samanta et al. (2012)
SVM Classifier	92.50% accuracy	(Junagade and Kulkarni; 2020)
VGG-16	91.7% accuracy in 8 classes	Biswas et al. (2021)
CNN	93% accuracy	Kishore et al. (2018)

From above table 2 it can be observed that the only evaluation method used is accuracy while comparing the performance of models. Understanding and analysing above literature, it can be understood that most of the techniques have been proposed in the area of computer vision. From above literature, it can be observed that the maximum number of classes classified are 8 whereas in this study, the number of classes are 10 and the performance of model will be evaluated on the basis of different evaluation methods.

3 Methodology, Design and Data Pre-processing Approach Used To Classify Dance Genres

3.1 Introduction

Generally deep learning projects does not have any pre-defined methodologies, the process to be followed depends on the quality of the dataset and the expected results. The process flow of the research is explained in the section In this section, a methodology approach used for classifying dance styles is discussed in section 3.2, dataset used for implementation is described in section 3.3, an exploratory analysis was carried out which is explained in section 3.4 followed by the pre-processing and feature extraction explained in detail in the section 3.5 with the brief explanation in section 3.6 of design process flow for of this study.

3.2 Methodology for Dance Video Classification

The process flow designed to implement as seen in the figure6 is according to the model applied. Firstly, raw data is been collected as video files with extension of .mp4 and

then starts data pre-processing as discussed in section 3.5, it is the main part of any implementation as results may vary depending upon the processing of raw data.

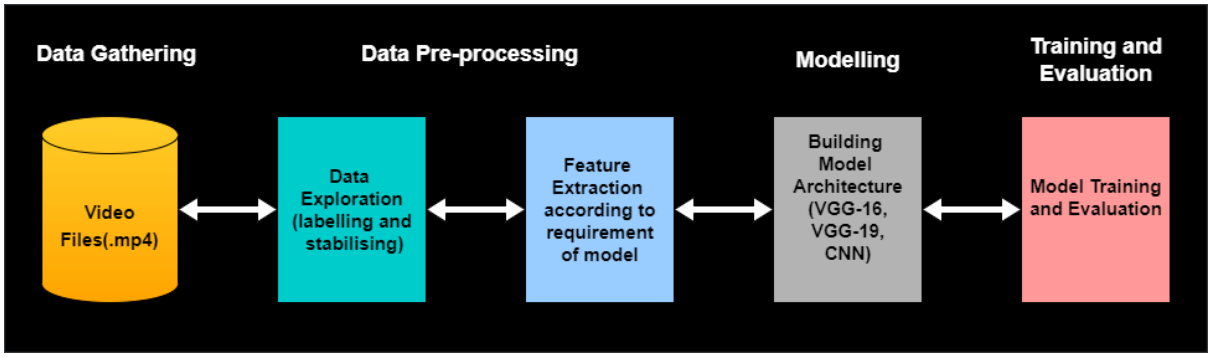

Figure 2: Methodology Approach

The proposed methodology consists of four steps as mentioned below:

- Data Gathering : Video files are collected from the source.
- Data Pre-processing : Exploring, storing and stabilising the data including feature extraction depending on the requirement of the model.
- Data Modelling : Building the architecture of models like VGG-16, VGG-19 and CNN model.
- Training and Evaluation : Training the data according to the need of individual model and evaluating the results on the test data.

3.3 Dataset Description

The dataset for this research has been collected from AIST Dance Academy. It is publicly available[1], consisting of 10 dance genres of street dancing like break dance, pop, locking, middle hip-hop, LA style hip-hop, house, waack, krump, street jazz and bellet jazz. Around 35 dancers have performed to make 13,980 videos, where each video has a standard playing time of around 20 seconds However, considering the scope and time-frame of the research project only a subset of the dataset is used. These videos consists of solo performances, group performances as well as battles between the groups. Videos with the solo performances has been considered for this study.

The AIST Dance Database has 10 street dancing styles with the total of 13,939 videos which includes 49 situational videos of 3 variety of situations and 13,890 videos of 10 types of Street dances. The situations like battle, solo and group dances are covered in the dataset, the most difficult videos to understand are battle dance videos as they have two different groups dancing in a single video. The database also has 60 musical pieces of 10 categories corresponding to Street Dance. These musical pieces can be used for audio classification or other audio processing. A choreographic variation is covered in these videos as 40 professional choreographers have performed to create this database. The choreographers includes 25 male and 15 female members. All dancers had experience of

[1] `https://aistdancedb.ongaaccel.jp/`

more than 5 years in dancing. The videos were recorded in colour as dancers mainly wore monotone clothing while performing. For 10 dance classes, 1080 were basic choreography dance video, 90 were group dance videos and 30 dance videos were recorded with motion camera. Situations like showcase, cypher and battle had 24,20 and 15 video respectively making total count of stuational videos as 49.

3.4 Exploratory Data Analysis

The data gathered should be analysed and investigated critically in order to get to know and understand the data deeply. To perform further analysis on the data, considering the domain of classification the distribution of classes present in the dataset should be analysed. There should not be large difference in the count of each class available in the data as seen in figure 3, the count of each class is represented by the bars and there is not much difference in the lengths of bars representing each class. This means classes in the dataset are nearly balanced. Otherwise, if classes are imbalance that means a large difference in the lengths of bars, then the dataset should be re-sampled by either under-sampling or over-sampling. In this case, data is almost balanced and so further processing can be performed on the dataset.

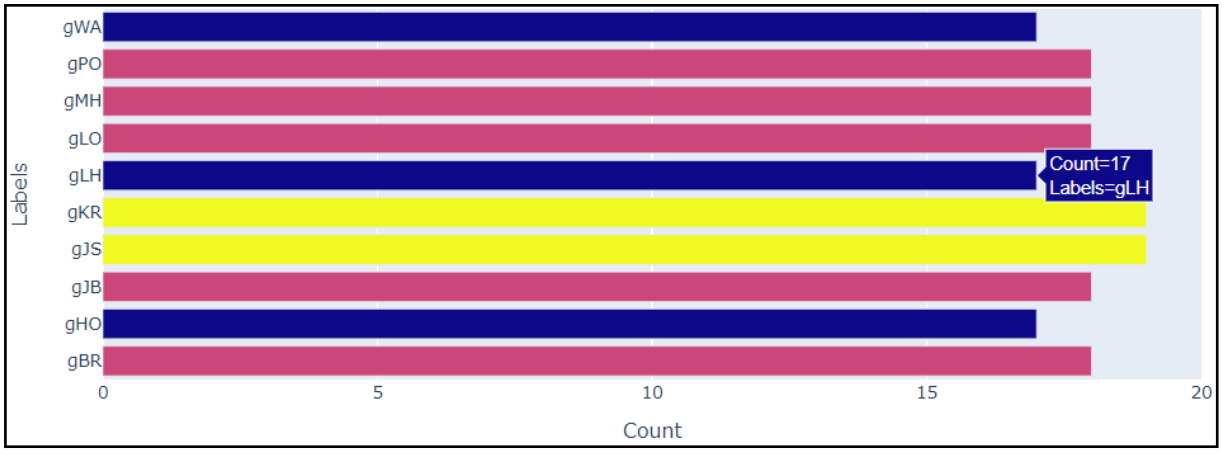

Figure 3: Data-Distribution

	video_name	tag
0	gBR_sFM_c01_d04_mBR0_ch01.mp4	gBR
1	gBR_sFM_c01_d04_mBR1_ch02.mp4	gBR
2	gBR_sFM_c01_d04_mBR2_ch03.mp4	gBR
3	gBR_sFM_c01_d04_mBR4_ch05.mp4	gBR
4	gBR_sFM_c01_d04_mBR4_ch07.mp4	gBR

Figure 4: Data-Analysis

To explore the data and its nature, pandas dataframes are used and so the raw data was stored as a dataframe While exploring the data, it is been observed that the video name consists of few components like the class it belongs to, dancer number, camera angle etc. As seen in figure 4 the video name is 'gBR_sFM_c01_d04_mBR0_ch01.mp4', then the class of video is 'BR' which means break-dance, comes under advanced dance type with dancer number 04, music id and choreography id. In this research, as videos are classified into dance styles only dance class has been taken into consideration.

3.5 Data Pre-processing and Feature Extraction

The pre-processing of the raw data depends on the model used to train and evaluate the results. As a part of pre-processing, all the videos were labelled with their corresponding dance styles as can be seen in table 3.5. A sequence of images run while playing a video, so in order to process with the video, it need to be cut down into frames. These frames can also be called as images of every seconds of the video. The names of videos in the dataset collected were stored in the text file. This text file was then used to label the generated frames of videos into their corresponding dance genre. The name of the videos are like 'gBR_sFM_c01_d04_mBR0_ch01.mp4' where 'gBR' is the Dance style of that particular video. Using pandas data-frame function, videos are captured in the data-frame and an empty list id created. This empty list will store the tags of videos. With the help of split() function the tag of every video from the video name is splitted and stored in the empty list.

Table 3: Labels for Dance Styles

Labels	**Dance Styles**
gBR	Break-Dance
gPO	Pop
gLO	Locking
gMH	Middle Hip-Hop
gWA	Waack
gLH	LA Style Hip-Hop
gHO	House
gKR	Krump
gJS	Street Jazz
gJB	Ballet Jazz

Feature extraction is the process where out of all frames generated from a video, only the ones which are necessary for the model to train and test are considered. For extracting features from the frames, the dimensions of the frames are set to (224,224,3) as for processing all the frames, they should be with same dimensionality. As seen in figure 5, two consecutive frames of an video are captured, the time difference among these two frames are around 5 seconds, thats the reason not much difference is been noticed in these two frames.

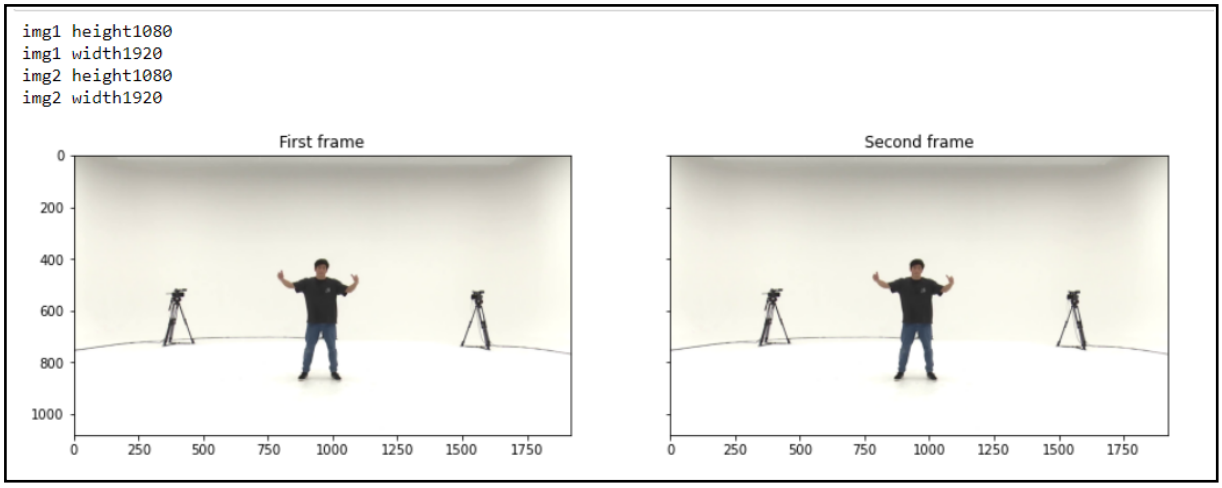

Figure 5: Consecutive Frames of Video

3.6 Design Process Flow

The process flow diagram as shown in Figure 6 represents the abstract level working of the video classification using deep learning, It is divided in two tiers namely presentation tier and business/logic tier. The process starts by collecting video classification dataset and the videos are in '.mp4' format, the raw .mp4 video names are stored in a text file for later retrieval.

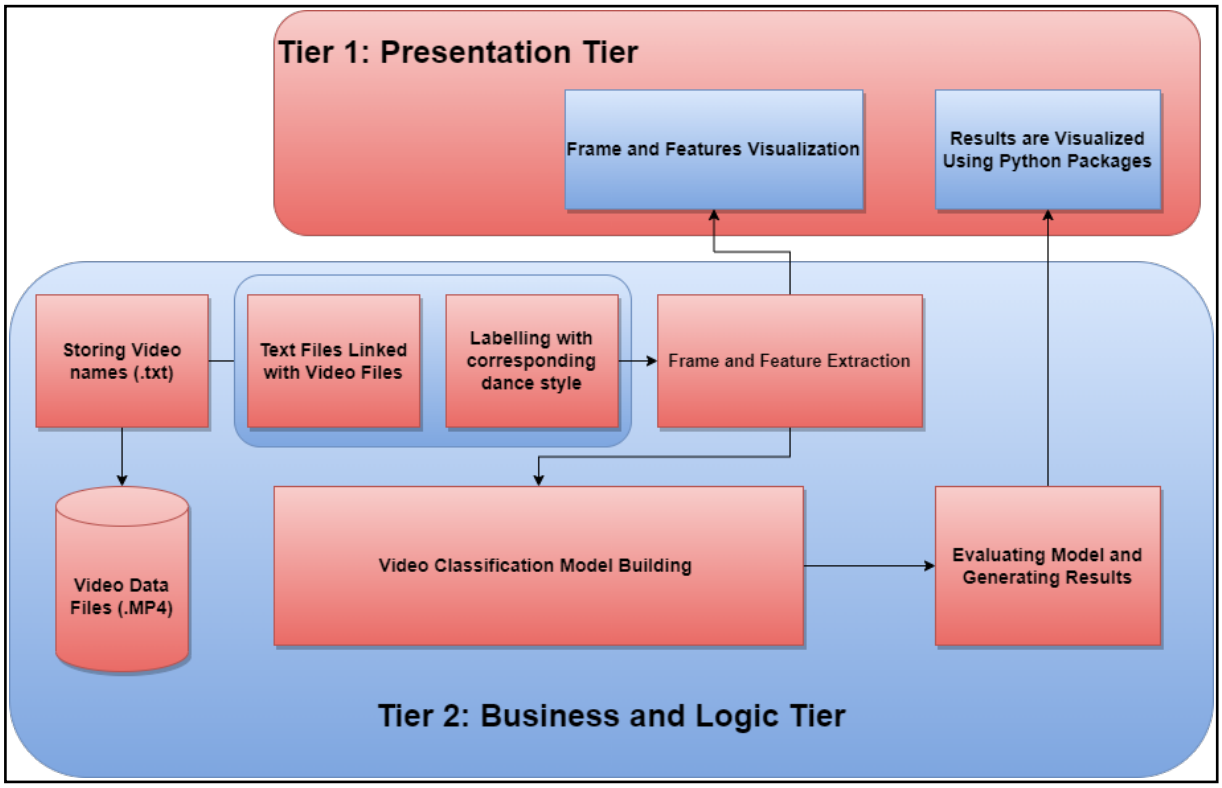

Figure 6: Design Process Flow

The next step consists of linking the text file containing name of the video files and videos followed by labelling the videos in their respective categories/genres, in this paper 10 distinct dance style classes are present.The next step is one of the most important

which includes frame and feature extraction. Videos can be understood as collection or arrays of images stacked together so for the model to learn patterns from the data each consecutive frames/images from the video file is extracted using opencv package of python. The extracted frames are then processed in order to extract features which are essentially collection of numpy arrays. The next step is to define the architecture of the model and preparing the model to be trained on the features extracted. Various hyper-parameters are tuned in order to increase the accuracy and decrease the training loss function of the model. Next step is to evaluate the trained model and generating results.The Presentation tier contains visualisations of feature, frames and the results generated for better understanding.

4 Implementation, Evaluation and Results of 4 Street Dancing Styles Classification Models

Implementation is the crucial phase of any study, it will be discussed in detailed in this section. Implementation is practically performing the proposed methodology with the models. The implementation, evaluation and results for VGG-16 model are discussed in section 4.1. Along with the explanation involving training of the model is discussed in 4.1.1 the evaluation methods used for evaluating model is explained in 4.1.2 and further results obtained are discussed in 4.1.3. And, the implementation of the model VGG-19 is explained in section 4.2.1 followed by the implementation of CNN in section 4.3.1.

4.1 Implementation, Evaluation and Results of VGG-16 Model

4.1.1 Implementation of VGG-16 Model

VGG16 is a convolutional neural network Model used for classification and detection in deep learning. From figure 7, the architecture of the model can be understood that the model consists of 5 dense layer, as it gets deep the density decreases. The first layer is defined with the density of 1024 decreasing to the number of classes which needs to be classified in the implementation that is 10.

Each deep learning classification model requires the data processing to be done in a certain way, the model applied i.e. VGG-16 which is used in this study to classify various dance genres also requires the data to be processed in a certain way for which the raw video data gathered was stored the names of all video files were retrieved and stored in a dataframe. After the names of all video were retrieved the tags of all the videos were extracted from the names of the video for e.g. 'gBR' represents break dance and 'gWA' represents waacking etc. Using openCV a python library which deals with computer vision, was used to read all the videos in a loop and extract the frames from the video at a certain frame rate.A frame rate is number of frames/images stacked in a second. The names of all these frames were then read in a data-frame for further processing. As because of computational limitations the frames cannot be processed in their original size and thus needs to be resized accordingly in this case, the resolution of 224 by 224 pixels were selected. After the rescaling of the images, each consecutive frames were then converted Into NumPy arrays. Normalization were then applied to the NumPy arrays for better optimization. The next step is to split the dataset into training and testing

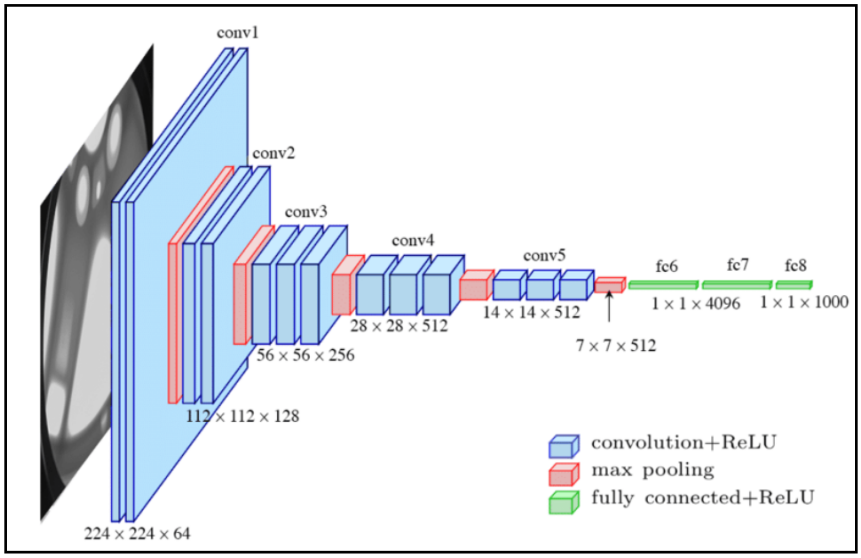

Figure 7: VGG-16 Model Architecture

along with separating labels and features which is done using train_test_split() method provided by sklearn package.

Training the Model : The model is implemented using keras VGG-16 model, the weights for the model defined are selected from 'imagenet'. Specific features for all the frames are extracted using the VGG-16 model, after feature extraction using VGG-16 model the shape of the features were changed to (7, 7, 512), after training features were extracted the features for the validation set were also extracted in the same manner.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 1024)              25691136
_________________________________________________________________
dropout (Dropout)            (None, 1024)              0
_________________________________________________________________
dense_1 (Dense)              (None, 512)               524800
_________________________________________________________________
dropout_1 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_2 (Dense)              (None, 256)               131328
_________________________________________________________________
dropout_2 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_3 (Dense)              (None, 128)               32896
_________________________________________________________________
dropout_3 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_4 (Dense)              (None, 10)                1290
=================================================================
Total params: 26,381,450
Trainable params: 26,381,450
Non-trainable params: 0
_________________________________________________________________
```

Figure 8: VGG-16 Model Summary

As the fully connected model takes only a single Dimensional input the shapes of the features were needed to be changed and were then normalized i.e. assigning the values

between 0 and 1 for each feature this enables the model to converge faster. A sequential architecture were defined for the deep learning classification model as seen in figure 8. 'Relu' activation function was used for the training of the model with multiple sequential dense layers and multiple dropout layers in between consecutive dense layers so as to avoid over-fitting. The final layer contains number of neurons equal to the classes we have to work with and hence in this case was 10 and 'Softmax' activation function was used for last dense layer. The model was now ready to be trained on the training dataset and validated on validation set, 'Adam' optimiser was used for the optimization and accuracy metrics was used in hyper-parameter tuning.

4.1.2 Evaluation

For the Evaluation of the trained model again the architecture was defined for VGG16 model with multiple dense layers and multiple dropout layers and using the same activation function. The weights which were stored during the training phase of the model were retrieved and loaded. The model was again compiled using 'Adam' optimiser and 'accuracy' as the metrics for the Evaluation. The next step was to load the test dataset and creating a data-frame of all the video names present in the dataset. A logic was implemented which uses a for loop to iterate over each video consecutively to predict the class of the video. First using opencv package of python the test videos were captured from the provided path and the data-frame which contained the names of the videos, the frames were extracted from the videos and read one video at a time followed by again rescaling the frame to a 224 by 224 pixel resolution as it will be computationally very costly to use higher resolution images. Frames were then converted in NumPy arrays of specific dimensions to feed into the trained model, and results were stored in a list. The Evaluation methods used for VGG-16 are :

- Accuracy : Accuracy is one of the metric on which the model performance is evaluated. It can be calculated as number of correct predictions divided by total number of predictions.

  $Accuracy = \frac{No.ofCorrectPredictions}{Totalno.ofPredictions}$

- Precision : It can be calculated as the number of videos labelled correctly belonging to their class divided by total number of videos labelled for the class.

  $Precision = \frac{No.ofTruePositives}{TruePositives+FalsePositives}$

- Recall : Recall can be evaluated as ratio of number of dance samples correctly labelled with their dance genre to the total number of dance samples.

  $Recall = \frac{\text{True}_{Positive}}{\text{True}_{Positive} + \text{True}_{Negative}}$

- F1_Score : It is also known as F_Score and can be calculated by considering both false positives and false negatives. Sometimes, F1 can be useful than accuracy.

  $F1_Score = \frac{2*(Precision*Recall)}{Precision+Recall}$

In figure 9, it demonstrates the graph between training loss and validation loss, as evaluation is performed on training and validation set. It can be observed that as the

number of epochs increases training loss decreases and validation loss is fluctuating during the process. The green line which denotes training loss was steady till the epoch 24 and increasing from epoch number 25 whereas if validation loss is observed that is the blue colour line, it is fluctuating and its peaks of losses were at epoch number 10 and 25. Around epoch 30, it can be seen that both validation and training loss are decreasing.

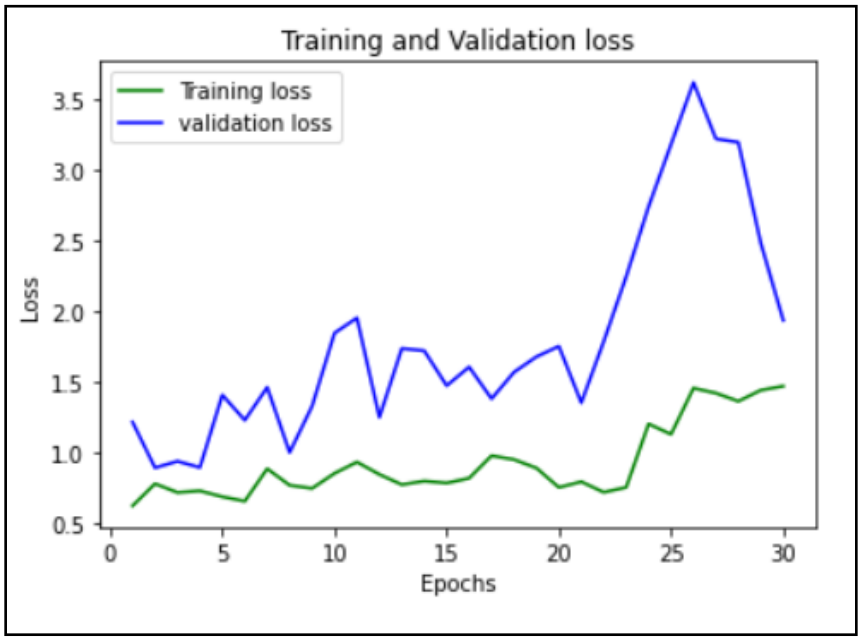

Figure 9: Plot for Training vs Validation Loss

The figure 10 shows the graph between training and validation accuracy denoted by green and blue coloured line respectively. It can be observed that the graph is uneven and starts with accuracy value of both training and validation at its peak, which is above 70%. The training accuracy has maintained value above 50% till the epoch number 25 and after that it starts decreasing gradually, when compared with validation accuracy, it forms a zig-zag pattern, the lowest value for validation accuracy was observed at epoch 26 and then it starts increasing.

The first measure used for evaluating the model was accuracy in which the trained model was able to achieve accuracy of 75.86 percentage.

4.1.3 Results

The accuracy resulted by this model is 47% when the model is trained on 15 epochs and the input frames are given in the batch size of 20 each. Similarly several modifications were made in order to get optimal accuracy. The final accuracy of 75.86 percentage was obtained when the input data was passed in the model. After fine-tuning of various hyper-parameters this accuracy was achieved.

The accomplishment of model can also be determined by the confusion matrix shown in figure 11. It can be observed that the x-axis and y-axis contains total number of classes starting with 0 till 9. Each class is mapped with every other classes, a diagonal pattern of numbers are formed where each class is mapped with vertical version of itself.

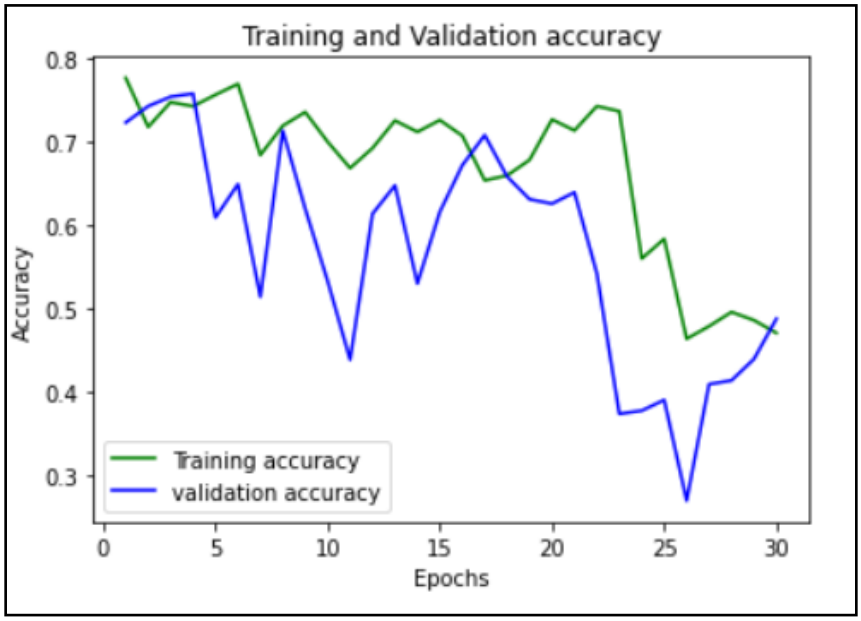

Figure 10: Plot for Training and Validation Accuracy

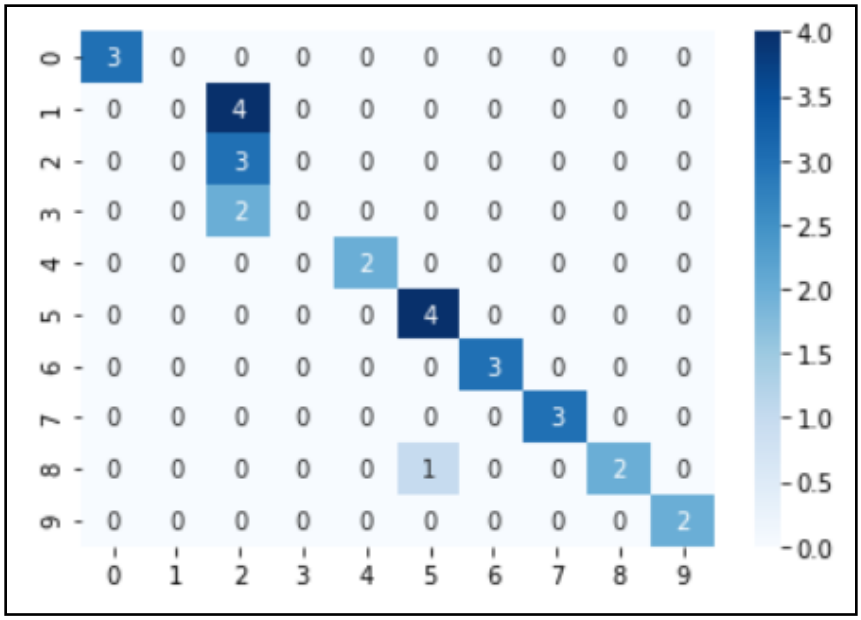

Figure 11: Confusion Matrix

For example, class0 at x-axis matches with class 0 of y-axis gives numerical digit 3, that means the 3 actual samples are matched with the predicted samples.

From figure 16, it can be seen that the actual and predicted lines are overlapping each other, the points where they are overlapping states that while predicting the Dance style, the model has assigned the test video with correct dance style. At points around 5 where the lines are separated by distance, model was not able to predict the actual Dance styles for the video.

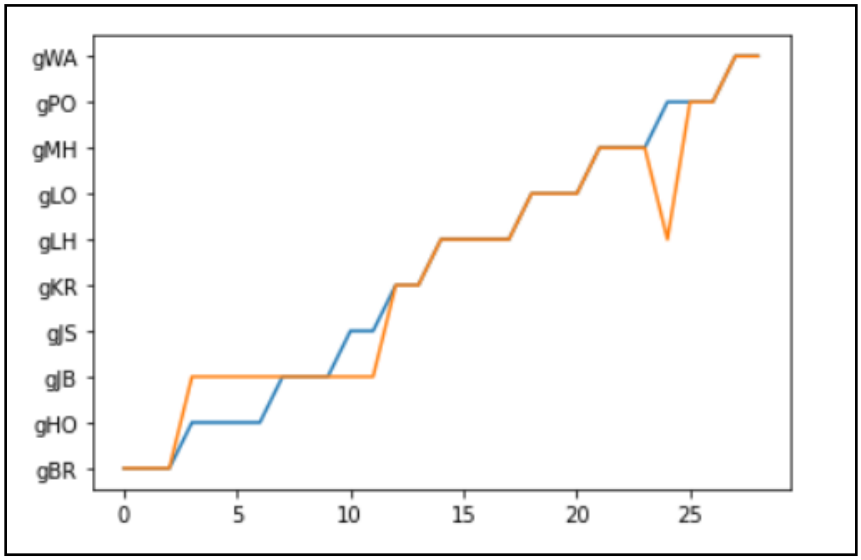

Figure 12: Plot for Actual Value vs Predicted Value

4.2 Implementation, Evaluation and Results of VGG-19 Model

4.2.1 Implementation of VGG-19 Model

VGG-19 is another one of the deep learning neural network model which can be used to perform deep learning tasks. In this research, it was used to perform video classification of various dance style classes. Feature extraction of the previously transformed numpy arrays of the video frames were done using VGG-19 predict function and were stored in different variables. The shape of the arrays were also changed in order to fit the data in the model. Features extracted data was then normalized in order to reduce computational complexity of the model. sequential model architecture of VGG-19 model was then defined by providing appropriate input shape and activation function as 'Relu'. **Training :** The training of the VGG-19 model started with defining a variable which can store the training weights from the training of the model on the dataset. After which model was compiled using optimiser as 'Adam', metric as 'accuracy' and loss function as 'categorical_crossentropy' as the set of selected hyper-parameters. The model was then trained on preprocessed data of video files using various epoch sizes such as 15, 30, 50 and 100 etc. in order to increase the validation accuracy of the model. A set of data which was split from training data was used for the validation of the model.The training weights were saved and different batch sizes were also tried in order to increase validation accuracy of the model.

4.2.2 Evaluation and Results of VGG-19 Model

To evaluate VGG-19 model, the 'test data' created by splitting the actual data was used which consists of videos of all the classes. While training the model, the accuracy achieved was upto 70% on validation data whereas when test data was given as input to the model, the final accuracy achieved was 68.96%. To mininmise the computational complexity of the model the frames are resized into 224 by 224 pixel resolution and converted to numpy arrays. These arrays were mapped with respective classes during testing of model with test data. After combinations of different batch_sizes, number of epochs, activation function and optimizers. The final accuracy of 68.95% achieved was

on activation function named 'relu'.

In figure 13, it can be observed that in the start of epochs the actual and predicted values were diverging and them gradually after sometime both lines starts to overlap each other that means predicted values match the actual value.

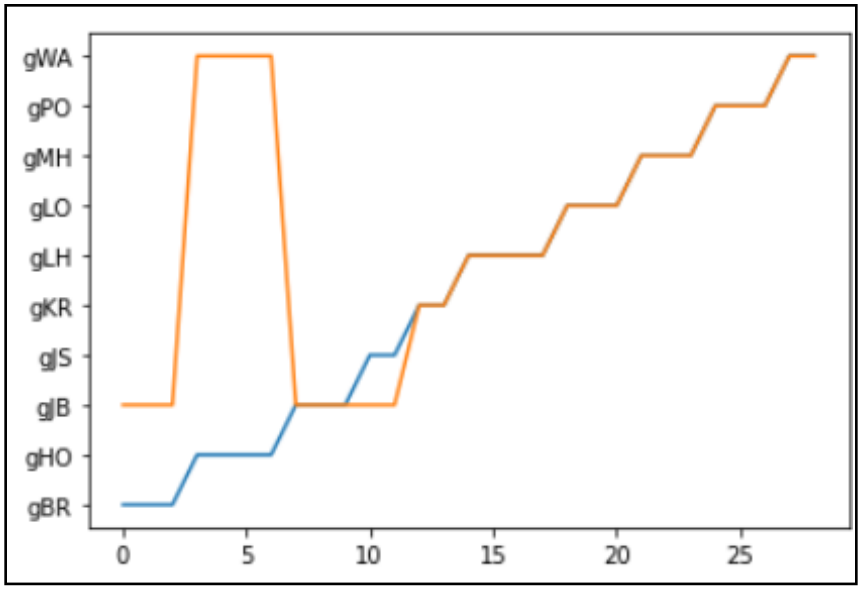

Figure 13: Accuracy of VGG-19 Model

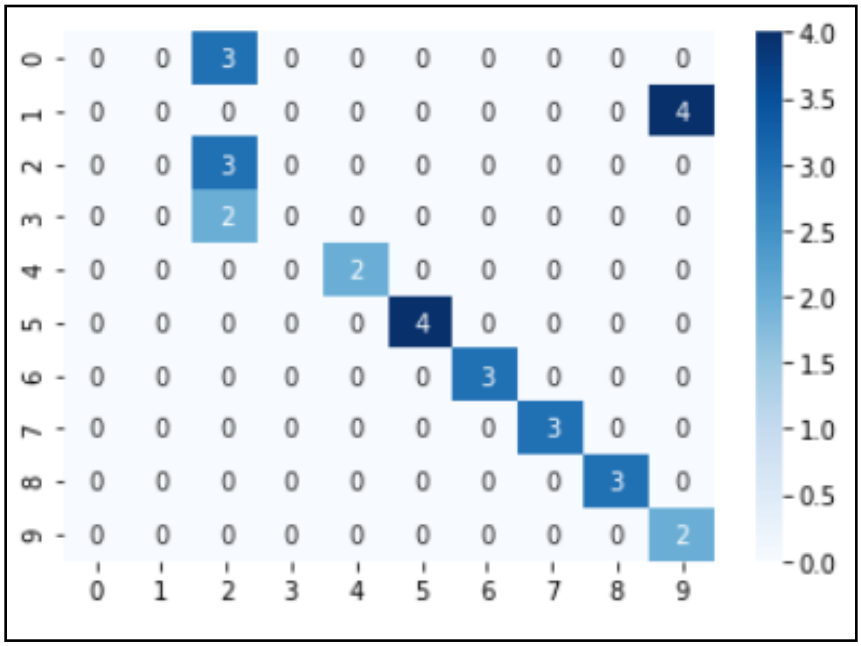

Figure 14: Confusion Matrix of VGG-19 Model

A summary of prediction results can be seen in confusion matrix figure 14, from class 4 to class 9 on the x-axis are mapping with class 4 to class 9 on y-axis, this indicates the total true-positive values obtained my the prediction of model.

4.3 Implementation, Evaluation and Results of CNN Model

4.3.1 Implementation of CNN

A CNN model is deep learning neural network model which can be used for variety of problems. In this study, it was used for solving video classification problem which falls under computer vision domain of deep learning, figure 15 describes the architecture of 2D convolutional neural network. As each model required data to be pre-processed before it can be fed to the neural network model same as the previous model, data was pre-processed. The extracted frames were loaded and rescaled so as to reduce the computational load, after which features were extracted and the frames of the videos were then transformed into numpy arrays and normalized that is assigned a values between 0 and 1. So that it can be fed to CNN model.

Training : The CNN model was imported from keras package and a sequential ar-

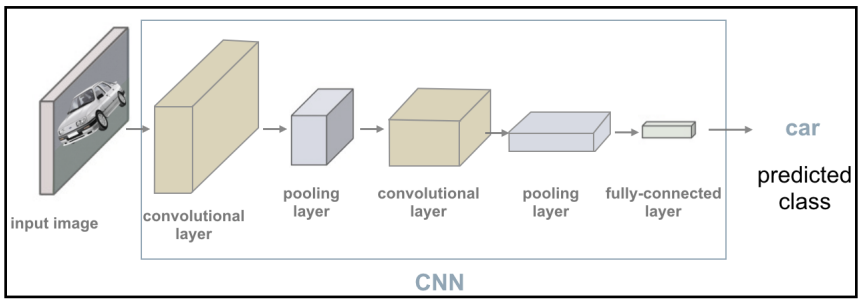

Figure 15: Architecture of CNN

chitecture was defined. In order to create and prepare proper architecture of the model, various layers were defined. The activation function 'Relu' was used for the training of the dataset on the model architecture. While compiling the model various optimizers were tried and tested like 'adam', etc. for the last training 'rmsprop' optimiser was selected and the model was then trained on normalised data with the batch size of 35 and various epoch sizes were also tried and tested. In this way, a lot of hyper-parameter optimization were tried in order to achieve maximum accuracy while training the data.

4.3.2 Evaluation and Results of Convolutional Neural Network

In order to evaluate the convolutional neural network model, the set of data which was kept from training called the 'test data' was used with different number of videos contained for each class. While training the model, it was able to achieve accuracy of upto 60% on the validation data, for testing the raw videos of test file were again converted to frames followed by rescaling the frames to 224 by 224 pixels resolution in irder to reduce the complexity of computation and then converted to numpy arrays, each array was associated wioth its respective class, while actually testing the model on the test data. However, it was only able to achieve accuracy of 27% which is not sufficient enough, in order to somewhat increase the accuracy various hyperparameter optimization were performed but accuracy did not increase significantly.

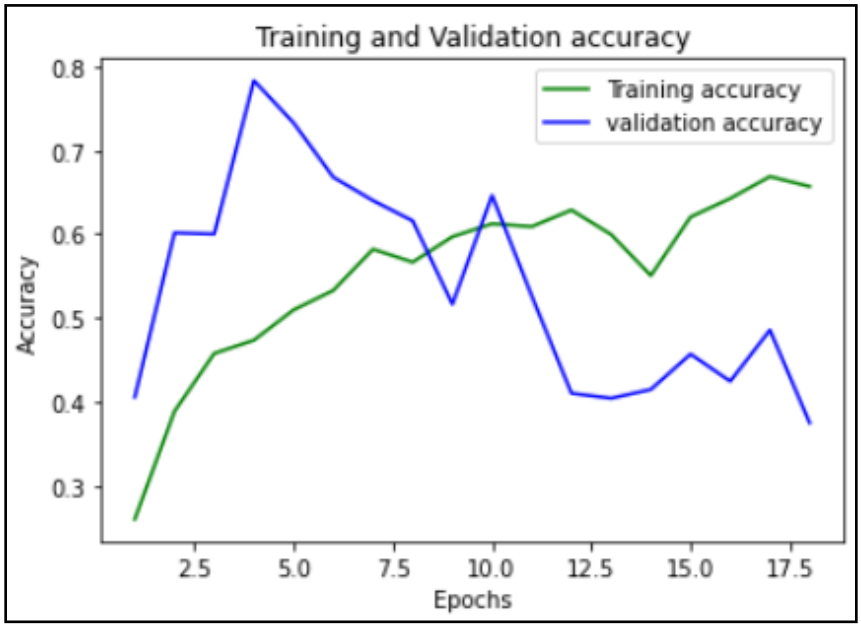

Figure 16: Accuracy Plot of CNN

5 Discussion and Comparison of Implemented Model with Existing Models

When comparing the performance of VGG-16 and VGG-19 models with existing studies, the accuracy evaluated is 75.86 percentage and 68.96 percentage respectively, whereas existing study performed by Biswas et al. (2021) obtained the accuracy of 91.7 percentage from VGG-19 and 89.6% from VGG-16 model. However, the point here should be noted that the study was performed for classifying 8 distinct classes whereas in this study 10 different classes were categorised. This proves that the number of classes plays an important role when comes to the performance of model.

Shuhei Tsuchida (2019) has performed the study on the same dataset, but with different objectives and models, they had applied SVM classifier and a method based on LSTM model to obtain project objectives. The LSTM based model was applied on 60 frames which resulted in 91% accuracy. As the frames are less the accuracy of model is more, but when the number frames input in the model increases the accuracy starts fluctuating. This can be taken into consideration that the number of frames or size of data affects the performance of model.

6 Conclusion and Future Work

The complexity of this research is that it is performed with the video dataset having 10 different classes. Performing this study helped to understand how deep learning Techniques can be helpful when classifying different Dance Styles and so it can be concluded that the research question stated in section 1.1 was solved with maximum accuracy of 75.8% which was achieved with VGG-16 model whereas VGG-19 and CNN models does not perform well giving the accuracy of 68.96% and 27% respectively. As deep learning has its own list of pre-trained models although applying and modifying these models depends on the type and nature of dataset used. VGG-16 and VGG-19 models outperformed

when compared to results achieved from CNN. The results evaluated can be deferred by the nature of dataset and the nature model used to process. All project objectives as stated in Figure 1.1 are successfully completed and are covered in the experiment. Critical examination of state-of-art methods helped in performing this research and achieving the optimal results.

While implementing the project, concepts of deep learning are thoroughly studied. Dealing with video data was a task which was successfully achieved by studying the various types of pre-processing methods and processing of data can also be done by pre-trained models. Concepts of Neural networks and working of model is understood in order to implement in the research.

Future Work : This study can contribute in the field of Video classification where researchers are intended to deal with complex video datasets. This study can be done again with the use of large number of video dataset. As in this research, only VGG-16 model performed well for the classification, other models applied in this study can be manipulated and applied on same dataset with minor modifications. It can be extended by applying various models and also changing the methods of pre-processing technique as enhancing pre-processing of data can also give great results.

Acknowledgement

I would like to thank my supervisor Dr. Catherine Mulwa for her guidance and supervision on this research project. I would like to acknowledge my parents for their support and trusting in me.

References

Bakalos, N., Protopapadakis, E., Doulamis, A. and Doulamis, N. (2018). Dance posture/steps classification using 3d joints from the kinect sensors, *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, pp. 868–873.

Biswas, S., Ghildiyal, A. and Sharma, S. (2021). Classification of indian dance forms using pre-trained model-vgg, *2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pp. 278–282.

Hu, Y. and Zhai, G. (2010). Classification based on svm of cultural relic videos' key frame, *2010 Sixth International Conference on Natural Computation*, Vol. 2, pp. 867–870.

Junagade, N. and Kulkarni, S. (2020). Human activity identification using cnn, *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pp. 1058–1062.

Kishore, P. V. V., Kumar, K. V. V., Kiran Kumar, E. and, S. A. S. C. S., Teja Kiran, M., Anil Kumar, D. and Prasad, M. V. D. (2018). Indian classical dance action identification and classification with convolutional neural networks, *Advances in multimedia*

2018: 1–10.
URL: *https://www.hindawi.com/journals/am/2018/5141402/*

Li, R.-C., Wu, X.-J., Wu, C., Xu, T.-Y. and Kittler, J. (2021). Dynamic information enhancement for video classification, *Image and Vision Computing* **114**: 104244.
URL: *https://www.sciencedirect.com/science/article/pii/S0262885621001499*

Maale, B. R. and Pushpanjali (2019). Recognition and classification of various dance forms by using svm classifier, *International Journal of Research in Advent Technology* .

Mao, F., Wu, X., Xue, H. and Zhang, R. (2019). Hierarchical video frame sequence representation with deep convolutional graph network, *Computer Vision – ECCV 2018 Workshops* p. 262–270.
URL: *http://dx.doi.org/10.1007/978-3-030-11018-5*$_2$4

Matsuyama, H., Aoki, S., Yonezawa, T., Hiroi, K., Kaji, K. and Kawaguchi, N. (2021). Deep learning for ballroom dance recognition: A temporal and trajectory-aware classification model with three-dimensional pose estimation and wearable sensing, *IEEE Sensors Journal* **21**(22): 25437–25448.

Md Faridee, A. Z., Ramamurthy, S. R. and Roy, N. (2019). Happyfeet: Challenges in building an automated dance recognition and assessment tool, *GetMobile: Mobile Comp. and Comm.* **22**(3): 10–16.
URL: *https://doi.org/10.1145/3308755.3308759*

Mohanty, A. and Sahay, R. R. (2018). Rasabodha: Understanding indian classical dance by recognizing emotions using deep learning, *Pattern Recognition* **79**: 97–113.
URL: *https://www.sciencedirect.com/science/article/pii/S003132031830030X*

Mohanty, A., Vaishnavi, P., Jana, P., Majumdar, A., Ahmed, A., Goswami, T. and Sahay, R. R. (2016). Nrityabodha: Towards understanding indian classical dance using a deep learning approach, *Signal Processing: Image Communication* **47**: 529–548.
URL: *https://www.sciencedirect.com/science/article/pii/S0923596516300844*

Naik, A. D. and Supriya, M. (2020). Classification of indian classical dance images using convolution neural network, *2020 International Conference on Communication and Signal Processing (ICCSP)*, pp. 1245–1249.

Nweke, H. F., Teh, Y. W., Al-garadi, M. A. and Alo, U. R. (2018). Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges, *Expert Systems with Applications* **105**: 233–261.
URL: *https://www.sciencedirect.com/science/article/pii/S0957417418302136*

Rahmani, H., Mian, A. and Shah, M. (2018). Learning a deep model for human action recognition from novel viewpoints, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**(3): 667–681.

Saikia, S. and Saharia, S. (2021). *The Sattriya Dance Ground Exercise Video Dataset for Dynamic Dance Gesture Recognition*, Vol. 1248 of *Advances in Intelligent Systems and Computing*.
URL: *www.scopus.com*

Samanta, S., Purkait, P. and Chanda, B. (2012). Indian classical dance classification by learning dance pose bases, *2012 IEEE Workshop on the Applications of Computer Vision (WACV)*, pp. 265–270.

Schonfeld, D. (2009). Motionsearch: Context-based video retrieval and activity recognition in video surveillance, *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 194–194.

Shin, K., Jeon, J., Lee, S., Lim, B., Jeong, M. and Nang, J. (2019). *Approach for video classification with multi-label on youtube-8m dataset*, Vol. 11132 LNCS of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Cited By :1.
URL: *www.scopus.com*

Shubhangi and Tiwary, U. S. (2017). *Classification of Indian classical dance forms*, Springer International Publishing, p. 67–80.

Shuhei Tsuchida, Satoru Fukayama, M. H. M. G. (2019). *AIST DANCE VIDEO DATABASE: MULTI-GENRE, MULTI-DANCER, AND MULTI-CAMERA DATABASE FOR DANCE INFORMATION PROCESSING*, ISMIR.

Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M. and Baik, S. W. (2018). Action recognition in video sequences using deep bi-directional lstm with cnn features, *IEEE Access* **6**: 1155–1166.

Venkatesh, P. and Babu, J. D. (2016). Automatic expression recognition and expertise prediction in bharatnatyam, *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1864–1869.

Ye, O., Li, Y., Li, G., Li, Z., Gao, T. and Ma, T. (2018). Video scene classification with complex background algorithm based on improved cnns, *2018 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, pp. 1–5.