TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

THAPATHALI CAMPUS

PROJECT NO.: THA076MSISE020

DEEP LEARNING-BASED POSE ESTIMATION FOR DYSTONIA SCORE PREDICTION

BY

SUSHANT GAUTAM

A PROJECT

SUBMITTED TO THE DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATICS AND INTELLIGENT SYSTEMS ENGINEERING

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

KATHMANDU, NEPAL

APRIL, 2022

Deep Learning-based Pose Estimation for Dystonia Score Prediction

by

Sushant Gautam

THA076MSISE020

Project Supervisor

Dr. Bishesh Khanal

A project submitted in partial fulfillment of the requirements for the degree of

Master of Science in Informatics and Intelligent Systems Engineering

Department of Electronics and Computer Engineering

Institute of Engineering, Thapathali Campus

Tribhuvan University

Kathmandu, Nepal

April, 2022

COPYRIGHT ©

DECLARATION

I declare that the work hereby submitted for Master of Science in Infomatics and Intelligent Systems Engineering (MSIISE) at the Institute of Engineering, Thapathali Campus entitled "**Deep Learning-based Pose Estimation for Dystonia Score Prediction**" is my own work and has not been previously submitted by me at any university for any academic award. I authorize the Institute of Engineering, Thapathali Campus to lend this project work to other institutions or individuals for the purpose of scholarly research.

Sushant Gautam
THA076MSISE020
April, 2022

RECOMMENDATION

The undersigned certify that they have read and recommend to the Department of Electronics and Computer Engineering for acceptance, a project work entitled "**Deep Learning-based Pose Estimation for Dystonia Score Prediction**", submitted by **Sushant Gautam** in partial fulfillment of the requirement for the award of the degree of "**Master of Science in Informatics and Intelligent Systems Engineering**".

Project Supervisor

Dr. Bishesh Khanal

NepAl Applied Mathematics and Informatics Institute for Research (NAAMII)

Director/Research Scientist

M.Sc. Program Coordinator

Er. Dinesh Baniya Kshatri

Assistant Professor

Department of Electronics and Computer Engineering, Thapathali Campus

April, 2022

DEPARTMENTAL ACCEPTANCE

The project work entitled "**Deep Learning-based Pose Estimation for Dystonia Score Prediction**", submitted by **Sushant Gautam** in partial fulfillment of the requirement for the award of the degree of "**Master of Science in Informatics and Intelligent Systems Engineering**" has been accepted as a genuine record of work independently carried out by the student in the department.

Er. Kiran Chandra Dahal
Head of the Department
Department of Electronics and Computer Engineering,
Thapathali Campus,
Institute of Engineering,
Tribhuvan University,
Nepal.

April, 2022

ACKNOWLEDGMENT

This project work would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study.

First of all, I would like to express my sincere gratitude to my supervisor, **Dr. Bishesh Khanal** along with **Dr. Nabin Koirala and Dr. Ajad Chhatkuli**, of **NAAMII** for providing invaluable guidance, insightful comments, meticulous suggestions, and encouragement throughout the duration of this project work. My sincere thanks also goes to the M.Sc. coordinator, **Er. Dinesh Baniya Kshatri**, for coordinating the project works, providing astute criticism, and having inexhaustible patience.

I am also grateful to my classmates and friends for offering me advice and moral support. To my family, thank you for encouraging me in all of my pursuits and inspiring me to follow my dreams. I am especially grateful to my parents, who supported me emotionally, believed in me and wanted the best for me.

Sushant Gautam
THA076MSISE020
April, 2022

ABSTRACT

Dystonia is a movement disorder that causes unusual movements and involuntary muscle contractions affecting some parts of the whole body. Selecting drugs and doses is a highly personalized process for dystonia, requiring frequent visits to the clinic, pointing toward the need for more systematic and objective methods of collecting patient data. A deep learning-based pose estimation algorithm can be a good candidate for aiding independent clinical assessment of dystonia as it has outperformed the classical approach to human pose estimation. The deep learning-based model can help patients and physicians assess the first symptoms of neurological diseases and build low-cost solutions not only for dystonia score prediction but also to monitor the progress of the disease. Pose estimation algorithms with convolution networks have already been shown to extract relevant information about the motor signals of Parkinson's disease from video assessments, and the calculated score correlates well with the clinical score. OpenPose algorithm was used for human pose estimation in videos of dystonia patients being clinically assessed to annotate body key points in the videos. This project explored the basic pipeline steps required to process the clinical videos, including spatiotemporal keypoints normalization. CNN successfully predicted neck dystonia scores to around the scores obtained from standard clinical assessment, leaving space for further validations and research with more data and methods.

Keywords: Deep Learning, Dystonia Score Prediction, Human Pose Estimation

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AR	Augmented Reality
BFMRDS	Burke-Fahn-Marsden Dystonia Rating Scale
CNN	Convolutional Neural Network
CT	Computed Tomography
DWI	Diffusion-weighted Imaging
GDI	Gini's Diversity index
GDS	Global Dystonia Severity Rating Scale
GPU	Graphics processing unit
HOG	Histogram of Oriented Gradients
IR	Infrared
JRS	Jankovic Rating Scale
LED	Light Emitting Diode.
LID	Levodopa-induced Dyskinesia
ML	Machine Learning
MRI	Magnetic Resonance Imaging
PD	Parkinson's Disease
PET	Positron Emission Tomography
PSF	Pictorial Structure Framework
RCNN	Region-Based Convolutional Neural Network
TWSTRS	Toronto Western Spasmodic Torticollis Rating Scale
URDS	Unified Dystonia Rating Scale
XYZ	Xkk ynkk Znkk

1 INTRODUCTION

1.1 Background

Dystonia is a movement disorder that causes repetitive or twisting movements and involuntarily muscle contractions affecting one part of the body (focal dystonia), multiple adjacent parts (segmental dystonia), or the whole body (general). Such muscular spasms can range from light to severe. The exact cause of the dystonia is unknown, but it could be due to altered nerve cell communication in some brain areas. Although medications can improve symptoms, surgery is used to deactivate or regulate certain regions of the brain and nerves for severe dystonia. People with dystonia experience several symptoms in various body parts, including the neck, eyelids, jaws, tongue, vocal cords, hands, and forearms. Dystonia may also be symptomatic of other diseases and conditions, including Parkinson's (PD). Levodopa, effective in neutralising or reducing motor symptoms, is a standard medication for managing PD but causes motor complications called levodopa-induced dyskinesia (LID) after long-term use, limiting its usefulness and causing twisting of parts of the body into abnormal positions: a kind of dystonia. Consequently, the drug prescription often comprises several drugs to minimise dyskinesia by maximising antiparkinsonian benefits. However, dystonia is a common early symptom of early PD but can occur at any stage. For people with dystonia, selecting medicines and doses is a personalised process that requires frequent clinic visits. They are often asked to keep notes of symptoms on paper. These two methods cannot fully represent the symptoms. As clinic visits are persistent, they will always involve the medical personnel. Instead, patients will choose to stay under existing medication instead of a clinical visit. Various rating scales record motor signs' characteristics (such as anatomical distribution, functional impact, and duration). However, these assessments require special training to administer and are time-consuming to execute. Moreover, the rater's experience can also significantly impact the rating. On the other hand, dystonia patients can be expected to note their symptoms in diaries, which previous studies have also shown. The interpretation of noted symptoms can be very different between patients and doctors. These problems point toward more systematic and objective data collection methods about the patient's state to provide more accurate information.

1.2 Motivation

Being a student of intelligent systems engineering and motivated by the ongoing revolution in automation, I believe that an automated assessment system could address the existing clinical practice problems related to dystonia. It would provide the neurologist with more helpful information that will help in the adjustment of the dose of the drug and enable patients to assess their symptoms by themselves frequently. Computerised evaluations can be more consistent than patient notes or even clinical assessments by a neurologist and could objectively measure motor symptoms. As a screening tool to classify whether someone has PD or dystonia, it could also be accelerating clinical trials involving LID interventions. If we could build an automated system and its parallel validity with current clinical scales for dystonia can be demonstrated, Patients can use this system to speed up a patient evaluation with objective assessment. This could also bring out new landscapes in dystonia diagnosis and reduce subjective bias. With resource constraints in developing countries like Nepal, doctors could also use this system to assess dystonia patients in remote areas with telemedicine. Distance and the need for frequent visits to clinics are always obstacles to medical access.

1.3 Problem Definition

Although portable wearable sensor-based (e.g., accelerometers, gyroscopes, etc.) systems have been proposed for automated assessment of dystonia, it is still questionable if they are clinically valuable and feasible. Alternatively, Engineers and innovators can use computer vision to build a solution for automatic dystonia assessment. Computer vision can extract information from images and videos. In today's world, computer vision applications are already a common thing, including face detection in mobile phones, AR-based video games, automatic vehicle number plate recognition, etc. Recent advances in hardware and algorithms have driven rapid progress in computer vision. For evaluating involuntary movements in dystonia, systems can use vision algorithms as they can be used to track body movements over time for predicting human poses. Cameras that are relatively cheaper and readily available than advanced sensors can be used for automated evaluation without even the need for direct body contact.

However, there is insufficient evidence that we can show the automated system and its parallel validity with current clinical scales for dystonia. Moreover, the quality of the pose estimation from the generic camera also is questionable.

1.4 Project Objectives

The research question to which this project attempts to answer is:

Are computer vision-based approaches capable of an automated non-obtrusive clinical assessment of dystonia? The project briefly explore the feasibility of current computer vision methods to track body movements for clinical dystonia assessment to answer this research question. With the hypothesis that current methods are capable of moment tracking, an algorithm is built to detect the presence of neck dystonia, and the features that could detect neck dystonia has also also be studied.

1.5 Scope of Project

There are multiple types of dystonia, some of which require focusing on specific body parts, such as cranial dystonia, which is characterised by solid muscle contractions of the face, mouth, and/or tongue. However, such a class of dystonia is not within the scope of this study because it requires a more complex model of a particular part of the body.

The project explores the possibility of using pose estimation models for dystonia prediction. Since there are a lot of varieties of models/architectures available, this project take considerations to choose of a model from the perspective of implementation. By no means can this project cope with the falsification of medical records, for example, a video where a patient acts as if he has a gait problem. This limitation is primarily due to the choice of input to the score prediction model, i.e., body poses. Limitations of the pose estimation algorithm directly impacts the performance of the project output. Since the quality of pose estimation is out of the project scope, priority is be given to the effective target implementation for the purpose instead of focusing on tuning the model for accuracy. 3D video capture methods would help to capture poses effectively. However, due to the nature of the available data, we are limited to 2D video, and the information is lost for every movement that is perpendicular to the camera plane resulting in errors in the measurement of joint angles are vital for dystonia.

1.6 Potential Project Applications

- The outcome for these kind of researches are handy for telemedicine where patients and medical personnel are separated by physical distance and various obstacles.
- The research will be helpful to the patients in their home environment for performing regular assessments, and the results could also be sent to their consulting neurologist as well as track the timely changes in the conditions.
- This research can develop system to supplement regular clinic visits by serving as a screening tool for early diagnosis for identifying the need for surgery.
- Outcome of the research can be an inspiration for other pose-based application areas not just limited to the medical domain.

1.7 Originality of Project

The implementation for dystonia score prediction is an original implementation. There have been notable efforts in relating PD scales to the outcome of models trained with pose-based approaches[32]; however, no considerable efforts have been made with dystonia. Understanding the feasibility of dystonia score prediction with pose-based computer vision methods is also a novel exploration. However, the project does not have any originality around the pose estimation method and have use the published procedure for the prediction pipeline,

1.8 Organisation of Project Report

The material in this project report is organised into seven chapters. After this introductory chapter introduces the problem topic this research tries to address, chapter 2 contains the literature review of vital and relevant publications on pose estimation and dystonia score prediction, pointing toward a notable research gap. Chapter 3 describes the methodology for the implementation of this project. Chapter 4 provides an overview of what has been accomplished, including profiling video datasets available, exploring multiple opportunities, and model training. Chapter 5 contains some crucial discussions on the used model and methods. Chapter 6 mentions pathways for future research direction for the same problem or in the same domain. Chapter 7 concludes the project shortly, mentioning the accomplishment and comparing it with the main objectives.

2 LITERATURE REVIEW

2.1 Technologies for Dystonia Assessment

The ongoing researches are in the direction to find the neuronal mechanisms that justify the pathophysiology of dystonia that could lead to developing new strategies for improving clinical management, including the accuracy of its diagnosis, the discovery of new therapeutic approaches, and prediction of the potential population at risk[27].

2.1.1 Vision-based Methods

So far, the assessment based on computer vision-based approaches for the symptoms of dystonia has been very limited. Multicolored suits were used to help segment the body based on previous work on vision-based gait analysis[16]. Background subtraction was used to detect walking participants frame by frame and then used the resulting silhouettes to distinguish patients with a gait problem[9]. Contact sensors and IR LED markers were used to segment records of finger tapping cycles[5]. Some finger-tracking studies required participants to hold their hands on the side of their heads so that face recognition could be used to approximate the position of their hands[21]. Several points were landmarked manually on the body and researchers tracked their motion during a communication task using nonrigid image registration. Lack of coordination between the limbs was related to Part IV of the UDysRS objective disability score for communication [34]. A diagnostic tool that can detect dystonia from MRI scans with high accuracy in less than one second has been developed using deep learning[39]. The latest pose estimation algorithms have proved to be effective in the visual-based assessment and in extracting vital information about PD motor signals from Parkinson's assessment videos and provide a baseline for the performance of future PD studies with deep learning[25].

2.1.2 Wearable Sensing

Wearable systems are popular technologies in the research of dystonia assessment. Movement dynamics are captured by accelerometers, gyroscopes, and magnetometers. Recent advances have been made in the fields of flexible materials, nanomanufacturing, and system integration, providing great opportunities for the development of flexible hybrid electronic devices for human health diagnosis and treatment[23]. Previous work with commercial wearable sensors shows the capabilities of daily movement recording at different body locations[18]. These devices, composed of rigid sensors

and multiple electronic components, require a complex device mounted on the body with adhesives and straps, preventing accurate body movement measurements. They are therefore not applicable for the detection of dystonia, which requires sensitive detection of subtle movements. Advanced material and system integration technologies have been introduced that allow soft, thin-film, and active-wireless bioelectronics[23].

2.1.3 Other Methods

As research tools, various methods of neuroimaging for explaining the brain organization of dystonia are used that include functional MRI(fMRI) to map the brain functional activity and networks, high-resolution structural MRI, and diffusion-weighted imaging (DWI) with tractography for the evaluation of brain structure organization, positron emission tomography (PET) with radiolabelled ligands for mapping neurons, and the pharmacological fMRI (ph-fMRI) for assessing the drug effects on brain function[36].

2.2 Pose Estimation

The DL-based approach has been able to outperform the classical approach to human pose estimation. In comparison to deep learning algorithms like CNNs, classical approaches are unable to capture the geometry and motion information of the human body[28]. The previous work on the estimate of human pose included the random forest implementation within a pictorial structure model to predict human joints. Composed of two components, discriminator and prior, the Pictorial Structure Framework (PSF) is a traditional method of estimating the human pose. The discriminator models the probability that a particular part will be present in a particular place. The prior is the modeling of the probability distribution over the position using the result of the discriminator. The overall objective was that the modeled position should be realistic. For an input image with a human body, PSF represents the human body as a set of coordinates for each body part in the image. Implementation of PSF was done with a nonlinear joint regression model, usually multi-layered random forests. When the image has clear and visible parts, these models worked pretty well, but they cannot capture and model hidden or invisible body parts from a specific angle. Feature-building methods such as contours, histograms, Histogram-oriented Gaussian (HOG), etc. were used to overcome these problems, however, classical models lacked precision, correlation, and generalization capabilities. Therefore, it was only a matter of time to adopt a better approach[1].

A pure encoder decoding network outputs heatmaps for each keypoint for an input image. Mask-RCNN is an efficient architecture that first predicts the boundary box of an image and then predicts the body position within the area of the image in the box. This approach also performs well in detecting multiple poses.

The Convolutional Pose Machines, also an encoder-decoder architecture and iterates the heatmaps prediction refinements using feature extraction and additional network layers[40]. It outputs a single heatmap, and its postprocessing includes identifying the exact pixel where the occurrence probability of a heatmap for each key point is the highest.

OpenPose[7] and PersonLab[30] are variations of the encoder-decoder architecture. In addition to the heatmaps, the output of the model also contains improvements to heatmaps in the form of short, medium, and long-range offsets. Capable of performing 2D real-time multi-person keypoint detection with 15, 18, or 25-keypoints, OpenPose additionally can perform body/foot keypoint estimation, including 6-foot keypoints.

2.2.1 3D Pose Estimation

Recovering a 3D human position from a series of frames is typically the same as recording body kinematics without the use of markers.The quantity of videos produced has exploded with the advancement of technology to capture videos, making it desirable to extract poses from a frame sequence from a video. Algorithms normally take the full movie as input and produce a series of poses after processing through multi stage pipelines. There are efficient algorithms that can analyze video sequences in real time.

However, there are lots of variables that causes people's form and look to alter substantially over time including background variation, occlusion, camera movement, quick motion, loose clothes, and lighting. These issued need to be addressed for efficient estimations.Inherent depth ambiguity is the main reason behind the 3D human posture estimation from a succession of monocular photos being so problematic. Many works use the picture sequence as input to decrease the ambiguity.

A video sequence's continue frames can present several images of the same person, with the regular movement and individual's structure and unchanging bone length.

Learning temporal relationships in frames are done with networks including MLPs[13], LSTMs[19], CNNS[31], TCNs[8] and GCNs[6]. Temporally smooth postures are developed either by penalizing pose-related parameters during training or by optimization of pose trajectories. To overcome the difficulty of 3D human posture estimation, several network designs have been investigated. LSTMs, even sometimes combined with noise models of the Kalman filter[12] and the sequence-to-sequence models are used for modeling temporal relationships of sequences with body parts in motion.

A recurrent 3D pose sequence machine (RPSM) could performs multi-stage refinement that could improve and captures long-range relationships across different body parts, for 3D posture estimation in order to gather rich temporal information ensuring that the anticipated posture sequence is consistent in time[26]. Propagating LSTM networks(p-LSTMs) used CNN to extract the 2D posture before reconstructing the 3D pose [24]. Applying the time restriction to features early in the network was found more successful than applying it to 3D pose predictions[35]. Sequence-to-sequence network made up of LSTM units with shortcut connections on decoder could predicts a succession of 3D poses relative to the root node using prior 2D poses as input which is saved as a fixed-size vector by the decoder. CNN-based structures also enforces temporal consistency in temporal sequences[19]. Spatio-temporal characteristics can be extarcted directly from Spatio-temporal volume of bounding boxes centered on the target frame using 3D CNNs instread of spatial CNNs to directly regress the 3D pose[38]. A temporal dilated convolutional to record long-term information predicts 3D poses using uplifted from 2D keypoint sequences obtained from 2D keypoint detectors[31]. Temporal PoseNet, a two-layer fully linked network with recurrent linear units (ReLUs) can learn complicated structural and motion signals and accepts a set number of neighboring poses as input and produces the appropriate posture[13]. To mitigate the influence of error-prone estimate of occluded joints, incomplete keypoints sequence of 2D poses as input was used in Occlusion-aware 2D temporal CNN. In order to train the 3D TCN and regularize the occluded joints, Cylinder Human Model was used to produce 2D-3D pose pairings with occlusion labels[8]. Graph network can capture multi-scale posture series data and learn a temporal constraint for the pose sequence by creating a local-to-global network to estimate the corresponding 3D pose sequences[6] .

2.3 Dystonia

2.3.1 Pathophysiology and Clinical Presentation

With unknown pathophysiology and exact causes, dystonia is a movement disorder characterized by intermittent or sustained muscle contractions that cause abnormal movements, postures, or both, often repeated that are usually patterned, twisting, and can also be quivering[3]. Associated with muscle overflow activation, dystonia is often caused or worsened by voluntary action. Affecting different muscle groups, focal dystonia is the most common form of dystonia. Treatment of this disorder is currently limited to the management of symptoms, usually through injections of botulinum toxin into the affected muscles[10].

Dystonia is usually classified into primary and secondary dystonia[33]. In contrast, in primary dystonia, other possible causes of dystonia, including acquired or neurodegenerative processes, are ruled out and dystonia is the only neurologic sign[15]. The exact cause of primary dystonia is not known. Even if there is no family history of dystonia, primary dystonia is mostly due to genetic contribution. It is classified further into early-onset and adult-onset forms. The early-onset form usually affects the extremities first and then spreads, becoming generalized in many cases. However, adult-onset dystonia generally remains focal or segmental or involves cranial, cervical, or brachial muscles. Secondary dystonia is caused by an environmental insult and certain identified causes like drug side effects, head injuries, or neurological diseases[29].

2.3.2 Treatment Options

Due to its heterogeneous nature, dystonia could be the result of many neurological disorders which could be treated. It is necessary to thoroughly examine for other diseases including Wilson's disease, hypoxia, Huntington's disease, traumatic brain injury, lipid storage disease, Lee's disease, and Parkinson's disease to rule them out[17].

Acute dystonic movements can be induced by several medications and a careful investigation of the patient's medication list must be performed to rule out those iatrogenic causes. The patient's list of drugs that were administered must be investigated to rule out the causes as several drugs can cause acute dystonic movements[20].

While evaluating for dystonia, several lab tests and studies should be taken into ac-

count, including blood chemistry, copper levels in the blood, ceruloplasmin, and liver functions[4]. Magnetic resonance imaging (MRI) and computerized tomography (CT) scanning of the brain in children and can identify hemorrhagic hypoxia or tumored lesions. Likewise, the slit-lamp examination of Kayser-Fleischer cylinders and the 24-hour urine copper test can also be useful to evaluate dystonia[4]. Genetic testing for genetic anomalies and genetic counseling is important for people with affected relatives or patients with primary dystonia before the age of 30[2].

2.3.3 Rating Scales

To help objectively evaluate dystonia and its reaction to therapeutic interventions, several clinical scales have been developed and are in use[37]. The Global Dystonia Severity Rating Scale(GDS) and Uniform Dystonia Diagnostic Scale (UDRS) developed by the Diagnostic Study Group is mainly used to evaluate generalized disorders. Burke–Fahn–Marsden Dystonia Rating Scale (BFMDRS) is also the commonly used scale. All of these three scales show good internal consistency and correlate well between themselves[37]. To evaluate cervical dystonia, the Toronto Western Spasmodic Torticollis Rating Scale (TWSTRS) was developed. The self-response scale Blepharospasm Disability Index (BSDI) and the Jankovic Rating Scale (JRS), which evaluates the severity and frequency of involuntary eye contraction in blepharospasm patients, are well correlated[14].

Global Dystonia Severity Rating Scale (GDS)

The Global Dystonia Severity Rating Scale (GDS) is a tool to assess the severity of dystonia. GDS is applied globally to 10 body regions. Global scores are an overall score for each area of the body, with separate scores for the right and left limbs (proximate and distant). Total points are the sum of all points for the body regions. The GDS has a maximum total score of 140 points[11].

The investigator rates the patient concerning all patients. The maximum disorder rating is recorded in case if the disorder changes during the examination,

The ten body regions to be tested are listed in the table2.1. Each body region is classified

Table 2.1: Global Dystonia Severity Rating Scale (GDS).

S.n.	Body Area	Ratings										Global Score
1	Eyes and upper face	1	2	3	4	5	6	7	8	9	10	
2	Lower face	1	2	3	4	5	6	7	8	9	10	
3	Jaw and Tongue	1	2	3	4	5	6	7	8	9	10	
4	Larynx	1	2	3	4	5	6	7	8	9	10	
5	Neck	1	2	3	4	5	6	7	8	9	10	
6	Shoulder and proximal arm (Right)	1	2	3	4	5	6	7	8	9	10	
	Shoulder and proximal arm (Left)	1	2	3	4	5	6	7	8	9	10	
7	Distal arm and hand including elbow (Right)	1	2	3	4	5	6	7	8	9	10	
	Distal arm and hand including elbow (Left)	1	2	3	4	5	6	7	8	9	10	
8	Pelvis and proximal leg (Right)	1	2	3	4	5	6	7	8	9	10	
	Pelvis and proximal leg (Left)	1	2	3	4	5	6	7	8	9	10	
9	Distal leg and foot including knee (Right)	1	2	3	4	5	6	7	8	9	10	
	Distal leg and foot including knee (Left)	1	2	3	4	5	6	7	8	9	10	
10	Trunk	1	2	3	4	5	6	7	8	9	10	
											Total Score	

between 0 and 10:

- 0: No dystonia present in that body area
- 1: Minimal dystonia
- 5: Moderate dystonia
- 10: Most severe dystonia

3 METHODOLOGY

3.1 Theoretical Formulations

3.1.1 Human Body Models

The human body is a complex and flexible non-rigid object with many different attributes such as kinematic structure, body shape, surface texture, and the positioning of body components. A human body model does not require incorporating all human body attributes but fits the requirements based on varying degrees of representations and application circumstances. Human pose estimation focuses heavily on human body models. Three different varieties of body models are in use:

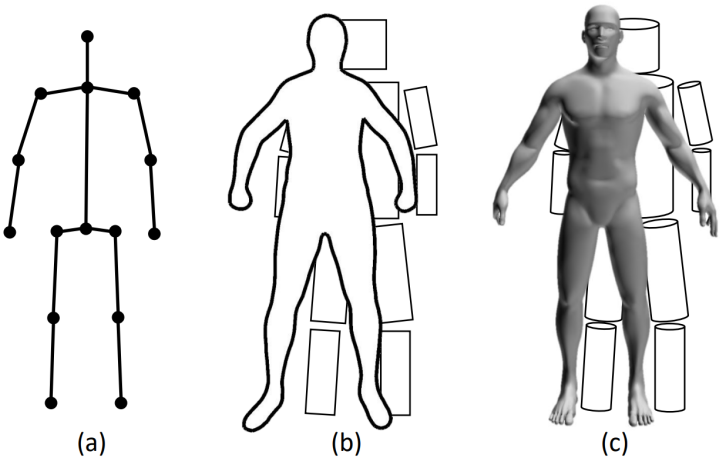

Figure 3.1: Human Body Models

Skeleton-based model: Also known as a stick figure or kinematic model as shown in figure 3.1a, a skeleton-based model is a fundamental and highly adaptable type that depicts a collection of joint positions as vertices and part-pair directions as edges to match the human body's skeletal structure. Such graph representations also model the restrictions or relationships of joints in the skeleton structure. Although extensively utilised in 2D and 3D pose estimations, it has several drawbacks: the lack of texture information as it is the oversimplified version.

Contour-based models: With overall width and contour information for the part-pairs and trunk, in the contour-based models, the rectangle of human silhouette borders indicates portions of the human body approximately as shown in figure 3.1a. It has been frequently utilised in earlier pose estimation techniques. The cardboard and active shape models are two typical shape-based models.

Volume-based models Typically, in 3D applications, volume-based models are used that employ geometry or mesh. Cylinders, conic sections, and other geometric forms were used to characterise bodily parts before meshes could be obtained readily through 3D scanning. Mesh-based models, as shown in figure 3.1a, are the most commonly used volume-based models.

3.1.2 Human Pose Estimations

Human pose estimation techniques evaluate the human body's complex geometry and kinematic information, mostly computer vision. The two common approaches: the classical approach and the deep learning-based approach to human pose estimations, are explored here with the intuitions about how conventional techniques failed to capture the geometry and kinematic information of the human body. However, deep learning algorithms such as CNNs widely dominated the field.

3.1.3 Classical methods to Human Pose Estimation

Classical methods generally refer to strategies and procedures that employ classical machine learning algorithms. The earlier effort to estimate human pose includes deploying a random forest inside a pictorial structure framework (PSF) which could estimate joints in the human body.

The pictorial structural framework, which is one of the earliest conventional approaches to measuring human position, is generally composed of two components:

- **Discriminator**: It identifies the body components by calculating the possibility of a specific body component existing at a particular place.
- **Prior**: It models the probability distribution over pose using the output from the discriminator, making sure that the modelled pose is realistic.

The PSF represents the human body as a collection of coordinates for each body component in a given input picture and employs nonlinear joint regressors, preferably a two-layered random forest regressor. They perform well when the input picture includes distinct and visible part-pair; however, they fail to capture and model when the part-pair is concealed or not visible from a specific perspective. Feature construction approaches

such histogram oriented gaussian (HOG), contours, histograms, etc., were applied to tackle these challenges. Even with those improvements, the classical model could not give good accuracy, correlation, and generalisation ability; therefore was replaced by deep learning-based approaches.

3.1.4 DL based Human Pose Estimation

New issues concealed as research and development began to take off in estimating human pose, without including tacking multi-person poses in the image. Although DNNs were quite effective at predicting a single human posture, they struggled to estimate multi-humans in the picture that may include considerable numbers of individuals in different situations. As the number of individuals increases, the interaction between poses increases, leading to computational difficulties. An increase in processing difficulties frequently leads to an increase in inference time in real-time. The researchers developed two approaches in order to solve these challenges:

- **Top-down**: Localise the people in the picture or video frame and then estimate the body components to compute the poses.
- **Bottom-up**: Estimate the components of the human body in the picture and then estimate the poses.

3.1.5 Top-down approach

Classical top-down techniques use observation from the picture to match it with the direct model. A priori human model is employed to represent the observations that are then continually updated. The models are often quite comprehensive and are capable of delivering the necessary information at any time. A computer programme handles the models' representations and updates during observations. Different kinematic constraints give these models the ability to manage occlusion, which is one of its most notable advantages. Various joints represent human body parts in the direct model, and sticks are used to connect these joints that depict a human body. The top-down method is highly problematic due to much inaccuracy in both human body localisation and body pose predictions. This problem of imprecise human detection is handled using a two-step methodology consisting of two different components: Person Detector and Single Person Pose Estimator (SPPE).

Person Detectors (for example, Symmetric Spatial Transformer Network (SSTN)) crop out the suitable area in the input, which simplifies the classification process, leading to more excellent performance. The bounding box output consisting of the exact area with a single human body visible is fed to the single person pose estimator to extract and estimate the human pose. Such a strategy could extract a high-quality single-person area from an erroneous bounding box by adding a person detector to the SPPE and improves classification performance by addressing invariance while offering a solid framework to predict human posture.

Another popular approach is using the mask segmentation (for example, Mask-RCNN) output provided by the network to recognise people in the supplied input image. As mask segmentation is quite exact in object detection, the human posture as an object is approximated pretty well. The person detection stage and keypoint detection stage are independent of each other.

3.1.6 Bottom-up approach

In this approach, the network first recognises the body parts or critical points in the picture and then maps the relevant vital points to generate pairings. Methods like DeepCut could concurrently perform the tasks of detection and posture estimation. The concept was to identify all potential body parts in the provided picture, then label them, such as a head, hands, legs, etc., followed by separating the body parts belonging to each individual. To constrain the final output to resemble a viable skeletal representation of the human, the network employs Integral Linear Programming (ILP) modelling to automatically arrange all the identified essential points in the provided input.

CNN has been chiefly used as fundamental architecture to extract patterns and representations from the provided input in recent years. OpenPose uses CNN for feature extraction. The output features of the extractor are then fed to two different networks. The first type of network predicts a set of confidence maps for each body component. Similarly, the second branch predicts Part Affinity Fields (PAFs), representing a degree of linkage between parts. Such information is used to prune the weaker linkages in the bipartite graphs.

The predictions from both branches and the features are concatenated for the following step to model a human skeleton based on the number of persons present in the input. The prediction is improved with the successive stages of the feature extractor.

3.2 Data analysis and Modelling

Machine learning automates pattern recognition processes to make predictions by learning the relationship between a set of data measured, *features* and *observations* and its corresponding results. For each example, a set of consistent n features $x \in \mathbb{R}^{1\times n}$ is measured with its corresponding result, y. While data is collected from several number of observations N, the measured $\mathbf{X}$ data are compiled into a *design matrix* $\mathbf{X} \in \mathbb{R}^{N\times n}$. For a single outcome measured for each observation, a *response vector* $\mathbf{y}$ is assembled for all outcomes as $\mathbf{y} \in \mathbb{R}^{N\times 1}$. Machine learning tries to learn the function or model of $f(\cdot)$ to solve $\mathbf{y} = f(\mathbf{X})$.

In supervised learning, the results of each observation are known before feeding to the learning algorithm. On the other hand, the observation results are not known in unsupervised learning.

3.2.1 Statistical Data Analysis

A dataset can be statistically analysed from two viewpoints: exploratory and confirmatory. In an exploratory analysis, the fundamental characteristics of the data are explored, which motivates the future analytical directions. This is generally the visualisation of the distributions and relationships between the underlying components in the dataset. Once a research goal and hypothesis are formulated based on exploratory analysis, the confirmatory analysis uses statistical hypothesis tests to prove or reject the hypothesis.

Correlation Analysis

Correlation analyses are exploratory analyses between random variable sets that determine the relationship or dependency quantified by a correlation coefficient that determines the typical variation of random variables. Pearson correlation coefficient is mainly used for continuous values, which analyses the covariance of the two measurements by their standard deviations as in the equation3.1.

Spearman's rank correlation, as quantified by the equation, 3.2is another famous coefficient that is less concerned about the similarity of two random variables. It evaluates the monotonic relationship between measurements meaning whether the increase in one measure corresponds to the increase in the other measurement.

X and Y in the equations are two different sets of random variables under examination, and $\bar{x}$ and $\bar{y}$ are their means. Both Pearson correlation coefficients and Spearman rank correlation are within the range [-1, +1]. The Pearson correlation coefficient of +1 indicates a linear relationship and is perfectly proportional, and the coefficient of -1 indicates inverse proportionality and a linear relationship. A value of zero correlation coefficient indicates no relationship between the two variables.

$$\begin{aligned} \mathrm{Corr}(X,Y)_{\mathrm{Pearson}} &= \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} \\ &= \frac{\sum_n^{i=1}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_n^{i=1}(x_i-\bar{x})^2 \sum_n^{i=1}(y_i-\bar{y})^2}} \end{aligned} \tag{3.1}$$

$$\mathrm{Corr}(X,Y)_{\mathrm{Spearman}} = 1 - \frac{6\sum_{i=1}^{n}(x_i-y_i)^2}{n(n^2-1)} \tag{3.2}$$

Hypothesis Testing

Hypothesis tests are a form of confirmation analysis performed to verify or reject hypotheses concerning data collection statistically. A *null hypothesis* is the assumption of expected discovery in the collected data, for example, if two samples are under consideration come from the same distribution. A hypothesis test generates a value of significance, called a *p-value*. A test result of a hypothesis is considered *significant* if it has not occurred by chance, as determined by the *level of significance*, a threshold p-value. The typical value for the level of significance used is 0.05, indicating a 5% risk of concluding from the test that there is a difference even if there is no real difference.

3.3 Mathematical Modeling

3.3.1 Body Poses using Part Affinity Fields

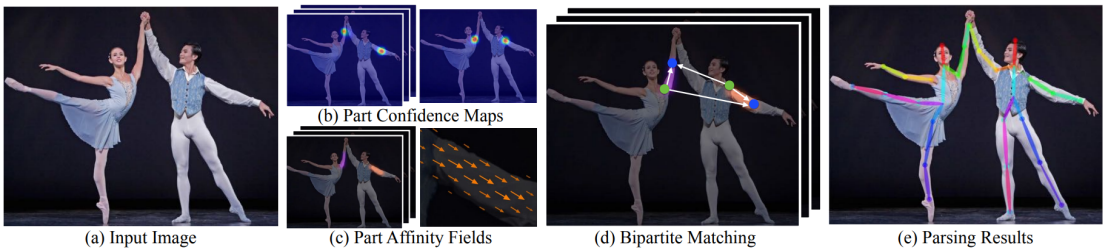

Figure 3.2: OpenPose Pipeline

OpenPose pipeline, as shown in 3.2 can output positions of multi-person 2D human body keypoints for the input of an arbitrary image size $w \times h$. The feature maps ϕ^1 processed by the CNN from the input image is fed to a network to obtain $\mathbf{S}$: a set of 2D confidence maps of body part locations and $\mathbf{L}$: 2D vector fields of part affinity fields (PAFs), $\mathbf{L}^1 = \phi^1(\mathbf{F})$ encoding the degree of connection between the parts. Confidence maps of body part locations: $\mathbf{L}$ is defined at equation 3.3 and and 2D vector fields of part affinity fields (PAFs): $\mathbf{L}$ is defined at equation 3.4.

$$\mathbf{S} = (\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_J), \text{ where } \mathbf{S}_j \in \mathbb{R}^{w \times h}, j \in \{1 \ldots J\} \tag{3.3}$$

$$\mathbf{L} = (\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_C), \text{ where } \mathbf{L}_c \in \mathbb{R}^{w \times h \times 2}, c \in \{1 \ldots C\} \tag{3.4}$$

J confidence maps are in $\mathbf{S}$, one per part and C vector fields are in $\mathbf{L}$ contains , one per part-pair encoding a 2D vector in each picture position in $\mathbf{L}_c$. A greedy algorithm uses these confidence maps: $\mathbf{S}$ and PAFs: $\mathbf{L}$ to output the 2D key points for each person detected in the picture at every stage as in equation 3.5.

$$\mathbf{L}^t = \phi^t \left(\mathbf{F}, \mathbf{L}^{t-1}\right), \forall 2 \leq t \leq T_P \tag{3.5}$$

where ϕ^t is the feature map at stage t out of total PAF stages T_P. Confidence maps detection is rerun after every T_P repetition beginning with the most recent PAF prediction as in equation 3.6.

$$\begin{aligned} \mathbf{S}^{T_P} &= \rho^t \left(\mathbf{F}, \mathbf{L}^{T_P}\right), \forall t = T_P, \\ \mathbf{S}^t &= \rho^t \left(\mathbf{F}, \mathbf{L}^{T_P}, \mathbf{S}^{t-1}\right), \forall T_P < t \leq T_P + T_C \end{aligned} \tag{3.6}$$

Out of the total confidence map stages T_C, for inference at t, the feature map is ρ^t. The original feature maps $\mathbf{F}$ are always concatenated with predictions from the previous stage to construct improved predictions at each subsequent step. The body component positions may be predicted from the PAF channel output. However, with no additional information, just many body parts cannot be assigned to the human body. Thus, revising affinity field predictions increases the confidence map results, whereas confidence map score revision has no such performance. Change among confidence map in two consecutive stages are barely noticeable visually as the confidence map results are forecasted with the slightly improving PAF predictions. Here, the ultimate objective to minimise is shown in equation 3.7.

$$f = \sum_{t=1}^{T_P} f_{\mathbf{L}}^t + \sum_{t=T_P+1}^{T_P+T_C} f_{\mathbf{S}}^t \tag{3.7}$$

where $f_{\mathbf{L}}^{t_i}$ and $f_{\mathbf{L}}^{t_k}$ are the loss functions at stage t_i and step t_k of the PAF branch and the confidence map branch respectively. At the end of each iteration, these geographically weighted $L2$ loss functions between the calculated values and the ground truth PAF: $\mathbf{L}_c^*$ as well as the ground truth part confidence map: $\mathbf{S}_j^*$, makes the network forecast PAFs of body parts first and then confidence map repetitively, and are expressed in equations 3.8.

$$\begin{aligned} f_{\mathbf{L}}^{t_i} &= \sum_{c=1}^{C} \sum_{\mathbf{p}} \mathbf{W}(\mathbf{p}) \cdot \left\| \mathbf{L}_c^{t_i}(\mathbf{p}) - \mathbf{L}_c^*(\mathbf{p}) \right\|_2^2, \\ f_{\mathbf{S}}^{t_k} &= \sum_{j=1}^{J} \sum_{\mathbf{p}} \mathbf{W}(\mathbf{p}) \cdot \left\| \mathbf{S}_j^{t_k}(\mathbf{p}) - \mathbf{S}_j^*(\mathbf{p}) \right\|_2^2 \end{aligned} \tag{3.8}$$

$\mathbf{W}$ is a binary mask that prevents penalisation of real positive predictions and also helps to tackle vanishing gradient problems during training. With the annotation absent at pixel p, $\mathbf{W}(\mathbf{p}) = 0$.

Body Part Detection with Confidence Maps

Annotated 2D keypoints are used to calculate the ground truth confidence maps $\mathbf{S}^*$, indicating the occurrence of certain body components at a given pixel. For each person, k in the image, there exist a peak in each confidence map corresponding to the associated visible part j in the ground truth location $\mathbf{x}_{j,k} \in \mathbb{R}^2$. In $\mathbf{S}_{j,k}^*$, any position $\mathbf{p} \in \mathbb{R}^2$ is given

by equation 3.9.

$$\mathbf{S}^*_{j,k}(\mathbf{p}) = \exp\left(-\frac{\left\|\mathbf{p}-\mathbf{x}_{j,k}\right\|_2^2}{\sigma^2}\right) \tag{3.9}$$

Where the peak dispersion is controlled by σ and various such confidence maps are max-aggregated to output ground truth confidence map $\mathbf{S}^*_j$ by the network as shown in equation 3.10.

$$\mathbf{S}^*_j(\mathbf{p}) = \max_k \mathbf{S}^*_{j,k}(\mathbf{p}) \tag{3.10}$$

Similarly, body parts candidates are predicted from confidence maps on test time by conducting non-maximum suppression.

Full-body pose using Part Affinity Fields

In the scenario of multi-human pose estimations with an unknown number of human bodies present in the image, we need the information on the association of body part-pairs per human body present. With only the positional information of the parts, a straightforward approach to establishing a belief about the connections can be misleading. Part Affinity Fields (PAFs) alleviate these constraints by also considering the orientation information over the area of support of the part-pair. Each PAF representation is a 2D vector field encoding the direction that points from one body part to the other in the corresponding pair.

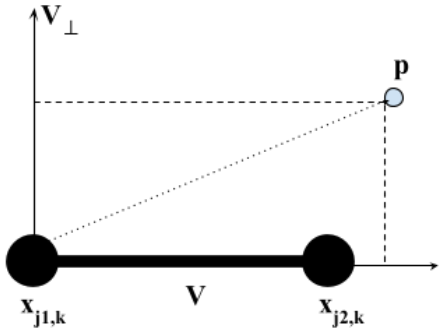

Figure 3.3: A point p around the body part-pair ($\mathbf{x}_{j_{21}}$, $\mathbf{x}_{j_2}$).

For a part-pair c consisting of body parts j_1 and j_2 with the ground truth positions as $\mathbf{x}_{j_1,k}$ and $\mathbf{x}_{j_2,k}$ respectively of one of the person k in the image, the ground truth PAF representation for a point $\mathbf{p}$ that lies precisely in the part-pair is a unit vector pointing from j_1 towards j_2 and is represented as $\mathbf{L}^*_{c,k}(\mathbf{p})$ and is zero-valued if the point $\mathbf{p}$ is not

on the part-pair as shown in the figure 3.3. This has been represented mathematically at equation 3.11.

$$\mathbf{L}^*_{c,k}(\mathbf{p}) = \begin{cases} \mathbf{v} & \text{if } \mathbf{p} \text{ on limb } c,k \\ \mathbf{0} & \text{otherwise.} \end{cases} \tag{3.11}$$

The collection of points on the part-pair is defined as those within a distance threshold of the line segment, i.e., those points p for which a relation shown in equation 3.12 holds.

$$0 \leq \mathbf{v} \cdot \left(\mathbf{p} - \mathbf{x}_{j_1,k}\right) \leq l_{c,k} \text{ and } \left|\mathbf{v}\perp \cdot \left(\mathbf{p} - \mathbf{x}_{j_1,k}\right)\right| \leq \sigma_l \tag{3.12}$$

where the limb width σ_l is a distance in pixels that controls spread, $l_{c,k} = \left\|\mathbf{x}_{j_2,k} - \mathbf{x}_{j_1,k}\right\|^2$ is the absolute length between parts in the pair, and $\mathbf{v}_\perp$ is a vector perpendicular to the unit vector $\mathbf{v}$ in the direction of the part-pair: $\mathbf{v} = \left(\mathbf{x}_{j_2,k} - \mathbf{x}_{j_1,k}\right) / \left\|\mathbf{x}_{j_2,k} - \mathbf{x}_{j_1,k}\right\|^2$.

Average-aggregation of the affinity fields of all persons in the picture gives the ground truth component affinity field $\mathbf{L}^*_c(\mathbf{p})$ as represented in equation 3.13.

$$\mathbf{L}^*_c(\mathbf{p}) = \frac{1}{n_c(\mathbf{p})} \sum_k \mathbf{L}^*_{c,k}(\mathbf{p}), \tag{3.13}$$

where at position $\mathbf{p}$, for all the k persons in the image, $n_c(\mathbf{p})$ is the count of non-zero vectors.

The confidence in the association between two potential part locations $\mathbf{d}_{j_1}$ and $\mathbf{d}_{j_2}$, can be calculated by sampling the predicted PAF, $\mathbf{L}_c$ along the connecting line segment, interpolating between the position $\mathbf{p}(u)$ between the two body parts with varying uniformly-spaced u during testing. These have been expressed mathematically in equation 3.14 and 3.15.

$$\mathbf{p}(u) = (1-u)\mathbf{d}_{j_1} + u\mathbf{d}_{j_2}. \tag{3.14}$$

$$E = \sum_{u=0}^{u=1} \mathbf{L}_c(\mathbf{p}(u)) \cdot \frac{\mathbf{d}_{j_2} - \mathbf{d}_{j_1}}{\left\|\mathbf{d}_{j_2} - \mathbf{d}_{j_1}\right\|_2} d_u \tag{3.15}$$

Tackling Multi-Person detection utilising PAFs

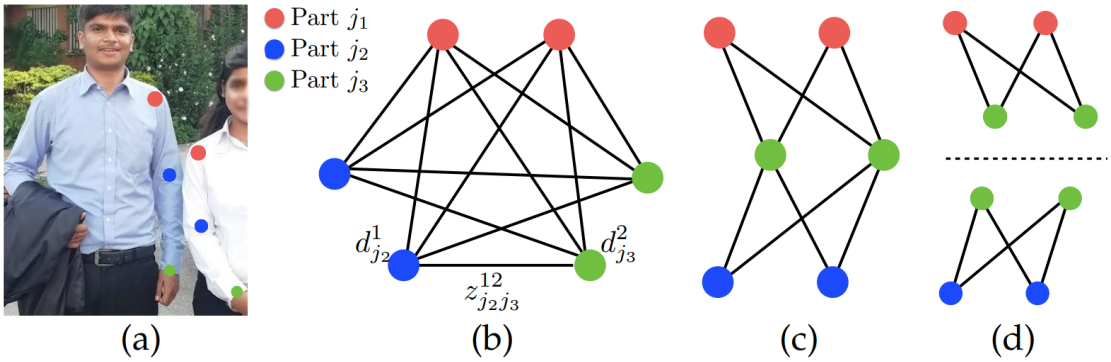

Figure 3.4: The complex problem of graph matching of poses.

Non-maximum suppression is applied on the detection confidence maps to get a discrete set of part candidate locations $\mathbf{D}_J = \left\{ \mathbf{d}^m_j : \text{ for } j \in \{1 \ldots J\}, m \in \{1 \ldots N_j\} \right\}$, where $\mathbf{d}^m_j \in \mathbb{R}^2$ is the corresponding location of the m-th detection candidate of body part j and N_j is the number of pair candidates for part j.

As shown in figure a, each candidate part location3.4a, will still be numerous pair candidates for the potential pair as in 3.4b. Line integral calculation on the PAF for a large number of each candidate to find the best pair is an NP-Hard K-dimensional matching problem. Because due to the broad receptive field of the PAF network, the pair-wise association scores inherently reflect global context, a greedy relaxation can be performed.

The aim is to discover the best assignment $\mathbf{d}^n_{j_2}$ for part $\mathbf{d}^m_{j_1}$ out of all possibilities and can be formulated as in equation 3.16;

$$\mathbf{Z} = \left\{ z^{mn}_{j_1 j_2} : \text{for } j_1, j_2 \in \{1 \ldots J\}, m \in \left\{1 \ldots N_{j_1}\right\}, n \in \left\{1 \ldots N_{j_2}\right\} \right\} \tag{3.16}$$

Variable $z^{mn}_{j_1 j_2} \in \{0, 1\}$ indicates whether two detection candidates are actual pairs.

Finding the optimum association $\mathbf{Z}_c$ for the c-th part-pair with consideration of only a single pair of components j_1 and j_2 boils down to a problem of maximum weight bipartite graph matching with the body part detection candidates as the nodes of the graph: $\mathbf{D}_{j_1}$ and $\mathbf{D}_{j_2}$ and all connections between the possible pairs of candidates are weighted by the part affinity aggregate E_{mn}. The match with the most significant weight

E_c for the given edges is the best candidate part. The Hungarian algorithm can be used to solve the following problem shown in equation 3.17.

$$\begin{aligned} \max_{\mathbf{Z}_c} E_c = \max_{\mathbf{Z}_c} \sum_{m \in \mathbf{D}_{j_1}} \sum_{n \in \mathbf{D}_{j_2}} E_{mn} \cdot z^{mn}_{j_1 j_2} \\ \text{s.t. } \forall m \in \mathbf{D}_{j_1}, \sum_{n \in \mathbf{D}_{j_2}} z^{mn}_{j_1 j_2} \leq 1, \\ \forall n \in \mathbf{D}_{j_2}, \sum_{m \in \mathbf{D}_{j_1}} z^{mn}_{j_1 j_2} \leq 1 \end{aligned} \tag{3.17}$$

The two constraints in the equation mandate that no two part-pair of the same kind (e.g. a pair of feet) share a node.

For multi-person human body estimations, calculating $\mathbf{Z}$ is still an NP-Hard K-dimensional matching problem. In the OpenPose algorithm, there are two relaxations customised to human body pose estimation. To construct a spanning tree skeleton, a minimum number of edges are used, as shown in figure 3.4c, which is quicker than optimising the fully connected graph. The matching is decomposed into a collection of bipartite matching subproblems, and the matching in neighbouring tree nodes is conducted separately, as in figure 3.4d. PAFs enables the system to model the link between adjacent tree nodes very well. Likewise, the CNNS trained with a broad receptive field have been able to consider the relationship between nonadjacent tree nodes as the predicted PAF is impacted by the PAFs from the nonadjacent tree nodes. Finally, the simplified formulation has been shown in equation 3.18.

$$\max_{\mathbf{Z}} E = \sum_{c=1}^{C} \max_{\mathbf{Z}_c} E_c \tag{3.18}$$

3.3.2 Machine Learning Models

Methods of dystonia prediction using machine learning are modelled as two types of problems: classification and regression tasks. Classification refers to predictive Modelling where a class label is predicted out of multiple categorical classes, i.e. discrete values, for a given example of input data. In contrast, if the outcome is continuous, the problem is modelled with regression, where ML tasks try to estimate the value of continuous results. On the other hand, when there are no accurate labels in the dataset, techniques such as clustering are generally used to group similar observations. Moreover, this falls under the category of unsupervised learning.

Regression

Since, in the case of the GDS scale, the output classes represent the degree of severity, a regression can be used to predict the severity. Due to its capability of providing outputs in absolute numbers, it has more expressive power than classification and can explain cases like when the prediction is between two levels. The linear regression model, say with p dependent variables, can be represented as the most straightforward regression technique as shown in 3.19.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \quad = \beta_0 + \boldsymbol{x}^T \boldsymbol{\beta} \tag{3.19}$$

where, $\hat{y}$ is the model prediction,

Here, $\boldsymbol{x} = [x_1, x_2, ..., x_p]$ are the feature values, $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_p]$ are the model weightings of each feature, and β_0 is the model intercept. The weights and intercept of the model are easily optimised by minimising the sum of the squares between the estimated response of the model, $\hat{y}$, and the true response, y as in the equation 3.20. The number of observations is assumed to be greater than the number of features to be solved algebraically.

$$\arg\min_{\boldsymbol{\beta}, \boldsymbol{\beta}_o} \sum_{i=1}^{N} \left(\beta_0 + \boldsymbol{x}^T \boldsymbol{\beta} - y_i\right) \tag{3.20}$$

Classification

In classification tasks, input data may be continuous or discrete, but the output must be one of all the possible classes. Classification problems can also be subdivided into binary and multiclass classification. In the context of dystonia score prediction, the binary classification problem, also called Logistic Regression, involves the detection of pathological motion, say in the GDS rating scale. To determine whether, above a threshold, a rating is considered pathological, and a rating below is considered normal, each rating can be binarised with such a thresholding technique.

For the Logistic Regression, the β coefficients representing model weights are determined by maximising the Log-Likelihood estimation of the training data. The mathematical formulation can be simplified to the form as shown in the equation 3.21 and can be solved by finding the *beta*, leading to the first derivative being zero3.22 . However, in practice, due to complications in doing so, approximations are made to find numerical solutions.

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{N} y_i \left(\beta_0 + \boldsymbol{x}^T \boldsymbol{\beta}\right) - \ln\left(1 + \beta_0 + \boldsymbol{x}^T \boldsymbol{\beta}\right) \tag{3.21}$$

$$\frac{\delta \mathbf{L}(\boldsymbol{\beta})}{\delta \boldsymbol{\beta}} = \sum_{i=1}^{N} \boldsymbol{x}_i \left(y_i - p\left(\boldsymbol{x}_i\right)\right) = 0 \tag{3.22}$$

Due to the nature of the GDS scale, multiclass classification problems can be naturally applied where we identify the severity ratings between 0 to 10. However, data with enough examples for each class must be present for effective learning, which limits the usability of such use-case.

3.3.3 Deep Neural Networks

Deep neural networks are the foundations of deep learning, often referred to as deep forward networks or multi-layer perceptions. Extended from artificial neural networks, where there are only a few hidden layers present in the network, the use of several hidden in deep neural networks has shown that the network is highly effective to map inputs

and outputs. Feature vectors are input to traditional deep neural networks, and when they pass through the first hidden layer, these features are converted to a more abstract representation by using nonlinear activation functions. The use of several hidden layers has shown that the network is highly flexible to map inputs and outputs.

Convolutional Neural Networks

CNN makes it possible to give raw inputs to neural networks without manually defining and extracting the features, as opposed to most other machine learning models. In fully connected DNNs, the traditional matrix multiplication procedure requires interaction between all inputs and outputs. However, by using fewer parameter representations, CNNs can avoid this inefficient process. A *kernel* whose dimensions are smaller than that of the input is used to provide sparse parameter representations such that each input unit only interacts with units of the same size as the kernel, enabling CNN to learn considerably fewer parameters. Using the same kernel for all input elements provides very efficient parameter storage by sharing parameters. This gives *translational invariability*: one of CNN's most potent properties. Consequently, a kernel trained to identify a particular aspect of input is suitable also to be used in an input where the particular aspect appears at a different place. *Pooling* is the widely used concept in CNN architecture that combines several outputs from a convolutional layer into a single output and also contributes to the translational invariance of convolutional weights. During training, pooling also reduces the number of parameters to be learned. It is performed by using one of several pooling functions available, including max pooling, average weight pooling, or mean pooling.

3.4 System Block Diagram

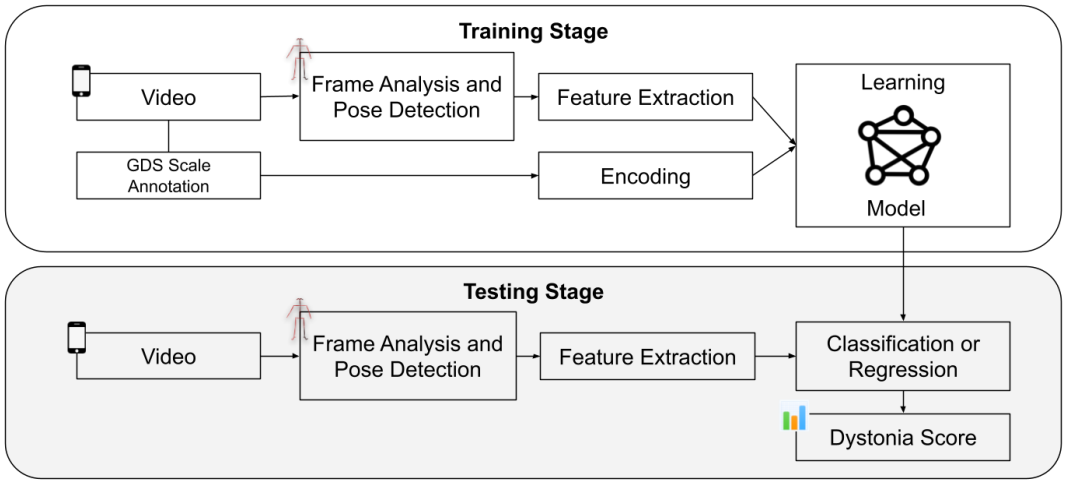

Figure 3.5: System Block Diagram

Figure 3.5 shows the system block diagram of the system. The workings of the different stages are explained in section 3.6.

From a broader perspective, the system involves two distinct stages, as shown in fig 3.5. The first stage involves training a network to learn to predict dystonia scores from the features that are obtained from the videos of the patients as the input. The raw videos segments are annotated, and the segments of interest are segmented. Such segments go through a pose estimation algorithm that annotates the human body poses in the video. Such annotations are not free from errors that need careful processing and are transformed into temporal features. The features are processed to get a standard form with their corresponding GDRS scores assigned to the patients during their evaluation. The dataset is used to train the network to be able to mimic the scores given by the domain experts.

The second stage is the testing stage, where the model is given the put of features obtained after processing the video. Such videos also could be unscored videos taken from a camera or one of the videos from the dataset. The input is processed by the system to output a score.

3.5 Description of Algorithms

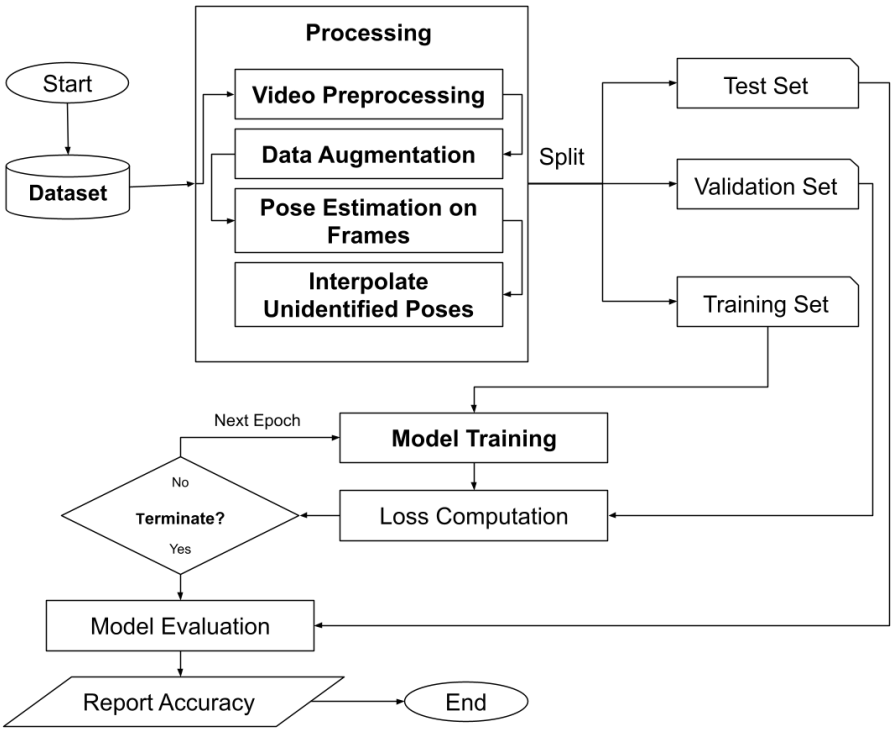

Figure 3.6: Flow Chart of Model Training

3.6 Elaboration of Working Principle

The proposed method is based on machine learning. Therefore, the system involves a training phase and a prediction phase. The output of pose estimation algorithms like OpenPose can be dumped into a JSON file composed of a set of key points. Figure 3.7 shows pose detection from 18 and 25-keypoint variants of OpenPose. A 25-keypoint variant is more accurate than 18 and is the default option for OpenPose body detection. For each video frame, OpenPose returns 25 body points in 2D image plane coordinates along with each detection's prediction confidence. The reported points are the estimated coordinates (x,y) pixels of the body key points.

Figure 3.7: Comparing poses from 18-keypoint COCO and 25-keypoint BODY_25 OpenPose varients.

Video segments where multiple persons are visible should be excluded as this adds complexity to person tracking. For each video, PoseNet is run to get the time series of 25-keypoint across all frames. Each set of coordinates in the time series is centred by subtracting the coordinates of the right hip in all videos. Scaling is done for uniformity by dividing by the Euclidean distance between the right hip and the right shoulder. Linear interpolation is used to fill the missing values in the timeframe as all images of the human body are not guaranteed to be detected ideally across all frames, and machine learning algorithms do not work with missing data.

In our case, the CNN model that used is a parameterised mapping of data from fixed-length time series (pose keypoints) to a result metric. Even without extensive feature engineering, CNNs perform well over conventional machine learning models. Even after being convinced with the architecture, a grid search over hyperparameters can help to tune the performance. Two popular loss functions: mean squared error (regression task) or cross-entropy (classification tasks), can be used to train the neural network models. After training a deep learning model in the testing stage to determine the dystonia score from input video clips, this model runs continuously through sliding window procedures on the video timeframe.

3.7 Instrumentation Requirements

The setup for the training stage is the machine learning workstation with a generic configuration. The scoring model needs to be trained in a GPU-enabled machine and exported such that it can be loaded and run without the presence of sophisticated setups and hardware, including a GPU. Since the dataset that used by the project already contains the video and by virtue of the project scope, a specialised camera for video recording purposes is not needed. However, for generic testing, videos taken from a generic smartphone can also be used. The trained model needs to be hosted in a mid-spec server that is capable of running inference over the trained dystonia scoring model with the demonstratable performance and low latency.

3.8 Dataset Explanation

Since this project tries to address a specific problem in the medical domain, the problem with data availability is the common problem these types of research face. However, due to the increasing usage of computer vision in the medical domain, there has been increasing interest in using pose detection algorithms to solve challenging problems. However, the video data used in the study are not available to the public due to laws that put restrictions on the dissemination of patient health information.

3.8.1 Dystonia Dataset

The data from which this whole project is inspired contains videos of dystonia patients recorded over the course of the past ten years.

Those videos are taken by medical personnel while assessing their dystonia patients. Each video is unique to a patient and is of their last visit to the hospital. Out of 453 cases, 100 videos are of generalised cases, 100 are of neck examinations, 100 are of upper face examinations, 24 are of upper leg/foot examinations, and 29 are of lower face examinations.

Along with the video recordings, patient details (including exam, onset, and tremor), medical history of the patient, BFM examination results as well as GDRS examination results are provided for each case. Since, in this project, the GDRS score is only of interest, it has only be used, and the rest has been ignored.

3.8.2 VideoGait-V1

This dataset was used in [22] and contained the body keypoint trajectories. As the video data cannot be shared publicly, these data are anonymised by OpenPose, meaning only pose data is available, including their corresponding labels that contain surgical decision, speed, cadence, flexion of the knee at max extension, and GDI, which are either annotated directly by trained medical personnel or obtained from optical motion capture. This dataset was used to implement and understand the methods in the paper, and the implementation helped a lot in understanding the flow they used.

3.9 3D Pose Estimation

The 3D human body pose estimation is a complex method. Facebook's VideoPose3D provided a comprehensive way to estimate 3D human body pose from a video captured with a single camera. Our videos explicitly needed to be processed through the following steps to get the 3D results, as shown in figure 4.9.

3.9.1 Video Pre-processing

For the tests, the video segment of various actions was cropped using the FFmpeg library and interpolated to 50 FPS using the minterpolate algorithm as well as NVIDIA's Optical Flow, and the results were compared visually.

3.9.2 Detectron for 2D keypoints estimation

As required by the flow of VideoPose3D, Facebook's Detectron2 was used to get 2D human pose estimations first. A function could export the result to a custom NumPy archive (.npz format).

3.9.3 Creating a Custom Dataset

Those 2D pose sequences information in the NumPy archives (.npz format) exported after using Detectron on the videos were used to create a custom standard dataset in the format that the library used using the provided function.

3.9.4 3D Inference and rendering

The custom dataset was fed to the 3D human body pose estimation function of VideoPose3D to get the 3D output inferred from the input. The library also provided a way to render the 3d video with MatplotLib library, which could be visualised easily.

3.10 Graph Based Methods

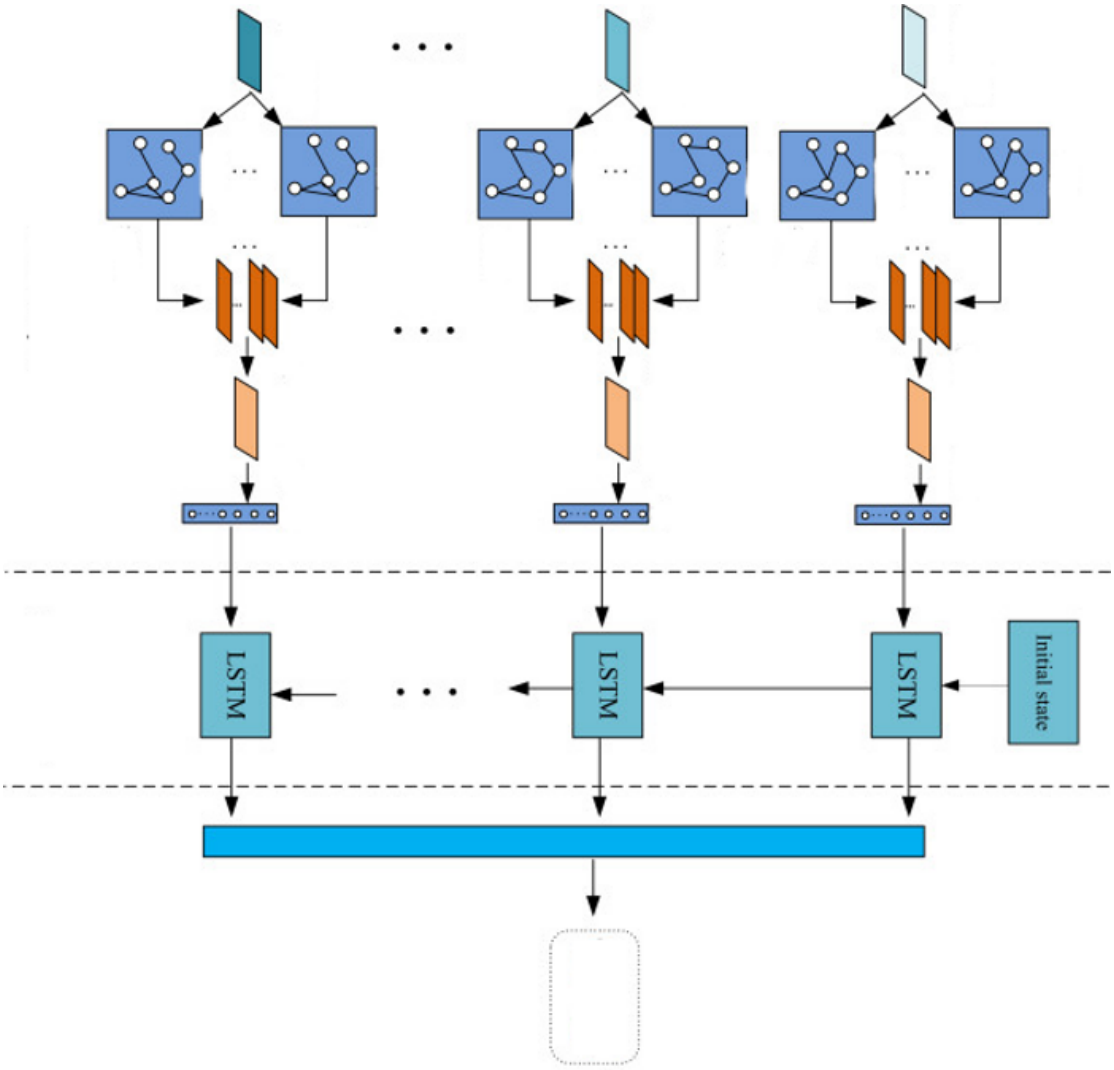

Figure 3.8: GC-LSTM on Human Body Poses

It is really intuitive to imagine the key points of the human body poses as nodes and the link between those nodes as edges of a graph, making it a perfect candidate for graph-based learning.

Each of the frames in the pose sequences is first passed through a Graph Convolution (GC) layer as shown in figure 3.8 which can model the structures in the pose representation. The output from the graph convolution is then fed to an LSTM layer which learns the temporal changes making it able to classify based on both the spatial and temporal information from the data. Each data in the dataset consists of input pose sequences of fixed length which represents a video segment considered.

3.11 Verification and Validation Procedures

The dataset is splitef into training, validation, and test sets. Regression tasks can be evaluated by determining the correlation between the predicted values and actual values of the test set for each model.

There is a clear boundary between validation and verification in software testing. However, verification and validation have their own meaning in the machine learning regime. Verification is the test of whether the model meets the specified mathematical description; the tests related to this are performed during each training step. Likewise, validation determines whether the model is accurately responding to the actual inputs or real-world applications, often termed cross-validation.

The standard metric used in the regression model is the Mean Absolute Error (MAE), which calculates the average magnitude of the error between the prediction and the actual value as in the equation 3.23. They can both calculate error rates as part of the verification stage and calculate model performance as part of the validation stage.

$$\text{MAE} = \frac{\sum_{i=1}^{N} |\hat{y}_i - y_i|}{N} \tag{3.23}$$

Another popular metric for regression is the Root Average Square Error (RMSE), which functions on the squared error between prediction and actual truth and is calculated as in the equation 3.24.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{N}} \tag{3.24}$$

The regression model metric uses the continuous model prediction, while the classification evaluation metric uses discrete model prediction. The popular verification metric for classification is log-loss. Some validation metrics to be used for classification are discussed below:

3.11.1 Cohen's kappa

Quadratic-weighted Cohen's kappa is a metric for the multiclass classification task that measures inter-annotator agreement on a classification problem. Equation 3.25mathematically represents Cohen's Kappa metric, with p_o representing the empirical probability of

consistency with the label assigned to any of the samples. p_e is the expected agreement when two annotations assign labels randomly and are estimated with a per-annotator before the class labels.

$$\kappa = (p_o - p_e)/(1 - p_e) \tag{3.25}$$

3.11.2 F-Score

This score uses the harmonic mean of *precision* and *recall*, where precision is the number of correct score classes retrieved by a search divided by the total number of score classes retrieved, and recall is the number of correct scores retrieved by a search divided by the total number of existing correct scores in the test set.

$$Precision = \frac{TP}{(TP + FP)} \tag{3.26}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{3.27}$$

$$F - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3.28}$$

The total number of true positives is TP, the total number of false positives is FP, and the total number of false negatives is FN.

4 RESULTS

The project concentrates on building a machine learning model capable of predicting dystonia scores from input videos of a patient. However, before building a machine learning model, the dataset's exploration, profiling, and annotation are necessary. Since the project has just completed the data annotation stage, the model training remains. The following things have been accomplished until now:

4.1 Interface

A web-based interface is built to interact with the data available. The interface helps to explore the data and can facilitate the segment tagging process.

After the model is hosted on an inference server, if necessary an option can be easily added on this interface to interact with the model for the end-user or demonstration purpose. The general pipeline to interact with the system is shown in the figure 4.2.

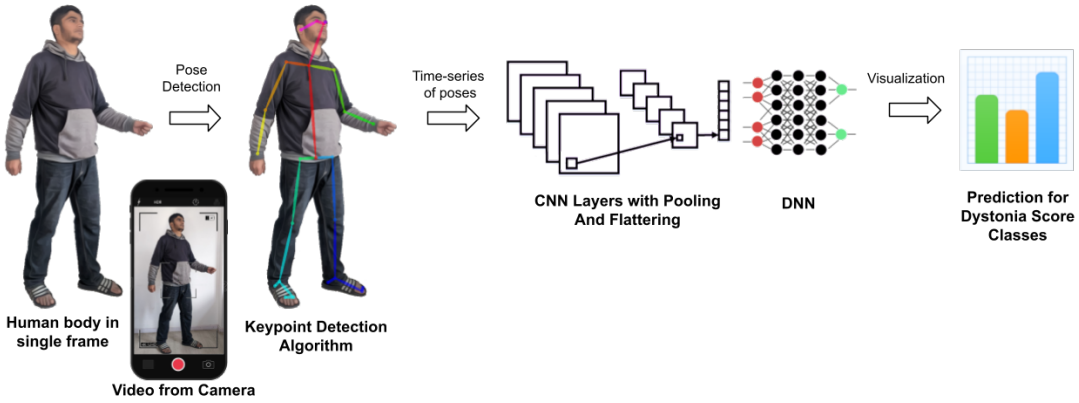

Figure 4.1: Proposed score prediction from the input pose sequences.

The Pose sequences extracted from a patient's video has been used to predict the dystonia score as shown in the figure 4.1 in the final system.

4.1.1 Web Tagging Tool

The interface has been built to be easily extended to tag video segments. Annotator can utilize the shortcut keys provided to tag annotate the start and end of the particular video segment. Since the video size is large, it is not easy to play them on Web Browsers. The tool is also assisted by a Video-on-demand service that can stream the video in the interface. Such a feature can be helpful even when multiple annotators work

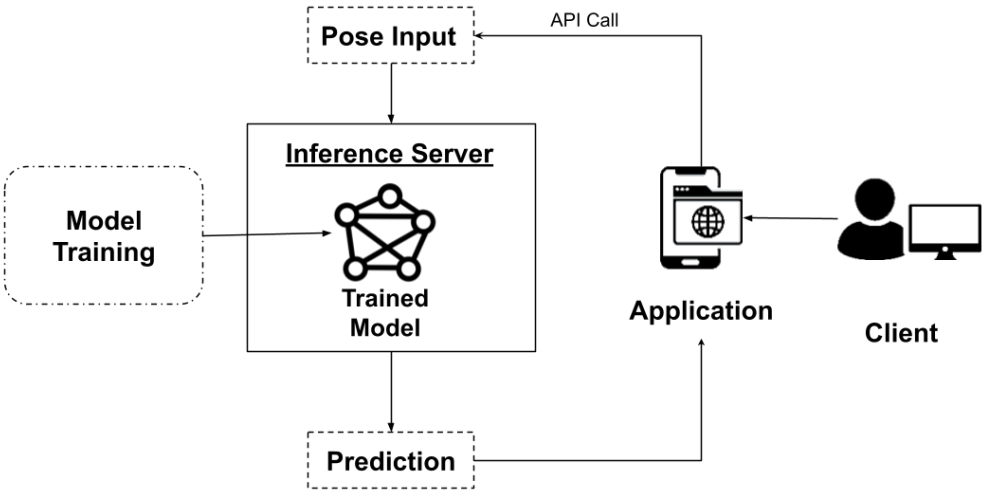

Figure 4.2: Expected Interface with ML model serving

simultaneously. The screenshot of the tool is shown in the figure 4.3.

4.1.2 OpenPose JSON Visualization

The tool also facilitated the visualization of JSON outputs predicted from the OponPose method. The key point predictions from OpenPose are overlayed on the streamable videos. The output is refreshed when the video is played to match the correct position. There is also an offset on the key points so that frame in the video and predicted key points could be compared side by side.

4.2 Selection of GDRD Score for Modeling

4.2.1 Full Body Assessment Segment

GDRS assessment contains numerous examinations and has multiple positions that patient has to be for the examination. The GDRS guidelines mention all these things. However, inspired by the previous research and by the available human body pose estimation algorithms, the segment with full-body visibility was initially chosen for the project. However, this idea was dropped due to multiple reasons:

- The videos were primarily taken in hallways where the patient could walk as per the assessment requirements. Most of the time, some people were walking in the hallway, and OpenPose could not inherently track the target person through changing scenes.

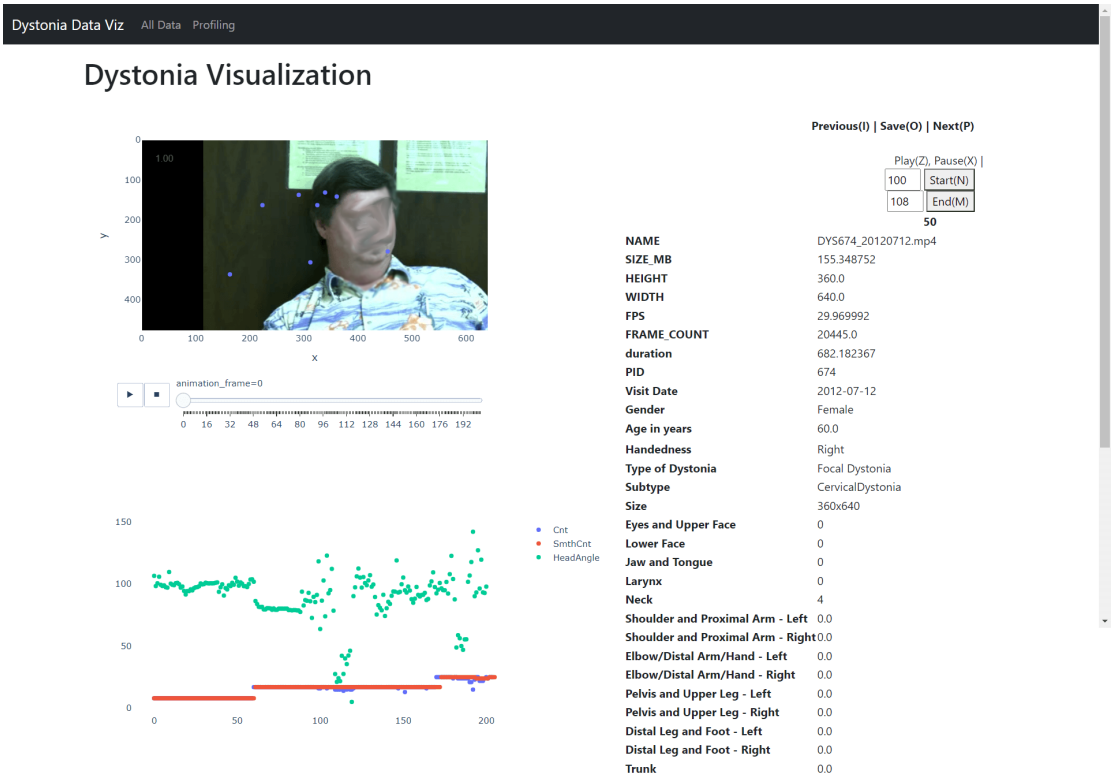

Figure 4.3: Screenshot of the Annotation Interface.

- The full body assessment mainly corresponded to the pelvis and leg part of the assessment. The score distribution as in figure 4.4clearly showed that there was an uneven distribution of the scores on such assessments on the dataset.
- Many videos were such that the annotator moved the camera with the patient as the patient walked.

To handle frames where multiple human bodies were detected, a simple queue was implemented. The coordinate of the few latest neck regions was continuously updated to the queue. Whenever multiple persons were detected, the one with its neck nearest to the coordinates in the queue was chosen. The verification of the working of this logic was manually checked on the videos. Although multi-person cases were handled perfectly with this simple idea, the idea of using video segments with full-body segments was discarded.

4.2.2 Neck Assessment Segment

Further analysis of the score distribution revealed that the data for neck assessment was the one with the most favourable distribution for the project. The most balanced class

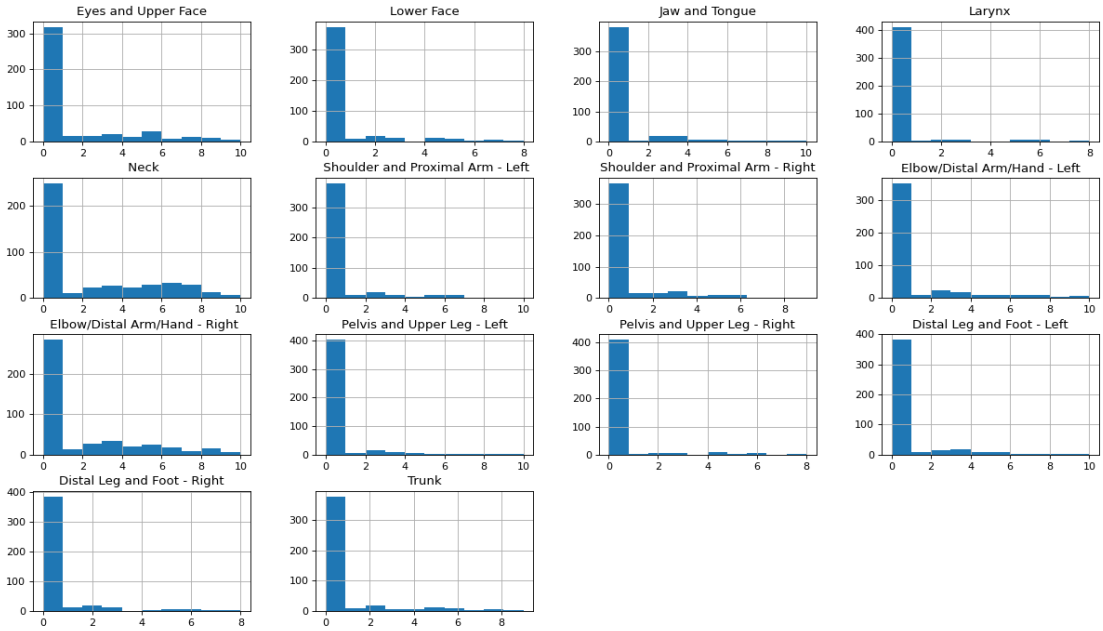

Figure 4.4: Distribution of the GDRS scores.

over the score range can be clearly seen in figure 4.5. A brief examination of the video segment of the neck assessment also revealed that it would favour the requirements. As per the GDRS protocol of neck assessment, the patient was to be seated in a chair facing directly towards the camera with the upper part of the body visible.

Although there were multiple assessments on the neck examination, it was decided to capture only the segments containing the three specific assessments. The patients needed to follow the following actions:

- First, they need to look to the far right and then to the left after the instruction from the examiner.
- Next, they should try their best to bring their right ear to the right shoulder. The same was repeated for the left side.
- Finally, they were asked to look up to the ceiling, come back to normal position and look down to the ground.

It was found that most of the time, the videos followed the order too. However, there were minor cases where orders were swapped by the examiners.

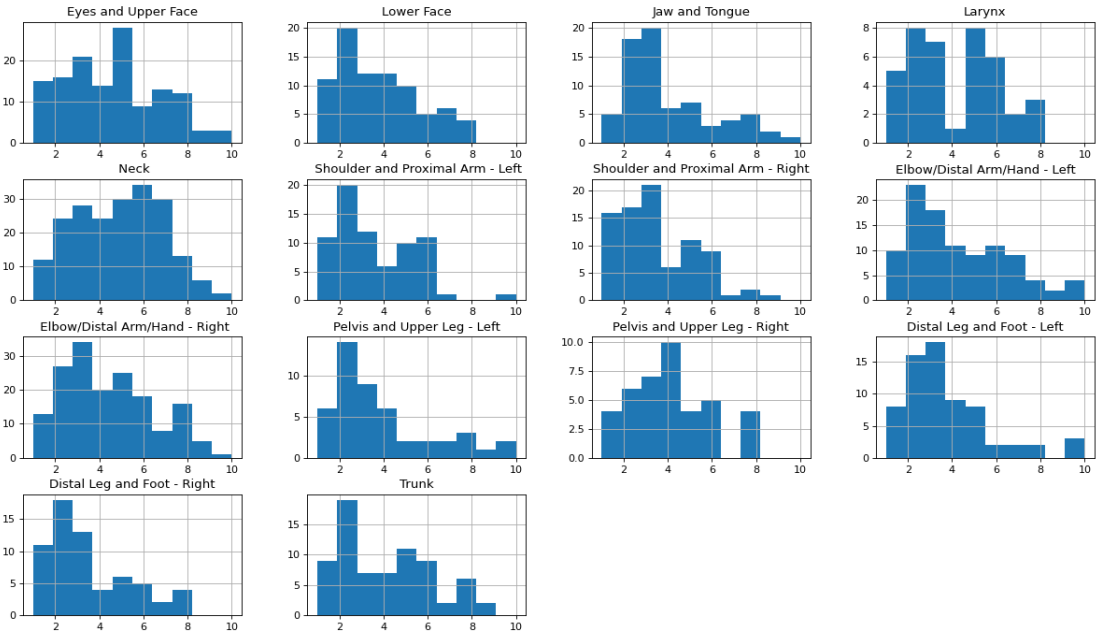

Figure 4.5: Distribution of the GDRS scores with Zero scores excluded.

4.3 Correlations between related GDRS Scores

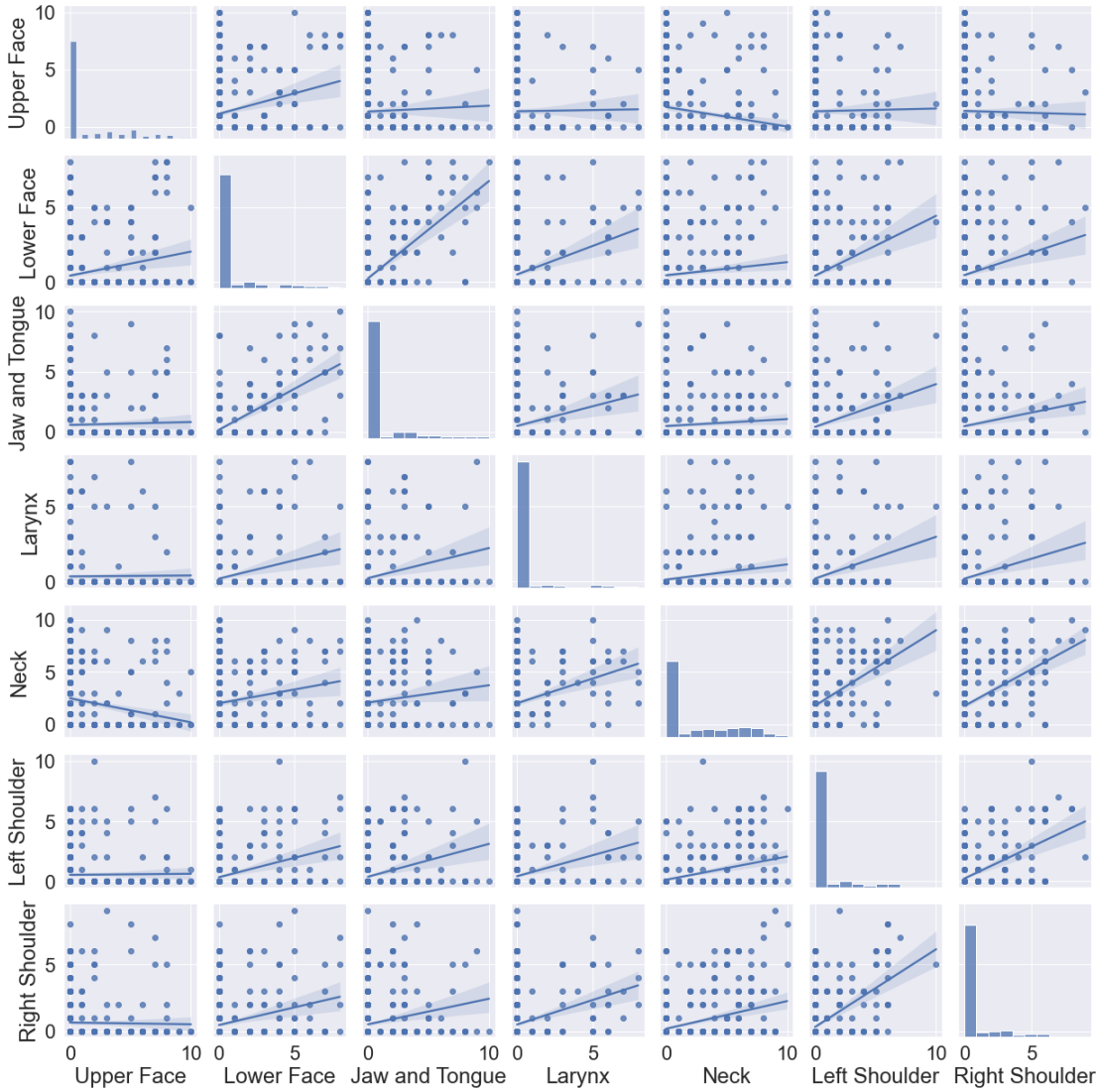

Figure 4.6: Correlations between related GDRS Scores.

The pairwise relationships analysis between GDRS scores from the assessment of the head and neck region was performed. The figure 4.6 shows a pairwise plot that depicts the relationship between six of the GDRS scores from the assessment of face, jaw/tongue, larynx, neck and shoulder. The plot shows that the scores for the Upper face (including the eye as well in the protocol) are the most uncorrelated out of the six scores analyzed. The score represents eye closure with squeezing, the cases with unable to open eyes within 10 seconds and intense forehead wrinkling. However, lower face soreness representing intense grimacing with extreme distortion of the mouth seems to have a correlation with other scores considered.

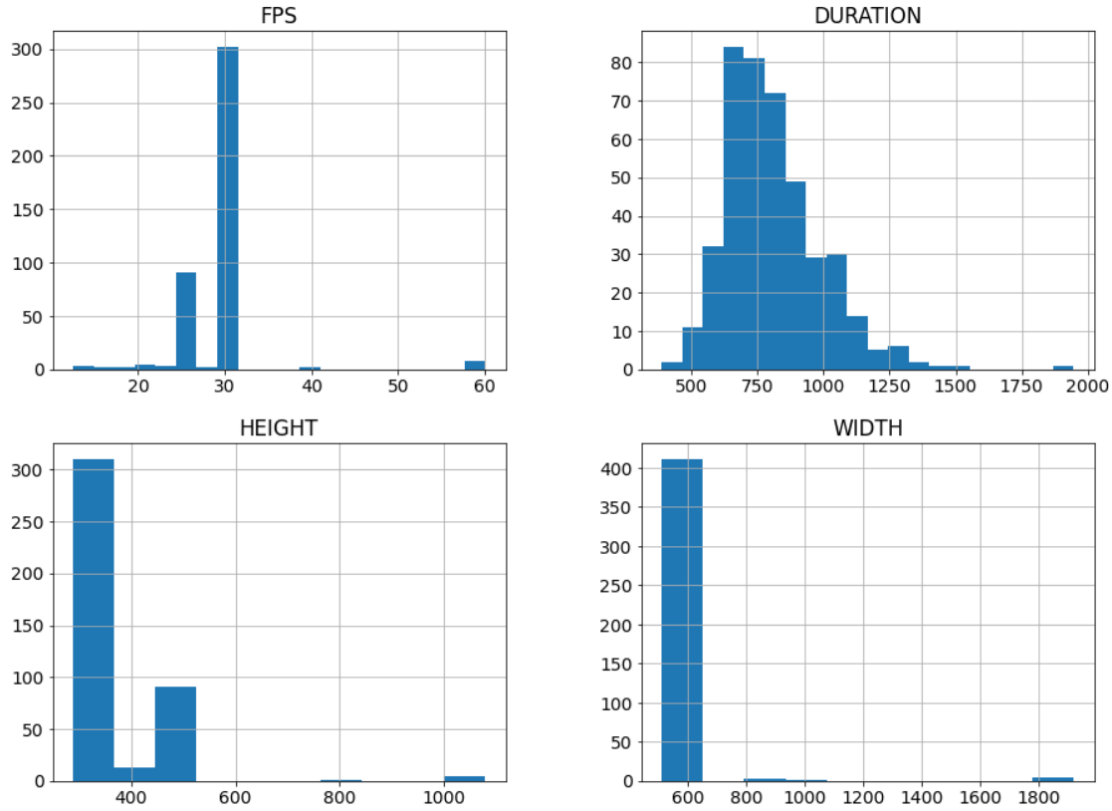

Figure 4.7: Histograms of FPS, duration in seconds, height, and width of the videos available.

The neck score showed a lower correlation comparatively with others in the set. Shoulders (and upper arms, distal arm and hands as well in the protocol) and neck scores had a good correlation depicting that people with the problem at the shoulders were more likely to get problems in the neck. Also, the left and right shoulders had strong correlations among themselves.

4.4 Applying OpenPose on the Videos

OpenPose worked on each frame of the video. Due to the fact that the videos were captured by different persons and with different devices, the FPS of the videos was different on each video. Also, the duration of the video ranged from eight minutes to thirty minutes. Such variability could have occurred due to multiple reasons which are out of the scope of the project. The variation in video sizes and FPS adheres to the availability of the video recording equipment at the time the video was taken. The histogram of FPS, duration in seconds, height, and width of the videos available are shown in figure 4.7.

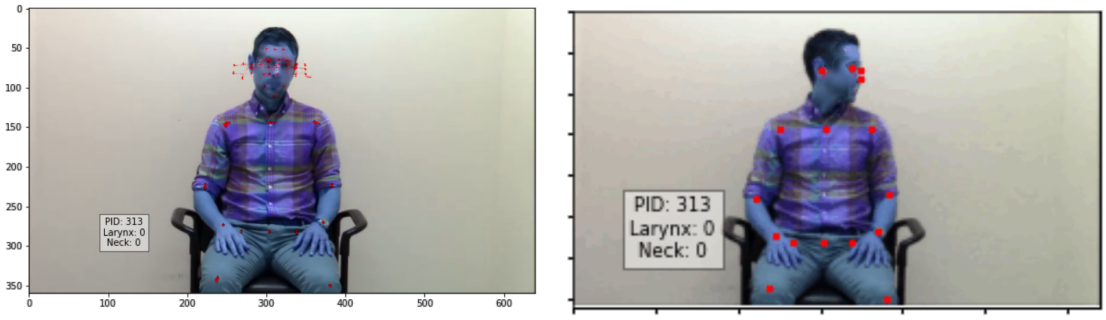

Figure 4.8: Left: All processed keypoint in a certain segment overlaid over patient's photo. Right: Side by side visualization of animated frames with overlaid poses at that particular frame.

Since using OpenPose on all videos was time-consuming and could take up to a couple of days, it was required to mark the segment of our interest so that we could run it efficiently. OpenPose Command-line API has options to specify the start frame, end frame, and the number of frames to skip. Frameskip attribute was set to 100, and OpenPose was run on the video, which was used to overlay over videos on the annotation interface. For the dataset creation stage, the frameskip value was discarded, and full frames were captured between the start and end frame number of the region of interest. Table 4.1 shows the types of keypoints predicted by the OpenPose, with the connection of each of them to the other part.

Table 4.1: List of OpenPose Keypoints used with their short-form and connection.

Index	**Keypoint**	**Full Name**	**Connected To**
0	Nose	Nose	1, 15, 16
1	Neck	Neck	0, 2 , 5, 8
2	RShoulder	Right Shoulder	1, 3
3	RElbow	Right Elbow	2, 4
4	RWrist	Right Wrist	3
5	LShoulder	Light Shoulder	1, 6
6	LElbow	Left Elbow	5, 7
7	LWrist	Left Wrist	6
8	MidHip	Mid Hip	9, 12
9	RHip	Right Hip	8, 10
10	RKnee	Right Knee	9, 11
11	RAnkle	Right Ankle	10, 22, 24
12	LHip	Left Hip	8, 13
13	LKnee	Left Knee	12, 14
14	LAnkle	Left Ankle	13, 19, 21
15	REye	Right Eye	0, 17
16	LEye	Left Eye	0, 18
17	REar	Right Ear	15
18	LEar	Left Ear	16
19	LBigToe	Left Big Toe	14, 20
20	LSmallToe	Left Small Toe	19
21	LHeel	Left Heel	14
22	RBigToe	Right Big Toe	23, 24
23	RSmallToe	Right Small Toe	22
24	RHeel	Right Heel	11

4.5 Video Normalization

One problem that was immediately realized while using OpenPose was that there was a number: FPS that was directly affecting the output size and processing time. Since there were videos with FPS variations, as shown in figure 4.7, it was needed to normalize the FPS settings. This decision was backed up by the fact that this would directly affect the temporal frame size that would be used in the model input.

On the other hand, the video size also was uneven, and it would be better if they could be normalized too. It was decided to normalize the video to 640x360. This particular choice was due to the fact that most of the videos were already around this size, as seen on the histogram in the figure. The videos were also re-encoded to Advanced Video Coding (AVC/H.264) format using Nvidia Encoder(NVENC) engine.

NVIDIA optical flow was used to normalize the frame rate to 30 FPS. NVIDIA optical flow SDK allows applications to use NVIDIA's optical flow engine. Calculated optical flows were used to increase frame rates. Frame rate up-conversion techniques (FRUC) are used to insert interpolated frames between the original frames. The interpolation algorithm generates intermediate frames using the optical flow between frames producing smooth transitions. FPS conversion was also inspired by the idea to examine 3D Video Pose Methods. Human3.6M model for Facebook's VideoPose3D, which is openly available, was trained on 50 FPS videos, and it was recommended to use 50 FPS videos for better prediction. However, using VideoPose3D turned out to be a bad idea which is explained in the discussion section.

4.6 3D Pose Estimation

The frame on the bottom left in figure 4.9 has a patient whose half of the body is only visible. The 3D output tried to fit the available data to the full human body skeleton. Also, the output 2D keypoints overlaid on the original frame clearly show that the Detectron2 mode used is trying to do the same. There are no options to handle such cases. Likewise, the top-right frame in figure 4.9 shows a doctor assessing the patient. The 3D output seems to have been generated from the key points of both the doctor and the patient, which can be inferred from the Detectron2 outputs overlaid on the original video frame.

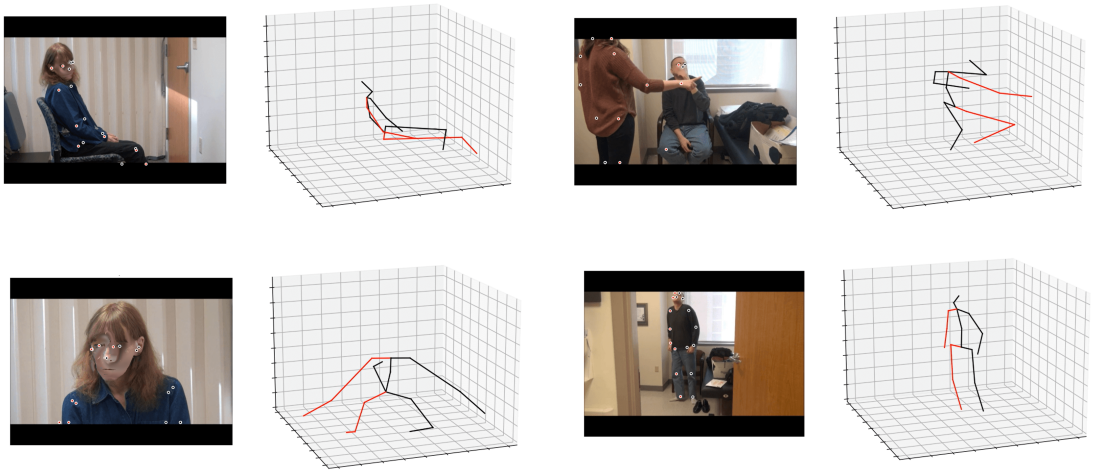

Figure 4.9: Four frames showing VideoPose3D Outputs. Original Video frame with an overlay of output from Detectron2 is shown on the left of each and the 3D prediction from VideoPose3D is shown on the right of each.

4.7 Using Graph Based Methods for Score Modelling

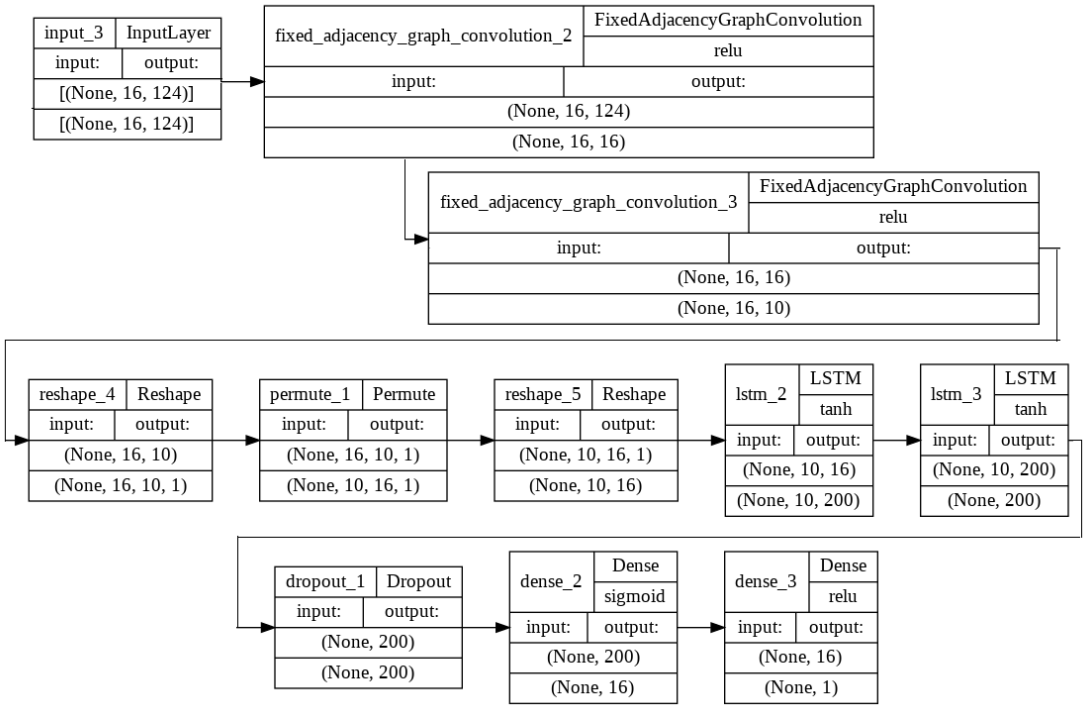

Figure 4.10: GC-LSTM used for Neck Dystonia Score prediction.

GC-LSTM network, as shown in 4.10 was trained to estimate neck dystonia score from the processed dataset. The dataset had pose locations extracted from around 400 videos, each normalized to 124 frames spatiotemporally from a variable-length segment that contained the neck assessment section. Thus, each input in the training set was a pair of

neck dystonia score (integer) and 124 frames with normalized 2D locations of 8 points of interest around the neck and head region (2*8, 124). As the number of training data was minimal, no validation was done. The initial objective was to see whether this network could model or overfit over the processes dataset with very little hope. The network failed to learn from the given dataset, and the loss fluctuated drastically within epochs. Slight variations with activation functions and layer number were done, but it showed no sign of improvements. The graphs and statistics have also not been included in this report as it was nothing interesting.

4.8 Multi-line plots of keypoint positions

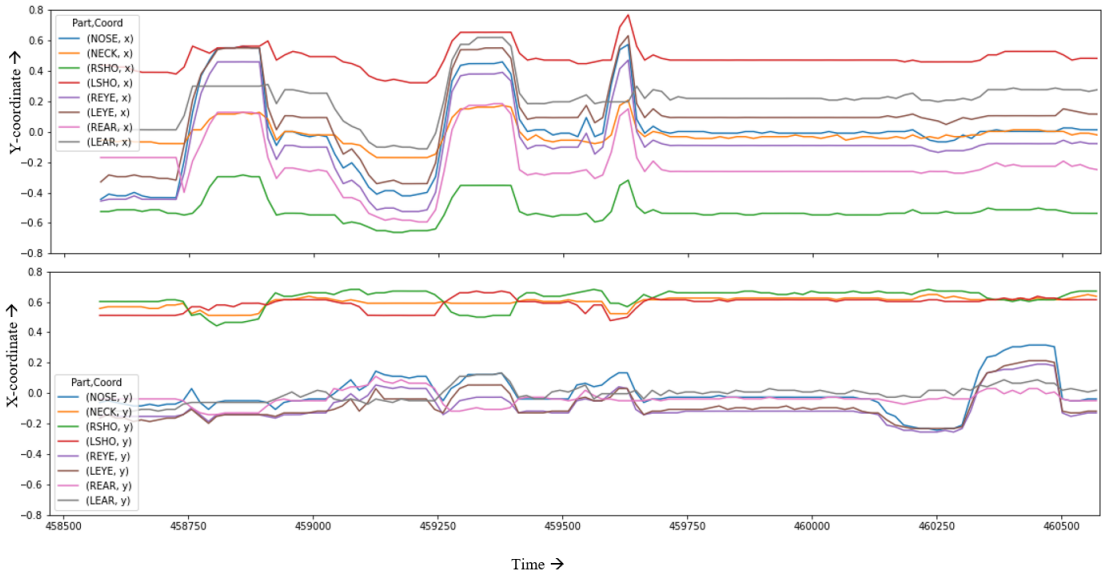

Figure 4.11: Multi-line plots of keypoint positions.

Figure 4.11 shows a multi-line plot representing temporal body pose data of eight body keypoint representing the head and neck region. The x-axis is the time domain, and the y-axis is the value on the particular axis the key point was. In this graph, the label on the x-axis is the original frame number of the video. This particular plot represents everything in a particular instance of data in both the spatial and temporal axis. The peaks and troughs represent extreme movement either in up/down or left/right direction. This plot was extremely helpful to visualize the data and the variations. Figure 4.12 shows multiplot with the spatial distribution of body points on the X-axis with the corresponding neck dystonia score represented by colours. The plot shows the absolute deviation in the pose coordinates within the video segment. A similar plot to visualize the same kind of variation in the Y-axis has also been presented in figure 5.14.

Figure 4.12: Spatial Distribution of body points in X-axis with the corresponding neck dystonia score represented by color.

5 DISCUSSION AND ANALYSIS

5.1 Model

The main motive of this project was to examine whether dystonia scores could be modelled with DL. However, the dataset for the project was not available while drafting the proposal. It turned out to be a very raw source and needed much effort in processing and tagging the segment of interest.

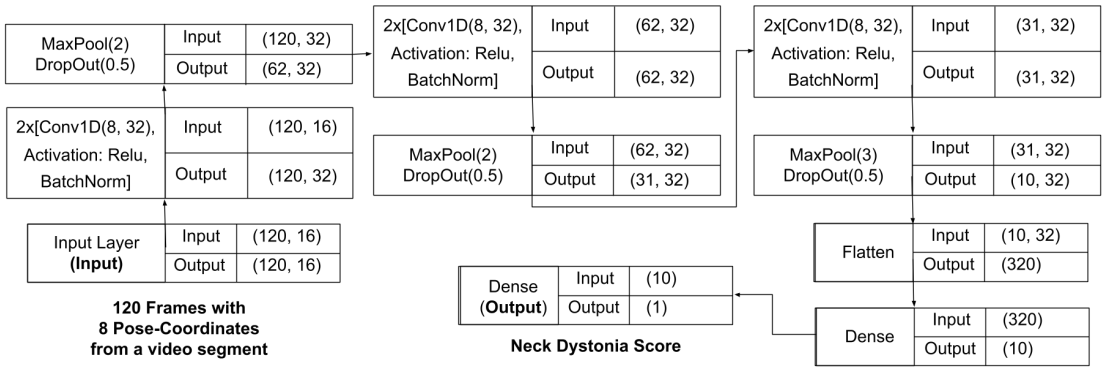

Figure 5.1: Block Diagram of Used CNN Architecture in the outcome.

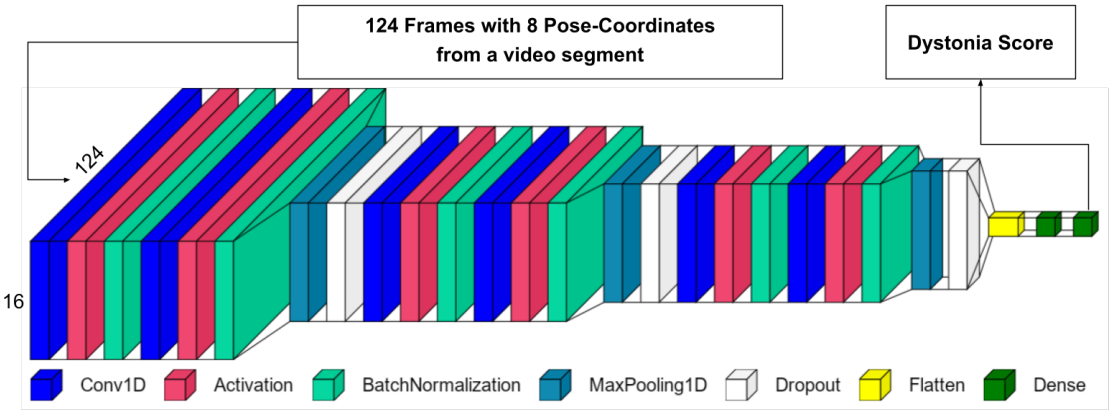

Figure 5.2: Layered View of CNN Architecture used.

Regarding the architecture, the CNN model used is the one shown in 5.1 and 5.2, with slight variations. This architecture has also been used in [22] for a similar task. The dimensionality of the number of output filters after each convolution was uniformly set to 32. The kernel size of each convolution block was set to 8, establishing the length of the 1D convolution window. ”Same” padding strategy was utilized with zeros uniformly distributed to the left/right or up/down of the input, resulting in output with the same height/width.

Root Mean Squared Propagation (RMSprop) was used with rho values of 0.9 and epsilon value of $1^{(}-8)$ with no decay parameter. To make the magnitude invariant to the number of components in the target, the mean square loss was modified by multiplying by the length of the weights. To avoid affecting the loss function, the weights were always normalized to a sum of one.

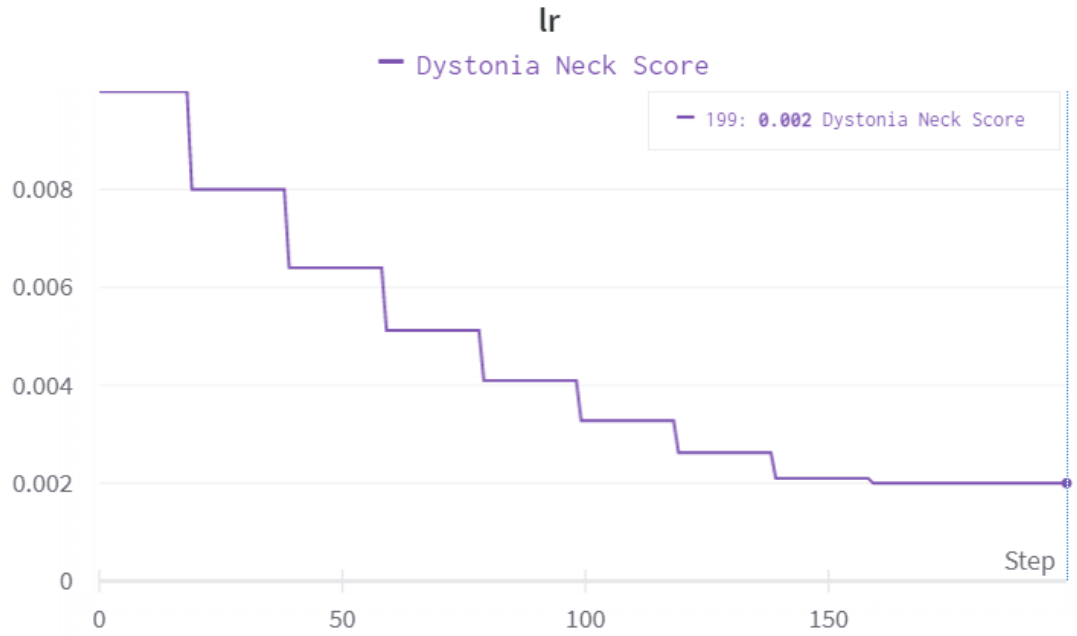

Figure 5.3: Learning rate variation with epoch on training a CNN with full data.

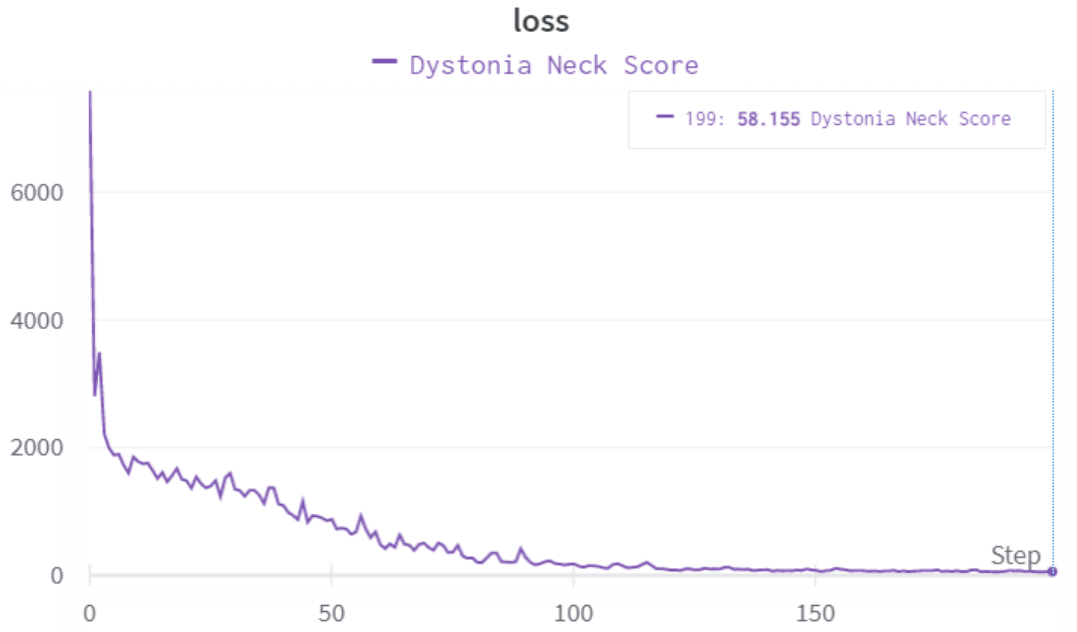

Figure 5.4: Training loss on training a CNN with full data.

Figure 5.3 represents the varying learning rate strategy used in training. For every epoch of 20, the rate was dropped to 0.8 from the initial learning rate of 0.01 and capped at 0.002. Figure 5.4 shows the decaying loss over the epoch.

5.1.1 Regularization Techniques

Some common regularization techniques were used for the regularization are discussed below:

Dropout: Dropout was added after each max-pooling operation after convolution. It was seen that it benefited the network to prevent over-fitting by reducing the quick dropping of loss. A 0.1 dropout rate was used, in which we randomly turned off part of a layer's weight at each training step by zeroing off the values.

L2 Regularization L2, commonly known as "weight decay," the most prevalent sort of regularization, was used with a value of $10^{-3.5}$. Using L2 regularization also helped train the network and had almost the same effect as the dropout layer. It fine-tuned the loss function by adding a penalty term that prevents the coefficients from fluctuating too much. As a result, the odds of over-fitting should reduce with such, however, due to a lack of data. Validation could not be performed to precisely measure how much these were beneficial.

Data Augmentation Although many techniques exist for augmenting images, it was different here. The input data with Spatio-temporal pose coordinates could not be much altered. However, horizon flipping was done on the coordinates by changing the sign of the x-axis coordinates. There is open space to find new techniques that could be explored to perform data augmentation on body pose data. It could facilitate the study of cases where it is not practical to get larger data sizes.

5.2 Model Evaluation

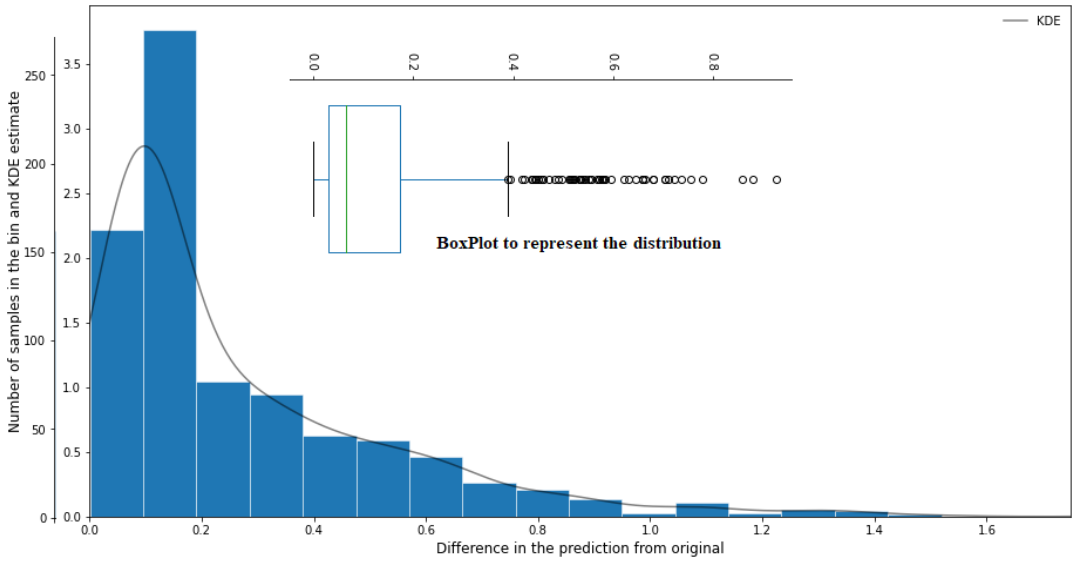

Figure 5.5: CNN model performance

Figure 5.5 shows the plot of the difference between the prediction of the model and the actual clinical score value of dystonia neck score data used in CNN training. Precision, Recall, and F1-Score are respectively 0.67, 0.61 and 0.63. The difference has been binned into twenty bins and plotted. A kernel density estimate (KDE) plot that uses a continuous probability density curve to describe the data has also been shown on the same plot that shows the Distribution of observed errors in a dataset. This indicated that CNN could model the parameter correctly, and the predicted output is very close to the actual clinical scores.

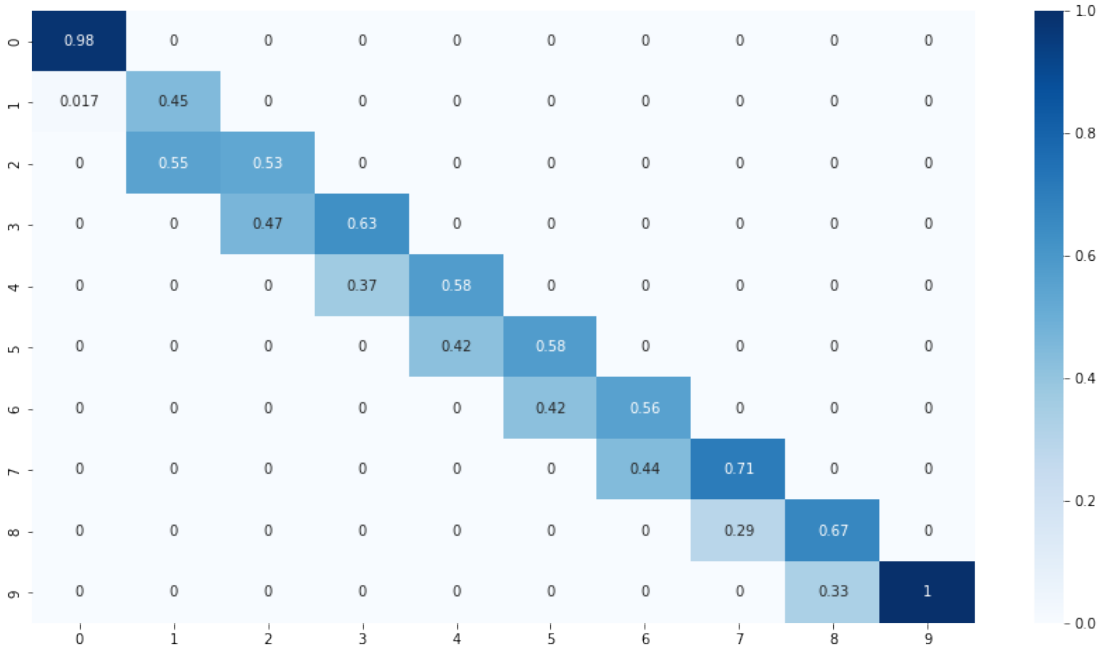

Figure 5.6: Confusion matrix showing results from CNN model per score class with Cohen's kappa $\kappa = 0.71$.

A confusion matrix is shown in figure 5.6 to summarize prediction outcomes in different neck dystonia score classes.

5.2.1 5-Fold Validation

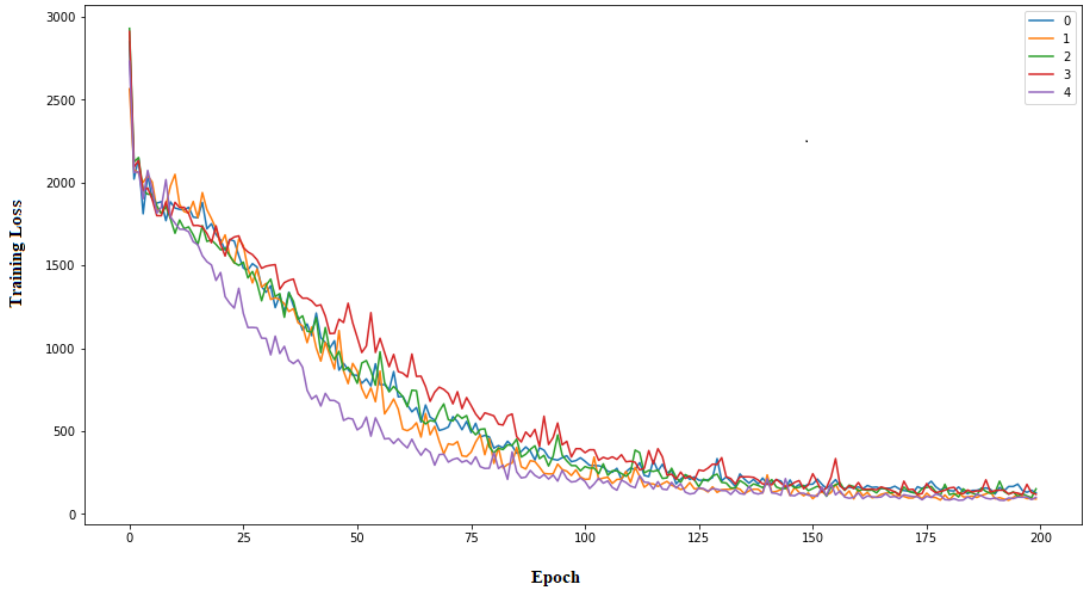

Figure 5.7: 5-fold validation on data.

Stratified 5-fold validation with a modified version of KFold validation has also been done on the data. The stratified folds are created by keeping track of the proportion of dystonia scores in each class. Figure 5.7 shows a plot of training losses of all five folds during over 200 epochs. It is seen that all the models almost converge around the same time.

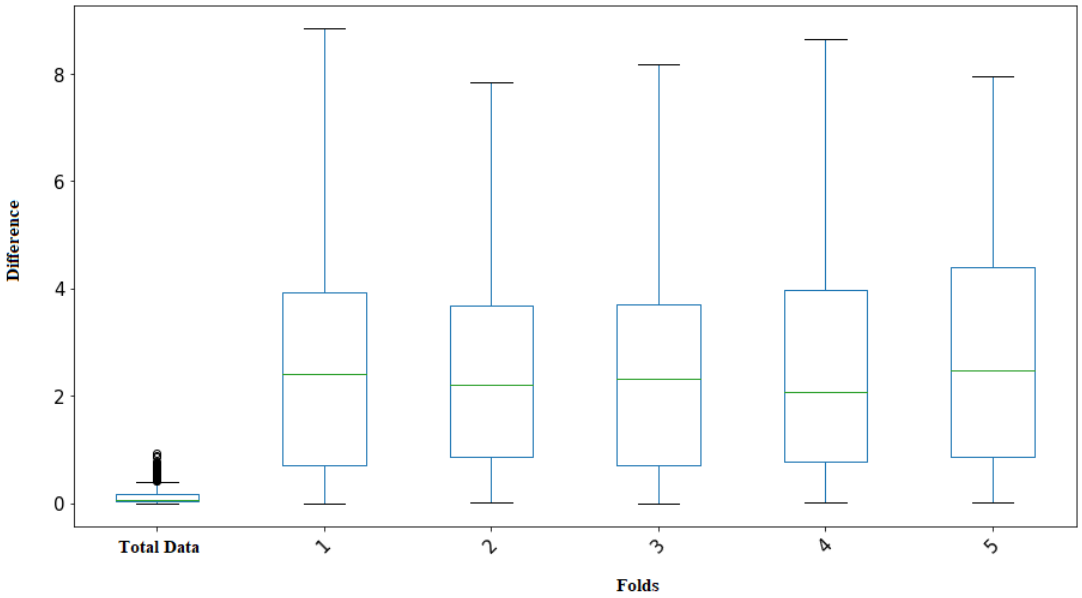

Figure 5.8: Box plot comparing difference in real and predicted values of 5 models from 5-fold training on data

Figure 5.8 compares the difference in the model after the 5-fold split. The first boxplot from the left is the model where all the data has been used for training, and the same data were used for prediction. The remaining boxplots represent the distributions of errors obtained from the difference between the actual clinical score and validation fold scores obtained on the model trained from each fold. The result has a vast difference, but this is expected because of the lower number of data in the critical score regime.

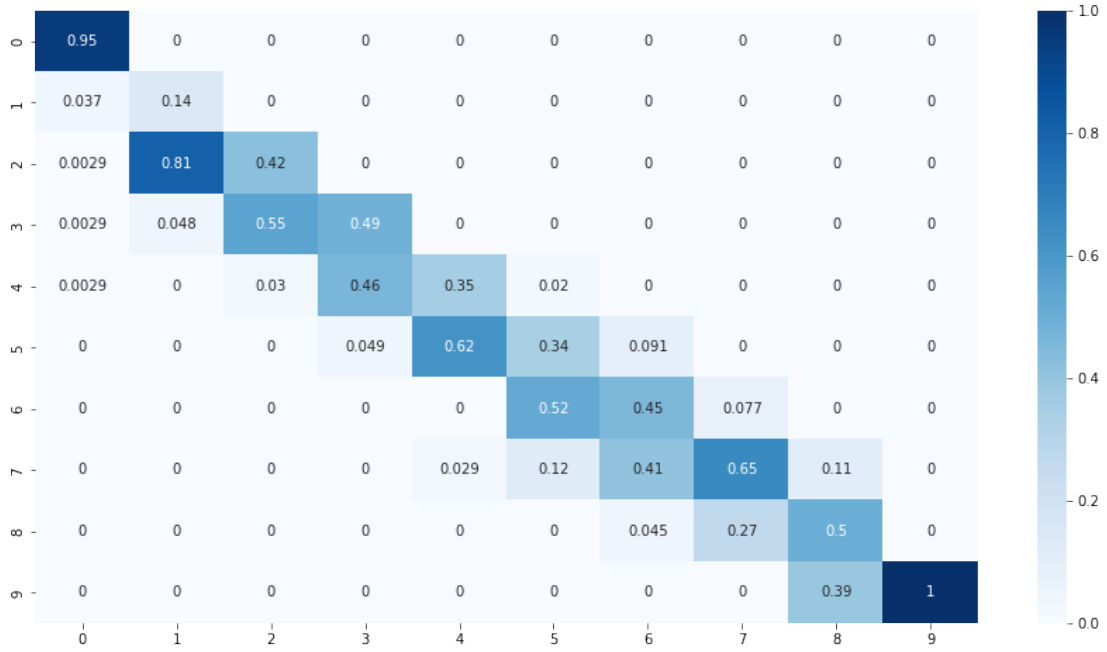

Figure 5.9: Confusion matrix showing results per score class of one of the models in the folds with Cohen's kappa $\kappa = 0.48$.

A confusion matrix of one of the folds is shown in figure 5.9 to summarize prediction outcomes in different neck dystonia score classes. Precision, Recall, and F1-Score are respectively 0.40, 0.32 and 0.32. This can be compared to the outcome in figure 5.6 which shows the outcome when complete data were used in both training and testing.

5.3 Using 3D Video Pose Methods

The nature of keypoint annotation of the human body naturally inspires the 3D methods. However, the videos from single cameras are not enough to reconstruct the original scene in 3D.

Although considerable efforts have been made on 3D human body estimation from a single camera, it turned out to be problematic in our cases due to the following reasons:

- VideoPose3D could not inherently handle multiperson cases, and the 3D results were directly affected by this. Although OpenPose could not track multiple people, it gave keypoint of all human bodies detected on the frame separately.
- The algorithm worked only when the entire human body was visible in the frame. Moreover, this was problematic due to the nature of our data and segments of

interest where the entire body was not visible.

- The variation in motion was not captured much in 3D output. Although the variations were not calculated/measured, it was visible from the visual observations.
- There was no key point to track the eye and ear.

Research in the direction to solve the problem faced with 3D human pose estimations are ongoing. Recent publication trend shows that considerable efforts are taken to run 3D predictions on video segments where the entire body is not visible. However, due to the lack of time and less potential seen with accurate estimations with a single camera, this direction was discarded and deemed inappropriate.

5.4 Using Graph-Based Approach for score prediction

Although GC-LSTM was used, it could not inherently model the processed data. This result was, in fact, not so surprising as the number of nodes used in our GCN layer was significantly smaller, whereas the application involving GCN has a more prominent number of nodes. Although the graph was also static, the result was not that the model over-fitted but quite the opposite. This could be justified with just a reason that the graph we used with eight body point coordinates was too sparse with less number of edges and could not even warm up the GCN.

5.5 Visualizing Data

5.5.1 Abnormality in the multi-line plot

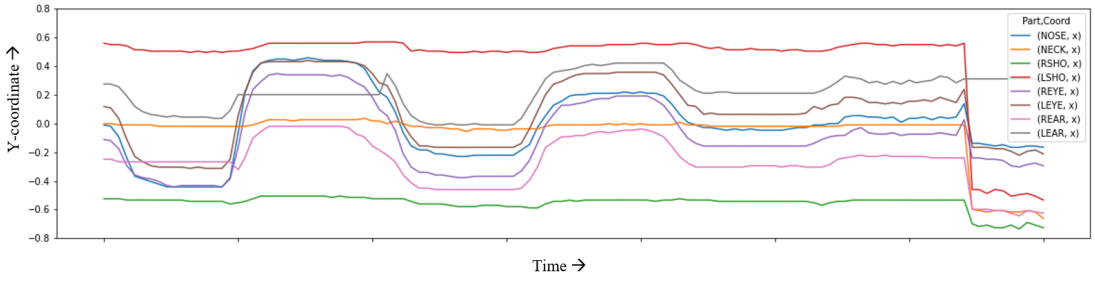

Figure 5.10: Scene Change during assessment on the segment of video depicted in multi-line plot

Figure 5.10 shows a particular case of the x-axis where a very drastic change is visible on the right side. This was discovered very late but was a significant finding due to the

limitation in the precision of video cropping applied where the frames were sampled at 1/100FPS. In a few videos, some extra frames were also cropped, which contained the assessment video in another scene. It was too late to correct this, and there were few cases only; however, these issues could have been solved by using higher precision on the range during video cropping.

5.5.2 Combined animating plot as visual tool

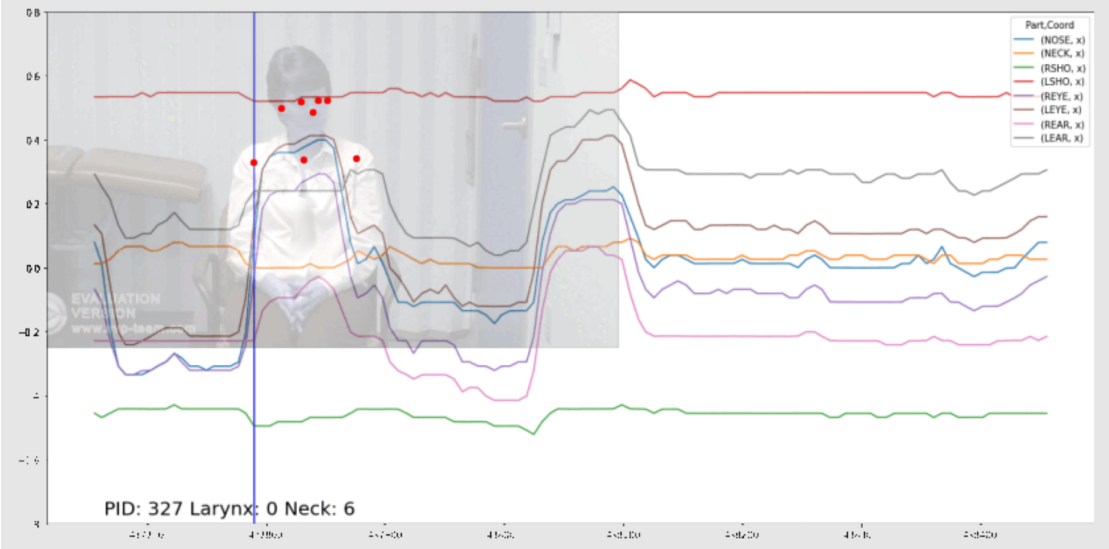

Figure 5.11: Combined animating plot with animating picture of patient and moving line to represent the data at that particular frame.

For aiding the visualization, a combined animating plot, as shown in figure 5.11 was rendered from the processed data. These graphics could be handy for doctors to access the patient instead of just looking at the patient.

Average Plots

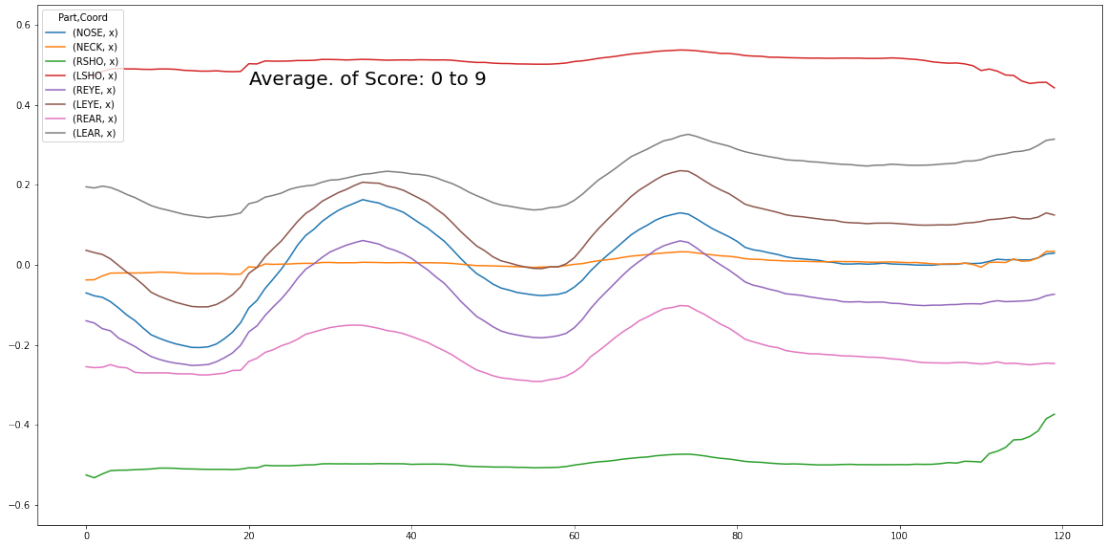

Figure 5.12: Average Multi-line plot for all the data.

Average multi-line plots like in figure like 5.12 were plotted to help understand the variation among the classes. However, due to the non-standardized location of assessment activities in the temporal space, the average plot could not be much help, even strongly for the lower score with very little data. If assessment actions were tagged and the position in the frames was standardized, the average plot would give the immediate visual difference between distinct scores.

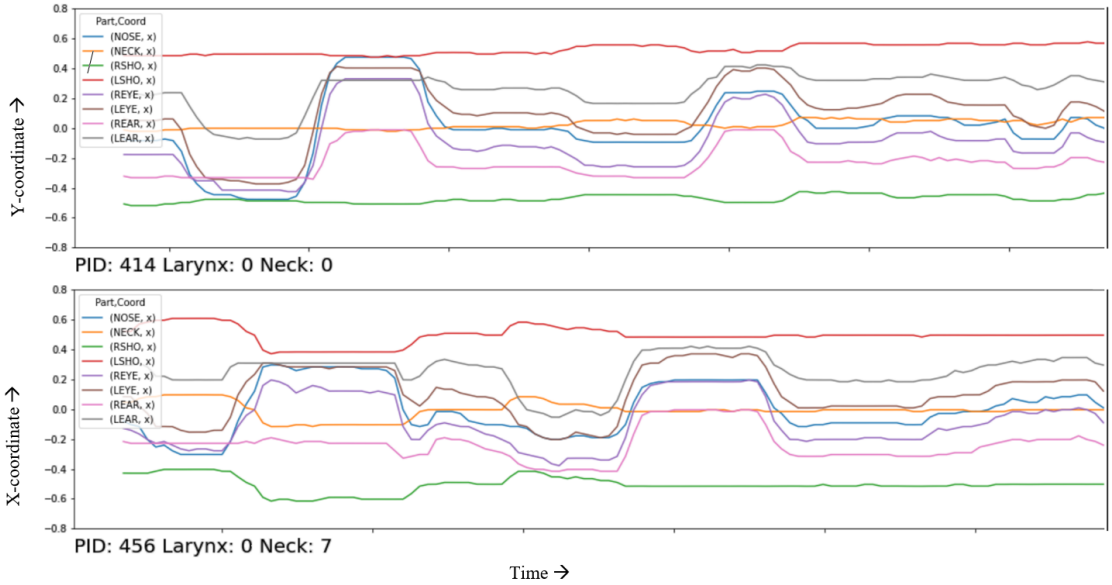

Figure 5.13: Multi-line plot for lower(up) versus higher (down) neck dystonia score.

Fig 5.13 shows the difference between the temporal variations of the body pose key points around the head and neck region in the patients with different neck dystonia scores.

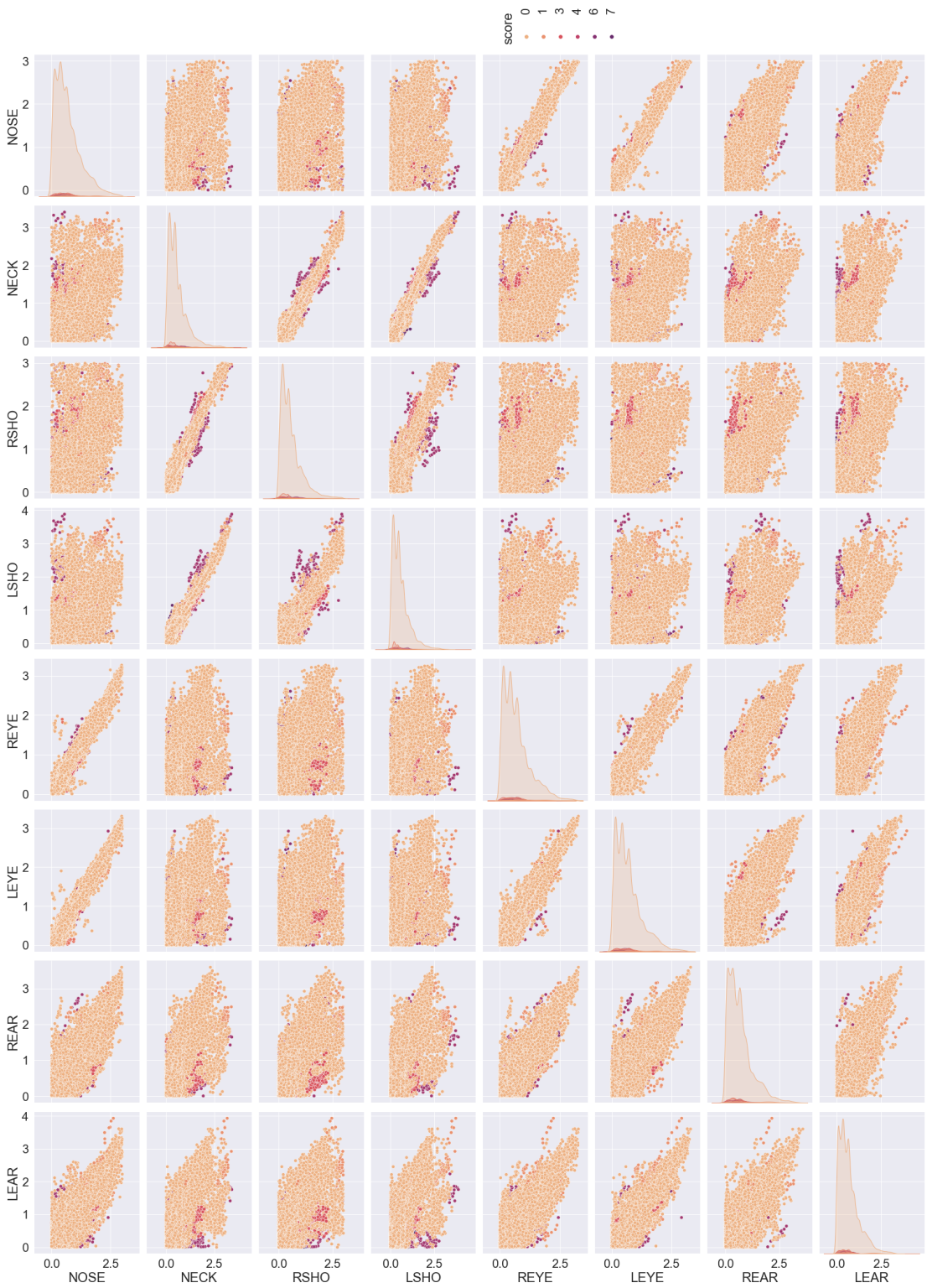

Figure 5.14: Spatial Distribution of body points in Y-axis with the corresponding neck dystonia score represented by color.

Figure 5.14 shows multiplot with spatial Distribution of body points in the Y-axis with the corresponding neck dystonia score represented by colours. A similar plot to visualize the same kind of variation in the X-axis has also been presented in figure 4.12. These visualizations clearly show that data for higher dystonia scores (coloured darker) are in the edges, thus justifying the possibilities of using an automated system to model such patterns.

The relation between the movement of the neck and shoulders in patients with high dystonia scores is captured in the figure 5.14 which indicated that patients with severe dystonia scores could not move their heads up and down (Y-axis) when asked. Figure 4.12 also shows some interesting phenomena in the shoulder and ear relationship: patients with higher scores moved their shoulders too when they were asked to turn their head.

6 FUTURE ENHANCEMENTS

This project depends on the OpenPose algorithm for human position estimation, which outputs 25 body key points. However, newer algorithms with newer architectures, including Transformer, have surpassed OpenPose for the same task. Therefore, using those algorithms would surely help. Also, the inherent handling of human motion tracking in the algorithm would reduce multiple processing steps, mainly in the case of numerous persons present in the same frame. Likewise, 3D based human essential point annotations methods that could inherently handle hidden body parts also could replace OpenPose. If multiple cameras are used to capture patient videos and if algorithms could convert those to 3D key points, that could open another avenue for the research with 3D key points.

Another significant limitation in this project has been due to data limitations. A newer dataset with videos of new patients would surely help run validations on the models. Also, new data collection techniques that aid the standardization of data collection as per the protocol would help collect standard datasets.

Multiple dimensions were necessary to explore, including the project action classification and scene recognition. A critical piece of information that has been unused in the dataset is the audio of the doctor instructing the patients in the video. Although in multiple languages, the audio can also be used as an input for action recognition, which would help better understand the scene and actions being taken automatically.

The score that was considered for modelling was only the neck dystonia score. However, since all the scores from the dystonia assessment assessing various body parts are already present in the dataset, the door is always open for modelling other scores with the appropriate video segment. A better approach can combine multiple models into a complete pipeline that could predict multiple scores from a given video of patients.

Although graph-based were not valuable for our case with few key points, they might give better results with full-body keypoint and careful modelling with minute modifications. This looks very intuitive when analyzing skeleton data, a graph in itself.

A simple approach was used to handle missing key points like the missing ears when

patients turned the head extreme right/left that filled the missing coordinate with the last value available. However, a better approach could replace linear or spline interpolation. Systems can use even machine learning-based techniques to estimate that gap. Also, annotations estimation models that could inherently handle such occlusions would be beneficial to eliminating these problems.

The precision of the start and end-points of the video segments was manually annotated; however, the custom tool made had very low accuracy in recording the position, creating problems on few annotations. Therefore, annotations should be done carefully so that the segment should not include the wrong scene.

CNNs were only used in this project; however, there are many varieties of ML models that we can indeed try on this problem. Although data-hungry, transformers are an excellent candidate for such a perspective.

7 CONCLUSION

This initiative aimed to solve the following research question:

Are computer vision-based approaches capable of an automated non-obtrusive clinical assessment of dystonia?

This project has contributed to the area of objective dystonia assessment in the following ways to achieve the goal:

- Use of one of the deep learning algorithms for human pose estimation in videos of dystonia patients being clinically assessed to annotate body key points in the videos.
- Explored basic pipeline steps required to process the clinical videos, including spatial and temporal normalization.
- Shown that one of the deep learning methods: CNNs, can predict neck dystonia scores leaving space for further research.

With the above contributions, this project has shown high hopes for computer vision-based approaches in an automated system that could perform a non-obtrusive clinical assessment of dystonia. Bringing these kinds of designs from concept to reality will need the cooperation of all stakeholders, including engineers, doctors, and patients. Nevertheless, this is essential to develop the technology foundation for an automated dystonia assessment.

APPENDIX A

A.1 Project Schedule

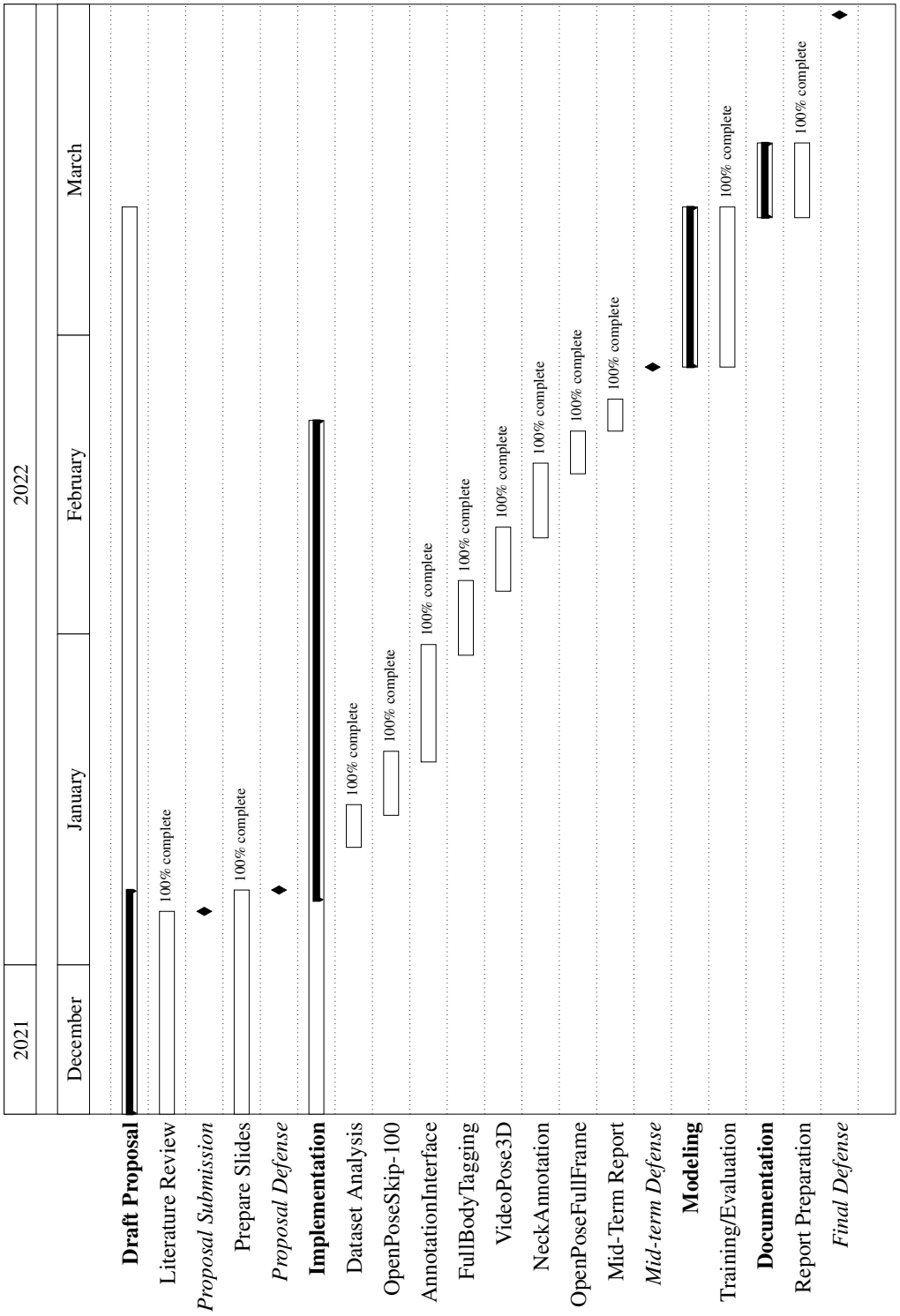

Figure A.1: Gantt Chart showing Expected Project Timeline.

A.2 Literature Review of Base Paper- I

Author(s)/Source: Łukasz Kidziński, Bryan Yang, Jennifer L. Hicks, Apoorva Rajagopal, Scott L. Delp & Michael H. Schwartz

Title: DNNs Enable Quantitative Movement Analysis Using Single-Camera Videos

Website: `https://www.nature.com/articles/s41467-020-17807-z`

Publication Date: August 2020	**Access Date: December, 2021**
Journal: Nature Communications	**Place:** n/a
Volume: 11	**Article Number:** 4054 (2020)

Author's position/theoretical position: Reknowned researchers from multiple research labs at Stanford, working on the intersection of computer science, statistics, and biomechanics.

Keywords: Parkinson disease; optical motion capture; CNN (from related research)

Important points, notes, quotations	**Page No.**
1. Single-event multilevel surgery prediction from CNN model correlated with GDI score.	**6**
2. Found derived time series parameter that improved model performance.	**7**
3. Used OpenPose for extracting time series of human body landmarks.	**7**
4. Special equipment, such as optical motion capture are used with ML models.	**2**

Essential Background Information: Quantitative evaluation of movement is currently only possible with expensive movement monitoring systems and highly trained medical staff.

Overall argument or hypothesis: DNNs can be used to predict clinically relevant motion parameters from a normal patient video.

Conclusion: DNN can help patients and clinicians assess the first symptoms of neurological diseases and enable low-cost monitoring of the progression of the disease.

Supporting Reasons

1. The GMFCS predictions were consistent with the assessments of doctors than of parents.

2. Neural Networks can reduce the cost of using optical motion capture devices.

3. Using DNNs does not require specialized training or equipment.

4. Technicians don't need to put markers on patients and use commodity hardware.

5. Smartphone cameras capture videos at sufficient resolution/quality for feeding to the model.

6. Gait quantification with commercial cameras aids quantitative movement analysis.

7. Generalizes well to a diverse impaired population and does not need to use hand-crafted features.

8. Multiple ML models were trained to predict gait parameters and tested.

Strengths of the line of reasoning and supporting evidence: Performance measures of using CNN for walking parameters were done and a Strong correlation was reported in the predictions of test sets.

Flaws in the argument and gaps or other weaknesses in the argument and supporting evidence: CNN needs lots of training examples. In the case of tasks with limited data available, feature engineering with other classical machine learning models might outperform CNNs.

A.3 Literature Review of Base Paper- II

Author(s)/Source: John Prince	
Title: Objective Assessment of Parkinson's Disease Using Machine Learning	
Website: `https://ora.ox.ac.uk/objects/uuid:fa35ec54-cb90-42f9-ae1a-cf1cf73f32e3`	
Publication Date: October, 2018	**Access Date: December, 2021**
Publisher or Journal: University of Oxford	**Place:** Department of Engineering Science
Volume: n/a	**Issue Number:** n/a
Author's position/theoretical position: PhD Student	
Keywords: Parkinson's disease, motor & non-motor learning, longitudinal phenotypes, digital biomarkers, smartphones, m-health (from related research)	

Important points, notes, quotations	**Page No.**
1. Digital sensors to objectively and quantitatively evaluate PD has been studied.	**188**
2. Wearable sensors can aid in regular clinical care on a large and diverse cohort.	**79**
3. Remote disease classification on the largest cohort of participants.	**134**
4. A dataset deconstruction technique with ensemble learning.	**135**

Essential Background Information: The current evaluation of PD is carried out infrequently due to infeasibility and needs clinical setting.

Overall argument or hypothesis: Digital wearable sensors have ability of performing objective disease quantification and can be effectively utlized to evaluate the PD patients remotely.

Conclusion: Clinical features derived from wearable sensors can perform disease classification and severity prediction on a diverse population.

Supporting Reasons

1. To overcome source-wise missing data, a novel methodology was used.

2. Identification of new longitudinal symptoms in motor and non-motor tasks.

3. Longitudinal behavior between motor and non-motor symptoms is studied.

4. Disease assessment in a remote environment using smartphones is investigated.

5. Using Convolutional neural networks improved classification.

6. Data were collected continuously and concentrated on the time when a tremor occurred.

7. ML model with a large cohort improved the remote classification of PD.

8. The parkinsonian tremor was differentiated from essential tremor with 96% accuracy.

Strengths of the line of reasoning and supporting evidence: Quantitative analysis of errors caused by the methods of imputation and automatic encoding was performed, which reveals the applicability of each technique.

Flaws in the argument and gaps or other weaknesses in the argument and supporting evidence: The remotely collected data set has not been clinically validated and is from a diverse population. There is a naive assumption that demographic data is accurate. The data collection and tests are focused on motion analysis only excluding other symptoms for PD. Many surveys also had binary-type questions.

A.4 Literature Review of Base Paper- III

Author(s)/Source: Li, Michael Hong Gang	
Title: Objective Vision-based Assessment of Parkinsonism and Levodopa-induced Dyskinesia in Persons with Parkinson's Disease	
Website: `https://tspace.library.utoronto.ca/handle/1807/77844`	
Publication Date: June, 2017	**Access Date: December, 2021**
Publisher or Journal: University of Toronto	**Place:** School of Graduate Studies
Volume: n/a	**Issue Number:** n/a
Author's position/theoretical position: Master's Student	
Keywords: Computer vision; Deep learning; Disease management; Health monitoring; Parkinson's disease	

Important points, notes, quotations	**Page No.**
1. Development of human pose estimation benchmark from the PD evaluation dataset.	**90**
2. Two DL methods have been tested for Pose Estimation	**22**
3. Video-based features as clinically actionable information for neuroscientists.	**58**
4. Markerless computer-based visual system to complement existing clinical practice	**94**

Essential Background Information: Computerized assessment can be a solution to the need of frequent automatic assessment of Parkinson's Disease signals without the help of a doctor.

Overall argument or hypothesis: Computer vision methods are capable of tracking the position and movement of the body in clinical PD assessment videos such that scores calculated can be coorelated with clinical socres.

Conclusion: The results show that the pose estimation algorithm can extract relevant information about the motor signals of Parkinson's disease from video assessments and the calculated scores correlates well.

Supporting Reasons

1. Models using movement features extracted from videos as input correlated to clinical ratings.

2. Could detect the presence of PD/LID and also predict its severities.

3. Objective movement features could bring a new scoring paradigm in PD assessment.

4. Evaluated latest human pose estimations algorithms in clinical assessment videos.

5. Exploration of the motion features that could be extracted from video analysis.

6. Identification of the important features of the movement for good model performance.

7. The regression model predicted severity with a high correlation with the clinical score.

8. Although consumer-grade video cameras were used, results are promising.

Strengths of the line of reasoning and supporting evidence: Objective evaluation has been done with strong evidence. E.g., evaluation of correlation coefficients has confirmed the results with a sufficient degree. Also, their results with benchmarking datasets have strong mathematical ground.

Flaws in the argument and gaps or other weaknesses in the argument and supporting evidence: Because of pose estimation from a single 2D image, information loss occurs when the patient is moving perpendicularly to the camera plane and largely influence the results.

REFERENCES

[1] Mahmoud Al-Faris, John Chiverton, David Ndzi, and Ahmed Isam Ahmed. A Review on Computer Vision-Based Methods for Human Action Recognition. *J. Imaging*, 6(6), Jun 2020.

[2] A. Albanese, M. P. Barnes, K. P. Bhatia, E. Fernandez-Alvarez, G. Filippini, T. Gasser, J. K. Krauss, A. Newton, I. Rektor, M. Savoiardo, and J. Valls-Solè. A systematic review on the diagnosis and treatment of primary (idiopathic) dystonia and dystonia plus syndromes: report of an EFNS/MDS-ES Task Force. *Eur. J. Neurol.*, 13(5):433–444, May 2006.

[3] Alberto Albanese, Kailash Bhatia, Susan B. Bressman, Mahlon R. DeLong, Stanley Fahn, Victor S. C. Fung, Mark Hallett, Joseph Jankovic, H. A. Jinnah, Christine Klein, Anthony E. Lang, Jonathan W. Mink, and Jan K. Teller. Phenomenology and classification of dystonia: a consensus update. *Movement disorders : official journal of the Movement Disorder Society*, 28(7):863, Jun 2013.

[4] Georg Becker, Daniela Berg, Michael Francis, and Markus Naumann. Evidence for disturbances of copper metabolism in dystonia: From the image towards a new concept. *Neurology*, 57(12):2290–2294, Dec 2001.

[5] Vladislava Bobić, Milica Djuric-Jovicic, Nathanael Jarrasse, Milica Jecmenica-Lukic, Igor Petrovic, Saša Radovanović, Natasa Dragasevic, and Vladimir Kostić. Spectral parameters for finger tapping quantification. *Facta universitatis - series: Electronics and Energetics*, 30(series: Electronics and Energetics):585–597, Jan 2017.

[6] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019.

[7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1):172–186, Jan 2021.

[8] Yu Cheng, Bo Yang, Bo Wang, Yan Wending, and Robby Tan. Occlusion-aware networks for 3d human pose estimation in video. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019.

[9] Chien-Wen Cho, Wen-Hung Chao, Sheng-Huang Lin, and You-Yin Chen. A vision-based analysis system for gait recognition in patients with Parkinson's disease. *Expert Syst. Appl.*, 36:7033–7039, Apr 2009.

[10] Leslie J. Cloud and H. A. Jinnah. Treatment strategies for dystonia. *Expert Opin. Pharmacother.*, 11(1):5, Jan 2010.

[11] Cynthia L. Comella, Sue Leurgans, Joanne Wuu, Glenn T. Stebbins, Teresa Chmura, and The Dystonia Study Group. Rating scales for dystonia: a multicenter assessment. *Mov. Disord.*, 18(3):303–312, Mar 2003.

[12] Huseyin Coskun, Felix Achilles, Robert DiPietro, Nassir Navab, and Federico Tombari. Long short-term memory kalman filters: Recurrent neural estimators for pose regularization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5524–5532, 2017.

[13] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. *Lecture Notes in Computer Science*, page 679–696, 2018.

[14] Giovanni Defazio, Mark Hallett, Hyder A. Jinnah, Glenn T. Stebbins, Angelo F. Gigante, Gina Ferrazzano, Antonella Conte, Giovanni Fabbrini, and Alfredo Berardelli. Development and validation of a clinical scale for rating the severity of blepharospasm. *Mov. Disord.*, 30(4):525–530, Apr 2015.

[15] Lucy Frucht, David L. Perez, Janet Callahan, Julie MacLean, Phillip C. Song, Nutan Sharma, and Christopher D. Stephen. Functional Dystonia: Differentiation From Primary Dystonia and Multidisciplinary Treatments. *Front. Neurol.*, 11, 2020.

[16] Richard Green, Ling Guan, and John Burne. Video analysis of gait for diagnosing movement disorders. *J. Electronic Imaging*, 9:16–21, Jan 2000.

[17] Grażyna Gromadzka, Beata Tarnacka, Anna Flaga, and Agata Adamczyk. Copper Dyshomeostasis in Neurodegenerative Diseases—Therapeutic Implications. *Int. J. Mol. Sci.*, 21(23):9259, Dec 2020.

[18] Eni Halilaj, Apoorva Rajagopal, Madalina Fiterau, Jennifer L. Hicks, Trevor J. Hastie, and Scott L. Delp. Machine learning in human movement biomechanics: Best practices, common pitfalls, and new opportunities. *J. Biomech.*, 81:1, Nov 2018.

[19] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 68–84, 2018.

[20] H. A. Jinnah. Diagnosis & Treatment of Dystonia. *Neurol. Clin.*, 33(1):77, Feb 2015.

[21] Taha Khan, Dag Nyholm, Jerker Westin, and Mark Dougherty. A computer vision framework for finger-tapping evaluation in Parkinson's disease. *Artif. Intell. Med.*, 60, Nov 2013.

[22] Łukasz Kidziński, Bryan Yang, Jennifer L. Hicks, Apoorva Rajagopal, Scott L. Delp, and Michael H. Schwartz. Deep neural networks enable quantitative movement analysis using single-camera videos - Nature Communications. *Nat. Commun.*, 11(4054):1–10, Aug 2020.

[23] Young-Tae Kwon, Yongkuk Lee, Gamze Kilic Berkmen, Hyo-Ryoung Lim, Laura Scorr, H. A. Jinnah, and Woon-Hong Yeo. Soft Material-Enabled, Active Wireless, Thin-Film Bioelectronics for Quantitative Diagnostics of Cervical Dystonia. *Advanced materials technologies*, 4(10), Oct 2019.

[24] Kyoungoh Lee, Inwoong Lee, and Sanghoon Lee. Propagating lstm: 3d pose estimation based on joint interdependency. *Lecture Notes in Computer Science*, page 123–141, 2018.

[25] Michael H. Li, Tiago A. Mestre, Susan H. Fox, and Babak Taati. Vision-based assessment of parkinsonism and levodopa-induced dyskinesia with pose estimation. *J. NeuroEng. Rehabil.*, 15, 2018.

[26] Mude Lin, Liang Lin, Xiaodan Liang, Keze Wang, and Hui Cheng. Recurrent 3d pose sequence machines. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 810–819, 2017.

[27] Andres M. Lozano, Nir Lipsman, Hagai Bergman, Peter Brown, Stephan Chabardes, Jin Woo Chang, Keith Matthews, Cameron C. McIntyre, Thomas E. Schlaepfer, Michael Schulder, Yasin Temel, Jens Volkmann, and Joachim K. Krauss. Deep brain stimulation: current challenges and future directions. *Nat. Rev. Neurol.*, 15(3):148, Mar 2019.

[28] Alexander Mathis, Steffen Schneider, Jessy Lauer, and Mackenzie Weygandt Mathis. A Primer on Motion Capture with Deep Learning: Principles, Pitfalls, and Perspectives. *Neuron*, 108(1):44–65, Oct 2020.

[29] Andrea H. Németh. The genetics of primary dystonias and related disorders. *Brain*, 125(4):695–721, Apr 2002.

[30] George Papandreou, Tyler Lixuan Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin P. Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *ECCV*, 2018.

[31] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019.

[32] John Prince. *Objective assessment of Parkinson's disease using machine learning*. PhD thesis, University of, Oxford, England, UK, Jan 2018.

[33] Suraj Rajan, Bonnie Kaas, and Emile Moukheiber. Movement Disorders Emergencies. *Semin. Neurol.*, 39(1):125–136, Feb 2019.

[34] Anusha Sathyanarayanan Rao, Benoit M. Dawant, Robert E. Bodenheimer, Rui Li, John Fang, Fenna Phibbs, Peter Hedera, and Thomas Davis. Validating an objective video-based dyskinesia severity score in Parkinson's disease patients. *Parkinsonism Relat. Disord.*, 19(2):232–237, Feb 2013.

[35] Helge Rhodin, Frederic Meyer, Jorg Sporri, Erich Muller, Victor Constantin, Pascal Fua, Isinsu Katircioglu, and Mathieu Salzmann. Learning monocular 3d human pose estimation from multi-view images. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[36] Kristina Simonyan. Neuroimaging Applications in Dystonia. *Int. Rev. Neurobiol.*, 143:1, 2018.

[37] Kirsty Stewart, Adrienne Harvey, and Leanne M. Johnston. A systematic review of scales to measure dystonia and choreoathetosis in children with dyskinetic cerebral palsy. *Dev. Med. Child Neurol.*, 59(8):786–795, Aug 2017.

[38] Bugra Tekin, Artem Rozantsev, Vincent Lepetit, and Pascal Fua. Direct prediction of 3d body poses from motion compensated sequences. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.

[39] Davide Valeriani and Kristina Simonyan. A microstructural neural network biomarker for dystonia diagnosis identified by a DystoniaNet deep learning platform. *Proc. Natl. Acad. Sci. U.S.A.*, 117(42):26398–26405, Oct 2020.

[40] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional Pose Machines. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732. IEEE, Jun 2016.