Dynamic Pose Diagnosis with BlazePose and LSTM for Spinal Dysfunction Risk Estimation

1st Rudramani Gyawali Singha
Dept. of Information Technology
RGIT, Mumbai University
Mumbai,India
rudramani.singha@gmail.com

2nd Manan Lad
Dept. of Information Technology
RGIT, Mumbai University
Mumbai,India
m616e616e.lad@gmail.com

3rd Gaurav Mukesh Shipurkar
Dept. of Computer Engineering
MPSTME, NMIMS University
Mumbai,India
gaurav.shipurkar4@gmail.com

4th Anish Rohekar
Dept. of Information Technology
RGIT, Mumbai University
Mumbai,India
rohekar.anish@gmail.com

5th Chandar Chauhan
Research & Development Dept
AI Lab, AZYO
New Delhi,India
chander.chauhan@azyo.io

6th Nilesh Rathod
Dept. of Information Technology
RGIT, Mumbai University
Mumbai,India
nilesh.rathod@mctrgit.ac.in

***Abstract*—The bodily movements of a person can be heavily restricted because of a spinal cord disorder. These disorders could be a result of various internal or external factors but we believe that a preliminary diagnostic for the magnitude of the spinal dysfunction can be determined by the movement of the patient itself. The behavioural patterns of a person doing specific tasks with spinal dysfunction varies from someone who does the same tasks with a healthy back. In this contribution, we demonstrate that this difference can be generalised and then implemented to predict the spinal health index of the patients with only but an input video of them performing predetermined tasks. Our system features a novel technique of abstracting dynamic heat-map as the sequential input to the custom LSTM model. Allowing us to outperform previous techniques by reaching 98.4% validation accuracy with a realtime computational complexity of ≤11 milliseconds.**

***Index Terms*—Bioengineering, Machine-Learning, Spine Diagnostics, Pose estimation, Dynamic Gesture Recognition**

I. Introduction

The spinal cord is responsible for electrochemical communication where the current travels across the nerves, sending signals to various parts of the body to establish a connection with the brain. A healthy spine always has a slight bend that assimilates the stress from the body movements. Recent studies have shown that spine curvature disorders like osteoporosis, spondylolisthesis, kyphosis, scoliosis, etc are increasing day by day in the world affecting the lives of many people disastrously. These disorders are responsible for the misalignment of the spinal cord causing it to bend more than normal. Most common spine disorder symptoms include swaybacks, pronounced buttock positions, problems in moving certain ways, a large gap between lower-back and floor when lying down, bending forward, hump / upper back curve more than usual, uneven shoulder blades and hip/waist positions, leaning towards one side, etc. depending on the severity of the conditions these symptoms vary from person to person.

These disorders if detected at an early stage, have a higher chance of getting treated through therapy and early observations of the doctor. Surgeries can be avoided and in adolescents, back braces can be applied to avoid further worsening of the back. Various exercise programs, electrical stimulation therapy, nutrition therapy are known to improve the spine condition with higher success rates in the early stages.

Gait analysis [13], a study used to identify posture-related or movement-related problems by analyzing the patient's leg movements through walking or running exercises. The proposed system however focuses on back postures and upper body movements with the help of a camera and no other external types of equipment are used, therefore, is different from the gait analysis [13].

In order to detect the spinal dysfunctions existing technologies rely on X-ray, CT scan and specialized UV tools. These technologies rely on costly machineries and cannot be performed in real time. We propose a temporal solution in order to solve the problem of detecting spinal disorder rather than image classification. Instead of using images in order to make inference, we utilize a 5 sec, 30 fps video. Human pose features throughout the video are used as input features for getting the results.

Our project's approach involves Google's new open-source project, Mediapipe [12]. Mediapipe [12], being a framework for building multimedia ML pipelines, contains a variety of pretrained multimodal(image, audio, video) models with high accuracy and speed. Blazepose [8], [12] is a pose detection model from Mediapipe which can track the whole body movement through 33 landmark points. From the body landmark key points detected by Blazepose [8] model, body posture of a person can be determined.

We are proposing a method that is based on the body posture and distorted movements of the back, that are the symptoms of spinal cord disorders. Utilizing this method we can then estimate the condition of the spinal cord. The risk factor can be calculated and accordingly, actions can be taken to prevent further harm.

II. Related Works

In order to analyze scoliosis and for spinal canal analysis; spinal cord detection, spine lumber detection, and spinal canal estimation and visualization are employed. In all the previous works spinal cord images are taken from CT scans, ultrasound black and white images, and X-rays images of the lateral spine are taken for detection and segmentation. Localization of the lumbar region and visualization of the ROI is necessary for the treatment and analysis of scoliosis patients.

The spinal cord detection task is broken down into sub-tasks like spine end detection and centerline delimitation [1]. like in paper [1] Object detectors like Alex-Net trained on Imagenet and fine-tuned for detecting cervical vertebra complex, sacral bone, and background are employed as SpineCNN. It is trained on a dataset consisting of CT scan images of 392 patients. In order to delimitate centerline, models like VGG16 + faster RCNN as CordCNN. After detection confidence is further calculated using mean curve distance of spinal axis and standard deviation of intestines. Unlike [1] which uses object detectors on CT scan images CAD detection algorithm [7] is used in the identification of spine midline, spinous processes, and inter laminar spaces in [2]. CAD algorithm works on high-contrast bone images where the bright bone surface is spinous/epidural space cross-section. The algorithm is used to map the bone surface to a 3d model of the lumbar segment.

Feature Extraction is used to identify bone structure [3], [6]. Two different techniques namely triangular analysis [3] Density-based ellipse-like object detection are proposed. [6] Focuses on ROI segmentation of lumbar spine region to locate valley areas using horizontal prediction method. foreground separation is done using entropy value as a threshold. binarized image is filtered for noise and triangular analysis is performed to identify ROI. [3] uses Ellipse-like object detection for vertebrae centroid detection with reference to facet. Image is preprocessed using Bi-Histogram Equalization with adaptive sigmoid functions and the improved image is filtered for noise. Vertebral objects are detected using the Multi-theta Gabor filter process and mean range candidate selection is applied to make point estimation using detected centroid. [3] uses feature detection techniques on X-ray images of the lateral spine in order to detect bone fractures and spondylosis. [4] also uses 695 generated spine models from pre-intra-op and post-op X-ray images.

Unlike other methods, landmarks are annotated by an orthopedist, and the model is passed as input to the proposed STCN (Spatial-Temporal Corrective Network) model [4]. [1] uses a combination of ResNet and VGG for spine detection, the STCN uses a full convolution network (FCN) and ResNet-101 in order to detect extract inter-vertebral disk. FCN is used to improve predictions. The target of the STCN model is to capture the varying extent of voxel regions. FC-ResNet and STCN in combination are used for disk extraction and landmark detection. This method was proposed for application in AVBGM surgery in scoliosis patients. The approaches heavily depend on object detectors and feature extraction algorithms in order to localize the lumber area. For feature extraction foreground separation and histogram-based thresholding is done finally using the centroid of detected features to draw the final spinal line.

The system is tested ideally in bright lighting conditions. Blaze pose used in the system if used in low light conditions will add abnormalities in the detection. System focuses on human postures and utilizes only the camera for its application, and therefore expects the user to be wearing light clothing. System aspects allow users to perform certain actions which are not always suitable for the users and may lead to incomplete results.

III. Proposed Work

A. System Architecture

Our implemented system, as described in Fig.1, can be divided into a series of 4 consecutive major steps to find the estimation of the spine's health.

The first step overlooks the input of the image sequence with each image being resized to $1280 * 720$ pixels. This image sequence is extracted from a live video stream or a pre-recorded saved video. Images extracted in the first step are stored in a sequence buffer of size N. Each image contains RGB (Red, Green, and Blue) channels of color information and is represented in an array of dimensions $1280 * 720 * 3$.

Each image array in this buffer has to iteratively feed into the implementation of BlazePose [8] landmark detection. BlazePose [8] is a pose landmark estimation pre-trained binary that is lightweight ($\leq$2 MB) and real-time ($\geq$60 Fps) while also being extremely accurate with over 94% validation accuracy on customized coco dataset. Using this model we are able to estimate the pose locator for our own N number of data points in the sequence buffer. At a time, a single image (x_t) from the image buffer pool is being pulled over, predicted upon, and then placed inside another buffer called landmark buffer. With each prediction, the landmarks of the image sequence are recorded into the database and optionally also drawn for the user and displayed as live feedback to their output screens. For the purpose of BlazePose [8], we are implementing it from the mediapipe framework [12] as it also has $\geq$50 other supplementary functions that further assist us in post-processing in the coming steps.

Secondly, each vector of predicted landmarks that has been recorded into the landmark buffer, represents the entirety of the original 1920*1080*3 dimensional image array from which it has been derived. The landmarks are a higher abstracted representation of the same pixel information that only stores and records information about the pose within the original image. This essentially is an advanced form of feature engineering but done using a deep neural network where all the unnecessary pixel information irrelevant to the pose is discarded efficiently with minimum manual effort. The average cost per execution of this computation is $\leq$ 48 ms when iterated over $\geq$ 20,000 images of standard size 1920*1080*3.

Every prediction consists of 33 total landmarks and each landmark consists of 3 coordinates. Namely, x, y, and z for

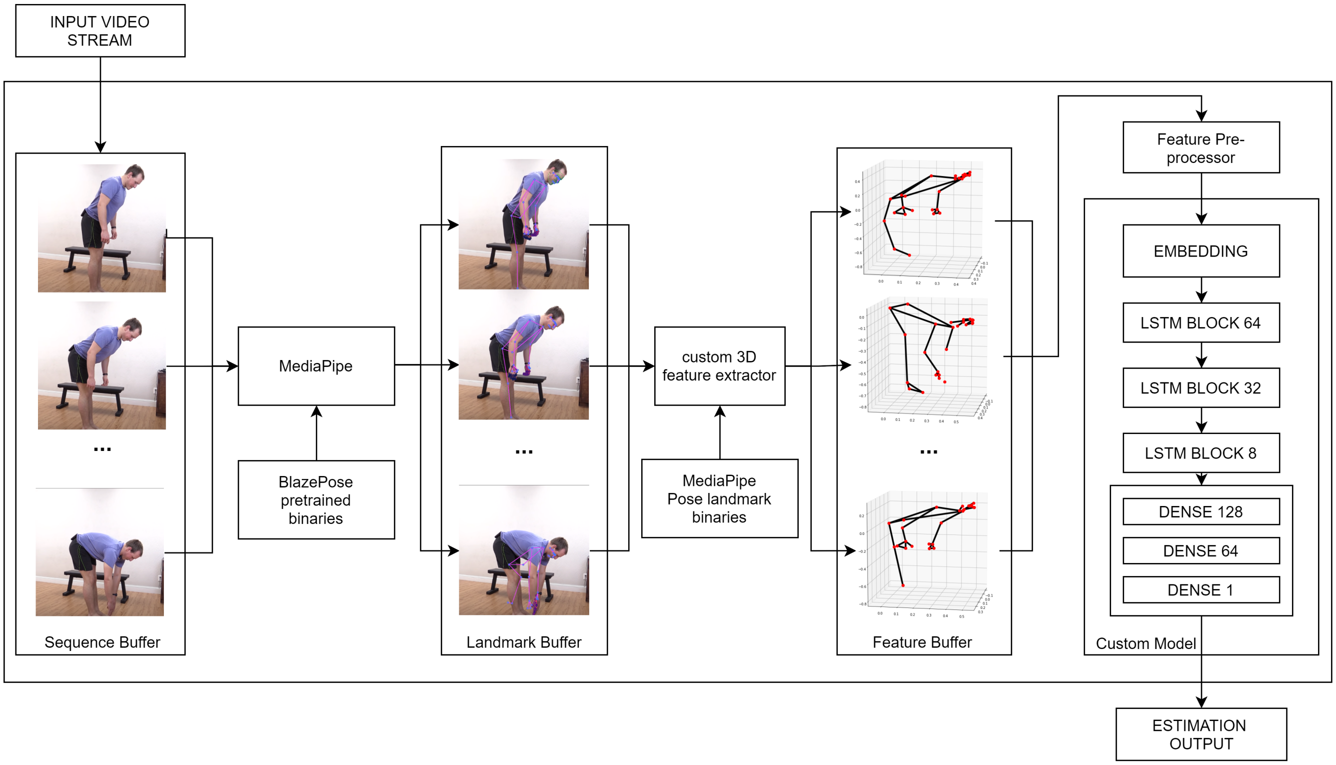

Fig. 1. The implemented system architecture for Spinal Dysfunction Risk Estimation

each dimension of the spatial domain. 33 landmarks represent a specific location on a human body that is visible. Therefore, in the next step, we can create a method that extracts all the features from the landmarks that are of confidence $\geq 60\%$ and place them in a special buffer. This process eliminates $\approx 40\%$ of all outliers that skew the data and introduce inaccuracies by "dirtying" the data.

The total number of images, $\sum x$, at a time in the system can only be

$$\sum x \leq N * \gamma \tag{1}$$

where γ is the number of sequence buckets in the whole system and N is the total size of the buffer.

Having all the 3-dimensional data extracted and cleaned, we can represent them in 3d space and pass it onto a feature pre-processor engine that flattens them all down to sequential format for embedding. The pre-processor ensures that all the input to the custom model matches the input dimensions correctly and normalizes all the data passing into it using min-max scaling. The pre-processor acts as a critical part of the system and with our experiments, we have noticed that with the pre-processor, our accuracy increases by $\approx 4\%$ - 5%

These data that are fed into the embedding layer of the custom model, have finally entered the last step that is the risk detection step of the entire system. Pose feature maps obtained from the blaze pose include 3D pose coordinates of the human body, which includes 193 coordinate values. These values are embedded to form LSTM [6] inputs. The custom model detects the risk by utilizing the LSTM [6] blocks to understand the relationship of the landmarks with respect to time, further detailed in section III-D. And the fully connected block, along with all its 193 densely connected neurons and activation functions, abstract the results of the LSTM [6] blocks and derive at a single-digit estimation for the risk by the final neuron in the last sequential layer.

The very final output neuron has to pass its prediction through the ReLU [10] activation function. ReLU [10] is described as

$$f(x) = \begin{cases} x & \text{if } x > 0, \\ 0.01x & \text{otherwise.} \end{cases} \tag{2}$$

where ReLU [10] for units $\geq$0 scale linearly but for values of x $\leq$0 scale with a decreased rate.

Using this final output, our implemented model is able to apply back-propagation and generalize even better as it goes through more datasets and iterations. And hence provide a confident, high accuracy of 93.7% accuracy on our validation set.

B. Gathering Ground Truth Training and Validation set

The strategy we have implemented is to crowdsource the data collection effort to a wide diversity of people within the education institution of RGIT, Mumbai University. A set of closely regulated 150 volunteers have been interviewed and shortlisted as valid candidates for the purpose of Ground Truth data generation. $\leq$30% of the applicant selected identify as female with the survey while the males constitute the rest

≥70% of the group. ≥97% of the candidates belong to the category of age group ≤25% while the remaining candidates average a calculated age of ≈ 43.4 years.

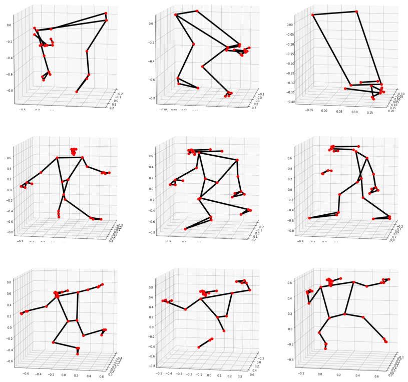

Fig. 2. Sample Inputs landmarks from the candidates

Fig. 2. represents sample inputs gathered from the qualified candidates after the successful completion of instructions given to each volunteer. Candidates were asked to record themselves doing 3 sets of pre-determined exercises while being in full view of the recording camera. Candidates were urged to ensure that the video is of sufficient quality of 1280*720 pixels resolution minimum.

The 3 exercises comprised of, first, ”touch your toes”, where the candidate had to start by standing still and then proceed to bend forward and touch your toes. Second, ”walk for 2 meters”, where they are supposed to walk normally and cover 2 meters in ≤ 60 seconds while also staying within the video frame throughout. Thirdly, ”jump as high as you can”, where the user has to jump while also being within the frame.

Every exercise is to be completed within the restriction of ≤ 60 seconds. And after each and every task, they are to rate what is the current status of their back on a scale of 0 - 100. where 0 represents ”most back pain/stress” and 100 represents ”least back pain/stress”.

These 2 data, the video sequences and the ratings from candidates are used as a single data point in the entire dataset that is used to train and validate the custom model architecture.

97% of the total selected candidates successfully completed the entire task and submitted the valid results for use in the experiment. Indexes are scaled from 0 to 1 where 0 being the best and 1 being the worst state of their condition with proportion to volunteer's input.

C. Landmarks detection and tracking with BlazePose

MediaPipe BlazePose [8] [12] is a real-time body pose tracker, it is used in pose estimation and in our case in order to get the posture of the patient with spinal cord distortion we use BlazePose pose detector. It is light weight as it was initially designed for light weight mobile devices. unlike other Posenets which relies only on heat map based approaches, it uses stacked hourglass architecture [9]. It extends the above architecture with encoder-decoder network architecture that predicts heatmaps of the joints and simultaneously other encoder regresses on ordinates of the joints. Initial pose alignment is used by the BlazePose detector, followed by tracker that detects key points, presence of a person and refined region of interests. BlazePose is trained on 60,000 images containing single or few people in it and 25K images containing only single person performing different exercises.

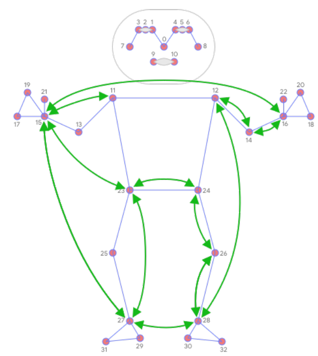

Fig. 3. BlazePose 33 keypoint topology

As seen in Fig. 3 BlazePose topology network consists of 33 human body keypoints, which allows us to obtain very precise data on the slight body movements.

TABLE I
BLAZEPOSE MODEL PERFORMANCE COMPARISON

	Pixel 3 CPU,FPS using XNNPack	***Pixel 3 GPU, FPS using TFLite GPU***
BlazePose lite	44	112
BlazePose full	18	69

Table I shows the performance comparison of the two BlazePose pose tracking models: full and lite on a mobile device (Pixel 3). As seen in the table the difference between the two models is speed or quality. While BlazePose lite model(2.7 MFlop, 1.3M Params) is lightweight and gives an amazing FPS performance, BlazePose full model(6.9 MFlop, 3.5M Params) is a bit heavier than lite for more average precision, sacrificing FPS performance in the process. Our system will use BlazePose lite simply for FPS performance advantage.

D. Memory Cell for Dynamic Gestures

Long Short Term Memory (LSTM) networks [6] are advanced Recurrent Neural Networks (RNN) [11] that we are utilising for dynamic gesture recognition from temporal image sequences. The image sequences make up a time series dataset that acts as an input to the LSTM block. Standard RNNs face the problem of vanishing gradient because of which they are not able to recall long term dependencies.

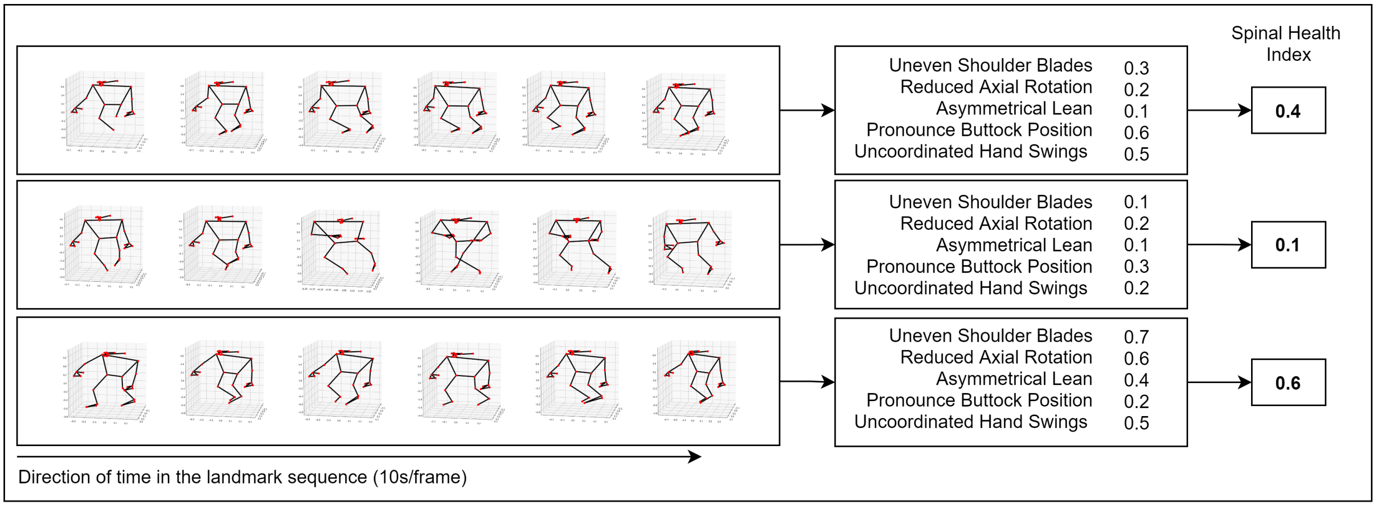

Fig. 4. Spinal Health Index estimation from series of extracted features

TABLE II
LSTM VS RNN

Networks	***comparison***		
	Accuracy	***Image***	***Dimension***
LSTM	87%	500	1920*1080
RNN	37%	500	1920*1080
LSTM	96%	100	720*480
RNN	92%	100	720*480

As seen in Table II, RNNs have less than 37% accuracy for a image sequence with over 120 images ($1920 * 1080$ pixels). LSTMs are developed with the goal of learning the long term dependencies. The LSTM block is a sequential feed forward network architecture which outperform the RNNs by 87% accuracy while validating on a data point of 500 images. The set of functions that define a LSTM block are described in equations (3), (4), and (5).

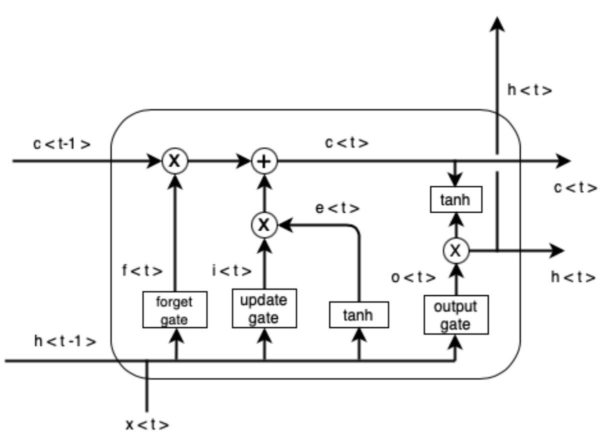

Fig. 5. LSTM Block

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i) \tag{3}$$

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f) \tag{4}$$

$$o_t = \sigma(W_o * [h_{t-1}, x_t] + b_o) \tag{5}$$

Each LSTM block has three gates namely forget gate (f_t), update/input gate (i_t), output gate (o_t) as shown in the above Fig.5. Each gate has its own significance in functioning of the block. Forget gate (f_t) (3) decides which information is to be excluded, it considers the output of the last LSTM block (h_{t-1}) and the input of the current block (x_t), then passes the output through a sigmoid function (σ) proceeding with whether to discard the information. The input/update gate (i_t) (4) decides which details to store in the cell state (c_t), the further tanh layer generates new candidate values which are further coalesced with the the output of the input/update gate to update the cell state (c_t). Output of the block is decided by the cell state (c_t) of the block. The sigmoid (σ) function present in the output gate (o_t) (5) decides what information is going to be passed further, resulting in final output.

And therefore, because of all the LSTM's advantage over RNN for these specific use case, we have opted to integrate it instead to our custom model architecture.

IV. Result Analysis

We are calculating Spinal Health index of people by analysing their movements. Fig.4 demonstrates a sample of candidates performing the task of ”walking” and how our implemented system derives that their specific spine index. Camera angle is fixed in a way that slant angle context of person moments is captured. For healthy person, asymmetrical lean while walking is usually low as compared to person with scoliosis as seen in Fig. 4. Person with healthy back has health index $\geq 0.1 \leq 0.5$ and constitutes to $\geq 79\%$ of the entire sample set. Candidates above this threshold receive red flags and are highly encouraged to seek professional help.

Person in the 3rd row of the Fig.4 is diagnosed with minor scoliosis and as evident from the high shoulder blade risk prediction of index 0.7. $\geq$80% Unstable body movements result in combination of problems like asymmetric lean index between 0.6 to 0.8 and reduced axial rotation index between 0.7-0.8 hence showing high co-relation.

Object Detection techniques like SpineCNN and CordCNN [1] on CT scan images and also detectors like STCN [4] and histogram [3] methods like BEASF proposed in [3] used to locate spinal cord and its dysfunctions on X-ray images do a good job detecting the ROI but they do not give damage estimation of the cord. The method proposed [2] requires specific alignment of the spinal cord for detection. Alternatively, Spine posture detection is carried out in [2] by using feature extraction on ultrasound samples and then mapping centroids detected from triangular analysis to a linear function. The centerline detection process employed requires the use of X-Ray images in order to estimate the spinal cord alignments. We use blaze pose to identify patient's movement behaviors which eliminate the requirement of imaging techniques like pre-intra-op and post-op x-ray in some cases [4]. Early detection and prediction of spinal cord health can be estimated by the combination of back, hip, and shoulder postures. In our experiment, $\geq$76% of the candidates with spinal cord disorders (114 candidates) show characteristics of restricted natural bodily movements like uneven shoulder blades resulting in unstable movements and sway backs causing irregularities while performing actions like bending down or lying down. In order to estimate the likeliness of spine disorder or spine injury in many cases, especially during sporting events, in our experiment, the athlete candidate's back injuries was monitored, and accordingly, severity was determined quickly with over $\leq$87% accuracy by monitoring athlete's movements and changes in bodily reflexes. In order to accurately visualize the vertebra, techniques like binarization and density-based estimations are used. We on the other hand use heat map-based joint estimation to quickly get the context reducing computational thresholds to less than $\geq$11ms from 643ms per iteration and increased validation accuracy on the same dataset from 74.3% to our method's 98.4% .

V. Conclusion

With this paper we have demonstrated the possibility of using BlazePose and LSTMS for the extraction of the user's spine condition. This technology can enhance diagnostic assistance tools by predicting the spine's health through various intermediary detection otherwise difficult for medical professionals to diagnose without movement analysis. Future work involves adding more intermediary diagnostics and larger dataset generation methods with more scalability to people with different medical backgrounds.

References

[1] R. Jakubicek, J. Chmelik, P. Ourednicek and J. Jan, "Deep-learning-based fully automatic spine centerline detection in CT data," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019, pp. 2407-2410, doi: 10.1109/EMBC.2019.8856528.

[2] A. J. Dixon, K. Owen, M. Tiouririne and F. W. Mauldin, "Computer aided detection of lumbar spine landmarks for ultrasound guided lumbar punctures and epidurals," 2017 IEEE International Ultrasonics Symposium (IUS), 2017, pp. 1-4, doi: 10.1109/ULTSYM.2017.8091806.

[3] S. Thamnawat, J. G. Ham and S. Rasmequan, "Region of Interest Identification on Low-Resolution Lateral Spine Radiography Image using Density-based and Ellipse-like Method," 2019 4th International Conference on Information Technology (InCIT), 2019, pp. 77-82, doi: 10.1109/INCIT.2019.8912046.

[4] W. Mandel, R. Oulbacha, M. Roy-Beaudry, S. Parent and S. Kadoury, "Image-Guided Tethering Spine Surgery With Outcome Prediction Using Spatio-Temporal Dynamic Networks," in IEEE Transactions on Medical Imaging, vol. 40, no. 2, pp. 491-502, Feb. 2021, doi: 10.1109/TMI.2020.3030741.

[5] W. Yookwan, K. Chinnasarn and B. Jantarakongkul, "Region of interest of human lumbar spine segmentation using geometric triangular analysis," 2018 International Workshop on Advanced Image Technology (IWAIT), 2018, pp. 1-4, doi: 10.1109/IWAIT.2018.8369775.

[6] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS 1 LSTM: A Search Space Odyssey," doi: 10.1109/TNNLS.2016.2582924.

[7] T. M, D. AJ, M. FW, S. D, and K. A, "Imaging Performance of a Handheld Ultrasound System With Real-Time Computer-Aided Detection of Lumbar Spine Anatomy: A Feasibility Study," Invest. Radiol., vol. 52, no. 8, pp. 447–455, Aug. 2017, doi: 10.1097/RLI.0000000000000361

[8] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device Real-time Body Pose tracking," Jun. 2020, Accessed: Jul. 08, 2021. [Online]. Available: https://arxiv.org/abs/2006.10204v1.

[9] A. Newell, K. Yang, and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9912 LNCS, pp. 483–499, Mar. 2016, Accessed: Jul. 08, 2021. [Online]. Available: https://arxiv.org/abs/1603.06937v2.

[10] A. M. Fred Agarap, "Deep Learning using Rectified Linear Units (ReLU)," Accessed: Jul. 08, 2021. [Online]. Available: https://github.com/AFAgarap/relu-classifier.

[11] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network," Phys. D Nonlinear Phenom., vol. 404, Aug. 2018, doi: 10.1016/j.physd.2019.132306.

[12] C. Lugaresi et al., "MediaPipe: A Framework for Building Perception Pipelines", arXiv [cs.DC]. 2019.

[13] M. W. Whittle, "Gait analysis," in The Soft Tissues, G. R. McLatchie and C. M. E. Lennox, Eds. Oxford, England: Elsevier, 1993, pp. 187–199.

[14] Rathod, N., & Wankhade, S. B. (2020). Improving Extreme Learning Machine Algorithm Through Optimization Technique. In Advanced Computing Technologies and Applications (pp. 157-163). Springer, Singapore.

[15] Joshi, H., Agarwal, V., Ghodke, A., Gupta, D., & Gaikwad, S. (2017). Proposal of chat based automated system for online shopping. American Journal of Neural Networks and Applications, 3(1), 1-4.

[16] S. Gaikwad, "Study on Artificial Intelligence in Healthcare," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021, pp. 1165-1169, doi: 10.1109/ICACCS51430.2021.9441741.

[17] A. Sungheetha, "3D image processing using machine learning based input processing for Man-Machine Interaction," Journal of Innovative Image Processing, vol. 3, no. 1, pp. 1–6, 2021.

[18] Vivekanadam B, "Evaluation of activity monitoring algorithm based on smart approaches," September 2020, vol. 2, no. 3, pp. 175–181, 2020.