

# Integrating gait and speech dynamics methodologies for enhanced stuttering detection across diverse datasets

Ravikiran Reddappa Reddy<sup>1,2</sup>, Santhosh Kumar Gangadharaih<sup>1</sup>

<sup>1</sup>Department of Electronics and Communication, East West College of Engineering, Visvesvaraya Technological University, Belagavi, India

<sup>2</sup>Department of Electronics and Communication Engineering, SJC Institute of Technology, Chickballapur, India

## Article Info

### Article history:

Received Nov 22, 2023

Revised Mar 6, 2024

Accepted Mar 15, 2024

### Keywords:

Adaptive graph topology

convolution network

Healthcare

Neural network

Speech impairment

Speech therapy

## ABSTRACT

Stuttering manifests as involuntary interruptions in the fluency of speech, often involving repetitions, prolongations, or blocks of sounds or syllables. These disruptions can significantly impact effective communication and psychosocial well-being. This research introduces a comprehensive system for speech impairment detection and gait analysis. Speech impairment, with a primary focus on stammer recognition, presents a multifaceted challenge in the field of speech processing. Stammers can manifest in various forms and detecting them accurately is a complex task. Our proposed methodology revolves around the development of StEnsembleNet, a neural network designed to learn spectral features at the frame level, enabling precise and efficient identification of speech impediments. Additionally, we extend our system's capabilities to the domain of gait analysis, leveraging a novel adaptive graph topology convolution network (AGT-ConvNet) for skeletal motion and visually enhanced topological learning to adapt to diverse visual environments and enhance the recognition of gait patterns. This research not only contributes to the field of speech therapy but also offers potential applications in healthcare and motion analysis.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## Corresponding Author:

Ravikiran Reddappa Reddy

Department of Electronics and Communication, East West College of Engineering

Visvesvaraya Technological University

Bangalore 560064, India

Email: ravikiranr\_12@rediffmail.com

## 1. INTRODUCTION

Stuttering is a speech disorder that arises from neurodevelopmental factors and is characterized by the impairment of speech sensorimotor functions. The condition is distinguished by the occurrence of involuntary utterances, interjections (such as the inclusion of sounds like "uh" or "uhm"), and core behaviors (blocks, repetitions, and prolongations, which are atypical elongations of speech sound segments) [1]. Based on research findings, individuals who encounter stuttering, commonly known as people who stammer (PWS), encounter various difficulties in their interpersonal and occupational engagements. Stuttering detection can be utilized to enhance the efficiency of automated speech recognition (ASR) systems for individuals with stuttering (PWS). The purpose of this enhancement is to improve the usability of voice assistants such as Alexa, Siri, and other similar platforms. Additionally, a discernible trend has been observed wherein users are increasingly engaging with voice assistants. However, it is common for these individuals to often overlook and encounter challenges when it comes to comprehending speech that displays stuttering.

The diagnosis of spinal dysraphism typically involves the utilization of specific hearing and brain scan tests. The implementation of the speech disfluency approach incurs significant costs and demands a substantial

commitment of effort from speech therapists [2]. Uncontrolled vocalizations are perceptible manifestations that contribute to the differentiation of various types of stuttering. A machine learning model was employed by several individuals to analyze auditory signals present in speech affected by stuttering. The primary aim of research on everyday behavior is to distinguish and classify distinct behaviors, including but not limited to walking, standing, sitting, running, and various others [3]. The assessment of the impact of action conversion on the performance of the segmentation recognition system holds significance, despite the system not being explicitly designed to handle transitional activities. Moreover, the routine activities performed by individuals often involve multiple shifts between various states of action, presenting a combination of intricacy and continuity. Everyday tasks often involve complex processes, where the complexity is demonstrated by the transitional steps that occur during state changes. The process of gait detection is predominantly dependent on machine learning and computer vision techniques. These techniques are employed to precisely identify individuals by analyzing their walking patterns. The main objective of the multidisciplinary field of gait recognition is to obtain a comprehensive understanding of an individual's behavior by analyzing the gait patterns [4].

This combines the principles and techniques of biomechanics, psychology, and advanced technology to achieve its goal. The assessment of an individual's stride can provide valuable insights into their cognitive and emotional state, physical well-being, and level of fatigue or stress. Contemporary computer vision systems and machine learning algorithms demonstrate the capability to precisely detect and analyze even the subtlest intricacies in motion [5]. The advancements achieved in this field have created new possibilities for diverse applications in the fields of security, healthcare, and psychological well-being. The technology, however, gives rise to privacy concerns as it is susceptible to fluctuations in accuracy caused by various environmental factors. The primary goal of studying gait is to analyze and evaluate movement patterns. It is important to consider the potential for exploring the correlation between interconnected motor patterns, such as gait, and their ability to provide indirect insights into various conditions, including stuttering. This can be accomplished by identifying abnormalities within the motor system. Gait identification and stuttering detection represent two prominent domains known as biometric and behavioral analysis. These domains offer valuable insights into the intricate interplay between physiological and neurological systems. During the first assessment, one might observe a discernible difference between an individual's walking style and the fluency of their spoken communication. Upon conducting a more detailed analysis, it becomes apparent that there exist potential intersections that necessitate additional investigation [6].

The occurrence of stuttering goes beyond mere speech errors, as it can provide valuable insights into an individual's broader neurological and emotional state. The occurrence of the condition often occurs due to emotional triggers, such as stress or anxiety. Just as emotions and neurological disorders have the potential to impact behavior, gait is a complex motor action with multiple facets. This inquiry seeks to determine if atypical walking patterns can serve as an indicator or predictor of stuttering occurrences. The underlying foundation of this relationship is grounded in the acknowledgment that intricate neural networks control both speech and movement. Hence, the identification of any disruption or deviation within any of these domains may indicate potential issues within the remaining domains [2]. The application of a comprehensive methodology that integrates data from speech patterns and gait analysis allows researchers to uncover concealed patterns that may not be immediately apparent when analyzing each domain independently. The integration of a multimodal strategy shows potential in facilitating the progress of innovative medical approaches, comprehensive patient monitoring, and early detection systems. The integration of gait identification and stuttering detection holds significant potential for advancing the understanding of human behavior and the underlying neurological processes that govern it. The ongoing multidisciplinary investigation is anticipated to generate significant insights into these particular domains.

The examination of bodily movements, specifically gait, offers significant insights into an individual's emotional and psychological conditions. The convergence of technology and psychology in the field of gait recognition provides a unique point for understanding human behavior. Stuttering, related to other medical conditions, can be identified through the analysis of speech patterns. One possible method for enhancing the understanding of stuttering involves integrating technologies that detect speech and gait. By analyzing the underlying factors that contribute to vocal abnormalities and the broader spectrum of motor system behaviors, it becomes feasible to acquire a more comprehensive comprehension of this particular condition. The union places significant importance on the fundamental interconnection between the mind and body. It provides not only technological advancements but also adopts a comprehensive approach to enhancing human well-being.

- Comprehensive speech impairment detection: The system is designed to detect various types of speech impediments, including stammers, effectively addressing the complexity of multiple stammer types and providing comprehensive speech impairment detection.

- Spectral learning for accurate detection: The methodology employs StEnsembleNet, a neural network that learns spectral features at the frame level, enhancing the system's ability to accurately identify speech impediments through spectral representations.
- Gait analysis adapted to visual environments: The methodology introduces enhanced topological learning, which enables the model to adapt to different visual environments and enhances its capacity to recognize and analyze gait patterns across varying conditions, improving gait analysis based on skeletal motion data.

## 2. RELATED WORK

The initial investigations in this field focused primarily on various spectrogram-based features, such as linear predictive cepstral coefficients (LPCCs), Mel-frequency cepstral coefficients (MFCCs), and similar characteristics [7]. Furthermore, there has been a notable rise in interest regarding the application of WT (wavelet transform) images for physiological signal processing applications, as evidenced by recent research. The images provide significant insights into the dynamics of time and frequency. The utilization of these methods allows for the acquisition of information regarding the fundamental factors contributing to stuttering, while also offering significant observations regarding the progressive characteristics of the condition as it develops over time. In 2019, Kuppevelt *et al.* [8] conducted a study in which they trained a multi-layer perceptron using 10-dimensional respiratory characteristics, focusing specifically on the block SD [9]. The researchers utilized a dataset consisting of 68 individuals who were Spanish speakers and originated from various countries in latin America to carry out their case study.

The binary deep learning classifier introduced by [10] in their study employs a combination of residual network and bidirectional long short-term memory (ResNet+BiLSTM) to accurately detect and classify six distinct types of disfluencies. Disfluencies can manifest in different forms, including prolongation, word repetition, sound repetition, phrase repetition, and false starts. Promising results were observed in a restricted subset of the University College London's Archive of Stuttered Speech (UCLASS) dataset, which consisted of 25 speakers. These results were achieved by utilizing spectrograms as input features [11]. In a specific study, Wang *et al.* [12] employed the convolutional long short-term memory (ConvLSTM) model to identify and detect six separate categories of stuttering. The categories mentioned consist of blocks, prolongations, sound repetitions, word repetitions, and interjections. The input of the model consists of 41-dimensional phoneme feature vectors, three-dimensional pitch characteristics, eight-dimensional articulatory features, and 40-dimensional MFCCs. The StutterNet classifier was introduced [13].

The classifier utilizes a singular multi-class time delay neural network (TDNN). The classifier can accurately identify and differentiate between core behaviors and fluent speech segments. The performance of the classifier demonstrates promising detection capabilities on a significant subset of UCLASS speakers when compared to state-of-the-art classifiers. The input format accepted by the model is limited to 20-dimensional MFCC. Smith and Weber [14] introduce "TranStutter," an innovative deep-learning model that leverages transformers. The model being presented eliminates the need for convolution and has been specifically designed to showcase outstanding performance in voice disfluency classification. The TranStutter system utilizes multi-head self-attention and positional encoding techniques to effectively capture intricate temporal patterns, distinguishing it from conventional methods. Considering the adoption of gait, there have been numerous works separately on gait recognition, wearable sensors are employed to monitor daily activities and lifestyle [15]. Task-specific machine learning techniques are employed to achieve accurate segmentation of recorded activities. The prioritization of identifying and analyzing essential actions is a critical process. The purposeful development of the segmentation system enables it to accurately identify and differentiate six primary actions: walking, ascending stairs, descending stairs, sitting, standing, and lying down. The actions mentioned are precisely detected and separated from a continuous signal. The initial stage of the process involves the segmentation of continuous signals. The next step involves the identification of transitional activities. The researchers employed the greedy Gaussian segmentation (GGS) technique to detect discontinuities that occur during action transitions. The XGBoost model was utilized to predict human activities in each frame.

The authors did not address the identification of transitional acts. The authors of the study have introduced a novel approach [16] to segment and identify continuous action sequences. The sequence can be classified into three primary categories: transitional activities, static acts, and dynamic actions. This study introduces a novel methodology for the identification of the gait cycle utilizing biometric data. The primary objective of this method is to accurately identify and categorize all dynamic activities that take place in a specified sequence. The achievement was accomplished by analyzing the characteristics of the gait signal during dynamic activities. The primary objective of the analysis was to assess the distribution of fluctuation points in two distinct types of activities, taking into account the disparity in signal conversion frequency between static actions and transition actions. The GaitParsing framework is a specialized methodology developed to study human semantic parsing and gait detection [17]. The main goal is to methodically examine the human body and categorize it into separate anatomical elements that have well-defined and comprehensive

structures. To fulfill this requirement, a dual-branch feature extraction network is employed. The network has been intentionally designed to efficiently handle both whole-body gait and individual body parts. One potential method for assessing self-occlusion in a gait sequence involves the use of a technique called self-occlusion frame evaluation. The main objective of this evaluation is to enhance the efficiency of gait frames that exhibit significant variations.

This paper presents a novel methodology for improving the functionality of human gesture and activity recognition (HGAR) [18], [19]. The proposed methodology involves using a newly developed smart belt to collect posture data. In addition, the process employs an advanced stacking design. The effectiveness of the intelligent framework outlined in this paper was evaluated through a series of experiments. The experiments involved the integration of multiple sensors to develop a holistic solution for the HGAR problem.

### 3. PROPOSED METHODOLOGY

We develop and design a system that is utilized to detect and understand people having challenges while speaking, specifically various kinds of stammers. The traditional method to resolve this issue is the development of various task classifiers, this further becomes complex as for every input the sound snippet could have various kinds of stammers. Considering the scenario and the scarcity of stammer data information, speech disruption analyzers are used to aid by decreasing the difficulty of every binary work. The main focus is to develop a specific system architecture that is trained specifically for the recognition and detection of various kinds of stammers for every trained situation, wherein collectively it can recognize various stammers and stutters. The proposed system (PS) aims to develop a network that has the ability of spectral learning at the frame stage expressions and temporal links. However, the system has to emphasize prominent regions of input for efficient learning of speech impediments as well as accurate performance. The proposed methodology also focuses on gait detection based on skeletal movements. It refers to a technique that involves studying the manner and patterns of an individual's walking. The proposed adaptive graph topology convolution network (AGT-ConvNet) for skeletal motion with the joint-focused learning filter enhances the gait profile by considering various visual environments. The proposed enhanced topological learning improves the graphical structures that include links between the joints, it is developed to aid a combined aggregation for graph-based convolutions in the next phase of the architecture.

#### 3.1. StEnsembleNet: Stuttering detection through ensemble deep learning architecture

The study proposes a network that utilizes spectrograms (MFCC) of sound snippet inputs, this network is termed as StEnsembleNet. Spectrograms are used in previous methods for the speech process, and ResNet is used for learning spectral frame stage expressions. A bidirectional long short-term memory is utilized when stammer variations have specific temporal as well as spectral features. Using classification of stammering as well as automatic speech detection is used to boost the performance of the proposed network.

In the spectrogram, colors depict amplitude for every filter at a particular frame, fewer amplitudes are depicted with blue and the colors yellow as well as green show increased amplitudes. The proposed StEnsembleNet aims model to learn efficient depictions for every spectrogram as input. In this case, convolutional neural networks are used along with excitation and squeeze models that enhance prior methods in various domains. The architecture of the ResNet is shown to have a resolution to rising gradient issues, these are challenges that arise during backpropagation. Normally, the rise in the depth of the model causes the weight gradient to reduce for every layer. Eventually, this could stop the gradients from altering or enlarging weights. Therefore, this prevents modifications in the model. The ResNet model resolves this issue by using shortcuts in the convolutional neural network segments that lead to the preservation of segments that can move the gradients for deep learning models.

Previous technologies used in this field such as excitation and squeeze are used for the classification of images that decrease the error. Each kernel in the layer of a convolutional neural network leads to an extra channel in the attribute map output. The excitation and squeeze segments increase the concentration of channel-based links in the convolutional neural network. These segments have two main processes. The collection of attribute maps for the length and width is performed by a squeezing process that results in a channel descriptor that has one dimension. The completely linked layers give channel-based lengths in the excitation process that is implemented in the initial attribute map. The architectures of ResNet as well as excitation and squeeze are utilized for effective learning of the spectral representations at the frame stage at the input, the proposed model consists of a ResNet-excitation and squeeze model. The model has eight segments of ResNet-excitation and squeeze. For every segment in the StEnsembleNet, there exists a convolutions layer of two-dimension size in batches of three, such that every layer is a result of rectified linear unit (ReLU) as well as batch normalization. An individual residual link has similar input as the segment's branch that is added before the ReLU activation function, however, it is after the excitation and squeeze. For every link, there exists

a layer of convolution after which batch normalization occurs. Global pooling is used for excitation and squeeze in the proposed model. The result is later made to enter two completely linked layers, the first layer has the activation ReLU function, followed by the sigmoid gait function in the next layer. The output is used for the evaluation of the major convolution layer utilizing channel-based multiplication.

The temporal connection must be learned by the model between depictions learned from spectrograms, hence in this case we use recurrent neural networks. Every long short-term memory has a memory cell that contains data that was present in prior cells that permits the model to learn and understand temporal links. The long short-term memory has the memory cell as a component that has various gates that collectively manage input data and the prior cell information that is used in the generation of the present cell as well as concealed states. Specifically, the input gate is denoted by  $j_u$ , the forget gate is expressed as  $g_u$  that is used in learning the data from every cell that will be stored in the present new cell is denoted as  $D_u$ . This is expressed in the (1) to (3).

$$g_u = \vartheta(X_g \cdot [i_{u-1}, y_u] + c_g) \quad (1)$$

$$j_u = \vartheta(X_j \cdot [i_{u-1}, y_u] + c_j) \quad (2)$$

$$D_u = j_u * \tanh \tanh (X_D \cdot [i_{u-1}, y_u] + c_D) + g_u * D_{u-1} \quad (3)$$

$$p_u = \vartheta(X_p \cdot [i_{u-1}, y_u] + c_p) \quad (4)$$

$$(D_u) * p_u \quad (5)$$

Considering the (1) to (3), the sigmoid function is represented as  $\vartheta$ , and a point-based multiplication is represented using the operator  $*$ . The present cell with  $p_u$  which is the output gate utilized for the generation of the concealed cell in the unit  $i_u$  that is formulated in the equation given (4) and (5). The memory cell as well as the concealed cell is moved to the next long short-term memory unit, this encourages the model in learning dependencies for the long term. The proposed model StEnsembleNet uses a bidirectional long short-term memory that has two types of long short-term memory that are contrary in direction, expanding the information from links of the future as well as the past. The multiplied outputs of the networks result in a single particular output. The proposed model has two bidirectional long short-term memory that are consecutive, where every cell has 500 concealed cells. Considering the dimension of the dataset, we avoid any issues of overfitting by masking the neurons produced for training the model where a sparse expression of the information is produced. The use of attention techniques as well as its diversity permits more emphasis on important parts of embedding. The techniques mentioned have been implemented recently in the domain of speech detection to focus on higher emotional features of stammers and are also utilized to enhance StEnsembleNet to emphasize particular regions of stammers that do not have fluent features. The proposed method StEnsembleNet utilizes global attention models that use the values of concealed cells.

### 3.2. Adaptive graph topology convolution network

To analyze the gait recognition, this research work presents an AGT-ConvNet to model the gait recognition across the dataset. The value of the final result in layer two of the bidirectional long short-term memory is given as  $i_u$  and  $D_u$  is used to express the contextual vector that is extracted by the attention model that generates the last classification of the proposed model which is denoted as  $\tilde{f}_u$ . The formulation is as given using the hyperbola tan function as given in (6).

$$\tilde{f}_u = \tanh \tanh (X_d(D_u; i_u)) \quad (6)$$

The contextual vector for the attention method has the added weights of all the concealed cell outputs of the encoder. A vector is used as a link for  $i_u$  as well as all the values of the concealed cells that are moved into a layer with softmax to depict the weights. This function uses the dot product in the attention method. This is formulated as given in the (7).

$$D_u = \sum_{j=1}^u i_j \left( \left( f_{i_u, j}^{i_u, j} \right) \left( \sum_{j'=1}^u f_{i_u, j'}^{i_u, j'} \right)^{-1} \right) \quad (7)$$

In the (6), the  $j$  – th bidirectional long short-term memory is denoted as  $i_j$ . The gait profile is enhanced using dual-split architecture of AGT-ConvNet for skeletal motion data  $J$  is considered as input information. The first phase of the architecture is based on the visual environments as well as sequential attributes. Visual-enhanced topological learning is introduced dynamically and denoted as  $H_{WB}$ . The enhanced

topological learning developed is used to aid a combined aggregation for graph-based convolutions in the next phase of the architecture. Cross-entropy error given as  $M_{DF}^{visual}$  is used to examine visual characteristic learning.

The second phase of the proposed architecture consists of vague skeletal features that are firstly obtained using a lightweight module which is a normal network of graphical convolution blocks that is expressed in (1). Further, the blocks of the AGT-ConvNet are gathered to filter the characteristic features, and specific clues are extracted. For every segment, a joint-focused learning filter is implemented to produce filters  $G_T$  and  $G_U$  considering the depth. Specifically,  $G_T$  along with  $H_{WB}$  from the visual enhanced topological learning is utilized for extraction spatially and  $G_U$  is utilized for the temporal system. Convolutions of dimensions 1 by 1 are implemented for interlinked channel data. However, the network of the graphical convolution process is given in the (8).

$$\begin{aligned} Y_T &= \text{Convolution}_{1 \text{ by } 1}(X_1, \sum_{l=1}^{L_T} H_{WB}^l(Y \otimes G_T^l)) \\ Y_U &= \text{Convolution}_{1 \text{ by } 1}(X_2, Y_T \otimes G_U) \end{aligned} \quad (8)$$

Considering the equations convolution based on depth is represented as  $\otimes$ ,  $H_{WB}^l$  belongs to  $S^{0 \times 0}$ ,  $G_T^l$  belongs to  $S^{0 \times D}$ ,  $G_U$  belongs to  $S^{L_U \times 0 \times D'}$ ,  $X_1$  belongs to  $S^{D \times D'}$ , and  $X_2$  belongs to  $S^{D' \times D'}$ . After the blocks of the AGT-ConvNet are processed, joint connection hierarchical mapping is implemented for the gait characteristic features to be mapped into six individual scales that are based on joint-linked architecture. There is a loss of three elements denoted as  $M_{three}$  and a circular loss denoted as  $M_{circle}$  which is implemented in the features in the output for training monitoring. The entire loss error function is calculated in (9).

$$M = \partial_1 M_{three} + \partial_2 M_{circle} + \partial_3 M_{DF}^{visual} \quad (9)$$

The (9) shows that loss error functions are balanced by hyperparameters  $\partial_1, \partial_2, \partial_3$ . Furthermore, features of the joint as well as the bone are combined for networks to identify the activities of humans, two distinct stages are used to separately extract features of the bone and the joint and the results of both are combined. The focus of features of bone is similar to that of joints except the input omits the coordinates value of adjacent joints. The first phase data is the features from joints and the second phase data is the features from bones. To simplify, the features of both phases are combined with the dimension of 12 by  $D_{output}$  and is implemented as the output.

### 3.3. Adaptive graph topology convolution network learning

It is observed that various parts of the body show various extent of difference as well as level of freedom considering the skeletal structure, the joint focused learning filter is utilized to explain the separate characteristics that are temporal and spatial for various gait series by developing custom filters. Here, two distinct phases form the temporal and spatial filters that are used for the extraction of spatial features as well as the acquisition of temporal features. The efficiency is expanded due to the use of depth-based filters.

The gait characteristics  $Y$  belongs to  $S^{U \times 0 \times D}$  is utilized for input and the enhanced temporal pooling is used to gain the output  $Y_q$  belongs to  $S^{U_q \times 0 \times D}$ , in this case  $U_q$  expresses the pooling dimension. Considering temporal as well as spatial phases, at first, we use convolutions that are temporal for learning context-based data at every joint. Furthermore, pooling that is temporal is applied in the spatial phase to gather global temporal data. Later, two fully linked layers along with a layer of batch normalization and an activation function ReLU are utilized for the construction of inter-linked channel communication and result in a filter  $g_t$  that has dimensions 1 by 0 by  $(L_t \text{ by } D)$ . Similarly, the filter is reshaped into dimensions  $L_t$  by 0 by  $D$  that uses batch normalization that omits small or large filter attributes to be used. The procedure that occurs at the spatial phase is evaluated using the (10).

$$\begin{aligned} g_T &= G(G(UQ(UD(Y_q)), X_3), X_4) \\ G_T &= \text{Batch Normalization(Reshaping}(g_t)) \end{aligned} \quad (10)$$

In (3), UD is used to denote convolution that is temporal, a completely linked layer is expressed as  $G$ , and UQ is used to express temporal pooling. Here,  $X_3$  belongs to  $S^{D \times \frac{D}{s}}$  that is used to decrease the size of the channel by  $s$  ratio and  $X_4$  belongs to  $S^{\frac{D}{s} \times D}$  gains back the size of the channel. For various spatial phases, the temporal part is utilized to explain the moving features that exclude the temporal pooling to retain a temporal architecture as well as implement two completely linked layers that have activation function ReLU for temporal size. The main aim is to achieve effective temporal links for various motions to investigate moving features. Furthermore, a process of normalization is implemented to guarantee that the feature distribution is stable. Finally, the process for the temporal phase is given as (11).

$$\begin{aligned} g_U &= G(G(UD(Y_q), X_5), X_6) \\ G_U &= \text{Batch Normalization}(g_U) \end{aligned} \quad (11)$$

Considering the (11),  $X_5$  belongs to  $S^{U_Q \times (\beta \times U_Q)}$  expands the temporal size by  $\beta$  ratio and  $X_6$  belongs to  $S^{(\beta \times U_Q) \times L_U}$  decreases it to a particular dimension  $L_U$ . There exist a few methodologies of gait on a visual basis that utilize various stage approaches to develop the local attributes of the model. These techniques show some similarities to joint focused learning filter. The joint-focused learning filter is compared to the various stage techniques, which yields the following results: i) the different stage technique extracts attributes of various sequences that use constant convolutions, in this case, joint joint-focused learning filter the attributes extracted are for every sequence; ii) the various stage technique utilizes convolutions that cannot be shared, where the attributes used are bigger than the convolutions. Whereas, for joint focused learning filter saves mostly half of the attributes and makes it more efficient; and iii) the various stage technique attains the features by partitioning manually, wherein the semantics alignment is not efficient. However, the joint-focused learning filter has features that are aligned well using inputs from the skeleton.

### 3.4. Visual feature extraction

The visual enhanced topological learning method uses inherent visual data for every sequence for learning to adapt to visual environments. Consider  $J$  as the initial gait sequence that is attained through the input. Initially, the method is applied for the extraction of visual features, after which global mean pooling is used to combine the global visual data. Later, a completely linked layer is used to gain the visual classifying vector  $g_w$  belongs to  $S^{L_w}$ , the visual count is denoted using  $L_w$ . In this case, the cross-entropy error for  $g_w$  results in a loss  $M_{DF}^{visual}$  that is utilized to monitor feature attribute learning. This assures effective visual prediction. Similarly, the function SoftMax is used to result in a normalized value vector  $\widetilde{g}_w$  belongs to  $S^{L_w}$ . A set of topologies that require to be learned are denoted as  $H_{set} = \{H_W^1, H_W^2, \dots, H_W^{L_w}\}$  that expresses the modification of visual enhanced ability, where  $H_W^j$  belongs to  $S^{L_T \times O \times O}$  represents the related topology gained at the  $j$  – th visual. For the generation of visually enhanced topology  $H_{WB}$ , the highest value index to be gained in  $g_w$ , which is given as (12). In the (5),  $ie_w$  is used to denote the sequence that has the highest encounters. Further, the related topology is chosen from that  $H_{set}$  by  $ie_w$  as given in (13).

$$ie_w = \text{highest arg arg } \widetilde{g}_w \quad (12)$$

$$H_1 = H_{set}[ie_w] \quad (13)$$

Considering (6), specific features are denoted using  $H_1$  to corresponding visuals. For the testing datasets, level-1 visual classification with high accuracies of 98% is shown to be reliable. Hence, for every sequence, known as the result of visual classification prediction, an accurate choice can be made for enhanced topology for the related visual. We assume that the initialization of learning attributes is done through topology, and updates can be made to modify the visual characteristics using the method of backward propagation. Although,  $H_1$  is not enough to express all kinds of intra-distinct existence for this visual, therefore an augmentation is introduced. We assume the distribution of information in  $\widetilde{g}_w$  depicts the sequential features, considering the information is in terms of linear weights for combination with topology for  $H_{set}$ . This is equated as given in (14).

$$H_2 = \sum_{j=1}^{L_w} \widetilde{g}_w^j H_W^j \quad (14)$$

$$H_{WB} = h_1 H_1 + h_2 H_2 + h_3 H_3 \quad (15)$$

Further,  $H_3$  is the constant topology that is used for the extraction of normal features. Finally,  $H_1, H_2, H_3$  is combined with  $h_1, h_2, h_3$  as the coefficients that result in  $H_{WB}$  given in (15). The joints of the body are linked dynamically using convolutions in the graph that incorporate data about the joints for a distant range as well as have a joint correlation with the locations. The links are dynamic as well as complicated which are modified effectively to enhance the capacity of cross-visual environments.

## 4. PERFORMANCE EVALUATION

The performance evaluation encompasses the Chinese Academy of Sciences Institute of Automation (CASIA-B) and Osaka University multi-view large population (OU-MVLP) datasets for gait recognition. Various metrics are utilized to assess the performance of gait recognition methods across different viewing angles. In the speech emotion recognition domain, a wide array of methods, including class imbalance handling approaches, are evaluated on the SEP-28k (Stuttering events in podcasts) dataset, considering metrics such as P, R, in, F1-Score, B, and total accuracy. These evaluations are compared with the state-of-the-art techniques

of the PS in addressing the unique challenges posed by gait and speech recognition tasks and the results are shown in the form of graphs and tables.

#### 4.1. Dataset details

**SEP-28k [6]:** The selection process involved identifying and including a total of 385 podcasts to form the SEP-28k stuttering dataset. The duration of the initial podcast recordings exhibits variability. The podcast adheres to a sequential framework, comprising a diverse range of segments that can span from 40 to 250 in number. Each segment is characterized by a predetermined duration of three seconds. The total number of segments obtained is 28,177. The SEP-28k dataset consists of two distinct label categories: stuttering and non-stuttering. The speech categories that do not demonstrate stuttering and are not pertinent to our study encompass incoherent speech, uncertain speech, absence of speech, low audio quality, and the presence of music. Stuttering is a speech disorder that encompasses various speech behaviors such as blocks, prolongations, repetitions, interjections, and fluent segments. The dataset comprises various speech types, which are as follows: 10.35 hours of fluent speech, 3.34 hours of interjections, 2.74 hours of repetition, 1.48 hours of prolongation, 1.75 hours of block, and an additional 1.48 hours of prolongation. After the completion of the tagging process, each voice fragment, which has a duration of 3 seconds, is subjected to downsampling. The downsampling operation reduces the frequency of the voice fragment to 16 kHz.

**CASIA-B [20]:** The CASIA-B dataset is employed for both the training and testing phases of the experiment in this study. The CASIA-B database, which is a comprehensive multi-view gait database, was compiled and initially released to the public in 2005. The study involved a cohort of 124 participants, with gait data being collected from 11 different perspectives. The perspectives were established by utilizing angles that spanned from 0 to 180 degrees, with a consistent interval of 18 degrees between each perspective. The CASIA-B system categorizes walking into three distinct scenarios: normal walking (NM), walking while carrying a bag (BG), and walking while wearing a coat or jacket (CL). The frame rate of the CASIA-B dataset is 25 frames per second (FPS). The dataset comprises a cumulative count of 854,000 training frames and 798,000 testing frames. The gait recognition methodology utilized in this study has undergone training and evaluation using the CASIA-B dataset.

**OU-MVLP [21]:** The OU-MVLP dataset is a well-established benchmark utilized for gait identification, widely acknowledged within the field. The dataset is the largest in terms of public availability, encompassing a total of 10,307 distinct participants. The dataset comprises a total of 14 perspectives, spanning from 0 to 90 degrees and from 180 to 270 degrees. Each view is linked to two sequences, specifically identified as 00 and 01. There are a total of 5,153 individuals employed specifically for testing purposes, with the remaining 5,154 subjects allocated for training. The sequences indexed as #00 and #01 are explicitly assigned as galleries and probes, respectively.

#### 4.2. Results for gait on CASIA-B and OU-MVLP dataset

The data in Figure 1 showcases a detailed analysis of different gait recognition methods at various orientation angles, each corresponding to a degree from 0° to 180°, the performance comparison of various gait recognition methods based on different metrics is carried out. The PS achieves remarkably high scores, consistently exceeding 99% in most metrics, which indicates its exceptional performance in gait recognition. Other methods like BiFusion [22] also perform well but fall slightly short in comparison with the PS. The results suggest that the proposed approach demonstrates superior effectiveness and accuracy in gait recognition compared to the other methods. In comparison with the existing system (ES) the accuracy value ranges from 96 to 99 in comparison with the PS which generates a higher value for various setting values.

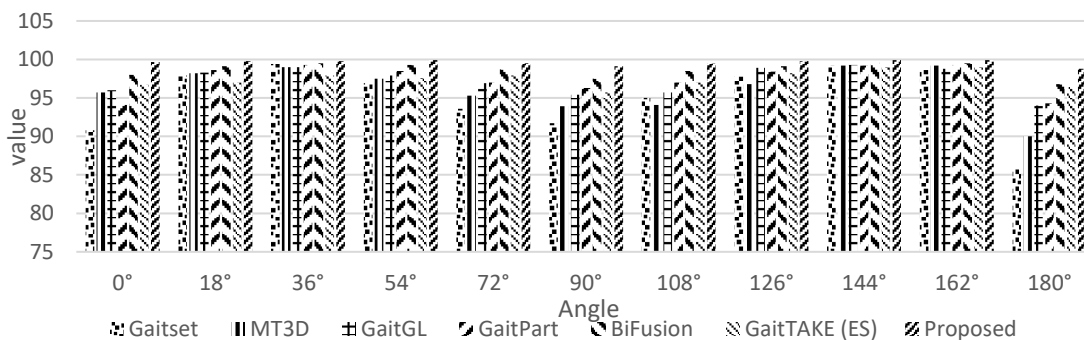


Figure 1. Accuracy comparison of state-of-art techniques with PS for NM#5-6



The provided data presents the mean performance of different gait recognition methods. In Figure 2 each method is compared along with the PS for the respective mean recognition score. The PS gives better performance depicting a mean score of 99.6. BiFusion [22] also demonstrates strong performance with a mean score of 98.7. Other methods like GaitGL [23], GaitPart [24], GaitTAKE (ES) [25], modular three-dimensional multispecies (MT3D) [26], and Gaitset [27] exhibit mean scores ranging from 95.7 to 97.5, reflecting their respectable recognition capabilities.

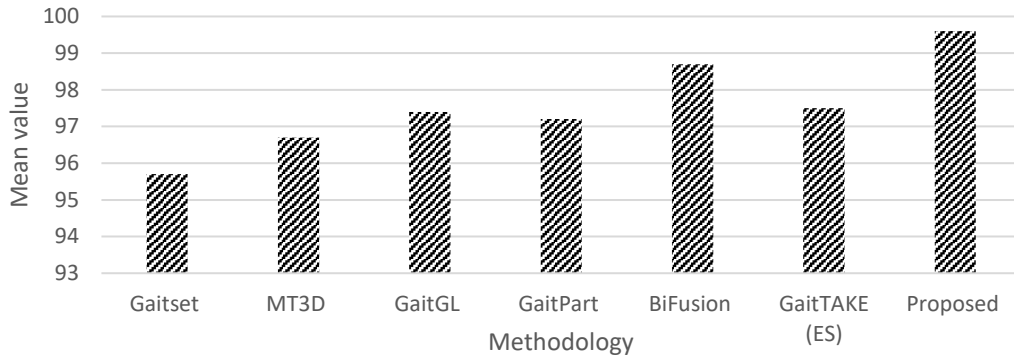


Figure 2. Mean value comparison for NM#5-6

In Figure 3 the results are displayed for BG#1-2, the PS consistently ensures better performance, achieving the highest recognition scores across most angles. This highlights its remarkable consistency and effectiveness in recognizing gait patterns, with scores reaching as high as 99.84%. BiFusion [22] also demonstrates strong performance, particularly at 0°, 18°, and 54° angles, with recognition scores above 97%. Meanwhile, GaitTAKE (ES) [25] delivers better results, especially at 36°, 54°, and 72° angles, with scores peaking at 99%. Other methods like MT3D [26], GaitGL [23], Gaitset [27], and GaitPart [24] offer varying performance, ranging from 84.9% to 98.8%. In conclusion, the PS gives better performance in comparison with the ES.

Figure 4 ensures the summary of the performance mean across various gait recognition methods, PS ensures best performance with a mean score of 98.86. BiFusion [22] also exhibits strong performance with a mean score of 96, signifying its effectiveness in gait recognition. GaitTAKE (ES) [25] follows closely with a mean score of 97.5, emphasizing its reliability in recognizing gait patterns. Methods such as MT3D [26] and GaitGL [27] demonstrate consistent performance with mean scores of 94.5, underscoring their suitability for gait recognition tasks. Gaitset [27] and GaitPart [24] deliver mean scores of 93 and 91.5, respectively. This data indicates that the PS stands out as the most effective gait recognition approach, with the highest mean score, suggesting its strong potential for applications in biometrics, security, or surveillance.

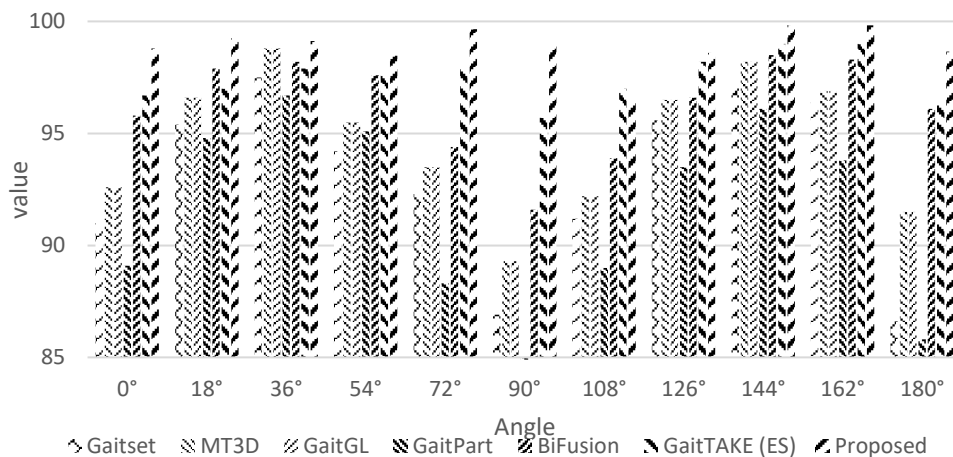


Figure 3. Accuracy comparison of state-of-art techniques with PS for BG#1-2

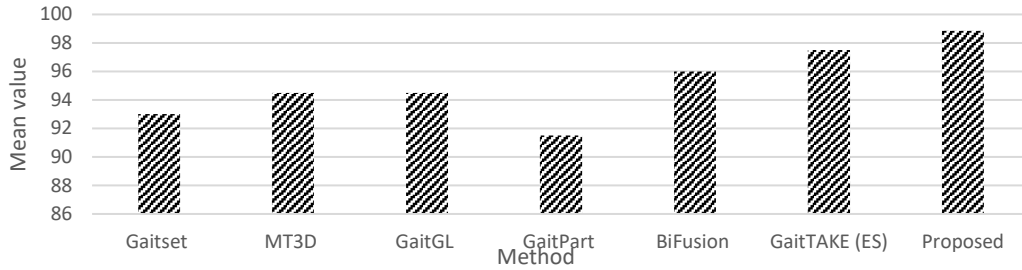


Figure 4. Mean value comparison for BG#1-2

In Figure 5 the analysis is carried out for CL#1-2 which shows that the PS consistently achieves the highest recognition rates across different angles, with values well above 90%. In comparison, some other methods, such as Gaitset [27] and MT3D [26], exhibit similar performance, maintaining recognition rates around 80% at most angles. GaitGL [23] demonstrates slightly better accuracy, with recognition rates exceeding 90% at specific angles. GaitPart [24] lags behind the rest, with recognition rates ranging from 66.5% to 86.9%. BiFusion [22] performs in an average manner, surpassing 95% accuracy at several angles. GaitTAKE (ES) [25] also achieves strong recognition rates, with a dip to 87% at 180°. In conclusion, the PS outperforms other methods, showcasing higher accuracy and robustness in gait recognition across various viewing angles.

Figure 6 presents a comparison of various gait recognition methods, focusing on their overall performance for the mean score. The PS showcases a mean recognition rate of 95.37%. It demonstrates the highest accuracy among all methods, indicating its effectiveness in gait recognition. GaitTAKE (ES) [25] and BiFusion [22] also perform exceptionally well, achieving recognition rates of 92.2% and 92.1%, respectively. GaitGL [23] follows closely with a recognition rate of 83.6%, surpassing Gaitset [27] and MT3D [26], which both exhibit a mean recognition rate of 81.5%. GaitPart [26] lags, with a mean recognition rate of 78.7%. The results suggest that the PS offers a significant advancement in gait recognition, outperforming its competitors in terms of accuracy and reliability.

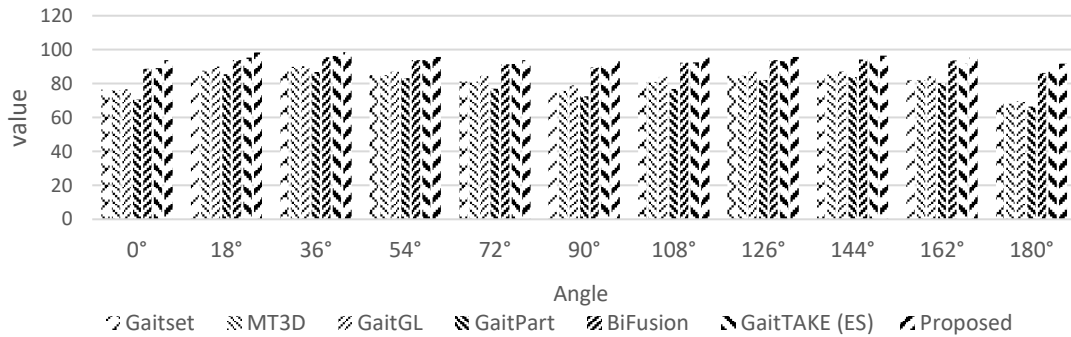


Figure 5. Accuracy comparison of state-of-art techniques with PS for CL#1-2

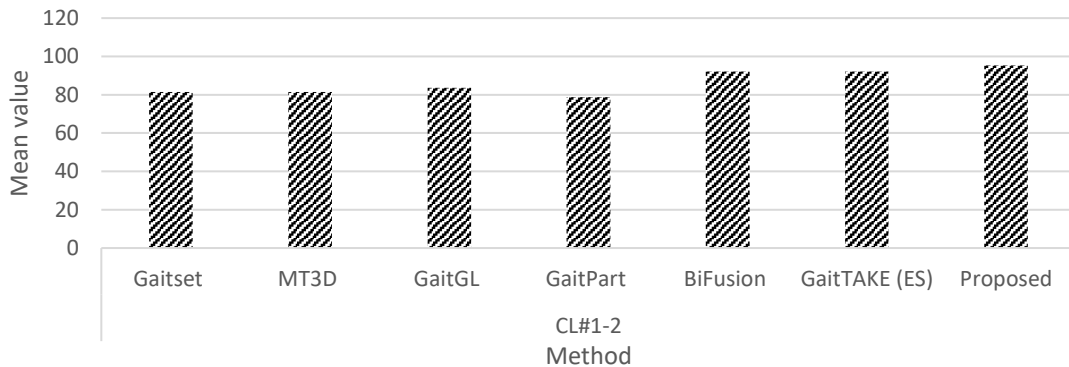


Figure 6. Mean value comparison for CL#1-2

**4.3. Results on OU-MVLP dataset for gait recognition**

The provided data presents a comparison of various gait recognition methods across different pose angles. The PS exhibits outstanding performance, consistently achieving the highest accuracy scores across various angles. It reaches an accuracy of 94.87% at a pose angle of 270°, showcasing its robustness and effectiveness. GaitTAKE (ES) [25] and BiFusion [22] also demonstrate strong performance, with high accuracy scores that remain competitive across various angles. In comparison, methods like GEINet [28], Gaitset [27], GaitPart [24], and GaitGL [23] exhibit lower accuracy scores, indicating their limitations in handling variations in gait poses. Figure 7 shows the accuracy comparison of state-of-art techniques with PS for the OU-MVLP dataset.

Figure 8 showcases a comparison of gait recognition methods in terms of their mean accuracy scores. The PS depicts a mean accuracy score of 94.36%, indicating greater performance in gait recognition. Following closely are GaitTAKE (ES) [25], BiFusion [22], GaitGL [23], and GaitPart [24], all of which exhibit strong mean accuracy scores ranging from 89.7% to 90.4%. Gaitset [27] also performs relatively well with a mean accuracy score of 87.1%. However, GEINet [28] lags with a considerably lower mean accuracy score of 35.8%.

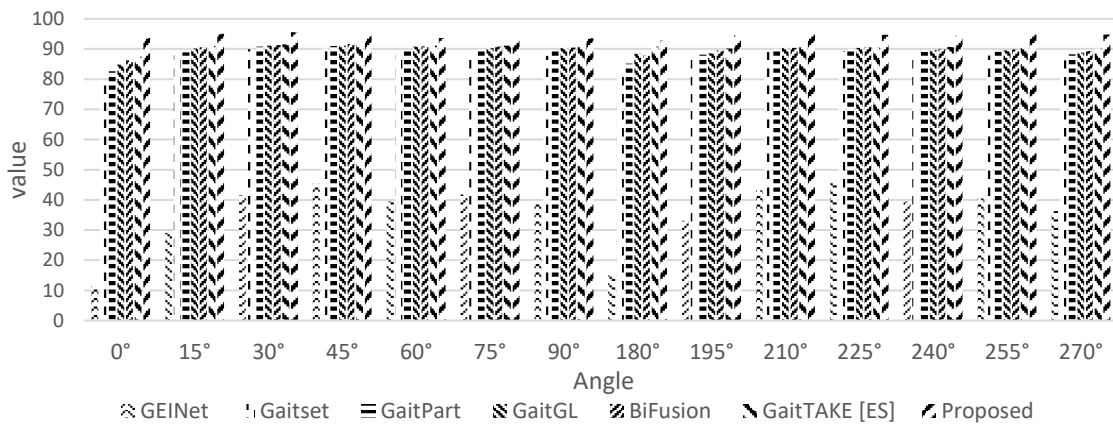


Figure 7. Accuracy comparison of state-of-art techniques with PS for OU-MVLP dataset

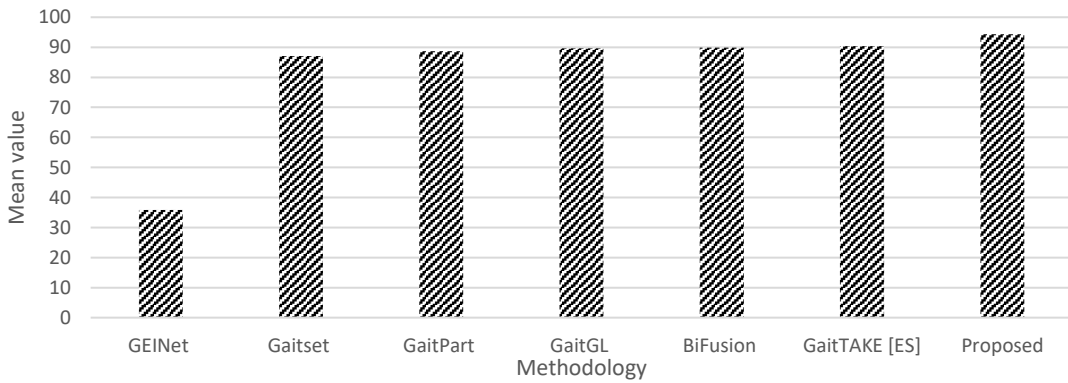


Figure 8. Mean value comparison for OU-MVLP

**4.4. Results for speech**

The data in Table 1 showcases the performance metrics of various speech emotion recognition methods. The analysis is carried out on various methods as ConvLSTM+FMFCC(BL1), ConvLSTM+Fphone (BL2), ConvLSTM+F0+MFCC(BL3), ResNet+BiLSTM (BL4) and ShutterNet (BL5). For class imbalance, ResNet+BiLSTM+WCE, MB ResNet+BiLSTM, StutterNet+WCE (StutterNet/WCE), MB StutterNet (StutterNet/MB), Mfrzenc, Mfrzenc, disf, Mfrzenc, fluent methods are evaluated for class imbalance methods. Notably, the PS exhibits better performance in terms of most metrics. It demonstrates that P, R, B, In, F1-Score, and accuracy with values of 74.02, 68.93, 60.24, 64.25, and 74.13, respectively. These results signify that PS performs better in recognizing and classifying speech emotions. Methods designed to handle class imbalance show improved performance, but the PS consistently outperforms across various metrics. Figure 9

shows the total accuracy for various state-of-art comparisons with PS for the SEP-28k dataset. Figure 10 shows the F1-score comparison for various state-of-art comparisons with PS for the SEP-28k dataset.

Table 1. Performance metrics of various speech emotion recognition methods

Method	R	P	B	In	F	TA	F1(%)
Baselines							
ConvLSTM + FMFCC (BL1) [6]	22.83	10.61	6.34	56.74	72.35	52.68	34
ConvLSTM + Fphone (BL2) [6]]	10.18	1.06	0.35	43.88	74.48	48.43	24
ConvLSTM + F0+MFCC (BL3) [6]	19.28	9.55	8.51	51.78	66.6	48.47	30.8
ResNet+BiLSTM (BL4) [5]	18.76	41.24	5.47	57.11	88.19	62.36	43.12
StutterNet (BL5) [2]	27.14	32.55	2.96	57.74	87.6	62.57	42.84
Class Imbalance							
ResNet+BiLSTM + WCE [5]	28.9	64.89	33.79	63.03	46.9	47.42	41
MB ResNet+BiLSTM [5]	34.79	30.19	5.92	49.26	75.47	55.62	39.2
StutterNet + WCE (StutterNet/WCE)	36.23	59.73	38.05	61.19	41.59	45.26	41.02
MB StutterNet (StutterNet/MB)	35.26	32.76	7.21	56.04	77.27	58.56	42.26
Mfrzenc	39.82	37.91	10.45	60.57	73.49	58.58	44.42
Mfrzenc, disf	29.25	45.85	18.11	56.88	74.49	58.18	44.8
Mfrzenc, fluent	31.15	27.62	5.01	57.64	73.64	55.83	38.6
Proposed	68.93	74.02	60.24	58.16	74.13	67.096	64.25

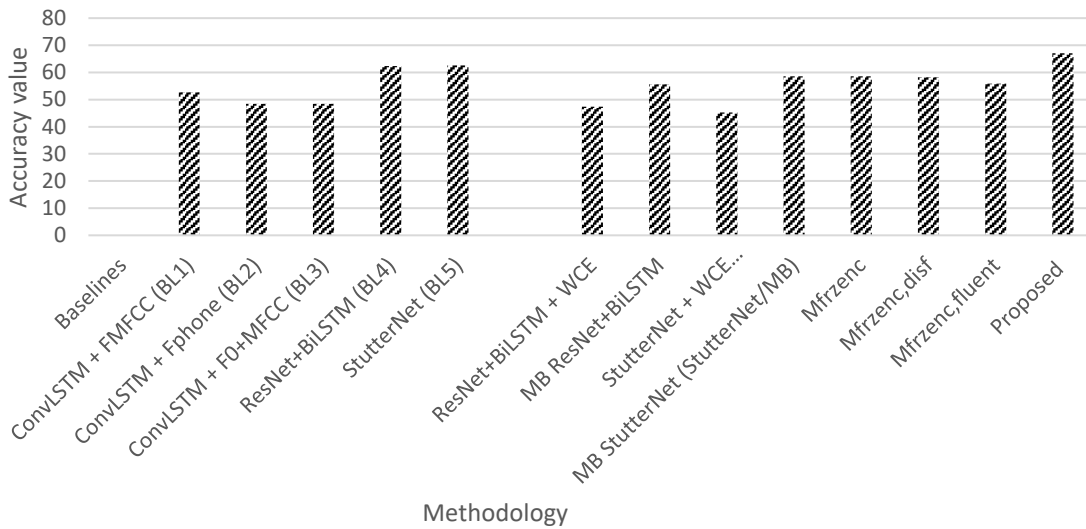


Figure 9. Total accuracy for various state-of-art comparisons with PS for the SEP-28k dataset

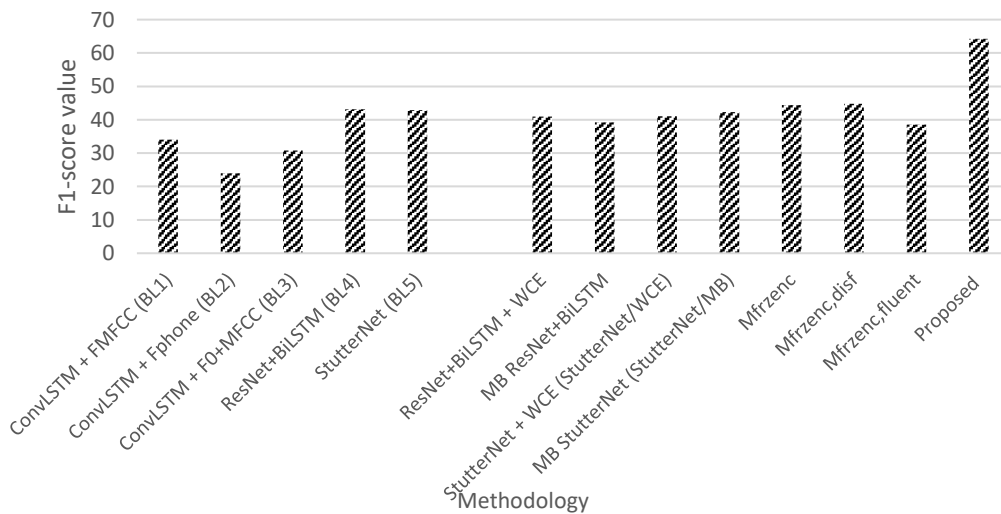


Figure 10. F1-score comparison for various state-of-art comparison with PS for Sep-28k dataset

## 5. CONCLUSION

In conclusion, the proposed multimodal deep learning approach for speech impairment detection and gait analysis offers a promising path toward enhancing healthcare and communication technologies. Through StEnsembleNet, remarkable accuracy is accomplished in identifying various stammer patterns and automatically detecting speech impairments. Simultaneously, the AGT-ConvNet for skeletal motion, augmented with visual enhanced topological learning, provides robust and adaptable gait analysis, offering a comprehensive view of human movement patterns. By uniting these two critical areas, our research breaks new ground in understanding and addressing speech disorders and gait-related health concerns. This synergy between healthcare and technology not only empowers diagnostic tools but also opens doors to innovative assistive devices, improving the quality of life for individuals with speech impairments and facilitating the study of human motion. This work signifies a step forward in bridging the gap between the healthcare and technology sectors for combining the gait phenome with speech to detect stuttering, ultimately benefiting a wide range of applications and individuals.




## REFERENCES

- [1] S. Oue, R. Marxer, and F. Rudzicz, "Automatic dysfluency detection in dysarthric speech using deep belief networks," in *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2015, pp. 60–64, doi: 10.18653/v1/W15-5111.
- [2] S. A. Sheikh, M. Sahidullah, F. Hirsch, and S. Ouni, "StutterNet: Stuttering detection using time delay neural network," in *2021 29th European Signal Processing Conference (EUSIPCO)*, IEEE, Aug. 2021, pp. 426–430, doi: 10.23919/EUSIPCO54536.2021.9616063.
- [3] K. Qian *et al.*, "A bag of wavelet features for snore sound classification," *Annals of Biomedical Engineering*, vol. 47, no. 4, pp. 1000–1011, 2019, doi: 10.1007/s10439-019-02217-0.
- [4] B. Villegas, K. M. Flores, K. J. Acuna, K. P. -Barrios, and D. Elias, "A novel stuttering disfluency classification system based on respiratory biosignals," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, Jul. 2019, pp. 4660–4663, doi: 10.1109/EMBC.2019.8857891.
- [5] T. Kourkounakis, A. Hajavi, and A. Etamad, "Detecting multiple speech disfluencies using a deep residual network with bidirectional long short-term memory," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2020, pp. 6089–6093, doi: 10.1109/ICASSP40776.2020.9053893.
- [6] C. Lea, V. Mitra, A. Joshi, S. Kajarekar, and J. P. Bigham, "SEP-28k: A dataset for stuttering event detection from podcasts with people who stutter," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Jun. 2021, pp. 6798–6802, doi: 10.1109/ICASSP39728.2021.9413520.
- [7] K. Basak, N. Mishra, and H. T. Chang, "TranStutter: A convolution-free transformer-based deep learning method to classify stuttered speech using 2D Mel-spectrogram visualization and attention-based feature representation," *Sensors*, vol. 23, no. 19, Sep. 2023, doi: 10.3390/s23198033.
- [8] D. V. Kuppevelt, J. Heywood, M. Hamer, S. Sabia, E. Fitzsimons, and V. V. Hees, "Segmenting accelerometer data from daily life with unsupervised machine learning," *PLoS ONE*, vol. 14, no. 1, pp. e0208692–e0208692, Jan. 2019, doi: 10.1371/journal.pone.0208692.
- [9] R. S. -Segundo, J. L. -Trueba, B. M. -González, and J. M. Pardo, "Segmenting human activities based on HMMs using smartphone inertial sensors," *Pervasive and Mobile Computing*, vol. 30, pp. 84–96, 2016, doi: 10.1016/j.pmcj.2016.01.004.
- [10] K. Li *et al.*, "Applying multivariate segmentation methods to human activity recognition from wearable sensors' data," *JMIR mHealth and uHealth*, vol. 7, no. 2, pp. e11201–e11201, Feb. 2019, doi: 10.2196/11201.
- [11] H. Geng, Z. Huan, J. Liang, Z. Hou, S. Lv, and Y. Wang, "Segmentation and recognition model for complex action sequences," *IEEE Sensors Journal*, vol. 22, no. 5, pp. 4347–4358, 2022, doi: 10.1109/JSEN.2022.3144157.
- [12] Z. Wang, S. Hou, M. Zhang, X. Liu, C. Cao, and Y. Huang, "GaitParsing: Human semantic parsing for gait recognition," *IEEE Transactions on Multimedia*, vol. 26, pp. 4736–4748, 2024, doi: 10.1109/TMM.2023.3325962.
- [13] A. N. Tarekgn, M. Sajjad, F. A. Cheikh, M. Ullah, and K. Muhammad, "Efficient human gait activity recognition based on sensor fusion and intelligent stacking framework," *IEEE Sensors Journal*, vol. 23, no. 22, pp. 28355–28369, Nov. 2023, doi: 10.1109/JSEN.2023.3319353.
- [14] A. Smith and C. Weber, "How stuttering develops: The multifactorial dynamic pathways theory," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 9, pp. 2483–2505, Sep. 2017, doi: 10.1044/2017\_JSLHR-S-16-0343.
- [15] V. Mitra *et al.*, "Analysis and tuning of a voice assistant system for dysfluent speech," in *Interspeech 2021*, ISCA, Aug. 2021, pp. 4848–4852, doi: 10.21437/Interspeech.2021-2006.
- [16] L. Verde, G. D. Pietro, and G. Sannino, "Voice disorder identification by using machine learning techniques," *IEEE Access*, vol. 6, pp. 16246–16255, 2018, doi: 10.1109/ACCESS.2018.2816338.
- [17] N. P. Narendra and P. Alku, "Dysarthric speech classification from coded telephone speech using glottal features," *Speech Communication*, vol. 110, pp. 47–55, 2019, doi: 10.1016/j.specom.2019.04.003.
- [18] C. Quan, K. Ren, and Z. Luo, "A deep learning based method for Parkinson's disease detection using dynamic features of speech," *IEEE Access*, vol. 9, pp. 10239–10252, 2021, doi: 10.1109/ACCESS.2021.3051432.
- [19] S. Alharbi, M. Hasan, A. J. H. Simons, S. Brumfitt, and P. Green, "A lightly supervised approach to detect stuttering in children's speech," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, ISCA, 2018, pp. 3433–3437, doi: 10.21437/Interspeech.2018-2155.
- [20] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *18th International Conference on Pattern Recognition (ICPR'06)*, IEEE, 2006, pp. 441–444, doi: 10.1109/ICPR.2006.67.
- [21] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition," *IPSA Transactions on Computer Vision and Applications*, vol. 10, no. 1, 2018, doi: 10.1186/s41074-018-0039-6.
- [22] Y. Peng, K. Ma, Y. Zhang, and Z. He, "Learning rich features for gait recognition by integrating skeletons and silhouettes," *Multimedia Tools and Applications*, vol. 83, no. 3, pp. 7273–7294, 2024, doi: 10.1007/s11042-023-15483-x.




- [23] B. Lin, S. Zhang, and X. Yu, "Gait recognition via effective global-local feature representation and local temporal aggregation," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2021, pp. 14628–14636, doi: 10.1109/ICCV48922.2021.01438.
- [24] C. Fan *et al.*, "GaitPart: Temporal part-based model for gait recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2020, pp. 14213–14221, doi: 10.1109/CVPR42600.2020.01423.
- [25] H.-M. Hsu, Y. Wang, C.-Y. Yang, J.-N. Hwang, H. L. U. Thuc, and K.-J. Kim, "Learning temporal attention based keypoint-guided embedding for gait recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 17, no. 3, pp. 689–698, May 2023, doi: 10.1109/JSTSP.2023.3271827.
- [26] B. Lin, S. Zhang, and F. Bao, "Gait recognition with multiple-temporal-scale 3D convolutional neural network," in *Proceedings of the 28th ACM International Conference on Multimedia*, New York, USA: ACM, Oct. 2020, pp. 3054–3062, doi: 10.1145/3394171.3413861.
- [27] H. Chao, Y. He, J. Zhang, and J. Feng, "GaitSet: Regarding gait as a set for cross-view gait recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 8126–8133, Jul. 2019, doi: 10.1609/aaai.v33i01.33018126.
- [28] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "GEINet: View-invariant gait recognition using a convolutional neural network," in *2016 International Conference on Biometrics (ICB)*, IEEE, Jun. 2016, pp. 1–8, doi: 10.1109/ICB.2016.7550060.

## BIOGRAPHIES OF AUTHORS



**Ravikiran Reddappa Reddy**    earned his Bachelor's of Engineering (B.E.) degree in Electronics and Communication Engineering (ECE) from VTU, Belagavi in 2009. He has obtained his master's degree (M.Tech.) in ECE from VTU in 2012. Currently, he is a research scholar at VTU, Belagavi, doing his Ph.D. in Electronics and Communication Engineering, and also working as Assistant Professor in SJC Institute of Technology, Chickballapur, Karnataka. He has attended many workshops and induction programs conducted by various universities. His areas of interest are image processing and signal processing. He can be contacted at email: ravikiran\_12@rediffmail.com.



**Santhosh Kumar Gangadharaih**    is a Professor and Principal in the Department of Electronics and Communication Engineering at East West College of Engineering, Bangalore with an experience of 15 years in teaching. He is qualified in Bachelor and Master Degrees in Electronics and Communication Engineering and Ph.D. in Electronics and Communication Engineering in the area of image processing. His areas of interest are image processing and signal processing. He can be contacted at email: skg2185@gmail.com.