Recognising British Sign Language using Deep Learning: A Contact-less and Privacy-Preserving Approach

Hira Hameed, *Student Member, IEEE*, Muhammad Usman, *Senior Member, IEEE*, Ahsen Tahir, *Member, IEEE*, Kashif Ahmad, *Senior Member, IEEE*, Amir Hussain, *Senior Member, IEEE*, Muhammad Ali Imran, *Senior Member, IEEE*, and Qammer H. Abbasi, *Senior Member, IEEE*

***Abstract*—Sign language is utilised by deaf-mute to communicate through hand movements, body postures and facial emotions. The motions in sign language comprise a range of distinct hand and finger articulations that are occasionally synchronized with the head, face and body. Automatic sign language recognition is a highly challenging area and still remains in its infancy compared to speech recognition after almost three decades of research. Current wearable and vision-based systems for sign language recognition are intrusive, suffer from limitations of ambient lighting and privacy concerns. To the best of our knowledge, our work proposes the first contactless British Sign Language (BSL) recognition system using radar and Deep Learning (DL) algorithms. Our proposed system extracts two dimensional spatio-temporal features from the radar data and applies state-of-the-art DL models to classify spatio-temporal features from BSL signs to different verbs and emotions, such as help, drink, eat, happy, hate and sad etc. We collected and annotated a large-scale benchmark BSL dataset covering 15 different types of BSL signs. Our proposed system demonstrates highest classification performance with a multiclass accuracy of up to 90.07% at a distance of 141 cm from the subject using the VGGNet model.**

***Index Terms*—RF sensing, Contactless monitoring, micro-Doppler signatures, British sign language, deep learning**

I. Introduction

OVER 430 million people on earth are living with some form of hearing impairment, who represent a significant 5% of the world's population. Hearing loss is expected to afflict roughly 700 million people by 2050 [1]. People who have trouble hearing rely on sign language to communicate, depending on their level of disability. Millions of people throughout the world use sign language to communicate, and it is critical for their social inclusion. Similar to spoken languages, sign language is used in different variants in different parts of the works. The importance of sign language with its origins is discussed in [6], [7], with examples from American, Japanese, Chinese, Arabic and British sign languages (BSL).

Automatic sign language recognition is a highly challenging field of research and the current progress still remains in its infancy compared to speech recognition. Many efforts to automate the sign language translation into speech/text have been made in the literature using sensor or vision-based approaches.

For sensor-based approaches, more than one sensors are generally adopted by researchers. [22] utilised 10 flex sensors to develop a single-hand data glove for hand gesture recognition. These flex sensors are attached to each finger's two joints. Patel et al. [23] proposed a data glove, which utilised 5 flex sensors, one for each finger to reduce the bulkiness of the wearable glove. However, flex sensors can only measure flexion of fingers and cannot obtain finger and hand movements or orientation. Wang et al. [24] proposed a gesture recognition system using data glove that utilised 5 flex sensors along with a three-axis accelerometer. The accelerometer was placed in the centre of the back of hand palm. The system recognises 50 Chinese sign language gestures in real-time with an accuracy of around 91%. However, wearable systems are intrusive, devices such as hand gloves are cumbersome to wear and limit human computer interaction to subjects which cannot carry or wear the sensors for continuous interaction.

Vision-based systems, such as Kinect sensor systems may perform fusion of depth data and color information for the recognition of static signs of sign language [8]. Coloured gloves are used in [9] along with hidden Markov models for automatic recognition of sign language. Authors in [10] presented a system for recognition of Thai sign language using a glove-based device equipped with flex sensors and gyroscopes. Furthermore, the authors classified the data using the K-nearest neighbor algorithm. Camera-based systems are widely used to recognize sign language, because of the wide availability of cameras and tools to comprehend video and images. For this reason, majority of the existing works that focus on designing sign language recognition (SLR) systems, generally utilise widely available two-dimensional (2D) video cameras [11]. One such work is presented in [12], where the authors propose a deep neural network with bidirectional recurrences and temporal convolutions. The paper aims at identifying sign language gestures with a focus on enhancing the frame-wise accuracy of gesture recognition. Interestingly, the work in [13] combines the data from a grayscale video with depth information and hand's skeletal joints. A convolutional neural network (CNN) based framework is designed that integrates the features coming from all three data sources. Raj et al. [18] utilised an artificial neural network (ANN) to classify the BSL signs on the image data coming from a standard 2-D camera, where images were examined using histogram of

Hira Hameed, Muhammad Usman, Ahsen Tahir, Muhammad Ali Imran, and Qammer H. Abbasi are with University of Glasgow, James Watt School of Engineering, Glasgow, G12 8QQ, UK ({Hira.Hameed, muhammad.usman, Ahsen.Tahir, muhammad.imran, qammer.abbasi}@glasgow.ac.uk).

Muhammad Usman is also with the School of Computing, Engineering and Built Environment, Glasgow Caledonian University, Glasgow, G4 0BA, UK (muhammad.usman@gcu.ac.uk).

Ahsen Tahir is also with Department of Electrical Engineering, University of Engineering and Technology, Lahore, PK (ahsan@uet.edu.pk).

Kashif Ahmad is with the Department of Computer Science, Munster Technological University, Cork, Ireland (kashif.ahmad@mtu.ie).

Amir Hussain is with the School of computing, Edinburgh Napier University, Scotland, UK (A.Hussain@napier.ac.uk).

oriented gradients (HOG), However, all camera-based systems have certain fundamental flaws, such as the video may be recorded and video/images raise privacy concerns in real-world applications. Furthermore, ambient lighting is an important requirement and a limitation of vision-based systems for SLR.

Contactless radio frequency (RF) sensing has been recently studied in the healthcare sector for its applications in activity monitoring and assisted living. The *"contactless"* feature of RF sensing obliterate the requirement of wearing and carrying devices. Furthermore, an added advantage of RF sensing is its ability to leverage existing communication infrastructure, such as Wi-Fi routers. Moreover, the radar-based sensing system can alleviate the limitations of camera-based systems by protecting users' privacy and providing immunity to variations in ambient lighting. From a privacy perspective, people do not vary to disclose their private data. The information they deliver by Sign language is confidential between people to people.

These radar systems exploit the Doppler signatures of the reflected wave to produce unique hand movements. Mccleary et al. [19] used a frequency-modulated continuous-wave (FMCW) radar to classify four BSL gestures with the help of deep CNN models. Similarly, a FMCW radar has been used in [20] to classify different words in American sign language with the help of generative adversarial networks (GAN). Feature-level fusion of RF sensor network data has been used for ML classification of American sign language [21]. Diverse hand and arm motions create distinct multi-path distortions in Wi-Fi signals, resulting in distinctive patterns in the time series of channel state information (CSI), which have been utilised for classification of sign language using kernel-based SVM [15]. Maclaughlin et al. [17] demonstrated that multiple radar antenna signals received at different azimuth angles can be utilized to differentiate between different subjects for sign language recognition with a fusion based CNN. The author of this study uses a deep learning architecture to solve the challenge of BSL fingerspelling alphabet recognition. The suggested work performs better than the current works in terms of precision (6%), recall (4%), and F-measure (5%). It reported better outcomes with 98.0% accuracy on webcam and BSL corpus dataset [28]. For overcoming the communication gap between speech-impaired and non-speech-impaired community members. The author creates a model of a deep learning model for predicting British Sign Language. CNN and LSTM employ vision-based data. The CNN model performed the best, achieving training and testing accuracies of 98.8% and 97.4%, whereas the LSTM model performed poorly, achieving training and testing accuracies of 49.4% and 48.8%, respectively [29]. This study presents the use of a four-beam patch antenna as a sensor node to evaluate the pill-rolling effect in Parkinson's disease using the S-band sensing technique at 2.4 GHz. The proposed system uses amplitude and phase information to efficiently distinguish between finger tremors and nontremors. It is determined that support vector machines have an accuracy of more than 90% on the collected dataset [30] .

Although RF sensing has been partly discussed in the literature to recognise sign language, the existence of a diverse dataset that includes samples from a wide range of subjects (diverse age and sex) and covers diverse number of classes is missing in the literature. To bridge this gap, this work focuses on recognising different gestures in BSL using micro-Doppler signatures of the data collected using a Radar sensor. Fifteen different types of Doppler signatures are considered that include verbs (Drink, Eat, Help, Stop and Walk), emotions (Sad, Happy, Hate, Depressed, and Confused) and Family Group (Family, Brother, Father, Mother and Sister). These categories of BSL signs include dynamic gestures wherein mobility or movements of the hands are used to represent various signs. An ultra-wideband (UWB) Radar, namely XeThru X4M03 was utilised for recording the dataset. We note that the dataset is recorded at two different distances and angles. These characteristics make the dataset a better choice for training and evaluation of ML algorithms for BSL signs recognition and translation. The recorded data is represented in the form of spectrograms and further spatio-temporal features are extracted using GoogleNet, squeezenet and VGGNet CNNs.

The main contributions of the work are summarized as follows:

- We propose a contact-less BSL recognition system that automatically recognises and translates BSL signs into verbs and emotions.
- We propose a contactless BSL recognition system that automatically recognises and translates BSL signs.
- We also collect a large-scale benchmark dataset containing a total of 1950 samples from 15 different types of BSL signs captured at distances of 141 and 154 centimetres. Moreover, the data samples are captured from two different angles. To ensure diversity, the data was collected from four deaf participants (1 male and 3 females) having ages between 16 and 82 years of age.
- We report experimental results of several state-of-the-art DL models on the dataset, which will provide a baseline for future research in the domain.
- In our contactless data, we got 90.07% accuracy across all fifteen classes using deep learning models.

The remaining sections of the paper are organized as follows: Section II discusses the adopted methodology including the data collection and annotation, pre-processing and deep learning algorithms. Dataset, evaluation metrics (matrices), experimental setup, results and Discussion are discussed in Section III. Conclusions and future insights are presented in Section IV.

II. Methodology

Figure 1 illustrates the methodology used in this study as a block diagram. The proposed framework is divided into three stages. In the first phase, we collected and annotated a large collection of BSL signs. In the second phase, we used signal processing techniques to extract spectrograms of various signs captured in the first phase. Finally, several DL models were employed to classify the signs.

In the next sub-sections, we provide a detailed description of each of the phases of the proposed methodology.

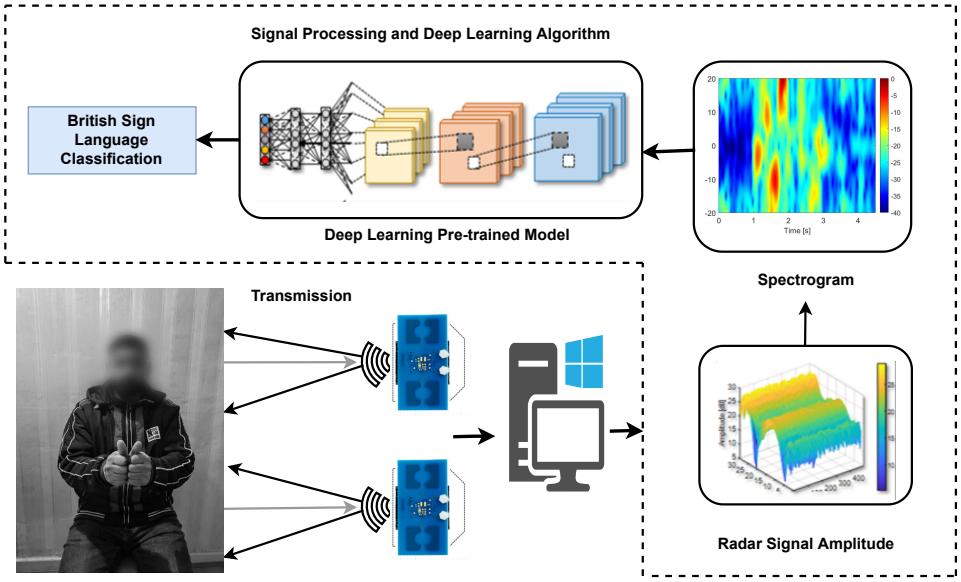
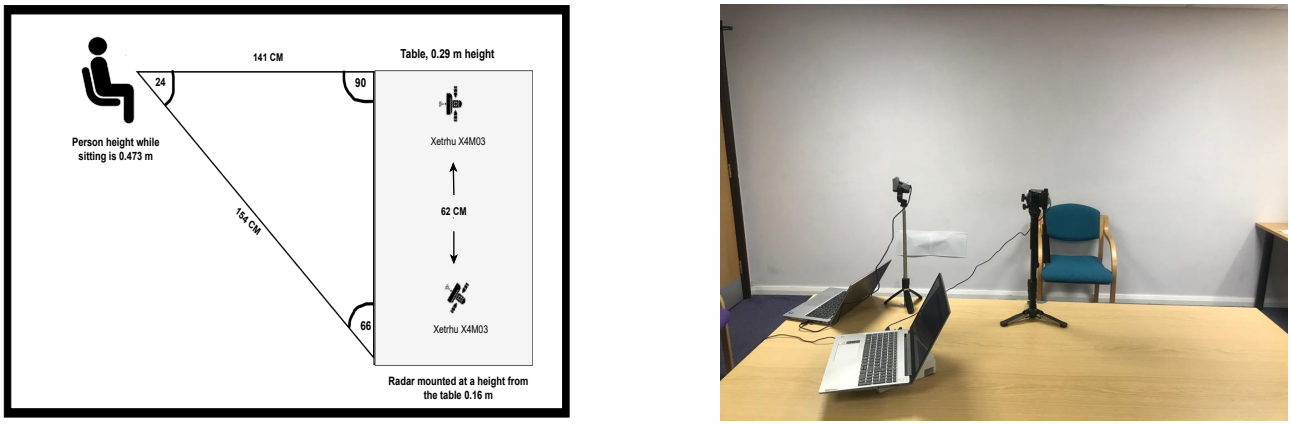


Fig. 1: A high level signal flow diagram of the proposed framework highlighting the UWB radar-based system, data collection, and the DL models for the classification of BSL signs.



(a) Measurable Experimental-Setup

(b) Real Experimental Setup

Fig. 2: Experimental setup of data collection using Xethru UWB radar sensor.

A. *Data Collection*

In this phase, we collected BSL sign data through UWB radar. Figure 2 provides an overview of the hardware setup of the radar-based BSL data collection system. To this aim, an ultra-wide band (UWB) radar sensor, namely Xethru X4M03, is used. The Xethru X4M03 is a UWB radar sensor with built-in transmitter (Tx) and receiver (Rx) antennas, providing a maximum detection range of 9.6 metres.

As shown in the figure, the radar sensor is placed on top of the screen of the laptop. The key parameter settings of the radar are indicated in Table I. In order to encompass different complexity levels in the dataset, the radar sensor was placed at two different distances including a distance of 141 and 154 centimetres from the subject (i.e., participant). Moreover, the sign gestures of the subject were recorded at two different angles. One was placed directly at front of the subject at a distance of 141cm, while the second was placed at an angle of 24^{o} from the subject at a distance of 154cm (see Fig. 2). The variations in the distance and viewpoint are expected to help in training distance and viewpoint invariant DL models that are able to recognize gestures of the subjects from different distances and angles. During the data collection, the body of the subject was in normal position with head and hands movements only. Moreover, the duration of each activity was set to 5 seconds involving the data collection of a single gesture from a single subject. Figure 3 provides visual illustration of the pronounced BSL[1].

Moreover, four deaf subjects/participants, one male and three females, participated in the data collection process. The reason to include more participants was to make the dataset more realistic and diverse. A total of 1950 data samples were collected during experiment for 15 different categories at distances of 141cm and 154cm. The details of the collected dataset are highlighted in Table II. In each experiment, a

[1]All the data and the photos for this research are being published with the consent of the participants.

(a) $Brother$ (b) $Sister$ (c) $Mother$ (d) $Father$ (e) $Family$

(f) $Confuse$ (g) $Depress$ (h) $Happy$ (i) $Hate$ (j) Sad

(k) $Walk$ (l) Eat (m) $Help$ (n) $Drink$ (o) $Stop$

Fig. 3: A visual illustration of the pronounced British Sign Language

Parameter	Value
Sensor	XeThru X4M03
Sensor range	9.6 metres
Sensor's distance from subject	141 cm and 154 cm
Sensor frequency	7.29GHz
Sensor bandwidth	1.5 GHz
Tx power	6.3dBm
Activity duration	5 seconds

TABLE I: Parameters configuration of radar software and hardware.

total of 975 data samples were collected from four deaf participants, where 15 samples were collected in each class. In particular, each participant repeated the speaking activity of each gesture 15 times with the radar. In this way, each participant contributed to collect 225 or 300 data samples in total for fifteen classes. In each case, A total of 975 spectrograms were categorized as fifteen BSL gestures, where

Classes	Experimental Dataset								
	141 CM				154 CM				Total
	Subject (S1)	Subject (S2)	Subject (S3)	Subject (S4)	Subject (S1)	Subject (S2)	Subject (S3)	Subject (S4)	
Brother	15	15	20	15	15	15	20	15	**130**
Sister	15	15	20	15	15	15	20	15	**130**
Mother	15	15	20	15	15	15	20	15	**130**
Father	15	15	20	15	15	15	20	15	**130**
Family	15	15	20	15	15	15	20	15	**130**
Confused	15	15	20	15	15	15	20	15	**130**
Depressed	15	15	20	15	15	15	20	15	**130**
Happy	15	15	20	15	15	15	20	15	**130**
Hate	15	15	20	15	15	15	20	15	**130**
Sad	15	15	20	15	15	15	20	15	**130**
Stop	15	15	20	15	15	15	20	15	**130**
Walk	15	15	20	15	15	15	20	15	**130**
Eat	15	15	20	15	15	15	20	15	**130**
Help	15	15	20	15	15	15	20	15	**130**
Drink	15	15	20	15	15	15	20	15	**130**
Total	**225**	**225**	**300**	**225**	**225**	**225**	**300**	**255**	**1950**

TABLE II: An overview of the data collected, number of subjects and the activities performed.

750 were utilized for training and 225 for testing.

B. Pre-Processing

This section describes the pre-processing steps carried out to extract the required spectrograms from the radar's data. In the beginning, the radar chip was configured via the XEP interface with x4driver. Data (were) was recorded from the module at 500 frames per second (FPS) (in the form of the float message data), where each value is a 32-bit floating point number. A loop was used to read the data file and save the data into a DataStream variable, which was mapped into a complex (range-time-intensity) range-time intensity (RTI) matrix. Thereafter, moving target indication (MTI) filter was applied to get the Doppler range map. Afterwards, the second MTI was used as a Butterworth 4^{th} order filter to generate the spectrograms using the following parameters: window length, overlap percentage, and fast Fourier transform (FFT) padding factor. In particular, a window length of 128 samples, and a padding factor of 16 was used. In addition, a range profile was created by first converting each chirp to an FFT. Afterwards, given a range bin, a second FFT was performed on a selected number of consecutive chirps [31]. Thereafter, spectrograms were generated leveraging a short time Fourier transform (STFT). The reason to exploit STFT is their ability to offer both temporal and frequency information, unlike Fourier transform (FT) that provides frequency information only. This is achieved through data segmentation wherein an FT was performed on each data segment. When the window length is changed, both the temporal and frequency resolutions are altered inversely. For instance, increasing one decreases the resolution of other parameter. Moreover, the level of Doppler details is mainly dependent on the sampling capability of radar's hardware. The greatest unambiguous Doppler frequency in radar is $F_{d,max} = \frac{1}{2}t_r$, where t_r represents the chirp time. The transmitted signal, T_s, can be represented by the following expression

$$T_s(t) = E\cos(2\pi ft), \tag{1}$$

where E is the reflection coefficient. At the receiver, the received signal, $R_s(t)$ can be represented by the given expression

$$R_s(t) = \acute{E}\cos(2\pi f(t - \frac{2D(t)}{c})), \tag{2}$$

where c is the speed of light and $D(t)$ is the distance of radar from the target location while performing hand's movements over time. Now 2 can be rewritten as The reflected signal can be expressed as $R_s(t)$,

$$R_s(t) = \acute{E}\cos(2\pi f(1 + \frac{2v(t)}{c})t - (\frac{4\pi D(\theta)}{c})), \tag{3}$$

where θ represents angle of the signal reflected off the target points to the direction of radar and $v(t)$ represents the point of target movement in front of the radar. The Doppler shift, f_d can be expressed as,

$$f_d = f\frac{2v(t)}{c}. \tag{4}$$

The reflected signal received at the receiver is a composite of many components, reflected from different moving parts of the body, such as head and hands. Each reflected component has unique characteristics of speed and acceleration. Let's assume that there are N number of unique reflected components, then 3 can be rewritten as the following expression.

$$R_s(t) = \sum_{k}^{N} E_k\cos(2\pi f(1 + \frac{2vk(t)}{c})t - (\frac{4\pi D_k(0)}{c})). \tag{5}$$

Accordingly, the combined Doppler shift comprises of an interaction of various Doppler individual shifts caused by unique movements of hands and head. Therefore, the successful classification of sign language gestures depends on unique characteristics of Doppler signatures due to different gestures. After obtaining the spectrograms of various Signs from the participants a dataset was constructed. As indicated in the high-level signal flow diagram in Figure 1, the dataset is consisted of two key modules: (i) System Training and (ii) System Testing. The proposed pre-trained DL classification algorithms were implemented on spectrograms to recognize British Sign language dataset.

C. Classification via Deep Models

For classification, the spectrograms generated in the previous step are fed into DL models. Three different pre-trained models are considered for this purpose: GoogLeNet, SqueezeNet and VGGNet. Our classification framework to differentiate in BSL signs/activities is mainly based on fine-tuning pre-trained models where multiple state-of-the-art CNN architectures pre-trained on ImageNet [5] are fine-tuned on the spectrogram images generated from the radar data. In fine-tuning the pre-trained models, we modify the top layers of the models to classify the collected data into fifteen considered classes The CNN architectures used in this work are described in detail in the following subsections.

1) GoogLeNet Model: GoogLeNet [4] is one of the state-of-the-art and commonly used CNN architecture for different image classification tasks [3]. The architecture is made up of 22 layers, including convolutional, pooling, inception, and fully linked layers. Six convolutional layers plus a pooling layer comprise the inception module. The module is made up of patches or filters of sizes 1 × 1, 3 × 3, and 5 × 5. These different-sized filters help in obtaining diverse patterns from the input image. At the output of each module, the feature maps produced from various filters are concatenated. Furthermore, 1 × 1 convolutions are performed prior to large filter convolutions. The use of a 1 × 1 convolution filter reduces the number of parameters that GoogLeNet requires.The hyper-parameter settings of GoogleNet are shown in Table III.

2) SqueezeNet Model: Our second pre-trained model is based on SqueezeNet architecture [2], which is composed of 18 layers. This architecture has shown comparable outcomes with fifty times fewer parameters, making it a better choice for applications with limited data and processing resources. Squeezenet accommodated three major strategies. In first, It reduces the squeeze layer from 3×3 filters to 1×1 filters. In the second strategy, It uses an expanded layer in which 1x1 and 3x3 filters were fed with fewer input parameters. The third technique down-samples late (with smaller stride values),

DL Model	Parameters	Settings
GoogleNet	Learning rate	0.0001
	Batch size	128
	Learning algorithm	Adam
	Loss function	Cross entropy
	Total epochs	100
	Iteration per epoch	500
	Elapsed time for 141CM	04:35:57
	Elapsed time for 154CM	06:15:21
SqueezeNet	Learning rate	0.0001
	Batch size	128
	Learning algorithm	Adam
	Loss function	Cross entropy
	Number of epochs	100
	Total epochs	500
	Elapsed time for 141CM	01:41:53
	Elapsed time for 154CM	02:03:28
VGG16	Learning rate	0.0001
	Batch size	16
	Learning algorithm	Adam
	Loss function	Cross entropy
	Total epochs	100
	Iteration per epoch	46
	Elapsed time for 141CM	41:32:11
	Elapsed time for 154CM	45:03:58

TABLE III: Parameter settings for the selected models

resulting in larger activation maps in the final layer, which improves accuracy. The parameter settings of SqueezeNet are shown in Table III.

3) VGGNet Model: Another pre-trained model is based on VGG16 architecture [27], which is composed of 16 layers. This architecture contains 138 million parameters, a 3x3 filter size with a stride of 1. Further the padding and max-pooling layer are same as a 2x2 filter with a stride of 2. The layers in this design are organized as follows: ReLU layers, convolutional layers, and max pool layers. ReLU provides efficient computing with faster learning. In the end, It has 3 fully connected layers and a softmax for output. The parameter settings of VGG16 are shown in Table III.

III. Experiments and Results

This section highlights the dataset description along with system evaluation using the discussed pre-trained DL models.

A. Dataset

The data collection and pre-processing phases described earlier resulted in a collection of spectrograms. In total, the dataset is composed of 1950 samples from 15 different categories/classes. These classes can be sub-grouped into three groups, namely (i) verbs, (ii) emotions, and (iii) family. The verbs group includes five classes namely Drink, Eat, Help, Stop and Walk. The emotions include Confused, Depressed, Happy, Hate and Sad while the final group is made of Family, Brother, Father, Mother, and Sister classes. Each of these fifteen classes contain an equal number of samples. Figure. 4 provides some samples images/spectrograms from the dataset.

The dataset has been divided into two subsets namely training and test set. The training set is composed of 1560 samples while the test set provides a total of 390 samples. The training and test sets have equal representation from all the classes and subjects.

B. Evaluation Matrics for Classification Model

In this work, the performance of the DL models in classification of BSL signs is evaluated in terms of weighted average accuracy, precision, recall, and F1 Score. F1 Score is one of the most commonly used metrics in the literature for classification, which is calculated using Equation 6. F1 Score is a combination of precision and recall, which are calculated using equation 7 and 8, respectively.

$$F1 = 2\frac{Precision.Recall}{Precision + Recall} \tag{6}$$

$$Precision = \frac{\sum TruePositive}{\sum TruePositive + \sum FalsePositive} \tag{7}$$

$$Recall = \frac{\sum TruePositive}{\sum TruePositive + \sum FalseNegative} \tag{8}$$

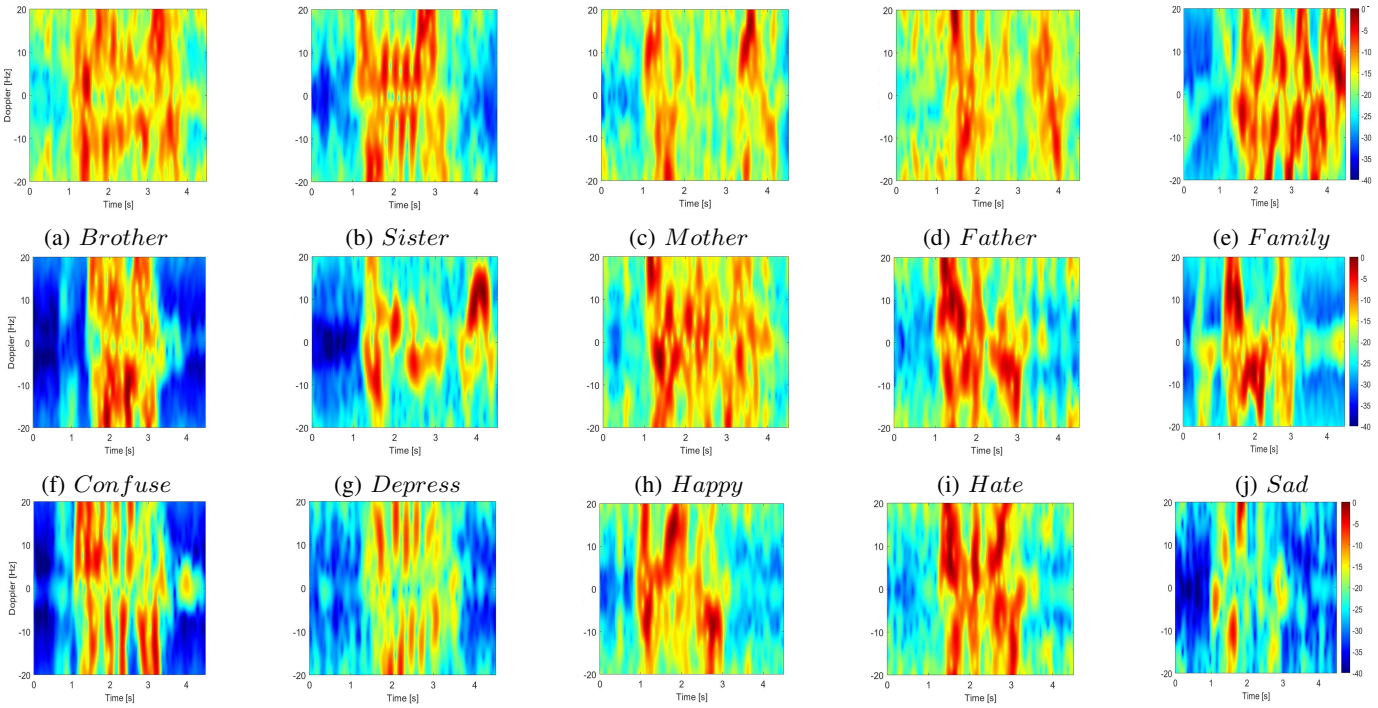
C. Results and Discussion

The objectives of the experimentation in this work are two-fold. On one side, we want to analyze the performance of different existing pre-trained models on the newly collected BSL dataset. On the other hand, we want to analyse the impact of variations in viewpoint and distances from the subject on the performance of BSL frameworks. Therefore, we conducted two different experiments by evaluating the performances of the models on spectrograms captured at distances of 141 and 154 centimetres.

The hyper-parameter settings for all the models are provided in Table III. All the models are fine-tuned on the dataset for a maximum of 100 epochs. Moreover, in all the experiments, fixed training and test sets are used. Our training and test sets are composed of 80% and 20% of the total data, respectively.

Table IV provides the experimental results of the experiments conducted at distances of 141 and 154 centimetres in terms of precision, recall, and F1 Score. As expected, overall better results, for all the models, are obtained when the radar sensor is placed at a distance of 141 centimetres compared to 154 centimetres. This is due to the reason that the 141cm radar is placed exactly in front of the subject and the micro-Doppler signature at this view points are more sensitive to hands movements as compared to what is received at 154cm radar for the same movements. As a result, the ML model better classify the hand movements as this viewpoint. As far as the performances of the pre-trained models is concerned, overall better results are obtained with VGGNet achieving an F1 Score of 0.87.

In order to better analyse the performances of the models, we also provide confusion matrices for each model at each distance in Figure 5. Figure 5a illustrates the confusion matrix of the GoogleNet model with 141 CM distance. It is worth noting that most of the signs are correctly identified with most of them having close to 100% accuracy. The lowest accuracy is 0.26% on *Brother* class, which is mostly confused with the class *Happy*. Similarly, the confusion matrix of the GoogleNet



(a) *Brother* (b) *Sister* (c) *Mother* (d) *Father* (e) *Family*

(f) *Confuse* (g) *Depress* (h) *Happy* (i) *Hate* (j) *Sad*

(k) *Walk* (l) *Eat* (m) *Help* (n) *Drink* (o) *Stop*

Fig. 4: Obtained spectrum's sample of 15 BSL signs

Model	141 CM					154 CM				
	Precision	Recall	F1 Score	Accuracy	95% CI	Precision	Recall	F1 Score	Accuracy	95% CI
GoogleNet	0.81	0.87	0.83	0.813	0.789-0.838	0.78	0.87	0.80	0.787	0.761-0.813
SqueezNet	0.75	0.83	0.77	0.75	0.723-0.777	0.69	0.76	0.71	0.693	0.667-0.718
VGG16	0.90	0.92	0.91	0.907	0.892-0.928	0.86	0.89	0.87	0.862	0.8404-0.884

TABLE IV: An evaluation of the pre-trained models in terms of macro-recall, macro precision, macro-F1-score, Accuracy and 95% confidence interval on the datasets captured at distances of 141 and 154 centimetres.

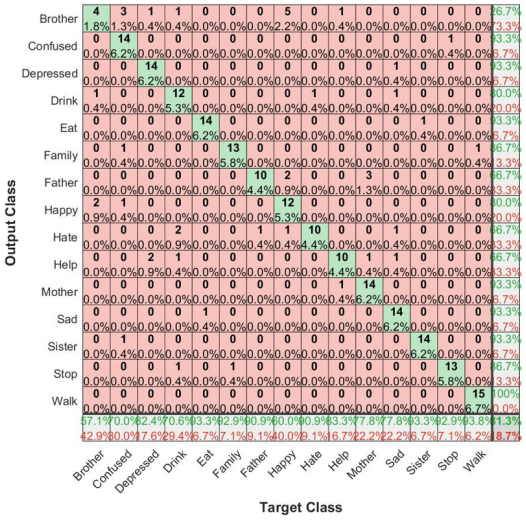
at a distance of 154 is presented in Figure 5b, where the classification accuracy is nearly 100% for all classes except the classes *Brother*, *Mother*, and *Sister*. The samples from the class *Brother* are mostly (around 0.13%) confused with the class *Happy*. The samples from the class *Mother* confused with the samples from classes *Father* and *Family*. This is due to the fact in both classes, both right and left-hand fingers move in the same manner. Similarly, the class *Sister* has resemblance with class *Drink* as the participants used their right-hand index finger and thumb to touch the nose.

Similarly, the confusion matrix of SqueezeNet with a distance of 141 centimetres is presented in Figure 5c. Here again, most of the test samples from all the BSL signs are correctly identified with an exception of *Brother*, which shows similarities with *Confused* class. This may be because of the reason that in both classes subjects rub their right-hand and left-hand with each other between the head. Furthermore, the confusion matrix of SqueezeNet at a distance of 154 centimetres is presented in Figure 5d. In this case, most of the signs are correctly identified with an exception of class *Brother* and *Mother*, which show similarities with the classes *Happy* and *Father*. Moreover, *Sad* gives a classification accuracy of 80% with only 20% matching/confusion with the class *Depressed*.

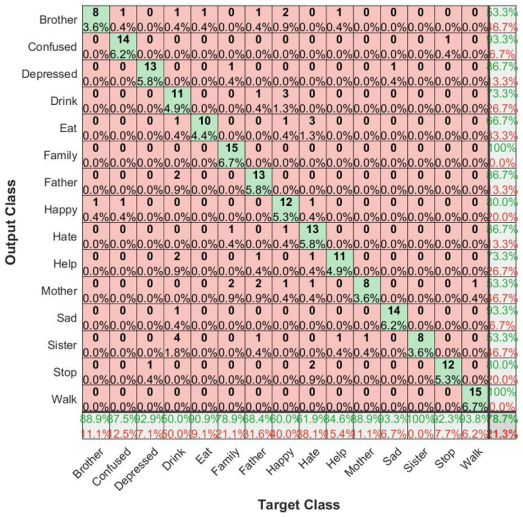
Lastly, the confusion matrix VGG16 with a distance of 141 centimetres is presented in Figure 5e. Here again, most of the BSL signs are correctly recognised with an exception of *Depressed*, which is similar to class *Sad*. This is because the way these two signs are performed are similar. In both classes the subjects face makes a downward movement. On the other side, *help* is similar to the *family* because both signs have upturned hands while performing the activity. Furthermore, the confusion matrix of VGG16 at a distance of 154 centimetres is presented in Figure 5f. In this case, VGG16 mostly classifies the considered classes correctly with few exceptions. The class *Confused* shows a resemblance with the two classes, namely *Family* and *Stop*, due to related movements. Despite, *Confused* class shows only 20% similarity with *Family* and 10% with the *Stop* class, producing 70% correct classification.

IV. Conclusion and Future Work

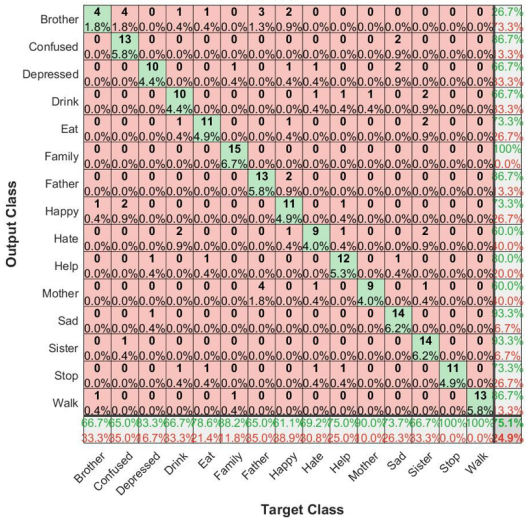
This paper presents BSL recognition framework in contactless and privacy-preserving manner. Off-the-shelf XeThru X4M03 UWB radar sensors are used combined with deep learning models. Fifteen most common gestures in BSL were studied, including verbs (drink, eat, help, stop, and walk), emotions (confused, depressed, pleased, hate, and sad), and



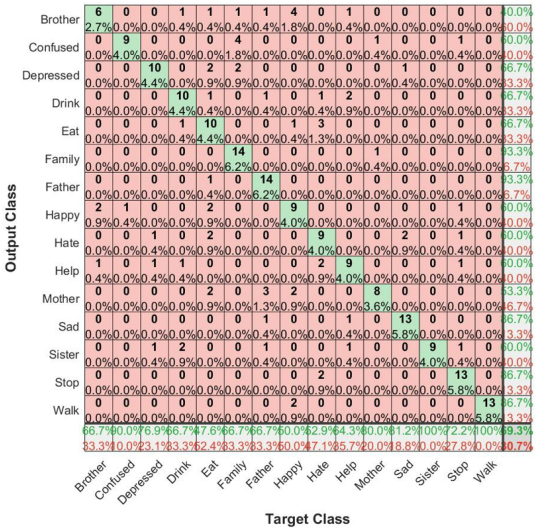
(a) Confusion matrix of GoogleNet at 141 CM



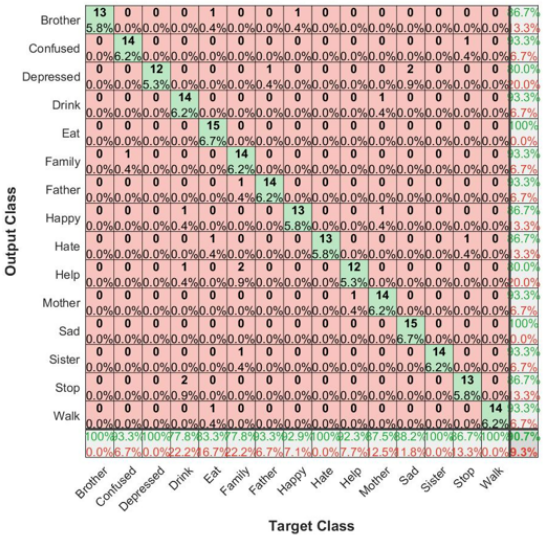
(b) Confusion matrix of GoogleNet at 154 CM



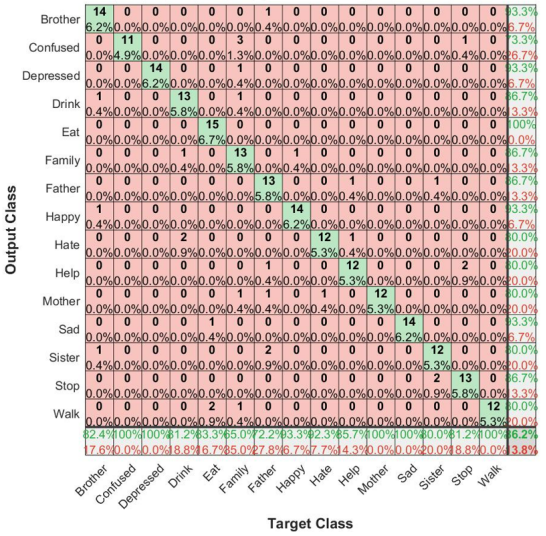
(c) Confusion matrix of SqueezNet at 141 CM



(d) Confusion matrix of SqueezNet at 154 CM



(e) Confusion matrix of VGGNet at 141 CM



(f) Confusion matrix of VGGNet at 154 CM

Fig. 5: The Confusion Matrices of all the models at two different distances.

family group (family, brother, father, mother, and sister). The experiment included four participants, three of whom were deaf, ranging in age from 16 to 82 years. Micro-Doppler unique features were maintained in the form of spectrograms for each class. Three deep learning models, GoogleNet, SqueezeNet, and VGGNet, were trained using these. Most of the gestures were correctly identified with 100% classification accuracy. The VGGNet model surpassed others, yielding an overall accuracy of 90.07% across all fifteen classes.

This preliminary work produced a fifteen-class BSL most common gestures dataset, which will be expanded for future research with additional vocabulary or sentences. We intend to strengthen our model by adding more users. The long-term goal is to create a real-time intuitive version of BSL that can be scaled to other sign languages and personalised for use by a range of end-users, such as deaf and blind children.

ACKNOWLEDGEMENTS

This work was supported in parts by Engineering and Physical Sciences Research Council (EPSRC) grants EP/T021020/1 and EP/T021063/1

REFERENCES

[1] WHO. (2021). Deafness and Hearing Loss, howpublished = https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss/.htm, note = Accessed: 1 April 2021.

[2] Iandola, F., Han, S., Moskewicz, M., Ashraf, K., Dally, W. & Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. *ArXiv Preprint ArXiv:1602.07360*. (2016).

[3] Ahmad, K. & Conci, N. How deep features have improved event recognition in multimedia: A survey. *ACM Transactions On Multimedia Computing, Communications, And Applications (TOMM)*. **15**, 1-27 (2019).

[4] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. Going deeper with convolutions. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 1-9 (2015).

[5] Deng, J., Dong, W., Socher, R., Li, L., Li, K. & Fei-Fei, L. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference On Computer Vision And Pattern Recognition*. pp. 248-255 (2009).

[6] Simons, G. & Fennig, C. Ethnologue: Languages of Honduras. (2017).

[7] Fenlon, J. & Wilkinson, E. Sign languages in the world. *Sociolinguistics And Deaf Communities*. pp. 5-28 (2015).

[8] Jadooki, S., Mohamad, D., Saba, T., Almazyad, A. & Rehman, A. Fused features mining for depth-based hand gesture recognition to classify blind human communication. *Neural Computing And Applications*. **28**, 3285-3294 (2017).

[9] Bauer, B. & Hienz, H. Relevant features for video-based continuous sign language recognition. *Proceedings Fourth IEEE International Conference On Automatic Face And Gesture Recognition (Cat. No. PR00580)*. pp. 440-445 (2000).

[10] Jitcharoenpory, R., Senechakr, P., Dahlan, M., Suchato, A., Chuangsuwanich, E. & Punyabukkana, P. Recognizing words in Thai Sign Language using flex sensors and gyroscopes. *I-CREATe2017*. **4** (2017).

[11] Mohandes, M., Deriche, M. & Liu, J. Image-based and sensor-based approaches to Arabic sign language recognition. *IEEE Transactions On Human-machine Systems*. **44**, 551-557 (2014).

[12] Pigou, L., Van Den Oord, A., Dieleman, S., Van Herreweghe, M. & Dambre, J. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *International Journal Of Computer Vision*. **126**, 430-439 (2018).

[13] Neverova, N., Wolf, C., Taylor, G. & Nebout, F. Multi-scale deep learning for gesture detection and localization. *European Conference On Computer Vision*. pp. 474-490 (2014).

[14] Li, B., Yang, J., Yang, Y., Li, C. & Zhang, Y. Sign language/gesture recognition based on cumulative distribution density features using UWB radar. *IEEE Transactions On Instrumentation And Measurement*. **70** pp. 1-13 (2021).

[15] Shang, J. & Wu, J. A robust sign language recognition system with multiple Wi-Fi devices. *Proceedings Of The Workshop On Mobility In The Evolving Internet Architecture*. pp. 19-24 (2017).

[16] Chong, T. & Kim, B. American sign language recognition system using wearable sensors with deep learning approach. *The Journal Of The Korea Institute Of Electronic Communication Sciences*. **15**, 291-298 (2020).

[17] MacLaughlin, G., Malcolm, J. & Hamza, S. Multi Antenna Radar System for American Sign Language (ASL) Recognition Using Deep Learning. *ArXiv Preprint ArXiv:2203.16624*. (2022).

[18] D RAJ, R. & JASUJA, A. British sign language recognition using HOG. *2018 IEEE International Students' Conference On Electrical, Electronics And Computer Science (SCEECS)*. pp. 1-4 (2018).

[19] McCleary, J., Garcia, L., Ilioudis, C. & Clemente, C. Sign Language Recognition using micro-Doppler and Explainable Deep Learning. *2021 IEEE Radar Conference (RadarConf21)*. pp. 1-6 (2021).

[20] Rahman, M., Mdrafi, R., Gurbuz, A., Malaia, E., Crawford, C., Griffin, D. & Gurbuz, S. Word-level sign language recognition using linguistic adaptation of 77 GHz FMCW radar data. *2021 IEEE Radar Conference (RadarConf21)*. pp. 1-6 (2021).

[21] Gurbuz, S., Gurbuz, A., Malaia, E., Griffin, D., Crawford, C., Rahman, M., Kurtoglu, E., Aksu, R., Macks, T. & Mdrafi, R. American sign language recognition using rf sensing. *IEEE Sensors Journal*. **21**, 3763-3775 (2020).

[22] Preetham, C., Ramakrishnan, G., Kumar, S., Tamse, A. & Krishnapura, N. Hand talk-implementation of a gesture recognizing glove. *2013 Texas Instruments India Educators' Conference*. pp. 328-331 (2013).

[23] Patil, K., Pendharkar, G. & Gaikwad, G. American sign language detection. *International Journal Of Scientific And Research Publications*. **4** (2014).

[24] Wang, X., Xia, M., Cai, H., Gao, Y. & Cattani, C. Hidden-markov-models-based dynamic hand gesture recognition. *Mathematical Problems In Engineering*. **2012** (2012).

[25] Lee, B. & Lee, S. Smart wearable hand device for sign language interpretation system with sensors fusion. *IEEE Sensors Journal*. **18**, 1224-1232 (2017).

[26] Fairchild, D., Narayanan, R., Beckel, E., Luk, W. & Gaeta, G. Through-the-wall micro-Doppler signatures. *Chen, VC, Tahmoush, D., Miceli, WJ (Eds.)*. (2014).

[27] Krishnaswamy Rangarajan, A. & Purushothaman, R. Disease classification in eggplant using pre-trained VGG16 and MSVM. *Scientific Reports*. **10**, 1-11 (2020).

[28] Kumar, K. DEAF-BSL: Deep lEArning Framework for British Sign Language recognition. *Transactions On Asian And Low-Resource Language Information Processing*. **21**, 1-14 (2022).

[29] Dhulipala, S., Adedoyin, F. & Bruno, A. Sign and Human Action Detection Using Deep Learning. *Journal Of Imaging*. **8**, 192 (2022).

[30] Shah, S., Yang, X. & Abbasi, Q. Cognitive health care system and its application in pill-rolling assessment. *International Journal Of Numerical Modelling: Electronic Networks, Devices And Fields*. **32**, e2632 (2019).

[31] Clemente, C., Balleri, A., Woodbridge, K. & Soraghan, J. Developments in target micro-Doppler signatures analysis: radar imaging, ultrasound and through-the-wall radar. *EURASIP Journal On Advances In Signal Processing*. **2013**, 1-18 (2013)