Indian Language Identification using Deep Learning

Shubham Godbole[1,*], *Vaishnavi* Jadhav[2,**], and *Gajanan* Birajdar[3,***]

[1]Department of Electronics Engineering, Ramrao Adik Institute of Technology, Nerul, Navi Mumbai, India

Abstract. Spoken language is the most regular method of correspondence in this day and age. Endeavours to create language recognizable proof frameworks for Indian dialects have been very restricted because of the issue of speaker accessibility and language readability. However, the necessity of SLID is expanding for common and safeguard applications day by day. Feature extraction is a basic and important procedure performed in LID. A sound example is changed over into a spectrogram visual portrayal which describes a range of frequencies in regard with time. Three such spectrogram visuals were generated namely Log Spectrogram, Gammatonegram and IIR-CQT Spectrogram for audio samples from the standardized IIIT-H Indic Speech Database. These visual representations depict language specific details and the nature of each language. These spectrograms images were then used as an input to the CNN. Classification accuracy of 98.86% was obtained using the proposed methodology.

Index Terms: Convolutional Neural Network (CNN), Spoken Indian Language Identification (SLID), Log Spectrogram, gammatonegram, IIR-CQT Spectrogram, Artificial Neural Network (ANN), Deep Learning

1 Introduction

Inventive systems like Siri and Google Assistant depend on Automatic Speech Recognition (ASR). In order to work appropriately the ASR frameworks expect users to manually indicate the proper input language. Traditional Language Identification (LID) systems use area explicit information for extracting hand-made features from sound samples[4].

Recently, Deep Learning and Artificial Neural Networks (ANN) are been considered the best in class for pattern recognition issues[25]. A variety of computer vision tasks like Image Classification, display better performance using Deep Neural Networks. LID can be characterized as the task of recognizing the spoken language in any given utterance.

As research in ASR advances, a LID system would be important for any multi-lingual speech recognition system. In case of multi-lingual speech recognition, the accuracy of the speech recognizer system can be improved by using LID setup at the front-end. It reduces complication by directly processing the speech over the identified language rather than running over several languages.

Particularly, in an Indian scenario where nearly every state has its very own language and each language having many lingos, designing LID system becomes vital. The languages and dialects used in India are categorized into distinct families: the significant ones are Indo-Aryan, Dravidian, Sino-Tibetan, Austroasiatic, Tai–Kadai and Great Andamanese languages [16]. This experimental study involves languages from two of the family groups. Marathi, Hindi, Bengali belong to the Indo-Aryan family whereas Tamil, Telugu, Kannada and Malayalam belong to the Dravidian family[4].

Neural networks are used to classify the different languages. The issue regarding LID can also be modeled as a pattern recognition task. Patterns from the given spectrogram can be identified by the trained ANN which helps it to classify the unfamiliar language specimen into a known category. Seven languages from IIIT-H Indic speech databases are used for evaluation[1].

In our methodology, WAV files from the database were converted into spectrogram visuals. Each sample is represented by separate visuals viz. Log Spectrogram, Infinite Impulse Response Constant-Q Transform (IIR-CQT) Spectrogram and Gammatonegram. Convolutional Neural Network (CNN) is used for classification of different languages. The classification accuracy rate obtained using CNN is 98.86%.

Formulation of the paper is done in the following manner: In section II wide range of methodologies and approaches for LID are summarized. Section 3 presents the architecture for LID using spectrogram visuals. Simulation results and comparison between the proposed method and available techniques are introduced in Section 4. Section 5 concludes the article.

2 Literature Survey

In LID various features and methods such as acoustic, phonotactic and prosodic along with different classifiers are used to differentiate the languages. In one of the previous works, sound samples from the dataset [1] were converted into spectrogram visuals which were then classi-

[*]e-mail: godboleshubham2@gmail.com
[**]e-mail: vaishnavijadhav1998@gmail.com
[***]e-mail: gajanan.birajdar@rait.ac.in

fied using CNNs. Considerable results were obtained using this methodology with an accuracy of 95.52% which was highest in comparison to different works done till date [2].

An approach based on Grey Wolf Optimizer (GWO) feature selection and ANN classifier was applied for LID. For extracting features from the spectrogram image LBPHF, CLBP and Wavelet are used. The GWO algorithm helps in extracting only the optimal features from the spectrogram. The ANN classifier then differentiates between languages using results obtained from the GWO algorithm. An accuracy of 97% was achieved using this approach [3].

Deepti Deshwal et al. [5] carried out research demonstrating feature extraction methods in LID. This article classifies multiple feature extraction approaches based on various highlights, the human speech production framework, spectral or cepstral research and also on the basis of different transforms. In addition, various techniques including noise compensation have also been mentioned. The study additionally exhibits that Mel-Frequency Cepstral Coefficients (MFCCs) intertwined with other features vectors and cleansing approaches gives better performance as in comparison with pure MFCC based feature extraction.

With a lot of languages spoken in India, building speech applications becomes a difficult task. For building a speech-based application, recognition of the language being spoken is one of the foremost steps. Bakshi Aarti et al. [4] have summarized the acoustic, phonetic and prosodic features used for SLID specifically for Indian Languages.

A study of acoustic features and different classifiers proposed by S. Jothilakshmi et al. [16] features a hierarchical approach to distinguish between 9 Indian Languages. The language is identified at first by segregating it to the family it belongs and then recognizing it from its' corresponding family. This approach is carried out using GMM based LID system using MFCC with delta and acceleration coefficients which results in the accuracy of 80.56%.

Another method for speech classification uses the human auditory system along with spectrogram features proposed by A.F.Pour et al. [7] clarifies that the basilar membrane filters are generally represented by a gammatone function which provides good approximation about the predecided responses.

A study represented by Nadia Jmour et al.[9] describes a learning approach on training CNNs for Image Classification. Here, Traffic Sign Image Classification System was developed with a total of 360 images, each of the neural network layers was considered and fine-tuning technique was used for experiments varying in the number of epochs. The highest accuracy of 93.33% was achieved by running 6 epochs.

Another analysis conducted by Raj Kumar G.A et al. [11] broadens the idea of a CNN. The algorithm proposes the use of CNN for recognition of Facial Emotions. A CNN Architecture designed for this problem results in a staggering accuracy of 90+% compared to other studies which could only give up to 48%.

Earlier, Automatic SLID systems were built using language dependent speech to-text synthesizers. Saikia R. et al. [17] presents the response of Language Independent speech to text transcribers for various Indian languages. This closely relates to the need for deep learning in order to overcome the lack of accuracy in traditional LID systems.

[15] Examines the use of spectral features that are extracted using conventional block processing, pitch synchronous analysis and glottal closure regions for Language Recognition Performance. The spectral features extracted from PSA and GC based regions showed better results on both IITKGP-MLILSC (27 Indian Languages) and OGI-MLTS speech databases.

The study by C. Madhu et al. [14] uses language dependent phonotactic features and prosodic information along with a multilayer feed forward neural network classifier for LID. Phonotactic feature vectors are obtained by numerically representing two consecutive syllables whereas Prosodic feature vectors are obtained by concatenating features of three consecutive syllables resulting into an accuracy of 72% and 68% respectively.

3 Proposed LID Algorithm using Deep Learning

3.1 Proposed Method

The process of taking an input and yeilding a class or a probability that the input belongs to a specific class is known as Image Classification. CNN is a part of Deep learning neural networks. They are normally used to examine images and are more often used in image classification. CNNs' are preferred over other image classification algorithms due to its' requirement of very less processing. In the convolution layer, the neurons get input from only a specific area of the preceding layer. Neurons in the dense layer receive input from every input in the earlier layer. The main function of CNN is to extract features from the images, this makes feature extraction by manual methods a redundant task. These features are learned on a set of training images and make the model very accurate for image classification tasks[11]. Algorithm for CNN:-

1. Initialization of the filters and weights with arbitrary values.

2. The training image is then sent as input to the network and it follows the forward propagation phase (i.e the convolution layer, pooling layer, and the fully connected layer) and gives the output chances for different classes.

3. Calculation of the absolute error at the output layer is computed.

4. Gradient descent helps calculate the error for the weights of the network. All the filter values, parameter values and weights are updated to reduce the

output error. The weights are then adjusted in such a way so as to influence the network in reducing the gross error. After updating the values, the network can now categorize the image accurately.

3.2 Log Spectrogram

A spectrogram is a visual portrayal representing a range of frequencies as they differ in time. When correlated with an audio signal, at times they are also called sonographs or voice grams. Music, sonar, radar are the fields which highly utilize the use of Spectrogram. These Spectrograms can be utilized to recognize speech. Various methods such as an optical spectrometer, use of bandpass filters, Fourier or wavelet transform are used to create a Spectrogram. The classic format of a graph represents time on one axis while the other axis represents frequency; the amplitude of a specific frequency at a particular time is denoted by colour which can also be called as the third dimension. There are numerous varieties of format, at times the vertical and horizontal axes are switched which can be either linear or logarithmic. Short Time Fourier transform was used to generate the spectrogram and the frequency axis was scaled logarithmically. Short Time Fourier Transform can be mathematically represented by the equation: [8]

$$X(l, w) = \sum_{a=-\infty}^{\infty} x[a]w[a-l]e^{-jwa} \tag{1}$$

Where w[l]: a specified window. l(time): Discrete variable, w(freq): Continuous variable. Typical graphical representation of Short Time Fourier transform is a Spectrogram which is characterized as follows:

$$Spectrogram = |X(l, w)|^2 \tag{2}$$

3.3 Gammatonegram

Gammatone filters resemble closely to the refining performed by our ear. These filters help in generating time-frequency surfaces dependent on a gammatone study, which may be utilized as a substitution for a traditional spectrogram. While spectrogram is conventionally used as a time-frequency interpretation, it has many differences in the methodology by which the ear perceives and analyzes the information and the approach by which the spectrogram does it[7]. The gammatone function is given as follows:

$$g(t) = t^{N-1}e^{-at}Cosw_0tu(t) \tag{3}$$

Where a:bandwidth parameter, w_0: center frequency, N: order of gammatone function.[6]

3.4 IIR CQT Spectrogram

3.4.1 Implementation of IIR CQT

Constant-Q transform (CQT) refers to a method that transforms a time-domain signal x(n) into time-frequency domain so as to obtain the middle frequencies of the frequency bins which are geometrically distributed and the Q-factors are equal. Although these are wavelet transforms due to comparitively high values of Q-factor the term CQT is preferred. The CQT transform X^{CQ} (l,m) of a discrete time-domain signal x(m) is given by:

$$X^{CQ}(l, m) = \sum_{j=m-[\frac{N_k}{2}]}^{m+[\frac{N_k}{2}} x(j)a *_l (j - m + \frac{N_k}{2}) \tag{4}$$

where l = 1, 2, . . . , K: indexes related to frequency bins which denote rounding towards negative infinity, $a*_l$(m): complex conjugate of a_l(m).

IIR filter with one zero and one pole is used on the spectrum of the signal. The zero is positioned at -1 in the Z-plane so as to force a time window that is zero in its extremes. The position of the pole is different for each frequency bin so as to obtain distinct time window widths. The recursive equation of the filter is:

$$y[l] = x[l] + x[l+1] + poles[l] * y[l-1] \tag{5}$$

3.5 Convolutional Neural Network

Convolutional Networks (ConvNets) are at present the most competent deep learning models for classifying image data. They are closely related to the neural network system from the human body. These models learn invariant features heirarchically. They initially differentiate low-level features and afterwards figure out how to perceive and merge these features to learn progressively convoluted examples. Distinctive levels of features are extracted by the various layers of the network. Neurons in each layer are are represented in 3 dimensions namely height, width, depth. An input image is generally represented as a matrix of pixels. A grayscale image has 2 measurements height and width. The color of the image is represented by depth.

3.5.1 Convolutional Layer

The convolutional layer is always the initial phase of a CNN. Extraction of features from the input image is the main function of the convolution layer. Convolution maps are generated by passing the input image through convolution kernels. Convolution layer needs a 3D input generally represented as [w1xh1xd1] where w1: width, h1: height and d1: depth, the outputs of neurons in this layers are determined by multiplying the weights and the regional matrix they are associated to in the input volume . The convolution maps now have new volume [w2xh2xd2] where, w2: new width, h2: new height and d2: new depth[9].

$$w2 = \frac{w1 - f + 2 * p}{s} + 1 \tag{6}$$

$$h2 = \frac{h1 - f + 2 * p}{s} + 1 \tag{7}$$

$$d2 = k \tag{8}$$

f : spatial extend of the filter. k : number of filters. p : zero padding. s : stride.

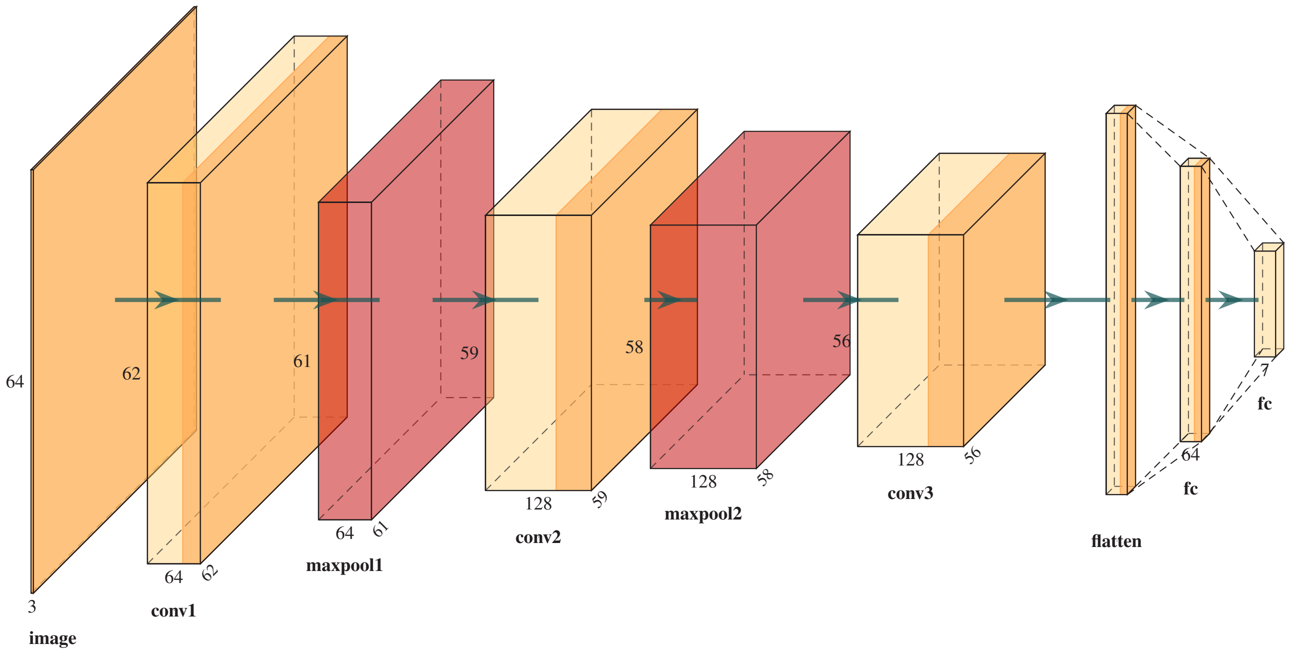

Figure 1: Architecture of Deep CNN.

3.5.2 Max Pooling Layer

Pooling dynamically reduces the size of the input representation. It helps to reduce the required parameters thereby reducing the computation power. This layer additionally helps to control over-fitting. Pool layer is embedded in the middle of two consecutive Conv layers, reducing the spatial dimensions of width and height.

Generally max pooling is used, the input image to this layer is subdivided into a set of areas that don't overrun on each other. The outputs are the maximum value of each area. They are obtained using MAX operation which reduces the number of parameters as well as optimzes the spatial dimensions. The volume of the Pool layer is [w2xh2xd2] and given by:

$$w2 = \frac{w1 - f}{s} + 1 \tag{9}$$

$$h2 = \frac{h1 - f}{s} + 1 \tag{10}$$

$$d2 = d1 \tag{11}$$

3.5.3 Flatten Layer

Flattening transforms a two-dimensional matrix of features into a single vector that can be further connected to the dense layer of the neural network.

3.5.4 Fully Connected Layer

The output from the earlier Layer is applied as an input to the fully connected layers. It combines the characteristics of CNN to categorize the image. It decides the classes introducing it in an output volume size of (1x1xk). As the name suggests each of the output neurons are connected to the earlier layer which results in classification of a particular category.

4 Experimental Results

4.1 Dataset

At present, the IIIT-H Indic speech database comprises of text and speech information in Bengali, Hindi, Kannada, Malayalam, Marathi, Tamil and Telugu. These languages were picked, as the absolute number of articles in every one of these languages sums up to be more than 10,000 and the speakers of these local languages were easily accessible. Every one of these dialects has a few lingos. As an underlying estimation, the discourse was recorded in the lingo in which the local speaker was comfortable with [1]. The samples in the dataset were converted into different spectrogram visuals viz. Log Spectrogram, Gammatonegram and IIR-CQT Spectrogram.

A key principle followed in developing a CNN is to manage the features spaced widely in the initial stage and then make it precise and deeper at the end. Generally, the number of channels are the same or increase while we advance through the layers in the CNN. The Activation shape can be calculated as (ceil(N+f-1)/s, ceil(N+f-1)/s, Number of filters), where 'N': input dimensions, 'f': filter size and 's': stride length. Activation value is calculated by taking the product of all the values in the dimensions of Activation Shape. The model used for CNN classification (Table 2) is as given below:

The performance of the proposed model can be evaluated from the confusion matrix (Table 1). It can be in-

Table 1: Confusion Matrix

	Bengali	Marathi	Telugu	Tamil	Malayalam	Kannada	Hindi
Bengali	2953	10	5	6	7	6	13
Marathi	7	2972	2	3	4	3	9
Telugu	1	2	2974	9	5	7	2
Tamil	2	3	9	2966	9	7	4
Malayalam	3	1	9	7	2972	7	1
Kannada	1	3	7	5	8	2974	2
Hindi	12	13	4	4	6	6	2955

Table 2: Architecture

Layer(type)	Output Shape	Parameters
conv2d_15 (Conv2D)	(None, 298, 138, 64)	1792
max_pooling2d_10 (MaxPooling)	(None, 149, 69, 64)	0
conv2d_16 (Conv2D)	(None, 147, 67,128)	73856
max_pooling2d_11 (MaxPooling)	(None, 73, 33, 128)	0
conv2d_17 (Conv2D)	(None, 71, 31, 128)	147584
flatten_5 (Flatten)	(None, 281728)	0
dense_10 (Dense)	(None, 64)	18030656
dense_11	(None, 7)	455
Total parameters: 18,254,343		
Trainable parameters: 18,254,343		
Non-Trainable parameters: 0		

ferred from the false positive values that Bengali, Marathi and Hindi share a common family and Kannada, Tamil, Telugu and Malayalam belong to the Dravidian family.

4.2 Results and Discussion

Using the CNN approach an accuracy of 98.86% is obtained.

Following are the graphs (fig.1 and fig.2) for the accuracy and loss of the model. Previous research on Indian LID demonstrates feature extraction and the use of different classifiers. Making modifications to the previous work and designing a CNN architecture that is deeper and which has an increased number of filters with a smaller filter size, exhibited better results.

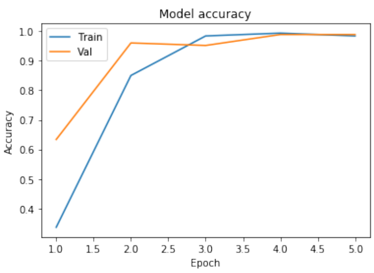

Figure 2: Model Accuracy

The Accuracy Graph (fig-1)initially demonstrates that the model is undergoing training. The training and validation datasets overlap at the end which determines that the model has a proper fit.

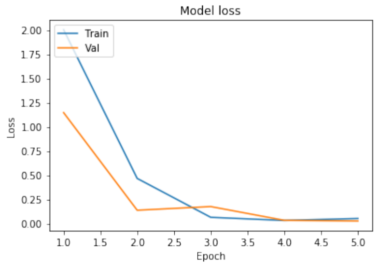

Figure 3: Model Loss

The Loss Graph (fig-2) depicts that the model has comparable performance on both training and validation datasets. Although, the model shows a high learning rate during the initial epochs, it settles to a point of stability with a minimal gap between the two final loss values.

5 Conclusion and Future Work

Identification of Language becomes a crucial part of Speech Recognition, especially in multilingual scenario. This article proposes an approach which works by visualizing the audio samples of each language with the help of spectrogram and a Deep Learning based LID system. The analysis was performed on seven Indian languages from the standardized dataset obtained from IIIT-H database . The experimental results demonstrate the classification accuracy using CNN to be 98.86%. A comparison between existing methods and studies was made depicting the proposed method (Table 3) to give the highest accuracy. As an extension to our work, a plan to implement the proposed method on a different dataset containing multiple languages can be executed. The languages can also be distinguished using a Multiview Deep Learning approach.

Table 3: Comparison of the proposed approach with existing algorithms.

Algorithm	Features	Languages	Accuracy
Verma et al.[12]	SVM	3	81%
Gupta et al.[13]	MFCC with random forest and SVM	4	92.6%
Madhu et al. [14]	Phonotactic and Prosodic	7	72% and 68%
Rao et al.[15]	MFCC,GMM	27	58.14%
S. Palanivel et al.[16]	MFCC,GMM,HMM and ANN	9	80.56%
Saikia et al.[17]	Transcription-based	7	96%
Himadri M et al.[2]	CNN	7	95.52%
Proposed	CNN	7	98.86%

References

[1] “IIIT-H Indic Speech Databases, IIIT Hyderabad, India”, *http://festvox.org/databases/iiit_voices/*

[2] Himadri Mukherjee, Subhankar Ghosh,Shibaprasad Sen, Obaidullah Sk Md, K. C. Santosh, Santanu Phadikar, Kaushik Roy, Neural Computing and Applications **31**, pp. 8483-8501 (2019)

[3] Amit A. Chowdhury, Vaibhav S. Borkar and Gajanan K. Birajdar, Journal of Experimental and Theoretical Artificial Intelligence **32**, pp. 111-132 (2019)

[4] Bakshi Aarti and Sunil Kumar Kopparapu, Sādhanā **43**, (2018)

[5] Deepti Deshwal, Pardeep Sangwan and Divya Kumar, Wireless Personal Communications **107**, pp. 2071–2103 (2019)

[6] A. Venkitaraman, A. Adiga, C. S. Seelamantula, Signal Processing (Elsevier) **94**, pp. 608-619 (2014)

[7] Aref Farhadi Pour, Mohammad Asgari and Mohammad Reza Hasanabadi, 4th International Conference on Computer and Knowledge Engineering (ICCKE), pp. 1-4 (2014)

[8] E. Jacobsen and R. Lyons, IEEE Signal Processing Magazine **20**, pp. 74-80 (2003)

[9] Nadia Jmour, Sehla Zayen, Afef Abdelkrim, International Conference on Advanced Systems and Electric Technologies (IC_ASET), pp. 1-6 (2018)

[10] Y. LeCun, B. Boser, J. S. Denker, Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Neural Computation **1**, pp. 541-551 (1989)

[11] Rajesh Kumar G A, Ravi Kant Kumar, Gaotam Sanyal, International Conference on Signal Processing and Communication (ICSPC), pp. 1-6 (2017)

[12] Verma, V.k., Khanna ,N., IEEE Students Conference on Engineering and Systems (SCES), pp. 1-5 (2013)

[13] Manish Gupta, Shambhu Shankar Bharti, Suneeta Agarwal, 4th International Conference on Power, Control and Embedded Systems (ICPCES), pp. 1-6 (2017)

[14] Madhu, C.,George, A. and Mary, L., IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES), pp. 1-6 (2017)

[15] Rao, K. S., Maity, S. Â. V., Reddy, R., International Journal of Speech Technology **16**, pp. 413-430 (2013)

[16] S. Jothilakshmi, V. Ramalingam, S. Palanivel, Digital Signal Processing (Elsevier) **22**, pp. 544-553 (2012)

[17] Saikia R, Singh SR, Sarmah P, International conference on Asian language processing (IALP), pp. 1-4 (2014)

[18] Neha Sharma, Vibhor Jain, Anju Mishra, Procedia Computer Science **132**, pp. 377-384 (2018)

[19] Paolo Arena, Adriano Basile, Maide Bucolo, Luigi Fortuna, Nuclear Instruments and Methods in Physics Research Section **497**, pp.174-178 (2003)

[20] Jaromir Przybyło, Mirosław Jabłoński, Computers and Electronics in Agriculture (Elsevier) **156**, pp. 490-499 (2019)

[21] Maier, A., Syben, C., Lasser, T., and Riess, C., Zeitschrift für Medizinische Physik **29**, pp. 86-101 (2019)

[22] Rashid, H., Zafar, N., Iqbal, M. J., Dawood, H., and Dawood, H., Procedia Computer Science (Elsevier) **147**, pp. 124-130 (2019)

[23] Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Xiang Li, S., Wang, T, Environmental Science and Technology **53**, pp. 4128-4139 (2019)

[24] Pan Zhou and Jiashi Feng., International Conference on Machine Learning (ICML), pp.1-10 (2018)

[25] Kim, T., Information Security and Assurance - 4th International Conference (ISA), CCIS **76** pp. 138-148 (2010)