Date of publication MAR-31, 2024, date of current version FEB-09, 2024
www.computingonline.net/ computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.23.1.3430

Speech Emotion Recognition using Hybrid Architectures

M. NORVAL[1], Z. WANG[2]
[1]Department of Electrical Engineering, University of South Africa, Johannesburg (e-mail: 36825050@mylife.unisa.ac.za)
[2]Department of Electrical Engineering, University of South Africa, Johannesburg (e-mail: wangz@unisa.ac.za)

Corresponding author: M. Norval (e-mail: 36825050@mylife.unisa.ac.za).

This research was supported partially by the South African National Research Foundation (Grants nos. 120106, 41951, and 132797) and the South African National Research Foundation Incentive (Grant no. 132159).

ABSTRACT The detection of human emotions from speech signals remains a challenging frontier in audio processing and human-computer interaction domains. This study introduces a novel approach to Speech Emotion Recognition (SER) using a Dendritic Layer combined with a Capsule Network (DendCaps). A Convolutional Neural Network (NN) and a Long Short-Time Neural Network (CLSTM) hybrid model are used to create a baseline which is then compared to the DendCap model. Integrating dendritic layers and capsule networks for speech emotion detection can harness the unique advantages of both architectures, potentially leading to more sophisticated and accurate models. Dendritic layers, inspired by the nonlinear processing properties of dendritic trees in biological neurons, can handle the intricate patterns and variabilities inherent in speech signals, while capsule networks, with their dynamic routing mechanisms, are adept at preserving hierarchical spatial relationships within the data, enabling the model to capture more refined emotional subtleties in human speech. The main motivation for using DendCaps is to bridge the gap between the capabilities of biological neural systems and artificial neural networks. This combination aims to capitalize on the hierarchical nature of speech data, where intricate patterns and dependencies can be better captured. Finally, two ensemble methods namely stacking and boosting are used for evaluating the CLSTM and DendCaps networks and the experimental results show that stacking of the CLSTM and DendCaps networks gives the superior result with a 75% accuracy.

KEYWORDS Emotion recognition; Artificial Intelligence; Dendritic Layer; Capsule Networks; Ensemble

I. INTRODUCTION

EMOTION recognition from audio signals plays a pivotal role in enhancing human-computer interaction and enabling machines to understand and react to human emotions. The field has gained significant attention due to the explosion of voice-activated systems, virtual assistants, and AI-driven customer support systems. However, despite substantial advancements, emotion recognition from audio signals remains challenging due to the inherent complexity and variability of human emotions, and the diverse acoustic characteristics they present. Recently, deep learning models have demonstrated superior performance in various domains, including image recognition, natural language processing, and speech recognition. Their ability to learn complex patterns and dependencies from raw data suggests that they may offer improved performance for emotion recognition from audio signals [1]. In the context of human-computer interaction and affective computing, the challenge of recognizing emotions from audio-speech signals remains a vital research problem. The task involves designing robust and efficient models capable of accurately discerning various emotional states conveyed through spoken language, accounting for factors such as speaker diversity, language variations, and the dynamic nature of emotions.

Researchers have ventured into the utilization of ensemble techniques for emotion detection, with a specific focus on employing random forest averaging, locally weighted naive bayes (LWNB), logistic classifiers, boosting-tree-based models, gradient boosting and majority-voting classifier [2] [3] [4] [5] [6]. Furthermore, Capsule Networks

have garnered attention in the domain of Speech Emotion Recognition (SER) [7], including Capsule Network hybrids [8] [9] [10] [11] [12]. While Dendritic layers [13] [14] [15] [16] [17] [18] [19] have found applications in diverse domains such as statistical learning, high-speed data streams, complex-valued neuron models, forecasting and hybrid configurations in conjunction with Recurrent Neural Networks (RNN). The dendritic neuron models use in SER have remained unexplored [20]. The study of a hybrid approach, DendCaps, for SER is a unique contribution to this paper, as such an approach has not been previously investigated.

The Dendritic Layer, inspired by the biological dendritic structures in the human brain, provides a mechanism for capturing complex hierarchical features in audio signals. This layer excels in modelling intricate relationships within speech data, enabling it to effectively extract nuanced emotional cues embedded in speech patterns. On the other hand, Capsule Networks [21] offer an elegant solution for handling spatial hierarchies and preserving spatial relationships, which are critical in understanding the nuances of emotion expressed through vocal intonations and cadences. Integrating the Dendritic Layer [22] with a Capsule Network creates a synergistic architecture that leverages the strengths of both approaches. The Dendritic Layer, inspired by the biological dendritic structures in the human brain, provides a mechanism for capturing complex hierarchical features in audio signals. This layer excels in modelling intricate relationships within speech data, enabling it to effectively extract nuanced emotional cues embedded in speech patterns. On the other hand, Capsule Networks offer an elegant solution for handling spatial hierarchies and preserving spatial relationships, which are critical in understanding the nuances of emotion expressed through vocal intonations and cadences. Integrating the Dendritic Layer with a Capsule Network creates a synergistic architecture that leverages the strengths of both approaches. From a research problem perspective, Speech Emotion Recognition accuracy is a critical performance metric that directly impacts the utility of SER systems and achieving high accuracy necessitates addressing challenges such as dataset diversity, model complexity, and the nuances of emotional expression in speech signals. The investigation of newer state-of-the-art hybrid models is required to bridge the gap and improve accuracy.

This paper aims to address this challenge by proposing innovative/state-of-the-art techniques and methodologies that advance SER accuracy. This paper proposes using a Dendritic layer combined with a Capsule neural network for SER. Combining the Dendritic Layer with a Capsule Network represents a promising advancement for SER. This combined framework is poised to enhance the accuracy and robustness of SER systems. Furthermore, this paper proposes ensemble [23] [24] methods to combine CLSTM [25] and DendCaps using stacking and boosting. The contributions of this paper are as follows: The presentation of a state-of-the-art model for SER. From a methodology perspective Acoustic Feature Extraction, Hybrid Approaches and Evaluation Metrics are used. Experiments show the superior accuracy of the hybrid model.

We organized the remaining parts of the paper as follows. Section 2 reviews existing literature on CLSTM architectures, Dendritic Layers, Capsule Networks, Training Data and Evaluation methods. Section 3 looks at the proposed system, data organization and training procedures. In Section 4, we present and discuss the experimental results. Section 5 provides a conclusion and closure to the article.

II. LITERATURE REVIEW

A. CONVOLUTIONAL LONG SHORT-TERM MEMORY

Convolutional Long Short-Term Memory (CLSTM) networks are a hybrid type of neural network architecture, combining the strengths of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. CNNs are excellent for extracting local and shift-invariant features, particularly from image and audio data, while LSTMs are designed to capture long-term temporal dependencies, making them ideal for time-series and sequence data. In a typical CLSTM [26] [27] architecture, CNN layers are applied first, extracting a rich set of spatial features from the input data. These spatial features, often in the form of high-level feature maps, are then fed as sequences into the LSTM layer. The LSTM layer analyses these sequences and captures the temporal dependencies between the features. The CLSTM architecture has found considerable success in tasks such as video classification and time-series prediction, where both spatial and temporal features are critical. It leverages the spatial feature extraction capabilities of CNNs and the ability of LSTM networks to model temporal dynamics, thereby providing a robust framework for spatiotemporal feature learning. Long Short-Term Memory (LSTM) architectures represent a significant breakthrough in the field of deep learning, specifically addressing the challenges associated with understanding temporal dynamics and depth [28] [29].

B. DENDRITIC NEURON LAYER

In recent years, the pursuit of biologically inspired artificial neural network architectures has gained significant traction. One of the intriguing developments in this domain is the incorporation of dendritic computations into artificial neuron models, leading to the formulation of "dendritic neuron layers." Traditional artificial neurons are largely based on point neurons, which do not capture the intricate dynamics of dendritic trees observed in biological neurons. Dendritic trees are known to integrate synaptic inputs in a complex, often nonlinear, manner, allowing for sophisticated input-output transformations that cannot be achieved by simple summation. By introducing dendritic neuron layers, neural networks can potentially harness this added computational power, enabling them to recognize and process patterns in data with greater nuance and efficiency. These layers, the-

oretically, bridge the gap between traditional deep learning models and the rich dynamics observed in biological neural circuits. However, while the dendritic neuron layer concept is promising, its practical applications, benefits over traditional architectures, and effective training methodologies are areas of active research and warrant deeper exploration [19].

A dendritic neuron model is an abstraction of biological neural computation that incorporates the nonlinear processing capabilities of dendritic trees in real neurons [20]. The input is received by dendritic trees. These are branched extensions of the neuron that receive input from other neurons. In computational models, dendrites can carry out independent computations before sending their signals to the neuron's soma (cell body). The results of dendritic computations are aggregated at the soma, which then decides the overall output of the neuron. Unlike the perceptron, which typically uses a single summation and activation function, the dendritic model allows for multiple layers of computation within a single neuron. Each dendritic branch can perform its computation, and their combined results are integrated at the soma. Learning in dendritic models can be more complex than in perceptrons, as it can involve adjusting weights in both the dendritic branches and the soma. The Dendritic Neuron can be seen in Figure 1. The dendritic neuron model captures more of the intricacies of biological neurons, allowing for complex computations within a single neuron. It can potentially model nonlinear computations that a single perceptron cannot.

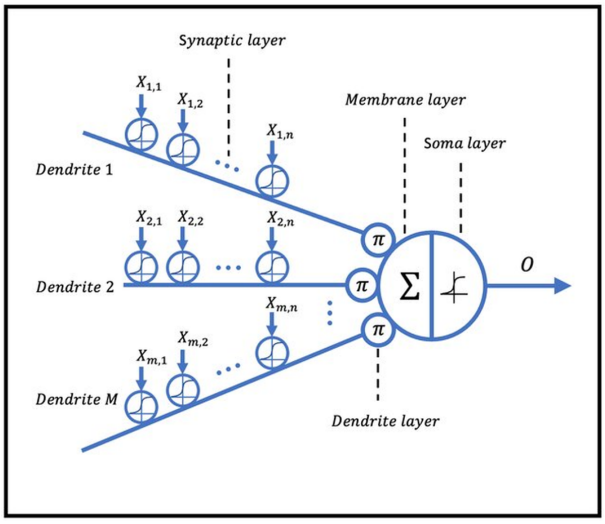


Figure 1. Dendritic Neuron Model [30].

When comparing the two both models aim to abstract neural computation, the perceptron offers a more simplified, linear approach, while the dendritic neuron model delves deeper into the nonlinear, hierarchical processing capabilities of biological neurons. The proof and verification for the abovementioned statement is that the perceptron applies a step function or threshold activation function, resulting in a piecewise linear decision boundary. The Dendritic models include more complex mathematical functions to capture nonlinearities and hierarchical processing. The model may involve spatial and temporal integration, capturing the nonlinear summation of inputs across dendritic branches. Nonlinear summation is especially better for SER because of the following aspects: Capturing Complex Relationships, Hierarchical Processing, Spatial Integration of Features, Temporal Dynamics, Handling Non-Linear Acoustic Patterns and Increased Model Expressiveness. A perception with a linear lacks these features. The basic perceptron formula is [31]:

$$y = f(w.x + b) \tag{1}$$

where:
y is the output.
f is an activation function, often a step function for the simplest perceptrons, but it could be a sigmoid, tanh, ReLU, etc. in modern networks.
w is the weight vector.
x is the input vector.
b is the bias.

The dendritic layer utilizes a multiplicative function to handle the output originating from numerous synapses in the synaptic layer. In this model, the synaptic layer receives output signals from other neurons, individually processing them using a sigmoid function. Subsequently, the dendritic layer employs a multiplication function to process the output signals from the synaptic layer. The membrane layer then processes the resulting signals from each branch in the dendritic layer through a summation function. Finally, the somatic layer processes the output signals from the membrane layer using another sigmoid function, ultimately producing the overall output signal of the entire dendritic neuron model.

Modelling the dendritic computation is more complex and less standardized, given that it's a newer area of exploration. However, a simplified version might look something like [19]:

$$y = f \sum_{i} g\left(w_i.x_i\right) + b) \tag{2}$$

f(·) Represents the activation function applied to the sum of the weighted inputs and bias. The somaś activation function.
$\sum_i$ This denotes the summation over all the inputs (i).
g This represents the activation function applied to each weighted input.

The choice of the activation function may depend on factors such as the characteristics of the data, the network architecture. Two widely used activation functions are:
1) Sigmoid Activation Function:

$$sigma(z) = \frac{1}{1 + e^{-z}} \tag{3}$$

2) Rectified Linear Unit (ReLU):

$$\text{ReLU}(z) = \max(0, z) \tag{4}$$

wi are weights associated with the i-th dendritic branch.
xi is input to the i-th dendritic branch. The output of a neuron is calculated as the weighted sum of its inputs plus a bias term.
Neuron Output:

$$\sigma\left(\sum_i w_i \cdot x_i + b\right) \tag{5}$$

The correctness of the formula $y = \sigma\left(\sum_i g(w_i \cdot x_i) + b\right)$ involves demonstrating that it accurately represents the desired computation, particularly in the context of neural networks. Here's an overview of the proof:

1) Spatial and Temporal Integration:
The summation term $\sum_i g(w_i \cdot x_i)$ represents the spatial integration of inputs across the dendritic tree. The optional activation function $g(\cdot)$ may introduce non-linearity, capturing more complex relationships in the input data.

2) Bias Term (b):
The bias term (b) allows for the adjustment of the overall activation level, accounting for factors not solely dependent on the synaptic inputs.

3) Activation Function (σ):
The sigmoid activation function $\sigma(z) = \frac{1}{1+e^{-z}}$ introduces non-linearity and squashes the output to the range (0, 1), making it suitable for binary classification problems.

4) Feasibility of Inputs:
The formula is feasible if the input to the sigmoid function is within a reasonable range. The choice of weights $(_i)$ and bias (b) during training ensures this feasibility.

5) Training and Adaptation:
The correctness is linked to the model's ability to adapt during training. The weights and bias are adjusted to minimize the difference between the predicted output and the actual target.

6) Gradient Descent:
The training process typically involves gradient descent or a similar optimization algorithm. Correctness is supported by the convergence of the optimization process, ensuring that the network parameters reach values that minimize the loss function.

The summation goes over all dendritic branches. It can be seen the perceptron computes a weighted sum of the inputs and then applies the activation function. On the other hand, each dendritic branch performs its own weighted sum and non-linear transformation before all the results are combined and processed by the soma. It's worth emphasizing that these are simplified representations. The actual formulas can be more complex, especially for the dendritic model which can incorporate various types of non-linearities, integration mechanisms, and more, reflecting our growing understanding of dendritic computation in biological neurons. It's also worth noting that the true behaviour and computation of real dendrites in biological neurons are still topics of active research, and artificial neural models might capture only certain aspects of this behaviour.

C. CAPSULE NETWORKS

In the domain of deep learning, Capsule Networks (CapsNets) has emerged as a novel architecture that aims to overcome the limitations inherent in traditional Convolutional Neural Networks (CNNs). Developed by Geoffrey Hinton and his team, Capsule Networks introduces the concept of "capsules"—a group of neurons that learn to recognize an object in the visual hierarchy and encapsulate its spatial and hierarchical relationships in the form of vectors. Unlike CNNs, which are highly susceptible to variations in orientation, scale, and pose, Capsule Networks are designed to maintain these hierarchical relationships, thereby achieving better performance in tasks requiring spatial understanding [32]. The distinctive routing-by-agreement mechanism enables capsules at one layer to send information to appropriate parent capsules in the layer above, replacing the function of pooling layers seen in CNNs. This mechanism allows for dynamic routing of information based on the data itself, which is a substantial departure from the static nature of traditional architectures. Consequently, Capsule Networks have shown promising results in various applications, from object classification to complex scene understanding, and they offer the possibility for more interpretable and robust models in the ever-evolving landscape of machine learning [7] [8] [33].

D. TRAINING DATA

The Wheel of Emotions, proposed by psychologist Robert Plutchik [34] in 1980, is a comprehensive model for understanding and categorizing human emotions. Plutchik's wheel is structured around eight primary emotion dimensions: joy versus sadness, trust versus disgust, fear versus anger, and surprise versus anticipation. These primary emotions can be combined in different ways to represent more complex emotional states [35]. The dataset used in this study is based on the 8 emotions in Plutchik's wheel. The dataset used is a custom-created Afrikaans speech corpora [36]. There are roughly 100 samples per category, so roughly 800 samples in total. The following features are extracted and trained on: Mel-frequency cepstral coefficients (MFCC), chromagram (Chroma), and Mel-frequency cepstrum (Mel), Contrast and German Tone Network (Tonnetz) [37] [38]. The figure below shows one training sample containing the extracted features.

E. EVALUATION

Upon the application of machine learning algorithms, it becomes imperative to employ performance evaluation met-

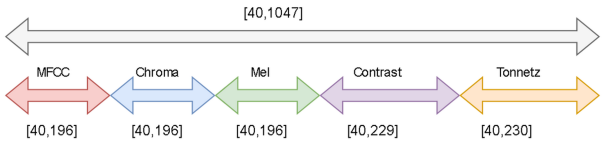


Figure 2. Extracted features.

rics to gauge their efficacy. These performance evaluation metrics constitute a crucial facet of the assessment process, with a multitude of metrics having been introduced in prior research endeavours, each catering to distinct aspects of algorithm performance. Consequently, the selection of an apt set of metrics is contingent upon the specific machine-learning problem at hand. Within the context of this paper, we employ a repertoire of established metrics tailored for classification problems. This approach allows us to glean invaluable insights into algorithm performance and facilitates a comprehensive comparative analysis. The following are evaluated in this research paper: Accuracy, precision, Recall, and F1-Score [39] [40]. Accuracy is a widely used deep-learning evaluation metric that quantifies the overall correctness of a model's predictions, representing the proportion of correctly classified instances out of the total. While useful for assessing overall model performance, it may be less informative in scenarios with imbalanced class distributions. Precision assesses the model's ability to make accurate positive predictions by measuring the ratio of true positive predictions to the total positive predictions. It is particularly relevant in applications where minimizing false positives is critical, such as medical diagnostics. Recall, also known as sensitivity or true positive rate, evaluates the model's capacity to capture all relevant positive instances by calculating the ratio of true positives to the total actual positives. It is essential in situations where missing positive cases can have significant consequences, like identifying diseases. The F1-Score is a balanced metric that combines precision and recall providing an extensive assessment of a model's performance, especially in imbalanced datasets. It represents the harmonic mean of precision and recall, favouring models that achieve both high precision and recall simultaneously. A confusion matrix is a tabular representation of a model's predictions, showcasing the count of true positive, true negative, false positive, and false negative predictions. It serves as the foundation for computing other evaluation metrics and offers insights into the types of errors a model makes, aiding in model diagnosis and improvement.

The respective formulas for the calculations can be seen below [41].

$$Accuracy = (\frac{TP + TN}{TP + TN + FP + FN}) \quad (6)$$

$$Precision = (\frac{TP}{TP + FP}) \quad (7)$$

$$Recall = (\frac{TP}{TP + FN}) \quad (8)$$

$$F1\ Score = (\ 2\ x\ \frac{\text{Precesion x Recall}}{\text{Precision + Recall}}) \quad (9)$$

III. METHODOLOGY

A. GAPS IN EXISTING SOLUTIONS

In the realm of speech emotion recognition (SER), significant advancements have been made in recent years, resulting in the development of robust models capable of discerning various emotional states from spoken language. However, despite these strides, there exist notable gaps and challenges that warrant attention and further research. One of these challenges is feature selection and extracting the right features that capture emotional content from audio signals. Another gap is that of high dimensionality where speech signals have a high dimensional space, making them computationally expensive to process and model. Finally, context awareness where emotions are often context-specific, but many models don't take contextual information into account.

B. JUSTIFICATION FOR CURRENT RESEARCH

Leveraging the synergistic potential of dendritic layers, capsule networks, and ensemble methods presents a compelling avenue for enhancing the robustness and interpretability of SER. Combining Capsule Networks (CapsNets) and Dendritic Neural Networks for audio emotion detection is intriguing. Both types of networks have characteristics that could be beneficial for the challenging task of emotion detection from audio signals.

C. PROPOSED SYSTEM

Based on the literature review limitations exist for emotion detection when using older architectures like CNNs and LSTMs. These architectures process audio in fixed-size chunks or frames. They might not capture long-term dependencies or subtle changes in emotion that occur over extended time intervals. Emotions in audio often depend on context, which can be challenging to capture with fixed-sized windows. Another challenging factor highlighted by the literature review is audio quality. In most real-world applications, audio signals are subjected to a range of distortions, noise, and variations in recording conditions. In this article, we propose training two networks. The first network is CLSTM. We train this network and record the results to get a baseline. The second network is a hybrid Dendritic [20] and Capsule network [8] DenCaps. This network is also trained and hyperparameters are adjusted for best accuracy. Next, ensemble methods [2] namely boosting and stacking are used to evaluate the two combined networks. A flowchart of the proposed system can be seen in Figure 3. The proposed system makes use of a custom-created Afrikaans speech corpora [36]. The following features are extracted: MFCC, Chroma, Mel, Contrast and Tonnetz. In terms of

evaluation Accuracy, F1-Score and a Confusion Matrix are used.

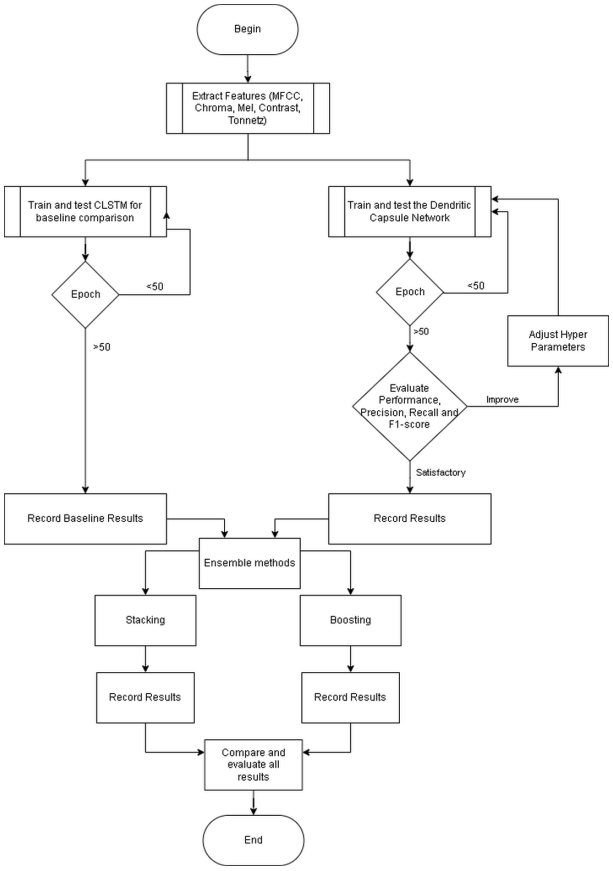


Figure 3. Proposed system using custom dataset.

D. DATASET PREPARATION

The dataset consists of eight emotion classes corresponding with the Plutchnik wheel of emotion. The data is placed into eight subfolders. Each contains 100 audio file samples. Audio files are processed, and the following are extracted: MFCC, Chroma, Mel, Contrast and Tonnetz. Data is split into validation and training samples using a 20% split. The following shape is fed into the networks (798, 1047, 40, 1). There are 1047 data points using 40 coefficients and 1 dimension.

E. TRAINING PROCEDURE

Google Colab was used with a High RAM V100GPU environment because of the sheer memory and processing requirements. Pre-trained models and training data were stored on Google Drive for fast access in Google Colab. We had to experiment with code examples from numerous GitHub repositories due to the extensive coding required for Dendritic Layers, Capsule NN and Ensemble methods. Some of the main libraries used are Tensorflow Keras, Numpy and sklearn.

1) CLSTM

The CNN LSTM hybrid CLSTM model consists of the following layers. Data is fed into a 1D Convolution layer followed by a Max Pooling layer and subsequently a dropout layer. The Dropout layer feeds into an LSTM layer which is flattened and then passed to a Relu-activated dense layer. Finally, another dropout layer. The model is compiled with a Categorical Cross-entropy loss function and an Adam optimizer. The model is trained using a batch size of 32 and 50 epochs. The CLSTM and DendCaps networks are saved in .tf Tensorflow format for later comparison. To save custom layers, the get_config () method needs to be overridden with the configuration parameters of the respective layer.

2) DendCaps

The DendCaps model consists of the following layers. Data is fed into two 2D Convolution layers, the first having a 128-kernel size and the second a 32-kernel size. Both Convolution layers use Relu activation. The output is fed into a max pooling layer, followed by a dropout and then a flattening layer. This output is fed into the custom Dentritic Layer. The output is sent to a Capsule Layer followed by a dense dropout and again a dense layer. The model is compiled with a Categorical Cross-entropy loss function and an Adam optimizer. The model is trained using a batch size of 32 and 50 epochs. The custom Dendritic layer is a custom class written in Python. The class takes two parameters namely units and segments. The units and segments are used to determine the shapes and sizes of the weight and bias tensors for the dendritic segments and the soma. The model was trained using various combinations of units and segments. The optimal segments were two. Unit wise the optimal value was 64. The custom Capsule layer is a custom class written in Python. The class takes three parameters namely the number of capsules (num_capsules), capsule dimensions (dim_capsule) and number of routings (routings). The layer takes input data, performs routing iterations, and produces output capsules as its result. The model was trained using various combinations of num_capsules, dim_capsule and routings. The optimal configuration was found to be num_capsules=10, dim_capsule=32, and routings=3.

3) Ensemble

Two Ensemble methods are implemented namely stacking and boosting. Stacking combines the predictions of multiple base models by training a meta-model on their outputs, often leading to improved prediction accuracy. This approach leverages diverse model architectures or algorithms to capture complex patterns in data, contributing to enhanced predictive performance. Boosting, on the other hand, is an ensemble learning method that sequentially trains a series of weak learners, each focusing on the data points that were misclassified by the preceding ones. Through this iterative process of re-weighting the data and combining weak learn-

Table 1. CLSTM Results

Test	Accuracy	Precision	Recall	F1Score	Loss
1	70%	0.84%	0.80%	0.80%	0.77%
2	69%	0.92%	0.77%	0.78%	0.82%
3	71%	0.82%	0.81%	0.81%	0.73%
4	72%	0.78%	0.82%	0.82%	0.72%

ers. To achieve this both the CLSTM and DendCaps .tf networks are loaded into memory and then processed.

a: Stacking

In the case of stacking a new stacked network is created using a custom Python class that Concatenates the two networks and adds a Softmax activation layer to it. The new model is compiled with a categorical_crossentropy loss function and Adam optimizer. The new model is now trained on an existing dataset using a batch size of 32 and 50 epochs.

b: Boosting

In the case of boosting a custom class is used to test both network architectures. The prediction for each input sample is determined by selecting the class with the highest summed probability. Outputs are aggregated to improve prediction accuracy.

IV. EXPERIMENTAL RESULTS

A. TEST DATA

The test data comprises randomly selected samples from the pool of 800 audio clips. This ensures a robust evaluation of the model's generalization capabilities. 160 clips are selected for validation and testing. The following evaluation metrics are used: Accuracy, Precision, Recall and F1 Score.

B. CLSTM

The baseline CLSTM results seen in Table 1 show good performance in terms of Accuracy, Precision, Recall, F1-Score and Loss. The highest accuracy reached is 72%, so this forms a solid baseline. The confusion Matrix seen in Figure 4 shows roughly four misclassified emotions, but overall shows a good distribution.

C. DENDCAPS

Looking at the DendCaps network results seen in Table 2 the highest accuracy achieved is 74%. In terms of the Confusion Matrix seen in Figure 5, the distribution is even, except for a few falsely classified labels.

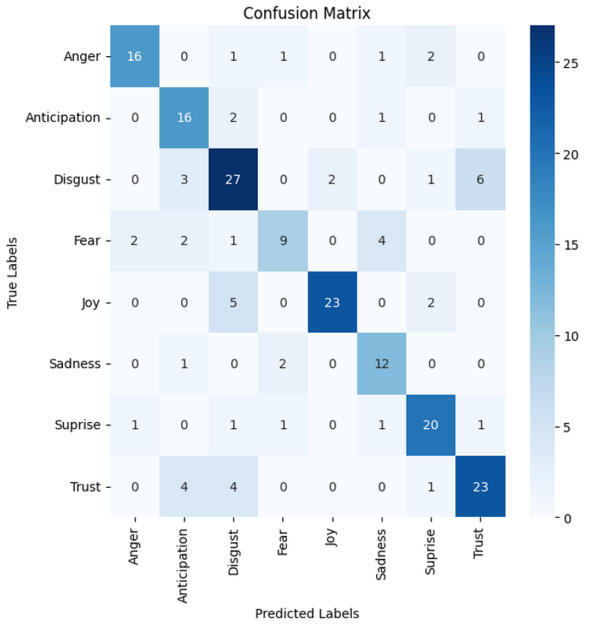


Figure 4. CLSTM Confusion Matrix.

Table 2. DendCaps Results.

Test	Accuracy	Precision	Recall	F1Score	Loss
1	68%	0.84%	0.80%	0.71%	0.67%
2	70%	0.82%	0.79%	0.79%	0.72%
3	73%	0.84%	0.78%	0.81%	0.63%
4	74%	0.81%	0.77%	0.75%	0.62%

D. ENSEMBLE

1) Stacking

The first Ensemble method results can be seen in Table 3. A high accuracy of 77% is achieved and the confusion matrix distribution is even. The highest false classification was for Disgust classified as trust. The confusion matrix can be seen in Figure 6.

Table 3. Stacking Results.

Test	Accuracy	Precision	Recall	F1Score	Loss
1	65%	0.84%	0.80%	0.71%	0.60%
2	67%	0.82%	0.79%	0.79%	0.61%
3	77%	0.84%	0.78%	0.81%	0.63%
4	75%	0.81%	0.77%	0.75%	0.63%

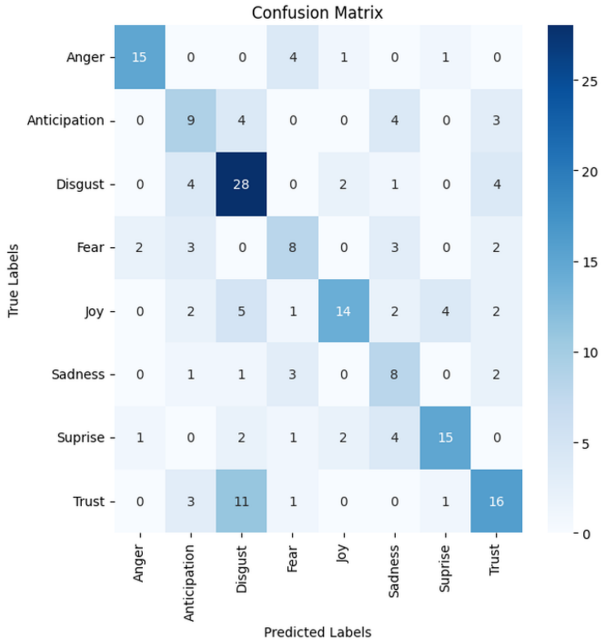


Figure 5. DenCaps Confusion Matrix.

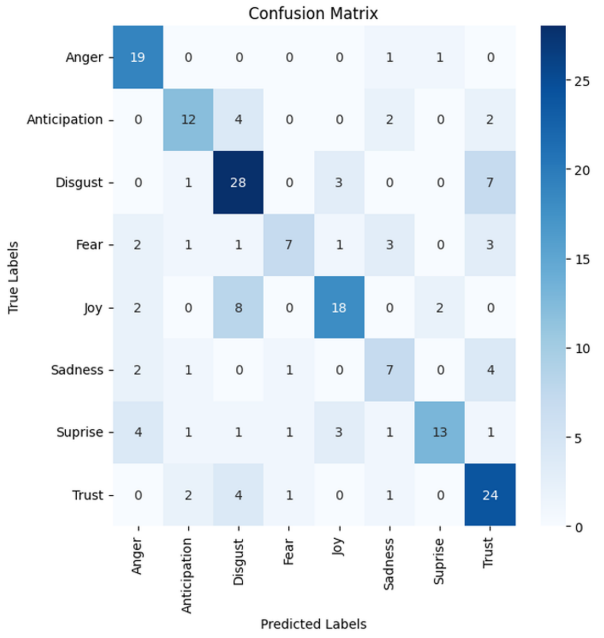


Figure 6. Stacking Confusion Matrix.

2) Boosting

For the second ensemble method namely boosting the results can be seen in Table 4. Accuracy wise it is 72%, which is decent. No confusion matrix can be formed because of the nature of the boosting method.

Table 4. Boosting Results.

Test	Accuracy	Precision	Recall	F1Score	Loss
1	72%	0.66%	0.9%	0.76%	0.87%
2	72%	0.67%	0.6%	0.63%	0.82%
3	72%	0.70%	0.72%	0.66%	0.83%
4	72%	0.81%	0.52%	0.63%	0.74%

E. DISCUSSION OF RESULTS

The CLSTM baseline has solid results for accuracy, Precision, Recall and F1-Score. Building on this, the DendCaps network outperforms it in terms of accuracy. The Precision, Recall and F1-Score are also very good. In Tests 3 and 4, the DendCaps model achieves higher recall values compared to the Conv LSTM model. This suggests that DendCaps is better at capturing instances of certain emotional states, possibly the ones with lower representation in the data. The DendCaps model consistently has lower loss values across all tests. Lower loss indicates better convergence and a better fit to the training data. In Test 3, the DendCaps model achieves a higher F1-Score, indicating a better balance between precision and recall, which can be important for tasks like speech emotion recognition. DendCaps may generalize better to the specific characteristics of the dataset used in these tests. Subsequently, these two network models (CLSTM and DendCaps) are saved and with the two ensemble methods, namely stacking and boosting, combined. Boosting yields decent results, but CLSTM stacked with DendCaps gives the best results. In terms of balanced performance, stacking consistently maintains a balance between precision and recall, which is crucial for tasks like speech-emotion recognition. While DendCaps may outperform in specific tests, Stacking provides a more balanced and consistent performance. Generilization-wise, Stacking shows the ability to generalize well to different aspects of the data across tests. It is not overly sensitive to variations in the dataset, as seen in Test 3, where it achieves the highest accuracy. For the competitive metrics, Stacking performs competitively in terms of accuracy, precision, recall, and F1-Score. It may not have the highest values in all tests, but it offers a well-rounded performance across different metrics. Lastly, Stacking has relatively stable loss values, indicating good convergence and robustness. Stacking emerges as the most favourable method among DendCaps, Stacking, and Boosting due to its balanced and consistent performance, achieving competitive accuracy, precision, recall, and F1-Score across various tests. Stacking's ability to generalize well to different aspects of the data while maintaining stability in loss values makes it a robust choice for speech emotion recognition. In terms of the various metrics, very high scores have been achieved ranging from 60% to 80% [42] using speech corporas like the well-known RAVDESS(Ryerson Audio-Visual Database of Emotional Speech and Song) [43], IEMOCAP (Interactive Emotional Dyadic Motion Capture) [44] and Emo-DB (Emotional Database) [45]. For the custom-created Afrikaans [36] speech corpora CNN,

RNN and LSTM produce lower and inferior results than CLSTM, DendCap and DendCaps Ensemble methods. One must bear in mind that the number of emotions, 7 for RAVDESS, 5 for IEMOCAP and 7 for EmoDB as opposed to 8 for the Afrikaans corpora plays a role. Another factor is the amount of samples. RAVDESS has 7356, EmoDB has 535, IEMOCAP 1,150 and Afrikaans 800 samples/audio clips. Taking these factors into account the model performance compared to other simulations set a high precedent and offers exciting research opportunities.

V. CONCLUSIONS

In this article, we investigated the performance and accuracy of using a Dendritic Layer Capsule network hybrid combined with ensemble methods. The aim of increased accuracy was achieved with this study. It is hypothesized that the model proved more effective because of the following. Hierarchical feature extraction, complex spatial relationships within speech data and capturing nuanced emotional cues. The ensemble methods and in this case stacking improved accuracy by increasing the overall robustness of the stacked model. The relevance of this is a more accurate model to be used by other researchers as well. Future work will be testing additional ensemble methods using DendCaps and CLSTM. We will also consider a Dendritic LSTM Capsule hybrid model. From our perspective, other model combinations and ensemble methods will build on the success of this model.

VI. ACKNOWLEDGEMENTS

This research was supported partially by the South African National Research Foundation (Grants nos. 120106, 41951, and 132797) and the South African National Research Foundation Incentive (Grant no. 132159).

References

[1] K. Hartmann, I. Siegert, D. Philippou-Hübner, and A. Wendemuth, "Emotion detection in hci: From speech features to emotion space," IFAC Proceedings Volumes, vol. 46, no. 15, pp. 288–295, 2013. [Online]. Available: http://dx.doi.org/10.3182/20130811-5-us-2037.00049

[2] M. Mohan, P. Dhanalakshmi, and R. S. Kumar, "Speech emotion classification using ensemble models with mfcc," Procedia Computer Science, vol. 218, pp. 1857–1868, 2023, iD: 280203. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050923001631

[3] R. A. M. T. H. A. A. O. Mohammad Subhi Al-Batah, Mazen Alzyoud, "Early prediction of cervical cancer using machine learning techniques," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 08, no. 04, pp. 357 – 369, 2022.

[4] T. Ahammad, "Risk factor identification for stroke prognosis using machine-learning algorithms," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 08, no. 03, pp. 282 – 296, 2022.

[5] F. A. Meaad Alrehaili, "Development of ensemble machine learning model to improve covid-19 outbreak forecasting," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 08, no. 02, pp. 159 – 169, 2022.

[6] S. O. A. E. Leen Al Qadi, Hozayfa El Rifai, "A scalable shallow learning approach for tagging arabic news articles," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 06, no. 03, pp. 263 – 280, 2020.

[7] X. Wu, S. Liu, Y. Cao, X. Li, J. Yu, dongyang dai, X. Ma, S. Hu, Z. Wu, X. Liu, and H. Meng, Speech Emotion Recognition Using Capsule Networks, 2019.

[8] I. Shahin, N. Hindawi, A. B. Nassif, A. Alhudhaif, and K. Polat, "Novel dual-channel long short-term memory compressed capsule networks for emotion recognition," Expert Systems with Applications, vol. 188, p. 116080, 2022, iD: 271506. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421014172

[9] X. Wu, Y. Cao, H. Lu, S. Liu, D. Wang, Z. Wu, X. Liu, and H. M. Meng, "Speech emotion recognition using sequential capsule networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3280–3291, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:243766683

[10] L. T. Van, Q. H. Nguyen, and T. D. T. Le, "Emotion recognition with capsule neural network," Computer Systems Science and Engineering, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:243997015

[11] J. Poncelet and H. V. hamme, "Multitask learning with capsule networks for speech-to-intent applications," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8494–8498, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:211146449

[12] X. Wu, S. Liu, Y. Cao, X. Li, J. Yu, D. Dai, X. Ma, S. Hu, Z. Wu, X. Liu, and H. M. Meng, "Speech emotion recognition using capsule networks," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6695–6699, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:146062005

[13] S. Gao, M. Zhou, Y. Wang, J. Cheng, H. Yachi, and J. Wang, "Dendritic neuron model with effective learning algorithms for classification, approximation, and prediction," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, pp. 601–614, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:51625695

[14] S. Gao, M. Zhou, Z. Wang, D. Sugiyama, J. Cheng, J. Wang, and Y. Todo, "Fully complex-valued dendritic neuron model," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, pp. 2105–2118, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:237432737

[15] Z. Wang, S. Gao, J. Wang, H. Yang, and Y. Todo, "A dendritic neuron model with adaptive synapses trained by differential evolution algorithm," Computational Intelligence and Neuroscience, vol. 2020, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:210954439

[16] J. Ji, C. Tang, J. Zhao, Z. Tang, and Y. Todo, "A survey on dendritic neuron model: Mechanisms, algorithms and practical applications," Neurocomputing, vol. 489, pp. 390–406, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247491484

[17] S. Gao, M. Zhou, Z. Wang, D. Sugiyama, J. Cheng, J. Wang, and Y. Todo, "Fully complex-valued dendritic neuron model," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, pp. 2105–2118, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:237432737

[18] Y. Yu, Z. Lei, Y. Wang, T. Zhang, C. Peng, and S. Gao, "Improving dendritic neuron model with dynamic scale-free network-based differential evolution," IEEE/CAA Journal of Automatica Sinica, vol. 9, pp. 99–110, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:239041270

[19] J. Ji, C. Tang, J. Zhao, Z. Tang, and Y. Todo, "A survey on dendritic neuron model: Mechanisms, algorithms and practical applications," Neurocomputing, vol. 489, pp. 390–406, 2022. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2021.08.153

[20] E. Egrioglu, E. Baş, and M.-Y. Chen, "Recurrent dendritic neuron model artificial neural network for time series forecasting," Information Sciences, vol. 607, pp. 572–584, 2022, iD: 271625. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025522005941

[21] J. M. Macías-Macías, J. A. Ramírez-Quintana, M. I. Chacón-Murguía, A. A. Torres-García, and L. F. Corral-Martínez, "Interpretation of a deep analysis of speech imagery features extracted by a capsule neural network," Computers in biology and medicine, vol. 159, p. 106909, 2023, iD: 271150. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482523003748

[22] H. H. Gul, E. Egrioglu, and E. Bas, "Statistical learning algorithms for dendritic neuron model artificial neural network based on sine cosine algorithm," Information Sciences, vol. 629, pp. 398–412, 2023, iD: 271625. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025523001792

[23] D. Morrison, R. Wang, and L. C. D. Silva, "Ensemble methods for spoken emotion recognition in call-centres," Speech Communication, vol. 49, no. 2, pp. 98–112, 2007, iD: 271578. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639306001713

[24] M. Alrehaili, F. Assiri, and K. Omari, "Development of ensemble machine learning model to improve covid-19 outbreak forecasting," Jordanian

Journal of Computers and Information Technology, vol. 8, no. 2, pp. 1–169, Jun 1, 2022. [Online]. Available: https://search.proquest.com/docview/2672019932

[25] A. Agga, A. Abbou, M. Labbadi, Y. E. Houm, and I. H. O. Ali, "Cnn-lstm: An efficient hybrid deep learning architecture for predicting short-term photovoltaic power production," Electric Power Systems Research, vol. 208, p. 107908, 2022, iD: 271091. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378779622001389

[26] R. Yang, S. K. Singh, M. Tavakkoli, N. Amiri, Y. Yang, M. A. Karami, and R. Rai, "Cnn-lstm deep learning architecture for computer vision-based modal frequency detection," Mechanical Systems and Signal Processing, vol. 144, p. 106885, 2020, iD: 272413. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888327020302715

[27] K. Adewole, A. Balogun, M. Raheem, M. Jimoh, R. Jimoh, M. Mabayoje, F. Hamza, A. Akintola, and A. Gbolagade, "Hybrid feature selection framework for sentiment analysis on large corpora," Jordanian Journal of Computers and Information Technology, vol. 7, no. 2, pp. 1–151, Jun 1, 2021. [Online]. Available: https://search.proquest.com/docview/2672361477

[28] T.-Y. Kim and S.-B. Cho, "Web traffic anomaly detection using c-lstm neural networks," Expert Systems with Applications, vol. 106, pp. 66–76, 2018. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2018.04.004

[29] S. Ravuri and A. Stolcke, "Recurrent neural network and lstm models for lexical utterance classification," 2015. [Online]. Available: http://dx.doi.org/10.21437/interspeech.2015-42

[30] S. Tao, Y. Todo, T. Zheng, B. Li, Z. Zhang, and R. Inoue, "A novel artificial visual system for motion direction detection in grayscale images," Mathematics, vol. 10, p. 2975, 2022.

[31] Y. Chen, L. Li, W. Li, Q. Guo, Z. Du, and Z. Xu, "Fundamentals of neural networks," pp. 17–51, 2024. [Online]. Available: http://dx.doi.org/10.1016/b978-0-32-395399-3.00008-1

[32] M. K. Patrick, A. F. Adekoya, A. A. Mighty, and B. Y. Edward, "Capsule networks – a survey," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 1, pp. 1295–1310, 2022, iD: 280416. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157819309322

[33] S. J. Pawan and J. Rajan, "Capsule networks for image classification: A review," Neurocomputing, vol. 509, pp. 102–120, 2022, iD: 271597. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231222010657

[34] R. PLUTCHIK, Chapter 1 - A GENERAL PSYCHOEVOLUTIONARY THEORY OF EMOTION, ser. Theories of Emotion. Academic Press, 1980, pp. 3–33, iD: 303393. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780125587013500077

[35] A. Mondal and S. S. Gokhale, "Mining emotions on plutchik's wheel," 2020. [Online]. Available: http://dx.doi.org/10.1109/snams52053.2020.9336534

[36] M. Norval and Z. Wang, "Creation of an afrikaans speech corpora for speech emotion recognition," 2022. [Online]. Available: http://dx.doi.org/10.1109/raai56146.2022.10092988

[37] V. Singh and S. Prasad, "Speech emotion recognition system using gender dependent convolution neural network," Procedia Computer Science, vol. 218, pp. 2533–2540, 2023, iD: 280203. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050923002272

[38] J. Gondohanindijo, E. Noersasongko, and D. R. M. Setiadi, "Multi-features audio extraction for speech emotion recognition based on deep learning," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 14, no. 6, /23/30 2023. [Online]. Available: https://thesai.org/Publications/ViewPaper?Volume=14&Issue=6&Code=IJACSA&SerialNo=23

[39] A. Maxwell, T. Warner, and L. A. Guillén, "Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—part 1: Literature review," Remote Sensing, vol. 13, p. 2450, 2021.

[40] D. Powers, "Evaluation: From precision, recall and f-factor to roc, informedness, markedness correlation," Mach.Learn.Technol., vol. 2, 2008.

[41] M. Vakili, M. Ghamsari, and M. Rezaei, Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification, 2020.

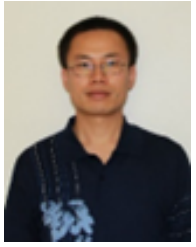
[42] D. Issa, M. Fatih Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," Biomedical Signal Processing and Control, vol. 59, p. 101894, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809420300501

[43] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," PloS one, vol. 13, no. 5, p. e0196391, 2018.

[44] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," Language resources and evaluation, vol. 42, pp. 335–359, 2008.

[45] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss et al., "A database of german emotional speech." in Interspeech, vol. 5, 2005, pp. 1517–1520.

MICHAEL NORVAL obtained his Masters degree in Electronic Engineering, University of South Africa, South Africa, in the year of 2019. He is a PhD student and his research interests are Deep Learning Neural Network and Speech Emotion Recognition.



ZENGHUI WANG received the Ph.D. degree in control theory and control engineering from Nankai University, China, in 2007. Currently, he is a Professor with the Department of Electrical and Mining Engineering, University of South Africa, Florida, South Africa. His research interests include model predictive control, nonlinear control, engineering optimization, image/video processing, artificial intelligence, chaos, and industry 4.0.