

Perspective Chapter: Emotion Detection Using Speech Analysis and Deep Learning

Alexander I. Iliev

Abstract

Speech reflects the sentiment and emotions of humans. People can identify the emotional states in speech utterances, but there is a higher chance of perception error, which is generally termed as human error to identify the proper emotion when only using speech signals. Thus, artificial intelligence plays an important role in the detection of emotion through speech. Deep Learning is the subset of Machine Learning (ML) and artificial intelligence through which speech signal processing can be performed and the detection of emotions can be accomplished using speech. In this chapter, the classifiers of Machine Learning and Deep Learning will be reviewed. From the comparison in various studies and performances we will conclude what methods work better than others. We will discuss the limitations of these approaches as well. Accuracy scores will be discussed for each proposed system.

Keywords: emotion recognition, emotional intelligence, speech analysis, deep learning, machine learning

1. Introduction

1.1 Paul Eckman

We are expressing our emotions through text, speech, songs, facial expressions, and body language. Recognizing emotions from text is gaining more and more popularity. Emotions expressed in terms of text are many times confusing and unexpected because they mostly depend on both context and language. The detection of human emotions from one's image or video recording then further classifying it into one of several emotion categories is still rather challenging task and many researchers proposed various methods to achieve this goal. The general framework of emotion detection through images is given in **Figure 1**. It may contain blocks such as: image preprocessing, face detection, facial landmark detection, feature vector creation, emotion classification, and finally the output of the system.

The first stage is used to remove noise present in the image, contrast adjustment and/or image resizing, if required. The second stage contains the detection of face from the given image and removing unwanted portion. In the third stage facial

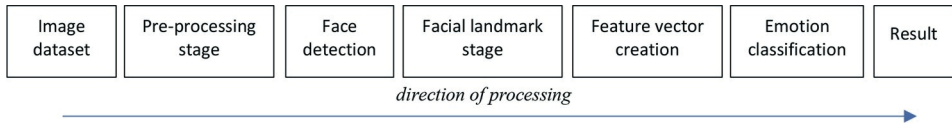


Figure 1.
General framework of emotion detection through images.

landmark detection takes place. This process includes eyes detection, nose detection, eyebrows detection, and hips detection. After detecting the landmarks, in the next stage feature vector is formed which is given to the last stage i.e., emotion classification. This stage outputs the category of the detected emotion in the given input image of a person.

Consistent with Don Hockenberry and Sandra E. Hockenberry, emotions can be thought of as a mentally complex phenomenon that involves three distinct psychological states, subject, response, and expression. But in 1972, Paul Ekman was able to classify these states into 6 different expressions that we call Emotions. He classifies them as:

- Fear
- Disgust
- Anger
- Surprise
- Happiness
- Sadness

He explains that emotions are a result of an automated response that is generated in reply to a speech or any form of information that has been relayed and was found important. These replies are influenced by our years of evolution and past experiences.

Simply put, emotions help us prepare for expected and unexpected events without even thinking about it. Although people are separated by various forms such as culture, language, and geographical boundaries, the 6 emotions mentioned here are what connect us and are believed to be expressed in the same way all over the world.

The image shown in **Figure 2** perfectly illustrates the different types of emotions that Paul Eckman has specified. There are various situations that can trigger any of the above-mentioned emotions. The situations can be:

- Any physical occurrence
- A social gathering
- Recurrence of nostalgic about any previous event
- Talking about an experience



Figure 2.
 Six different emotional states from top left to bottom right: Joy, anger, disgust, sadness, surprise, and fear [source: <https://www.theatlantic.com/>].

1.2 Robert Plutchik

Although these being the common factors, it does not stop here. It may vary person to person and an emotion expressed by one individual may or may not be the same as for another individual for the same given situation. However, Psychologist Robert Plutchik differentiated emotions in a more complex way. He derived the Plutchik model as shown in **Figure 3**.

In 1980, in reply to his early 2D model, Plutchik developed a 3D model of emotions. The wheel can be sought as a map to explain the various complexities associated with every emotion. He explains, emotions start out as simple, but depending on an individual's ethnic and socio-cultural background, can branch to various forms of other high-level emotions. Further, he divides basic emotions in pair of two. They are:

- Sad and Joy
- Disgust and Trust
- Anticipation and Surprise
- Anger and Fear

Taking the help of the 3D wheel illustrated above, we can now divide these emotions to more complex ones.

- Joy + Anticipation = Optimism (Opposite: Disapproval)
- Trust + Joy = Love (Opposite: Remorse)
- Fear + Trust = Submission (Opposite: Contempt)
- Surprise + Fear = Awe (Opposite: Aggression)

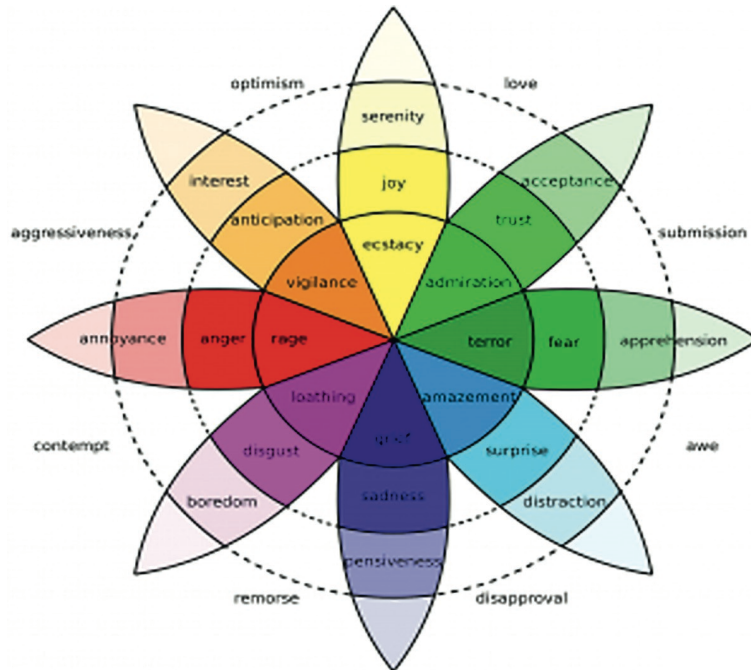


Figure 3.
Robert Plutchik 2D model [source: spring.co.uk].

- Surprise + Sadness = Disapproval (Opposite: Optimism)
- Sadness + Disgust = Remorse (Opposite: Love)
- Disgust + Anger = Contempt (Opposite: Submission)
- Anticipation + Anger = Aggressive (Opposite: Awe)

This wheel of emotion can be considered as a good starting point, but it also has its own limitations.

1.3 The importance of emotions

1.3.1 Expression of feelings

Brain performs some processes which help us associating a type of emotion to a kind of experience we are facing. From the smell of fresh baked bread to a late night horror show, various stimuli can elicit different emotional responses. This is one of the most important tasks the brain performs. It is also the main reason that we feel comfortable with certain situations and react accordingly. For example, listening to an old song makes us feel nostalgic and happy whereas listening to a sad song makes us sad. In today's world, it is very much said that to work and feel healthy, expressions of emotions are necessary. There are many benefits associated with it like,

- Helps solving long-standing problems,

- Decision-making gets easier,
- Depression is eased off,
- Anxiety reduces.

Failure to express emotions may have some detrimental effects too.

- A state of flight or fight.
- It puts a stress on our body.
- Increases heart rate and makes us depressed or anxious.

In research that was carried out in 2013–2014, the authors [1] help us understand how a particular individual is feeling on a given day of the week as well as in given time of the day. The participants particularly were given a questionnaire and told to be filled whenever they were experiencing a given set of emotions at any given time and day. Unsurprisingly, the participants faced the maximum number of emotions from 7 AM to 8 PM. This can be regarded to the fact that most of us are active during this time. Among emotions, Joy was particularly experienced by majority of the population followed by Love and Anxiety.

In another research, authors only focus on teenagers from United States who use social media. There are many comments and observations that can be inferred from the study, but the main points are as follows: around 25% of the teenagers felt less lonely while using social media apps whereas 21% felt more popular while using them. 20% of the people felt confident with themselves whenever they used various social media apps.

Both research papers give us a good idea of the important role emotions play in today's world. Further, we discuss about how globally important emotions are across cultures.

1.3.2 Emotions as a tool for globalization

A case study carried out in 2008 [2], explained various modes of communication between the Filipino staff and their Australian clients. Data generated was in the form of phone calls collected over period of months talked about how the cultural difference could have adverse effects on the success ratio of the staff and the approval ratings of the clients. Ultimately, it was concluded that if there was a good understanding of cultures between the staff, the ratings would have been better.

Due to globalization, there has been a lot of interaction between various cultures. It becomes utmost important to carefully understand what the person in front is speaking to formulate response based on that. Misunderstanding may lead to unfortunate consequences.

Historically, Speech has been the greatest form of communication. World leaders, and great orators have time and again used speech to motivate and inspire the crowd. Emotions play a vital role in such situations and can have a positive impact on the crowd.

1.3.3 Emotional intelligence

Emotional intelligence is the ability to comprehend, organize emotions positively, to ease off the stress, communicate in a better way, relate to others, and fight difficult situations and

relax conflicts. Emotional Quotient also is an important factor in building strong relationships, which brings success at home and office. It also helps to achieve personal milestones. According to a <https://www.helpguide.org/> from July 2021 [3], emotional intelligence is mainly influenced by 3 factors.

- **Self-Management:** This is the ability to control strong feelings and manage our emotions in a better way.
- **Self-Awareness:** Able to understand our own emotions and how it has various effects on thoughts and processes.
- **Managing Relationships:** Can develop good and strong relationships which are long lasting.

1.4 Emotional Process

Another report published at University of Alabama [4], assesses emotion based on 3 factors.

1.4.1 Subjective experiences

Emotions are typically associated with past experiences and can be triggered by certain stimuli. Whether it is a familiar smell leading to a happy emotion or the loss of a loved one causing a more complex emotional response, individuals are always expressing one emotion or another. Moreover, the same experience can evoke an array of emotions across different individuals.

1.4.2 Physiological responses

Everyone has, at one time or another, felt their heart beating fast in situations such as waiting for expected results, which is often expressed as fear. When entering a new relationship, the feeling is often described as butterflies flying in the stomach.

1.4.3 Behavioral responses

This part talks about the actual expression of any emotion. They include anything from Smile, a sigh or laugh depending on situations. There has been numerous research undertaken that explains how facial expressions are universally expressed in the same way.

These expressions are very important to show how an individual is feeling to others, and they are also essential for one's wellbeing.

2. Emotion vs. mood

Affect is termed as general keyword used to describe a broad range of feelings experienced by people. It is a superficial concept that has both emotions and moods

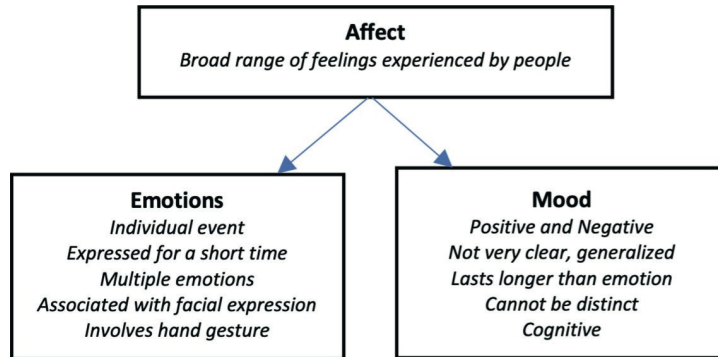


Figure 4.
Emotions vs. mood.

associated with it. Emotions are just some strong feelings experienced or expressed at someone. Mood is less intense and does not need a context to be true.

Many researchers have said that emotions are temporary as compared to moods. For example, if something goes wrong and not as expected by an individual, he/she will get angry but that anger feeling fades quickly. But if there is a bad mood, it will stay on for a longer period.

Emotions can be triggered by a situation or something someone has said, causing an individual to feel a range of emotions such as anger, happiness, or sadness. These are contextualized as strong feelings. Whereas mood is necessarily not expressed at a person. They might be the result of emotions not being resolved quickly. They generally happen when you lose concentration over a certain thing. For example, if a colleague questions you on the way an individual spoke to a client, that individual might get angry. You show emotion at a specific thing, but as that dissipates, your emotion gets generalized.

As **Figure 4** shows, affect is a highly generalized term. There are significant differences between Mood and Emotions. This image also describes how emotions and mood can influence each other. If an emotion is intense and stable enough, it can transform into a mood. Getting a job make us happy, and this feeling can last for several days. Similarly, if you are experiencing a positive or negative mood, you may feel strong positive or negative emotions in response to situations that arise at that moment. Some aspects to be considered for emotion are:

- **Intensity:** Individuals give varied feelings for same emotional stimuli. There may be some people who almost never show any feelings. Those people rarely get angry. Also, there would be people who are highly emotional and show emotions all day long.
- **Frequency and Period:** Emotions cannot be expressed for a long period of time. Also, the emotional demands are too hard to maintain.

3. Speech analysis and properties

3.1 Properties of speech signal

The resonance architecture of the auditory tract, specifically the two bottom resonances known as formants, can be easily examined by drawing an “envelope” over the spectrum.

This involves drawing a continuous line immediately above the spectrum, as shown in the picture on the right [5]. As a result, researchers receive the spectral envelope that describes the macro-shape of a voice signal's spectrum and is frequently used to describe speech signals. The basic frequency of a spoken transmission, or its lack thereof, conveys a great deal of information. Voiced and unvoiced parts of speech are those that have or do not have a vibration in the vocal cords. Researchers classify phonemes into voiceless or voiced categories based on their predominance. A speech signal with its spectral properties is shown in **Figure 5a** and **b** [6].

As mentioned in [6], when developing a speech processing system, data will be used for:

- a. *Speech analysis* in order to observe the signal from speech production process. This way we can identify different properties of speech that can help us improve performance and increase our comprehension of features importance.
- b. *Use Machine Learning* so that we can successfully train any given model for specific automation and smart development of subsequent systems.
- c. *Performance evaluation* of the speech system. This stage is vast and can take different shapes and forms as it is developed for various needs. Therefore, the speech corpus must be such that can be used both in training, validation, and testing for the specific environment for which it is intended to be used.

A sample, limited-vocabulary speech recognition data for speech commands can be found in [7].

3.2 Speech linguistic structure

Several linguistic features of speech, including consonants and phrases, have written language equivalents. Nevertheless, it is critical to distinguish between the two: Speech signals always are uninterrupted and non-categorical, but written language is made up of discrete category pieces. This really is owing to the motor activity in speech generation, which acts in real time and at a limited pace while being constrained by physiological and neurophysiologic restrictions [8–10].

Therefore, the generated voice signal is similarly constant in time and frequency.

Furthermore, since voice conveys additional data in regards of how items are spoken and the qualities of the speaker, recorded language is deficient in these areas. Written language, on the other hand, employs lexical and grammatical techniques, as well as unique characters, to distinguish finer-grained interpretations, such as communicating emotional content or distinguishing inquiries from assertions. The speech structure is shown in **Figure 6** [6].

3.3 Speech waveforms

The speech signals are categorized as audio or sound pulses that can travel from one place to another regarding the internal energy in it. The waveform of speech signals contains different components such as the frequency, magnitude, phase etc. In the present environment, we are particularly concerned with waveforms processing and analysis in

digital systems. As a result, researchers would always presume that the acoustic voice signals were caught by a microphone and translated to digital format.

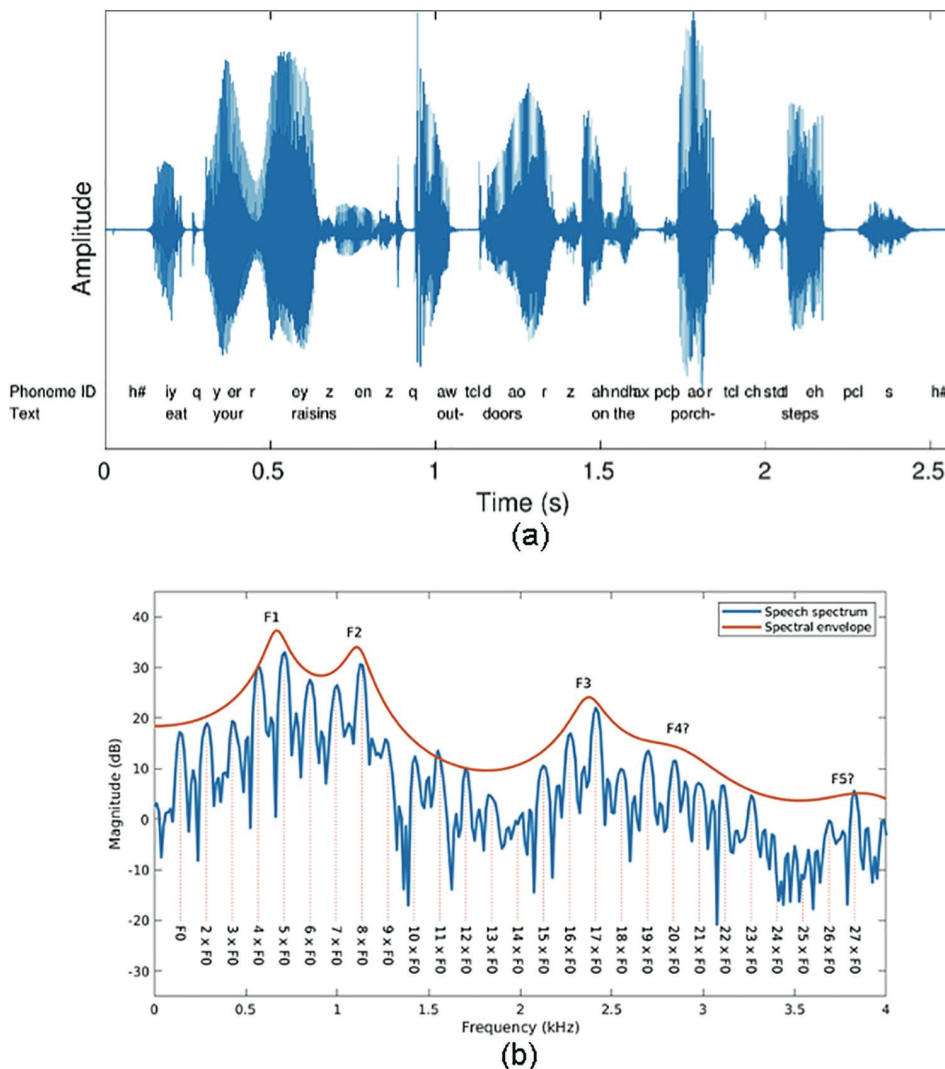


Figure 5.
 (a) Speech signal in time domain [6] (b) Formants in a speech signal [6].

The consonants in speech signals contain crucial information, that ranges from 300 to 3500 Hz, implying that a lower threshold for sampling frequency is roughly 7 or 8 kHz [11]. Most power, though, stays below 8 kHz, implying that wide band, that really is, a sampling frequency of 16 kHz, is enough for most applications.

3.4 Speech signal windowing

Windowing of speech signals involves dividing or sub-sampling speech pulses into several short segments [12]. Windowing functions are seamless operations that return to zero at the edges. With the application of the windowing process, the audio signals can be

truncated into pieces from which the overall features and properties of the speech signals can be identified rather than considering a long speech pulse. A simple window is shown in **Figure 7** [6].

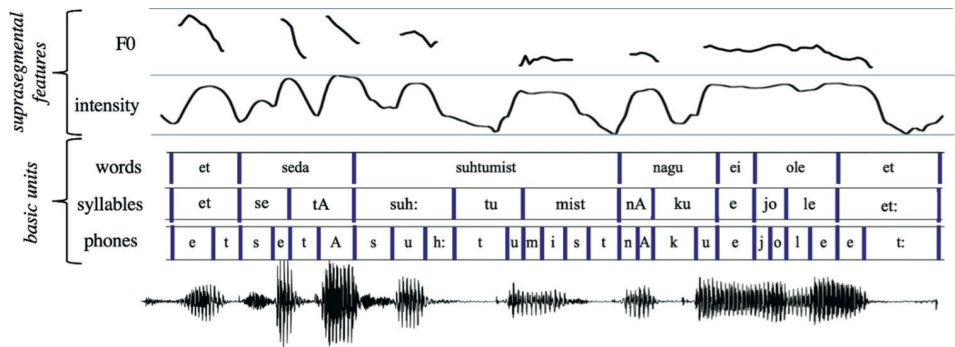


Figure 6.
Hierarchical organization of speech in terms of phones, syllables, and words [6].

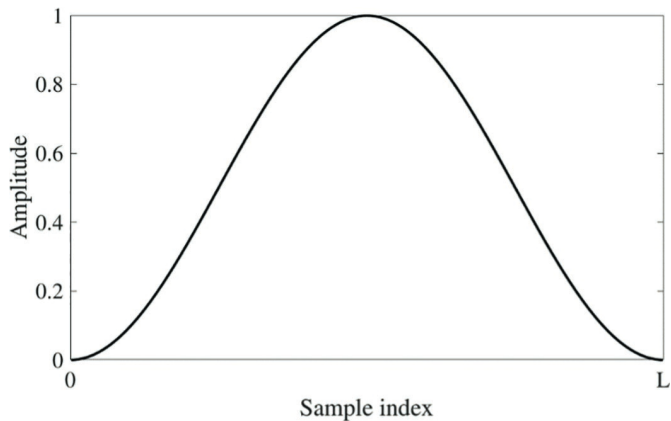


Figure 7.
Speech signal window [6].

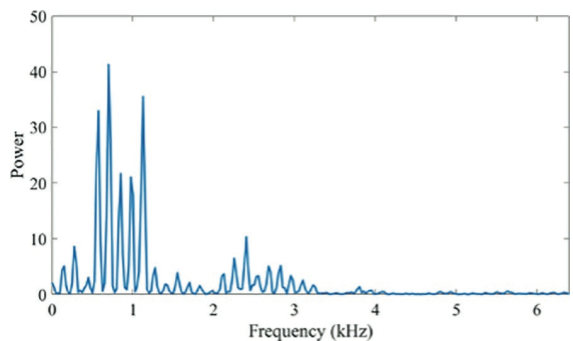


Figure 8.
Windowed speech signal [6].

In **Figure 8**, we see a window applied over a speech sample with its resulting signal in time and frequency domains respectively.

3.5 Speech spectrogram

Speech signals, on the other hand, are non-stationary indicators. When we convert a spoken phrase to the frequency response, then get a spectrum that is an aggregate of all phonemes in the phrase, although we typically want to view the spectrum of every single phoneme independently [12]. Researchers can concentrate on signal qualities at a certain point in time by dividing the information into shorter parts. This type of segmentation already was covered in the windowing part. One of the most utilized techniques in speech processing and analysis is the Short-Time Fourier Transform (STFT). It illustrates how frequency components change over time. One of the advantages of STFTs is that its characteristics have a simple physiological and understandable explanation.

3.6 Speech cepstrum and Mel-frequency cepstral coefficient (MFCC)

Cepstrum refers to the properties of the speech signals that help detect the information of the speech signal [12]. It can be obtained by extracting the features from the speech signals so that the essential components of the speech audio can be identified [6]. This component can be extracted using the application of speech processing and the Mel-Frequency Cepstral Coefficient (MFCC) method where the cepstral components are achieved through numerical data. This is essential in Machine Learning and Deep Learning as the algorithms require numerical features to classify and detect the type of speech and to identify the emotions.

4. Speech emotion detection using machine learning

According to Qing and Zhong [13], the rise of big data handling in recent times, coupled with the continual improvement of computers' computational power and the ongoing improvement of techniques, has led to significant advancements in the field. Also, with the advancement of artificial intelligence studies, individuals are not always content that the computer does have the same problem-solving abilities as the human mind. Still, they also wish for a much more humanized artificial intelligence with the same emotions and character. It may be utilized in students' learning to recognize students' feelings in real time and analyze them appropriately and in intelligent human-computer interaction to detect the speaker's emotional shifts in real time. Researchers primarily investigate the Mel-Cepstral Coefficient settings and K-Nearest Neighbor algorithm (KNN) for speech signals and implement MFCC extraction of features using MATLAB and emotion classification using the KNN method. The CASIA corpus is utilized for training and validation, and it eventually achieved 78% accuracy. As per Kannadaguli and Bhat [14], humans see feelings as physiological changes in the composition of consciousness caused by various ideas, sentiments, sensations, and actions. Although emotions vary with an individual's familiarity, they remain consistent with attitude, color, character, and inclination. Researchers employ Bayesian and Hidden Markov Model (HMM) based techniques to study and assess the effectiveness of speaker-dependent emotion identification systems. Because all emotions may not have the same prior probability, researchers must calculate the conditional probability by multiplying the pattern's chances by each class's previous distribution and dividing by the pattern's likelihood function derived by summing its potential for all categories. An emotion-based information model is constructed using the acoustic-phonetic modeling technique to voice recognition. Following that, the template classifier and pattern recognition are built using the three probabilistic methodologies in Machine Learning.

As described by Nasrun and Setianingsih [15], emotions in daily language are often associated with feelings of anger or rage experienced by an individual. Nevertheless, the fact that action is predisposed as a property of emotions does not necessarily make things simpler to describe terminologically. Speech is a significant factor in determining one's psychological response. The Mel-Frequency Cepstral Coefficient (MFCC) approach, which involves extracting features, is commonly used in human emotion recognition system that are based on sound inputs. Support Vector Machine (SVM) is a novel data categorization approach developed in the 1990s. SVM is guided Machine Learning, frequently used in various research to categorize human voice recognition. The RBF kernel has been the most often used kernel in SVM multi-Class. This is because SVM employs the Radial Basis Function (RBF) seed to improve accuracy. This report's most incredible accuracy ratio was 72.5%.

According to Mohammad and Elhadeif [16], emotion recognition in speech may be defined as perceiving and recognizing emotions in human communication. In other respects, speech- emotion perception means communicating with feelings between a computer and a human. The proposed methodology comprises three major phases: signal pre-processing to remove noise and decrease signal throughput, feature extraction using a combination of Linear Predictive Rules and 10-degree polynomial Curve fitting Coefficients over the periodogram power spectrum feature of the speech signal, and Machine Learning that utilizes various machine learning algorithms and compares their overall accuracy to determine the best accuracy. Several of the causes are that the recognition approach selects the best elements for a method to be powerful enough to distinguish between different emotions. Another factor is the variety of languages, dialects, phrases, and speaking patterns. As per Bharti and Kekana [17], speech conveys information and meaning via pitch, speech, emotion, and numerous aspects of the Human Vocal System (HVS). Researchers suggested an outline that recognizes sentiments using Speech Signal (SS) with the highest average accuracy and effectiveness when compared to techniques such as Hidden Markov Model and Support Vector Machine. The detection step can be easily implemented on various mobile platforms with minimal computing effort, as compared to previous approaches. The ML model has been trained successfully using the Multi-class Support Vector Machine (MSVM) approach to distinguish emotional categories based on selected features. In machine learning, Support Vector Machines (SVMs) are popular models used for classification and regression analysis. They're especially known for their effectiveness in high-dimensional spaces. However, traditional SVMs are inherently binary classifiers. When there are more than two classes in the dataset, adaptations like MSVMs are used, which can handle multi-class classification problems. The MSVM classification was used to extract features Gammatone Frequency Cepstral Coefficients (GFCC) and remove elements to achieve a high success rate of 97% on the RAVDESS data set (ALO). The GFCC is a feature extraction method used often in the field of speech and audio processing. The GFCC features try to mimic the human auditory system, capturing the phonetically important characteristics of speech, and are robust against noise. Whenever extracted features using MFCC are applied to existing databases, all classifiers achieve an accuracy of 79.48%.

As described by Gopal and Jayakrishnan [18], emotions are a very complicated psychological phenomenon that must be examined and categorized. Psychologists and neuroscientists have performed extensive studies to analyze and classify human emotions over the last two decades. Emotional prosody is used in several works. The goal of this project was to develop a mechanism for annotating novel texts with appropriate emotion. With the SVM classifier, a supervised method was used. The One-Against-Rest technique

was utilized in a multi-class SVM architecture. The suggested approach would categorize Malayalam phrases into several emotion classes such as joyful, sad, angry, fear, standard, etc., using suitable level data with an overall accuracy of 91.8%. Throughout feature vector choice, many aspects such as n-grams, semantic orientation, POS-related features, and contextual details are analyzed to determine if the phrase is conversational, or a question.

5. Predictive visual analysis of speech data using machine learning algorithms

Goyal and Rathore [19] state that there is a vast amount of digital and social media data available on the internet, including platforms like Twitter, LinkedIn, message boards, blogging sites, customer groups, and feedback on products. In the modern environment, product feedback has become quite vital. The specific approaches, such as Max Monitoring, SVM, Naive Bayes, logistic regression, or KNN classification, can be used to detect faces and classify the displayed mood. The approaches mentioned here are utilized for document-level categorization. In this study, we apply the ME-based text technique to examine the level sense of neural networks using recursive nerve sensor networks that evaluate emotion at the phrase level. According to the investigation, real. emotion, including audio and video, is identified by specific words with a senti-ment component. This study lays the groundwork for future research to concentrate on phrases that influence decision-making and discover generic public mass assessments.

According to Ali et al. [20], there are numerous methods to define microaggressions (MAs). Derogatory stereotype representations that are “put setbacks” by such an offender often are subtle, shocking, frequently reflexive, and nonverbal encounters. Computerized MA detection introduces a new and difficult topic in Natural Language Processing (NLP) research and sentiment evaluation. A better understanding of how machine algorithms are constructed and how MAs are categorized in writing can help to enhance our understanding of sentiment in documents. Regarding MA identification, the outcomes of the two classification tests were encouraging. The characteristics recovered from the annotated dataset reinforce this point, with phrases/attributes that can be seen as slightly racist, such as darkish and minority, being picked in the first 20. There are not very harsh or obscene words/blasphemes on the listing. While employing these parameters, advancements in classification accuracy have also shown encouraging outcomes, with similar trends throughout all systems examined.

As stated by Tasha et al. [21], an emotion detection element could be added to spoken conversation systems. It can be used as a component of the interaction program’s architecture to impact the interaction program’s responsiveness to the customer’s verbal statements or to improve the user interface in those other ways. The traditional GMM technique has the worst efficiency of 38%. Correspondingly, using a DNN or an ELM to assign various values to distinct characteristics improves accuracy to 48 and 51.6%. The DNN-ELMK algorithm continues to be the highest performer, with 57.9% accuracy. Researchers measured the effectiveness of numerous GMMbased algorithms that calculate the statistics of the complete speech first and then do classification, to our cutting-edge DNN-ELMK technique, which conducts a variety of segments first and then computes statistics. Ultimately, GMM-based algorithms cannot match the accuracy of the DNN-ELMK approach when showing emotion identification on 0.25-second components.

6. Speech emotion detection using deep learning

According to Tariq et al. [22], speech emotion detection has received a lot of interest in recognizing people's emotional states. Speech is an excellent mode of communication for identifying the speakers and many sorts of feelings. Researchers created an IoT system that predicts patients' moods in real time. Researchers used the SED model on actual data.

They discovered that female audio has a higher accuracy of 78% than male performers, who have an accuracy of 71% owing to the purity of their voices. They saw that their network was running quite well. We employed a 2D CNN model using Peak, RMS, and EBU normalization and data augmentation approach to train and test speech emotion detection. They discovered that combining the normalization and augmentation approaches acquired the greatest accuracy, superior to state-of-the-art techniques in audio-based emotions categorization and forecasting.

As stated by Zhang et al. [23], being one of the greatest natural forms of human interaction, speech signals include not just explicit language information but also implicitly paralinguistic data about the speaker. The suggested technique is tested on four available datasets: the Berlin database of the RML audio-visual dataset, German emotional speech (EMO-DB), the eNTERFACE05 audio-visual dataset, and the BAUM-1 s audio-visual set of data. Researchers offer a new automated emotional feature learning technique that combines DCNNs with DTPM. A DCNN is being used to train discriminative segment-level characteristics from triple channels of log Mel-spectrograms, analogous to RGB picture representations. DTPM is intended to combine learned segment-level elements into a universal utterance-level feature extraction for emotion identification.

According to Singh and Sharma [24], sentiments are “strong sentiments arising from one's surroundings, mood, or interactions with others.” Emotions are the most important aspect of human communication in everyday life. Deep Learning approaches surpass shallow classifiers because external classifiers can only acquire high-level characteristics, but Deep Learning methods can develop insights through low-level information. There is a significant increase in accuracy from the SVM approach (86.75) to the LSTM approach (91.75). The CNN design with a two-layer deep network produces the greatest results (95.4%). Generally, SVM has the greatest outcomes in shallow structures. Deep Learning feeds on large amounts of data and would operate considerably better with large amounts of data.

As stated by An and Shi [25], due to urbanization and industrialization, social rivalry is rising in economic building and automation, leading to a dramatic increase in different psychological sources of stress, mental diseases, and mental health issues. A CNN architecture consists of six layers, including two convolution layers, two max-pooling layers, and two fully connected layers. The negative repercussions of a mental health issue are often dramatic and result in outrageous conduct. The testing set included 198 statements written by students with negative feelings and individuals with neutral feelings.

The findings indicated that the training model's accuracy was 78.4%, the test set's accuracy was 70.5%, and the final model's experimental outcomes were more than 70.5%.

According to Mokonyane and Sefira [26], Deep Learning is a Machine Learning approach that mimics the human brain's operations in the analysis of organized and unstructured information for application areas such as translation software, voice recognition, object identification, and many others. Deep Learning techniques may think of making judgments without human intervention. Researchers discovered that the Sigmoid Kernel fails to meet state-of-the-art accuracy, coming in last at 58%, followed by polynomial, linear, and RBF kernels, which achieved state-of-the-art accuracy at 81%, 85%, and 88%, correspondingly. After evaluating the models, deep neural networks outperformed

Machine Learning methods in emotional speaker recognition from voice signals, achieving the most excellent accuracy of 92% and beating state-of-the-art designs.

As stated by Qidwai and Al-Meer [27], human emotion is crucial in humanhuman interaction since it conveys the person's unspoken mood. A CNN is typically composed of three layers: convolution, pooling, and fully linked layers. In CNN, the feature extractors are the convolution and pooling layers. A series of filters are convolved with the input picture in the convolution layer to extract features like vertical or horizontal lines. Emotion recognition garnered interest in human-centered design as computer technology advanced, particularly in Human-Computer Interaction. The suggested model obtains an overall accuracy of 81% on unseen data. Accordingly, it recognizes positive and negative feelings with 87% and 85% accuracy. The accuracy at distinguishing neutral emotion, on the other hand, is just 51%.

According to Gunathilake et al. [28], recently, text-to-speech synthesis has been challenging since the voices produced by these algorithms sound robotic and thus are easily distinguished from human agents. Despite extensive study into creating naturalsounding voices, delivering an emotional speech is a reasonably young topic. Expressive TTS has several uses, such as supporting the visually challenged, and emotion recognition from text is a critical module in this process. The technique of detecting emotions begins with identifying what emotions are. Researchers combine bi-directional longshort memories with an attention layer for higher prediction accuracy. To enhance future outcomes, researchers use text preparation. Researchers run the tests on three different data sets, as well as the algorithms are graded based on their classification results.

As stated by Lin and Yang [29], because of the rapid progress of Machine Learning and deep understanding in recent years, studies on using these technologies to aid in elderly care and children have become widespread. This research offers an intelligent system for distinguishing emotional sounds such as laugh, weep, scream, wail, or sigh to help caretakers comprehend the needs of the elderly and kids. They can receive appropriate care more rapidly. Empirical mode segmentation is utilized to improve the identification and recognition of emotional sounds. Furthermore, deep ensemble learning is used to address the issue of overfitting. Experiments reveal that the suggested technique has a classification accuracy of 91.6%, which is significantly higher than without using EMD. Researchers think that this technology will improve the care of the aged and young. According to Wani and Guna Wan [30], speech emotion recognition is an expanding topic of research currently, and as a result, multiple researchers have developed different technologies in this field. This procedure is required to categorize a voice signal to detect a specific mood. Many people strive to discover aspects of voice signals that range from efficient to salient to discriminative. The algorithms were fed spectrograms created from the speech dataset. As the number of test epochs increased from 500 to 1200 and 1500, the efficiency of both models improved. The provided model Depthwise Separable Convolutional Neural Networks (DSCNN) surpasses the current state-of-the-art model CNN by a wide margin. DSCNN achieved an accuracy of 87.8%, while CNN attained an accuracy of 79.4%. DSCNN is a variant of the standard convolutional neural network (CNN) and is part of the family of convolutional networks which are designed to be more efficient.

In a DSCNN, the convolution operation is split into two separate operations aiming to reduce the model's complexity and size. The first is a depthwise convolution which applies a single filter per input channel. The second is a pointwise convolution, a simple 1x1 convolution, which is used to build new features through computing linear combinations of the input channels.

This structure is a key element of several efficient and compact network architectures such as MobileNet developed by Google, which is used for tasks like object detection and

image segmentation on mobile and embedded devices where computational resources are limited. More work is required to enhance the provided architecture for convincingly recognizing emotions.

7. Contributions and summary

As evident from the facts given, it is clear how important emotions are in our day to day lives. Globalization across the world has accelerated the need for better understanding of emotions. The interactions between various cultures at such a fast pace underlines the importance of expressing ourselves and our intentions to the crowd for maximum effect. Good orators and leaders have always used emotions to drive home their views and have a positive impact on their followers.

Numerous research has been undertaken to understand how important emotions are. They are explained by the multiple case studies discussed in the previous chapters. It is worth to note that emotions can have both positive and negative impact on our health. Thus, researchers advice that an individual should always express emotions from time to time to avoid depression and anxiety. In professional life, understanding your colleague from other country or culture helps in better communication and leads to increased success.

Therefore, we can see how emotions have played a vital role in all aspects of life. Going forward, we see the emergence of AI and how it helps to understand and express emotions even better.

The main contributions of this chapter are that it summarizes the latest state of the art in emotion recognition through various use case in different environments. A short summary is provided next:

8. Machine learning

Machine learning is a branch of AI that makes applications to have better and accurate predictions without hard coding it to do so. It uses historical data to predict new values. Common use cases include recommendation engines, credit detection, fraud, predictive maintenance.

It is necessary as it gives extended view of various behaviors and patterns as well as supporting new products. Almost all corporate giants like Google, Microsoft, IBM, Uber use ML as a core part of its operations. It has become so important to gain advantage over other companies, ML is seen as a solution. There are various types of Machine Learning:

- *Supervised Learning*: In this method, labeled historical data is given based on which models train and learn to associate various results. Then a user provides unseen data and based on metrics, we understand how accurate the model is.
- *Unsupervised Learning*: This type is totally opposite to Supervised learning. The model trains on unlabeled data and learns to associate results on its own. It looks for meaningful results and gives out relevant output.
- *Semi-Supervised Learning*: This is a combination of the two previous types. Some labeled data maybe be given as well as unlabeled data. In this, the model is free to make its own assumptions and give out results.

- **Reinforcement Learning:** The model starts by making a few predictions and outcomes. The user based on what results it receives gives out incentives or positive and negative reactions. The models take in the feedback given by the user and the future predictions are reliant on such feedbacks.

9. Machine learning use cases

The supervised machine learning can be used for a variety of tasks and will require labeled data to give outputs. The various tasks are:

- Binary Classification
- Multi-Class Classification
- Regression Technique
- Ensembling Technique

The unsupervised machine learning does not require labeled data. They analyze unlabeled data to identify patterns that can be utilized to classify data into various categories. Almost all Deep Learning techniques are unsupervised learning techniques. The various tasks for which unsupervised learning technique could be used are.

- Clustering
- Dimension Reduction techniques
- Associate

Reinforcement learning has a set of rules to accomplish a particular goal. DSs use algorithms with positive rewards, meaning when a model performs an action which leads to the goal, it gets a reward and when it performs badly, it leads to punishment. Such learning is often used in areas like.

- Resource management
- Videogames
- Robotics

As we can see, ML is used for a variety of reasons. One of the most famous examples is the recommendation engine employed by Netflix. Based on search results and movies an individual has seen, Netflix is able to suggest some other movies or shows for that individual. Other ways we can use ML is:

- *Customer Relationship Management systems (CRM):* Based on importance of certain notifications, it tells the sales team to answer to important notifications first. A more complex system can also suggest the type of response.

- *Business Intelligence (BI)* and its analytic sellers use ML to analyze important data, see the trends and various deviations.
- *Chatbots*: They usually employ supervised as well as unsupervised learning techniques to give out curated results to the customers coming on the websites.

Advantages and Disadvantages of ML methods are summarized in **Table 1**.

As ML grows in demand, new techniques and applications will surface. Today's models need intense work before it gets optimized for one task. Some researchers are performing various operations to make ML models that are flexible and inexpensive with requiring low infrastructure. It will be not quick but once achieved, can pave way for more accurate and better results. A ML model generally goes through a common lifecycle as explained below.

- Data Collection
- Choosing a ML technique
- Finetuning
- Final model

9.1 Deep learning

It is a ML technique that makes computer learn things that a human can do naturally. It is the driving force behind many tasks which could be termed as complex for machines. Tasks such as self-driving cars, recognizing stop signal, drive in a straight

Advantages	Disadvantages
Help analyze customer behaviors.	Expensive
Customized product	High salaries for DS
Primary source for products	High infrastructure

Table 1.

Advantages and disadvantages of ML.

line and avoid collision. It also can be used for various control devices. DL can achieve better accuracy than classic ML algorithms. This helps in meeting various customer demands. Recent advancement has been so much that it outperforms humans in tasks like classification of objects.

Despite being discovered earlier, it has only achieved success in recent times. The reason being

- It requires large amount of labeled data.
- It requires a very high computing power.

There are various types of Deep Learning networks available in the market. Some of them are listed below.

- *Feed Forward Neural Network*: It is a basic kind of network in which data flows from one layer to another. It has only one kind of layer or just a hidden layer. There are no

backpropagation techniques available. The weight sum is fed as an input to the next layer.

- *Radial-Basis Function (RBF) Neural Networks*: They have more than one type of layer. In such networks, the distance between any point to the center is calculated and passed as an input to the next layer.
- *Multi-Layer Perceptron*: It has multiple layers and used to classify non-linear data. They have fully connected layers.
- *Convolutional Neural Network (CNN)*: It has n number of layers. It can have more than one convolutional layer and is very deep with few parameters.
- *Recurrent Neural Network (RNN)*: An output from a particular neuron is fed as an input to the same node. It helps in getting better output. It has memory storage and utilizes past results to optimize future outcomes.
- *Modular Neural Network*: Such networks are a collection of smaller neural networks. Combination of smaller networks leads to a big neural network and all networks work independently to achieve results.
- *Sequence to Sequence models*: There are generally a combination of RNN networks. It works on encoding and decoding.

9.2 Deep learning use cases

- Speech Recognition
- Image Recognition
- Natural Language Processing
- Recommender Systems
- Customer Relationship Management systems

Advantages	Disadvantages
Features are finetuned automatically	Require large amount of data
Same network used for various tasks	Expensive
Flexible and can be adapted for various problems.	No tool to formulate correct neural network models

Table 2.
Advantages and disadvantages of DL.

Advantages and Disadvantages of Deep Learning methods are summarized in **Table 2**. The Deep Learning lifecycle is like the one used for Machine Learning.

- Collection of Data
- Creation of model

- Training of model
- Deploying

Author details

Alexander I. Iliev^{1,2,3}

1 Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria

2 SRH University Berlin, Charlottenburg, Germany

3 UC Berkeley, Berkeley, California, USA

*Address all correspondence to: ailiev@berkeley.edu

IntechOpen

© 2023 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Trampe D, Quoidbach J, Taquet M. Emotions in everyday life. *PloS One*. 2015;10(12):e0145450 [2] Owens A. A Case study of crosscultural communication issues for Filipino call centre staff and their Australian customers. In: 2008 IEEE International Professional Communication Conference. Montreal: IEEE; 2008. pp. 1-10 [3] Jeanne Segal PM. 2021. articles. Retrieved from: <https://www.helpguide.org/articles/mental-health/emotionalintelligence-eq.htm#> [4] Australia, U. The Science of Emotion: Exploring The Basics Of Emotional Psychology. 2019. Retrieved from: <https://online.uwa.edu/news/emotional-psychology/> [5] Backstrom T. Speech Production and Acoustic Properties. Aalto University; 2021. Available from: <https://speechprocessingbook.aalto.fi/> [6] Aalto. Speech Processing. [Online]. 2020. Available on Jan.10.2023 at: <https://wiki.aalto.fi/display/ITSP/Introduction+to+Speech+Processing> [7] Warden P. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. DOI: 10.48550/arXiv.1804.03209 [8] Blanding M. The role of emotions in effective negotiations. 2014. Retrieved from: <https://hbswk.hbs.edu/item/the-role-of-emotions-in-effectivenegotiation> [9] Pavelescu LM, Petrić B. Studies in second language learning and teaching. 2018. Retrieved from: <https://pressto.amu.edu.pl/index.php/sslft> [10] Raisanen O. Linguistic Structure of Speech. Aalto University; 2021. Available from: <https://speechprocessingbook.aalto.fi/> [11] Backstrom T. Waveform. Aalto University; 2022. Available from: <https://speechprocessingbook.aalto.fi/> [12] Backstrom T. Windowing. Spectrogram and the STFT, Cestrum and MFCC: Aalto University; 2019. Available from: <https://speechprocessingbook.aalto.fi/> [13] Qing Z, Zhong W. Research on speech emotion recognition technology based on machine learning. In: 7th International Conference on Information Science and Control Engineering (ICISCE). 2020. pp. 1220-1223 [14] Kannadaguli P, Bhat V. A comparison of Bayesian and HMM based approaches in machine learning for emotion detection in native Kannada speaker. In: IEEMA Engineer Infinite Conference (TechNet). 2018. pp. 1-6 [15] Nasrun M, Setianingsih C. Human emotion detection with speech recognition using Mel-frequency cepstral coefficient and support vector machine. In: International Conference on Artificial Intelligence and Mechatronics Systems (AIMS). 2021. pp. 1-6 [16] Mohammad OA, Elhadeif M. Arabic speech emotion recognition method based on LPC and PPSD. In: 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM). 2021. pp. 31-36 [17] Bharti D, Kekana P. A hybrid machine learning model for emotion recognition from speech signals. In: International Conference on Smart Electronics and Communication (ICOSEC). 2020. pp. 491-496 [18] Gopal GN, Jayakrishnan R. Multiclass emotion detection and annotation in Malayalam Novels. In: International Conference on Computer Communication and Informatics (ICCCI). 2018. pp. 1-5

- [19] Goyal N, Rathore SS. Predictive visual analysis of speech data using machine learning algorithms. In: 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE). 2020. pp. 69-73 [20] Ali O et al. Automated Detection of Racial Microaggressions Using Machine Learning. IEEE; 2020
- [21] Tasha IJ, Wang Z-Q, Godin K. Speech emotion recognition based on Gaussian mixture models and deep neural networks. In: 2017 Information Theory and Applications Workshop (ITA). 2017
- [22] Tariq Z, Shah SK, Lee Y. Speech emotion detection using IoT based deep learning for health care. In: 2019 IEEE International Conference on Big Data (Big Data). 2019. pp. 4191-4196
- [23] Zhang S, Zhang S, Huang T, Gao W. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. IEEE Transactions on Multimedia. 2018:1576-1590 [24] Singh V, Sharma K. Empirical analysis of shallow and deep architecture classifiers on emotion recognition from speech. In: 2019 6th IEEE International Conference on Cyber Security and Cloud Computing (Cloud)/ 2019 5th IEEE International Conference on Edge Computing and Scalable Cloud (Edgecam). 2019. pp. 69-73
- [25] An H, Shi D. Mental health detection from speech signal: A convolution neural networks approach. In: International Joint Conference on Information, Media, and Engineering (IJCIME). 2019. pp. 436-439
- [26] Mokonyane TB, Sefira TJ. Emotional speaker recognition based on machine and deep learning. In: 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC). 2020. pp. 1-8 [27] Qidwai U, Al-Meer M. Emotional stability detection using convolutional neural networks. In: IEEE International Conference on Informatics, IoT, and Enabling Technologies (Ikot). 2020. pp. 136-140
- [28] Gunathilake S, Raj U. Emotion detection using Bi-directional LSTM with an effective text pre-processing method. In: 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). 2021. pp. 1-4
- [29] Lin Y-Y, Yang J-Y. Use empirical mode decomposition and ensemble deep learning to improve the performance of emotional voice recognition. In: IEEE 2nd International Workshop on System Biology and Biomedical Systems (SBBS). 2020. pp. 1-4
- [30] Wani TM, Guna Wan TS. Speech emotion recognition using convolution neural networks and deep stride convolutional neural networks. In: 6th International Conference on Wireless and Telematics (ICWT). 2020. pp. 1-6