ISLAMIC UNIVERSITY OF LEBANON
FACULTY OF ENGINEERIING

DEPARTMENT OF GRADUATE STUDIES

MASTER RECHERCHE

Designing building a speech emotion recognition system using Machine Learning techniques

Salah Khoudair Alwan

2023 - 2024

Acknowledgments

This master thesis was completed with the motivation and encouragement of many people. I wish to express my sincere gratitude to them.

I would first like to genuinely thank my advisor Pr. XXXXX for her continuous support, guidance and valuable advice.

I am grateful to all my professors of XXXXX and all members of my laboratory XXXXX .

Finally, I am indebted to my parents for developing my skills in order to understand the true value of education. I am also thankful for my family and friends, for their love and support.

Abstarct

....

....

Keywords : Software Engineering, Artificial Intelligence(AI), Machine Learning(ML), Deep Learning(DL), Neural Network(NN), Classification, speech emotion.

Contents

1 Introduction 1

1.1 Opening Section 1

1.2 Background 2

1.3 Statement of the problem 2

1.3.1 State the problem 2

1.3.2 Justify the problem 3

1.4 Rationale 3

1.5 Research aims 3

1.6 Research objectives 3

1.7 Research questions 4

1.8 Scope 4

1.9 Significance 5

1.10 Structure of the document 5

2 Machine Learning For Speech Emotion Recognition System 6

2.1 Introduction 6

2.2 Machine Learning 6

2.2.1 Supervised Learning 7

2.2.2 Unsupervised Learning 8

2.2.3 Semi-supervised Learning 9

2.2.4 Reinforcement Learning 10

2.3 Machine learning methods 10

2.3.1 Support Vector Machine (SVM) 10
2.3.2 Random Forest (RF) 12
2.3.3 K-Nearest Neighbor (KNN) 13
2.4 Deep Learning methods 14
2.4.1 Convolutional Neural Network (CNN) 14
2.4.2 Long Short-term Memory (LSTM) 15
2.5 Hybrid methods: CNN-LSTM 16
2.6 Traditional ML techniques for previous studies 17
2.7 Conclusion 19

3 Materials and Methods/Methodology 20
3.1 Intoduction 20
3.2 Research design 20
3.3 Research philosophy 20
3.4 Research type 20
3.5 Research strategy 21
3.6 Time horizon 21
3.7 Sampling strategy 21
3.8 Data collection methods 21
3.9 Data analysis methods 21
3.10 Conclusion 21

Bibliography 25

List of Figures

2.1 Machine learning categories 7
2.2 Supervised Learning method 8
2.3 k-means clustering of raw speech signal 9
2.4 Support Vector Machine (SVM) 11
2.5 Random Forest 13
2.6 K-Nearest Neighbor (KNN) 14
2.7 Convolutional neural network 15
2.8 Long Short-term Memory (LSTM) 16
2.9 Combination between convolutional neural network and Long Short-term Memory 17

List of Tables

Chapter 1

Introduction

1.1 Opening Section

Communication between humans through spoken language has been an essential component of civilization as a basis for the exchange of data. However, the emergence of Emotional Voice Recognition (ERS) is the fastest way to connect and communicate human-machine interaction (HMI). ERS plays an important role in real-time applications of HMI [Kwon 2019]. Especially, in the field of human-machine interaction, growing interest in recent years due to the use of low-cost Internet and social media occur semantic gaps. Many researchers work in this area to make a machine understand the state of an individual's speech to identify or analyze the emotional condition of the speaker. The new generation of Artificial Intelligence (AI) systems has achieved considerable success through the use various machine learning techniques to accurately detect the speaker's emotions while speaking Therefore, popular methods of machine learning are proposed to solve the correct human emotion such as deep learning [Jain et al. 2020].

Deep learning methods have become more popular, as a major problem with deep learning, data availability, although still a challenge, has become less impactful. Additionally, deep learning methods have proven efficient, while offering high accuracy and a low error rate in detection speech. These approaches manage to reduce the loss of information when detect human emotion, overcoming limitation of other approaches.

In this study, we propose an advanced algorithm for improving the accuracy of speech emotion recognition. To accomplish this, we first apply noise reduction techniques to the speech samples using filters. Next, we use various classification techniques to accurately identify the emotional state of the speaker. By combining these two steps, our algorithm is able to generate desirable results for speech emotion recognition.

1.2 Background

Speech refers to the vocal communication and one of the most important mediums of communication for humans. It is a complex signal that contains information about the speaker, language, and emotions, etc. However, emotions are a crucial area of study in cognitive sciences like psychology and neuroscience because they play a key role in human cognition. Speech Emotion Recognition (SER) is a system for identifying the emotions of various audio samples. Speech recognition of emotions has moved from a niche into an important component of HMI [Khalil et al. 2019], it can can be described as a technology in which the type of speech emotion can be deduced from the characteristics of the emotional signal by computer processing.

1.3 Statement of the problem

In this section, to explain the problematic of our thesis for speech emotion recognition. The goal of this section is to discussing the main issues and provide a comprehensive understanding of why this problem is important to address.

1.3.1 State the problem

Quantifiable emotional recognition using sensors from voice signals is an emerging field of research in HMI that applies to multiple applications such as behavior assessment, virtual reality, and human-reboot interaction, etc. The problematic is due to the low accuracy of the speech emotion recognition systems using in the literature.

1.3.2 Justify the problem

Nowadays, people are communicating more with each other through social media. However, trust and engagement in the online world are weaker and lower than in the face-to-face world. This can include difficulties in determining which emotions to detect, how to define emotions in a way that is applicable across cultures and languages, and addressing the potential negative consequences of using emotion detection in certain contexts, such as privacy concerns and bias in decision-making.

1.4 Rationale

The advent of correct emotional speech recognition models could significantly improve the user experience in systems involving human-machine interactions. For example, researchers used artificial intelligence (AI) to detect recognition of speech emotion.

1.5 Research aims

This research aims to improve the speech emotion recognition rate using different machine learning classification techniques. One important pre-processing step is to remove the noise from speech samples using filters. Other steps could be applied such as features extraction and features selection to build the classification techniques in order to detect the correct emotion. Emotions could be Anger, Happiness, Sad and Neutral, etc.

1.6 Research objectives

This research goal is use speech emotion recognition to facilitate natural interaction with machines through direct voice interaction rather than using traditional devices as inputs to understand spoken content and facilitate listener reaction.

1.7 Research questions

What are the most effective pre-processing techniques, specifically noise removal using filters, for improving the accuracy of speech emotion recognition?

How do different machine learning classification techniques, such as decision trees, support vector machines, and deep neural networks, perform in speech emotion recognition?

How can the proposed algorithm be used to improve real-time applications of human-machine interaction?

How can the proposed algorithm be tested and validated in recognizing multiple emotions such as Anger, Happiness, Sad, and Neutral, etc.?

1.8 Scope

Speech emotion detection in machine learning is a field of research that aims to develop algorithms and systems to automatically recognize emotions in speech. This technology has a wide range of potential applications, including:

- Human-computer interaction: Emotion detection in speech can be used to improve the user experience in virtual assistants, chatbots, and other interactive systems by making them more responsive to the user's emotional state.
- Mental health: Emotion detection in speech can be used to monitor and support individuals with mental health conditions, such as depression and anxiety, by detecting changes in their emotional state.
- Marketing and advertising: Companies can use emotion detection in speech to gain insights into how consumers respond to their products, services, and advertising campaigns.
- Public safety: Emotion detection in speech can be used to monitor the emotional state of pilots, air traffic controllers, and other critical operators in safety-critical systems.

Overall, Speech emotion detection in machine learning is a rapidly growing field with the potential to have a significant impact on a wide range of industries and applications.

1.9 Significance

Currently, numerous machine learning methods and literature studies have demonstrated the potential to detect the emotions of the human voice. But, there is still a research gap between the identification of appropriate voice signals and ML patterns. Moreover, different studies have used different data sets and machine learning methods, but they still face the challenge of achieving the best results.

1.10 Structure of the document

This report is structured as follows:

- Chapter 1: introduces the basic concepts of Speech emotion recognition.
- Chapter 2: explains the basic methods in machine learning and presents an overview of studies related to Speech emotion recognition for machine learning.
- Chapter 3: explains the exact steps of the proposed model.
- Chapter 4: demonstrates the results of the experimental study, and it compares with other methods.
- Chapter 5: summarizes our dissertation in the conclusion statement and proposes future work to be done in Speech emotion recognition.

Chapter 2

Machine Learning For Speech Emotion Recognition System

2.1 Introduction

Machine learning is an intriguing answer to speech emotion recognition that ensures speed and efficiency. In particular, a machine learns to carry out a new task in order to maintain a specific performance measure based on previous experience. [Mahesh 2020].

In this chapter, we will go through different Machine Learning (ML) specially Deep Learning (DL) approaches, some helpful architectures, several performance measurements and features that may be applied to this issue, and a review of works on speech emotion recognition for machine learning.

2.2 Machine Learning

Machine learning (ML) is a part of the wider area of Artificial Intelligence (AI). ML deals with the creation of algorithms and statistical models that allow a system to learn from data and make prediction without being explicitly programmed. It entails employing mathematical and computational approaches to comprehend patterns and correlations in data and then making predictions or judgments based on this comprehension [Samuel 1988].

In this section we present the basic concepts of machine learning techniques. ML algorithms can be divided into four categories: supervised, unsupervised, semi-supervised, and reinforcement learning, as shown in figure 2.1. We concentrate mainly on the supervised ML technique which is relevant to the context of this thesis.

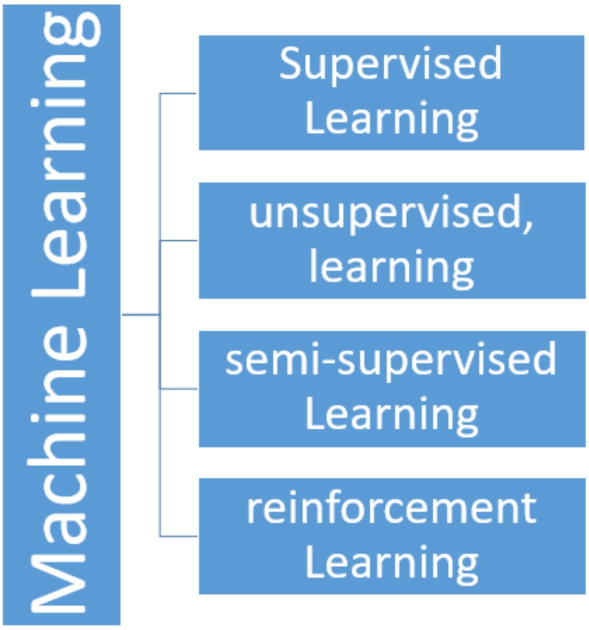

Figure 2.1: Machine learning categories.

2.2.1 Supervised Learning

Supervised learning [Singh et al. 2016] is the first category of ML. This method of learning trains a computer to recognize specific patterns using labeled data. For instance, tagged photographs might be used to train a computer vision system to distinguish things if one wished to develop one. This kind of ML is useful for a variety of tasks, including reading handwritten writing, forecasting stock prices, and spotting possible fraud. Shown example in figure 2.2 The concept is to use a model that learns and then predicts the output depending on the input variables. Typically, supervised learning consists of two types of tasks:

- The method that predicts the class of an input is known as classification. In a classification problem, output is a category variable.
- Regression is used to anticipate the value of a continuous variable, such as a home price.

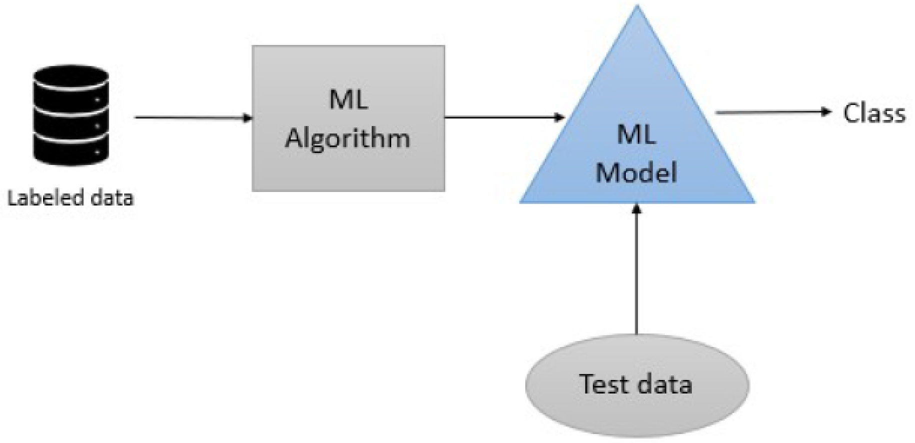

Figure 2.2: Supervised Learning method

In section 2.3, we present the classification task and its common algorithms that we will use it later.

2.2.2 Unsupervised Learning

Unsupervised learning [Usama et al. 2019] is the second kind of ML. Unlabeled data is used in this form of learning to find patterns in the data. Unsupervised ML, for instance, might be used to divide a set of clients into several categories. This kind of learning is beneficial in a variety of situations, including data analysis, customer segmentation, and product recommendation.

Unsupervised learning is classified into two types: clustering and dimensionality reduction. Clustering methods put related data points together, whereas dimensionality reduction algorithms minimize the amount of features in a dataset while maintaining critical information.

Clustering is a method for grouping similar data points together. In this case, the data points would be speech samples and the goal is to group the speech samples that express similar emotions. For example Use the k-means algorithm to group speech samples into k clusters, where k is the number of emotions you want to detect (e.g. happy, sad, neutral, etc). shown in figure 2.3.

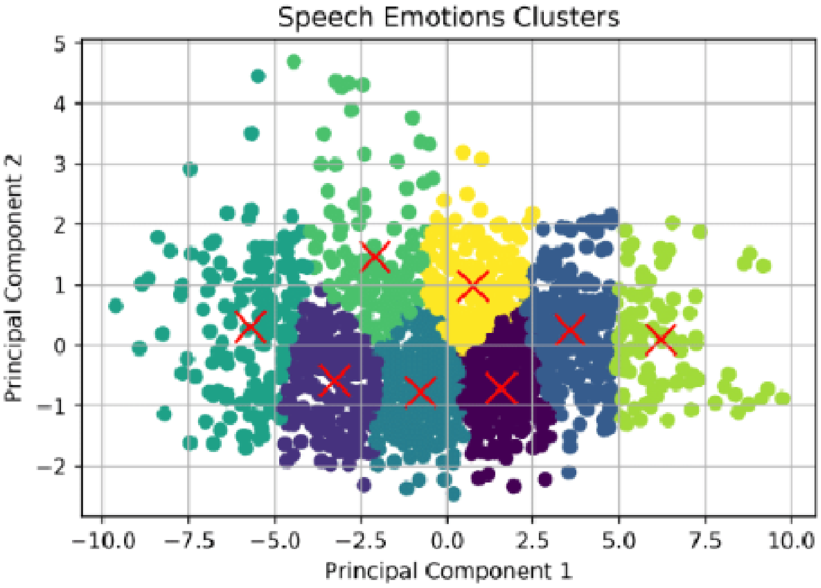

Figure 2.3: k-means clustering of raw speech signal. [Sefara 2019]

2.2.3 Semi-supervised Learning

Semi-supervised learning is a ML approach which combines supervised and unsupervised techniques. It is used when there is a small amount of labelled data and a large amount of unlabelled data available. The goal is to utilize the labeled data to learn about the data's underlying structure, and then use that information to label the remaining unlabeled data [Zhu 2005] [van Engelen and Hoos 2020].

Semi-supervised learning methods can produce equivalent or even better outcomes than supervised learning, and they can even employ less or unlabeled data to improve model performance. Self-training, co-training, and multi-view learning are some prominent semi-supervised learning techniques.

2.2.4 Reinforcement Learning

Reinforcement learning (RL) is a form of ML that focuses on teaching decision-making agents [Chen 2016]. It is used to teach an agent how to interact with its surroundings in order to maximize some concept of cumulative reward.
At each stage, the agent receives input in the form of rewards or penalties and learns to pick activities that will result in the largest cumulative reward over time. The decision-making process of the agent is based on a policy that maps an observation of the environment to an action. The objective is to determine the optimal strategy that maximizes overall reward.
Reinforcement learning has been utilized in many areas, including game play, robotic control, and decision making under uncertainty. Q-learning, SARSA, and the actor-critic algorithm are some popular RL algorithms.

2.3 Machine learning methods

Machine learning can be used to construct a model based on sample data. ML has been applied in many area fields and provide good performance in solving regression, classification, and clustering problems. In this section we represent machine learning methods such as: Support Vector Machine, Random Forest, K-Nearest Neighbor, Deep Learning, and hybrid methods.

2.3.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a kind of supervised learning algorithm which can be used for classification or regression tasks [Cutler et al. 2012]. The main idea behind SVM clasifier is to find the best hyperplane that separates the different classes of data in a feature space. The hyperplane that is chosen is the one that maximizes the margin, which is the distance between the hyperplane and the closest data points from each class [Yu and Kim 2012].

The decision hyperplane with the greatest separation between the closest points in these two classes is the best one, as shown in Figure 2.4.

An SVM uses data points to identify the optimal hyperplane, represented as a line

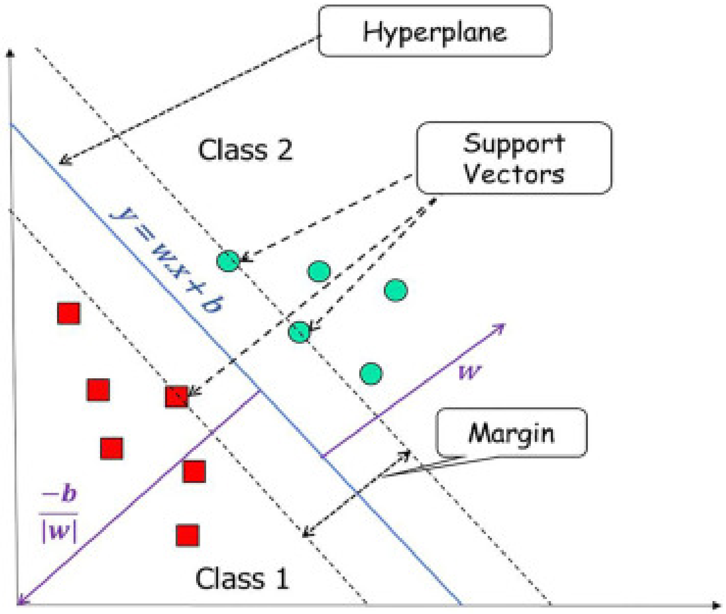

Figure 2.4: Support Vector Machine (SVM) [Rani et al. 2022]

in two-dimensional space, that separates the different classes . This line serves as the decision hyperplan, where data points on one side are classified as one class (red) and those on the other side are classified as another (green). The optimal hyperplane is the one that maximizes the distance, or margins, between the two classes. It is the hyperplane that is farthest away from the closest data points of each class.

Support Vector Machine (SVM) regression [Yu and Kim 2012] is a variation of the standard SVM algorithm that is used for solving regression problems, rather than classification problems. The main idea behind SVM regression is to find the best fit line (i.e., the line that best approximates the underlying pattern in the data) that maximizes the margin between the data points and the line. In SVM regression, instead of trying to find a hyperplane that separates the data points into different classes, the goal is to find a line that best fits the data while minimizing the error. The SVM algorithm finds the line that maximizes the margin between the data points and the line, while also minimizing the error.

In the context of emotion speech recognition, an SVM classifier is a popular technique used in speech emotion recognition. SVM are known for their ability to handle high-dimensional and non-linearly separable data, which makes them well-suited for speech emotion recognition tasks. It can be trained on a dataset of speech recordings labeled with the corresponding emotions (e.g. happy, sad, angry). To implement SVM classification in speech emotion recognition, researchers typically begin by collecting and preprocessing a dataset of speech samples, which may include labeling the samples with their corresponding emotional category. Next, the researchers will choose a kernel function to transform the data into a higher-dimensional space where it is easier to find a linear boundary. Once the data is preprocessed and the kernel function is chosen, the SVM algorithm is trained using the dataset. The algorithm finds the optimal hyperplane that maximally separates the different emotional categories. The trained model can then be evaluated using metrics such as accuracy, precision, and recall. If the performance is not satisfactory, the model's hyperparameters can be adjusted and the training and evaluation steps can be repeated until an acceptable level of performance is achieved. Once the model is fully trained, it can be used to predict the emotional category of new speech samples. SVM has been shown to achieve high performance and robustness against variations in speaker, environment, and other factors when applied to speech emotion recognition tasks.

2.3.2 Random Forest (RF)

Random forest uses a set of decision trees for classification tasks. The algorithm creates multiple decision trees during the training phase, and for each new input. RF classifies the input based on the majority vote of the decision tree [Cutler et al. 2012]. Random Forest is a popular method used for speech emotion recognition, which involves the automatic recognition of the emotional state of a speaker based on their speech.

In this context, Random Forest is used to classify speech segments into different emotional categories such as happy, sad, angry, etc. The algorithm works by first extracting features from the speech signal. These features are then used to train a set of decision trees using a random subset of the data. During the testing phase, a new speech segment is passed through the trained random forest, and each decision tree makes a prediction. The final prediction of emotion is made by majority voting predictions from all decision trees. The use of Random Forest in speech emotion recognition has been shown to achieve high performance and robustness against variations in speaker, environment, and other factors. shown figure 2.5.

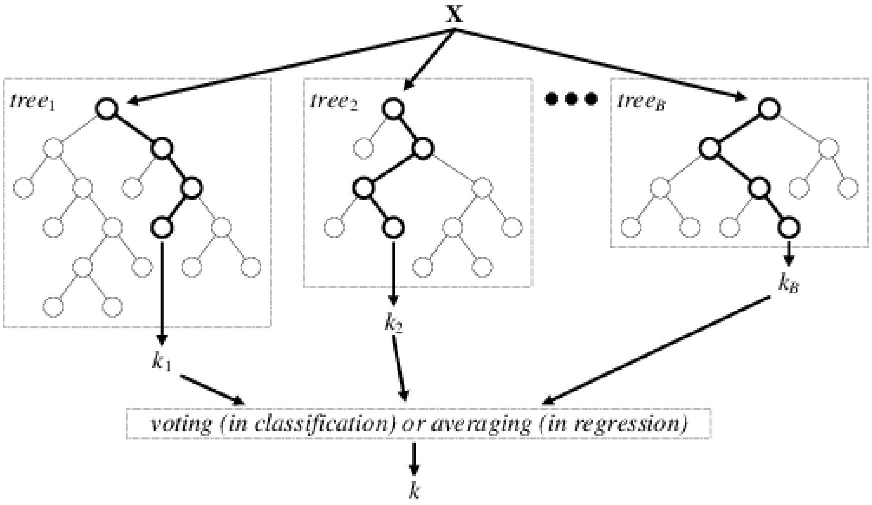

Figure 2.5: Random Forest.[Verikas et al. 2016]

2.3.3 K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) is a simple supervised machine learning algorithm that can be used to perform speech emotion recognition tasks. The basic idea behind the KNN algorithm is that it classifies a new speech segment according to the majority class of its k nearest neighbors among the pre-labeled speech segments [Peterson 2009].

In the training phase, the algorithm stores the feature vectors and corresponding emotional class labels of the speech segments. During the testing phase, for a new speech segment, the algorithm calculates the distance between the new speech segment and all the stored speech segments and finds the 6 nearest speech segments based on these features. Then it assigns the emotional class label to the new speech segment based on the majority class among 6 nearest speech segments [Venkata Subbarao et al. 2022]. shown figure 2.6.

K-NN classification can be a simple yet powerful algorithm for speech emotion recognition tasks, it is easy to implement and computationally efficient.

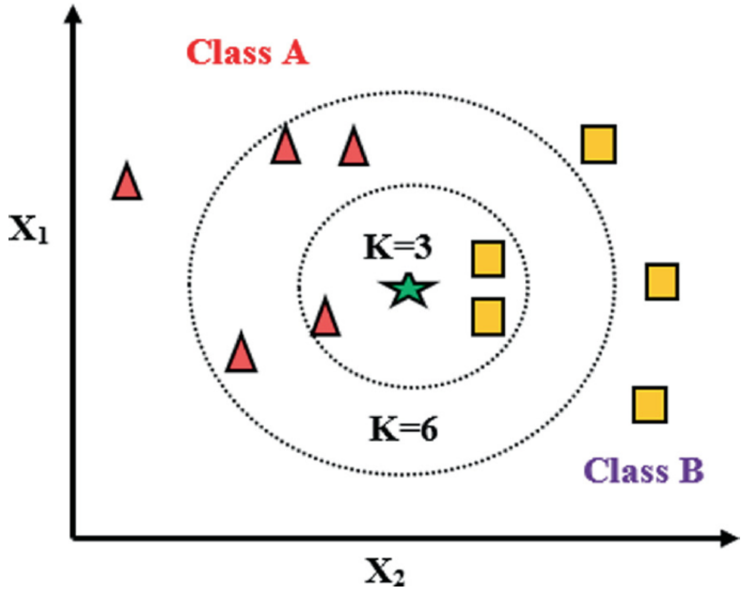

Figure 2.6: K-Nearest Neighbor (KNN). [Venkata Subbarao et al. 2022]

2.4 Deep Learning methods

Deep learning (AD) is a subset of machine learning which is based on artificial neural network (ANN) that is characterized by its multiple hidden layers between the input and output layers. In the following, we mainly focus on Convolutional Neural Network (CNN), and Long Short-term Memory (LSTM).

2.4.1 Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is a kind of deep learning algorithm that are particularly well-suited for image and video classification tasks. A CNN is a network of neurones that is composed of several layers.: convolutional layer, pooling layer and fully connected layer.

The Convolutional layers are designed to learn and extract features of the input. Then, the pooling layer serves to reduce spatial resolution and increase strength of features, and the fully connected layer are used to classify the input [Phung and Rhee 2019]. Figure 2.7 shows an example of CNN architecture.

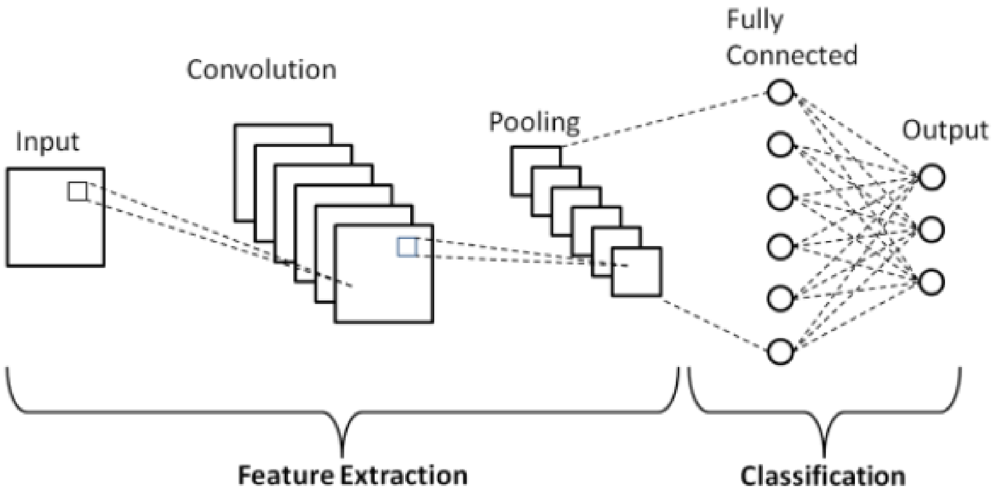

Figure 2.7: Convolutional neural network.[Phung and Rhee 2019]

CNN can be used for speech emotion recognition tasks. In this context, the CNN training process is done by showing the network a large dataset of labeled speech segments and updating the network's parameters in order to minimize the classification error. In the testing phase, a new speech segment is passed through the trained CNN along with the extracted features, and the final output is a probability distribution over the different emotional classes. The most probable class is considered as the final prediction.

CNNs have been shown to be effective in speech emotion recognition tasks when combined with other feature extraction methods and provide a robust and accurate model for the recognition of emotional states from speech.

2.4.2 Long Short-term Memory (LSTM)

The Long Short-Term Memory (LSTM) classification architecture is a kind of Recurrent Neural Network (RNN) which is designed to handle sequential data and the long-term dependencies problem. The basic structure of LSTM network for classification tasks includes multiple layers: input layer, LSTM layers, and output layer [Graves 2012]. Figure 2.8 shows an example of LSTM architecture.

Long Short-term Memory (LSTM) can be used for speech emotion recognition tasks. RES is the automatic recognition of a speaker's emotional state based on their speech. The input layer takes in the feature vectors extracted from the speech segments. These features are then passed to the LSTM layers where the network learns to classify the speech segments based on the emotional content. The output of the LSTM layers is then passed to the output layer, which produces the final output, which is a probability distribution over the different emotional classes. In this context, LSTM can be used to classify speech segments into different emotional categories such as happy, sad, angry, etc.

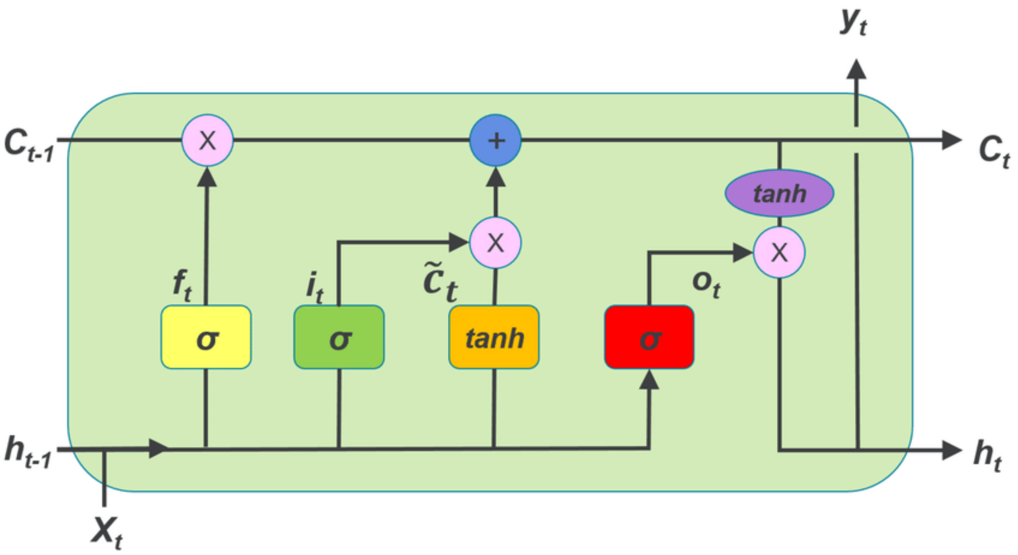

Figure 2.8: Long Short-term Memory (LSTM) [Ismail et al. 2018].

2.5 Hybrid methods: CNN-LSTM

Hybrid methods are a combination of multiple ML techniques that are used together to improve the performance of a specific task. Hybrid methods can be categorized into several types such CNN and LSTM . Hybrid methods have been used in various applications such as speech emotion recognition.

The architecture of a combination between a CNN and LSTM network typically consists of several layers, including an input layer, multiple CNN layers, multiple LSTM layers, and an output layer, shown figure 2.9.

The input layer takes in the input sequence, in the case of speech emotion recognition it's the feature vectors extracted from the speech segments. These features are then passed to the CNN layers where the network learns to extract features from the input. The output of the CNN layers is then passed to the LSTM layers where the network learns to classify the speech segments based on the emotional content. The LSTM layers are responsible for handling the sequential data and the problem of long-term dependencies. The output of the LSTM layers is then passed to the output layer which produces the final output, which is a probability distribution over the different emotional classes [Zhao et al. 2019].

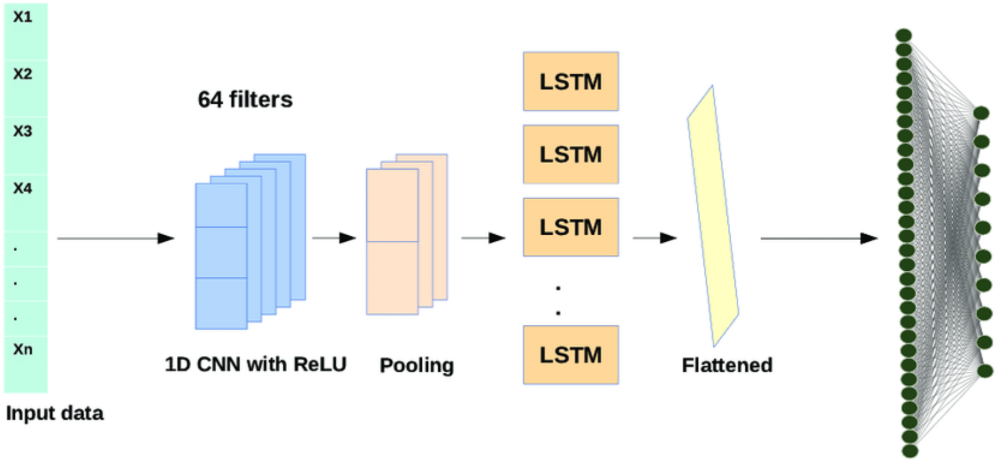

Figure 2.9: Combination between convolutional neural network and Long Short-term Memory.[Hamad et al. 2020]

2.6 Traditional ML techniques for previous studies

In recent years, several researchers have been a notable success of machine learning methods to predict speech emotion recognition for example K-Nearest Neighbor, Random Forest, Support Vector Machine, and Deep Learning, etc.

In this section, we discuss the related works for machine learning in speech emotion recognition.

[Kanth and Saraswathi 2015] proposed a model based on SVM. A binary Support Vector machine used Correlation Based Feature Selection (CFS) with the Sequential Forward Selection (SFS). Binary SVM is constructed along with a Multiclass SVM for efficient Speech Emotion Recognition. Then the combination between Binary SVM and multiclass SVM achieves an overall accuracy (acc) equal to 87.86%.

[Reddy and Vijayarajan 2020] proposed a medel for emotion recognition system to examine based on multi-algorithms fusion. In this model,there are used Berlin emotional database and Telugu database to extract Mel Frequency Cepstral Coefficients (MFCC) and Discrete Wavelet Transform (DWT) features features. Then, used SVM and KNN algorithms for classifies different states of emotion. the model achieves overall acc 94%.

[Chen et al. 2012] presented SER technologies and categorization models. They used MFCC and classifiers like SVM, ANN, and PCA to extract the attributes. In this situation, PCA was used to reduce the dimension. A hybrid mix of characteristics and classifiers was used to perform four scale comparability tests. Fisher obtains 85% acc when paired with SVM. As a consequence, PCA eliminates linked qualities, decreases fitting issues, increases algorithm dependability, and enhances visualisation.
[Iliou and Anagnostopoulos 2010] proposed a model called SVM-MLP-PNN for SER and used Berlin database. Then, the MFCC pitch feature extraction is used in this case. The study's main finding is that PNN has a 94% accuracy rate in recognizing emotions based on whole speech distinction.
[Cai et al. 2015] used MLP method for accelerating learning in automated speech detection technologies. They suggest using different feature extraction models, such as GMM, and FMLLR, in conjunction with the MLP classifier. In this model, MLP used a singular significance methods to decomposition, which reduces the dimensionality of the MLP to improve performance. This can help improve the accuracy and efficiency of the speech detection system.

[Chavhan et al. 2010] proposed Support Vector Machine (SVM) classifiers to analyze Speech Emotion Recognition (SER), and various extraction features such as fundamental frequency, ZCR zero crossing rate, linear prediction coefficients, and mel-frequency cepstral coefficients. The study revealed that MFCC had the highest linear kernel consistency across all databases. Berlin database achieves

89.80%, Japan databases achieves 93.57%, and Thai databases achieve 98% accuracy(acc). The study found that SVM is useful in situations where the number of measurements is greater than the number of samples and is relatively stable in memory.

2.7 Conclusion

In this chapter, we pointed out different methods of machine learning for speech emotion recognition in the first part and we gave some different proposed methods in the literature used to detect speech emotion recognition in the second part. The next chapter details our proposed approach based in hybrid methods.

Chapter 3

Materials and Methods/Methodology

3.1 Intoduction

..

3.2 Research design

..

3.3 Research philosophy

..

3.4 Research type

..

3.5 Research strategy

..

3.6 Time horizon

..

3.7 Sampling strategy

..

3.8 Data collection methods

..

3.9 Data analysis methods

..

3.10 Conclusion

..

Bibliography

[Cai et al. 2015] Cai, C., Xu, Y., Ke, D., and Su, K. (2015). A fast learning method for multilayer perceptrons in automatic speech recognition systems. *Journal of Robotics*, 2015.

[Chavhan et al. 2010] Chavhan, Y., Dhore, M., and Yesaware, P. (2010). Speech emotion recognition using support vector machine. *International Journal of Computer Applications*, 1(20):6–9.

[Chen 2016] Chen, J. X. (2016). The evolution of computing: Alphago. *Computing in Science & Engineering*, 18(4):4–7.

[Chen et al. 2012] Chen, L., Mao, X., Xue, Y., and Cheng, L. L. (2012). Speech emotion recognition: Features and classification models. *Digital signal processing*, 22(6):1154–1160.

[Cutler et al. 2012] Cutler, A., Cutler, D. R., and Stevens, J. R. (2012). Random forests. In *Ensemble machine learning*, pages 157–175. Springer.

[Graves 2012] Graves, A. (2012). Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45.

[Hamad et al. 2020] Hamad, R. A., Yang, L., Woo, W. L., and Wei, B. (2020). Joint learning of temporal models to handle imbalanced data for human activity recognition. *Applied Sciences*, 10(15):5293.

[Iliou and Anagnostopoulos 2010] Iliou, T. and Anagnostopoulos, C.-N. (2010).

Svm-mlp-pnn classifiers on speech emotion recognition field-a comparative study. In *2010 Fifth International Conference on Digital Telecommunications*, pages 1–6. IEEE.

[Ismail et al. 2018] Ismail, A. A., Wood, T., and Bravo, H. C. (2018). Improving long-horizon forecasts with expectation-biased lstm networks. *arXiv preprint arXiv:1804.06776*.

[Jain et al. 2020] Jain, M., Narayan, S., Balaji, P., Bhowmick, A., Muthu, R. K., et al. (2020). Speech emotion recognition using support vector machine. *arXiv preprint arXiv:2002.07590*.

[Kanth and Saraswathi 2015] Kanth, N. R. and Saraswathi, S. (2015). Efficient speech emotion recognition using binary support vector machines & multiclass svm. In *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pages 1–6. IEEE.

[Khalil et al. 2019] Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., and Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 7:117327–117345.

[Kwon 2019] Kwon, S. (2019). A cnn-assisted enhanced audio signal processing for speech emotion recognition. *Sensors*, 20(1):183.

[Mahesh 2020] Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9:381–386.

[Peterson 2009] Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, 4(2):1883.

[Phung and Rhee 2019] Phung, V. H. and Rhee, E. J. (2019). A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets. *Applied Sciences*, 9(21):4500.

[Rani et al. 2022] Rani, A., Kumar, N., Kumar, J., and Sinha, N. K. (2022). Machine learning for soil moisture assessment. In *Deep Learning for Sustainable Agriculture*, pages 143–168. Elsevier.

[Reddy and Vijayarajan 2020] Reddy, A. P. and Vijayarajan, V. (2020). Audio compression with multi-algorithm fusion and its impact in speech emotion recognition. *International Journal of Speech Technology*, 23(2):277–285.

[Samuel 1988] Samuel, A. L. (1988). Some studies in machine learning using the game of checkers. ii—recent progress. *Computer Games I*, pages 366–400.

[Sefara 2019] Sefara, T. J. (2019). The effects of normalisation methods on speech emotion recognition. In *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, pages 1–8. IEEE.

[Singh et al. 2016] Singh, A., Thakur, N., and Sharma, A. (2016). A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1310–1315. Ieee.

[Usama et al. 2019] Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K.-L. A., Elkhatib, Y., Hussain, A., and Al-Fuqaha, A. (2019). Unsupervised machine learning for networking: Techniques, applications and research challenges. *IEEE access*, 7:65579–65615.

[van Engelen and Hoos 2020] van Engelen, J. E. and Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, 109:373–440.

[Venkata Subbarao et al. 2022] Venkata Subbarao, M., Terlapu, S. K., Geethika, N., and Harika, K. D. (2022). Speech emotion recognition using k-nearest neighbor classifiers. In *Recent Advances in Artificial Intelligence and Data Engineering*, pages 123–131. Springer.

[Verikas et al. 2016] Verikas, A., Vaiciukynas, E., Gelzinis, A., Parker, J., and Olsson, M. C. (2016). Electromyographic patterns during golf swing: Activation sequence profiling and prediction of shot effectiveness. *Sensors*, 16(4):592.

[Yu and Kim 2012] Yu, H. and Kim, S. (2012). Svm tutorial-classification, regression and ranking. *Handbook of Natural computing*, 1:479–506.

[Zhao et al. 2019] Zhao, J., Mao, X., and Chen, L. (2019). Speech emotion recognition using deep 1d & 2d cnn lstm networks. *Biomedical signal processing and control*, 47:312–323.

[Zhu 2005] Zhu, X. J. (2005). Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison, Department of Computer Sciences.