University of Essex
Department of Computer Science and Electronic Engineering

CE902-7-SU: PROFESSIONAL PRACTICE AND RESEARCH METHODOLOGY

Comparative speech emotion recognition use in multiple models

Supasun Khumpraphan

Registration number:2110366

Supervisor: Dr Cunjin Luo

June 30, 2022

Colchester

Contents

List of Figures

CHAPTER 1

abstract

Communication is essential in our day-to-day life. Because it is used in conversations and increases relationships with people and causes understanding The key to living with others and being successful in anything such as work, friends and family is a thorough understanding of other people's feelings and thoughts. in order to adjust behavior according to the situation to suit the mood of the conversationalist correctly and appropriately so there is no conflict. Sometimes they talk without seeing their faces or tones. and understand only the sentence patterns such as sounds, letters, words, phrases, and sentences that may cause the listener not fully to understand which speaker's feelings are complicated to understand. However, We can fix all these problems and fix them effectively if we can predict the mood of the person speaking. using existing data or past information to predict mood and feelings, and understand the conversation partner's mood accurately It enhances a good relationship with the interlocutor and accurate prediction modeling. It will help make better decisions. Reduce the risk of people hating and increase the number of people who love us, thus creating machine learning or deep learning that can interpret the emotions of the person speaking. Can do a lot of benefits, such as a call center system, we can know the mood of the caller whether they are angry or irritated. It makes us understand customers more about how satisfied or dissatisfied customers are with us and build AI that can react more effectively to the user than bring a simple monotone voice, so speech emotion recognition to classifier the speaker's mood is very necessary. In most research surveys on speech emotion recognition and different aspects of speech recognition, it has only one model was used in the study

and does not compare to use in several models, although there are many models capable of speech emotion recognition such as Extreme Gradient Boosting Classifier, support vector machine, Random Forest, Convolutional neural network, and Decision Tree. this is the reason why this research is very important because it can give us a full understanding of speech emotion recognition in all models. To help develop the model more effectively, and accurately in speech emotion recognition, we will compare the 5 models by the accuracy of each model measured using a confusion matrix.

CHAPTER 2

Introduction

Mind-reading has been studied since ancient Greece, around 384-322 B.C., with prominent philosophers Plato and Aristotel trying to understand and explain nature [7]. human expression But it is still difficult to be able to read a human mind. until now Technology is increasingly playing an important role in human life. Technology has long been associated with human life. For example, telephones. These technologies have come to help facilitate and begin to change the way of living of human beings until today, humans need to rely on technology in almost every activity and age range From birth to old age. the importance of these Technology has therefore been developed to be closer to humans, to have a greater understanding of humans in order to bridge the gap in communication between them. If the computer can Able to perceive human emotions It will make it possible to choose to respond to humans more accurately and more appropriately. Computers can react with regard to emotions as well. resulting in humans feeling friendly and I felt a lot more fun too. However, it is still difficult for computers to understand human beings very intricately computers have to use artificial intelligence to help. recognition and perception of Computers are gaining attention and are constantly being developed, such as the recognition of Human motion recognition, Facial recognition, and Speech recognition, but the difficult part is Emotion recognition, which is diverse and expressive in many ways. in addition, there are also similar patterns in certain moods. Speech emotion recognition was also used in this study to identify different emotions by Sound characteristics, such as pitch, and frequency, some words use different pitches. The resulting meaning is different. Therefore, this is why we will study and

develop speech emotion recognition from various models in order to try to understand the tone of the voice as much as possible in order to better understand the human emotion.

Figure 2.1: Structure of the Speech Emotion Recognition System [1].

2.1 Feature extraction

In this research, speech signals will extract the characteristics (Feature extraction) of each sound unit that is different. the learning process will do a Memorization of all sound attributes in each group to compare when the segmentation begins. If the audio signal has the same characteristics or is close to any group, it is able to determine to which group the signal belongs. Feature extraction also helps to reduce the amount of data to be analyzed and processed without the need for data. All audio signals are compared. What is analyzed will be only the features that are extracted only and retain the important features of the data [20].

2.2 Feature selection

The present study used F-Score, a method of selecting correlated features for segmentation, and ranking the feature according to descending correlation to select redundant features. Issuing the F-Score then adopts statistical and creates classification and modeling methods. by specifying High scores for attributes where data points are very distant from other groups and The data within that group must also be similar [10].

2.3 Classification

in the emotional classification process The dimensions of all attributes and those obtained from the selection go into the learning process. to classify emotions We used five models: Support Vector Machine, Decision Tree, Random Forest, Extreme Gradient Boosting, and Convolutional Neural Network, and the data was segmented into 10 folds cross-validation .

CHAPTER 3

Related work

In the research involved in analyzing sound emotions, the Artificial Neural Network and Support Vector Machine have long been the fundamental methods that have been used frequently. In Supervised Learning, Rajisha, Sunija and Riyas (2016) used the Support Vector Machine and Artificial Neural Network model analysis to classify moods: Angry, Happy, Sad, and Neutral [22]. this research selected the sound feature that was extracted Mel Frequency Cepstral Coefficients, Short Time Energy and Pitch. The Artificial Neural Network model was capable of discriminating more accurately than SVM at 88.4 percent and 78.2 percent, respectively, on the Malayalam language dataset developed by the research team. Have the average person speak 20 sentences each, with the tone of voice conveying the four moods assigned to them. Accuracy values can be viewed separately for each emotion as shown in (Fig. 3.1).

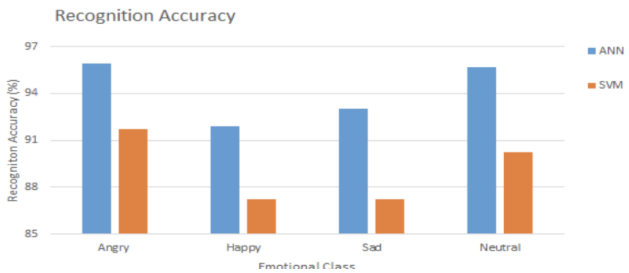

Figure 3.1: Structure of the Speech Emotion Recognition System [3].

In another study on Issa's analysis of sound mood, Demirci and Yazici (2020) selected five characteristics of sound: 40 Mel-frequency Cepstral Coefficients (MFCCs), 12 attributes

Chromagrams, 128 Mel-scaled spectrograms, 7 attributes spectral contrast, and finally 6 Tonnetz attributes for a total of 193 attribute size vectors. structure The basic consists of CNN1D attached to the Dense Layer or Fully connected layer before going through the Softmax classifier as shown in (Fig. 3.2).

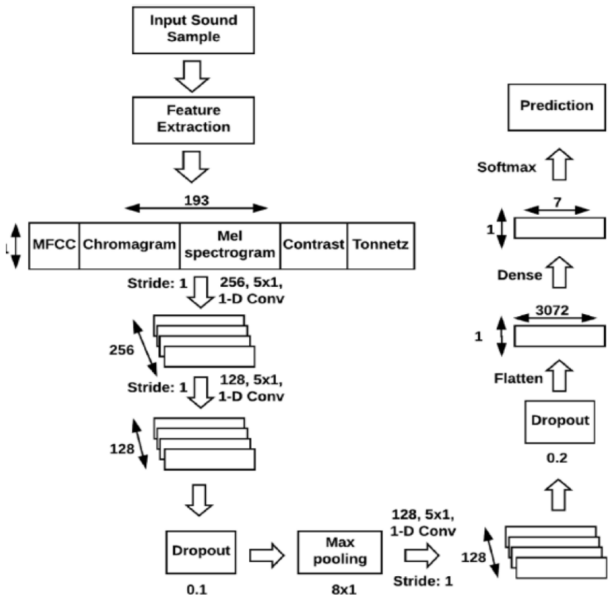

Figure 3.2: Structure of the Speech Emotion Recognition System [8].

Based on the above model structure(Fig. 3.2), when used to train with EMO-DB, which contains all 7 emotion classes: anger, sadness, happiness, boredom, stress, fear, and disgust, and measure performance with a test set, the model's accuracy is 82.86 percent In addition to this research, several models are presented with the use of Ensemble Methods by creating multiple models and using them. combined to help decide the end result.

Ramakrishnan's research uses classification, feature, database as well as normalization of signals and then preprocessing. He also suggested some that are not related to speech emotion recognition, but rather that it is necessary to do [21].

A study by Yun Jin(2014) presented a way to increase independent speech recognition rate using feature selection [9]. and feature integration methods based on multiple kernel learning (MKL). Initially, MKL was used to select features. Second, these preselected features are integrated at the kernel level. In the last step, the entire kernel will Combine, and the result will be a merged kernel. from the results of the experiment with the database Berlin, which consists of 7 emotions, has the highest recognition rate at 83.10 percent.

Research by Dipti D. Joshi and M.B. Zalte(2013) proposes an emotional recognition system

for Marathi speech, one of the Indian languages. Within the research features extracted from Speech consists of audio power, pitch, format, and discrete wavelet transforms, extracting vector-based features. Classifiers use SVMs in Identifying emotions such as angry, sad, happy, and normal emotional states [18].

According to a study by Jun-Seok Park and Soo-Hong Kim (2014), in the study, the use of fractal features in speech recognition systems Fractals are used to represent nonlinearity and properties. self-similarity of speech signals. A technical support vector machine is used as a classifier and recognition. The sound database is used as a database Berlin Standard, and the result is a recognition rate of about 77 percent[16].

CHAPTER 4

Background

4.1 Machine Learning

Machine Learning is an algorithm that allows computers to learn on their own from a set of input data. in order to be able to anticipate events and help humans make decisions machine learning has two types of learning: Supervised, which is learning from a data set with real answers to problems, and learning without an instructor. Unsupervised is learning from a series of unanswered data sets [13].

4.2 Support Vector Machine (SVM)

SVM is one of the algorithms that have a supervised learning model. It is responsible for making decisions, making a plan or decision boundary to isolate different classes, and must get the most margin by margin. margin is the perpendicular distance between the line and the point closest to the line (support vector). The SVM must to the line to be separating a hyperplane with proper class separation and optimization of the hyperplane for maximum margin [15].

4.3 Random Forest

Random Forest is one of a group of models of Ensemble Learning is a prediction by using several models together to make decisions. In the ensemble, there are two more ways to the ensemble, namely Bagging and Boosting. Random Forest uses the Bagging method to train in Parallel with each model receiving a set. Randomly dissimilar data, which is a subset of the entire data set. Each Random Forest model is a decision tree that is obtained by training different data. When each model makes a prediction, it uses the majority vote to decide the final outcome (Fig. 4.1)

4.4 Neural Network

A Multilayer Perceptron neural network is one of the supervised learning algorithms that simulate the structure of the nervous system of living things.

1) Input Layer is a layer that is responsible for receiving input data. It has the same number of Node or Neuron as the feature number, which is obtained by doing feature extraction on the audio data.

2) Hidden Layer is a layer that is hidden in the middle of the Input Layer and Output Layer. It is responsible for the processing. There can be many layers. In each hidden layer, there is a processor called Neuron. The number of neurons can be any number, for example The addition of layers and the number of neurons will affect the function of the model.

3) Output Layer is the layer on which we will use the calculated data. The number of nodes in this layer depends on the format of the output. Here, there are 5 nodes equal to the mood to be classified: normal, anger, happiness, sadness, and irritated, which each line Linked between neurons will have a weight. different which is obtained by learning from the Training data as shown (Fig. 4.2)

4.5 Convolutional Neural Network (CNN)

Convolutional neural networks are classified as a special neural network structure in which the learning capability. It extracts its own attributes (Feature Engineering) from the input data. Instead of using a normal activation function, CNN uses a convolution and pooling

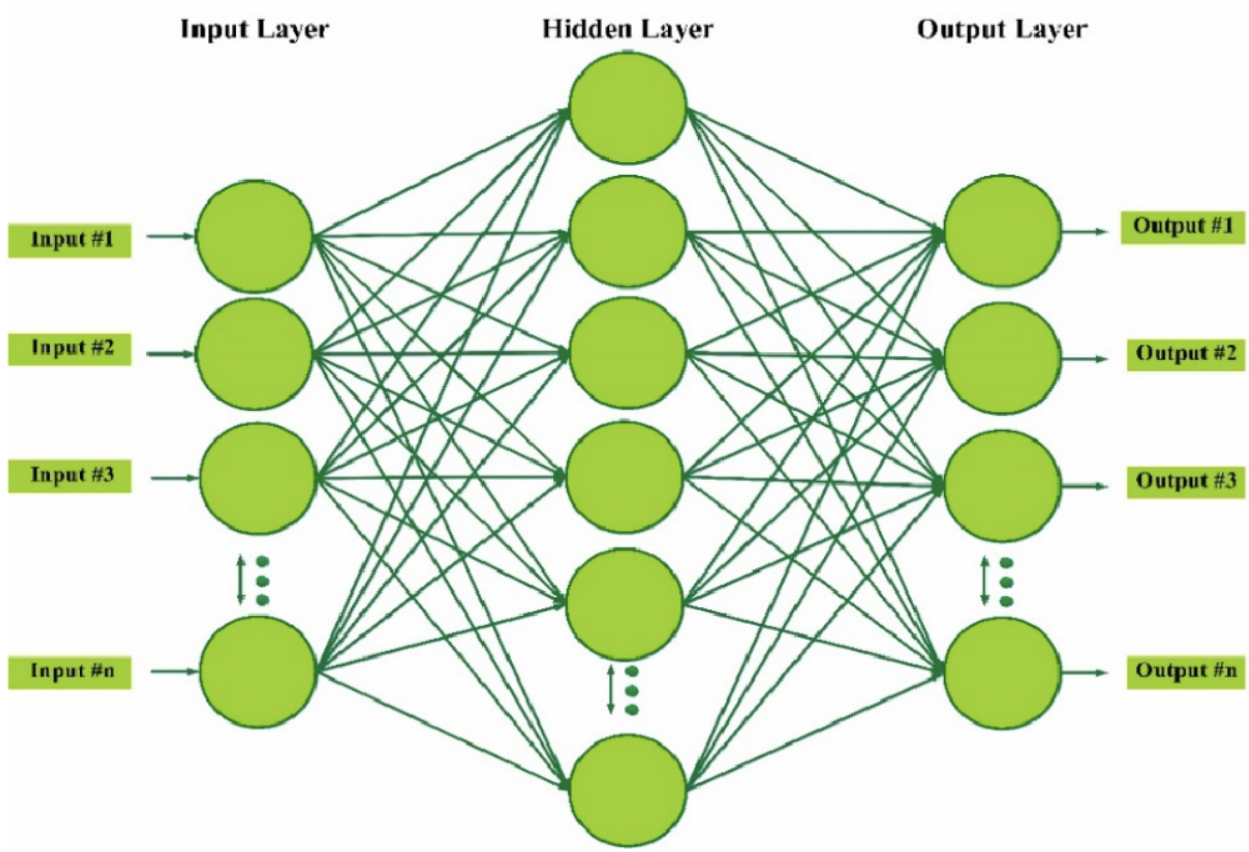

Figure 4.1: structure of Neural Network [12].

function instead. Convolution is an operation between two 1D data or between two 2D data. One will be Input and the other will be Kernel or Filter that will convolution to get a feature map. Pooling is filtering values to reduce their size. or dimension (Dimension) of the data. Pooling has many types, here will be an example. Max Pooling is a filter that finds the maximum value in the area where the kernel of max pooling is grafted to represent that area (Fig. 4.3)

4.6 Decision Tree

The Decision Tree is a rule-based model that creates an if-else rule from each feature is value without an equation to direct the relationship between feature and target. The important thing in creating a Decision Tree is choosing splits, each feature's value must be minimized. The value of the cost function is minimal [14].

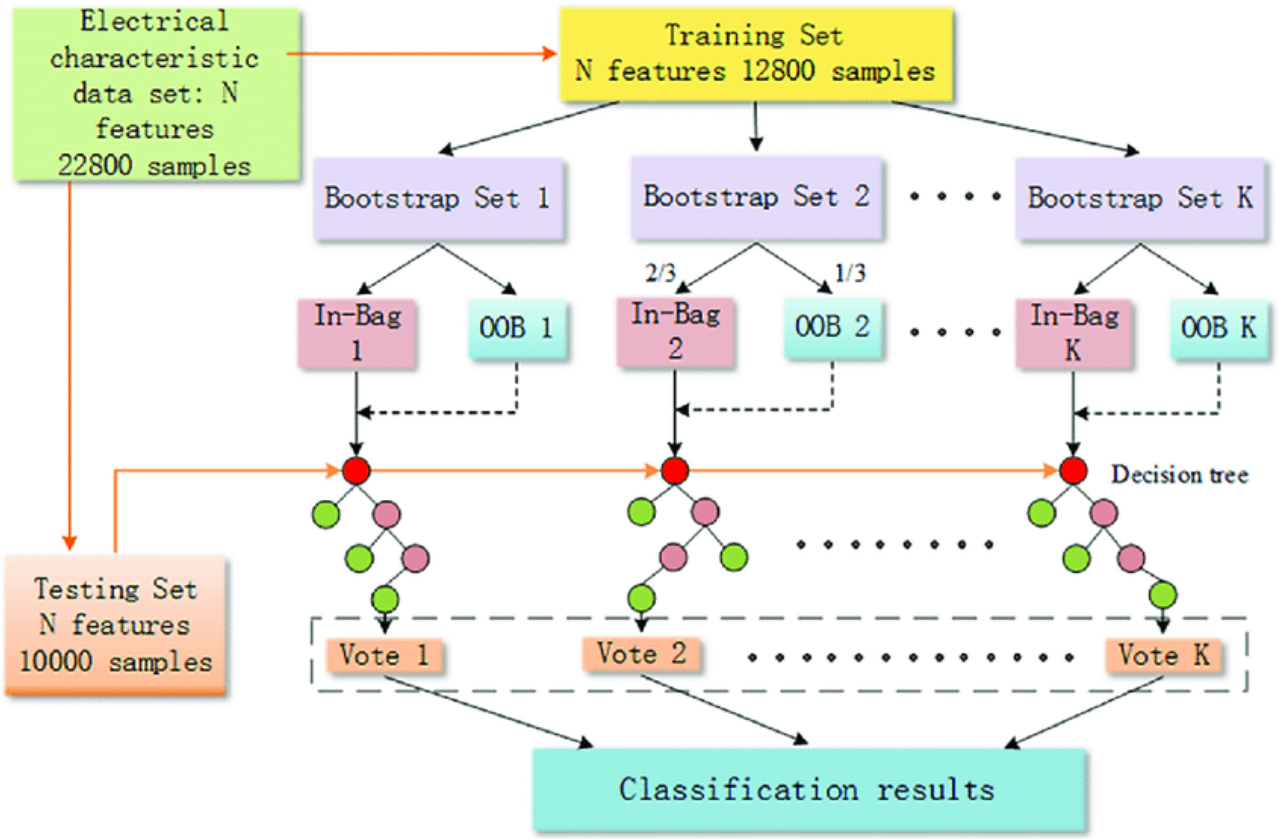

Figure 4.2: the structure of Random Forest [11].

4.7 Extreme Gradient Boosting

XGBoost, or Extreme Gradient Boosting, is a model that trains multiple Decision Trees in succession, where each decision tree learns from the errors of the previous tree, making prediction accuracy more and more accurate. When the learning of the tree continues until it is deep enough, the model stops learning when there are no more error patterns from the previous tree to learn [4].

CHAPTER 5

Research Questions

5.1 State the research question/topic you are addressing?

5.1.1 What methods are used to improve the model?

5.1.2 How will the impact of this research affect the world??

5.1.3 What are the key challenges of this research??

5.1.4 What do you use to measure accuracy in a model?

5.1.5 What impact will this research create on the AI industry?

5.2 Research gap

Speech emotion recognition has been around for a long time. Research and programming with code Much improved and more accurate but few research compares the model. However, this research aims to explore and study to compare Speech emotion recognition in all 5 models and improve to find the best parameter and model.

5.3 Hypothesis

As far as the history of speech emotion recognition, people tend to trust psychologists more in consulting our feelings, and AI opponents are viewed as an unstable tool. However, in the past decade and into the foreseeable future. That perspective has changed. People are starting

to trust AI more, they have to wonder if psychologists will reveal our secrets. But for AI, no secrets will be revealed. Speech emotion recognition has improved abilities and accuracy. In the future will replace psychologists. We hope that the Speech emotion recognition we are researching will increase our research knowledge and that the 5 models we will be working on will be better than any previous research.

5.4 limitation of work

Speech emotion recognition is currently more than 90 percent accurate, but it's an ideal number that means users can hear clearly without any background noise at all with a good microphone. But in the real world, this is not the case. There are some challenges to face that are unique to VUI(VOICE USER INTERFACES) design. These challenges as a VUI designer are beyond the designer's control. Aside from waiting for technology to improve design is the realization of imperfections: noise. One of the toughest challenges for ASR instruments is noise management. This includes static noises such as those heard while driving on a freeway or sitting in a busy restaurant or near a fountain. It may also include sounds that are produced when the user speaks, for example. the sound of a dog or Frying pan while cooking in the kitchen Other challenges include surrounding speech. (When a user talks to a friend or colleague while the app is listening) the TV in the background or several people talking at the same time. As mentioned earlier. there are many ways to deal with these challenges, so the best thing to do is to remember the user. When the system doesn't understand the user, what VUI can do is follow the techniques described previously. to help alleviate this problem as much as possible Sometimes the app will guess what the problem is and direct the user to move to a less noisy environment, closer to the microphone, and so on, but it's too dangerous to annoy the user [19]. with these tips Instead, the focus is on providing assistance through escalation of error behavior and offering other ways than audio to get users to continue. The device only responds to its own voice. But determining who's speaking is still a challenge for VUI if users are in the middle of browsing (Hey Siri, can you tell me the top-ranked California?) and the co-workers of the VUI are browsing. Used to start a computer conversation, how do you know who's listening? This problem may have to wait for technology. But it is still important to be aware of this and that the speaker is a child. It is more difficult for the ASR tool to recognize accurately. This is partly because children have

shorter vocals and higher-pitched vocals, and there is very little information available for that type of speech. Another reason is that young children are more likely to stutter, pause, and replay when designing an app for kids. may have to be extra careful. restaurants in Walnut Creek,

CHAPTER 6

Methodology

This research flowchart explains the steps we will do in the Speech Emotion Recognition model.

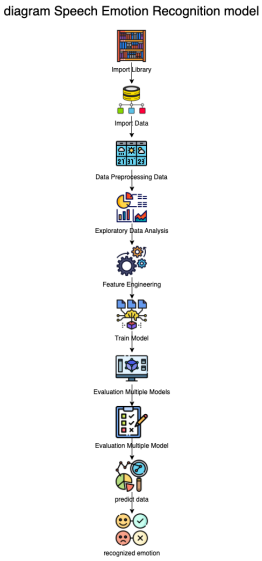

Figure 6.1: flowchart of the project.

6.1 Dataset

In this study, the Speech and Speech Emotion Recognition dataset was used to classify English speech emotions for six moods: sadness, anger, disgust, fear, happy, and neutral. 7442. Audio file developed by DMITRY BABKO[9]

6.2 Feature Extraction

Feature Extraction is a very important aspect of AI. We have to clean data and select the appropriate features to make the model work as efficiently as possible. Irrelevant or improper feature use may result in lower accuracy or overfitting. for feature Extraction, we will use Librosa library and will extract each audio file as a vector5 The set consists of

1) Mel-frequency cepstral coefficients (MFCC)

2) Chromagram

3) Mel-scaled spectrogram

4) Spectral contrast

5) Tonnetz

Studies have shown that the human ear They do not have the ability to perceive linear frequencies because research has shown that humans are better at distinguishing between lower frequencies than higher frequencies. 500 and 1000 Hz easily But we can hardly tell the difference between 10,000 and 10,500 Hz. Even though the distance between the two pairs is 500 Hz, the distance perception of the human ear holds. that they are not equal This led to the introduction of a new unit of sound level, called the Mel-frequency cepstral coefficients scale, where equidistant values in the Mel-frequency cepstral coefficients scale are equidistant in the human perception. The new unit was created because the human ear was assumed to be a reliable voice recognition machine. Mel-scaled spectrogram and MFCC have long been popular features in speech analysis and are still popular today. because it can be used to distinguish the identity However, they cannot differentiate between Pitch classes and Harmony (McFee et al., 2015). Chromatograms need to be included as a feature because they are relevant. With all 12 Pitch classes, for example, Class C counts for every note C regardless of the octave. same class, That is to say, to include all spectral data relating to that Pitch class into a single coefficient. Ultimately, the density of each of the 12 pitch classes is obtained. In addition to the Chromagram, Tonnetz is also found to be a characteristic that can be used to represent the Pitch class and the Harmony. Studies have shown that it is actually a 12-dimensional Chromagram. It was adjusted to show 6 dimensions instead (Harte et al., 2006). Spectral contrast is a study on music classification (Jiang et al., 2002). In this study, MFCC was compared with Spectral. contrast as a modeling feature. As a result, models built with the Spectral contrast feature are more efficient in accuracy. The last of the 5 vectors will

be concatenated into one vector.

6.3 Normalization

the attribute data used to train the model is distributed over a fairly wide range, it must be done. adjustments in the same range Suitable for use in the Train further easier In this work. The mean and std values used to optimize the Feature data are derived from the Train dataset. Both values must be stored for adjustments in the test dataset as well as for implementation [17].

6.4 train test split

We have to separate train and test because if we don't separate the data. Models can perform precisely with never-before-seen data. but will cause Overfitting. we train 80 percent and test 20 percent of the total dataset, through feature extraction and normalization. The past will be used to create each type of model.

6.5 Train model

Train all 5 models: Extreme Gradient Boosting Classifier,support vector machine, Random Forest,Convolutional neural network, Decision Tree. At this stage, we will start a Train or teach the model using a training set, and optimization will be involved. to find the optimal parameters for our ML model. Therefore,

6.6 Tuning model

After trying to train the model with default parameters, we use gridsearchcv. to find the best parameter of every model to extract the potential of each model To come out effectively, we have to adjust hyperparameter tuning until the desired accuracy is obtained. Hyperparameters suitable for that type of model that will result in a model with high accuracy or reducing the loss to the lowest value [25].

CHAPTER 7

Evaluation

Implementation of the model requires that the model's performance be measured first. which is generally a type of problem classification.

Classification uses a Confusion Matrix table where the number of rows and columns is equal to the number of answer classes.

For example, a basic table with two answer classes, True and False, would be a table. 2x2 as shown in (Fig 7.).

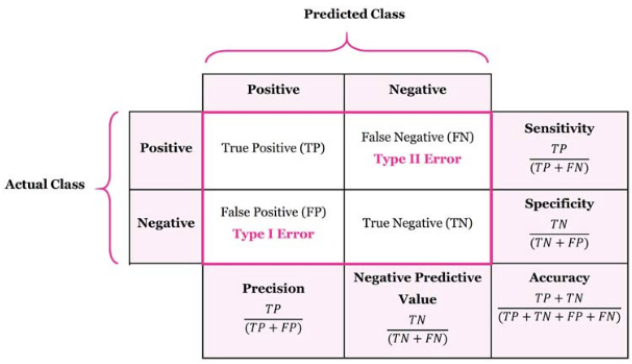

Figure 7.1: calculation of confusion maxtrix [23].

The table will consist of

1)True Positive is what the model predicts to be true. and the answer says it's true

2)True Negative is what the model predicts is not true. And the answer says it's not true.

3)False Positive is what the model predicts to be true. But the answer says that it's not true.

4)False Negative is what the model predicts is not true. But the answer said it was true.

which can be calculated as the accuracy or precision, precision, recall, and the F1-score or F-measure values are as follows:

1) Accuracy is a value that tells how accurately a model can make predictions [24].

2) Precision is obtained by comparing True Positive with False Positive [2].

3) Recall is accuracy value caused by comparing True Positive value with False Negative [6].

4) F1-score is the mean between Precision and Recall [5].

CHAPTER 8

Project Plan

8.1 JIRA

There are many types of Task Management software out there, but JIRA stands out for its ease of searching and customization. Users can design data collection. and the steps of the work are free Assign licenses notification style and connect to other software systems including a large number of extensions (add-ons) to use, we brought JIRA to apply to this work. The reason why we choose Jira:

Clarity creates clear communication. Using a variety of communication channels such as telephone, e-mail, and scattered information. Centralized user communication, exchanged information, forward work, and update job status in a single system. Help to find information accurately and quickly.

Transparency creates transparency in work. Organizational damage is sometimes caused by the concealment of mistakes and the refusal to openly disclose information. Or there may be condemnations that affect the morale of the team. Generally, it is difficult to investigate the root cause of problems retrospectively, but JIRA clearly records the activity of the work. This makes it easier to detect the source of the error.

Accountability Create a sense of participation and responsibility with multiple levels of work we may not understand common goals. Using JIRA will give members an overview of their work. easy job tracking Allows each person to see the status of his or her job responsibility and the impact on the work of others. encourage them to realize the importance

of their own work

Measurability Create measurable performance assessments. The success or failure of a task should be measurable, where JIRA can collect and summarize actual data to measure the effectiveness of different teams. It is precise information for supervisors or executives to assess team performance. This is why we decided to use Jira.

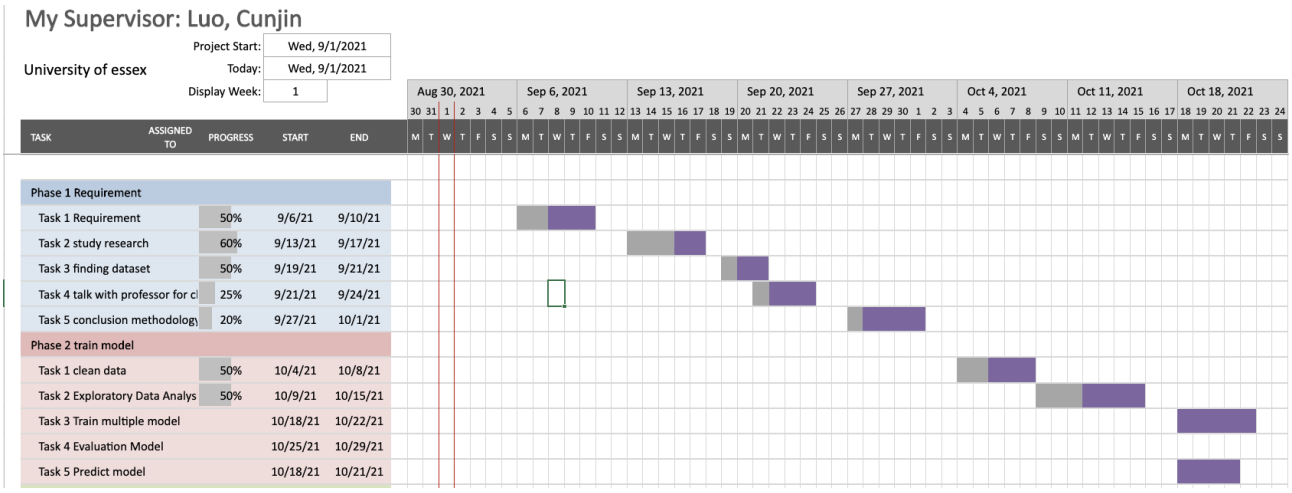

Figure 8.1: flowchart of the project.

8.2 Choosing a LIFE CYCLE of work

There are two types of life cycles: waterfall must be clearly planned and specified before the beginning of the project. Time, work, budget, people, and other factors in each waterfall phase are like waterfalls. When the water has flowed down The water will not flow back up to the top again. And the customer will receive the software only after the final waterfall has flown. Therefore, whenever an error or problem is encountered at the start of the Test stage, either through a misunderstanding of the requirements or due to any changes This made it quite difficult to fix and cost a lot of money. Most of them will be aware of these problems when the project is finished. making it not suitable for doing in this project form So we will use the second method, agile. Instead of planning everything Target and dash forward in one go. Then change to plan and work little by little. Break the goal into smaller pieces. and evaluate whether to continue Are you going in the right direction? If not, then hurry to find a solution and move on, so that when you encounter a problem, it will be easier to fix and manage. So when it comes to Agile, this song has to come up.

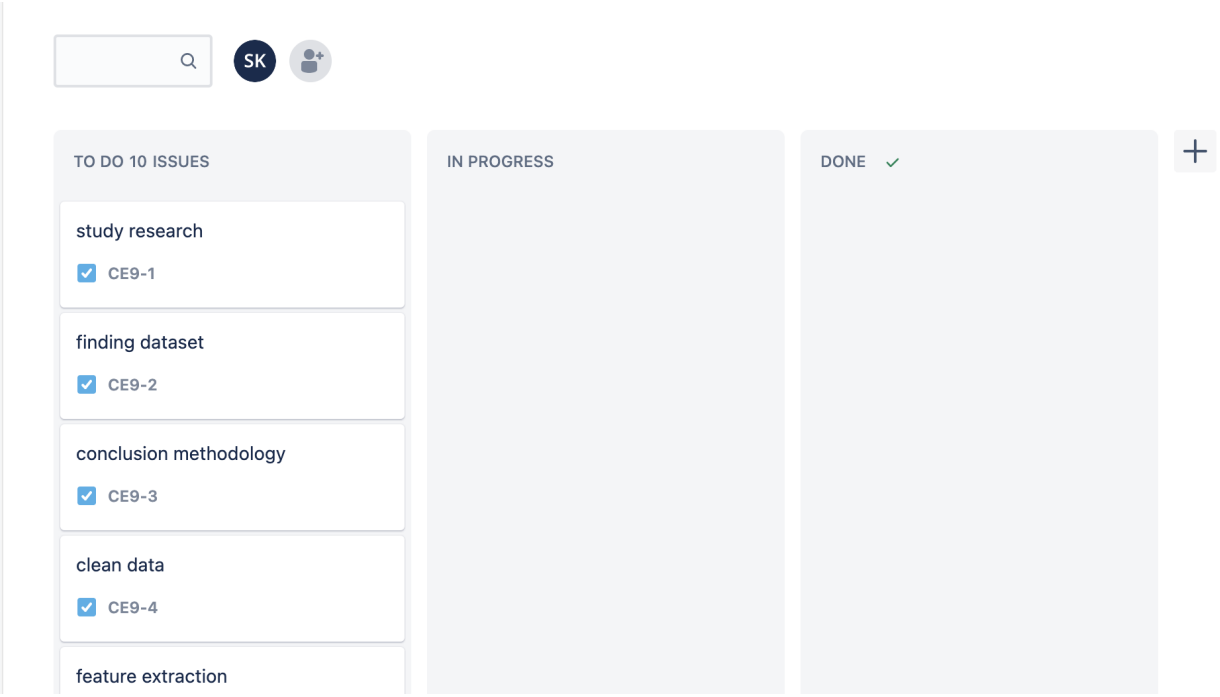

Figure 8.2: jira.

Bibliography

[1] Humaid Alshamsi, Veton KÃ«puska, and Hazza Alshamisi. 2018. Automated Speech Emotion Recognition App Development on Smart Phones using Cloud Computing. (05 2018). DOI:http://dx.doi.org/10.9790/9622-0805027177

[2] Danilo Bzdok and Andreas Meyer-Lindenberg. 2018. Machine learning for precision psychiatry: opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 3, 3 (2018), 223–230.

[3] Bharathi Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Sherly Elizabeth, and John McCrae. 2022. DravidianCodeMix: sentiment analysis and offensive language identification dataset for Dravidian languages in code-mixed text. *Language Resources and Evaluation* (02 2022), 1–42. DOI:http://dx.doi.org/10.1007/s10579-022-09583-7

[4] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, and others. 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2* 1, 4 (2015), 1–4.

[5] Davide Chicco and Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics* 21, 1 (2020), 1–13.

[6] Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*. 233–240.

[7] PM Dunn. 2006. Aristotle (384–322 BC): philosopher and scientist of ancient Greece. *Archives of Disease in Childhood-Fetal and Neonatal Edition* 91, 1 (2006), F75–F77.

[8] Dias Issa, M. Fatih Demirci, and Adnan Yazici. 2020. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control* 59 (2020),

101894. DOI:http://dx.doi.org/https://doi.org/10.1016/j.bspc.2020.101894

[9] Yun Jin, Peng Song, Wenming Zheng, and Li Zhao. 2014. A feature selection and feature fusion combination method for speaker-independent speech emotion recognition. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4808–4812. DOI:http://dx.doi.org/10.1109/ICASSP.2014.6854515

[10] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. 2017a. Feature selection: A data perspective. *ACM computing surveys (CSUR)* 50, 6 (2017), 1–45.

[11] Ke Li, Nan Yu, Pengfei Li, Shimin Song, wu Lei, Yang Li, and Meng Liu. 2017b. Multi-label spacecraft electrical signal classification method based on DBN and random forest. *PLoS ONE* 12 (05 2017). DOI:http://dx.doi.org/10.1371/journal.pone.0176614

[12] Mehdi Mirzaey, Mohammad Behdad, and Yousef Hojatpour. 2017. Applications of Artificial Neural Networks in Information System of Management Accounting.

[13] Tom M Mitchell and Tom M Mitchell. 1997. *Machine learning*. Vol. 1. McGraw-hill New York.

[14] Anthony J Myles, Robert N Feudale, Yang Liu, Nathaniel A Woody, and Steven D Brown. 2004. An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society* 18, 6 (2004), 275–285.

[15] William S Noble. 2006. What is a support vector machine? *Nature biotechnology* 24, 12 (2006), 1565–1567.

[16] Jun Park and Soo Kim. 2014. Emotion recognition from speech signals using fractal features. *International Journal of Software Engineering and its Applications* 8 (01 2014), 15–22. DOI:http://dx.doi.org/10.14257/ijseia.2014.8.5.02

[17] Pere Pujol, Dusan Macho, and Climent Nadeu. 2006. On real-time mean-and-variance normalization of speech recognition features. In *2006 IEEE international conference on acoustics speech and signal processing proceedings*, Vol. 1. IEEE, I–I.

[18] Sonali T Saste and SM Jagdale. 2017. Emotion recognition from speech using MFCC and DWT for security system. In *2017 international conference of electronics, communication and aerospace technology (ICECA)*, Vol. 1. IEEE, 701–704.

[19] Ben Shneiderman. 2000. The limits of speech recognition. *Commun. ACM* 43, 9 (2000), 63–65.

[20] Urmila Shrawankar and Vilas M Thakare. 2013. Techniques for feature extraction in speech recognition system: A comparative study. *arXiv preprint arXiv:1305.1145* (2013).

[21] Ramakrishnan Srinivasan. 2012. *Recognition of Emotion from Speech: A Review*. DOI: http://dx.doi.org/10.5772/39246

[22] AP Sunija, TM Rajisha, and KS Riyas. 2016. Comparative study of different classifiers for Malayalam dialect recognition system. *Procedia Technology* 24 (2016), 1080–1088.

[23] Sofia Visa, Brian Ramsay, Anca L Ralescu, and Esther Van Der Knaap. 2011. Confusion matrix-based feature selection. *MAICS* 710, 1 (2011), 120–127.

[24] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[25] Dong Yu, Li Deng, and George Dahl. 2010. Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition. In *Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning*. sn.