Received: December 12, 2021. Revised: March 5, 2022.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Study of Machine Learning Based Stressed Speech Recognition System

Barlian Henryranu Prasetio[1]* **Edita Rosana Widasari[1]** **Fitra Abdurrachman Bachtiar[1]**

[1]*Faculty of Computer Science, Brawijaya University, Indonesia*
* Corresponding author's Email: barlian@ub.ac.id

Abstract: The nonverbal communication processes a critical parcel. In some cases verbal communication is incapable since the speaker does not utilize non-verbal communication well at the same time. Non-verbal communication which falls in unconscious emotion is important in determining in function of cognition, language comprehension, and decision making. However, a little research studied in this area. Many years, researchers are amazed by the reliability of Mel-Frequency Cepstral Coefficients (MFCC) feature extraction technique in recognizing stressed speech. In this paper, we propose a simple feature extraction technique that effective but strong enough to recognize stressed speech. There are the speech energy and frequency. We attempted a basic approach to classify unbiased or stretch on female and male discourse. The highlight extraction is based on control and recurrence. This investigate utilized 10 female and 10 male discourse datasets. There are 5 classification strategies utilized. The classification models are Neural Arrange, k-Nearest Neighbour, Bolster Vector Machine, combination of NN-k-NN and combination of NN-SVM. Test comes about approved utilizing k-fold cross-validation strategy. The tests are assessed utilizing R-index to distinguish whether the highlights contributing to the push discourse acknowledgment. Based on exploratory comes about, the number of inputs impacts the esteem of R-index. In general, combining the Neural Organize and Back Vector Machine is the most excellent classification strategy by appearing stretch acknowledgment rate of 85% precision.

Keywords: Stress recognition, Speech, Machine learning, Deep learning, Neural network, Support vector machine.

1. Introduction

Communication is the most prerequisite in human life. The capacity to communicate influences how we oversee conduct, social communication and make individual choices. There are two sorts of communication: verbal and non-verbal. Verbal communication possesses an expansive parcel due to the reality, thoughts, considerations, or choices, more effectively verbally conveyed. The nonverbal communication possesses a vital parcel. Some of the time verbal communication is incapable since the speaker does not utilize non-verbal communication well at the same time. It can offer assistance communicators to advance fortify the message conveyed. One frame of non-verbal communication is vocal. Vocal or paralanguage could be a nonverbal component in a discourse to be specific, how to talk such as a tone of discourse, tone of voice, boisterous or powerless, talking speed, sound quality, sound, and others. In expansion, vocal data can too be communicated in feeling [1].

Feelings in nonverbal communication can be assembled into two fundamental categories: cognizant and oblivious [2]. Conscious communication is a communication when a person thinks about the verbal communication before the person communicate. Emotions that are consciously expressed are more easily recognizable. Some examples of conscious communications are anger, sadness, and happiness. Meanwhile, unconscious communication is a communication when a person does not think about nonverbal message during communication. Unconscious communication often happens naturally, and it displays emotion through nonverbal behaviours. Examples of unconscious emotion are stress and depression. Unconscious emotion is difficult to recognize compared to conscious emotion. The person showing

unconscious emotion do not realize when showing nonverbal behaviour.

Push acknowledgment is nearly indistinguishable to passionate acknowledgment [3], but stretch has another approach since stretch is an oblivious feeling. Emotions in nonverbal communications is important in human daily life. Previous study suggest that unconscious emotions are important in three aspects that is well-functioning of cognition, language comprehension, and decision making [4]. Previous studies have explored how emotions can be recognized from speech and how prosody stems from human brain [5]. Nonverbal communication plays an important role in one's daily life. It emphasizes the stress during one's speech. Unconsciousness emotion is not yet fully explored in any scientific attention even though the unconscious emotion plays critical role in speech understanding and communication [6]. Thus, this study is an attempt to explore about unconscious emotion specifically in speech emotion recognition. Machine learning can be helpful in identifying nonverbal communication, particularly related to emotion.

Existing literature has reported studies on stress based on speech information. [7] conduct a study to detect one's stress level based on conversation. The information used as the input in their study was tone change, viewed from the speech pitch and level. The result of stress recognition using sound level and pitch exhibit an accuracy rate of higher than 90%. The previous study also implements stress and non-stress recognition using speech signal. The study used speech waveform and glottal waveform. The extraction process was performed using mean shift clustering in order to distinguish the extracted Bispectral Feature. This study applying k-NN, ELM, and PNN/GRNN and it was found that neural network-based method can properly recognize stress and it exhibited better performance compared to k-NN. Another study attempts to detect stress through speech analysis by using two algorithms, Artificial Neural Network and Support Vector Machine [8]. The study applied Mel-Frequency Cepstral Coefficients (MFCCs) to extract the speech feature. MFCC is the most widely used technique for speech recognition because of its ease of implementation and flexibility [9]. The study reported a 90% accuracy in several data sets. Previous study detects speech emotion recognition using MELBP and Deep Belief Network (DBN) [10]. MELBP used for feature extraction and the DBN used for classification purpose. The result of their study shows best accuracy of 72.14%

Despite the contribution, the studies described previously also exhibit several limitations in terms of the recognition approach. The study conducted by [7] did not use machine learning and relied on threshold. This method causes difficulties in quick stress recognition. In addition, the use of threshold is considered unsuitable as the threshold changes when different data for stress recognition is employed. Meanwhile, in the work by [8] they have not considered combining the algorithm. These studies show that Neural Network, k-NN, and SVM algorithms exhibit satisfactory performance to recognize stress.

This paper presents a simple feature extraction and classification approach in recognizing stress. The shortcomings from the previous research will be considered in this study such as, machine learning utilization and combining the recognition model. Single model and combined model will be presented to recognize stress in speech. In addition, different with other works that use MFCC as the main feature extraction technique, our work uses a simple and efficient feature extraction technique. This technique results a low-dimension of feature but adequate to recognize the stressed speech.

2. Stress recognition approach

A speech feature needs to be extracted first prior to the modelling step. Feature extraction is important step as this process could influence the result of the prediction task. In the next section describes the type of measurement and how the speech could be extracted.

2.1 Measurement type

The push level can be measured utilizing the highlight examination. Sorts of push estimations and meddling degree can be seen in Fig. 1 [11].

The foremost precise stressed level estimation is employing a physiological parameter test. These estimations can be performed by a blood test or introducing an anode on the brain. Estimation of physiological parameters or therapeutic examination

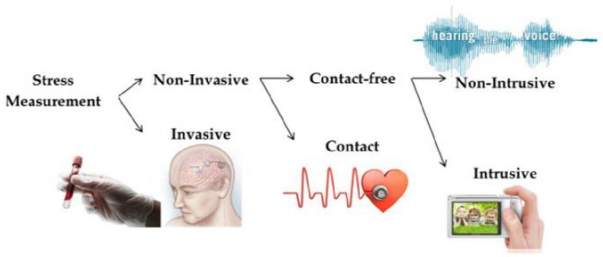


Figure. 1 The types of stressed measurements and intrusiveness degree

is called an obtrusive strategy [12]. Obtrusive estimations tend to require estimations from inside the body, such as cut or entry point. The invasive method considered as expensive as this method mostly need a surgical action. Non-invasive estimations can be made by physical contact but from exterior the body, such as utilizing heart rate sensors or measuring the skin utilizing EEG sensors. Whereas this strategy is non costly, be that as it may, the subject is cognizant that he/she is being measured.

Contact-free measurement is more convenient to analyse stress. However, if we use sensors on the subject body, it can cause the subject disturbed and can cause other effects. Another approach in measuring stress is to use a camera. Be that as it may, it may still be irritating since the subject feels directed [13]. The foremost likely push estimation is the non-intrusive method. The non-intrusive strategy could be a free-contact and does not aggravate the subject. This permits the member to feel ease when the test conducted and re-enact a common condition. This estimation can be done by conducting a voice investigation of the subject.

2.2 Speech and stress

From the over categorization it can be seen that discourse flag investigation is one of a conceivable strategy to recognize a stretch [7]. Discourse has been examined as a include to watch human push. Stretch can influence human physiques [14], such as human discourse, the eyes, and the brain. Push can influence a discourse flag when the speaker is beneath focused. Stress is one of product of unconscious emotion where the emotion comes out naturally without any control from the person itself. Fig. 2 appears the distinction between unbiased sound control and beneath stretch as re-enacted utilizing MATLAB Instruments. Fig. 2 speaks to the unbiased and push discourse control. As can be seen in Fig 2. the most extreme control and the frequencies have been changed where x axis is frequency (f) and y axis is magnitude (|P(x)|). In this way, it can be concluded that an individual who is beneath push might influence his discourse generation [15].

3. Speech stress recognition

Several steps need to be conducted to detect stress in speech. In our study, four basic steps are performed which starts from Speech Sampling, Pre-Processing, Highlight Extraction, and Design Classification. The workflow graph is based on the well-established and broadly utilized design acknowledgment approach [14]. Specifically, in the Pattern Classification single and combination model will be conducted to get the highest recognition accuracy. Fig. 3 shows the overall steps to detect stress. Each of the process is explained in the following sub-section.

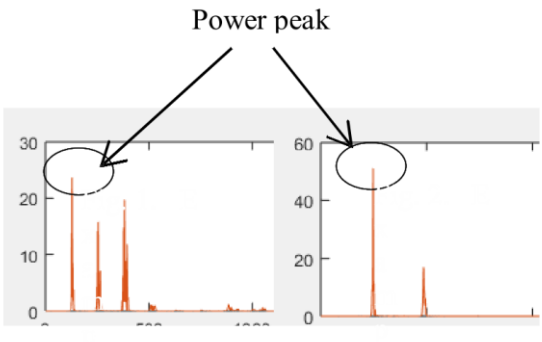


Figure. 2 Power signal for neutral and stress speech

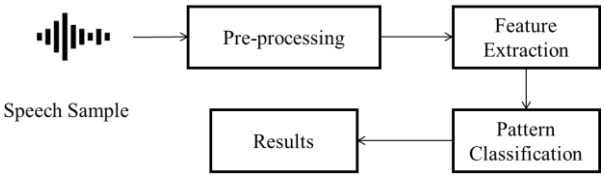


Figure. 3 The stress recognition system diagram using speech

3.1 Speech sample

In this research, we collected data of speech. There are 20 subjects participated in this study which consist of 10 male and 10 female. Each of the subjects is asked to pronounce 10 words that is, hello, okay, hey, come on, bye, see you, good, well, morning, and go. Each of the word is uttered in normal intonation. All the recorded speech is then labelled to neutral and stress speech.

3.2 Pre-processing

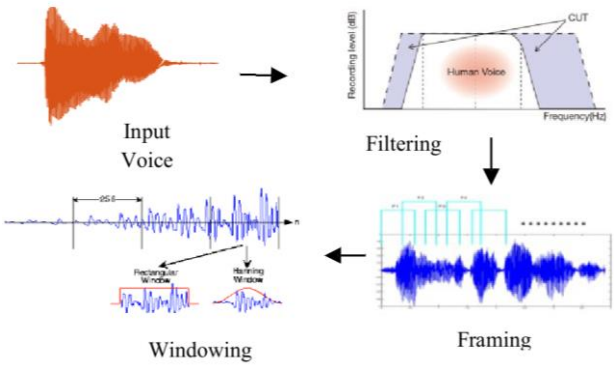


Figure. 4 The pre-processing process steps

In the next step is pre-processed the speech sample. In the pre-processing step there are three activities conducted which is Filtering, Framing, and Windowing. Each of the step is performed sequentially. The pre-processing steps can be seen in Fig. 4.

The sifting handle points to smother the tall recurrence parts amid the human voice generation instrument. The suppression of the frequency parts can be calculated using Eq. (1).

$$y[n] = \sum_{i=0}^{N} b_i . x[n-i] \quad (1)$$

where $x(n)$ is input signal, $y(n)$is output signal, N is the filter order, b_i is the value of the impulse response for $0 \leq i \leq N$ of an Nth-order FIR filter.

The surrounding handle points to fragment into a little outline with a length of 10ms with 50% cover with the outline another to be nonstop. The windowing handle points to decrease discontinuities at the starting and conclusion of each by increasing by the Hamming window within the time space which can be calculated utilizing Eq. (2).

$$y(n) = x(n).w(n) \quad (2)$$

where $x(n)$ is input signal, $y(n)$ is output signal, n is number of frames, and $w(n)$ is the Hamming window can be calculated using Eq. (3).

$$w(n) = 0.54 - 0.46\ \cos\frac{2\pi n}{N-1}, \quad 0 \leq n \leq N-1 \quad (3)$$

3.3 Feature extraction

In this step, feature extraction is conducted to get the feature that represent stress emotion. Stressed is a negative emotion. As a result, stress influences energy of the person who produce it. Previous study suggested that our thoughts, feelings, and emotions are related to the physical vibration frequencies [16]. Therefore, in this study energy will be calculated as the feature that represent emotion. The energy and frequency of stress can be calculated using Eqs. (4) to (6) as follows:

$$T = fft(x, N) \quad (4)$$

$$P = \frac{abs(T)^2}{N} \quad (5)$$

$$[Gy, Gx] = \max(P) \quad (6)$$

where x is speech signal, N is length(x), G_y is maximal power, G_x is Frequency (Powermax).

The feature extraction process is based on the speech partition in a small interval known as the frame. The speech signal will be taken feature power and frequency for each frame [17-19]. The result of feature extraction yields two features that is energy and frequency.

3.4 Pattern classification

The strategy of discourse classification is exceptionally differing. It depends on how we extricate the discourse highlight to be classified. The stretch classification calculations can be connected to distinctive inputs [20]. Accordingly, this study will examine three single model and two combined model where the combination model used Neural Network model as this model shows highest recognition in the previous study.

3.4.1. Neural network model

Neural Arrange (NN) gets input from exterior within the frame of design vectors and signals [21]. This input is scientifically decided by the documentation (xn) for n the number of inputs. Each input is duplicated by the fitting weight (w_{ij}). Weight is data utilized by Neural Systems to unravel an issue [22, 23]. The weights within the neighbourhood are updated as shown in Eq. (7).

$$w_{ij}\,(n+1) = w_{ij}\,(n) + \delta(n)[x(n) - w_{ij}\,(n)] \quad (7)$$

as a rule, the weights speak to the control of interconnection between neurons inside the Neural Arrange. The predisposition (δn) is utilized as a viper for the weighted input entirety, not 0. The limit esteem is determined from the output to reach at the required esteem since the real yield esteem

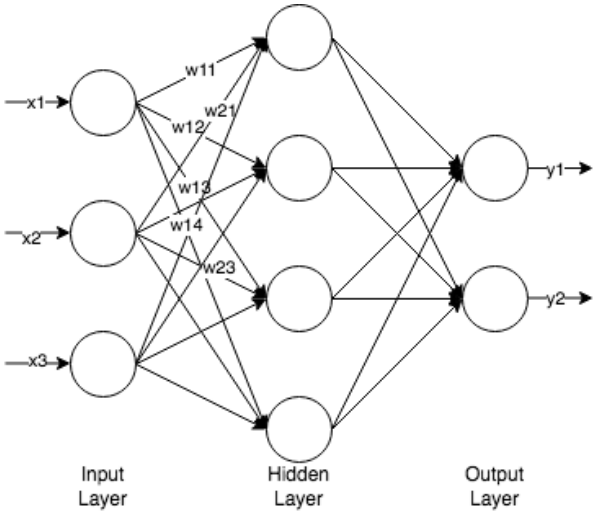


Figure. 5 The illustration of neural network architecture

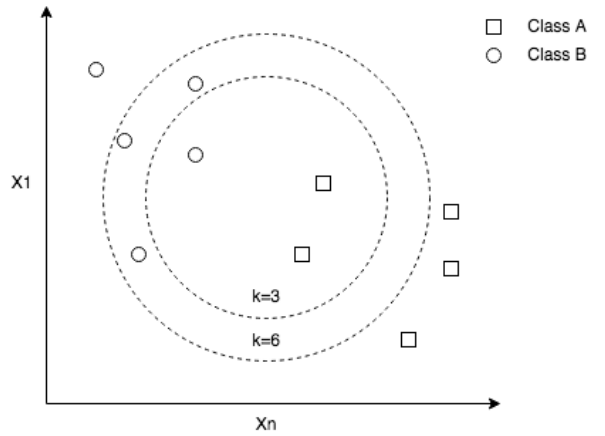


Figure. 6 The illustration of k-nearest neighbour algorithm

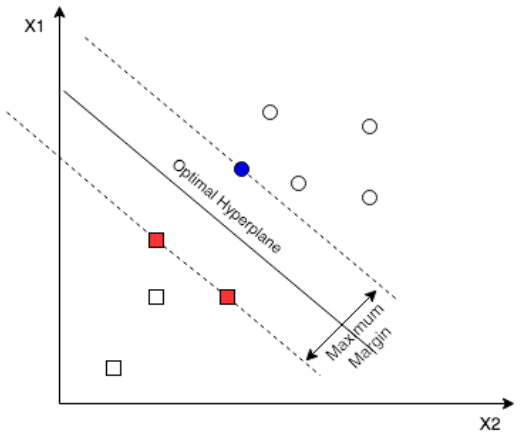


Figure. 7 The illustration of support vector machine algorithm

(yn) ranges from to limitlessness. This segment is called the actuation work.

This work contains the exchange work utilized to decide the required esteem. The exchange work can be either direct or nonlinear [24, 25]. The outline of Neural Organize engineering can be seen in Fig. 5.

3.4.2. k-nearest neighbor model

k-Nearest Neighbor (k-NN) could be a strategy to classify objects based on learning information closest to the protest. Learning information is anticipated into many-dimensional spaces, where each space speaks to a highlight [26, 27]. k-NN performs capacity of highlight vectors (xn) and test preparing information classification. This segment is partitioned into a few classes agreeing to the reason of classification. The point (*) in space is considered as course A since the point is closest to the center of lesson A. As a run the show, the neighbor isolated was calculated based on Euclidean isolated [28]. The k-NN features a few models. The models of k-NN depend on the number of neighbors (Fine k=1, Medium k=10, Coarse k=100) and remove network (Cosine= cosine separate framework, Cubic=Minkowski) utilized. The Outline of k-Nearest Neighbor calculation can be seen in Fig. 6.

3.4.3. Support vector machine model (SVM)

Back Vector Machine (SVM) may be a method that produces estimates, inside the case of classification and backslide. The SVM method tries to find the finest classifier/hyperplane work among an unbounded number of capacities to apportioned two objects. The ideal hyperplane may be a hyperplane that found halfway between two sets of objects of two classes [29, 30]. Finding the foremost amazing hyperplane is proportionate to maximizing edges or scattering between two sets of objects on particular classes. Within the occasion that a hyperplane supporter of the lesson, the edge between the two classes can be calculated by finding the evacuate between the two hyperplane-supporters of the two classes. The SVM encompasses a few models. The SVM models depend on portion work (Straight, Polynomial, Gaussian) utilized. The outline of the Bolster Vector Machine calculation can be seen in Fig. 7.

3.4.4. Hybrid model

The hybrid model is proposed by combining two model. The base model to be used is the Neural Network model. Neural Network chosen as the base model due to the ability to recognize stress with high accuracy by more than 90%. The base model is combined with two machine learning model, that is k-NN and SVM. Both of these models also shown to be able to recognize stress with relatively high recognition rate but below Neural Network model.

3.4.5. Validation model

In this stage, the adequacy of the proposed framework is approved employing a cross-validation strategy. The cross- approval strategy is one method to approve the exactness of a show built on a specific dataset. The improvement of a demonstrate ordinarily points to anticipate and classify unused data. The information utilized within the show improvement prepare is called preparing information. The information will be utilized to approve the demonstrate is called the test data.

One of the well-known strategies could be a k-Fold cross-validation. In this strategy, the dataset is isolated into a few bunches of k at arbitrary. In

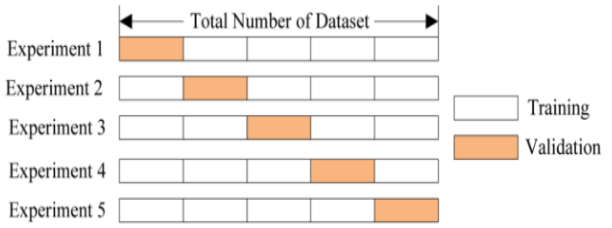


Figure. 8 The illustration of the 5-fold cross-validation

expansion, each bunch has the same number of individuals [31]. At that point a few exploratory K-times were conducted, each try utilizing k-group information as information testing and utilizing the other of the gather as preparing information. The outline of the 5-fold cross approval can be seen in Fig. 8.

Fig. 8 appears that the outline for 5-Fold Cross-Validation. The number of information set is separated into 5 bunches. 4 bunches are utilized as preparing information and 1 bunch as information testing. It is run by the number of bunches. The precision of the test comes about can be taken the normal esteem of all the tests. In this ponder a few k-Fold Cross Approval is utilized with k = 2, 4, 5, 10, 20.

4. Result and discussion

In this area will be depicted in more detail around the test of highlight extraction and classification strategy that has been utilized. The highlight extraction segment displayed designs of control and the frequencies of female and male speech. There are 5 classification methods used in this study that is, Neural Network (NN), k-Nearest Neighbor (k-NN), and Support Vector Machine (SVM), combination of NN-k-NN and combination of NN-SVM. The detail of each result is described in the next sub-section.

4.1 Feature extraction

Based on the test comes about, the discourse highlight extraction is taken from the greatest control and recurrence when the control is maximized. Control and recurrence information has been normalized. Sound tests are taken from 10 female and 10 male participants. In each of the participant two types of labels will be collected that is stress and no stress. For each of the label, the participant will be sampled 10 times. Thus, in total there are 20 samples for one subject. The summary of the features can be seen in Table 1.

In Table 1, we will see that 6/10 female discourse control has been diminished and 8/10 female discourse recurrence has been expanded. In the meantime, the 9/10 male discourse control has been diminished and 5/10 male discourse recurrence has been expanded. This extraction information shows that stress power in female and male participant tend to decrease. However, most female participant showing an increase in the stress frequency while male participant half of the participant shows frequency increases.

4.2 Classification

4.2.1. Neural network result

In this inquire about, we built the framework

Table 1. The voice feature extraction

Word	Female				Male			
	Neutral Power	Stress Power	Neutral Freq	Stress Freq	Neutral Power	Stress Power	Neutral Freq	Stress Freq
hello	0.1497	0.3228	0.1275	0.2412	0.0018	0,0034	0.9915	0.2627
okay	0.0661	0.544	0.2029	0.2990	0.0029	0.0043	0.2881	0.2797
hey	0.0135	0.0669	0.2147	0.2108	0.0031	0.0030	0.8983	0.9034
Come on	0.0739	0.0174	0.1333	0.3951	0.0019	0.0056	0.2593	0.8559
bye	0.9949	0.4439	0.4549	0.6755	0.0065	0.0096	0.2983	0.2915
See you	0.0830	0.0887	0.1559	0.3167	0.0022	0.0045	0.2136	0.2492
good	0.1979	0.1372	0.2216	0.2765	0.0032	0.0039	0.2949	0.3864
well	0.0651	0.1165	0.3029	0.5716	0.0043	0.0082	0.3051	0.6441
morning	0.3691	0.1825	0.3676	0.3324	0.0052	0.0047	0.3559	0.3051
go	0.1509	0.0038	0.2441	0.9990	0.0035	0.0143	0.8458	0.3000

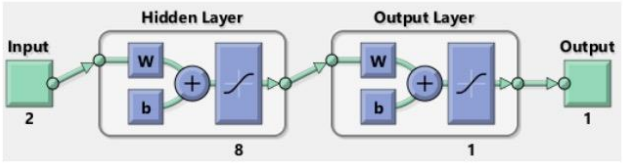


Figure. 9 The neural network architecture

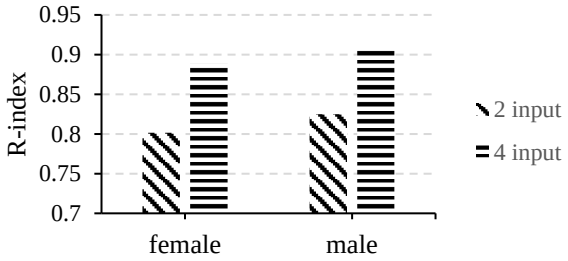


Figure. 10 The R-index of neural network

utilizing MATLAB tool compartment. Within the to begin with explore, we utilize NN with 3 organize layers which comprise of inputs layer, a covered-up layer, and a yield layer. The input layer is nourished by the extricated include within the past step. The arrange covered up layer has 8 neurons. In conclusion, the arrange target yield layer is unbiased or stretch. The organize was prepared to run 1000 times ages. The Neural Arrange engineering can be seen in Fig. 9.

In the Neural Network models, we propose two experiment scenarios. Different scenarios are proposed to maximize the Neural Network model. The primary explore comprises of 2 scenarios that are 2 inputs and 4 input systems. The primary situation employments 2 inputs: the most extreme control and frequencies when greatest control. The moment situation is 4 inputs, i.e., most extreme control, frequencies when greatest control, least control and recurrence when negligible control. Early identification of the relation of the input to the output in examined using R-index. The R-index shows the relationship between the outputs and targets of NN. The result of input-output relation can be seen in Fig 10.

Fig. 10 appears that the number of NN inputs influences the R-index esteem. NN with 4 inputs has an R-index increment of 8% compared to 2 inputs.

In the moment try, we utilized the same engineering with 4 input and approving utilizing k-fold approval strategy to know the framework exactness execution in a few folds. The results show that different k-fold yield in different recognition results. The best accuracy is 70% when the fold is 2 and 4 for both stress recognition for male and female. Whereas the lowest recognition is in fold 10

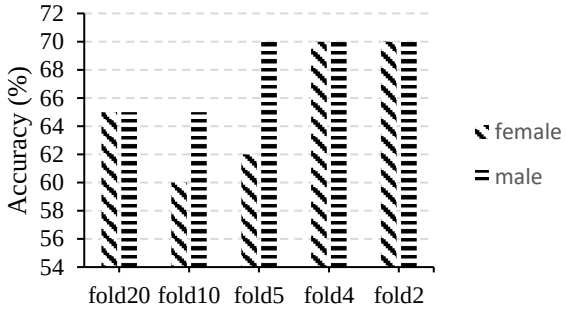
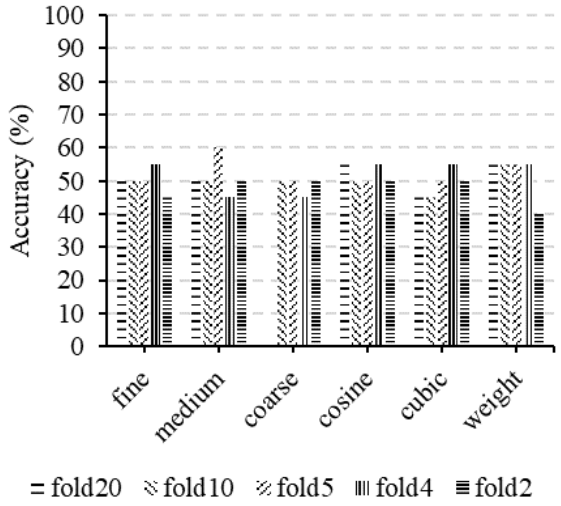


Figure. 11 The accuracy of neural network

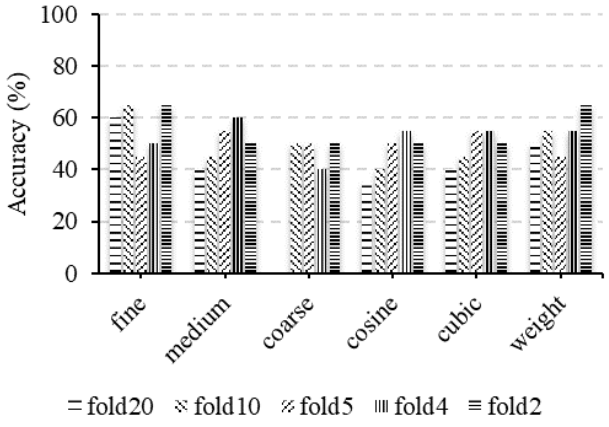
for female and fold 10 and 20 for male. The average stress recognition for female and male is around 60%. Fig. 11 shows the results for each fold.

4.2.2. k-NN result

The third test tried the exactness of the k-NN calculation with 4 inputs, i.e., greatest control, frequencies when the control is maximized, least

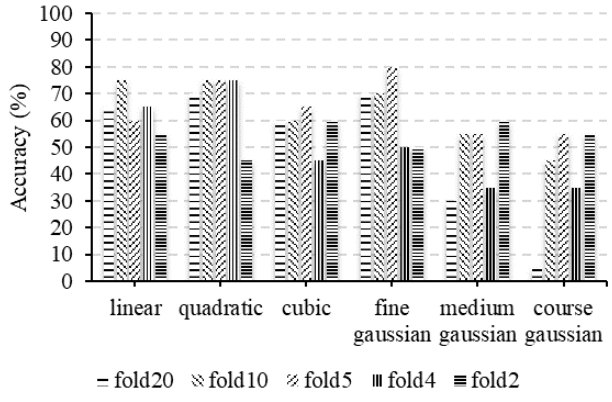


(a)

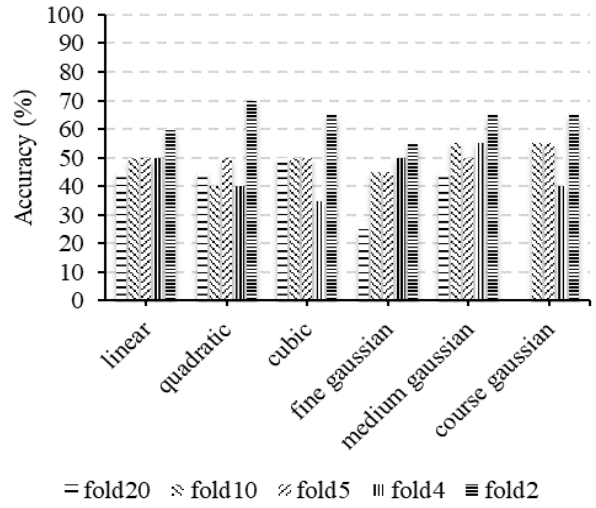


(b)

Figure. 12 The accuracy of *k*-NN algorithm: (a) Female and (b) Male



(a)



(b)

Figure. 13 The accuracy of SVM algorithm: (a) Female and (b) Male

control, and recurrence when the control is minimized. The test was performed with 6 k-NN models, i.e., fine, medium, coarse, cosine, cubic and weighted. Each test was endorsed utilizing cross-validation methodologies by changing the regard of overlay. The exactness of the k-NN calculation can be seen in Fig. 12.

The stress recognition results for female data shows relatively stable across different types of k-NN models. In addition, the results show relatively low for most of the model. There are two noticeable model that shows highest among all model that is medium and weighted k-NN. In the k-NN with medium number of k-neighbours using 5-fold shows the highest accuracy of 60%. On the other hand, in the weighted model shows that most of all fold shows relatively stable accuracy by showing 55% accuracy except for 2-fold showing the lowest accuracy by 40%. The graph of stress recognition for female respondent can be seen in Fig. 12(a).

The stress recognition results for male data shows fluctuates across different types of k-NN models. Overall, the recognition results show lower accuracy compared to the female respondent's data. However, some of the models shows highest accuracy such as k-NN with low k-neighbours, and Weighted k-NN. The k-NN model with low number of k-neighbours have two-fold that yield more than 60 % accuracy that is, fold 10 and fold 2. The other model, Weighted k-NN, has only one-fold showing more than 60 % accuracy that is fold 2.

4.2.3. SVM results

The fourth explore has tried the precision of SVM calculation with 4 inputs, i.e., most extreme control, frequencies when most extreme control, least control and recurrence when negligible control. The test was performed with 6 SVM models, i.e., coordinate, quadratic, cubic, fine Gaussian, medium Gaussian, coarse Gaussian. Each test was endorsed utilizing cross validation-method by changing the regard of overlay. The precision of SVM calculation can be seen in Fig. 13.

The stress recognition for female and male respondents varies among SVM models the models. For female data stress recognition, the highest accuracy is obtained by SVM Fine Gaussian using 5-fold by showing accuracy of 80% while the lowest accuracy is shown by SVM Coarse Gaussian with 20-fold yield in 5 % accuracy. With regard to stress recognition, the male data showed a fluctuating accuracy, where the lowest accuracy was found in SVM Fine Gaussian model with 20 folds, while the highest accuracy rate (i.e.,70%) was found in SVM Quadratic model with 2 folds.

4.2.4. Hybrid NN-k-NN results

The fifth explore is combining between NN and k-NN. The engineering can be seen in Fig. 14. Fig. 14 appeared that the design has 2 NN. Each NN has 2 inputs. The primary NN is control and recurrence when maximize. The moment NN is control and recurrence when minimize. Each yield of NN is as input the k-NN. The exactness of combining between NN and k-NN can be seen in Fig. 15.

The results for both female and male stress recognition data shows higher accuracy compared to a single model. Furthermore, in the stress recognition for both female and male data has increase around 10-20% compared with single model. Combination of NN-k-NN medium number of k, cosine, and cubic can reach 90% accuracy in some of the fold. Compared to those three models the cosine model shows more stable results in different number of folds. The stress recognition

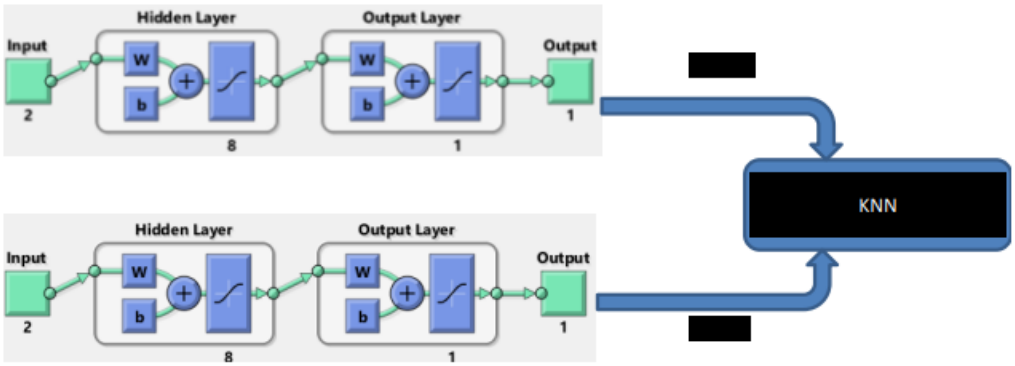


Figure. 14 Combining architecture between NN and KNN

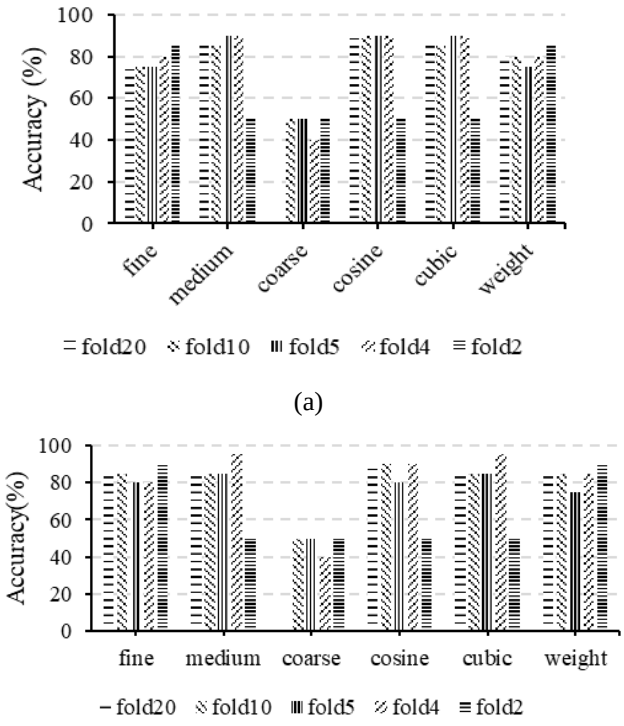


Figure. 15 The accuracy of combining between NN and KNN: (a) Female and (b) Male

results for male data shows better recognition compared to the female data. Some of the model in specific fold able to reach more than 90% accuracy which is achieved by NN-k-NN Medium and Cubic.

4.2.5. Hybrid NN-SVM results

The 6th test is combining between NN and SVM. The design is same with Fig. 14. In any case, we changed the k-NN with SVM. The exactness of combining between NN and SVM can be seen in Fig. 16.

The NN-SVM model shows relatively high recognition results for both female data and male data. The stress recognition for female data shows more stable accuracy across model and fold ranging from 70%-90%. The model that shows a relatively stable in all fold is NN-SVM Coarse Gaussian. This model shows 90% accuracy for fold 2, 5, 10, 20. Other model that able to reach 90% accuracy are

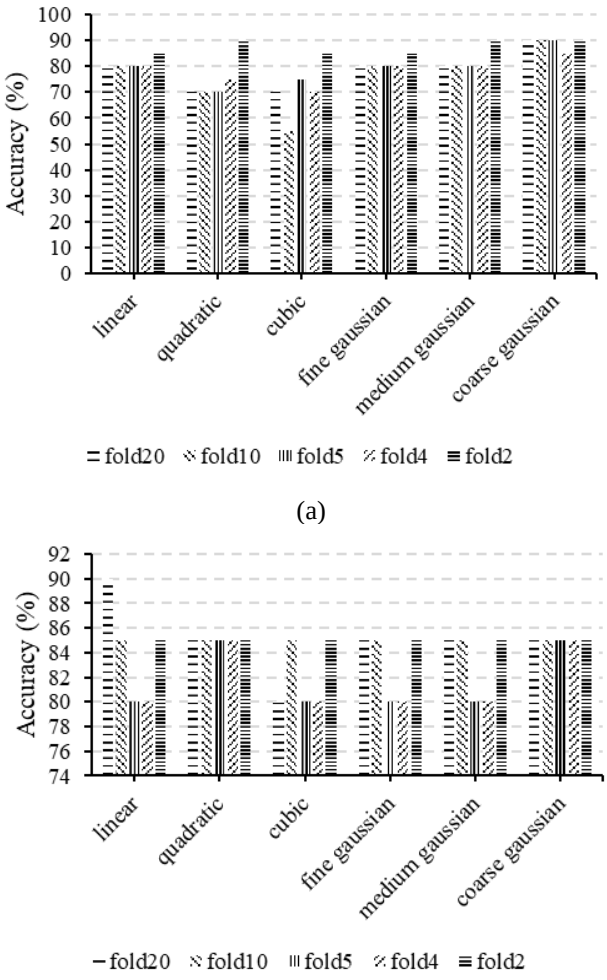


Figure. 16 The accuracy of combining between NN and SVM: (a) Female and (b) Male

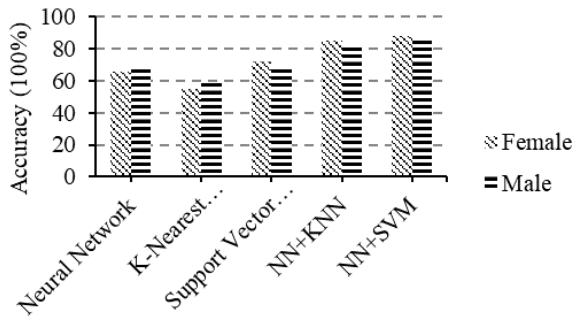


Figure. 17 The average classification method performance

Table 2. Comparison result of our work in term of accuracy

Reference	Accuracy (%)	
	SVM	NN
[32]	20,83	20,83
Our work	65,5	62,5

NN-SVM Medium Gaussian with 2-fold and NN-SVM Quadratic with 2-fold. The stress recognition for male data also shows stable recognition rate with less variability ranging from 80% - 90%. Notably, only one model could achieve 90 % accuracy that is NN-SVM Linear with 20-fold.

To summarize the result Fig. 17 takes the average of 10 runs of the model. Generally, all classification methods show fairly good results for both female and male data. However, compared to all the model the combination model shows higher accuracy for both female and male data. The combination model shows 80 % accuracy for both female and male stress recognition. The high recognition result may happen due to the amplification of the additional model to the base model.

In 2018, [32] presented speech analysis to detect stress by exploring the present of MFCC as feature extraction. The work of Tomba, et al [32] obtained that without MFCC, the accuracy score is low. To evaluate the effectiveness of our work, we compare the accuracy score of [32] and our proposed method. Table 2 shows the comparison of our work with [32] in term of accuracy without MFCC.

As showed in Table 2, our work outperforms the baseline method [32] in term of accuracy without MFCC feature. It shows that with a simple feature extraction, our work is effective and efficient enough for detecting stress in speech.

5. Conclusion

In this paper, we have built a framework that can recognize stretch through discourse. The framework is built employing a straightforward approach of discourse examination. The discourse power and recurrence ended up the most calculate include extraction. By and large, the contrasts between impartial and stretch are the control will diminish whereas the recurrence will increment. This investigate utilized 10 female and 10 male respondents and test their discourse. The framework tried 5 classification strategies. The strategies are the Neural Organize, k-Nearest Neighbour, Support Vector Machine, and combination of NN-k-NN and NN-SVM. All of the models were approved utilizing k-fold cross-validation. Based on the test comes about, the number of input highlights impacts the R-index esteem for the Neural Arrange classification strategy. By and large, combining the Neural Organize and Bolster Vector Machine is the leading classification strategy by appearing 85% exactness.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Theory and conceptualization, B.H.P. and E.R.W.; data requirement, B.H.P. and E.R.W.; methodology, B.H.P. and E.R.W.; software design and development, B.H.P.; validation, B.H.P., E.R.W. and F.A.B.; formal analysis, B.H.P. and E.R.W.; investigation, B.H.P.; writing–original draft preparation, B.H.P.; writing–review and editing, B.H.P., E.R.W. and F.A.B.; visualization, B.H.P. and E.R.W.; supervision, E.R.W. and F.A.B.

References

[1] S. Paulmann, D. Furnes, A. M. Bøkenes, and P. J. Cozzolino, “How psychological stress affects emotional prosody”, *PLoS One*, Vol. 11, No. 11, p. e0165022, 2016.

[2] J. Nicholson, K. Takahashi, and R. Nakatsu, “Emotion recognition in speech using neural networks”, *Neural Comput. & Appl.*, Vol. 9, No. 4, pp. 290-296, 2000.

[3] E. L. Shifflett, "Pitch, Frequency and Duration:Using Phonetic Correlates to Determine the WordLevel Stress Pattern of Tunisian Arabic”, *Undergraduate Honors Theses*, Paper 432, 2011

[4] L. Perlovsky and F. Schoeller, “Unconscious emotions of human learning”, *Phys. Life Rev.*, Vol. 31, pp. 257-262, 2019.

[5] S. A. Kotz and S. Paulmann, "Emotion, language, and the brain", *Lang. Linguist. Compass*, Vol. 5, No. 3, pp. 108-125, 2011.

[6] D. A. Sauter, F. Eisner, A. J. Calder, and S. K. Scott, "Perceptual cues in nonverbal vocal expressions of emotion", *Q. J. Exp. Psychol.*, Vol. 63, No. 11, pp. 2251-2272, 2010.

[7] N. Matsuo, S. Hayakawa, and S. Harada, "Technology to detect levels of stress based on voice information", *Fujitsu Sci. Tech. J*, Vol. 51, No. 4, pp. 48-54, 2015.

[8] K. Tomba, J. Dumoulin, E. Mugellini, O. A. Khaled, and S. Hawila, "Stress Detection Through Speech Analysis", In: *Proc. of the 15th International Joint Conference on e-Business and Telecommunications*, Vol. 1 pp. 394-398, 2018.

[9] S. Ali, S. Tanweer, S. S. Khalid, and N. Rao, "Mel Frequency Cepstral Coefficient: A Review", In: *Proc. of the 2nd International Conference on ICT for Digital, Smart, and Sustainable Development*, 2021

[10] S. N. Mohammed and A. K. A. Hassan, "Speech Emotion Recognition Using MELBP Variants of Spectrogram Image", *Int. J. Intell. Eng. Syst.*, Vol. 13, No. 5, pp. 257-266, 2020, doi: 10.22266/ijies2020.1031.23.

[11] M. Hagmuller, E. Rank, and G. Kubin, "Can stress be observed by analyzing the human voice", In: *Proc. of 3rd Eurocontrol Innovative Research Workshop*, 2004.

[12] M. Hagmüller, E. Rank, and G. Kubin, "Evaluation of the human voice for indications of workload-induced stress in the aviation environment", *EEC Note*, Vol. 18, No. 06, 2006.

[13] H. Gao, A. Yüce, and J. P. Thiran, "Detecting emotional stress from facial expressions for driving safety", In: *Proc. of IEEE International Conference on Image Processing*, pp. 5961-5965, 2014.

[14] X. Liu, "Voice stress analysis: Detection of deception", *Master's Thesis Univ. Sheffield, Dep. Comput. Sci.*, 2005.

[15] J. A. C. García, A. J. M. Cantero, I. M. G. González , S. L. Arroyo, and M. M. Monge, "Towards Human Stress and Activity Recognition: A Review and a First Approach Based on Low-Cost Wearables", *Electronics*, Vol. 11, pp. 1-30, 2022.

[16] M. Benzeghiba, R. D. Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review", *Speech Commun.*, Vol. 49, No. 10-11, pp. 763-786, 2007.

[17] J. H. L. Hansen and S. Patil, "Speech under stress: Analysis, modeling and recognition", *Speaker Classification I*, pp. 108-137, 2007.

[18] J. Basavaiah, C. Patil, and C. Patil, "Robust feature extraction and classification based automated human action recognition system for multiple datasets", *Int. J. Intell. Eng. Syst.*, Vol. 13, No. 1, pp. 13-24, 2020, doi: 10.22266/ijies2020.0229.02.

[19] N. R. P. Pullaiah, D. Venkatasekhar, P. Venkatramana, and B. Sudhakar, "Detection of Breast Cancer on Magnetic Resonance Imaging Using Hybrid Feature Extraction and Deep Neural Network Techniques", *Int. J. Intell. Eng. Syst.*, Vol. 13, No. 6, pp. 229-240, 2020, doi: 10.22266/ijies2020.1231.21.

[20] F. Yasmeen and A. Rahim, "Detection of Stress by Voice & Textual Information Analysis", *Project Report CSE6339 Winter*, 2014, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.700.3964&rep=rep1&type=pdf

[21] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine", In: *Proc. of the Interspeech*, 2014.

[22] Xenonstack, "A Medium Corporation", 2017. https://medium.com/@xenonstack/overview-of-artificialneural-networks-and-its-applications-2525c1addff7 .

[23] R. F. Rachmadi, I. K. E. Purnama, S. M. S. Nugroho, and Y. K. Suprapto, "Family-aware convolutional neural network for image-based kinship verificatio", *Int. J. Intell. Eng. Syst.*, Vol. 13, No. 6, pp. 20-30, 2020, doi: 10.22266/ijies2020.1231.03.

[24] T. Næs and E. Risvik, *Multivariate Analysis of Data in Sensory Science*, 1996.

[25] A. N. Sihananto and W. F. Mahmudy, "Rainfall forecasting using backpropagation neural network", *Journal Inf. Technol. Comput. Sci.*, Vol. 2, No. 2, 2017.

[26] L. Gao, J. Song, X. Liu, J. Shao, J. Liu, and J. Shao, "Learning in high-dimensional multimedia data: the state of the art", *Multimed. Syst.*, Vol. 23, No. 3, pp. 303-313, 2017.

[27] D. Syauqy, H. Fitriyah, and K. Anwar, "Classification of Physical Soil Condition for Plants using Nearest Neighbor Algorithm with Dimensionality Reduction of Color and Moisture Information", *Journal Inf. Technol. Comput. Sci.*, Vol. 3, No. 2, pp. 175-183, 2018.

[28] L. Y. Hu, M. W. Huang, S. W. Ke, and C. F. Tsai, "The distance function effect on k-nearest neighbor classification for medical datasets",

Springerplus, Vol. 5, No. 1, pp. 1-9, 2016.

[29] I. Carmichael and J. S. Marron, "Geometric insights into support vector machine behavior using the KKT conditions", *ArXiv Prepr. arXiv1704.00767*, 2017.

[30] M. G. L. Putra, W. Ariyanti, and I. Cholissodin, "Selection and Recommendation Scholarships Using AHP-SVM-TOPSIS", *Journal Inf. Technol. Comput. Sci.*, Vol. 1, No. 1, pp. 1-13, 2016.

[31] C. Ieracitano, F. Pantò, P. Frontera, and F. C. Morabito, "A neural network approach for predicting the diameters of electrospun polyvinylacetate (PVAC) nanofibers", In: *Proc. of International Conference on Engineering Applications of Neural Networks*, pp. 27-38, 2017.

[32] K. Tomba, J. Dumoulin, E. Mugellini, O. A. Khaled, and S. Hawila, "Stress Detection Through Speech Analysis", In: *Proc. of International Joint Conference on e-Business and Telecommunication*, pp. 394-398, 2018.