Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)
Vol. 10, No. 3, September 2024, pp. 507-519
ISSN: 2338-3070, DOI: 10.26555/jiteki.v10i3.29022

Analysis Kernel and Feature: Impact on Classification Performance on Speech Emotion Using Machine Learning

Jutono Gondohanindijo[1,2], Edi Noersasongko[1], Pujiono[1], Muljono[1]
[1]Faculty of Computer Science, Dian Nuswantoro University, Semarang 51031 Indonesia
[2]Faculty of Technics and Informatics, AKI University, Semarang 50136 Indonesia

ARTICLE INFO

Article history:

Received June 08, 2024
Revised July 16, 2024
Published August 05, 2024

Keywords:

Kernel Classifier;
Feature Engineering;
Dataset;
Analysis;
Speech Emotion Recognition

ABSTRACT

The main objective of this study is to test the machine learning kernel's selection against the characteristics of the data set used, resulting in good classification performance. The goal of speech emotion recognition is to improve computers' ability to detect and process human emotions in order to improve their ability to respond to interactions between people and computers. It can be applied to feedback on talks, including sentimental or emotional content, as well as the detection of human mental health. One field of data mining work is Speech Emotion Recognition. One of the important things in data mining research is to determine the selection of the kernel Classifier, know the characteristics of datasets, perform Engineering Features and combine features and Corpus Datasets to obtain high accuracy. The research uses analysis and comparison methods using private and public datasets to detect speech emotions. Experimental analysis was done on the characteristics of datasets, selection of kernel classifiers, pre-processing, feature and corpus datasets fusion. Understanding the selection of a classifier kernel that matches the characteristics of the dataset, engineering features and the merger of features and datasets are the contributions of this investigation to improving the accuracy of the classification of speech emotion data. For models with the selection of kernels that match the characteristics of their datasets, this study gave an increase in accuracy of 12.30% for the private dataset and 14.80% for the public dataset, with accuracies of 100.00% and 74.80% respectively. Combining features and public datasets provides an increase in accuracy of 33.62% with an accuracy of 73.95%.

Corresponding Author:

Muljono, Faculty of Computer Science, Dian Nuswantoro University Semarang 51031 Indonesia
E-mail: muljono@dsn.dinus.ac.id

1. INTRODUCTION

In daily life, human speech is one of the most prevalent ways that information is transferred [1]. Human speech can convey a range of information in casual conversation, including emotion. In order to establish discussion in social situations, it is essential to understand the other person's emotional state. When one knows the kind of feeling a person is experiencing, one can treat them and their attitude toward them accordingly [2].

Making interaction processes more sophisticated is a result of the development of information technology and the growing demand for human computer interaction or HCI. Speech is a basic yet powerful medium for advanced interaction [3], [4]. Because consumers implicitly treat computers like humans in general, the development of spoken command systems must also consider the user's emotional state [5]. Therefore, the availability of emotion data references is necessary for the development of advanced Human Computer Interaction (HCI) systems, as it forms the foundation for the creation of artificial intelligence (AI) systems that are capable of identifying human emotions.

In general, the diversity of the tribes does not correspond or balance with the availability of emotional information, making it impossible for all feelings to be conveyed equally [6]. However, when compared to the

emotion identification procedure in spoken speech, not all speech-based emotion detection models perform well. The number of balanced classes (balance), the machine learning technique employed, the feature extraction procedure, and the dataset utilized all have a significant impact on the outcomes of speech-based emotion recognition [7].

In machine learning, the characteristics of datasets refer to the various aspects that describe the data used to train models [8]. Understanding these characteristics is crucial because they have a significant impact on the performance of models and techniques used to process and analyze data. Some characteristics of datasets are: Datasets size, Dimensionality, Data Types, Data Distribution, Noise and Outliers, Missing Values.

By understanding the characteristics of datasets, we can choose and apply the right techniques for preprocessing, model selection, and evaluation that will improve the accuracy and reliability of machine learning models.

The use of kernel functions allows the algorithm to operate in a larger feature space without having to explicitly calculate the coordinates of the datasets . In this way, we can handle nonlinear relationships between data without having to perform explicit transformations [9]. There are several types of kernel functions: Linear, Polynomial, Radial Base Function, Sigmoid.

For example, if we use the hyperparameter setting of the RBF kernel in SVM to separate data that cannot be separated linearly, the algorithm will transform the data into a higher dimensional space where the data can be separated by a hyperplane. The SVM with the RBF kernel is used to classify iris data. The rbf kernel allows the model to handle data complexity that may not be linear in the original feature space [10].

Data mining operations, which begin with data collection, pre-processing, classification, and evaluation, are not independent of speech emotion identification processes [11]. In data mining research, selecting kernel classifiers, understanding dataset characteristics, performing engineering features, and combining features and datasets to generate characteristics that can identify classes are all crucial [10], [12], [13], [14], [15], [16]. The impact of kernel classifiers on data properties, feature engineering, and the combination of features and datasets to achieve high accuracy will all be analyzed in this study.

K-Nearest Neighbor (kNN), Naïve Bayes (NB), Neural Network (NN), Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), and Deep Neural Network (DNN) are a few examples of machine learning techniques used in categorization [17].

The data goes through phases of extraction and engineering features before classification. Five features—MFCC, Chroma, Mel-Spectogram, Contrast, and Tonnetz—are used in feature extraction [48]. Use the Principal Component Analysis technique (PCA) for feature engineering. Two datasets and a number of features were combined in an effort to increase accuracy [18].

Pre-processing, or the extraction of audio properties based on Mel Frequency Cepstral Coefficients (MFCC), Chroma, Mel-Spectrogram, Tonnetz, and Contrast, is how identification is accomplished. The implementation of Principal Component Analysis, or PCA, is the next step. The procedure of identification using DNN and machine learning comes next. The datasets utilized come from the RAVDESS (Ryerson Audio-Visual Database of Emotion Speech and Song) and the Indonesian Private Datasets Speech Emotion and Public Data. The evaluation's findings were then examined using the metrics of accuracy, precision, and recall on the confusion matrix table.

The research's contributions are as follows: 1. Examining how the classifier kernel's application affects classification performance, 2. Examining how the Feature Engineering Algorithm is used in relation to the properties of the Dataset and its Classifiers 3. Merging the datasets and features to increase precision. The associated study, specifically on Speech Emotion Recognition (SER), will be discussed in the next section. The research approach will next be explained in the third section. The experiment's results and their analysis will then be shown in the following section, and the study's conclusion will be found in the fifth section.

2. LITERATURE REVIEW

The goal of Speech Emotion Recognition (SER) is to build a method that uses various machine learning and data processing techniques to identify emotions in speech. According to the findings of Singh and Goel's literature review [19], there is a great need for research in the SER, particularly in practical applications. The research's findings lead to the conclusion that a number of factors affect the performance and outcomes of the SER during development, including the kinds of classifiers used in model training, the available datasets used in model development, and the procedure for extracting features pertinent to the type of emotion.

Many studies have attempted to construct models connected to SERs; Chowdary and Hemanth's research [20] was one of the proposals. The study suggested building SER with a convolutional neural network (CNN) and the RAVDESS dataset. The extraction of MFCC features from the REVDESS dataset marked the start of

the study phase. Subsequently, the feature extraction result was verified using 1642 training data and 810 test data. The CNN model based on Conv1D was trained using the training data as a training benchmark. Testing data was then used to assess the model. The accuracy performance of the model was 71.35%, according to the evaluation findings.

Additionally, Iqbal *et al*. [21] proposed using Artificial Neural Networks (ANN) to recognize speech emotion. The study used a speech-based emotion detection dataset called the Berlin Database of Emotion Speech, or Berlin EmoDB. Many feature extractions are used, such as formant, amplitude, pitch, and frequency. In the meantime, the Bayesian Regularized (BRANN) classifier model is ANN based. The model's accuracy performance was 95% according to the evaluation. Multi-features approaches such MFCC, Cross Zero Rate, Root Mean Square (RMS) [22], Chroma, Mel-spectrogram, Contrast, and Tonnetz [23], [24] were also used in several investigations. One way to enhance the classifier model's effectiveness in identifying different emotion types is to employ multiple features [25].

The distribution of data, class variety, dimensionality, feature correlation, linear and non linear patterns, outliers, and noise are some characteristics of datasets that might influence the choice of machine learning models and methods [8].

Comprehending the characteristics of these datasets is crucial for machine learning and model selection methodologies. For datasets with linear separation, for instance, an SVM with a linear kernel may perform well; however, datasets with more complicated patterns might benefit more from an SVM with a non linear kernel [26]. Selecting a classifier kernel that aligns with the dataset's features is essential for detecting speech emotions since it influences the classification model's efficacy and precision [9] .

Support vector machines (SVM) and k-Nearest Neighbors (kNN), two machine learning algorithms that are sensitive to data dimensions, can be employed with the PCA Engineering Feature technique since it solves high-dimensional problems on datasets [27]. PCA can assist in reducing feature redundancies in datasets with high rates of redundancy, which will enhance the performance of machine learning algorithms, particularly in models that are susceptible to multicolinearity issues [28].

PCA addresses the issue of multicolinearity by extracting the primary components that are orthogonal and appropriate for datasets where the features are highly correlated with each other and have different scales. It is also sensitive to outliers in datasets and performs best on datasets with a linear relationship between features [27].

It's crucial to keep in mind that while PCA performs admirably on datasets with linear feature relationships, this does not imply that PCA cannot produce insightful results on datasets with non linear structures. Even in cases where there is not a perfectly linear relationship between the characteristics, PCA can occasionally uncover helpful structures or offer good representation [28].

In addition, additional methods like manifold learning or non linear dimensionality reduction can be applied if the dataset has a highly non linear structure. As a result, the particulars of the dataset and the intended analysis's goal determine which approach is best. It is a way to apply PCA to datasets that are non linear.

The pre-processing algorithms chosen must take into account the demands of the job at hand as well as the properties of the data. Assumptions and requirements vary throughout machine learning algorithms, and appropriate pre-processing can enhance model performance. Comprehending the correlation between pre-processing and machine learning is crucial, as is carrying out trials to ascertain the optimal amalgamation for a given undertaking [29], [30].

Additionally, Kumala and Zahra advocated using other ways [7]. The study suggested using cross-corpus methods to identify Indonesian speech emotions. The study makes use of several datasets, including the Surrey audio-visual expressed emotion (SAVEE), the Ryerson Audio-Visual Database for emotion speech and song (RAVDESS), and the Berlin Database of Emotion Speech (Berlin EmoDB). As a result, there are three corpora: two English-speaking and one German-speaking. Teager Energy and MFCC techniques were applied during the feature extraction procedure. The study found that the accuracy performance may be improved by 2.09% in the Teager-MFCC combo and 4.16% in MFCC when using the Support Vector Machine (SVM) classifier. Furthermore, in terms of emotion recognition, the three corpora that were tested demonstrated high compatibility with the Indonesian language.

The goal of this inquiry is to better understand how to select a classifier kernel that matches the features of the dataset, how to engineer features, and how to integrate features and datasets to increase the accuracy of speech emotion data categorization [31].

In order to identify the emotion classes in human speech, the study will examine the following: the use of datasets, feature engineering, selecting the kernel classifier, merging datasets, merging features associated with data mining activities, and performing the process using several machine learning techniques, including Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), k-Nearest Neighbor (kNN),

Naïve Bayes (NB), Neural Network (NN), Stochastic Gradient Descent (SGD), and Deep Neural Networks (DNN).

3. METHODS

This section describes the Research Methodology used as shown in Fig. 1, which contains 4 main stages: Datasets collecting, pre-processing and feature extraction, Machine Learning Classification and Evaluation, which will be described in the next section.

The research proposes emotion identification based on Indonesian private speech dataset and public dataset RAVDESS by conducting experimental analysis of the use of kernel settings tailored to the characteristics of the datasets, use of PCA engineering features and experiments of merging several features and merging two public datasets from RAVDESS. The identification process uses eight types of Classifiers : Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), k-Nearest Neighbour (kNN), Naïve Bayes (NB), Neural Network (NN), Stochastic Gradient Descent (SGD) and Deep Neural network (DNN).

The dataset will be tested for its correlation with the selected kernel settings as well as the effects of the combination of many features on the speech emotion dataset. Experiments were also conducted by combining two corpus Ravdess datasets to see improved accuracy in the identification of speech emotion classes. Evaluation is performed using the Confusion Matrix to measure the performance of parameters such as accuracy, precision and recall.

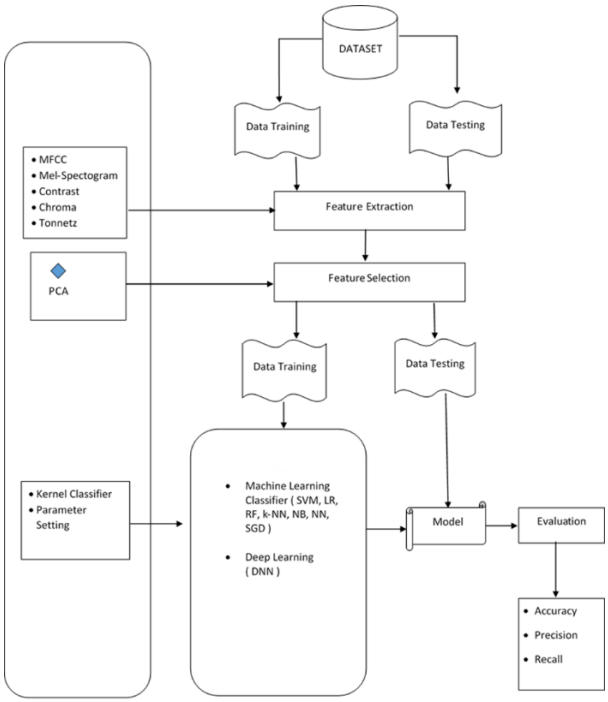

Fig. 1. Research Methodology

In Fig. 1, it can be seen that private and public datasets will be extracted audio features using several types of acoustic extraction methods namely Mel Frequency Cepstral Coefficients (MFCC), Chroma, Mel-Spectrogram, Tonnetz, and Contrast. On these acoustical features, PCA (Principal Component Analysis) techniques were applied to determine the impact of transformation and data reduction on the model produced.

After a pre-processing phase, the data can be grouped into training sets and testing sets using split validation. The training set will be used by Machine Learning as the model training reference data. After the training model is obtained, the next step is to test it using the testing set as the test data. From the testing using the test set, the prediction of the emotion type will be obtained by the trained Machine Learning model, where the results will be transformed into the confusion matrix table as the reference table defining the performance parameters of the proposed model. The parameters used to determine the performance of the proposed model are accuracy, precision, and recall.

3.1. Dataset

Private datasets are taken from a speech recording process consisting of two actors with male and female genders. A sample was taken with an actor pronouncing emotion words covering emotions of pleasure, sadness, disgust and anger. Each class of emotions is pronounced in as many as four classes of words, namely the words 'one', 'two', 'three' and 'four'. There are 640 data records produced from the process of recording this private dataset with Indonesian pronunciation [32]. The private dataset was then subjected to a data cleaning procedure that included noise reduction, file duration normalization (839 ms) with a sampling frequency of 44.1 kHz, and Channel Mono in order to prepare it for use.

The study made use of both public and private data sources, including the Ryerson Audio-Visual Database of Emotion Speech and Song, or RAVDESS [33]. The collection is a multi-modal database made up of distinct audio and video recordings that depict various emotional states. There are 2452 audio data in RAVDESS, of which 1440 are for voice and 1012 are for music. Eight categories of emotions are also applied to the data: fear, surprise, calm, happy, sad, angry, neutral, and disgust [34].

3.2. Feature Extraction

3.2.1. Mel Frequency Cepstral Coefficients (MFCC)

One kind of extraction feature that is frequently utilized in audio files is the Mel Frequency Cepstral Coefficients (MFCC) [35]. It is widely advised against using MFCC to recognize monosyllabic words on audio without identifying the speakers [36]. The pre-emphasis phase, which is the amplification of the audio signal at high frequencies, is when the process of extracting the MFCC feature from an audio file begins. The following stage involves converting windowing results into MFCC by using the Fast Fourier Transform, Mel Filter Bank, and Discrete Cosine Transform.

3.2.2. Chroma

Chroma is a feature that takes an audio with an emphasis on music and extracts a feature with a tone [37]. In the form of a basic feature, this feature can distribute fluctuations in the audio's tone level. A chromagram constructed from twelve (twelve) tone levels is the outcome of the Chroma feature [38].

3.2.3. Mel-Spectogram

In order to overcome the constraints of human ability to discriminate frequency values at a high level, the Mel-Spectrogram is an extraction of audio features [38]. The Mel-Specrogram is used in this study to extract frequency information, particularly for recognizing the types of emotions the performers are expressing.

3.2.4. Tonnetz

In addition to concentrating on the class of harmony and tone on an audio, Tonnetz is a derivative feature extraction of Chroma [39].

3.2.5. Contrast

In contrast Using the spectral values of the peak and valley of each sub-band, Contrast is one of the properties extracted from audio that may be used to determine the energy ratio of sound [40].

3.3. Features Engineering

After the feature extraction process is completed, the dataset is given the engineering feature process using PCA to reduce data that is correlated with each other. According to the function, this PCA will be used to reduce the amount of data or feature that correlates. Large correlations and data dimensions will reduce the performance of the classifier [27].

PCA or Principal Component Analysis is a statistical method that is widely used in data processing processes such as dimension reduction, data compression, and feature extraction [28]. PCA is conceptually capable of identifying new variables based on the principal component, where the value is linearly the result of a combination of the original features used. Simply put, the PCA will project a new feature or variable whose representation is the same as the original features in which the number of components can be adjusted. In this study, PCA was focused on dimension reduction to reduce the number of features in the extraction result as well as improve the representation of the value of the feature.

3.4. Machine Learning

To conduct the process of identifying groups of emotion classes, this study uses some Machine Learning as a comparison of the performance of the classification.

Support Vector Machine (SVM), is a superior method in Machine Learning because it has the advantage in terms of relatively short data training time and is simpler to use in the classification process for both linear and non linear datasets [41], [42] so SVM is one of the most efficient methods.

The Logistic Regression (LR) method is a statistical approach that is often used in statistical methods to predict probability by matching data. LR is a stable algorithm that has been used in various research and is one of the methods that is powerfully applied in Machine Learning [43].

Random Forest (RF) is one of the methods in Machine Learning used to solve classification and regression problems. RF capacity is resistant to overfitting, capable of handling non linear data and has the advantages of efficient processing [44].

The kNN algorithm (k-Nearest Neighbors) is an approach in the classification process that involves identifying objects whose location is closest. KNN can also provide solutions to classification and regression problems [45].

Naive Bayes (NB) is one of the classic methods of Machine Learning introduced by Thomas Bayes. NB has the advantages of being easier to implement, more time-efficient, requiring low computational processes and being able to handle big data [46].

A Neural Network (NN) is an algorithm used in various types of research, such as pattern recognition, speech recognition and various classification problems [47].

Stochastic Gradient Descent (SGD) is an optimization algorithm used to train models, especially in the context of machine learning and deep learning. At each iteration, SGD updates the model parameters by minimizing the gradient of the loss function against one data point randomly. (stokastik). SGD is sensitive to data scaling so it requires good data preprocessing and normalization [48].

Deep Neural Network (DNN) is used as one of the deep learning (DL) methods built on the basis of Neural Networks. DNN is an improvement to the conventional neural network method by adding some depth such as additional hidden layers on the input and output layers [49]. This method is generally used to predict or classify data according to its class. In this study, the DNN structure used consisted of one dense layer as an input and one dense layer as an output with each activation being 'ReLu' and 'Softmax'. Then there were 3 three hidden dense Layers.

3.5. Performance Evaluation

Several measuring instruments, including accuracy, precision, and recall, were used in this study to evaluate the classification techniques. The degree to which the classification model accurately predicts each class is known as accuracy, and it may be computed using Eq. (1). Eq. (2) can be used to determine precision, which is a metric used to assess how accurate the model's positive predictions are out of all the positive predictions made by the classification results. In the meantime, recall—which is determined by Eq. (3)—is helpful in assessing how well the procedure can distinguish between all genuine positive instances and all positive cases that have already occurred.

$$Accuracy = \frac{TN + TP}{FP + FN + TN + TP} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

In the above equation, it is known that TP (True Positive) is the sum of test data that is predicted to be true as a positive class, and TN (True Negative) is the sum of the test data that is predicted to be right as a negative class. These four values can be derived from the Confusion Matrix table.

4. RESULTS AND DISCUSSION

In this study, the experiment was carried out using Orange Tools and Python Programming. Orange tools are used on Private Data and Public Data because it is more practical to use many of the kernel Classifiers' Settings. Public data is used in Python programming to get variations in the use of Machine Learning, especially Deep Neural Network and for the flexibility of setting the Merge of Features and Datasets.

The research uses the Indonesian Private Speech Emotion and RAVDESS public datasets. From the data, it was then extracted using several feature extraction techniques such as MFCC, Mel-Spectrogram, Chroma,

Contrast, and Tonnetz. From the extraction results, a total of 193 features were obtained, consisting of MFCC 40 features, the Chroma produced 12 features, and the Mel-Spectrogram, the Contrast and the tonnetz each produced 128 features, 7 features, and 6 features.

From the features obtained, the correlation test process is performed using the Pearson correlations test. The correlation test results are shown in Table 1 for the Private dataset and Table 2 for the Public dataset. In the Private Dataset there are 805 features correlating with the correlation value 'r' between 1 > r >= 0.6 or 4.34% of the total of 18,528 correlated features. In the Public Dataset there are 654 features correlating with the correlation value 'r' between 1 > r >= 0.6 or 5.53% of the total of 18,528 correlated features.Data that has a correlation between features will reduce system performance because of the same value or weight, so one of the correlated features can be eliminated [27]. Through Engineering Features using PCA will generate new features and eliminate interrelated features.

Table 1. Private Dataset Feature Correlation

Private Data Before PCA		
Corr. Range	**Number Corr.**	**Percentage Corr.**
r >= 0.9	13	0.07
0.9 > r >= 0.8	93	0.5
0.8 > r >= 0.7	259	1.4
0.7 > r >= 0.6	440	2.37
r < 0.6	17723	95.66
Total	**18528**	**100**

Table 2. Public Dataset Feature Corellation

Public Data Before PCA		
Corr. Range	**Number Corr.**	**Percentage Corr.**
r >= 0.9	50	0.27
0.9 > r >= 0.8	152	0.82
0.8 > r >= 0.7	185	1
0.7 > r >= 0.6	267	1.44
r < 0.6	17874	96.47
Total	**18528**	**100**

The Private Data and Public Data feature correlation after the PCA value is low or close to zero for each feature. After performing the pre-processing phase by performing Engineering Features, the next process is the classification of data into groups of emotion classes using the Classifier or Machine Learning by conducting experiments using the selection of the Kernel Classifiers.

In Private Datasets, the test results are shown in Table 3 and Fig. 2. As seen in the Table and Figure, in general the linear setting results in a relatively higher or equal accuracy value when compared to the non linear setting, this indicates a more suitable linear kernel setting for use in this private dataset or, in other words, the private data has a linear property against the Classifier that will predict its class.

Table 3. Kernel Private Dataset

Machine Learning	Linear		Non Linear		Setting		Characteristic
	- PCA	+ PCA	- PCA	+ PCA	Linear	Non Linear	
SVM	**100.00**	96.80	97.70	96.90	Linear	Sigmoid	Dual
Logistic Regression	99.20	**99.60**	98.70	99.60	Ridge, Weak	Lasso, Half	Regresi
Random Forest	99.00	94.10	**99.20**	97.00	Shallow, 10 Tree	Deep, 30 Tree	Dual
kNN	99.20	**99.80**	99.20	99.80	Euclidean, By Distance	Euclidean, By Distance	Distance
Naïve Bayes	**91.70**	87.70	91.70	87.70	Default	Default	Probabilistic
NN	**100.00**	97.80	**100.00**	97.40	SLP	MLP	Dual

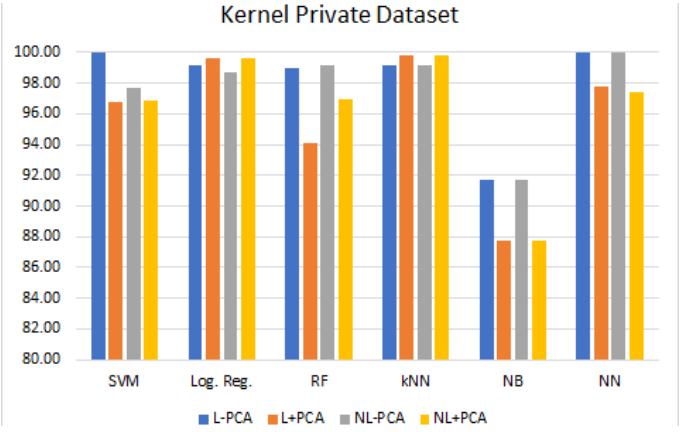

Fig. 2. Kernel Private Dataset

The linear definition of a dataset differs from the linear concept of a feature. Linear features are features that have a linear relationship,whereas a Linear Dataset is how a Classifier modeled class classifications on the dataset whether linear or non linear [8].

The highest accuracy values are seen on SVM of 100.00% for the linear setting without using PCA (L-PCA), and also on NN of 100.00% for the Linear without PCA setting (L - PCA) and the Non Linear setting without PCA (NL-PCA).

On LR, the highest value is 99.60% for setting Linear without PCA (L-PCA) and Non Linear using PCA (NL+PCA). This shows an adaptive LR algorithm for linear and non linear classification [43].

In Random Forest, there was a 99.20% increase in non linear PCA (NL-PCA) kernel settings. Random forest worked well on both linear and non linear classification processes [44].

kNN is relatively constant, with the highest accuracy value of 99.80% for the linear setting (L+PCA) and the non linear setup (NL+PCA) [45].

For NB, the relative remains the same as its accuracy, both for linear and non linear settings, as is the case with LR and kNN.

NN has the highest accuracy values in the linear (L-PCA) and non linear (NL+PCA) settings. This corresponds to the characteristics of NN that are adaptive to the classification of classes as linear and non linear [47]. The PCA function in Private Dataset does not always increase accurability in the settings of the Linear or Non Linear Kernel, it corresponds with the PCA characteristics that create new features and reduce data dimensions, so there is a possibility that the original feature is missing and thus decreases the precision [28].

It can be said that the highest accuracy value is grouped on the linear kernel setting of this Private Dataset. It can also be said that the private dataset is linear to the Classifier or matches using the Linear setting on the Classifier.

In the setting of the Kernel Classifier for Public Data, as shown in Table 4 and Fig. 3, it appears that the non linear setting produces the highest accuracy value compared to the linear setting. So it can be said that this public data set has non linear properties to its classification of the Classifiers [50].

However, kNN's accuracy increases with linear kernel settings, which is still possible because k-NL works well on linear and non linear classification processes [45].

See the highest accuracy value produced by NN with a value of 74.80% and setting non linear kernels without PCA (NL-PCA). NN also works well on both linear and non linear classifications [47].

Table 4. Kernel Public Dataset

Machine Learning	Linear		Non Linear		Setting		Characteristic
	- PCA	+ PCA	- PCA	+ PCA	Linear	Non Linear	
SVM	60.00	60.50	61.40	**63.00**	Linear	RBF	Dual
Logistic Regression	60.30	62.50	**63.00**	62.50	Lasso, Half	Lasso, Half	Dual
kNN	**64.20**	63.00	63.00	63.00	Manhattan, By Distance	Euclidean, By Distance	Distance
NN	73.00	64.30	**74.80**	67.30	SLP	MLP	Dual
SGD	60.10	61.60	**62.20**	61.60	Default	Hinge, 0,001, Ridge	Gradient

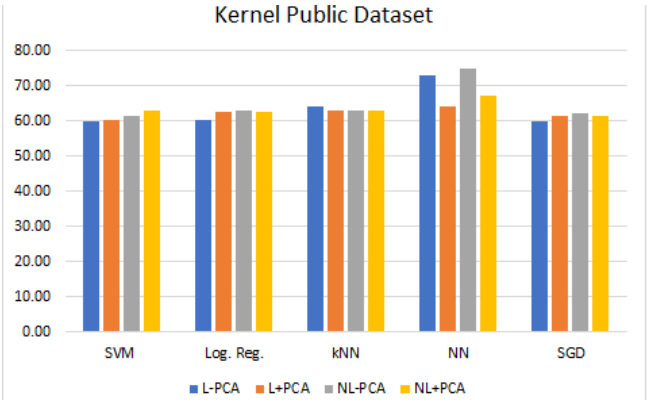

Fig. 3. Kernel Public Dataset

PCA functions on public datasets do not necessarily increase the accuracy of the linear or non linear kernel settings, they correspond to the PCA characteristics that create new features and reduce the size of the data, so there is a possibility that the original features are missing and thus decrease the accuracy [28].

To improve accuracy, in addition to using the appropriate kernel settings, as described earlier, another effort could be made to combine features as seen in the experimental results described in Table 5 and Fig. 4.

Table 5. Combining Features Speech Dataset

Feature	Speech	
	- PCA	+ PCA
MFCC	60.29	53.57
Mel+Chroma	47.05	40.33
Mel+MFCC	60.08	64.07
5 Feature	62.81	**63.23**

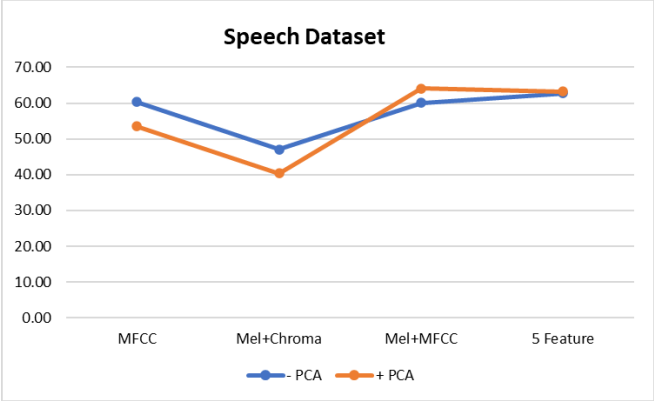

Fig. 4. Combining Features Speech Dataset

On Single Speech Datasets, Table 5 and Fig. 4, you can see a combination of features using 1 feature, 2 features (two times) and 5 features. Feature combination does not always increase accuracy, such as when Mel-Spektofgram features are combined with MFCC features resulting in a decrease in value compared to using the MFCC feature alone. MFCC precision alone is 60.29% and MFCC+Mel-Spectogram combined feature values are 60.08% with processes without PCA. While in processes with PCA there is a decline, from MFCC alone of 53.57% to 40.33% when using a Mel-Spectogram in combination with Chroma. However, in this single Speech Data, using a variety of combined features, there is generally an increase with the highest precision value of 63.23% for combining five features and the PCA process.

The PCA function on a Public Dataset merger does not necessarily increase the accuracy of the data, it is in line with the PCA characteristics that create new features and reduce the dimension of the data, so there is a possibility that the original feature is missing and thus decreases the precision [28].

In combined datasets that combine a Speech Dataset combined with a Song Dataset or a combination of two Datasets to enhance accuracy, as shown in Table 6 and Fig. 5, combined feature combination does not always increase accuration, such as when the Mel-Spectogram feature is combined with the MFCC feature, resulting in a decrease in value compared to using only the MFCC function. The MFCC accurate value alone is 69.01 and the combined MFCC+Mel-Spectrum feature value is 68.64% with a PCA-free process. While in processes with PCA, there is a decline, from MFCC alone 68.39% to 52.71% when using the Mel Spectogram in combination with Chroma. However, in this combined dataset using a variety of combined features, it generally experiences an increase with the highest accurate value of 73.95% for the integration of 5 features with the PCA process.

Table 6. Combining Features Speech and Song Dataset

Feature	Speech and Song	
	- PCA	+ PCA
MFCC	69.01	68.39
Mel+Chroma	58.02	52.71
Mel+MFCC	68.64	72.83
5 Feature	69.25	**73.95**

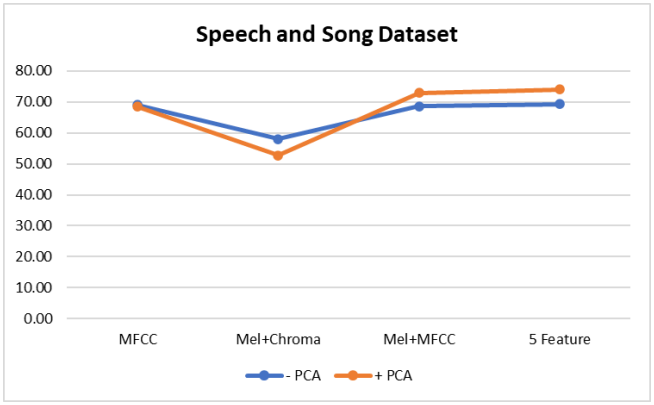

Fig. 5. Combining Features Speech and Song Dataset

Overall, the performance evaluation of the public datasets using the proposed DNN model was able to produce the highest accuracy of 73.95% followed by a precision of 72.36% and a recall value of 72.07%. The results also surpassed the results of previous studies, where the study also used the RAVDESS dataset as shown in Table 7.

Table 7. Comparison of Performance Results with Previous Study

No.	Work	Dataset	Feature	Classifier	Accuracy Result
1	Chowdary and Hemanth [16]	RAVDESS	Mel Frequency Cepstral Coefficients (MFCC)	CNN	71.35%
2	D. Issa et. al [46]	RAVDESS	Mel Frequency Cepstral Coefficients (MFCC), Chroma, Mel Spectogram, Constrast, Tonnetz	DCNN	71.61%
3	Damodar et. al [47]	RAVDESS	Mel Frequency Cepstral Coefficients (MFCC)	CNN	72.00%
4	**Proposed Method**	**RAVDESS**	**Mel Frequency Cepstral Coefficients (MFCC), Chroma, Mel Spectogram, Constrast, Tonnetz**	**Principal Component Analysis (PCA) Deep Neural Network (DNN)**	**73.95%**

5. CONCLUSION

Classification performance is determined by or associated with the selection of the corresponding kernel classifier settings, also related to the dataset's characteristics. Engineering features and dataset embedding contribute to increasing accuracy values as well.

From the experimental results, on the Private Dataset, the selection of the appropriate kernel will result in the highest accuracy value of 100% on the SVM Classifier with a linear kernel setting without PCA. The highest value in the Private datas*et al*so exists on the NN Classifiers with a 100.00% accurate value using the kernel without the PCA setting. On the Public Dataset, the matching kernel selection will yield a maximum accuracy value of 74.80% for the NN classifier with the Non Linear Kernel Setting without PCA. There is an improvement of 12.30% in accuracy for the Private Data, calculated from its highest accuracy value (100.00%) minus its lowest value (87.70%). For the Public Data, there is an increase of 14.80% in accuracy, computed from its top accuracy (74.80%) minus the lowest value (60.00%). The combination of features and datasets on the Public Dataset will also increase the accuracy value by 33.62%, from the higher accuracy obtained from 73.95% minus the 40.33%. For further research, we can explore tests with sound features taken from spectroscopic images. We can also combine text datasets to combine features.

REFERENCES

[1] T. Puri, M. Soni, G. Dhiman, O. Ibrahim Khalaf, M. alazzam, and I. Raza Khan, "Detection of Emotion of Speech for RAVDESS Audio Using Hybrid Convolution Neural Network," *J. Healthc. Eng.*, vol. 1, p. 8472947, 2022, https://doi.org/10.1155/2022/8472947.

[2] N. Ahmed, Z. Al Aghbari, and and S. Girija, "A systematic survey on multimodal emotion recognition using learning algorithms," *Intell. Syst. with Appl.*, vol. 17, p. 200171, 2023, https://doi.org/10.1016/j.iswa.2022.200171.

[3] M. Egger, M. Ley, and S. Hanke, "Emotion Recognition from Physiological Signal Analysis: A Review," *Electron. Notes Theor. Comput. Sci.*, vol. 343, pp. 35–55, 2019, https://doi.org/10.1016/j.entcs.2019.04.009.

[4] A. Rizal and I. Istiqomah, "Lung Sounds Classification Based on Time Domain Features," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 8, no. 2, pp. 318–325, 2022, http://dx.doi.org/10.26555/jiteki.v8i2.24007.

[5] R. L. Soash, "Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places," *Collect. Manag.*, vol. 24, no. 3–4, pp. 310–311, 1999, https://doi.org/10.1300/j105v24n03_14.

[6] A. M. Badshah *et al.*, "Deep features-based speech emotion recognition for smart affective services," *Multimed. Tools Appl.*, vol. 78, no. 5, pp. 5571–5589, 2019, https://doi.org/10.1007/s11042-017-5292-7.

[7] O. U. Kumala and A. Zahra, "Indonesian Speech Emotion Recognition using Cross-Corpus Method with the Combination of MFCC and Teager Energy Features," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 4, pp. 163–168, 2021, https://doi.org/10.14569/IJACSA.2021.0120422.

[8] Y. Deldjoo, T. Di Noia, E. Di Sciascio, and F. A. Merra, "How Dataset Characteristics Affect the Robustness of Collaborative Recommendation Models," *SIGIR 2020 - Proc. 43rd Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 951–960, 2020, https://doi.org/10.1145/3397271.3401046.

[9] Q. Zhang, E. C. C. Tsang, Q. He, and Y. Guo, "Ensemble of kernel extreme learning machine based elimination optimization for multi-label classification," *Knowledge-Based Syst.*, vol. 278, p. 110817, Oct. 2023, https://doi.org/10.1016/j.knosys.2023.110817.

[10] O. B. Victoriano and A. C. Fajardo, "Multi-label Learning Linearity in Ensemble of Pruned Set," *ACM Int. Conf. Proceeding Ser.*, pp. 17–21, 2019, https://doi.org/10.1145/3394788.3394922.

[11] J. Olufemi Ogunleye, "The Concept of Data Mining," *IntechOpen*, pp. 1–20, Mar. 2022, https://doi.org/10.5772/intechopen.99417.

[12] A. R. S. Parmezan, H. D. Lee, N. Spolaôr, and F. C. Wu, "Automatic recommendation of feature selection algorithms based on dataset characteristics," *Expert Syst. Appl.*, vol. 185, 2021, https://doi.org/10.1016/j.eswa.2021.115589.

[13] Z. Fengyu, Zhao., Liqun, Gao., Zhouran, "The application of machine learning regression algorithms and feature engineering in practical application.," *10th International Conference on Information Systems and Computing Technology (ISCTech)*, pp. 259–263, 2022, https://doi.org/10.1109/ISCTech58360.2022.00048.

[14] M. Xu, F. Zhang, and W. Zhang, "Head Fusion: Improving the Accuracy and Robustness of Speech Emotion Recognition on the IEMOCAP and RAVDESS Dataset," *IEEE Access*, vol. 9, pp. 74539–74549, 2021, https://doi.org/10.1109/ACCESS.2021.3067460.

[15] D. Mamieva, A. B. Abdusalomov, A. Kutlimuratov, B. Muminov, and T. K. Whangbo, "Multimodal Emotion Detection via Attention-Based Fusion of Extracted Facial and Speech Features," *Sensors*, vol. 23, no. 12, 2023, https://doi.org/10.3390/s23125475.

[16] L. Muflikhah, F. A. Bachtiar, D. E. Ratnawati, and R. Darmawan, "Improving Performance for Diabetic Nephropathy Detection Using Adaptive Synthetic Sampling Data in Ensemble Method of Machine Learning Algorithms," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 10, no. 1, p. 123, 2024, https://doi.org/10.26555/jiteki.v10i1.28107.

[17] V. Kumar and M. L., "Deep Learning as a Frontier of Machine Learning: A Review," *Int. J. Comput. Appl.*, vol. 182, no. 1, pp. 22–30, 2018, https://doi.org/10.5120/ijca2018917433.

[18] L. Sun, J. Chen, K. Xie, and T. Gu, "Deep and shallow features fusion based on deep convolutional neural network for speech emotion recognition," *Int. J. Speech Technol.*, vol. 21, no. 4, pp. 931–940, 2018, https://doi.org/10.1007/s10772-018-9551-4.

[19] Y. B. Singh and S. Goel, "A systematic literature review of speech emotion recognition approaches," *Neurocomputing*, vol. 492, pp. 245–263, Jul. 2022, https://doi.org/10.1016/j.neucom.2022.04.028.

[20] M. Kalpana Chowdary and D. Jude Hemanth, "Deep Learning Approach for Speech Emotion Recognition," in *Lecture Notes on Data Engineering and Communications Technologies*, pp. 367–376, 2021, https://doi.org/10.1007/978-981-15-8335-3_29.

[21] M. Iqbal, S. A. Raza, M. Abid, F. Majeed, and A. A. Hussain, "Artificial Neural Network based Emotion Classification and Recognition from Speech," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 12, pp. 434–444, 2020, https://doi.org/10.14569/IJACSA.2020.0111253.

[22] B. Pragati, C. Kolli, D. Jain, A. V. Sunethra, and N. Nagarathna, "Evaluation of Customer Care Executives Using Speech Emotion Recognition," in *Machine Learning, Image Processing, Network Security and Data Sciences*, pp. 187–198, 2023, https://doi.org/10.1007/978-981-19-5868-7_14.

[23] S. Jothimani and K. Premalatha, "MFF-SAug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network," *Chaos, Solitons & Fractals*, vol. 162, p. 112512, Sep. 2022, https://doi.org/10.1016/j.chaos.2022.112512.

[24] S. Patra, S. Datta, and M. Roy, "Analysis on Speech-Emotion Recognition with Effective Feature Combination," in *OITS International Conference on Information Technology (OCIT)*, IEEE, Dec. pp. 1–5, 2022, https://doi.org/10.1109/OCIT56763.2022.00018.

[25] A. Baird, "Extending Multimodal Emotion Recognition with Biological Signals," in *Proceedings of the 1st*

International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop, pp. 7–7, Oct. 2020, https://doi.org/10.1145/3423327.3423512.
[26] D. Oreski, S. Oreski, and B. Klicek, "Effects of dataset characteristics on the performance of feature selection techniques," *Appl. Soft Comput. J.*, vol. 52, pp. 109–119, 2017, https://doi.org/10.1016/j.asoc.2016.12.023.
[27] T. Kurita, "Principal component analysis (PCA)," *Comput. Vis. A Ref. Guid.*, pp. 1–4, 2019, https://doi.org/10.48550/arXiv.1503.06462.
[28] M. Ringnér, "What is principal component analysis?," *Nat. Biotechnol.*, vol. 26, no. 3, pp. 303–304, 2008, https://www.nature.com/articles/nbt0308-303.
[29] F. Mustofa, A. N. Safriandono, A. R. Muslikh, and D. R. I. M. Setiadi, "Dataset and Feature Analysis for Diabetes Mellitus Classification using Random Forest," *J. Comput. Theor. Appl.*, vol. 1, no. 1, pp. 41–48, 2023, https://doi.org/10.33633/jcta.v1i1.9190.
[30] M. Feurer, A. Klein, K. Eggensperger, J. T. Springenberg, M. Blum, and F. Hutter, "Auto-sklearn: Efficient and Robust Automated Machine Learning," *Advances in neural information processing systems*, pp. 113–134, 2019, https://doi.org/10.1007/978-3-030-05318-5_6.
[31] V. M. Praseetha and S. Vadivel, "Deep learning models for speech emotion recognition," *J. Comput. Sci.*, vol. 14, no. 11, pp. 1577–1587, Nov. 2018, https://doi.org/10.3844/jcssp.2018.1577.1587.
[32] J. Gondohanindijo *et al.*, "Comparison Method in Indonesian Emotion Speech Classification," *Proc. - 2019 Int. Semin. Appl. Technol. Inf. Commun. Ind. 4.0 Retrosp. Prospect. Challenges, iSemantic 2019*, pp. 230–235, 2019, https://doi.org/10.1109/ISEMANTIC.2019.8884298.
[33] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS One*, vol. 13, no. 5, p. e0196391, May 2018, https://doi.org/10.1371/journal.pone.0196391.
[34] J. Gondohanindijo, Muljono, E. Noersasongko, Pujiono, and D. R. M. Setiadi, "Multi-Features Audio Extraction for Speech Emotion Recognition Based on Deep Learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 6, pp. 198–206, 2023, https://doi.org/10.14569/IJACSA.2023.0140623.
[35] R. M. Hanifa, K. Isa, and M. Mohamad, "Comparative Analysis on Different Cepstral Features for Speaker Identification Recognition," *IEEE Student Conf. Res. Dev. SCOReD 2020*, no. September, pp. 487–492, 2020, https://doi.org/10.1109/SCOReD50371.2020.9250938.
[36] S. Ajibola Alim and N. Khair Alang Rashid, "Some Commonly Used Speech Feature Extraction Algorithms," *From Nat. to Artif. Intell. - Algorithms Appl.*, pp. 2-19, Dec. 2018, https://doi.org/10.5772/intechopen.80419.
[37] J. V. T. Abraham, A. N. Khan, and A. Shahina, "A deep learning approach for robust speaker identification using chroma energy normalized statistics and mel frequency cepstral coefficients," *Int. J. Speech Technol.*, no. 0123456789, 2021, https://doi.org/10.1007/s10772-021-09888-y.
[38] U. Garg, S. Agarwal, S. Gupta, R. Dutt, and D. Singh, "Prediction of Emotions from the Audio Speech Signals using MFCC, MEL and Chroma," *Proc. - 2020 12th Int. Conf. Comput. Intell. Commun. Networks, CICN*, pp. 87–91, 2020, https://doi.org/10.1109/CICN49253.2020.9242635.
[39] S. Sen, A. Dutta, and N. Dey, "Speech Processing and Recognition System," *Audio Process. Speech Recognit.*, pp. 13–43, 2019, https://doi.org/10.1007/978-981-13-6098-5_2.
[40] S. Bhattacharya, S. Borah, B. K. Mishra, and A. Mondal, "Emotion detection from multilingual audio using deep analysis," *Multimed. Tools Appl.*, vol. 81, no. 28, pp. 41309–41338, 2022, https://doi.org/10.1007/s11042-022-12411-3.
[41] Nurhanna, "Multi-class Support Vector Machine Application in the Field of Agriculture and Poultry : A Review," *Malaysian Journal of Mathematical Sciences,* vol. 11, pp. 35-52. no. 11, pp. 35–52, 2017, https://einspem.upm.edu.my/journal/fullpaper/vol11sfeb/3.%20Nurhanna.pdf.
[42] F. S. Gomiasti, W. Warto, E. Kartikadarma, J. Gondohanindijo, and D. R. I. M. Setiadi, "Enhancing Lung Cancer Classification Effectiveness Through Hyperparameter-Tuned Support Vector Machine," *J. Comput. Theor. Appl.*, vol. 2, no. 2, pp. 179–189, 2024, https://doi.org/10.62411/jcta.10106.
[43] H. Rianto and R. S. Wahono, "Resampling Logistic Regression untuk Penanganan Ketidakseimbangan ZClass pada Prediksi Cacat Software," *Journal of Software Engineering*, vol. 1,no. 1, pp. 46–53, 2015, https://www.romisatriawahono.net/lecture/rm/paper/Rianto%20-%20Resampling%20Logistic%20Regression%20untuk%20SDP%20-%202015.pdf.
[44] A. Sarica, A. Cerasa, and A. Quattrone, "Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease : A Systemic Review," *Frontiers in aging neuroscience*, vol. 9, pp. 329, 2017, https://doi.org/10.3389/fnagi.2017.00329.
[45] D. Jadhav, S. and H. P. Channe, "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification TEchniques," *International Journal of Science and Research (IJSR)*, vol. 5, no. 1, pp. 1842-1845, 2016, https://www.ijsr.net/archive/v5i1/NOV153131.pdf.
[46] A. P. Wibawa *et al.*, "Naive Bayes Classifier for Journal Quartile Classification,"*IJES,* vol. 7, no. 2, pp. 91–99, 2019, https://doi.org/10.3991/ijes.v7i2.10659.
[47] H. Ramchoun, M. Amine, J. Idrissi, Y. Ghanou, and M. Ettaouil, "Multilayer Perceptron : Architecture Optimization and Training," *Multilayer Perceptron : Architecture Optimization and Training*, vol. 4, no. 1, pp. 26-31, 2016, https://doi.org/10.9781/ijimai.2016.415.

[48] V. L. Abeykoon, G. C. Fox, and M. Kim, "Performance optimization on model synchronization in parallel stochastic gradient descent based SVM," *Proc. - 19th IEEE/ACM Int. Symp. Clust. Cloud Grid Comput. CCGrid 2019*, pp. 508–517, 2019, https://doi.org/10.1109/CCGRID.2019.00065.
[49] J.-T. Chien, "Deep Neural Network," *Source Sep. Mach. Learn.*, pp. 259–320, 2019, https://doi.org/10.1016/B978-0-12-804566-4.00019-X.
[50] Z. Han, "Speech Emotion Recognition Based on Deep Learning and Kernel Nonlinear PSVM," *Proceedings of the 31st Chinese Control and Decision Conference, CCDC,* pp. 1426–1430, 2019. https://doi.org/10.1109/CCDC.2019.8832414.

BIOGRAPHY OF AUTHORS

Jutono Gondohanindijo is a lecturer of the S1 Study Program TI (Technic Informatics) at the Faculty of Technics and Informatics, AKI University Semarang, Indonesia. He graduated with a Bachelor of Computer (Local) from Budi Luhur, Jakarta, in 1984. He got the M.Kom degree (Master of Computer) from STTIBI Jakarta, Indonesia, in 2001. He is currently taking Doctoral Program S3 Computer Science at Dian Nuswartoro University Semarang and has done special data mining research on Speech Recognition. Email: jutono.gondohanindijo@unaki.ac.id.

Edi Noersasongko is the Rector of the Dian Nuswantoro University Semarang, whereas his last academic position is Professor. History of education: graduated Bachelor of Computer (Local) at Informatic and Computer College, Jakarta, 1983; graduated Bachelor of Computers (State) at Informatic and Computer College, Jakarta, 1993; graduated Master of Computer at Technology Information Benarif Indonesia College, Jakarta, 1995; graduated Doctor of Economic (S3) Merdeka University Malang, 2005; and an Honorary Doctorate in Educational Information Technology, Teknikal Malaysia Melaka University, 2012. Another outstanding success is qualified entrepreneurship. Email: edi.noer@research.dinus.ac.id.

Pujiono is an Assistant Professor in the Faculty of Computer Science at Dian Nuswantoro University, Semarang, Indonesia. He received his Doctoral degree from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, in 2018. He joined the Mathematics tent at PRIMAGAMA, Semarang, in 2000. He received his M.Kom (Informatics) from STTIBI Jakarta, Indonesia, in 2001, and he has published research papers in reputed international journals and conferences. His current research interests include Modeling Under Water Image Processing, Computation and Mathematics. Email: pujiono@dsn.dinus.ac.id.

Muljono holds a Doctor of Electrical Engineering degree from the Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, in 2016. He received his M.Kom (Informatics) from STTIBI Jakarta, Indonesia, in 2001 and his B.Sc. (Mathematics) from Universitas Diponegoro (UNDIP) in 1996. He is currently an associate professor at the Informatics Engineering Department at Dian Nuswantoro University, Semarang, Indonesia. His research includes artificial intelligence, machine learning, data mining, data science and natural language processing. He has published over 90 papers in international journals and conferences. He can be contacted at: muljono@dsn.dinus.ac.id.