ANALYSIS OF VARIATION OF FEATURE EXTRACTION METHODS IN THE CLASSIFICATION OF AL-QUR'AN MAQAM USING MACHINE LEARNING

MUHAMMAD AHMAD AGIL ALAYDRUS[1],AMALIA ZAHRA[2]

[1]Computer Science Department, BINUS Graduate Program – Master of Computer Science

[2] Computer Science Department, School of Computer Science Bina Nusantara University

E-mail: [1]muhammad006@binus.ac.id, [2]amalia.zahra@binus.edu

ABSTRACT

As the book of religion with adherents of more than 1.9 billion people in the world and 86.7% of Indonesia's population, the Qur'an has earned the title of being the 'book' that is most widely read in the world. Al-Qur'an is the holy book that guides Muslims in life and religion. The Koran is read every day by a Muslim, for example during prayer services. Until now, the process of learning and memorizing the Qur'an has never stopped; the development of applied technology to support the learning process is also intensively carried out. Until now, studies that have focused on reciting the Qur'an have not made maqām the main focus. In contrast to recitation, discussions and learning applications can be found on various platforms in various forms. Tajweed does indeed play an important role as a basis for reciting the Koran, but limited information and support systems in maqām learning often prevent someone from learning it. This is the background of this research to move in presenting a system that can assist the maqām learning process. This classification system was built with the hope that maqāmat learners can detect the types of maqām independently, both from personal readings and voice recordings of other people. Thus, users can verify whether the chanting tone of the Al-Qur'an recitation is appropriate, or find out the maqām from
someone else's reading recording.

Keywords: *MFCC, SDCC, Maqam, and CNN*

1. INTRODUCTION

As the book of religion with adherents of more than 1.9 billion people in the world and 86.7% of Indonesia's population, the Qur'an has earned the title of being the 'book' that is most widely read in the world [1][2]. Al-Qur'an is the holy book that guides Muslims in life and religion. The Koran is read every day by a Muslim, for example during prayer services. Until now, the process of learning and memorizing the Qur'an has never stopped; the development of applied technology to support the learning process is also intensively carried out
[3][4][5].

Throughout the world, it is common to read and memorize the Al-Qur'an using the original language of writing, Arabic, following the rules of recitation known as tajwid [6]. In practice itself, the Qur'an is often recited using unique melodic tones called maqām (in the plural it is called maqāmat, from Arabic). Maqām was born from the art culture of the Arabian peninsula which has its characteristics [7][8]. There are 8 types of maqām: Bayat, Saba, Hijaz, Nahawand, Rast, Seka, Ajam, and Kurd, each of which has different patterns, tones, and intervals. Each maqāmat chant can evoke different emotions and feelings so many Muslims use it to beautify readings and help the process of memorizing the Qur'an [9][10].

Until now, studies that have focused on reciting the Qur'an have not made maqām the main focus. In contrast to recitation, discussions, and learning applications can be found on various platforms in various forms [11][12][13]. Tajweed does indeed play an important role as a basis for reciting the Koran, but limited information and support systems in maqām learning often prevent someone from learning it. This is the background of this research to move in presenting a system that can assist the maqām learning process.

Accommodating this need and assisting the learning process, this research raises the issue of creating a "Maqām Classification System in Automatic Al-Qur'an Recitation Using Machine Learning". Research developments in the field of speaker recognition, speaker verification, and language classification show that a combination of feature extraction such as the Mel-frequency Cepstral Coefficient (MFCC) and its derivatives combined with model building using machine learning (classifier algorithms, neural networks, deep learning) can produce results that are highly satisfactory' in its model predictions [14][15][16].

In the research [17], the feature extraction used was MFCC, and the deep learning models tested were convolutional neural network (CNN), long-short term memory (LSTM), and a combination of CNN+LSTM. The three models were trained with a learning rate of 0.0001 and a batch size of 64, and each model was trained with a different epoch (iteration). CNN with 35 epochs, LSTM with 50 epochs, and CNN+LSTM with 37 epochs. After testing the results show that the accuracy with 5-Fold cross-validation on the CNN model is 74.6%, the LSTM model is 77.5% and CNN+LSTM is 75.8%. The test accuracy of each model is 72.1%, 79.3%, and 76.4% resulting in an F1-score for each model of 0.76, 0.78, and 0.79.

However, this research only relies on one feature extraction method, namely MFCC. Although MFCC is known to have better performance than other feature extraction methods (PLP, LPC, DWT) and is suitable to be combined with machine learning modeling, system performance diagnosis also needs to be done with a variety of feature extraction methods before entering the model training stage. MFCC itself can still be processed into more specific features. First, the MFCC components can be processed further using the first and second derivative factors (delta and double-delta MFCC).

After that extraction, MFCC is recalculated using SDC to become SDCC. SDCC is expected to be able to capture wider speech dynamics factors in the language recognition system and capture dialect diversification [18]. Each maqām has its characteristics in the choice of melody, pitch intervals, and emphasis on intonation. This can significantly affect the features of each class, thus affecting the performance of the model and system created as a whole.

This research will focus on variations in feature extraction methods and their impact on the system performance. State-of-the-art selected feature extraction methods: MFCC and SDCC (shifted delta cepstral coefficients). The modeling method chosen is CNN, using a system architecture that is equated with is [17] as a reference. The metrics used to evaluate the system are cost average (Cavg, loss), model accuracy (acc), and F1 score [16][19]. The results will be directly compared against the main referral as a complementary result.

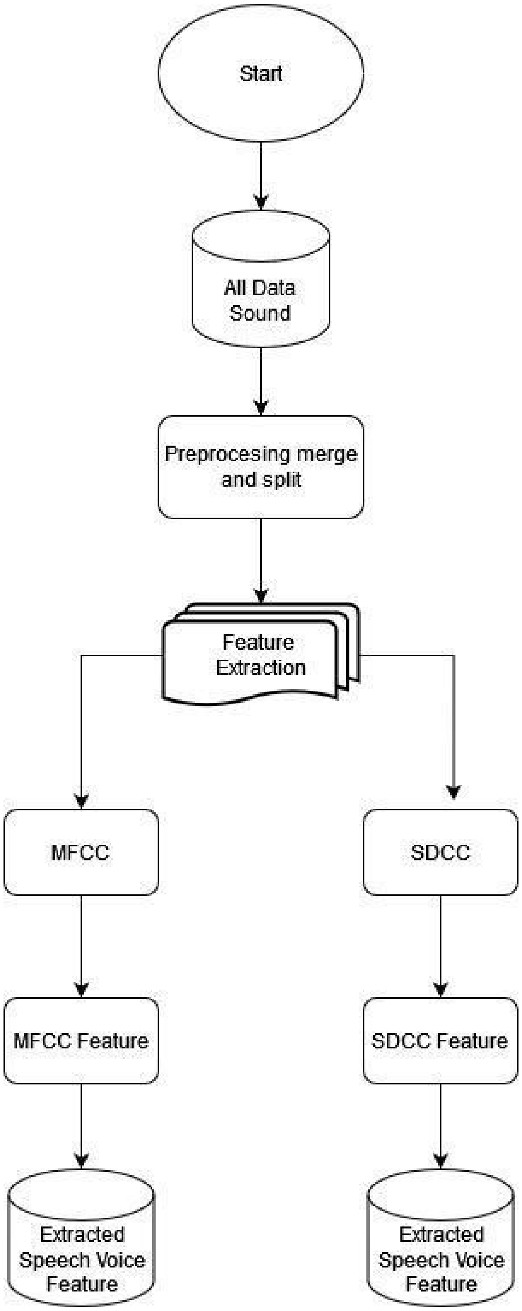
In building a computing system, it is necessary to pay attention to the tradeoff between system performance and the complexity of the system being built. The addition of processing steps, iterations, and more complex input data can increase processing resource requirements. In this study, the total processing time and computational resources required for each alternative solution will be taken into account, to create effective and efficient applicable solutions.

Often maqāmat teaching is limited to being carried out by a teacher orally, without any other methods for checking readings or supporting facilities. This classification system was built with the hope that maqāmat learners can detect the types of maqām independently, both from personal recitals and voice recordings of other people. Thus, users can verify whether the chanting tone of the Al-Qur'an recitation is appropriate, or find out the maqām from someone else's reading recording.

This paper will focus on building a similar maqam classification system in [17] using MFCC and comparing its performance with one using SDCC features extraction. This comparison will tell us whether SDCC features extraction is more suitable than MFCC in maqam classification. This will also tell us if the relationship of several consecutive audio files is promising as a parameter for maqam classification. This paper will not be focusing in

MFCC and SDCC parameters and there will be no tuning for optimal performance from each method.

Figure 1: Diagram Of Research Flow Design Preparation Is Carried Out. In This Study, Feature Extraction


Start
All Data Sound
Preprocesing merge and split
Feature Extraction
MFCC
SDCC
MFCC Feature
SDCC Feature
Extracted Speech Voice Feature
Extracted Speech Voice Feature

itself stage and feature extraction

2. RESEARCH METHOD

This research is divided into 4 stages: database preparation, feature extraction and processing, model training, and result comparison. Figures 1 and 2 show the design flow of this research work.

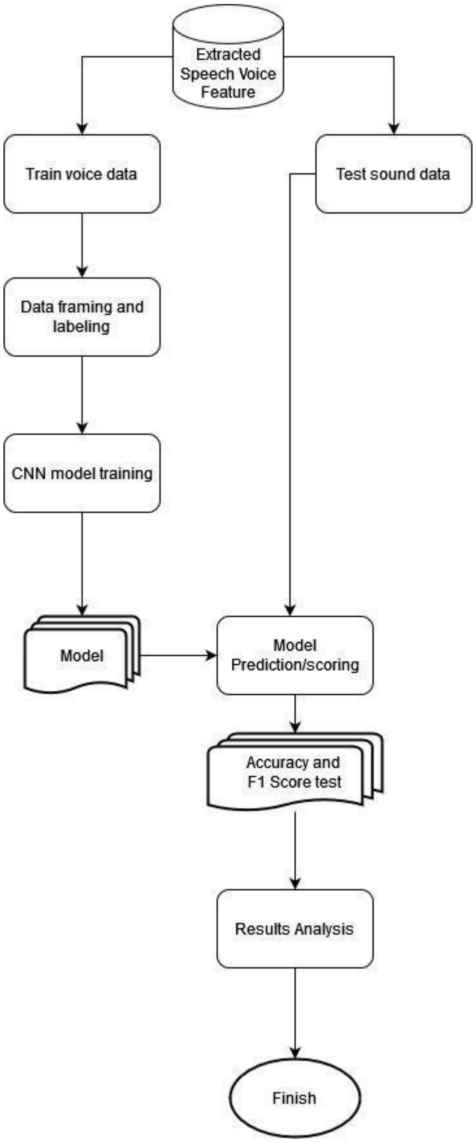


Figure 2: Modeling and evaluation

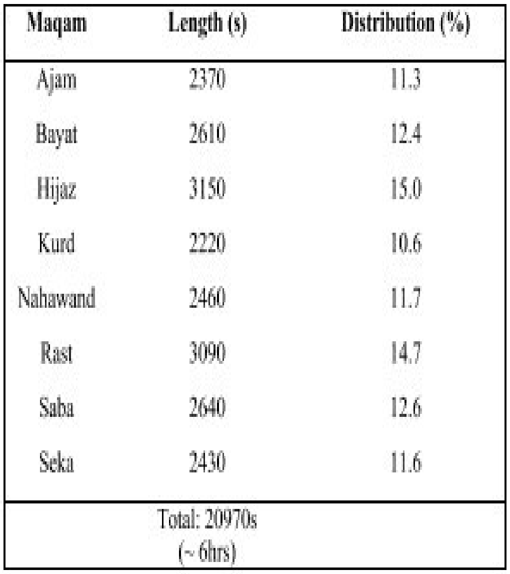
First, all audio files that has been prepared is going through VAD to clean the files from background noises. If the data used is fairly clean and has little noise, the VAD process can be ignored. After the VAD process, feature extraction

is divided into two, namely by using MFCC and SDCC. Furthermore, from the extraction results, it will produce extracted speech features, then the extracted speech features are split into 2 data, namely testing data and training data. In the voice training data, the results of the extracted speech features are framing and labeling, then are feed into a CNN model. This training model will then be validated using 5-fold CV and then tested using the split testing data. From validation and testing, we will be obtaining the F1 scores and also accuracy from each model to see its effectiveness in maqam classification.

2.1 Database

The database for recording maqāmat readings of the Qur'an itself was obtained from the research of Shahriar and Tariq [17], named Maqam478 data. Composed of 478 audio files, each 30 to 60 seconds long. Each class has 50 to 70 different reading variations of the audio files. The entire reading of the Koran was read by 2 people: Sheikh Bandar Baleela (Imam Masjid al-Haram) and Alm. Sheikh Muhammad Ayyub (Imam of the Nabawi Mosque). Each file will be labeled according to the type of maqām used.

All files are in .wav (lossless audio) format. The total duration and distribution of the amount of data from each class are summarized in Figure 3. The uncompressed file of all the data used has a total size of 3.78 GB. During the classification training process, 80% of the data will be taken as training data and 20% as test data. The training data also includes validator data in it, a small part (~ 10 – 20%) that is sampled serves as a reference for validating the model training conditions.



Maqam	Length (s)	Distribution (%)
Ajam	2370	11.3
Bayat	2610	12.4
Hijaz	3150	15.0
Kurd	2220	10.6
Nahawand	2460	11.7
Rast	3090	14.7
Saba	2640	12.6
Seka	2430	11.6
	Total: 20970s (~ 6hrs)	

Figure 3: Total duration of each maqām class and data distribution.

2.2 Feature Extraction

At this stage, the basis of acoustic features is extracted using the MFCC and SDCC methods. Before extracting, the data is conditioned first by removing parts that do not have significant information (empty or too noisy). This process is called voice activity detection (VAD). However, because the sample data used was fairly clean, the VAD process does not have a significant impact on the audio files. Before the data is extracted, the audio files are merged and then split with a duration of 30 seconds. In the MFCC extraction method, 1 9 mel-coefficient + 1 energy coefficients were extracted. The 20 delta MFCC and 20 double-delta MFCC from the mel-coefficient derivatives are also calculated so that a total of 60 MFCC features are obtained. In the MFCC+SDCC configuration with the N-d-P-k 7-1-3-7 parameter scheme, 7 static MFCC components are extracted plus 49 SDCC components, resulting in a total of 56 SDCC features [20].

2.2.1 MFCC extraction

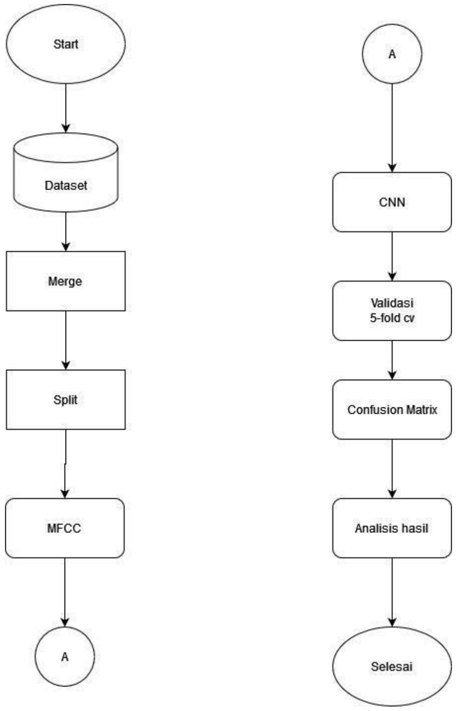


Figure 4: Extraction process MFCC

MFCC is an acoustic feature extraction method based on audio frequencies spectrum. These features are obtained by transforming the original windowing processes audio data to a frequency domain using discrete fourier transform and then onto Mel filters. The logarithmic values are then obtained from the weighted Mel results to get its energy readings and then inverted to time domain using discrete cosine transform. This Mel filter are said to be able to envelope human voice spectrum. MFCC coefficient obtained will be able to accommodate human ear critical frequencies in the high frequency. In the next process, a split and merge process is carried out, because the data used to be processed by the algorithm must have a consistent length, therefore the data is processed first by merging and splitting until it becomes a .wav file which has a duration of 30 seconds. The merged and split data results in a .wav file which will later be used for MFCC extraction. Furthermore, the data is extracted using MFCC. Then a 5-fold cross-validation process is carried out to find the best combination of data, which could be in terms of accuracy, precision, error, and others. After the 5-fold cross-validation process, the data is evaluated and processed to determine the level of prediction accuracy.

2.2.2 SDCC extraction

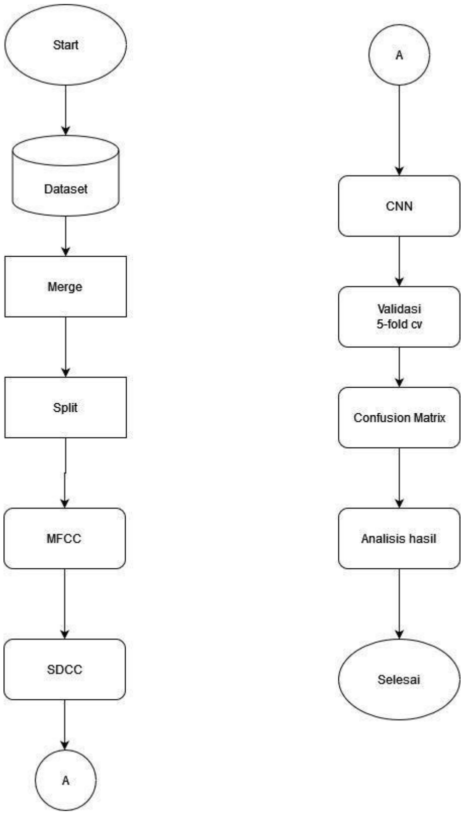
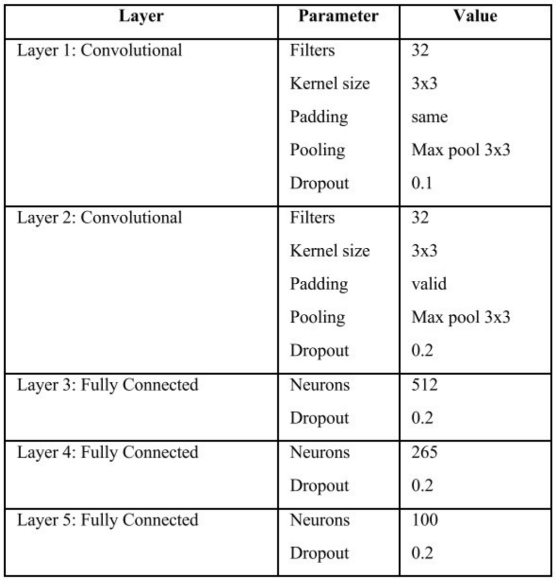


Figure 5: extraction process SDCC

Voice data is cleaned using VAD. In the next process, a split and merge process is carried out, because the data used to be processed by the algorithm must have a consistent length, therefore the data is processed first by merging and splitting until it becomes a .wav file which has a duration of 30 seconds. The merged and split data results in a .wav file which will later be used for MFCC extraction. Furthermore, the data is extracted using MFCC. Furthermore, the MFCC data is processed using SDC so that it becomes SDCC extraction. Then a 5-fold cross-validation process is carried out to find the best combination of data, which could be in terms of accuracy, precision, error, and others. After processing the data is evaluated and processed to determine the level of prediction accuracy.

2.3 Model Classifier Training

CNN construction and the separation between training and test data (train_test_split) were carried out using the Keras toolbox package. The CNN architecture used can be seen in Table 6. During training, the batch size used is 64, with an epoch of 35. All programming is written in python.



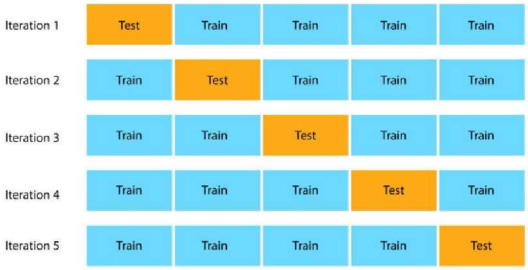
Layer	Parameter	Value
Layer 1: Convolutional	Filters	32
	Kernel size	3x3
	Padding	same
	Pooling	Max pool 3x3
	Dropout	0.1
Layer 2: Convolutional	Filters	32
	Kernel size	3x3
	Padding	valid
	Pooling	Max pool 3x3
	Dropout	0.2
Layer 3: Fully Connected	Neurons	512
	Dropout	0.2
Layer 4: Fully Connected	Neurons	265
	Dropout	0.2
Layer 5: Fully Connected	Neurons	100
	Dropout	0.2

Figure 6: CNN architecture parameters.

2.4 K-fold Validation

In this stage, the test was carried out k times. k-Fold validation is used to estimate prediction error in evaluating model performance. The data is divided into k subsets of almost equal numbers. In each iteration, one of the subsets will be used as training data and testing data [21].

After initial training, we will be validating the model using 5-fold cross validation. Using this, we will be able to see if the model is overfitted and not effectively predicting new inputted data across each fold. Figure 4 illustrate what happens with our training data on validation. On every iteration, we can observe the accuracy and F1 score of each iteration to detect overfitting and also underfitting early. If this happens, we can tweak the hyperparameters to better improve the model training. When the model passed our validation, it will then be retrained using the whole training data and tested using test data.



Iteration 1	Test	Train	Train	Train	Train
Iteration 2	Train	Test	Train	Train	Train
Iteration 3	Train	Train	Test	Train	Train
Iteration 4	Train	Train	Train	Test	Train
Iteration 5	Train	Train	Train	Train	Test

Figure 7: 5-fold cross validation illustration

After validation and testing, the results will be subject to evaluation and comparison with the original model. As in [17], we will be using classification accuracy and F1 score for evaluation. These values from both extraction method will be compared and the result will become base for further analysis and research. The accuracy and F1 score of each model will be obtained using Keras toolkit and the cross validation will be done using scikit-learn.

3. RESULTS AND COMPARISON

In this section, the result of the experiment done will be explained. The validation using 5-fold CV will also be discussed.

2.1 Training Validation

After the initial CNN training, we validate the model using 5-fold CV. The result of this validation can be found in Table 1.

Table 1: MFCC training 5-fold CV

	1	2	3	4	5
Step 1	76.42%				
Step 2		81.42%			
Step 3			72.85%		
Step 4				77.85%	
Step 5					82.01%

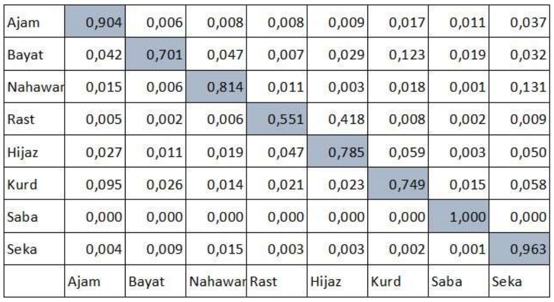
The results obtained with MFCC extraction is not overfitted or underfitted in any step. The results averaged 78.11% accuracy. The highest accuracy obtained is in step 5 which is 82.01%.

Table 2: MFCC training 5-fold CV

	1	2	3	4	5
Step 1	79.76%				
Step 2		92.4%			
Step 3			90.24%		
Step 4				83.13%	
Step 5					84.72%

The results obtained with SDCC extraction is not overfitted or underfitted in any step. The results averaged 86.05% accuracy. The highest accuracy obtained is in step 5 which is 92.4%. We can already see that SDCC extraction gives a better accuracy performance even in validation stage.

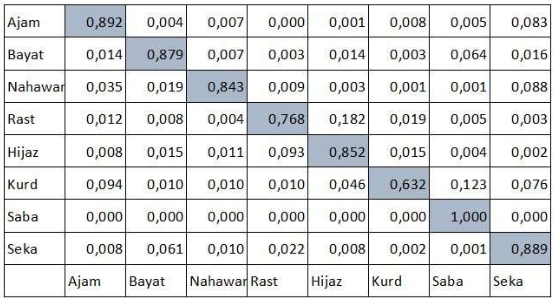
2.2 Testing Result



Ajam	0,904	0,006	0,008	0,008	0,009	0,017	0,011	0,037
Bayat	0,042	0,701	0,047	0,007	0,029	0,123	0,019	0,032
Nahawar	0,015	0,006	0,814	0,011	0,003	0,018	0,001	0,131
Rast	0,005	0,002	0,006	0,551	0,418	0,008	0,002	0,009
Hijaz	0,027	0,011	0,019	0,047	0,785	0,059	0,003	0,050
Kurd	0,095	0,026	0,014	0,021	0,023	0,749	0,015	0,058
Saba	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000
Seka	0,004	0,009	0,015	0,003	0,003	0,002	0,001	0,963
	Ajam	Bayat	Nahawar	Rast	Hijaz	Kurd	Saba	Seka

Figure 8: Confusion matrix MFCC

Figure 8 contains the confusion matrix from MFCC extraction model test. As we can see, the result is decent except on Rast maqam which yields 0,551. The best classification accuracy is obtained from Saba maqam with 100% accuracy. The test accuracy of MFCC extraction model is 80.8% with 0.7821 F1 score.



Ajam	0,892	0,004	0,007	0,000	0,001	0,008	0,005	0,083
Bayat	0,014	0,879	0,007	0,003	0,014	0,003	0,064	0,016
Nahawar	0,035	0,019	0,843	0,009	0,003	0,001	0,001	0,088
Rast	0,012	0,008	0,004	0,768	0,182	0,019	0,005	0,003
Hijaz	0,008	0,015	0,011	0,093	0,852	0,015	0,004	0,002
Kurd	0,094	0,010	0,010	0,010	0,046	0,632	0,123	0,076
Saba	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000
Seka	0,008	0,061	0,010	0,022	0,008	0,002	0,001	0,889
	Ajam	Bayat	Nahawar	Rast	Hijaz	Kurd	Saba	Seka

Figure 9: Confusion matrix SDCC

Figure 9 contains the confusion matrix from SDCC extraction model test. In this model, the worst performing class is Kurd with 0.632. The best performing class is still Saba with 100% accuracy. The average accuracy of SDCC model is 84.43% with 0.8560 F1 score.

Table 3. Test results comparison

Model	5-Fold CV Accuracy (%)	Test Accuracy	F1-Score
MFCC	78.11	0.808	0.7821
SDCC	**86.05**	**0.8472**	**0.8560**

Table 3 contains the comparison between MFCC extraction model and SDCC extraction model. As we can see, SDCC performance beats MFCC performance in every aspect. The classification accuracy is improved by around 4% and the F1 score increase by 0.08 points.

4. CONCLUSIONS

Integrating SDCC extraction to the original system in [17] give a positive improvement in classification accuracy and F1 score performance. This indicates that SDCC is a better feature extraction for maqam classification. This also indicates that the relationship between consecutive audio frames is an important parameter in classifying maqam. Although the improvement is not very significant, we believe that with enough tuning SDCC can open up possibilities not only in maqam classifications but in audio classifications in general.

5. SUGGESTIONS

For future research, we would like to see SDCC to be combined with other audio features such as spectral and also temporal features to see its performance in maqam classification. Another suggestion is to conduct research with only SDCC but with even larger datasets to verify its effectiveness in maqam or audio classifications.

REFERENCES:

[1] Hackett, C., Grim, B., Stonawski, M., Skirbekk, V., Potančoková, M., And Abel, G., "The Global Religious Landscape", 2012, Pew Res. Center, Washington, Dc, Usa, Tech. Rep.

[2] Https://Web.Archive.Org/Web/20211019173310/Https://Data.Kemenag.Go.Id/Agamadashbo Ard/Statistik/Umat, 20:28:55 07/07/2023.

[3] Oktaviani, D., Bijaksana, M. A., And Asror, I. (2019). Building A Database Of Recurring Text In The Quran And Its Translation. Procedia Computer Science, Vol. 157, Pg. 125-133, Issn 1877-0509, Https://Doi.Org/10.1016/J.Procs.2019.08.149

[4] Hackett, C., Grim, B., Stonawski, M., Skirbekk, V., Potančoková, M., And Abel, G., "The Global Religious Landscape", 2012, Pew Res. Center, Washington, Dc, Usa, Tech. Rep.

[5] Santoso, H. B., Schrepp, M., Hasani, L. M., Fitriansyah, R., And Setyanto, A. (2022) The Use Of User Experience Questionnaire Plus (Ueq+) For Cross-Cultural Ux Research: Evaluating Zoom And Learn Quran Tajwid As Online Learning Tools. Heliyon, Vol. 8, 11, E11748, Issn 2405-8440, Https://Doi.Org/10.1016/J.Heliyon.2022.E11748.

[6] Hassan, S. S. And Zailaini M. A. (2013). Analysis Of Tajweed Errors In Quranic Recitation. Procedia - Social And Behavioral Sciences, 103, Pp 136 – 145.

[7] Nelson, K. (2001) The Art Of Reciting The Qur'an. Cairo, Egypt: American Univ. Cairo Press.

[8] Globerson, E., Elias, T., Kittany, N., And Amir, N. (2016). Pitch Discrimination Abilities In Classical Arab-Music Listeners. Elsevier, Applied Acoustics Vol. 126, Pp 120–124. Http://Dx.Doi.Org/10.1016/J.Apacoust.2015.09.0 03.

[9] Touma, H. H. (1971). The Maqam Phenomenon: An Improvisation Technique In The Music Of The Middle East. Ethnomusicology, Vol. 15, No. 1 (Jan., 1971), Pp. 38-48 (11 Pages). University Of Illinois Press. Https://Doi.Org/10.2307/850386.

[10] Ḥisāmpūr, S. And Jabbāra, A. (2010). A Study Of The Qur'ānic Musical Modes (Maqāmāt) In The Recitation Of Several Qārīs (Reciters). Quran And Hadith Studies, 42(2), -. Doi: 10.22067/Naqhs.V42i2.11816.

[11] Touma, H. H. (1996). The Music Of The Arabs. Milwaukee, Wi, Usa: Hal Leonard Corporation

[12] Alkhatib, H. H., Mansor, E. I., Alsamel, Z., And Albarazi, J. A. (2020). A Study Of Using Vr Games In Teaching Tajweed For Teenagers. Interactivity And The Future Of The Human-Computer Interface. Hershey, Pa, Usa: Igi Global, Pp. 244–260.

[13] Ibrahim, N. J., Yusoff, Z. M., And Razak, Z. (2011). Improve Design For Automated Tajweed Checking Rules Engine Of Quranic Verse Recitation: A Review. Quranica-Int. J. Quranic Res., Vol. 1, No. 1, Pp. 39–50.

[14] Tandel, N. H., Prajapati, H. B., And Dabhi, V. K. (2020). Voice Recognition And Voice Comparison Using Machine Learning Techniques: A Survey. 2020 6th.

[15] Bai, Z. And Zhang X. (2021). Speaker Recognition Based On Deep Learning: An Overview. Neural Networks 140 (2021) 65–99, Elsevier.

[16] Lee, K. A., Li, H., Deng, L., Hautamäki, V., Rao, W., Xiao, X., Larcher, A., Sun, H., Nguyen, T. H., Wang, G., Sizov, A., Chen, J., Kukanov, I., Poorjam, A. H., Trong, T. N., Xu, C.-L., Xu, H., Ma, B., Chng, E. S., And Meignier, S. (2016): The 2015 Nist Language Recognition Evaluation: The Shared View Of I2r, Fantastic4 And Singams. Interspeech 2016, Isca, Isca, 3211–3215. Https://Doi.Org/10.21437/Interspeech.2016-624

[17] Shahriar, S. And Tariq, U. (2021). Classifying Maqams Of Qur'anic Recitations Using Deep Learning. Ieee Access Vol. 9, 117271-117281, 2021. 10.1109/Access.2021.3098415

[18] Li, H., Ma, B., And Lee, K. A. (2013): Spoken Language Recognition: From Fundamentals To Practice. Proceedings Of The Ieee, 101(5), 1136–1159. Https://Doi.Org/10.1109/Jproc.2012.2237151

[19] Mclaren, M., Nandwana, M. K., Castán, D., And Ferrer, L. (2018): Approaches To Multi-Domain Language Recognition. The Speaker And Language Recognition Workshop (Odyssey 2018), Isca, Isca, 90–97. Https://Doi.Org/10.21437/Odyssey.2018-13

[20] P. A. Torres-Carrasquillo Et Al., "The Mitll Nist Lre 2009 Language Recognition System," 2010 Ieee International Conference On Acoustics, Speech And Signal Processing, Dallas, Tx, Usa, 2010, Pp. 4994-4997, Doi: 10.1109/Icassp.2010.5495080.

[21] Mardiana L., Kusnandar D And Satyahadewi N. (2022).Analisis Diskriminan Dengan K Fold Cross Validation Untuk Klasifikasi Kualitas Air Di Kota Pontianak. Buletin Ilmiah Mat. Stat. And Terapannya (Bimaster)