Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/v8i2.12912
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Music Genre Classification using K-Nearest Neighbor and Mel-Frequency Cepstral Coefficients

Tika Pratiwi[1*)], Andi Sunyoto[2)], Dhani Ariatmanto[3)]
[1,2,3)]Amikom Yogyakarta University, Indonesia
[1)] tika.pratiwi@students.amikom.ac.id, [2)] andi@amikom.ac.id, [3)] dhaniari@amikom.ac.id

Submitted : Aug 21, 2023 | **Accepted** : Feb 29, 2024 | **Published** : Apr 1, 2024

Abstract: Music genre classification plays a pivotal role in organizing and accessing vast music collections, enhancing user experiences, and enabling efficient music recommendation systems. This study focuses on employing the K-Nearest Neighbors (KNN) algorithm in conjunction with Mel-Frequency Cepstral Coefficients (MFCCs) for accurate music genre classification. MFCCs extract essential spectral features from audio signals, which serve as robust representations of music characteristics. The proposed approach achieves a commendable classification accuracy of 80%, showcasing the effectiveness of KNN-MFCC fusion. Nevertheless, the challenge of overlapping genres, particularly rock and country, demands special attention due to their shared acoustic attributes. The inherent similarities between these genres often lead to misclassification, hampering accuracy. To address this issue, an enhanced feature engineering strategy is devised, leveraging deeper insights into the subtle nuances that differentiate rock and country music. Additionally, a refined KNN distance metric and neighbor selection mechanism are introduced to further refine classification decisions. Experimental results underscore the effectiveness of the refined approach in mitigating genre overlap issues, significantly enhancing classification accuracy for rock and country genres. This study contributes to the advancement of music genre classification techniques, offering an innovative solution for handling overlapping genres and demonstrating the potential of KNN-MFCC synergy in achieving accurate and refined genre classification.

Keywords: Music; Genre; Genre Classification; KNN; Mel-Frequency Cepstral Coefficient

INTRODUCTION

Music genre classification has arisen as an intriguing and complicated study topic in the quickly expanding digital age. The enormous impact of technology on the music industry, including music creation, distribution, and consumption, has created new obstacles and opportunities for comprehending and categorizing musical compositions. However, with the large number of types and genres of music available, it can be difficult for users to find music that suits their tastes. Automatic classification of music genres can help solve this problem. Using machine learning techniques and artificial intelligence, it can predict the appropriate musical style for a song by analyzing musical characteristics such as pattern, rhythm, pitch, and harmony.

A study broadcast networks for music classifications using CNN with dataset GTZAN, Homburg, FMA give the best result 90%. Although the attractive classification performances were presented in tackling the GTZAN and Ballroom based datasets, it remains unclear whether the broadcast networks offer competitive generalization abilities on a broader class of music classification settings. This research using confusion matrix for model evaluation and for preprocessing using Short Time Fourier Transform (STFT). At current, the classification performance in HOMBURG and FMA datasets are still low, suggesting that broadcast networks still lack the desirable generalization abilities. Furthermore, we plan to explore different representations of audio files, although spectrograms are convenient and straight forward to use in MGC, there is potential to explore other representations such as scalograms and wavelet scattering transforms (Ahmed Heakl et al., 2022).

Web-based Music Genre Classification for Timeline Song Visualization and Analysis using FF, LSTM Network, Naïve Bayes, SVM with GTZAN dataset. Achieves a mean average precision (AP) of 0.314 and an average AUC score of 0.959. In particular, the FF trained on the unbalanced audio set yields the best AP / AUC

* Author corresponding

e-ISSN : 2541-2019
p-ISSN : 2541-044X

scores with values of 0.465 / 0.930. The main problem we face when comparing predicted genres with the information in online services is how to identify a positive (and a negative) match. When trying to compare them with categorizations from other sources, such as online music platforms, because there is no standard or formal way of defining genres (Castillo, et al., 2021).

From the several problem statements in the literature review above, the performance of classification in the HOMBURG and FMA data sets is still low, indicating that broadcasting networks still lack the desired generalization ability, lack a variety of datasets used, and some genres are still wrong and overlap, such as rock and country (Jaime Ramirez Castilo et al., 2021).

The K-nearest neighbors technique will be used because it has produced the best results for this problem in a number of studies. A well-liked machine learning approach for regression and classification is K-Nearest Neighbors. It forecasts data points according to similarity metrics, or the separation between them.

According to prior research, it may be characterized by the existence of several existing musical genres; however, there are difficulties in distinguishing these genres, particularly between the rock and country genres, where there is confusion and overlap in recognizing the genres of some songs. The difference between this and other studies is that the researcher uses different datasets and methodology.

To address this issue, the purpose of this research is to create a music genre categorization system based on the K-Nearest Neighbors (KNN) method and the MFCC (Mel-Frequency Cepstral Coefficients) preprocessing technique. This research is expected to contribute to enhancing the accuracy of music genre classification while also describing the various sound characteristics of each genre, hence strengthening the separation between rock and country genres.

LITERATURE REVIEW

In the digital age, the classification of music genres has become an exciting and complex topic of research. Prior to doing our own research in this subject, we conducted a thorough literature review to obtain insights and information from previous studies. We give an overview of significant research projects that have investigated music genre classification, putting light on their techniques, conclusions, and prospects for future improvement in this review. In a study by (Ahmed Heakl et al., 2022) , investigated broadcast-based neural networks in their study. Their goal was to improve localization and generalization with a small parameter set (about 180k). To do this, they studied 12 alternative broadcast network variants, looking at aspects including block configuration, pooling method, activation function, normalization, label smoothing, channel interplay, LSTM block inclusion, and beginning schema variant. Computational tests were carried out using datasets such as GTZAN, Extended Ballroom, HOMBURG, and Free Music Archive (FMA), with features extracted manually from audio signals using Short Time Fourier Transform (STFT). Notable results include a GTZAN dataset accuracy of 90.0% and Extended Ballroom accuracy of 93.1%, but lower figures for other datasets, highlighting the need for enhanced generalization. Similar research was also conducted by (Jaime Ramirez Castilo et al., 2021) , investigated Music Information Retrieval (MIR), including genre classification and other topics. They created a web application that retrieves and categorizes songs from YouTube. The study confronted the difficulty of comparing genre predictions across multiple categorical representations by leveraging datasets such as GTZAN and employing methods such as Feed Forward (FF), LSTM Network, Recurrent Neural Network (RNN), Naive Bayes, and SVM. Using a confusion matrix, the FF and LSTM algorithms outperformed the others, with an average precision (AP) of 0.314 and an AUC score of 0.959. Notably, the FF technique achieved the best AP/AUC values when trained on an imbalanced audioset.

Research conducted by (Rui Yang et al., 2020) addressed the problem of identifying music genres on mobile devices without the need for expertise in handcrafted feature extraction. They pioneered the parallel recurrent convolutional neural network (PRCNN) approach, reaching 92% accuracy on the GTZAN dataset and 92.5% on another. The PRCNN method combines feature extraction and classification, resulting in robust feature representation. The benefits of employing the Short-Time Fourier Transform (STFT) to translate time domain information into frequency domain information for increased music analysis were highlighted in this study. Next, research by (Safaa Allamy et al., 2021) used 1D CNN for music genre categorization in an architectural study, achieving an accuracy of 80.93% with the GTZAN dataset. The discovery paves the path for future research with larger datasets, such as the LMD Latin music dataset and the MSD dataset, to improve accuracy and robustness even more. Research conducted (Lvyang Qiu et al., 2021) classify music genre classification, deep bidirectional transformers were combined with a predictive encoder technique. Their study used the CNN and MPE algorithms on the Lakh MIDI dataset, with a maximum accuracy of 94%. However, issues with label ambiguity persisted. Researcher (Jash Mehta et al., 2021), used transfer learning on log-based MEL spectrograms to classify music genres. Using the GTZAN dataset, they achieved accuracies ranging from 71.30% to 79.00% using several architectures such as Resnet34, Resnet50, Alexnet, and VGG16. The study ran into concerns with genre overlap. Other research conducted by (Tianhao Qiao, 2021), also investigated sub-spectrogram segmentation vs. temporal

* Author corresponding

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/v8i2.12912
e-ISSN : 2541-2019
p-ISSN : 2541-044X

attention mechanisms with the ESC-50 dataset, attaining a maximum accuracy of 86.4% using the CNN approach. Research by (Yu Su et al., 2019) They used a two-stream CNN based on decision-level fusion for environmental sound classification, attaining an accuracy of 95.2% with the LMCNet architecture. The study emphasized the importance of better approaches for speech recognition. Based on the literature analysis, an expanded feature engineering method is built to address this issue, employing deeper insights into the subtle details that define rock and country music. To further refine the categorization decision, an improved KNN distance metric and neighbor selection process are introduced.

In 2019, NASSIF et al. reviewed 174 papers which were developed ASR models based on deep learning models where all the deep learning models were fed by extracting specific features. Moreover, all the papers (174) have been published between 2006 to 2018. They found that most of the researchers, which was about 69%, still use MFCCs as a feature to feed the machine learning models. (A. B. Nassif, I. Shahin, I. Attili, M. Azzeh and K. Shaalan, 2019)

METHOD

KNN Method

The most common classification method is K-Nearest Neighbor (KNN), however that method be used to for prediction and forecasting by similarity. KNN has a number of parameters, including robustness in the face of noisy training data and efficacy when dealing with large training datasets. In KNN, which is an example of instance-based learning, the categorization for classified new data is checked to the training set using a more common training data. Its distance measure method is used to calculate the size of an object and can be used to determine similarity by evaluating the distance between the two objects.

a. Analysis of Feature Extraction

The first stage of feature extraction is implemented on audio files. The audio file should go through four stages in the feature extraction stage: frame blocking, signal windowing, signal transformation, and counting feature. A number that describes the amplitude of the audio file.

1. Frame Blocking Analysis

There in frames layer blocked signals, the input audio file will be divided into signals made up of multiple frames. The number of frames the audio track would be split into it and the data rate will define the duration of the a frame made up of numerous samples. This research implemented a frame size of 50 milliseconds with no repetition. In this frame, the blocked signals remains constant, that implies it does not move. These technique produces an output that was separated into multiple frame. The waveform is a starting to change moment and relatively non waveform. In order to have stable acoustic features, speech must be evaluated for a sufficiently short length of time. As a result, when the audio signals was considered toward being stable, audio analytics must be performed in smaller sections.

Short-term spectral structures are usually collected every ten seconds in 20-ms frames. Advance the time window every 10 ms to track the temporal features of individual speech sounds, and the 20 ms analysis window is usually enough to provide good spectral resolution while simultaneously being short enough to detect important temporal aspects. Each speech sound in the input sequence is generally centered at some frame using the overlapping analysis. Each frame has a window applied to it to taper the signal towards the frame boundaries. Hanning or Hamming windows are utilized in the majority of circumstances. While conducting the DFT on the signal, this is done to boost the harmonics, smooth the edges, and lessen the edge effect. Short-term spectral measurements are typically performed every ten seconds in 20-ms frames.

Advance the time window every 10 ms to track the temporal features of individual speech sounds, and the 20 ms analysis window is usually enough to provide good spectral resolution while simultaneously being short enough to detect important temporal aspects. Each speech sound in the input sequence is generally centered at some frame using the overlapping analysis. Each frame has a window applied to it to taper the signal towards the frame boundaries. Hanning or Hamming windows are employed in the majority of circumstances. While performing the DFT on the signal, this is done to boost the harmonics, smooth the edges, and lessen the edge effect.

2. Window Signal Analysis

The signal will approach the windowing stage once it has been separated into multiple frames. The purpose of windowing is to eliminate the impact of blocking frame discontinuities. The Blackman Window, Window Rectangle, and Hamming Window are three types of windowing procedures. Hamming window is a type of window that has a small side lobe and a large main lobe, resulting in a more smooth windowing that eliminates the impacts of discontinuities. The signal on the frames that have been windowed is the result of this stage.

3. Signal Transformation Analysis

At this point, the original signal will be converted to a time frequency. Each windowed frame is converted into magnitude spectrum by applying DFT. The objective of changing the signal in a domain is to calculate the features of an input music, where the features are the frequency characteristics of the music.

* Author corresponding

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/v8i2.12912
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Classification Music Type

The categorisation of musical genres will be the final process. The K-Nearest Neighbor method is used to classify music (KNN). The previously collected features will be compared to the training data that is currently accessible. The euclidean distance of each feature to the training data will be calculated, and the feature with the lowest distance to the training data will be chosen. We will acquire some form of music as a candidate k-parameters from the results of the euclidean distance computation. The music input will be categorized according to the type of music that the majority of the candidates prefer.

Machine Learning

Machine learning is a field of artificial intelligence that focuses on developing computational algorithms and models that enable computer systems to learn from data and experience, without the need to be explicitly programmed. Its purpose is to enable computers to recognize patterns in data, make predictions, make decisions, or perform certain tasks automatically.

1. How machine learning works generally involves the following steps:
2. Data Collection: First, relevant data relating to the problem or task to be completed is collected. This data can be numbers, text, images, audio, or any other appropriate type of data.
3. Pre-processing Data: Collected data often needs to be pre-processed to clean, remove noise, or fill in missing values. This involves steps such as normalizing, coding, or transforming data so that it can be used in machine learning models.
4. Data division: The data is then divided into two main parts: training data and testing data. The training data is used to train the model, while the test data is used to test the extent to which the model being trained is able to produce accurate predictions.
5. Model Selection: In this step, an appropriate machine learning model is selected based on the task or problem to be solved. There are many types of machine learning models that can be used, including regression, classification, clustering, artificial neural networks, and more.
6. Model Training: In this stage, the machine learning model is trained using the training data. This process involves optimizing the parameters and weights in the model in order to produce accurate predictions or outputs. The model is given training data and iteratively adjusts itself to improve performance.
7. Model Evaluation: After the model has been trained, its performance is evaluated using discrete test data. Commonly used evaluation metrics include accuracy, precision, recall, or the ROC curve, depending on the type of problem and the type of model used.
8. Model Application: Once the model is deemed good enough, it can be applied to new data or real situations to make predictions or make decisions based on the inputs provided.
9. Maintenance and Update: Machine learning models usually require regular maintenance and updating. This may involve retraining the model with new data to maintain performance or updating model parameters and weights as changes occur.

This process provides an overview of how machine learning works. However, keep in mind that there are different approaches, techniques, and algorithms in machine learning, and the process can vary depending on the type of problem you want to solve or the task you want to perform.

Machine and deep learning are research areas in multidisciplinary fields that constantly evolve due to the advances in data analytics research in the age of Big Data, Cloud digital ecosystem, etc. The effects of new computing resources and technologies combined with increasing data sets are changing many research, health, and industrial areas. As technology advances, novel solutions are sought in many areas to address complex problems, presenting data mining projects with a significant challenge in deciding which tools to choose (Patrick Schneider, Fatos Xhafa, 2022).

MFCC

Nowadays, many feature extraction techniques are available in a variety of fields based on the characteristics of the raw data. In most of the fields, finding harmonics and sidebands of signal in both time and frequency domain are important to any pattern recognition system. Power spectrum using Fast Fourier Transform (FFT) is used to capture the harmonics and sidebands of the signal in the time domain. While cepstrum; such as Mel Frequency Cepstrum Coefficient (MFCC), Gamma Tone Cepstrum Coefficient (GTCC), is capable to extract harmonics and sidebands of the spectrum version of the signal (B. Liang, S. D. Iwnicki and Y. Zhao., 2013) .

Recently, the MFCC features have been widely used for this purpose; for example, Yusuf & Hidayat evaluated two well-known features which are 13 MFCCs and Discrete Wavelet transformation when fed to the kNN. The obtained result shows that the performance of the 13 MFCCs outperformed DWT (S. A. Alodia Yusuf and R. Hidayat, 2019).

* Author corresponding

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/v8i2.12912
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Mel-frequency cepstral coefficients (MFCC) is a feature extraction method that is commonly used in speech processing and speech recognition. MFCC is used to convert the audio signal into a compact numerical representation, which represents the spectral characteristics of the sound signal.

KNN Method

KNN is an instance-based or lazy learning algorithm, which means it doesn't do the learning process explicitly. Instead, it stores all training data as "instances" in attribute space. The KNN algorithm classifies new data by finding the K-nearest neighbors of the data in the attribute space. If K = 1, this algorithm is referred to as KNN with classification based on one nearest neighbor. The main process in KNN is calculating the distance between new data and each training data using distance metrics (eg Euclidean distance or Manhattan distance). After the distance is calculated, KNN selects the K nearest neighbors depending on the specified K value. For classification problems, the most common label of the selected neighbors is used as a prediction for the new data. For regression problems, this algorithm can calculate the average or median value of the selected target neighbor values as predictions.

RESULT

While achieving an 80% accuracy in music genre classification, particularly in the presence of overlapping genres like country and rock, is a commendable achievement, there are several avenues for further research and development that could potentially lead to even higher accuracy levels. As technology advances and new methodologies emerge, the field of music genre classification can continue to evolve to meet the challenges posed by genre overlap.

spectral_bandwidth_mean	spectral_bandwidth_var	rolloff_mean	...	mfcc16_var	mfcc17_mean	mfcc17_var	mfcc18_mean	mfcc18_var	mfcc19_mean	mfcc19_var	mfcc20_mean	mfcc20_var	label
1972.744388	117335.771563	3714.560359	...	39.687145	-3.241280	36.488243	0.722209	38.099152	-5.050335	33.618073	-0.243027	43.771767	blues
2010.051501	65671.875673	3869.682242	...	64.748276	-6.055294	40.677654	0.159015	51.264091	-2.837699	97.030830	5.784063	59.943081	blues
2084.565132	75124.921716	3997.639160	...	67.336563	-1.768610	28.348579	2.378768	45.717648	-1.938424	53.050835	2.517375	33.105122	blues
1960.039988	82913.639269	3568.300218	...	47.739452	-3.841155	28.337118	1.218588	34.770935	-3.580352	50.836224	3.630866	32.023678	blues
1948.503884	60204.020268	3469.992864	...	30.336359	0.664582	45.880913	1.689446	51.363583	-3.392489	26.738789	0.536961	29.146694	blues

Figure 1. Data

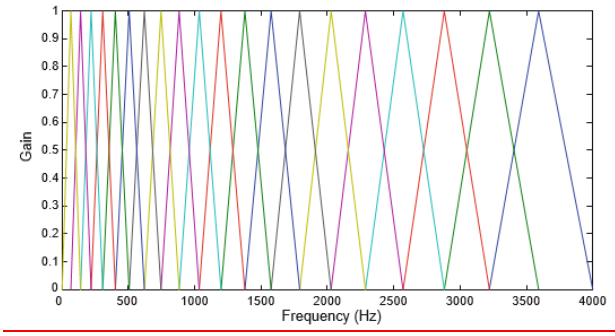

Figure 2. Apendix

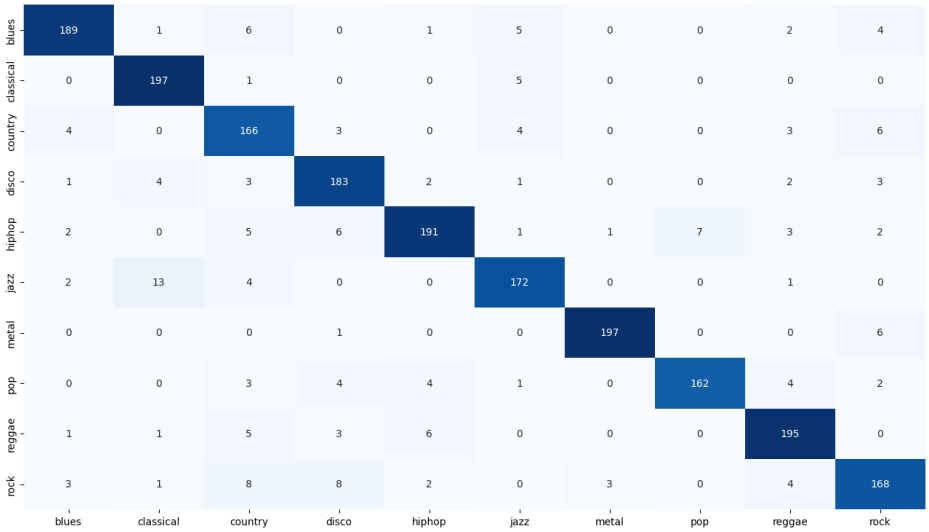

Figure 3. Confusion Matrix

* Author corresponding

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/v8i2.12912
e-ISSN : 2541-2019
p-ISSN : 2541-044X

DISCUSSIONS

In this section shows the performance carried out using the KNN method showing the values of precision, recall, and F1-Score where the model is better and more accurate in classifying music genres and the results of the rock and country genres do not overlap like the problem the several problem statements in the literature review above, the performance of classification in the HOMBURG and FMA data sets is still low, indicating that broadcasting networks still lack the desired generalization ability, lack a variety of datasets used, and some genres are still wrong and overlap, such as rock and country, which can be seen in the confusion matrix table.

Table 1. Classification Report

	Precision	Recall	F1-Score	Support
Blues	0.94	0.91	0.92	208
Classical	0.91	0.97	0.94	203
Country	0.83	0.89	0.86	186
Disco	0.88	0.92	0.90	199
Hiphop	0.93	0.88	0.90	218
Jazz	0.91	0.90	0.97	192
Metal	0.98	0.97	0.93	204
Pop	0.96	0.90	0.92	180
Reggae	0.91	0.92	0.97	211
Rock	0.88	0.85	0.87	197
Accuracy			0.91	1998
Macro Avg	0.91	0.91	0.91	1998
Weighted Avg	0.91	0.91	0.91	1998

CONCLUSION

Based on the research that has been done, it can be concluded that the KNN method with accuracy is 5 can be used to categorize music genres. maybe in the future, further research can be compared with other methods. Input data audio can be categorize into type of music because of the features with training data.

Training data of music were analyzed in total records of up to 100 data in 10 genres of music. wav, sample songs of up to 100 data that are not in agreement with the type of music. But, we just take 2 genre especially in country and rock to solve problem about overlapping genre. We use 80% of the data for training and 20% for testing. The accuracy of using the K-Nearest Neighbor algorithm with the k value is 5 to classify the type of music success is 0.91.

REFERENCES

Castillo, J. R., & Flores, M. J. (2021). Web-based music genre classification for timeline song visualization and analysis. IEEE Access,9,18801–18816.

Heakl, A., Abdelgawad, A., & Parque, V. (2022). A Study on Broadcast Networks for Music Genre Classification. http://arxiv.org/abs/2208.12086.

Qiu, L., Li, S., & Sung, Y. (2021). Dbtmpe: Deep bidirectional transformers-based masked predictive encoder approach for music genre classification. Mathematics, 9(5), 1–17.

Yang, R., Feng, L., Wang, H., Yao, J., & Luo, S. (2020). Parallel Recurrent Convolutional Neural Networks-Based Music Genre Classification Method for Mobile Devices. IEEE Access, 8, 19629–19637. https://doi.org/10.1109/ACCESS.2020.2968170

Mehta, J., Gandhi, D., Thakur, G., & Kanani, P. (2021). Music Genre Classification using Transfer Learning on log-based MEL Spectrogram. Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021, 1101–1107. https://doi.org/10.1109/ICCMC51019.2021.9418035

Liu, C., Feng, L., Liu, G., Wang, H., & Liu, S. (2019). Bottom-up Broadcast Neural Network For Music Genre Classification. http://arxiv.org/abs/1901.08928

Allamy, S., & Koerich, A. L. (2021). 1D CNN Architectures for Music Genre Classification. http://arxiv.org/abs/2105.07302

B. Liang, S. D. Iwnicki and Y. Zhao, "Application of power spectrum cepstrum higher order spectrum and neural network analyses for induction motor fault diagnosis", *Mech. Syst. Signal Process.*, vol. 39, no. 1, pp. 342-360, Aug. 2013.

A. B. Nassif, I. Shahin, I. Attili, M. Azzeh and K. Shaalan, "Speech recognition using deep neural networks: A systematic review", *IEEE Access*, vol. 7, pp. 19143-19165, 2019.

S. A. Alodia Yusuf and R. Hidayat, "MFCC feature extraction and KNN classification in ECG signals", *Proc. 6th Int. Conf. Inf. Technol. Comput. Electr. Eng. (ICITACEE)*, pp. 1-5, Sep. 2019.

* Author corresponding

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/v8i2.12912
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Patrick Schneider, Fatos Xhafa, in Anomaly Detection and Complex Event Processing over IoT Data Streams, 2022

Sun, B., Chen, H.: A survey of nearest neighbor algorithms for solving the class imbalanced problem. Wirel. Commun. Mob. Comput. **2021**.

Agarwal, Y., Poornalatha, G.: Analysis of the nearest neighbor classifiers: a review. Advances in Artificial Intelligence and Data Engineering: Select Proceedings of AIDE **2019**, 559–570

Yuan, B.-W., Luo, X.-G., Zhang, Z.-L., Yu, Y., Huo, H.-W., Johannes, T., Zou, X.-D.: A novel density-based adaptive k nearest neighbor method for dealing with overlapping problem in imbalanced datasets. Neural Comput. Appl. **33**(9), 4457–4481.2021

Jayaram Subramanya, S., Devvrit, F., Simhadri, H.V., Krishnawamy, R., Kadekodi, R.: Diskann: Fast accurate billion-point nearest neighbor search on a single node. Adv. Neural Inf. Process. Syst. **32** .2019

* Author corresponding