open access
Published online on: https://journal.uii.ac.id/jurnalsnati/
Check for updates
DOI: 10.20885/snati.v3.i3.42
JURNAL SNATI Sains, Nalar, dan Aplikasi Teknologi Informasi
Jurnal Sains, Nalar, dan Aplikasi Teknologi Informasi
Vol. 3 No. 3 (2024) 137 - 142 ISSN Media Electronic: 2807-5935

Clustering Analysis of Chess Portable Game Notation Text

Feri Wijayanto
Department of Informatics, Faculty of Industrial Technology, Universitas Islam Indonesia, Yogyakarta, Indonesia
feri.wijayanto@uii.ac.id

Abstract

Chess is a game that requires a high level of intelligence and strategy. Generally, in order to understand complex move patterns and strategies, the expertise of chess masters is required. With the rapid development in the field of machine learning, the digitization of chess game recordings in Portable Game Notation (PGN) format, and the availability of large and widely accessible data, it is possible to apply machine learning techniques to analyze chess games. This research studies the use of text clustering algorithms, specifically hierarchical clustering and K-means clustering, to categorize chess games based on their moves. To perform the analysis, we used 100 chess games that use certain openings such as French Defense, Queen's Gambit Declined, and English Opening. In implementing hierarchical clustering, single, average, and complete linkage methods are used. As an evaluation metric, we use the original openings of the analyzed games for comparison. As a result, our findings show that hierarchical clustering with single linkage is less effective. On the other hand, the average and complete linkage methods, as well as K-means clustering, successfully identify clusters corresponding to the original openings. Notably, K-means clustering showed the highest accuracy in clustering chess games. This research highlights the potential of machine learning techniques in uncovering strategic patterns in chess games, paving the way for deeper insights into game strategies.

.

Keywords: clustering analysis; text analysis; hierarchical clustering; k-means

1. Introduction

Chess is a rule-based game that requires a high level of intelligence and strategy. To be an expert in the game of chess requires not only an understanding of the rules but also a proven strategy that can only be acquired through extensive experience and study. Generally, this knowledge of strategy is only the domain of chess masters, whose experience enables them to recognize and execute complex patterns and plans.

On the other hand, the development of digital technology has made the experiences of chess masters accessible to the public. The chess games played by these masters can be recorded and analyzed. Chess game databases are now abundant and available in text form, especially in the form of Portable Game Notation (PGN) files, the standard format for documenting chess games [1]. The digitization process of chess games provides a wide opportunity to explore and study the move patterns in chess games. This opportunity is of course supported by developments in the field of machine learning, especially in the field of text analysis. With the availability of PGN files as text data, we can apply text analysis methods to uncover hidden patterns in chess game moves, thus enabling a deeper understanding of the underlying strategies.

In recent years, text clustering has made significant progress, allowing researchers to sort and study large amounts of text data in many fields. This method has been used successfully in areas like language processing [2],[3], social media analysis [4,5], and biology [6], helping to find patterns, themes, and trends in big sets of data. Chess, with its abundant data, is a good candidate for this type of analysis. While much research has been done on different aspects of chess, including some use of machine learning for developing chess engines [7,8], using text clustering to study chess game data and uncover strategic patterns remains relatively underexplored. However, some of works are already in the literature. Reid et.al. [9] focused on applying transfer learning to the Maia model to model the behavior of chess players. Meanwhile, Raghav [10] tried to classify players based on two attributes, namely player rating and opening code. With a different focus, this research attempts to classify and recognize chess game patterns based on the sequence of moves played.

This research aims to explore the tactical knowledge and pattern recognition possessed by chess masters through data analysis. To this end, I apply text clustering techniques to move sequences in chess games that have been recorded in PGN format. Specifically, this research wants to see if clustering

Received: 17-06-2024 | Revised: 30-06-2024 | Accepted: 01-07-2024 | Published: 05-07-2024

algorithms, in this case hierarchical clustering and k-means clustering can cluster PGN text data well. Furthermore, by applying text clustering to the PGN texts, this research is projected to gain a deeper understanding of strategic knowledge, especially related to chess opening selection.

2. Research Methods

2.1. Data Collection

To do the analysis, I use games that is recorded in https://www.chessgames.com/, a website that provides a comprehensive chess games database. The website claims on having more than 1,6 millions recorded games. The analysis only uses chess games that use certain openings. These openings are including:

- French Defense (Advance Variation, Steinitz Variation)
  - 1. e4 e6 2. d4 d5 3. e5 c5 4. dxc5
- French Defense (Advance Variation, Paulsen Attack)
  - 1. e4 e6 2. d4 d5 3. e5 c5 4. c3 Nc6 5. Nf3
- French Defense (Advance Variation, with 6.a3)
  - 1.e4 e6 2.d4 d5 3.e5 c5 4.c3 Nc6 5.Nf3 Qb6 6.a3
- Queen's Gambit Declined
  - 1.d4 d5 2.c4 e6
- English opening (symmetrical variation, main line)
  - 1 c4 c5 2 Nc3 Nc6 3 g3 g6 4 Bg2 Bg7 5 Nf3 Nf6 6 O-O O-O 7 d4

The selection of openings is arbitrary, but I have opted for three openings that are variations of the French defense. These selections will make the gameplay of these openings somewhat similar. Thus, it creates a challenge to categorize the games that use these three openings. To perform the analysis, I have only used the movement notation from the PGN files, excluding any additional metadata or comments.

I only used the most recent 20 chess games recorded for each opening so the total data to be analyzed is 100. Here is the order of the data:

- French Defense (Advance Variation, Steinitz Variation) data nr. 1-20
- French Defense (Advance Variation, Paulsen Attack) data nr. 21-40
- French Defense (Advance Variation, with 6.a3) data nr. 41-60
- Queen's Gambit Declined data nr. 61-80
- English opening (symmetrical variation, main line) data nr. 81-100

2.2. Text Preprocessing

As an initial stage in text clustering, the data needs to be prepared for further processing. Before entering the tokenization stage, the data obtained from PGN files needs to be cleaned and normalized. The order numbers of movement sequences, annotations, and unnecessary text are cleaned up leaving only the movement notation. Then, this movement notation needs to be standardized to maintain the consistency of the analysis results.

Having a cleansed data, we then perform the tokenization stage. The process is performed by splitting the set of moves into individual tokens. For example, a set of moves "1.e4 e6 2.d4 d5 3.e5 c5 4.c3 Nc6 5.Nf3 Qb6 6.a3" is split into individual tokens: ["e4", "e6", "d4", "d5", "e5", "c5", "c3", "Nc6", "Nf3", "Qb6", "a3"].

2.3. Feature Extraction

The next stage is feature extraction. This stage aims to represent text data into vector form (vectorization). At this stage we apply the vector space model by creating a Document-Term Matrix (DTM). DTM is a matrix whose rows represent documents (chess game) and columns represent unique terms (moves in chess). Basically, the DTM shows the frequency of moves in each chess game.

In a way, this DTM provides a representative numerical representation of the chess game, capturing the frequency and distribution of moves. Furthermore, the resulting Document-Term Matrix will be the input for the chosen clustering algorithm.

2.4 Clustering Algorithm

In this study, we use two clustering algorithms to analyze the sequence of moves in a chess game, namely Hierarchical Clustering [11] and K-Means Clustering [12,13]. In general, any clustering algorithm can be applied after performing feature extraction. However, since our objective is to observe game patterns and strategies, for practicality and explainability, we used a simple clustering algorithm. We consider that in case of success using these two algorithms, the results using advanced algorithms will be better.

On the one hand, hierarchical clustering will be very helpful in looking at the similarity level of chess games. This clustering algorithm starts by calculating the similarity level of the chess game and followed by merging the entities based on the proximity level into one cluster. This is repeated until all data points are merged into one cluster. For the agglomeration method, we use single linkage, complete linkage, and average linkage.

On the other hand, K-Means clustering can be used as a comparison to the results obtained by hierarchical clustering with the same number of clusters. Unlike hierarchical clustering which is based on the similarity level of data points, the K-Means clustering algorithm is based on the distance of data points to the cluster center. So, although ideally these two algorithms would produce the same clusters, they often produce slightly different results. In this study, we use a value of k = 5 (5 clusters) for the K-Means clustering algorithm, in accordance to the number of openings we use.

In this research, we use the concept of similarity in comparing document texts. Similarity is calculated by calculating the correlation of document text attribute values. While the distance computation for hierarchical clustering is done by converting the similarity value using the following formula [14],

$$x = 1 - \|x\|$$

As for k-means clustering, Euclidean distance is used to compute the distance of documents.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

2.5 Evaluation Metrics

As mentioned above, the dataset for this analysis is chess games that have been categorized based on the type of opening used. Hypothetically, the result of this clustering will be close to the initial category. Therefore, external validation will be conducted for evaluation. The initial categories of game data will be used to assess the performance of clustering algorithms, i.e. hierarchical clustering and k-means clustering.

3. Results and Discussions

Having 100 chess game sequences using five different openings, I applied two clustering algorithms, hierarchical clustering and k-means clustering.

3.1. Hierarchical Clustering

In this study three linkage criteria were used, single, average and complete linkage. The results of clustering by applying single linkage are shown in Figure 1. In Figure 1, it can be seen that there are only two obvious clusters formed. The cluster in the small rectangle on the right consists of most chess games that use the Queen's Gambit Declined opening and few chess games that use the French defense opening, Steinitz variation. The cluster in the large rectangle on the left consists of chess games that use various openings, French defense (advanced variation, Paulsen attack), French defense (advanced variation, with 6. a3), and English opening (symmetrical, main line).

The results of clustering using average linkage are shown in Figure 2. Unlike single linkage which only produces two clusters, clustering using average linkage produces five clusters. However, none of the clusters actually have the full number of members as the original members. The first cluster from the right consists entirely of chess games that use the French defense opening (Steinitz variation), however, there are two games that do not belong to this cluster. The second cluster from the right is a mixture of chess games that use the French defense (advanced variation, Paulsen attack) and French defense (advanced variation, with 6. a3) openings. However, there are chess games that use one of these two openings but are not included in the cluster

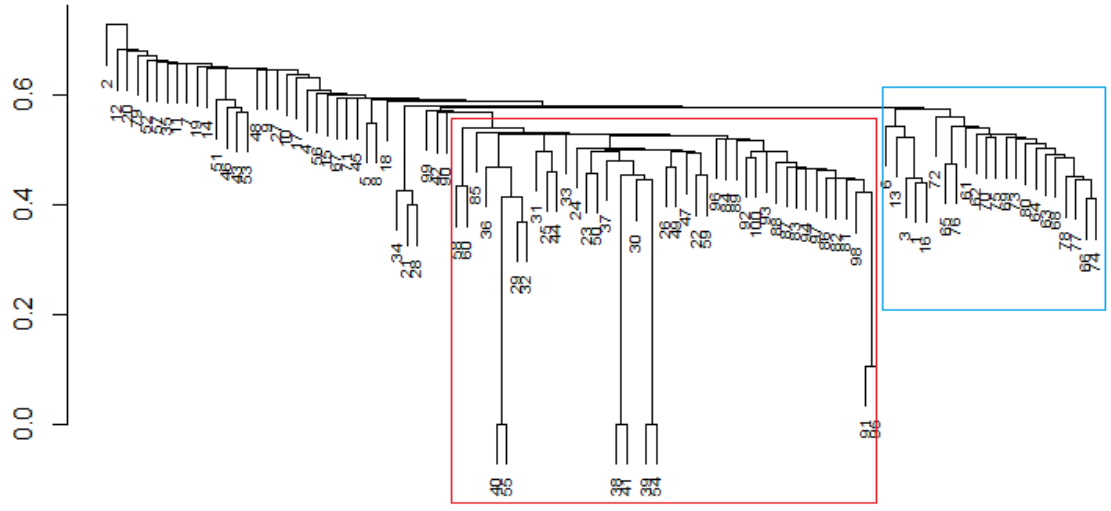

Figure 1. Dendrogram of hierarchical clustering result using single linkage criteria.

.

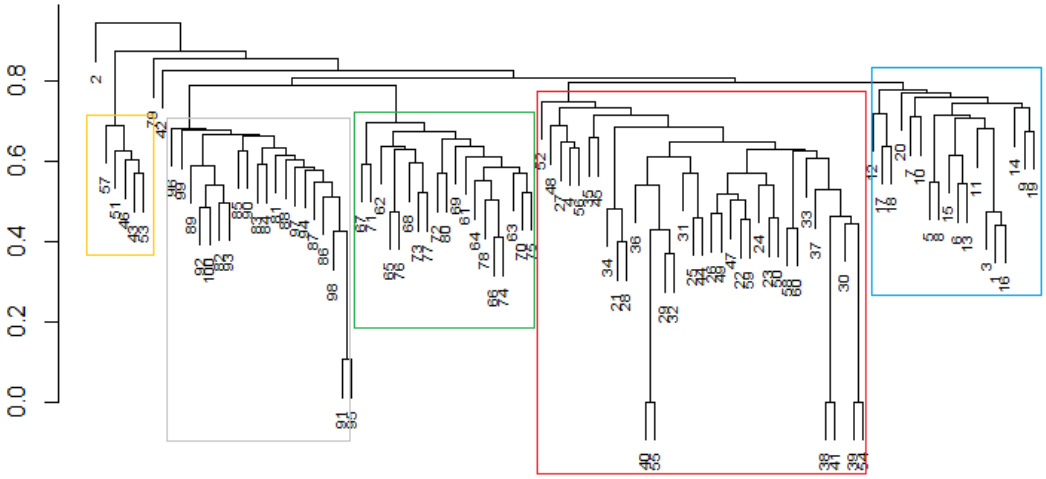

Figure 2. Dendrogram of hierarchical clustering result using average linkage criteria.

The third cluster from the right consists entirely of chess games that use the Queen's Gambit Declined opening, however, there is one game that is not included in the cluster. The fourth cluster from the right includes chess games that use the English opening (symmetrical variation, main line), but there is one chess game that applies the French defense (advanced variation, with 6. a3) that is included in the cluster. The leftmost cluster is a small cluster of five chess games that use the French defense opening (advanced variation, with 6. a3). Interestingly, there are three chess games that are considered different from the other games, one game with the French defense opening (Steinitz variation), one game with the French defense (advanced variation, with 6. a3), and the other using the English (symmetrical variation, main line).

The result of clustering using complete linkage is shown in Figure 3. There are two clusters that perfectly group chess games that use the same opening. The leftmost cluster successfully groups chess games that use the English opening (symmetrical variation, main line) and the second cluster from the left successfully groups chess games that use the Queen's Gambit Declined opening. Meanwhile, the three clusters on the right are a mixture of chess games that use the French defense. However, chess games that use the French defense opening (Steinitz variation) are more concentrated in the rightmost cluster.

3.2. K-means Clustering

In addition to hierarchical clustering, k-means clustering was also applied for comparison. The results of k-means clustering can be seen in Table 1. Table 1 consists of five columns and each column represents the cluster formed. Cluster formation is arbitrary and cluster members are listed in the first row. Like hierarchical clustering that uses complete linkage, k-means clustering succeeds in producing two clusters that group chess games that use the same opening perfectly. Cluster 1 consists of chess games that use the English opening (symmetrical variation, main line) and Cluster 2 consists of chess games that use the Queen's Gambit Declined opening. Although, Cluster 3 manages to collect all the chess games that use the French defense opening (Steinitz variation), there is one game that uses the French defense opening (advanced variation, with an opening of 6.a3). Clusters 4 and 5 contain chess games that use the French defense (advanced variation, Paulsen attack) and the French defense (advanced variation, with an opening of 6. a3).

Table 1. The result of k-means clustering

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
81-100	61-80	1-20, 42	27, 43, 46, 51, 53, 57	21-26, 28-41, 44, 45, 47-50, 52, 54, 55, 56, 58-60

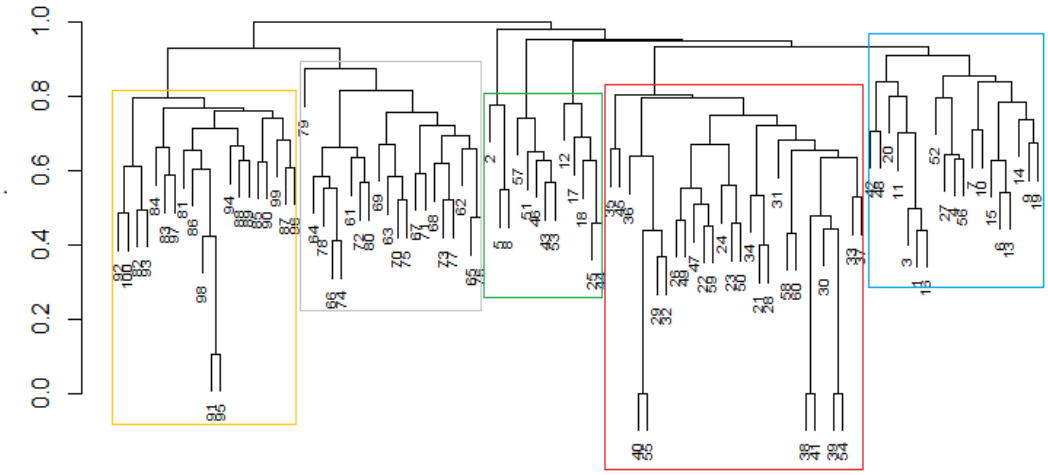

Figure 3. Dendrogram of hierarchical clustering result using complete linkage criteria.

3.3. Discussions

It is interesting to see that the results of the four clustering methods differ from each other. This is of course not only because of the different clustering methods, but also because of the different ways to determine the degree of closeness between clusters. And once again, it will result in different clusters as well. However, despite these differences, all clustering methods agree that chess games that use French defense variations, whether French defense (Steinitz variation), French defense (advanced variation, Paulsen attack), and French defense (advanced variation, with 6. a3), are similar to each other. This is evidenced by the failure of all clustering methods to perfectly group the three openings. This is natural as these three openings are variations of the same opening, the French defense. This similarity is seen in Figure 1, where the cluster in the large rectangle is a mixture of games using the French defense (advanced variation, Paulsen attack), and French defense (advanced variation, with 6. a3) openings. The second cluster from the right in Figure 2 consists of games that use the French defense (advanced variation, Paulsen attack), and French defense (advanced variation, with 6. a3) openings and even in Figure 3, chess games that use the French defense variation are scattered across the three clusters on the right.

Despite the similarities between the French defense opening variations, the other two openings, Queen's Gambit Declined and English (symmetrical variation, main line), are two openings that have different move choices compared to the French defense which starts with 1. e4. The Queen's Gambit declined opening starts with 1. d4 and the English (symmetrical variation, main line) opening starts with the move 1. c4. Of the four clustering methods, only hierarchical clustering complete linkage and k-means clustering perfectly recognized the differences. Although not perfect, hierarchical clustering average linkage also managed to recognize the difference. Meanwhile, hierarchical clustering single linkage can be considered a failure in distinguishing chess games based on the opening.

On the other hand, the failure of single linkage hierarchical clustering can also be an interesting knowledge. It is a public secret that although a chess game starts with different moves, it can progress to the same position. Moreover, with the characteristic of single linkage that combines adjacent data points, it can be said that this clustering method looks more at the similarity between games rather than the clustering itself. This result then allows us to see more details about the similar games.

Of all the clustering methods implemented, k-means was the most successful in clustering this chess game based on its original opening category. K-means clustering was even able to recognize the difference between French defense (advance variation, Steinitz variation) and French defense (advance variation, Paulsen attack and with 6.a3). This is very likely due to the characteristics of k-means which prioritizes the proximity of data points to the cluster center. This is similar to hierarchical clustering complete linkage which ensures that a data point is close to a cluster by ensuring that this data point is close enough to the distance of its furthest data point. Hence, hierarchical clustering complete linkage produces a clustering similar to k-means, but due to the high similarity of games using French defense, this method fails to recognize the difference.

4. Conclusion

Chess is a game across the centuries that requires tactics and strategy to play. However, apart from developing chess engines, there are not many

applications of machine learning in this field especially for learning tactics and strategies in chess games. With the availability of a database of chess games in text form, text analysis methods can be applied to uncover the knowledge hidden in chess games. From the results of the experiments conducted, the computation of similarity is more suitable for text rather than directly computing the distance between texts. It will also make it easier to interpret the relationship between texts.

In conclusion, the research plan conducted in this study is quite successful in clustering and understanding the similarity of chess games. K-means clustering and hierarchical clustering complete linkage performed well in clustering chess games and hierarchical clustering single linkage is helpful in discovering the similarity of chess games. While hierarchical clustering average linkage is somewhere in between. This further opens up the potential of using machine learning methods to explore the hidden knowledge in chess games.

Despite its success, this research has potential drawbacks. The use of simple algorithms such as hierarchical clustering and k-means will most likely underperform advanced algorithms such as DBSCAN [10], BIRCH [15], support vector clustering [16], etc. It is possible to apply and compare these algorithms in future research. Furthermore, as with data diversity, it would be interesting to see the results analyzed with a larger amount of data and a more diverse opening.

Reference

[1] Tim Mann's Chess Pages, "PGN: Portable Game Notation," https://tim-mann.org/Standard. Accessed: Jun. 10, 2024.

[2] R. Janani and S. Vijayarani, "Text document clustering using Spectral Clustering algorithm with Particle Swarm Optimization," Expert Systems with Applications, vol. 134, pp. 192-200, 2019. https://doi.org/10.1016/j.eswa.2019.05.030.

[3] H. Kim, H. K. Kim, and S. Cho, "Improving spherical k-means for document clustering: Fast initialization, sparse centroid projection, and efficient cluster labeling," Expert Systems with Applications, vol. 150, p. 113288, 2020. https://doi.org/10.1016/j.eswa.2020.113288.

[4] H. Rehioui and A. Idrissi, "New clustering algorithms for twitter sentiment analysis," IEEE Systems Journal, vol. 14, no. 1, pp. 530-537, 2019. https://doi.org/10.1109/JSYST.2019.2912759

[5] S. Riaz, M. Fatima, M. Kamran, and M. W. Nisar, "Opinion mining on large scale data using sentiment analysis and k-means clustering," Cluster Computing, vol. 22, pp. 7149-7164, 2019. https://doi.org/10.1007/s10586-017-1077-z

[6] I. Manipur, I. Granata, L. Maddalena, and M. R. Guarracino, "Clustering analysis of tumor metabolic networks," BMC Bioinformatics, vol. 21, pp. 1-14, 2020. https://doi.org/10.1186/s12859-020-03564-9

[7] M. Patel, H. Pandey, T. Wagh, A. D. Hujare, and R. Dangi, "Vecma: An advance chess engine," in 2022 IEEE Pune Section International Conference (PuneCon), 2022, pp. 1-6.

[8] H. Zhang and T. Yu, "AlphaZero," in Deep Reinforcement Learning, H. Dong, Z. Ding, and S. Zhang, Eds. Singapore: Springer, 2020. doi: 10.1007/978-981-15-4095-0_15.

[9] R. McIlroy-Young, R. Wang, S. Sen, J. Kleinberg, and A. Anderson, "Learning models of individual behavior in chess," in Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 1253-1263. https://doi.org/10.1145/3534678.3539367

[10] Raghav, Kuldeep, and Laxmi Ahuja. "Chess Opening Analysis Using DBSCAN Clustering and Predictive Modeling." In 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), pp. 1-5. IEEE, 2024.

[11] G. Gan, C. Ma, and J. Wu, Data Clustering: Theory, Algorithms, and Applications. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2020.

[12] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," Information Sciences, vol. 622, pp. 178-210, 2023. https://doi.org/10.1016/j.ins.2022.11.139

[13] A. Subayu, "Penerapan Metode K-Means Untuk Analisis Stunting Gizi Pada Balita: Systematic Review", SNATI, vol. 2, no. 1, pp. 42–50, Jul. 2022.

[14] M. Meyer and K. Buchta, "proxy: Distance and Similarity Measures," R package version 0.4-27, Comprehensive R Archive Network (CRAN), 2023. [Online]. Available: https://cran.r-project.org/web/packages/proxy/. [Accessed: Jun. 10, 2024].

[15] I. Pauletic, L. N. Prskalo and M. B. Bakaric, "An Overview of Clustering Models with an Application to Document Clustering," 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2019, pp. 1659-1664, https://doi.org/ 10.23919/MIPRO.2019.8756868.

[16] M. Babaei, S. M. Muyeen and S. Islam, "Identification of Coherent Generators by Support Vector Clustering With an Embedding Strategy," in IEEE Access, vol. 7, pp. 105420-105431, 2019, https://doi.org/10.1109/ACCESS.2019.2932194.