JISS
Journal of Indonesian Social Science

Vol. 5, No. 4, April 2024
E-ISSN: 2723-6692
P-ISSN: 2723-6595
http://jiss.publikasiindonesia.id/

Implementation of Vector-Based Melody Extraction for Plagiarism Detection Using Szymkiewicz-Simpson Coefficient

Nindyo Artha Dewantara Wardhana, Agung Mulyo Widodo*,
Gerry Firmansyah, Budi Tjahjono
Universitas Esa Unggul, Jakarta Selatan, Indonesia
Email: agung.mulyo@esaunggul.ac.id
Correspondence: agung.mulyo@esaunggul.ac.id*

KEYWORDS
Communication Sociology; Communication; Mass Media

ABSTRACT
Plagiarism is topical within the music industry. It is filled with circumstances such as the potential of massive losses coupled with a “false-positive” court ruling due to the blurred line of plagiarism factor. This research aims to solve the gray line of music plagiarism by exploring the potential of the Szymkiewicz-Simpson coefficient toward musical aspects of music. Melody and Rhythm are chosen as the main features to focus on in the research. MIDI files of music involved in court cases are used as data for the study, with limitations put on what cases can be used for the research. Using a threshold range of 0.1 to 0.25, detection accuracies for melodic plagiarism range from 45% to 60%, while rhythm plagiarism ranges from 60 to 65%. This shows that the algorithm of plagiarism detection has a tendency to detect non-plagiarism cases and is more effective towards rhythm plagiarism detection rather than melodic plagiarism detection against existing plagiarism cases.

1. Introduction

Music is a universally recognized art of auditory expression of emotion. Along with the development of technology, one's method of expression using music has become increasingly affordable. As a result, the music industry has experienced significant revenue. It was reported by the International Federation of Phonographic Industry (IFPI) that in 2022, revenue from music recordings globally will reach US$ 26 billion. This is a 9% increase from the previous year (2021) and is the 8th consecutive year of growth (Cameron, 2020; International Federation of Phonographic Industry, 2023; Savage et al., 2021).

Since music is a medium for expression, it is also vulnerable to being copied by others. Therefore, music plagiarism has become a hot topic in the music industry. In practice, music plagiarism is not black and white. It is because music is the result of ideas and creativity inspired by other works. As a result, the approach taken in music plagiarism cases is generally conducted on a case-by-case basis and evaluated by music experts, giving complexity to the determination of music plagiarism claims (Cameron, 2020; Gjorgjioska & Gligorovski, 2023; Pidhayna, 2022).

In court cases, court settlements for music plagiarism can range from the payment of a fixed amount to the assignment of royalties on the music to the plaintiff (in percentages, where 100% is the maximum). One such case involved "I do not give a fuck" by Tulisa Contostavlos, who sued "Scream and Shout" by Will.i.am and Britney Spears in 2012. The decision of this case ordered will. i.am and Britney Spears paid 10% of the royalties from the song to Tulisa, who made her a co-writer (Maine, 2018). Other than this case, there is also the case of The Rolling Stones' song "The Last Time" with The Verve's "Bittersweet Symphony." This plagiarism case took place in 1999, with the ruling that The Verve should make The Rolling Stones the songwriter, as well as 100% of the song's royalties. Although the decision was reversed in 2019, during this time, The Verve suffered a loss of US$5 million (Spencer, 2023; Tsioulcas, 2019).

The case of “Bittersweet Symphony” introduces the possibility of a “false positive” plagiarism case, signifying that a case may be ruled as plagiarism even though it is not necessarily similar. Such a fact sparked a debate within the case of "Blurred Lines" by Robin Thicke and Pharrell Williams, who was accused of plagiarizing "Got to Give It Up" by Marvin Gaye's estate. From the public perspective, Robin Thicke’s music was not a direct plagiarism towards Marvin Gaye. However, instead, it took inspiration from Marvin Gaye’s music, which is in the same genre. This particular case incites discussion on the line of plagiarism between two pieces of music (Stempel, 2018).

The fact that "false-positive" musical plagiarism is possible, coupled with the fact that the potential for significant losses and a seemingly gray line of plagiarism, lowers the confidence of musicians to write new music (Rolling Stone, 2020). Moreover, the problem's root cause lies in the question of “What is considered plagiarism in music?” and “Where does the line of plagiarism is drawn?”. It is worth noting that cases of music plagiarism are determined by music experts in court, which means it is not 100% objective. With that in mind, the need to define an objective metric for music plagiarism verdicts is paramount to prevent “false positive” musical plagiarism.

This research aims to explore the potential of music plagiarism detection using similarity detection with the Szymkiewicz-Simpson coefficient, commonly known as the "Overlap Coefficient." The features used for this research will focus on the melodic and rhythmic elements of the music. This research is expected to contribute to the development of music plagiarism detection research by providing melodic and rhythmic-based metric solutions to music plagiarism.

This paper will divide the structure into five distinct sections. It started with an overview of facts, problems, research motivation, and contribution of the research that will be laid out in Section 1. Afterward, a review of related literature regarding the research will follow in the next section. The section afterward will discuss the systematic use of the research method, which will comprise data collection, pre-processing, formula, and evaluation. The following section will showcase the results of the research and provide a discussion of the results. In the final section, a conclusion will be drawn, and the potential of further research will be detailed.

2. Materials and Methods

From the preceding, a solution for research is proposed. The proposed solution is a vector-based algorithm for melody and rhythm elements using similarity coefficients. Melody is chosen because previous music plagiarism studies have used melody as the main feature for determining music similarity. In contrast, rhythm is selected because this feature is rarely used as a factor in deciding music similarity but is considered to affect the results of music plagiarism identification.

The evaluate melody and rhythm simultaneously, the method will use vectors as feature containers. To be able to process the vector data and output the similarity information, the chosen algorithm must be able to compare the two music to be checked. Similarity coefficients were selected because they support the comparison of two data sets and can be used with vector data.

The Szymkiewicz-Simpson coefficient (Overlap Coefficient) was chosen as the similarity coefficient. The selection of this coefficient is considered to provide optimal output due to its better performance compared to other similarity algorithms such as Jaccard, Sorensen-Dice, Kulczynski, Otsuka-Ochiai, or Braun-Blanquet. Szymkiewicz-Simpson, compared to Kulczynski, Otsuka-Ochiai, and Sorensen-Dice, have approximately the same accuracy. Szymkiewicz-Simpson was chosen because of its slightly higher accuracy than the other three based on research by Korepanova et al. (2020). In addition to its higher accuracy, the Szymkiewicz-Simpson algorithm is also considered more sensitive to data similarity between two given samples (Gianino et al., 2021; Hanley et al., 2022).

To test for plagiarism, a concrete case of musical plagiarism is needed. This court case must have the following criteria: (1) the public can see the results of the court. Generally, the results of this court will be reported by the media so that the results can be considered as open to the public; (2) the case must have been closed, for cases that are still ongoing will be excluded because the results can change at any time; (3) a closed settlement will be excluded, this is because generally this settlement is carried out not based on whether or not there is plagiarism, but based on the reluctance of the accused party to go to court.

Data Collection

In addition to information about the plagiarism case, it is also necessary to search for music data from the case. The music data must be in MIDI (.mid) format so that it can be processed by the algorithm naturally. There are various data sources for MIDI files on the internet. Generally, MIDI file data sources spread across the internet provide MIDI files that are transcriptions from third parties, not from the musicians directly. However, sources like Musescore (https://musescore.com) provide original copies of the music directly for some music. In addition to original copies from the musicians directly, transcriptions from various third parties are also available.

The MIDI files taken from sources like this, several criteria must be met. Criteria such as (1) The music can be searched on MIDI file data source platforms spread across the internet. If one of the pieces of music involved in the plagiarism case cannot be found, then the plagiarism case cannot be used in the music plagiarism detection test; (2) Both pieces of music involved in the plagiarism case must be obtained from the same source. If one of the music involved is not found in that source while the other one can be found, then both music will be searched in another MIDI file provider platform; and (3) The musician's original copy should be favored. If the musician's original copy is not found, then a third-party copy may be used as a substitute.

From the previously described criteria, 20 (20) cases of music plagiarism were found to fit the requirements and ranged in the years from 1966 to 2020.

Data processing

First, the tracks from the collected MIDI files need to be cleaned. MIDI files usually contain more than one audio track that corresponds with different kinds of melodies and rhythms. To use the MIDI file for this research, the main melody track must be selected first, and tracks other than the main melody must be removed.

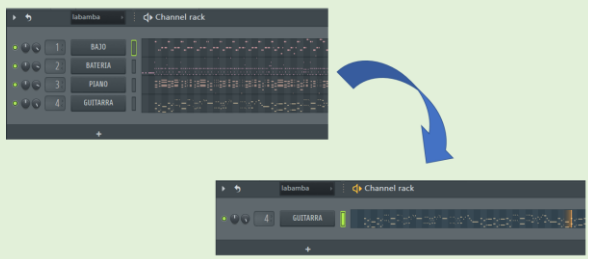

Figure 1 Visualization of MIDI Track Cleaning

After the MIDI tracks were processed, the melody's time signature was extracted from the MIDI files by slicing. Measure slicing is done by iteratively pulling two measures from each piece of music. These measures will be compared with each other in terms of the music involved in the plagiarism case.

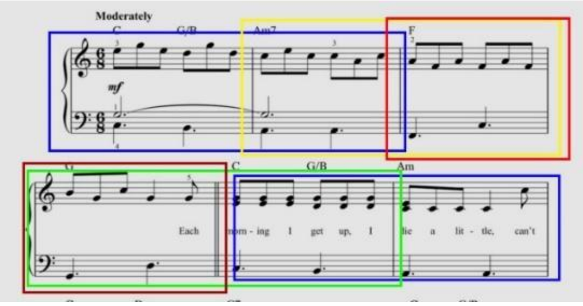

Figure 2 Example of Measure Slicing

Once the measures to be examined have been separated, they will be extracted for their melodies. The melody of each beat will be represented as a vector that summarizes information such as the pitch, duration, octave, and interval of each note. To calculate the interval of a note, the formula interval (ni, ni+1) = ni+1 - ni is used, where ni represents the note being played, and ni+1 is the following note. If one of the notes is a resting note, the interval (ni, ni+1) is equal to 0. The note duration representation used is in decimal form according to the rhythm presentation.

The Szymkiewicz-Simpson formula detects the similarity between the separated vectors. Each melody vector is evaluated individually, resulting in a degree of similarity between 0 and 1, with 0 being very different and one being very identical (De Prisco et al., 2017; Korepanova et al., 2020).

Note that duration needs further attention as a form of rhythm. Previous studies on music plagiarism have paid little attention to rhythm as a calculation object. Commonly used note durations are 4/4 (complete note), 2/4 (half note), 1/4 (quarter note), 1/8 (eighth note), 1/16 (sixteenth note), and so on (Rohrmeier, 2020; Schuitemaker et al., 2020).

Formula

The algorithm to be used to process the existing vectors is the Szymkiewicz-Simpson coefficient. Each melody vector will be evaluated one by one, resulting in a fuzzy degree consisting of 0 and 1, with 0 being very different and one being very identical (De Prisco et al., 2017).

$$oc(A,B) = \frac{|A \cap B|}{(A,B)} \quad (1)$$

$oc(A,B)$	:	Overlap Coefficient on Vector A and B
A	:	Vector A
B	:	Vector B

We already have the melody vector and rhythm vector available from the previous step. These two vectors can be applied to the Szymkiewicz-Simpson similarity coefficient to get a better similarity coefficient. We have the following equation to measure melodic and rhythmic similarity. Note that this equation needs to be applied for each available time signature selected by the sliding window in the previous step. Here is the implementation of the Szymkiewicz-Simpson similarity coefficient on the melody vector and rhythm vector.

$$oc(Vm_A, Vm_B) = \frac{|Vm_A \cap Vm_B|}{min\ (Vm_A\ , Vm_B)} \quad (2)$$

$oc(Vm_A, Vm_B)$	:	Overlap Coefficient for Melodic Vector A and B
Vm_A	:	Melodic Vector A
Vm_B	:	Melodic Vector B

$$oc(Rm_A, Rm_B) = \frac{|Rm_A \cap Rm_B|}{(Rm_A\ , Rm_B)} \quad (3)$$

$oc(Rm_A, Rm_B)$	:	Overlap Coefficient for Rhythm Vector A dan B
Rm_A	:	Rhythm Vector A
Rm_B	:	Rhythm Vector B

Once the similarity results are found, they need to be averaged. The use of averaging here serves to combine the similarity values generated from the previous calculations. By applying the Szymkiewicz-Simpson coefficient to the averaging formula, the following equation is generated. Where "a" is equal to the minimum number of notes in Melody Vectors A and B. "b" is equal to the least number of comparisons of Melody Vectors A and B. This equation will also be applied to rhythmic similarity.

$$\underline{oc(Vm_A, Vm_B)} = \frac{\sum_a^b \quad oc(Vm_A,Vm_B)}{b-a+1} \tag{4}$$

$\underline{oc(Vm_A, Vm_B)}$	:	The average equation for Melodic Vector Overlap Coefficient
$\sum_a^b \quad oc(Vm_A, Vm_B$	:	Sum of Melodic Vector Overlap Coefficient
a	:	Starting Index
b	:	End Index

$$\underline{oc(Rm_A, Rm_B)} = \frac{\sum_a^b \quad oc(Rm_A,Rm_B)}{b-a+1} \tag{5}$$

$\underline{oc(Rm_A, Rm_B)}$	:	The average equation for the Rhythm Vector Overlap Coefficient
$\sum_a^b \quad oc(Rm_A, Rm_B$	:	Sum of Rhythm Vector Overlap Coefficient
a	:	Starting Index
b	:	End Index

Algorithm

To support the research conducted, a program was designed to drive the calculations to be performed. To achieve the expected testing, algorithms are formed as a basis for program design. The following are the main algorithms used in the designed program. The algorithm will receive two sets of measures of the music to be compared. The time signature of both music will be sliced by six notes because, based on research conducted by (Schuitemaker et al., 2020), six sequences of melody are considered plagiarism. However, in practice, melodies only sometimes consist of 6 or more sequences. Therefore, this algorithm provides a failsafe for conditions where the number of notes in the melody is less than six sequences.

Table 1 Algorithm of Szymkiewicz-Simpson Coefficient for Plagiarism Detection

	Szymkiewicz-Simpson Algorithm
1	Load data a and b
2	Find the minimum value of the size of a and b
3	If minimum value <= 6, then slicing frame size = minimum value / 2 rounded up.
4	Each piece of data in a will be compared with each piece of data in b.
5	The data of a and b will be sliced equal to the frame sizes each.
6	If there is an overlap between a and b, overlap + 1. If overlaps have previously been detected, skip to the next step.
7	If a and b are still not empty, then go back to step 5

8	overlap / minimum value, resulting in the degree of similarity.

To access the algorithm showcased above, another algorithm is designed to load the data that will be processed. The following algorithm will perform the plagiarism detection module algorithm and evaluate the detection.

Table 2 Algorithm of Plagiarism Detection Process

	Plagiarism Detection Algorithm
1	Set threshold variable number
2	Track True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) for each melodic and rhythm detection
3	Load court case data and associated MIDI files
4	Examine court case data
5	Input MIDI files related to court case data into the plagiarism detection module algorithm, which generates melody and rhythm plagiarism detection values.
6a	If melody/rhythm plagiarism detection value < threshold, detection is considered non-plagiarism.
6b	If melody/rhythm plagiarism detection value > threshold, detection is considered non-plagiarism.
7a	If detection results in plagiarism and the court case outcome is plagiarism, then TP + 1
7b	If detection results in plagiarism and the court case outcome is non-plagiarism, then FP + 1
7c	If detection results in non-plagiarism and the court case outcome is non-plagiarism, then TN + 1
7d	If detection results in non-plagiarism and the court case outcome is plagiarism, then FN + 1
8	If there are still court cases that have not been examined, return to Step 4
9	Using TP, FP, TN, and FN values for melody/rhythm, calculate the confusion matrix consisting of accuracy, precision, recall, and F-measure.

Evaluation

The result of this music similarity calculation algorithm will show a value between 0 and 1, where 0 indicates that there is nothing similar, and 1 indicates that both music are the same. Due to the lack of research on music plagiarism detection algorithms, the threshold value for plagiarism detection still needs concrete value. This threshold serves as a boundary between plagiarism and non-plagiarism decisions, where detection values smaller than the threshold will be inferred as non-plagiarism, and detection values more significant than the threshold will be inferred as plagiarism. In an effort to find the optimal threshold value for this algorithm, the range value for the threshold value of the similarity calculation result will be determined first. The range value is taken from the average melody plagiarism detection value, which is then spread twice up and twice down using an interval of 0.05.

Testing is carried out on all lawsuit case data collected previously. After the test results have been collected, they will be evaluated. The evaluation will be done through an accuracy test, precision test, recall test, and F-measure test to determine the algorithm's performance in music plagiarism detection.

3. Result and Discussion

For the test conducted in this research, a range of threshold points is determined. The range value is taken from the average melodic plagiarism detection value, which is then expanded twice upwards and twice downwards using an interval of 0.05. The result of the average value of musical plagiarism detection is 0.1665. Using 0.05 intervals, the range values for the threshold points were set as 0.1 and 0.15 for downward values of 0.1665 and 0.2 and 0.25 for upward values. Therefore, it was determined that the threshold values to be used for algorithm evaluation were thresholds from the range of 0.1 to 0.25, with an interval of 0.05.

The melody plagiarism test is conducted on the pitch interval data in the melody of the examined song. From the test, information will be collected in the form of the number of True Positive, False Positive, True Negative, and False Negative. The four pieces of information obtained will be processed to obtain the accuracy value (how precise the plagiarism detection results are), precision value (the tendency of the algorithm to detect a case as a plagiarism case), recall value (the tendency of the algorithm to detect a case as non-plagiarism), and F-measure value (the balance value between precision and recall).

In testing 20 cases, the following results were obtained.

Table 3 Result of Melodic Plagiarism Detection

Case Year	Plaintiff	Suspect	Melody Detection Value	Court Verdict	Detection Result on Thresholds			
					0.1	0.15	0.2	0.25
1966	"Sweet Little Sixteen"	"Surfin' U.S.A." (1963)	0.04	1	X	X	X	X
1968	"All Day and All of the Night" (1964)	"Hello, I Love You" (1968)	0.23	1	√	√	√	X
1971	"He's So Fine" (1963)	"My Sweet Lord" (1970)	0.49	1	√	√	√	√
1973	"Speedy Gonzales" (1962)	"Crocodile Rock" (1972)	0.43	1	√	√	√	√
1988	"Just Another Night"	"Just Another Night" (1985)	0.12	0	X	√	√	√
1988	"Run Through the Jungle" (1970)	"The Old Man Down the Road" (1984)	0.03	0	√	√	√	√
2010	"Kookaburra" (1932)	"Down Under" (1980)	0.44	1	√	√	√	√
2012	"Baby I'm Yours" (2010)	"Treasure" (2012)	0.07	1	X	X	X	X
2013	"Got to Give It Up"(1977)	"Blurred Lines" (2013)	0.02	1	X	X	X	X
2015	"Oops Up Side Your Head" (1979)	"Uptown Funk" (2014)	0.52	1	√	√	√	√
2015	"Takin' Me to Paradise" (1983)	"The Most Beautiful Girl in the World" (1994)	0.03	1	X	X	X	X
2015	"I Won't Back Down" (1989)	"Stay With Me" (2014)	0	1	X	X	X	X

2016	"Taurus" (1968)	"Stairway to Heaven" (1971)	0.09	0	√	√	√	√
2017	"Playas Gon' Play" (2001)	"Shake It Off" (2014)	0.31	0	X	X	X	X
2018	"The Man Who Can't Be Moved" (2008)	"Say You Won't Let Go" (2016)	0.04	1	X	X	X	X
2018	"Let's Get It On" (1973)	"Thinking Out Loud" (2014)	0.01	0	√	√	√	√
2018	"Seven Nation Army" (2003)	"Toy" (2018)	0.05	1	X	X	X	X
2018	"Oh Why" (2015)	"Shape of You" (2017)	0.18	0	X	X	√	√
2019	"Holly Wood Died" (2006)	"Lucid Dreams" (2018)	0.07	0	√	√	√	√
2020	"Sunrise" (2018)	"Pray for Me" (2018)	0.16	0	X	X	√	√
True Positive					5	5	5	4
False Positive					7	7	7	8
True Negative					4	5	7	7
False Negative					4	3	1	1
Accuracy					45.0 %	50.0 %	60.0 %	55.0 %
Precision					42.0 %	42.0 %	42.0 %	33.0 %
Recall					50.0 %	62.0 %	88.0 %	88.0 %
F-Measure					45.0 %	50.0 %	56.0 %	48.0 %

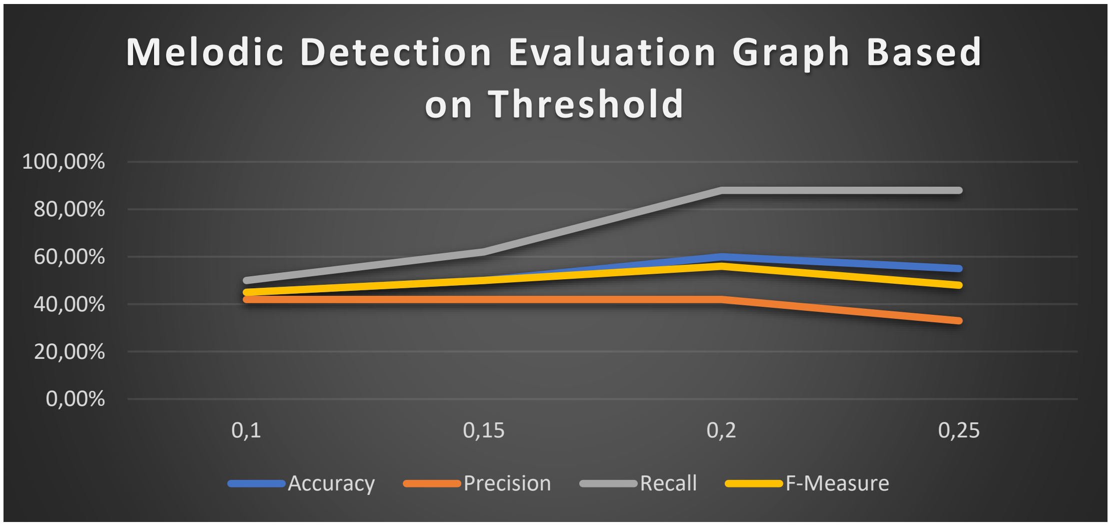

Figure 3 A graph of Melodic detection performance

Aside from a table showcasing the result of Melodic Detection and the evaluation for each threshold point, a graph depicting the performance evaluation of melodic plagiarism detection for each threshold point is shown.

In addition to melodic testing, the data was also tested rhythmically. Rhythmic plagiarism testing is performed on the pitch duration data in the melody of the examined song. From the tests conducted, information will be collected in the form of the number of True Positive, False Positive, True Negative, and False Negative. The four pieces of information obtained will be processed to obtain the accuracy value (how precise the plagiarism detection results are), precision value (the tendency of the algorithm to detect a case as a plagiarism case), recall value (the tendency of the algorithm to detect a case as non-plagiarism), and F-measure value (the balance value between precision and recall).

In testing 20 cases, the following results were obtained.

Table 4 Result of Rhythm Plagiarism Detection

Case Year	Plaintiff	Suspect	Rhythm Detection Value	Court Verdict	Detection Result on Thresholds			
					0.1	0.15	0.2	0.25
1966	"Sweet Little Sixteen"	"Surfin' U.S.A." (1963)	0.21	1	√	√	√	X
1968	"All Day and All of the Night" (1964)	"Hello, I Love You" (1968)	0.36	1	√	√	√	√
1971	"He's So Fine" (1963)	"My Sweet Lord" (1970)	0.55	1	√	√	√	√
1973	"Speedy Gonzales" (1962)	"Crocodile Rock" (1972)	0	1	X	X	X	X
1988	"Just Another Night"	"Just Another Night" (1985)	0.08	0	√	√	√	√
1988	"Run Through the Jungle" (1970)	"The Old Man Down the Road" (1984)	0	0	√	√	√	√
2010	"Kookaburra" (1932)	"Down Under" (1980)	0	1	X	X	X	X
2012	"Baby I'm Yours" (2010)	"Treasure" (2012)	0.3	1	√	√	√	√
2013	"Got to Give It Up"(1977)	"Blurred Lines" (2013)	0	1	X	X	X	X
2015	"Oops Up Side Your Head" (1979)	"Uptown Funk" (2014)	0.73	1	√	√	√	√
2015	"Takin' Me to Paradise" (1983)	"The Most Beautiful Girl in the World" (1994)	0.47	1	√	√	√	√
2015	"I Won't Back Down" (1989)	"Stay With Me" (2014)	0.44	1	√	√	√	√
2016	"Taurus" (1968)	"Stairway to Heaven" (1971)	0.43	0	X	X	X	X
2017	"Playas Gon' Play" (2001)	"Shake It Off" (2014)	0.01	0	√	√	√	√

2018	"The Man Who Can't Be Moved" (2008)	"Say You Won't Let Go" (2016)	0.13	1	√	X	X	X
2018	"Let's Get It On" (1973)	"Thinking Out Loud" (2014)	0	0	√	√	√	√
2018	"Seven Nation Army" (2003)	"Toy" (2018)	0	1	X	X	X	X
2018	"Oh Why" (2015)	"Shape of You" (2017)	0	0	√	√	√	√
2019	"Holly Wood Died" (2006)	"Lucid Dreams" (2018)	0.23	0	X	X	X	√
2020	"Sunrise" (2018)	"Pray for Me" (2018)	0.43	0	X	X	X	X
True Positive					8	7	7	6
False Positive					4	5	5	6
True Negative					5	5	5	6
False Negative					3	3	3	2
Accuracy					65.0 %	60.0 %	60.0 %	60.0 %
Precision					67.0 %	58.0 %	58.0 %	50.0 %
Recall					62.0 %	62.0 %	62.0 %	75.0 %
F-Measure					65.0 %	60.0 %	60.0 %	60.0 %

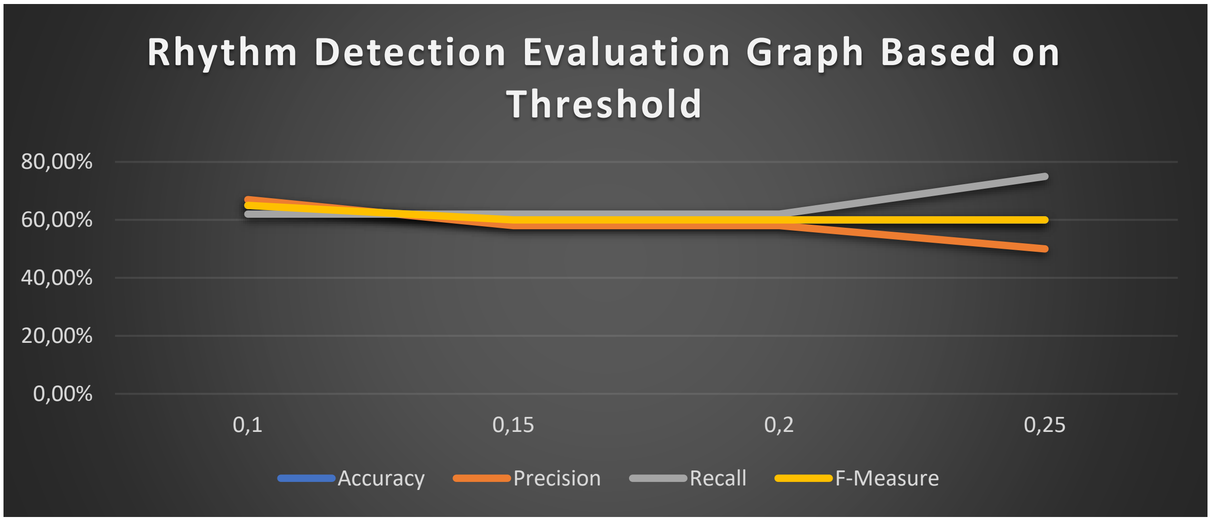

Figure 4 A graph of Rhythm Detection Performance

Apart from than showing a table of the rhythm detection result and the evaluation for each threshold point, a graph depicting the performance evaluation of the rhythm plagiarism detection for each threshold point is shown.

Based on the results of plagiarism detection testing using the Szymkiewicz-Simpson coefficient on the dataset used, the results of similarity detection of two pieces of music melodically and rhythmically are obtained. The results of this similarity detection were then tested against the limit value between 0.1 and 0.25. This test is conducted to find which limit value is the most optimal as a threshold value between plagiarism and non-plagiarism.

From the tests conducted, information was obtained regarding the precision and recall values of melodic data. From this information, it can be concluded that the algorithm tends to detect non-plagiarism cases optimally compared to plagiarism cases. The high recall value indicates this compared to the precision value.

For tests conducted on melody data, it is found that the accuracy is in the range of 45% to 60%, with the most optimal accuracy achieved by a limitation value of 0.2. From testing the melody, it is known that the precision value is in the range of 33% to 42%, while the recall value is in the range of 50% to 88%. From the results obtained by testing, the accuracy produced by melody detection could be better. This can be interpreted into two conclusions. The first conclusion is that the Szymkiewicz-Simpson coefficient is not effective enough to detect melodic musical plagiarism. The second conclusion is that there are extreme metric differences in the determination of melodic plagiarism in court.

The second test was conducted on rhythm data. From the test, it was found that the accuracy was in the range of 60% to 65%. The most optimal accuracy was achieved by using a constraint value of 0.1. From testing the rhythm data, information was obtained that the precision value was in the range of 50% to 67%, and the recall value was in the range of 62% to 75%. From these results, the detection of rhythm plagiarism between music has a reasonably high accuracy. Compared to melodic detection, this shows that using Szymkiewicz-Simpson coefficients as an algorithm for musical plagiarism detection is more effective for rhythmic data than melodic data.

4. Conclusion

In this research, a method for music plagiarism detection is investigated. This plagiarism detection method uses the Szymkiewicz-Simpson similarity coefficient, also known as the Overlap Coefficient. This research was designed by taking court cases as the standard for plagiarism detection and optimizing the detection performance based on these cases. For this research, 20 court cases were used for algorithm evaluation. This research uses MIDI files as the primary data. The results of this research show that melody detection of existing plagiarism cases is less than optimal, with accuracy ranging from 45% to 60%. This plagiarism detection tends to detect non-plagiarism cases due to its higher recall value than precision value. Meanwhile, rhythm plagiarism detection against existing plagiarism cases has a more optimal performance than melodic plagiarism detection, with accuracy ranging from 60% to 65%. Rhythm plagiarism detection has balanced precision and recall values. The better performance of rhythm detection compared to melody detection leads to the conclusion that this music plagiarism detection algorithm using Szymkiewicz-Simpson coefficients is more effective for rhythm plagiarism detection. In this research, MIDI files are used as the primary data for testing with the aim of obtaining complete and accurate melody information. The MIDI file is cleaned, and only one track is left for the main melody. However, in practice, plagiarism in music generally does not focus on just one type of melody. Further research into the implementation of this algorithm can take the form of research into the inclusion of other melody tracks and their impact on the performance of this algorithm. Besides melody and rhythm, lyrics are another factor of concern in music plagiarism. Lyrics are excluded in this research due to their complexity, such as their textual nature. Not all music has lyrics, and it may consist of various languages. For future research, the use of this algorithm on music lyrics data can be further investigated. In addition to the development in the testing method, further development can be done on the amount of data used for testing. The test in this study used 20 court cases, the number of which can affect the performance evaluation of this algorithm. This can be further investigated by lifting some of the limitations given at the time of data collection.

5. References

Cameron, S. (2020). *An Economic Approach to the Plagiarism of Music*. Springer International Publishing. https://doi.org/10.1007/978-3-030-42109-0

De Prisco, R., Malandrino, D., Zaccagnino, G., & Zaccagnino, R. (2017). Fuzzy Vectorial-based Similarity Detection of Music Plagiarism. *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–6. https://doi.org/10.1109/FUZZ-IEEE.2017.8015655

Gianino, M. M., Savatteri, A., Politano, G., Nurchis, M. C., & D. Pascucci, G. (2021). Burden of COVID-19: Disability-Adjusted Life Years (DALYs) across 16 European countries. *European Review Medical Pharmacologi Sciences*, *25*(17), 5529–5541. https://doi.org/10.26355/eurrev_202109_26665

Gjorgjioska, E., & Gligorovski, V. (2023). Code of Ethics in Scientific Papers: The Plagiarism Issue. Knowledge International Journal, 61(1), . *Knowledge - International Journal*, *6*(1), 319–323.

Hanley, H. W. A., Kumar, D., & Durumeric, Z. (2022). No Calm in the Storm: Investigating QAnon Website Relationships. *Proceedings of the International AAAI Conference on Web and Social Media*, *16*, 299–310. https://doi.org/10.1609/icwsm.v16i1.19293

International Federation of Phonographic Industry. (2023). *Global Music Report 2023*. IFPI: International Federation of Phonographic Industry. https://www.ifpi.org/wp-content/uploads/2020/03/Global_Music_Report_2023_State_of_the_Industry.pdf

Korepanova, A. A., Oliseenko, V. D., & Abramov, M. V. (2020). Applicability of Similarity Coefficients in Social Circle Matching. *2020 XXIII International Conference on Soft Computing and Measurements (SCM)*, 41–43. https://doi.org/10.1109/SCM50615.2020.9198782

Maine, S. (2018). *Tulisa has won her legal battle against Will.i.am and Britney Spears: The five-year court battle over "Scream And Shout" has come to an end*. NME Networks. https://www.nme.com/news/music/tulisa-will-i-am-britney-spears-lawsuit-2281544

Pidhayna, A. (2022). Plagiarism: Problems and Influence on Music World. *Інноваційні Тенденції Підготовки Фахівців в Умовах Полікультурного Та Мультилінгвального Глобалізованого Світу. Київський Національний Університет Технологій Та Дизайну*, 140–142.

Rohrmeier, M. (2020). Towards A Formalizatio of Musical Rhythm. *Ismir*, 621–629.

Rolling Stone. (2020). *How Music Copyright Lawsuits Are Scaring Away New Hits*. Rolling Stone. Https://Www.Rollingstone.Com/pro/Features/Music-Copyright-Lawsuits-Chilling-Effect-935310/

Savage, P. E., Loui, P., Tarr, B., Schachner, A., Glowacki, L., Mithen, S., & Fitch, W. T. (2021). Music as a coevolved system for social bonding. *Behavioral and Brain Sciences*, *44*, e59. https://doi.org/10.1017/S0140525X20000333

Schuitemaker, N., Adriaans, F., & Dotlacil, J. (2020). *An Analysis of Melodic Plagiarism Recognition using Musical Similarity Algorithms*. https://www.youtube.com/watch?v=oOlDewpCfZQ

Spencer, A. (2023, September 28). *The Hit Song That Cost The Verve $5 Million After Their Lawsuit With The Rolling Stones*. The Things. https://www.thethings.com/the-verve-song-lawsuit-bittersweet-symphony-with-the-rolling-stones/

Stempel, J. (2018, March 22). *Marvin Gaye family prevails in "Blurred Lines" plagiarism case*. Reuters. https://www.reuters.com/article/us-music-blurredlines-idUSKBN1GX27P/

Tsioulcas, A. (2019, May 23). *Not Bitter, Just Sweet: The Rolling Stones Give Royalties To The Verve*. Npr News. https://www.npr.org/2019/05/23/726227555/not-bitter-just-sweet-the-rolling-stones-give-royalties-to-the-verve