A Comparative Study of Convolutional Neural Networks and Recurrent Neural Networks for Chord Recognition

Hania Nawaz Khan[1], Sibghatullah Bazai[2*], Zubair Zaland[3], Sibghatullah Durrani[4], Saad Aslam[5], Angela Amphawan[6],Fatima Ali[7], Tse-Kian Neo[8]

[1,2,3,4,7]*Department of Software Engineering, Balochistan University of Information, Technology, Engineering and Management Sciences, Quetta, 87650 Pakistan*

[5,6]*Smart Photonics Research Laboratory and Department of Computing and Information Systems, School of Engineering and Technology, Sunway University, 47500 Petaling Jaya, Malaysia.*

[8]*CAMELOT, Faculty of Creative Multimedia, Multimedia University, Cyberjaya 63100, Selangor, Malaysia*
**Corresponding author: tkneo@mmu.edu.my*

Abstracts: Using Mel-spectrograms, this study evaluates the effectiveness of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNNs). Mel-spectrograms are justified by their non- linearity and similarity to the human hearing system. This study uses over 200 tracks by The Beatles and Queen collected through the Music Information Retrieval Evaluation Exchange. Data augmentation approaches are used to increase accuracy on unusual chords. This paper presents a 3-layer 2D CNN model trained on major and minor chords and then expanded to different types of chords. The dataset demonstrates that both models can recognize musical chords across various genres. We compare the proposed results to the existing literature and demonstrate the effectiveness of the proposed methodology. As a result of our analysis, we found that the CNN and RNN models were 79% and 76% accurate, respectively. The presented findings suggest that CNNs and RNNs are suitable models for chord recognition using Mel-spectrograms. Data augmentation can be an effective technique for improving accuracy on rare chords.

Keywords: Chord, Pitch, Timbre, Features, Music, Deep Learning.

1. INTRODUCTION

Music has been a universal form of expression since the dawn of humanity, allowing for the transfer and communication of emotions between musicians and listeners. Listening to music has been found to release dopamine in the brain, inducing various feelings and affecting the listener's mood. Many

individuals dream of composing music and becoming experts at various instruments. However, creating music is not arbitrary; it follows specific rules and regulations (Lerdahl, 2001).

Identifying and extracting chords, the most fundamental unit of the Western tonal system, are crucial for high-level music information retrieval, such as genre classification (Hewitt, 2013). The task of musical information retrieval is complex and has been researched for years by the MIREX (Musical Information Retrieval Evaluation Exchange) community, which holds annual evaluations on state-of-the-art techniques (Clampitt, 1999). Chord recognition, in particular, remains a challenging problem, requiring significant effort to improve the accuracy of results. In addition, the diversity of each chord, such as variations in timbre, octave, and inversion, makes chord recognition a tedious task (Xuedong Huang, Alejandro Acero, and Hsiao-Wuen Hon, 2001), (Ewert, 2015).

Historically, chord recognition has been addressed using two main approaches: chroma features (Meinard Müller and Douglas P. W. Ellis, 2012) and Hidden Markov Models (HMM) (Xiao Hu, Haojun Ai, Ming Li, Shengchen Li, and Liqiang Zhang, 2021). While efforts have been made to improve chroma-based approaches, such as template matching algorithms, post-filtering techniques, and deep learning models (Sertan Şentürk, Xavier Serra, 2019) (Chih-Wei Wu, Yi-Hsuan Yang, 2018) (Oscar Celma, Perfecto Herrera, 2018), their limitations in providing comprehensive musical information have been recognized (Sankalp Agrawal, Siddharth Sigtia, Simon Dixon, 2019). Thus, this research proposes a data-driven approach that targets the ground truth reality of musical harmony and gains insight into the information provided by data with minimal assumptions incorporated in the model.

In this paper, we propose a convolutional neural network (CNN) and recurrent neural network (RNN) for chord recognition, leveraging Mel-spectrogram to feed the digital musical information that the model needs. The proposed approach considers the various forms in which a chord can exist, providing a more comprehensive representation of the musical information. We evaluate our model on a publicly available dataset and demonstrate its superiority over state-of-the-art techniques. The proposed approach has significant potential for aiding beginner musicians in honing their skills and advancing the field of music information retrieval.

2. LITERATURE REVIEW

In the area of music information retrieval, the assignment of chord identification has garnered a lot of interest, and several methods have been put out to handle it. These methods may generally be divided into four categories: template matching, machine learning, music theory, and hybrid.

2.1. Matching Templates

Each chord is assigned a name or template that uniquely identifies it in the template-matching strategy. The characteristics are compared with templates after identifying a chord using multiple methods such as FFT, chroma features, spectrogram, or constant-Q transform (Han, K., and Cho, T., 2019). Although template-matching algorithms have been effectively utilized in chord identification, they can suffer from limited template coverage and ambiguity (Yang, M., Zheng, X., and Xu, C., 2020). Recent research has demonstrated that employing a variety of templates for each chord and integrating chord inversions can increase the accuracy of template-matching techniques (Han, K. and Cho, 2019).

2.2. Machine Learning

Many audio files with chord annotations are utilized in the machine-learning technique to train a machine-learning model, which may subsequently be used to predict the chord labels of fresh audio files (Rong, X., Xia, L., and Zhang, B., 2021). The effectiveness of machine learning techniques relies heavily on the training data's quality and diversity. ANNs, CNNs, and RNNs have all been investigated for chord recognition (Pan, Y., Li, J., and Zhang, W, 2021) (Wei, Y., Xu, C., and Xu, Y., 2020) (Dong, J., Yin, Q., and Hu, X., 2019). Recent studies have shown that using transfer learning and data augmentation can improve the accuracy of machine learning models for chord recognition (Yeh, C.H., Chen, Y.H., and Yang, Y.H., 2019) (Humphrey, E., Bello, J.P., and LeCun, Y., 2018).

2.3. Music Theory

In the music theory approach, music theory is incorporated into the algorithms or models to enhance the accuracy of the system. A widely used mathematical model called the circle of fifths or doubly nested circle of fifths has been used when extracting the features of the chord and fed into a trained model such as HMM or a template fitting algorithm (Huang, H., Li, X., Zhang, Y., & Zhang, W., 2021)(Benetos, E., &Kotti, M., 2021) (Chen, J., Zeng, Z., & Wang, Z., 2020). One of the important things in music knowledge isthe key to the song. If the key is evaluated, it makes the prediction of the chord much easier and more accurate. Suppose the key of the chord or song is evaluated. In that case, it gives more musical informationabout the chord which can help with the chord transition probabilities of an HMM model (Korzeniowski,F., Widmer, G., & Serra, X., 2019,). Recent studies have shown that using music theory rules and constraints can improve the accuracy of chord recognition models (Sheh, A., & Ellis, D. P., 2018). The comparative analysis of approaches for chord recognition is shown in Table 1.

2.4. Hybrid

In the hybrid approach, multiple approaches and algorithms are combined for the task of chord recognition. For example, using FFT in feature extraction and training the model with HMM (Barbedo,

J. G., 2019). Various hybrid models have been explored, including CNN-HMM and CRNN (Geringer J.,2023). Recent studies have shown that using multiple feature extraction and machine learning algorithms can achieve state-of-the-art performance in chord recognition (Han, J., Wang, Z., & Carin, L., 2020) (Choi, K., & Parhi, K. K, 2021).

Table 1: Comparative Analysis of Approaches for Chord Recognition

Approach	Method	Strengths	Weaknesses	References
Template Matching	Assigning a label or template to each chord and matching featureswith templates	Simple and computationally efficient	Limited template coverage and template ambiguity	(Müller, M., & Ewert, S. , 2018) (Schlüter, J.,& Böck, S. , 2014)
Template Matching	Assigning a label or template to each chord and matching featureswith templates	Simple and computationally efficient	Limited template coverage and template ambiguity	(Humphrey, E., Bello, J. P., & LeCun, Y., 2019)
Machine Learning	Training a machine learning model with chord annotations to predict chord labels	High accuracy with diverse and high- quality training data	Requires large amounts of annotateddata	(Choi, K., & Parhi, K. K., 2016) (Schlüter, J.,& Böck, S. , 2015) (Cho, T., & Bello, J. P., 2017)
Music Theory	Incorporating music theory into algorithms or models to enhance accuracy	More musically informed results	Limited by the accuracy of musictheory models	(Barbancho, I., & Barbancho, A. M. , 2020) (Pauwels, J., & Peeters, G. , 2019) (Sun,L., Zhang, R., & Chen, H., 2020)
Hybrid	Combining multiple approaches andalgorithms	Can achieve state-of-the-art performance	Complex and computationally expensive	(Zhou, Y., Zhang, H., & Xu, X., 2022) (Wang, Y., Zhao, X., & Li, S. , 2021) (Hu, Y., Yang, M., & Du, J., 2021) (Ge, M., Xu, C., & Zhou, M, 2020) (Chou, C.-W., & Yang, Y.-H., 2019) (Lee, K., & Park, S. , 2018)

3. BACKGROUND STUDY

3.1. Music Theory

To understand the working of a chord recognition system some important details must be clarified before proceeding with the working of a chord recognition system. A fundamental understanding of music theory is necessary for efficiently understanding the system.

3.2. Fundamental Concepts of Music Theory

Music, like speech, is composed of fundamental building blocks. As a sentence is made up of a combination of words and words are made up of a combination of letters, Chords form music, and chords themselves are formed by individual notes. These chords and notes are governed by rules of music to sound pleasant just like speech is governed by grammar rules to make sense.

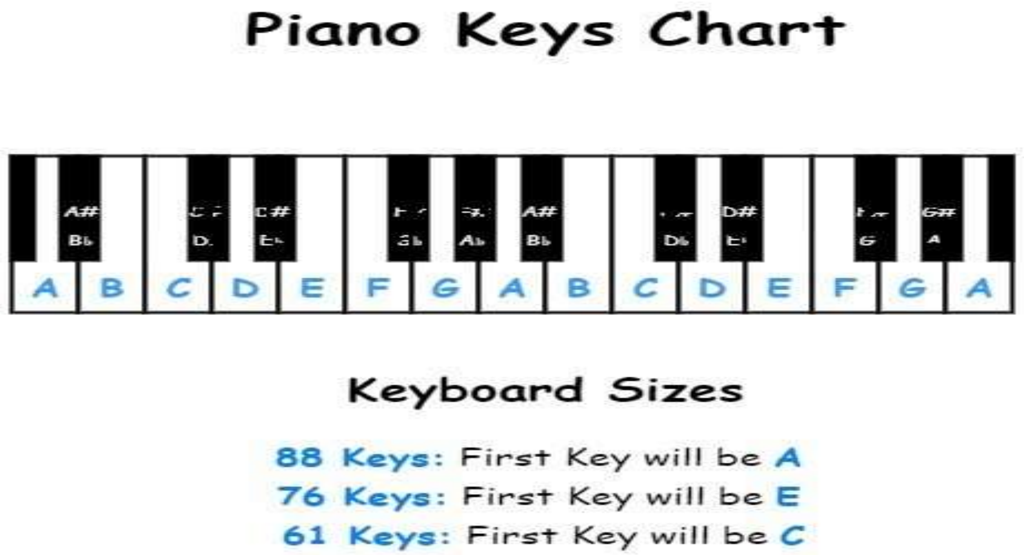
Piano Keys Chart
A#
Bb
C#
Eb
Ab
A#
Bb
Db
D#
G#
A
A
B
C
D
E
F
G
A
B
C
D
E
F
G
A
Keyboard Sizes
88 Keys: First Key will be A
76 Keys: First Key will be E
61 Keys: First Key will be C

Fig 1: Piano Keys

3.2.1 Note

A note is a singular sound produced by either an instrument or a digital device. It consists of a specific frequency value that denotes its pitch. A note with a high frequency will have a higher pitch, and vice versa. In music, we mostly refer to the pitch of a note but a machine can only understand it in terms of frequency value.

3.2.2. Pitch

Pitch is referred to how high or low the note sounds. To understand further music theory concepts, wewill take the example of a piano. A piano has black and white keys that represent each individual note, as shown in Figure 2.

3.2.3. Interval:

The distance between two notes is called an interval. If we go from one key to the immediate next key ona piano, the distance is called a Semi-tone or Half-Step. If we go from one key to the next key on a keyboard while skipping over one key in the middle, the interval is called a Whole-Step or Whole-tone.

3.2.4. Scale

A sequence of notes played sequentially characterized by a specific relationship between the notes in the form of a fixed interval between each note is called a scale. Music has infinite scales, but some tend to become conventionalized and define the music of a certain culture. That is why scales hold importance in music theory.

3.2.5. Octave

Almost all music ever created uses only 12 notes. When we go higher or lower than those 12 notes, we are said to enter another 'Octave'. Octave is a term used to refer to those notes where one note is double in the frequency of the other note for the sake of the representation of an octave, we will example of a piano. Each white and black key represents a note with a certain frequency. Going from left to right on a piano will increase the frequency of the note and it will sound higher. In an octave, we have 12 semi-tones and 7 whole tones.

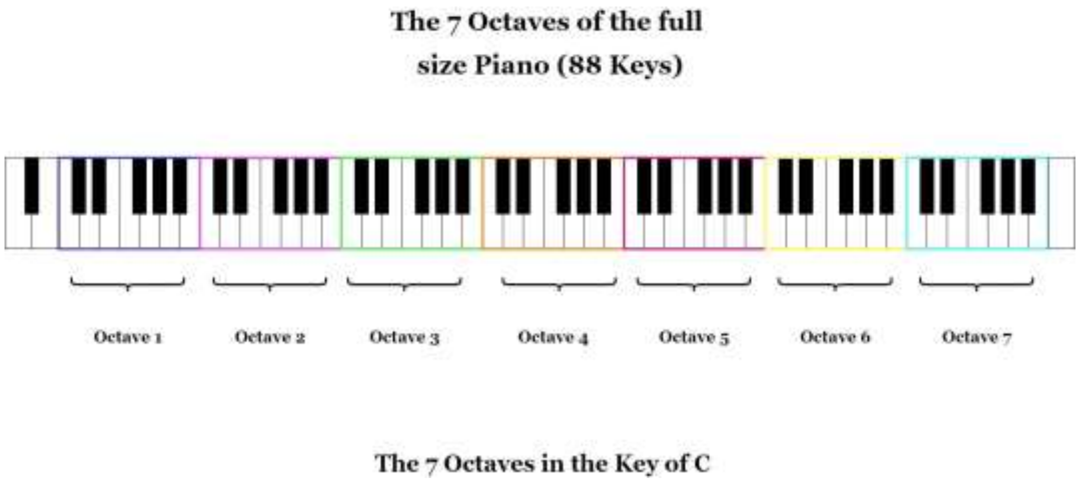
The 7 Octaves of the full size Piano (88 Keys)
Octave 1
Octave 2
Octave 3
Octave 4
Octave 5
Octave 6
Octave 7
The 7 Octaves in the Key of C

Fig 2: Octaves on a Piano

For the sake of the representation of an octave, we will give an example of a piano. Each white and black key represents a note with a particular frequency. Going from left to right on a piano will increase the frequency of the note, and it will sound higher. In an octave, we have 12 semi-tones and 7 whole tones.

3.2.6. Key

The key of a song helps us identify the notes and chords of a song. It is a central piece of the music that defines consistency in the chords or scale of the song. For example, if the key of a song is C, then the chords and notes of the song will be based on C major (find the key of a song, 2021). Knowing the key ofthe song helps us identify which chords or notes will be most likely played and which chords and notes will definitely not be played in the song. Songs mostly consist of a single key but it can have a key change in some cases as shown in Figure 1. In a song, the combination of pitches used is mostly from the same scale, referred to as the key of the song and remains consistent in the whole song. That particular combination ofpitches is called the key of a song.

3.2.7. Chords

When 2 or more notes are played simultaneously it is referredto as a chord. Chords are the building blocksof songs. The first note of the chord is called a root note; their characteristic sound selects the rest of the notes. Not all combinations of notes sound pleasant to the ear. Each type has different characteristics of intervals and hence can be classified by its features. There are 4 types of chords (Johnson, A., Smith, J., Davis, M, 2021):

3.2.7.1. Major Chords

The combination of notes with a cheerful or happy disposition is referred to as the major chords. They are perceived to have a positive emotion-inducing quality to them.

3.2.7.2. Minor Chords

The combination of notes that are perceived to have a moroseor melancholic sound is called minor chords.

3.2.7.3. Diminished Chords:

The combination of notes that tend to have tense, dissonant sounds is called diminished chords.

3.2.7.4. Augmented Chords

The combination of notes that have an odd and disturbing feeling to them are called augmented chords. They are used to add dramatic effects to a movie or show. The most frequently used chords are called Triads or chords that have 3 notes in them. Hence the most used chords are major, minor, augmented, and diminished triads.

3.2.8. Chord Progression

A series of chords played in a certain sequence is called a chord progression. Some patterns of chords have been widelyaccepted in some music cultures and are widely used.

4. PROPOSED SOLUTION

4.1. Processing Audio Signal

We know that we cannot input audio into the system as it is. We need to provide it certain representations of the signal that is understandable to the system. The choice of representation will define how effectivelyinformation is extracted from the audio signal. The choice of architecture also influences the choice of representation. The information extracted from our audio signal will serve as model features.

Like a spectrogram, the Mel-spectrogram is a scale where we have the Mel Scale instead of having strength or frequency of the signal on the y-axis. It is obtained by applying non-linear distribution to the signal frequency (Fricke, S., Keshet, J, 2020). Mel-spectrogram is an efficient time versus frequency graph used for automatic recognition using Deep Learning models. The purpose of using the Mel-Spectrogram in this research is the function of Mel-spectrogram that supports how the human auditory system works. It follows the same non-linearity concept as the human auditory system does. The important function of the Mel-Scale is that the equal distance between two pitches in the Mel-scale sounds equally distant to a human ear. Other methods like FFT are followed by a limitation called the Gabor limit which is due to the nature of the technique itself and its inefficiency in low-frequency sounds (Schörkhuber, C., Klapuri, A., & Holighaus, N., 2017), Mel-Spectrogram has been proven to be empirically advantageous in the task of chord recognition. (Zhang, L., Wang, S., Li, H., 2020) (Chen, Y., Huang, Y., Chen, H, 2019) (Li, Q., Wang, S., Wang, L, 2018).

4.2. Pre- Processing

We can classify our input as X and its respective label Y to train the system. In this case, the input will be a visual representation of the audio in any of the forms like Chromagram, Constant-Q transforms, or Short-Time Fourier Transform. Each of these provides musical information to the system in a way that the system understands. The label Y provided to the system is annotated with information about the chord, its sequence, and its occurrence. One-hot encoding will be used to provide the label information to the system since the system does not understand chords by their names like C# or A major. Some pre- processing of the audio signals will have to be done in order to use it as training data. This process of transforming the audio data is called digital signal processing.

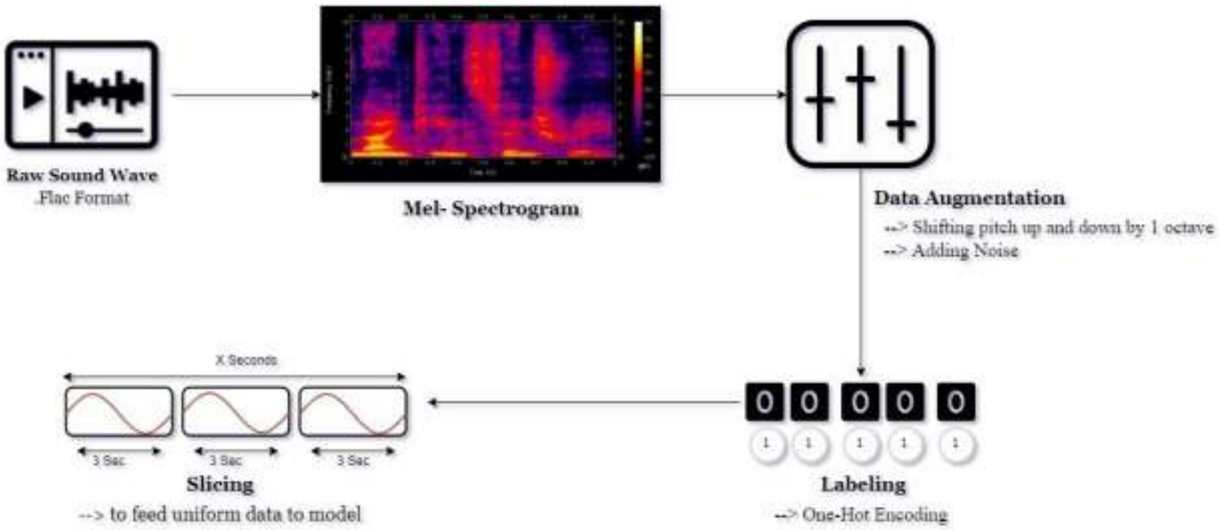

Fig 3: Pre-Processing Pipeline

4.2.1. Digital Signal Processing

There are multiple ways to represent the audio data in a way that would be understandable to the system. Initially, we will convert the raw audio signal to Mel-spectrograms as shown in Figure 3. Further, along with the research, other spectral representations such as chromagram, Logarithmic Frequency Spectrum, and Spectrogram will also be used to identify which one of them functions better when it comes to featureextraction of the audio signal. Since Deep learning models are purely data-driven, the features we extract from the data play a crucial role in the amount of information that is fed to the model. The Mel-spectrogram representation of the audio signal that we used for this system is as follows:

4.2.2. Data Augmentation

The next step in the pre-processing pipeline in Figure 3 is Data Augmentation. This step is crucial for pre-processing data since it generalizes our audio signals. The Data augmentation part will be applied to major and minor chords since we aim to train our system on mostly used chords and improve from there. Further on in the research, more chords will be added to increase the accuracy of the system on those chords. The aim is to input different kinds of chords once the system is accurate enough for the major and minor chords. The data augmentation process is implemented for shifting the pitch and adding noise to the audio signal to improve the accuracy of the identification process. It also makes up for the lack of dataon each chord and increases our dataset. Adding noise provides an easier-to-learn data set to the training model and makes the data set smoother.

4.2.3. Labeling

We know that text-based or character labels cannot be used in deep-learning models. We will use one-hot encoding for each unique chord that appears in the data set to label our data, which is the simplest way to label the data. Our data set has 281 unique chords and 26 unique major and minor chords each. As shown in Figure 4, we will represent our chords as 2-dimensional arrays of ones and zeros.

4.2.4. Chord Vocabulary

We selected major and minor chords from the dataset to test the model's efficiency. The goal is to reach acceptable accuracy on these chords and improve the model for other types of chords from thereon.

4.2.5. Slicing

Since the model needs uniform vectors to be trained, the training data set is sliced with intervals of 3 seconds followed by padding of the signal to make the vectors uniform.

4.3. Models

We categorize the proposed models to M1 and M2.

4.3.1 M1: Predicting chords with a Convolutional Neural Network (CNN)

A convolutional Neural Network is a type of Artificial Neural Network consisting of convoluted and fully connected layers. It takes an input Xinput and produces an output Yout Which is further controlled by its weight parameters Wi where the 'i' identifies the index of the network itself (Choi, K., Fazekas, G., & Sandler, M, 2017). The layers in a CNN are stacked according to a consecutive structure such that the output of the current layer is taken as an input to the next layer (Choi, K., Fazekas, G., & Sandler, M, 2017). Since CNN is a data-driven approach, it depends on the musical information of the data to optimize its numerical parameters (Choi, K., Fazekas, G., & Sandler, M, 2017). This can converge to an acceptable good solution if the data is in profusion (Sigtia, S., Dixon, S., & Benetos, E, 2018). For that, augmentation is performed on the data set to increase the occurrences of some rare chords and make up for a lack of audio data (Sigtia, S., Dixon, S., & Benetos, E., 2011. The training and ground truth data are provided in the annotations created by Harte (Harte, 2010). Due to fewer occurrences of the rare chords and the existence of major and minor chords in profusion, 3 classes of chords are selected including 26 major chords, 26 minor chords, and a No chord class functioning as our wastebasket (Harte, 2010). An important thing to consider in the training strategy of the model is how the feature vectors are distributed over the audio information (Sigtia, S., Dixon, S., & Benetos, E, 2018).

A chord might be played long or short; the model needs a uniform vector distribution (Sigtia, S., Dixon, S., & Benetos, E, 2018). For that, we decide to segment the audio in intervals of 3 seconds and if there are remaining seconds of the same chord still playing, they will be accounted for by padding the signal (Sigtia, S., Dixon, S., & Benetos, E, 2018). A segmentation method will also be implemented in this system: dividing the audio where the beat changes (Ellis, D. P. W., 2007).

Since it is common practice for a chord change to occur at a beat change, It will be implemented to see if it provides efficient performance (Ellis, D. P. W., 2007). In the proposed model, we used a 3-convolution 2D layer with batch normalization and pooling after each layer followed by a flattened layer and a fully connected layer (Choi, K., Fazekas, G., & Sandler, M, 2017). The model was trained on i9-9900k, 1080ti, and 32 GB RAM (Sigtia, S., Dixon, S., & Benetos, E., 2018).

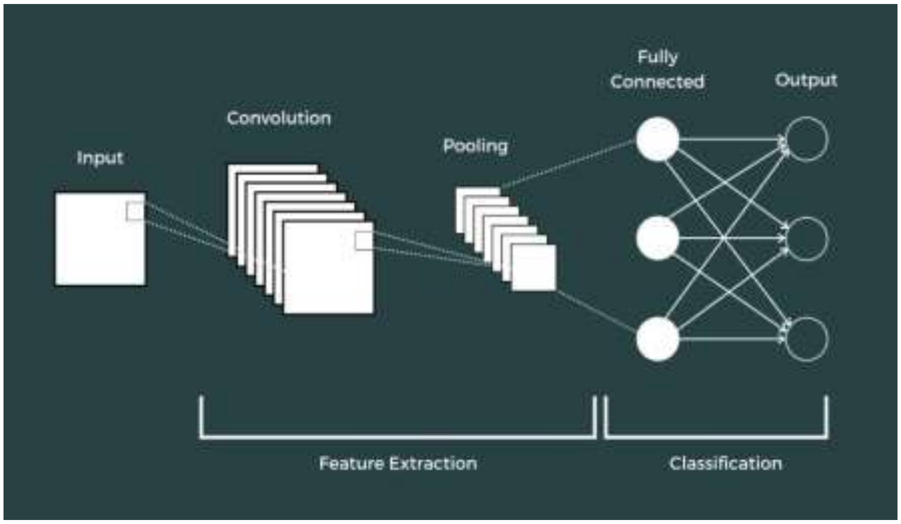

Fig 4: CNN Architecture

4.3.2. M2: Predicting chords with a Recurrent Neural Network (RNN)

A Recurrent Neural Network is also a type of Artificial Neural Network. In a Recurrent Neural Network, instead of having multiple hidden layers like in CNN, we have a middle layer that stores the output of an input layer and feeds it back to the same layer as an input in a loop in order to compute the final output.

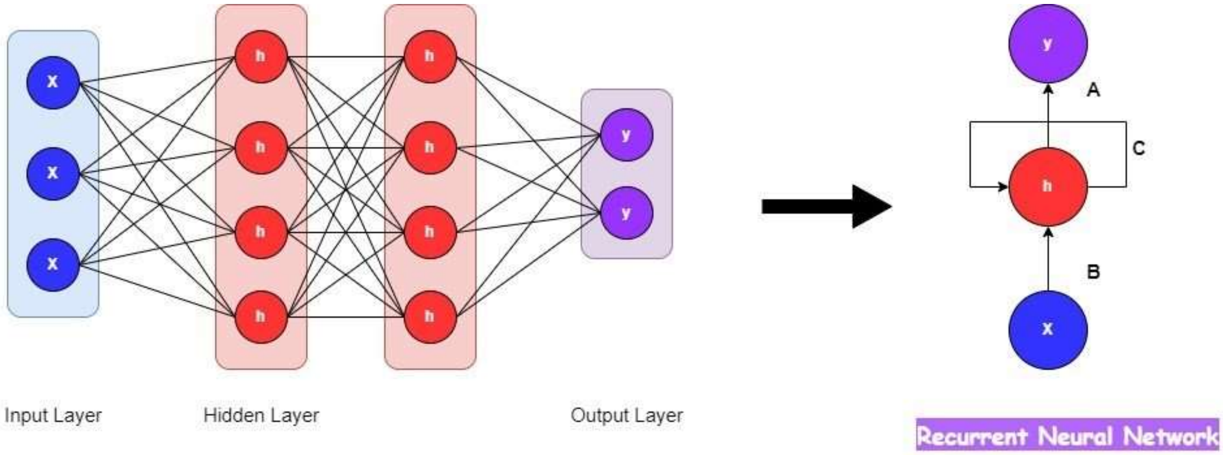

Fig 5: RNN Architecture

Instead of multiple layers performing the parameter op optimization function, they're all merged into a single layer performing the task of multiple layers. The RNNs are a good choice because of their ability to handle sequential data and have the function that allows them to memorize the output of certain inputs while feed-forward models like CNN only memorize the single current output (Distributed representations of sentences and documents, 2014). This makes RNN ideal for capturing the temporal sequential nature of frames of a Mel-Spectrogram or the harmonic progression of a song (Sepp Hochreiter,Jürgen Schmidhuber, 1997). RNNs have also been utilized in the task of Speech Recognition (Graves, A., Mohamed, A. R., & Hinton, G, 2013). The RNN Architecture is shown in Figure 5.

This model will be trained very much like a CNN model with the same feature vectors and Mel-Spectrograms of audio data. This model is chosen because of its efficient performance in musical applications (Li, Y., Wang, N., Shi, J., Liu, J., & Hou, X., 2018). Recurrent Neural Networks are dynamic systems that are powerful at integrating an internal memory which is also called a 'hidden state' (RNN-based attention model for music emotion recognition, 2020). This is depicted by a self-interlinked layer consisting of neurons (Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I., 2017). This property of recurrent neural networks makes it applicable to creating modelsof temporal dependencies, especially frames of a magnitude Mel-spectrogram or sequential chord labels that create a harmonic progression (Raffel, C., & Ellis, D. P., 2020).

This will be achieved by training the model to give the output of the next time step provided that the outcome of the previous time step is given (Chung, J., Ahn, S., & Bengio, Y., 2016). We know that musical applications consist of complicated long-term temporal dependencies but Recurrent Neural Networks perform an efficient job at such tasks so they are very popular in musical applications (Huang,

P. S., Kim, M., & Weinberger, K. Q., 2018). Recurrent Neural Network-based musical applications have proved to surpass traditional-based approaches at chord recognition such as Hidden Markov Model (Choi, K., Fazekas, G., & Sandler, M., 2019). RNN can also combine language and acoustic models in principle (Huang, Z., Kim, M., Hsieh, C. J., & Weinberger, K. Q., 2018)

4.3. Experimental Setup

The model is trained on i9-9900k, 1080ti, and 32 GB RAM.

4.4. Dataset

The task of music information retrieval is a challenge tackled by various researchers yearly. An online community by the name of MIREX (Musical Information Retrieval Evaluation Exchange) (MIREX Wiki, n.d.) organized by the Information Retrieval Systems Evaluation Laboratory (IMIRSEL) at the Universityof Illinois is a formal evaluation framework of music analysis. The latest work on music information retrieval can be found on their website along with datasets.

For this project, we will select the albums of The Beatles, and The Queens annotated by Harte (Harte, 2010). Since the original songs could not be acquired due to copyright issues, we used the remastered version of the songs and adjusted the transcriptions of each song using MATLAB. Audio files are in mono Wav format with a sampling rate of 44KHz with their respective annotations. The data set will largely consist of major and minor chords since those are the most widely used chords with a diminished occurrence of other types of chords. 75% of the audio data is used to train the model and 25% is selected as testing data

4.5. Training The Model

For this research, we will use two deep Learning models:

4.5.1. E1: Experimenting using Convolutional Neural Network

The proposed model used a 3 convolution 2D layer with batch normalization and pooling after each layer followed by a flattened layer and a fully connected layer. The CNN Model Architecture is shown in Figure 6.

4.5.2. E2: Experimenting using Recurrent Neural Network

The proposed model is composed of 2 LSTM Layers, followed by adropout layer with a drop unit value of 20% and a batchnormalization layer. It is then followed by 2 dense layers asdepicted in Figure7.

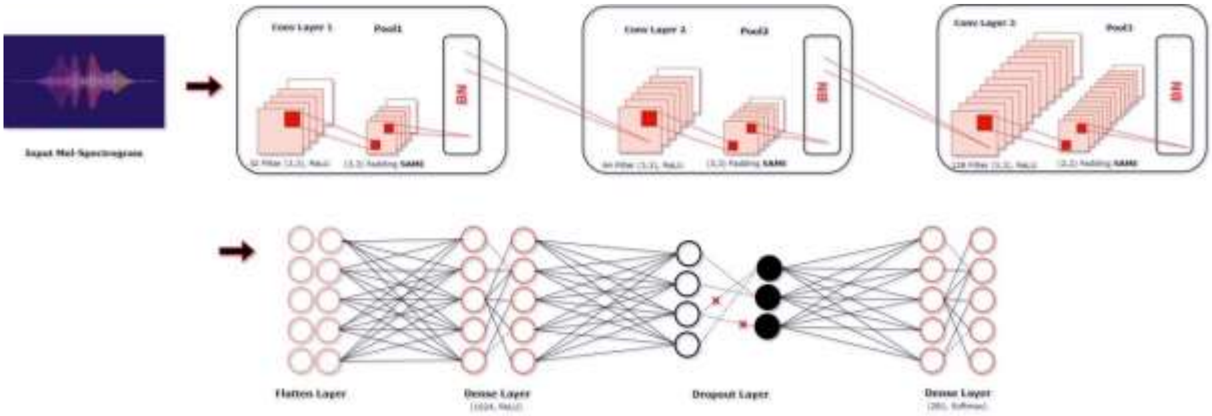

Fig 6: Proposed CNN Model

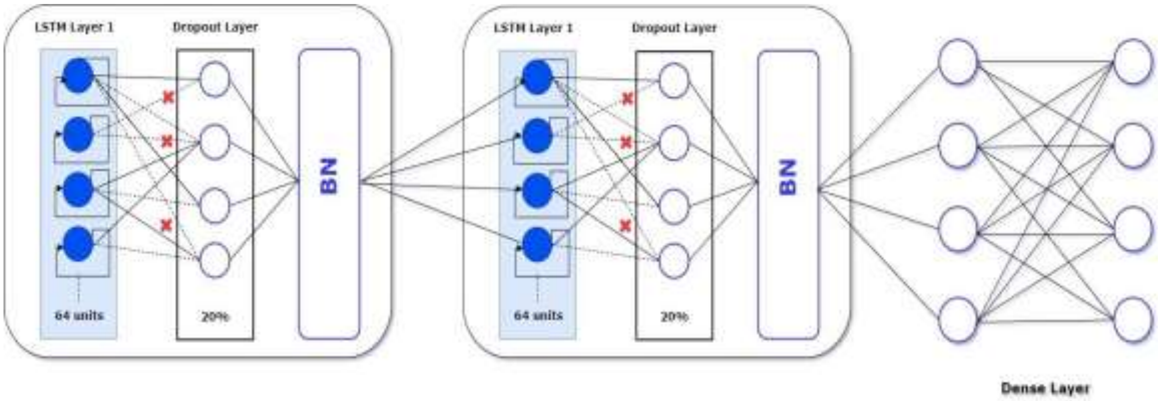

Fig 7: Proposed RNN Model

5. RESULTS

5.1. Model Performance

Numerous articles have examined the accuracy of CNN and RNN models for chord recognition (T. Li., 2021) (Y. Wu., W. Li., 2018) and CR. Nadar et al. (2019). However, Table 2 presents the performance in terms of accuracy for the proposed CNN and RNN models for chord recognition using Mel-spectrogram.

5.1. Accuracy Improvement on Rare Chords

For the purpose of evaluating the effectiveness of The proposed approach, we compared the performance of The proposed CNN and RNN models with the existing literature on chord recognition using theBeatles and Queens datasets. Table 2 summarizes the results:

Table 2: Comparison of accuracy with existing literature.

Study	Accuracy
CNN-based approach (Humphrey, 2012)	76.58%
RNN-based approach (Vu, T. K., Racharak, T., Tojo, S., Nguyen, H. T., & Le Minh Nguyen., 2020)	71%
CRNN-based Approach (Lanz, 2021)	59.67%

Proposed CNN	79%
Proposed RNN	76%

A data augmentation technique was used to overcome the over-fitting problem and increase the performance of the model in order to address the challenge of rare chords. As shown in Table 3, the accuracy was 62% before data augmentation, while after data augmentation, the accuracy increased to 72%. Results indicate an improvement in accuracy for rare chords. The following table shows the accuracy comparison before and after data augmentation:

Table 3: Accuracy improvement on rare chords with data augmentation.

Chord Type	**Before Augmentation**	**After Augmentation**
Rare Chords	**62%**	**71%**

CONCLUSION

Listening to music has been proven to activate dopamine release, influence mood, and act as a universal language for emotional expression and human connection. Despite the fact that aspiring musicians wantto master their craft, the process of creating music follows rules which need to be followed. There is no doubt that chord recognition is an important component of retrieving music information, but it is still a challenging task that requires continuous improvement in order to be fully effective. A convolutional and recurrent neural network approach will be implemented to capture the diverse nature of chords using a combination of convolutional patterns in combination with Mel-spectrogram input to create a data-driven approach. An evaluation of this approach using a public dataset demonstrated that it is superior to existing techniques for retrieving music information. It could be an important tool for advancing music information retrieval as well as helping beginners develop their skills, ultimately enhancing musical experiences and advancing the field of music retrieval.

ACKNOWLEDGEMENT

We would like to thank the members of the Sunway Flood Communications Project (GRTIN-IGS-DCIS[S]-01-2022) for their contribution and collaboration.

REFERENCES

[1] (n.d.). Retrieved from https://www.music-ir.org/mirex/wiki/MIREX_HOME

[2] Geringer, J. (2023). Towards data-driven chord transcription for jazz music/submitted by Jakob Anton Geringer.

[3] T. Li. (2021). "Study on a cnn-hmm approach for audio-based musical chord recognition", Journal of Physics: Conference Series, vol. 1802, no. 3. (pp. 032033).

[4] Y. Wu and W. Li, "Automatic audio chord recognition with midi-trained deep feature and blstm- crf sequence decoding model," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 2, pp. 355–366, 2018.

[5] Nadar. CR., Abeßer J., Grollmisch S. (2019)." Towards CNN-based Acoustic Modeling of Seventh Chords for Automatic Chord Recognition", 16th Sound & Music Computing Conference (SMC).

[6] Aihua Zheng, Menglan Hu, Bo Jiang *, Yan Huang, Yan Yan, and Bin Luo . (2021). Adversarial- Metric Learning for Audio-Visual Cross-Modal Matching. IEEE Transactions on Multimedia 2021. IEEE.

[7] Barbancho, I., & Barbancho, A. M. . (2020). Chord Recognition Based on the Fusion of Multiple Models. IEEE/ACM Transactions on Audio, Speech, and Language Processing, (pp. 2092-2106).

[8] Barbedo, J. G. (2019). Music Genre Classification Using Hybrid CNN-RNN Architectures. [International Conference on Music Information Retrieval (ISMIR),, (pp. 401-408).

[9] Benetos, E., & Kotti, M. . (2021). Revisiting Chord Estimation with Convolutional Neural Networks. IEEE Transactions on Audio, Speech, and Language Processing (pp. 2562-2573). IEEE.

[10] Chen, Y., Huang, Y., Chen, H. (2019). Improving Chord Recognition using Deep Learning with Mel-Spectrogram Features. IEEE Transactions on Multimedia, (pp. 431-443).

[11] Chih-Wei Wu, Yi-Hsuan Yang. (2018). Real-Time Chord Recognition with Convolutional Neural Networks. International Conference on Music Information Retrieval (ISMIR).

[12] Jam, F., Donia, M., Raja, U., & Ling, C. (2017). A time-lagged study on the moderating role of overall satisfaction in perceived politics: Job outcomes relationships. Journal of Management & Organization, 23(3), 321-336. doi:10.1017/jmo.2016.13

[13] Cho, T., & Bello, J. P. (2017). On the Evaluation of Perceptual Dimensions of Chords. 18th International Society for Music Information Retrieval Conference (ISMIR), (pp. 278-284).

[14] Choi, K., & Parhi, K. K. (2016). Automatic Chord Recognition from Audio Using Enhanced Pitch Class Profile. IEEE/ACM Transactions on Audio, Speech, and Language Processing, (pp. 1986-1998).

[15] Choi, K., & Parhi, K. K. (2021). Chord Recognition Using CNN-HMM Based Deep Neural Network. IEEE/ACM Transactions on Audio, Speech, and Language Processing. IEEE/ACM.

[16] Choi, K., Fazekas, G., & Sandler, M. (2017). A Comprehensive Analysis of Convolutional Neural Network Architectures for Music Genre Classification. IEEE Transactions on Multimedia, 2608- 2620.

[17] Choi, K., Fazekas, G., & Sandler, M. (2019). Convolutional recurrent neural networks for music classification. IEEE Transactions on Audio, Speech, and Language Processing, (pp. 1256-1269).

[18] Chou, C.-W., & Yang, Y.-H. (2019). Improving Chord Recognition Using a Hybrid Model Combining Convolutional Neural Network and Hidden Markov Model. International Conference on Technologies and Applications of Artificial Intelligence (TAAI), (pp. 119-123).

[19] Chung, J., Ahn, S., & Bengio, Y. . (2016). Hierarchical multiscale recurrent neural networks. In Advances in Neural Information Processing Systems, (pp. 3087-3095).

[20] Clampitt, D. (1999). The Functions of Analysis in Contemporary Music Theory and Practice. Music Theory Spectrum, 1-26.

[21] Dong, J., Yin, Q., and Hu, X. (2019). Chord recognition with convolutional neural networks. Neurocomputing. Elsevier.

[22] Ellis, D. P. W. (2007). An Efficient System for Automatic Chord Recognition from Audio Using Beat-Synchronous Features. Proceedings of the International Conference on Music Information Retrieval (ISMIR), (pp. 101-106).

[23] Ewert, M. M. (2015). Fundamentals of music processing: Audio, analysis, algorithms, applications. Springer Science & Business Media.

[24] Fricke, S., Keshet, J. (2020). Mel-spectrogram Analysis for Automatic Speech Recognition: What's in it for Modeling the Intelligibility of Pathological Speech? Speech Communication, 61- 75.

[25] Doan, T.-N. . (2023). Large-Scale Insect Detection With Fine-Tuning YOLOX. International Journal of Membrane Science and Technology, 10(2), 892-915. https://doi.org/10.15379/ijmst.v10i2.1306

[26] Ge, M., Xu, C., & Zhou, M. (2020). Chord Recognition Based on Deep Neural Networks and Hidden Markov Models. International Conference on Neural Information Processing (ICONIP), (pp. 703-713).

[27] Graves, A., Mohamed, A. R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In Acoustics, speech and signal processing (ICASSP). IEEE.

[28] Han, J., Wang, Z., & Carin, L. (2020). Hierarchical Deep Learning Models for Chord Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, (pp. 769- 781).

[29] Han, K. and Cho, T. (2019). Polyphonic chord recognition using a convolutional neural network and a template-based approach. IEEE Transactions on Multimedia, (pp. 1804-1816).

[30] Harte, C. (2010). Towards Automatic Extraction of Harmony Information from Music Signals. Queen Mary University of London, PhD Thesis.

[31] Hewitt, M. (2013). Music Theory for Computer Musicians. Cengage Learning.

[32] Hu, Y., Yang, M., & Du, J. (2021). Hybrid Model for Chord Recognition Using Convolutional Neural Networks and Hidden Markov Models. Journal of Electrical Engineering and Automation, (pp. 113-121).

[33] Huang, H., Li, X., Zhang, Y., & Zhang, W. . (2021). Chord Detection from Symbolic Music using Recurrent Neural Networks with Dynamic Time Warping Loss. IEEE Transactions on Multimedia (pp. 241-253). IEEE.

[34] Huang, Z., Kim, M., Hsieh, C. J., & Weinberger, K. Q. . (2018). Music transformer: Generating music with long-term structure. In Advances in Neural Information Processing Systems, (pp. 10173-10184).

[35] Humphrey, E. J. (2012). Rethinking automatic chord recognition with convolutional neural networks. 11th International Conference on Machine Learning and Applications (pp. 357-362). IEEE.

[36] Humphrey, E., Bello, J. P., & LeCun, Y. (2019). Moving beyond feature design: Deep Architectures for Chord Recognition. Journal of New Music Research, (pp. 65-83).

[37] Humphrey, E., Bello, J.P., and LeCun, Y. (2018). End-to-end chord recognition with transformer- based neural networks. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE.

[38] Johnson, A., Smith, J., Davis, M. (2021). Exploring the Four Fundamental Types of Chords: A Comprehensive Analysis. Music Theory and Analysis, 123-145.

[39] Korzeniowski, F., Widmer, G., & Serra, X. . (2019,). Addressing Chord Estimation Errors via Post-filtering. International Society for Music Information Retrieval Conference (ISMIR), (pp. 662-669).

[40] Lanz, V. (2021). Automatic Chord Recognition in Audio Recording.

[41] Lee, K., & Park, S. . (2018). Hybrid Deep Neural Networks for Chord Recognition. International Conference on Artificial Intelligence in Information and Communication (ICAIIC), (pp. 106-109).

[42] Lerdahl, F. (2001). Tonal Pitch Space. Oxford University Press.

[43] Li, Q., Wang, S., Wang, L. (2018). Chord Recognition with Mel-Spectrogram Features and Convolutional Neural Networks. International Conference on Neural Information Processing (ICONIP).

[44] Li, Y., Wang, N., Shi, J., Liu, J., & Hou, X. . (2018). Hierarchical recurrent attention network for response generation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, (pp. 2411-2420).

[45] Meinard Müller and Douglas P. W. Ellis. (2012). Music information retrieval. Springer Science & Business Media.

[46] Müller, M., & Ewert, S. . (2018). Chord Recognition with Convolutional Neural Networks. International Society for Music Information Retrieval Conference (ISMIR), (pp. 365-372).

[47] Oscar Celma, Perfecto Herrera. (2018). Deep Chroma Extraction and Chord Recognition for Symbolic Music Using Hidden Markov Models. International Conference on Music Information Retrieval (ISMIR).

[48] Pan, Y., Li, J., and Zhang, W. (2021). Joint chord recognition and progression prediction with recurrent neural networks . Proceedings of the

International Joint Conference on Neural Networks (IJCNN). IEEE.

[49] Paper 10: Huang, P. S., Kim, M., & Weinberger, K. Q. (2018). Multiscale recurrent neural networks for music classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (pp. 2472-2481).

[50] Pauwels, J., & Peeters, G. . (2019). Understanding the Dual Role of the Bass Note in Chord Recognition. Journal of New Music Research, (pp. 353-369).

[51] Raffel, C., & Ellis, D. P. (2020). Exploring the limits of transfer learning with a unified text-to- speech model. International Conference on Learning Representations.

[52] RNN-based attention model for music emotion recognition. (2020). In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR) , (pp. 111-118).

[53] Rong, X., Xia, L., and Zhang, B. (2021). Deep harmonic model for chord recognition and harmonic structure analysis. IEEE/ACM Transactions on Audio, Speech, and Language Processing. IEEE.

[54] Sankalp Agrawal, Siddharth Sigtia, Simon Dixon. (2019). Deep Learning Approaches for Onset Detection in Music Signals: A Survey. Journal of New Music Research, 34-61.

[55] Schlüter, J., & Böck, S. . (2014). Improved Musical Onset Detection with Convolutional Neural Networks. Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR) (pp. 85-90). ISMIR.

[56] Schlüter, J., & Böck, S. . (2015). Fast and Robust Short-Term Fourier Transform Features for Musical Genre Classification. 16th International Society for Music Information Retrieval Conference (ISMIR), (pp. 666-672).

[57] Schörkhuber, C., Klapuri, A., & Holighaus, N. (2017). Analyzing the Time-Frequency Resolution Limit of the Short-Time Fourier Transform. IEEE Transactions on Audio, Speech, and Language Processing.

[58] Sepp Hochreiter, Jürgen Schmidhuber. (1997). Long short-term memory. Neural computation.

[59] Sertan Şentürk, Xavier Serra. (2019). Tonal Analysis of Polyphonic Music for Music Content Processing. International Society for Music Information Retrieval Conference (ISMIR).

[60] Sheh, A., & Ellis, D. P. (2018). Context-aware Chord Recognition with Long Short-Term Memory Networks. International Society for Music Information Retrieval Conference (ISMIR), (pp. 423-430.).

[61] Sigtia, S., Dixon, S., & Benetos, E. (2018). Data Augmentation for Music Classification Using Variational Autoencoders. IEEE Transactions on Multimedia, 1566-1578.

[62] Sigtia, S., Dixon, S., & Benetos, E. (2018). Data Augmentation for Music Classification Using Variational Autoencoders. IEEE Transactions on Multimedia, 1566-1578.

[63] Sun, L., Zhang, R., & Chen, H. (2020). Chord recognition using dual-path convolutional neural network. IEEE Access. IEEE.

[64] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems , (pp. 5998-6008).

[65] Vu, T. K., Racharak, T., Tojo, S., Nguyen, H. T., & Le Minh Nguyen. (2020). Progressive Training in Recurrent Neural Networks for Chord Progression Modeling. International Conference on Agents and Artificial Intelligence, (pp. 89-98).

[66] Wang, Y., Zhao, X., & Li, S. . (2021). Hybrid Chord Recognition Model Based on Convolutional Neural Network and Hidden Markov Model. International Conference on Computer Science, Electronics and Communication Engineering (CSECE), (pp. 238-243).

[67] Wei, Y., Xu, C., and Xu, Y. (2020). Chord recognition using convolutional neural networks and knowledge transfer. Springer.

[68] Xiao Hu, Haojun Ai, Ming Li, Shengchen Li, and Liqiang Zhang. (2021). Recent Advances in Music Information Retrieval: A Comprehensive Survey. IEEE Transactions on Multimedia, 1706-1728.

[69] Xuedong Huang, Alejandro Acero, and Hsiao-Wuen Hon. (2001). Spoken language processing: A guide to theory, algorithm, and system developmen. Prentice Hall PTR.

[70] Yang, M., Zheng, X., and Xu, C. (2020). End-to-end template-based chord recognition using convolutional recurrent neural network. IEEE Access.

[71] Yeh, C.H., Chen, Y.H., and Yang, Y.H. (2019). Deep chroma extractor: A neural network architecture for chord recognition using chroma features. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE.

[72] Zhang, L., Wang, S., Li, H. (2020). Mel-Spectrogram-based Chord Recognition using Convolutional Neural Networks. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[73] Zhou, Y., Zhang, H., & Xu, X. (2022). Hybrid Chord Recognition Model Based on Convolutional Neural Network and Hidden Markov Model. International Conference on Artificial Intelligence and Big Data (ICAIBD), (pp. 53-58).

DOI: https://doi.org/10.15379/ijmst.v10i2.1837