Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Genre Classification of Saint Yared Qum Zema Using a Convolutional Neural Network

Birku Litgeb Aschenek

birku_litgeb@dmu.edu.et

Debre Markos University

Alehegn Ayanaw Getaneh

Debre Markos University

Getaneh Awoke

Debre Markos University

Article

Keywords: Music Information Retrieval, Genre Classification, Convolutional Neural Networks, Qum Zema, Ethiopian Sacred Music

Posted Date: September 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10826774/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Music Information Retrieval is one research area in Machine learning that emphasizes on extraction of valuable features from audio datasets for classification. The study shows the tasks of automatic genre classification of the greatest Composers of Saint Yared's Qum Zema, which is a form of Ethiopian sacred music sung in vocal voice. The three forms of zema genres are Geez, Ezil, and Araray, which have high intra-class similarity in classification. In this study, we have built a model called Saint Yared’s Qum Zema Classifier model to categorize audio data into three classes. The model took audio data and transformed it into spectrogram images. The total number of data used in this study were 1,555 audios, each audio segments in 10 equal seconds, and transformed into spectrogram images. The model uses ReLU activation function and Adam optimizer, which results a training accuracy of 98% and a testing accuracy of 88%. The model has the best result as compared with ResNet, VGGNet, and AlexNet in testing performance and model efficiency. The total average of the loss rate for the model was 1.01. This study provide base for further research in Ethiopian religious music and shows how deep machine learning can be used to preserve and classify intangible cultural treasures.

1. Introduction

Automatic music classification is a main application of artificial intelligence under the broader field of Music Information Retrieval (MIR). It include the sub categorical tasks like genre classification, instrument recognition, and mood detection via audio analysis (Purwins et al., 2019). The main concern of this study focuses on a specific subtask of music information retrieval, which is automatic genre categorization of Saint Yared's Qum Zema into three main genres using a convolutional neural network.

The Ethiopian Orthodox Tewahdo Church (EOTC) defines the meaning of Zema as a sacred melody or chant, which has a tactical and standard manner of singing that produces a spiritually pleasing sound (Kassaye, 2018). In the 6th century, Zema composer Saint Yared prepared Zema by systematizing this tradition into three primary forms, namely Geez, Ezil, and Araray. Zema is said to be Qum zema when it is sung using only vocal sounds, without instrumental accompaniment.

Saint Yared is named as the Ethiopian Re'ese Liqawnt (head professor) of Zema composers who was born from his father, Abyude or Isaac, and his mother, Christina or Tawuklia, in Axum in 505 E.C. (Kassaye, 2018). Figure 1 presents the Saint Yared (□□□□□, 2008).

The motivation of this study is twofold. First, the transmission of this knowledge occurs primarily via traditional church schools (*Zema Bet*), and expertise is concentrated among a diminishing number of scholars. Second, most of the flock often cannot distinguish such variations of these genres. This work aims to bridge this knowledge gap using modern machine learning techniques.

The main goal is to use the convolutional neural network (CNN) to categorize audio parts of Qum Zema into the appropriate genres (Geez, Ezil, and Araray). A spectral image converted from an audio signal is

used as an input to the model by removing the way of the manual acoustic feature extraction method (Peffers et al., 2007), which enables the network to learn unique features directly from the dataset.

2. Related Work

Numerous methods of machine learning have been used in significant studies on music information retrieval. Early work often relied on manually engineered acoustic features combined with classifiers like Support Vector Machines (SVM) (Costa et al., 2011) and k-Nearest Neighbors (k-NN) (Ali and Siddiqui, 2017). More recently, deep learning, particularly CNNs, has become the state of the art for audio classification tasks by learning features directly from raw data or spectrograms.

CNNs have been successfully applied to music instrument recognition from spectrograms (Jawaherlalnehru and Jothilakshmi, 2019), music emotion recognition (Bilal Er and Aydilek, 2019), and general sound classification (Khamparia et al., 2019). However, these studies often focus on Western music or general sounds and do not address the specific challenges of vocal-only, culturally unique music like Qum Zema, which exhibits high intra-class similarity.

We have observed from the related work that several studies have limitations such as a lack of labeled data for niche domains a dependence on manual feature extraction or a disregard for vocal sounds. Our work addresses these gaps by focusing specifically on the classification of such sacred chants using a CNN on spectrogram representations, automating the feature learning process for this unique audio corpus.

3. Methodology

3.1. Research Method

Our study follows the approaches of a research methodology called Design Science Research Methodology (Peffers et al., 2007), which highly emphasizes creating and evaluating an artifact. Therefore, the research that is designed in this study is the SYQZC classification model used to solve an identified problem.

3.2. Dataset Preparation

There was no well-prepared dataset to conduct our research Qum Zema was available. Data was collected from scholars at traditional EOTC schools (*Zema Bet*). To prepare well-organized datasets, we have collected the audio-recorded data from Zema Bet scholars, including chants Wudasie Maryam, Ankeste Birhan, Mestegab, Selamta, Tsome Diggua, and Diggua. Before preprocessing, the length of the data was so long. To prepare a well-defined spectral image, each long-sized recorded data was chunked into 10-second equal clips to ensure uniform input size. Every piece of segmented audio is converted into spectrogram images, which visualize the signal's frequency spectrum over time. The final prepared dataset consisted of 1,555 spectrograms across the three classes (Araray: 595, Ezil: 539, Geez: 421).

The dataset split into a ratio of 70% for training and 30% for testing. Table 1 presents the total numbers of datasets.

Table 1
Total numbers of datasets

Class	Number of audio (.wav)	Time	Spectrogram image
Araray	595	10 sec	595
Ezil	539	10 sec	539
Geez	421	10 sec	421
Total	1555	15550 sec	1555

3.2.1 Audio Data Collection and Ethical Considerations

The audio dataset was collected directly from Zema scholars who participated as human subjects and provided authentic Saint Yared Qum Zema performances for this study. All methods involving human participants were conducted in accordance with relevant ethical guidelines and regulations. The research protocol was reviewed and approved by the Institutional Review Board/Ethics Committee of Debre Markos University (Approval/Reference No. DMU-IRB-2026-6341), and informed consent was obtained from all participants before recording. The participants were informed about the purpose of the study and the intended academic use of the recordings. The collected audio data were anonymized and prepared for preprocessing and genre classification. The dataset is available from the corresponding author upon reasonable request, subject to applicable consent and privacy conditions.

3.3. Spectrogram Transformation

The audio collected from experts has different lengths. In order to make it more recognizable and uniform, we split each audio into 10-second segments. The dimensions of the spectrogram image, which becomes an input for the first layers of convolutional neural networks, were 128 × 128 × 3 sized with RGB, and the graphical depiction of the signal of an image in terms of frequencies at various times. It uses a 1024 sample of window size and the Short Time Fourier Transform. This transformation changes the 1dimensional audio signal into a 2 dimensional image (time vs. frequency), where amplitude (loudness) is represented with the intensity of the colors, which makes it an appropriate input for CNNs that use images. Figure 2 presents the waveform of the audio data, while Fig. 3 presents the spectrogram image of the sample audio data.

3.4. Model Architecture: SYQZC

The proposed Saint Yared Kum Zema Classifier (SYQZC) is a CNN architecture designed for spectrogram images as an input.

From the model shown in Fig. 4, primarily audio data is collected from Zema scholars in WAV format; in addition, we have used open-source recorded data annotated by the scholars. We applied the spectral gating techniques to reduce the background noise that may occur in our data. We tried to split the audio into equal 10-second segments to make them more recognizable. To generate the best quality of spectral image from the audio, we have used different parameters: window size, hop size, FFT size, amplitude scaling, and frequency scale.

The built SYQZC model takes images as final input to CNN layers, which are converted spectral images obtained from the audio files. From the input image, features are extracted with the first layers of the CNN called convolution, and the dimension reduction of the image is performed with the second layers named as pooling, and finally the pixels of the spectral image are transformed into vector form with fully connected layers and classified with SoftMax. The classification is performed with the training phase starting from the starting points of reading audio data and similarly for the testing phase.

The SYKZC model is designed to classify St. Yared Qum Zema types into three main genres. Initially the model takes input audio data, and this data needs further preprocessing; this includes noise reduction and segmentation to have uniform segments of audio with time interval and size. Each segmented form of audio is transformed into spectrogram images.

4. Results and Discussion

4.1. Experimental Setup

For the experimental setup used to design the SYQZC model, we used the software package, library and, computer of an Intel ® core™ i5-4300U CPU and 4 GB of RAM. The model trained for 100 epochs, a batch size of 32, and a starting or initial learning rate of 0.001 (1e-3). The data was partitioned into a training and testing dataset; 70% of the data is assigned for training the model, and 30% of the data is allotted for testing.

4.2. Results

When we come to the result obtained from the experiment, the SYQZC model obtained 98% training accuracy and 88% testing accuracy. The overall average loss rate for the model is 1.01. This model accuracy is obtained when the model is trained with the absence of background noise from the audio data, with texture feature extraction, and using dropout rate at initial stages. This model has better accuracy performance as compared to the remaining related convolutional neural network models.

In this study, we tried to compare the training and testing results of the model to evaluate the performance with different activation functions. As shown in Table 2, the ReLU activation function yielded the best performance compared to Sigmoid and Tanh.

Table 2
Model comparison with different activation functions.

Model Name	Activation Function	Training Accuracy	Testing Accuracy	Loss Rate
SYQZC	**ReLU**	**98%**	**88%**	**1.014**
	Sigmoid	96%	84%	0.875
	Tanh	95%	84%	0.875

The SYQZC model was compared with several well-known CNN architectures, as summarized in Table 3. While ResNet achieved a slightly higher training accuracy (99%), its testing accuracy is more or less low. The result obtained from each model, as compared with the SYQZC mode1 in some extent, the testing accuracy of our model has a higher result as shown (88% vs. 85%). Furthermore, SYQZC is significantly more efficient, with a much smaller model size (8.73 MB) and faster execution time.

Table 3
Comparison of SYQZC with other CNN models.

Model Name	Training Accuracy	Testing Accuracy	Loss Rate	Model Size (MB)
SYQZC (Proposed)	98%	88%	1.014	8.73
ResNet	99%	85%	0.9814	25.49
VGGNet	95%	75%	1.9965	745.31
AlexNet	98%	82%	1.2620	343.56

Figure 5 presents the comparison results of the SYQZC model with different optimizers and Fig. 6 presents the SYQZC model with Adam optimizers and different activation functions.

The training loss and accuracy of the SYQZC model is shown in Fig. 7, while Fig. 8 shows the accuracy rate of training and test of SYQZC model.

Figure 9 presents the loss rate of training and testing of SYQZC model.

4.3. Discussion

The results obtained from the experiment confirm that a convolutional neural network can effectively learn to distinguish between the audio Zema dataset that characterizes the three genres of Saint Yared's Qum Zema. The model’s result is a high testing accuracy of 88% as compared with other machine learning models, but its testing accuracy rate doesn't go above it because of the challenging nature of the dataset, which consists only of vocal chants with high intra-class similarity.

The overall performance of the built model, resulting in 98% training accuracy and 88% testing accuracy, which has relative greatest over larger architectures like ResNet (99% training accuracy and 85% testing

accuracy), VGGNet (95% training accuracy and 75% testing accuracy), and AlexNet (98% training accuracy and 82% testing accuracy). The SYQZC model had the best performance when applied in image processing and DSP tasks, thanks to its sound design approach. This algorithm was equally suitable for the nature of the spectrogram data with a moderate number of samples, hence lowering the chances of overfitting. Our model trained steadily and effectively thanks to the use of ReLU and the Adam optimizer.

5. Conclusion

Generally, the study conducted to classify Saint Yared's zema, which are sung with only vocal voice in three main classes, and to make it successful, we applied a convolutional neural network. The initial data were audio files passed via different preprocessing techniques. We have observed the ability of CNNs in image processing to automate feature extraction and classification in order to transform audio segments into spectrograms. This approach is feasible as evidenced by the developed SYQZC models high testing accuracy of 88%.

This work has both practical and methodological significance. Practically, it provides a tool to support in the preservation and education of Saint Yared’s sacred musical tradition. Methodologically, it establishes a roadmap for applying deep learning to other forms of traditional and culturally significant music.

6. Future Work

Future work will focus on:

- Applying both acoustic feature extraction methods and visual representation of audio data may maximize the accuracy and performance of the classifier model.
- Increasing the dataset also has a great impact on the classifier model. When the number of input data increases, the ability of the classifier model becomes better.
- Applying data augmentation techniques on the audio signal, such as addition of a noise, using different loudness ranges, time stretching, and pitch shifting.
- Adding textual information (features) other than audio and video, such as metadata found in Zema itself, such as the singer’s name could, give us additional information.

Declarations

Competing interests:

The authors declare that they have no competing interests

Conflicts of interest:

The authors declare that they have no conflict of interest

Ethical Approval:

We confirm that informed consent was obtained from all participants before their participation and recording, after they were informed of the study's purpose, recording procedures, and intended academic use of the recordings.

Funding:

Authors declare no funding for this research

Author Contribution

B.L.A. conceived and designed the study, collected and prepared the audio dataset, developed the methodology, conducted the experiments and data analysis, and drafted the manuscript. A.A.G. contributed to the research design, methodology, and critical revision of the manuscript. G.A. contributed to supervision, validation, interpretation of the results, and critical revision of the manuscript. All authors reviewed and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Acknowledgements:

Not applicable

Data Availability

The data used in this study are available from the corresponding author upon reasonable request

Code availability:

Not applicable

References

1. Peffers et al. A Design Science Research Methodology for Information Systems Research. *J. Manage. Inform. Syst.*, 52–55. (2007).
2. Ali and Siddiqui. Automatic Music Genres Classification using Machine Learning. *Int. J. Adv. Comput. Sci. Appl.*, 337–343. (2017).

3. Bilal Er and Aydilek. Music Emotion Recognition by Using Chroma Spectrogram and Deep Visual Features. *Int. J. Comput. Intell. Syst.*, 1622–1632. (2019).
4. Costa et al. Music Genre Recognition Using Spectrograms. In proceedings of the Internatinal Conference on System, Signals, and Image Processing., 151–154. (2011).
5. Jawaherlalnehru abd Jothilakshmi. Music Instrument Recognition from Spectrogram Images Using Convolution Neural Network. *Int. J. Innovative Technol. Exploring Eng.*, **8**(9). (2019).
6. Kassaye. Education of the Ethiopian Orthodox Church:Personal Reflection on Nibab Bet and Zema Bet. *J. Ethiop. Church Stud.*, 112. (2018).
7. Khamparia et al. *Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network* 1–9 (IEEE access, 2019).
8. Purwins et al. *Deep Learning for Audio Signal Processing* 206–219 (IEEE JOURNAL OF SELECTED TOPICS OF SIGNAL PROCESSING, 2019).
9. □□□□□, □. (2008). □□ □□ □□□□ □□□. □□□ □□□: □□□□ □□□□.

Figures



Figure 1

Saint Yared (□□□□□, 2008)

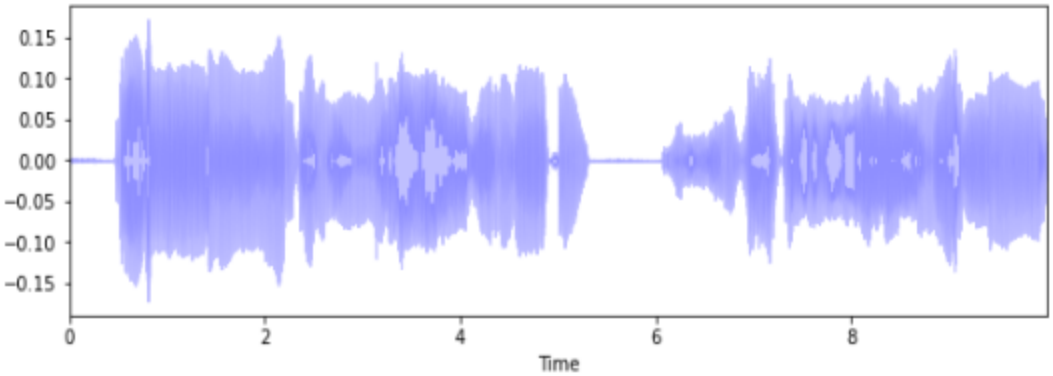


Figure 2

Waveform of audio data

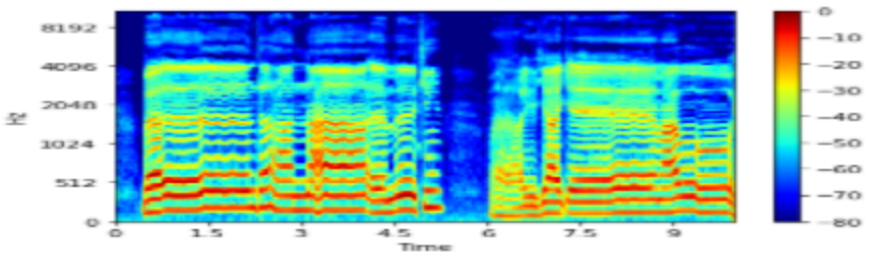


Figure 3

Spectrogram image of sample audio

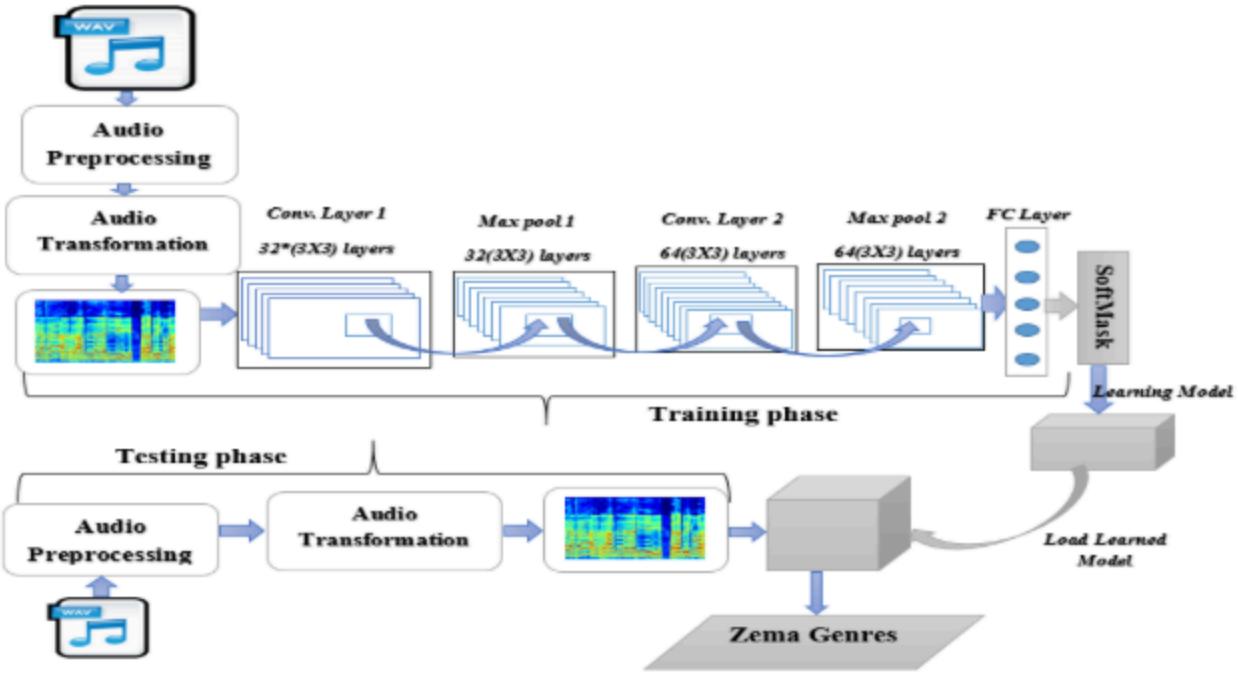


Figure 4

Designed SYQZC Model

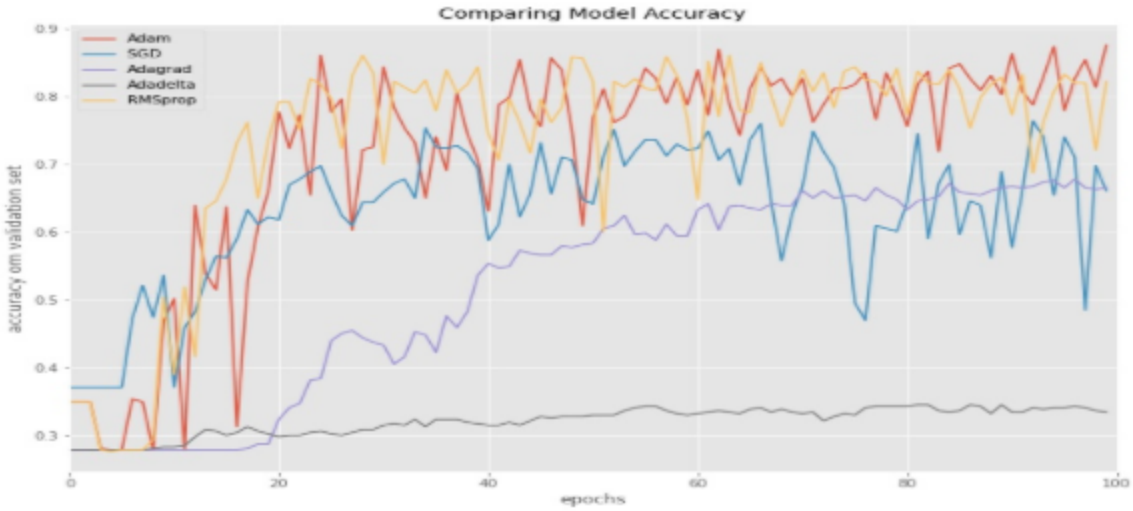


Figure 5

Comparison results of the SYQZC model with different optimizers

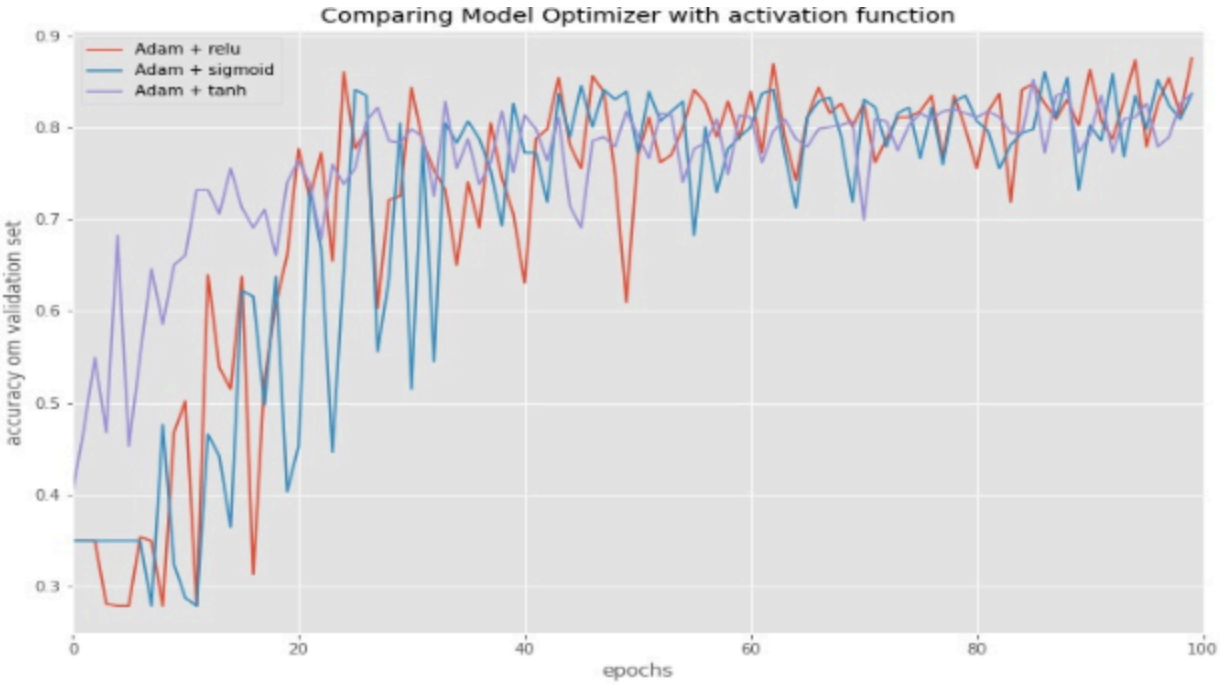


Figure 6

SYQZC model with Adam optimizers and different activation functions

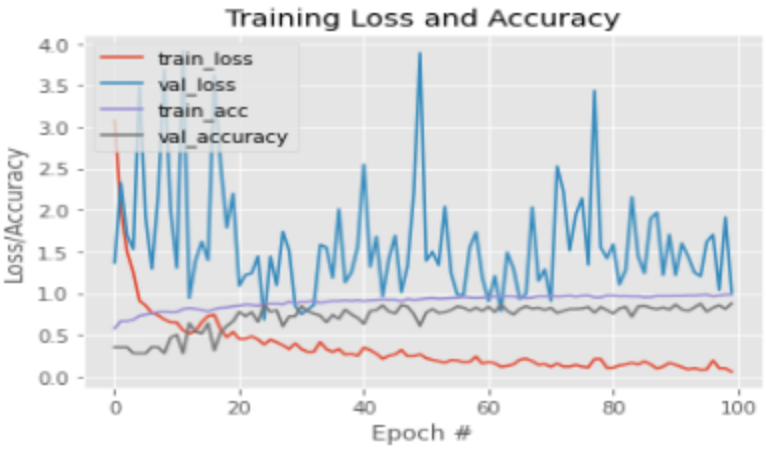


Figure 7

Training loss and accuracy of SYQZC model



Figure 8

Accuracy rate of training and test of SYQZC model

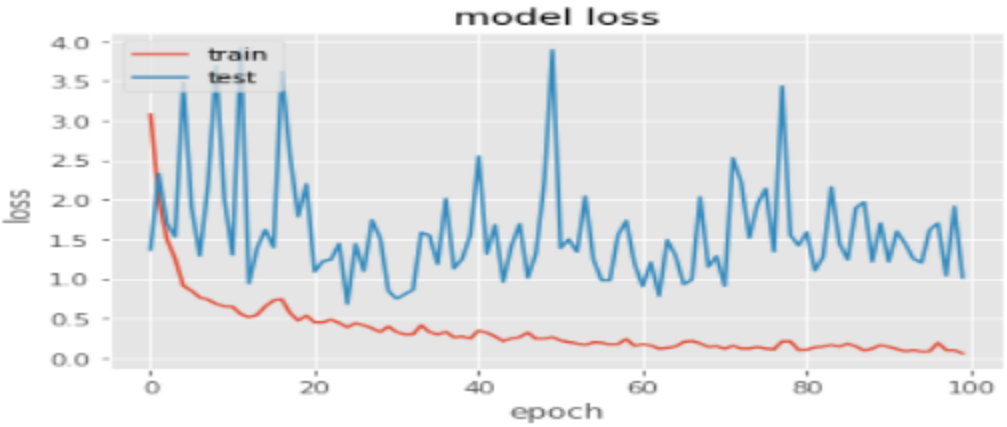


Figure 9

Loss rate of training and testing of SYQZC model