DSpace Institution

DSpace Repository **http://dspace.org**

Computer Science thesis

2021-07

SAINT YARED KUM ZEMA CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORK

BIRKU, LITGEB ASCHENEK

http://ir.bdu.edu.et/handle/123456789/12650

Downloaded from DSpace Repository, DSpace Institution's institutional repository

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

SCHOOL OF GRADUAT E STUDIES

FACULTY OF COMPUTING

MSc. THESIS ON

SAINT YARED KUM ZEMA CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORK

BY

BIRKU LITGEB ASCHENEK

JULY , 2021

BAHIRDAR , ETHIOPIA

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

SCHOOL OF GRADUATE STUDIES

FACULTY OF COMPUTING

SAINT YARED KUM ZEMA GENRES CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORK

BY

BIRKU LITGEB ASCHENEK

A Thesis submitted to the School of Graduate Studies of Bahir Dar Institute of Technology, BDU in a Partial Fulfillment of the Requirements for Degree of Master of Science in Software Engineering in the Faculty of Computing

Advisor: Mekonnen Wagaw (PhD)

July, 2021

Bahir Dar, Ethiopia

© 2021

BIRKU LITGEB

ALL RIGHTS RESERVED

Acknowledgment

First of all, my gratitude goes to our almighty GOD who supported us to perform every sequence of activity in this study and gave the bread of Saint Angels for Saint Yared. Next to this my gratitude goes to Dr. Mekonnen Wagaw, who gave me his full support of advices writing from the day of topic selection up to this point. I would also like to thanks Memhir Libsework Alemayehu, Memhir Serstedngle Wudu and Memhir Mengstu Fekadie who support me by giving their full dedication on St Yared zema compositions. I wish to thank Memhir Abrham Misganaw for his support from the early time of this research work up to the last with recording the audio zema with different time. My special respect goes to Yeshanew Ale and all my friends who support me until the completion of this work in several ways. At last my special gratitude goes to my family who are always with me by supporting me in all ups and downs.

Abstract

Machine learning approaches are applied in different fields of disciplines. The approach used in each area is implemented with a supervised or unsupervised learning method. The new and rapidly growing research area has emerged with the digitalization of music, called Music Information Retrieval (MIR), which emphasizes the extraction of information from music audio and musical notes. This recent technology focus on the categorization of the given audio music into several classes based on its characteristics. It is a researchable area which includes genre classification, song identification, chord recognition, sound event detection, and mood detection.

Zema defined as tactical shouting to produce a sweet song with zema notation for listeners. Zema classification is one category of MIR which is defined as the technique of grouping audio zema into appropriate classes. The first composer of spiritual melody was St. Yared with three Zema forms. These forms are Geez, Ezil, and Araray. He given six compositions of zema and stated its own features. Kum Zema is one of his compositions which is sung with only vocal sound, no instruments are used like that of Kebero, Tsinatsil, Mekuamia.

The main things which initiated us to conduct this study was most of the flocks as well as some disciples who passed with traditional school are not identified each zema genres properly. The knowledge gap between modern education and traditional on zema genres. Most study were carried out on classifying the data which doesn€t have inter as well as intra similarity between the dataset. The dataset is prepared from the recorded audio Zema taken from experts. Each audio zema segmented into an equal size of 10 seconds. The segmented audio Zema is changed into a visual representation form called a spectrogram.

We applied a convolutional neural network for classification, because it has better performance in image processing. So, the spectrogram with a specified size becomes an input for CNN, and each layer of the network filters the image. Features are also extracted from the spectrogram and finally, the SoftMask classifier classifies the input audio into three classes. The research method we used is experimental and the result obtained from our model, SYKZC, is 98% training accuracy and 89% testing accuracy.

Key words: Zema, Gubaebet, Aryam, Geez, Ezil, Araray, CNN and Genres

Dedicated

To Ethiopian orthodox tewahdo church, Ethiopian Traditional schools, Traditional school Scholars and their disciples.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ANN	Artificial Neural Network
CNN	Convolutional Neural Network
DIP	Digital Image Processing
DSP	Digital Signal Processing
E.C	Ethiopian Calendar
EOTC	Ethiopian Orthodox Tewahdo Church
FC	Full Connected
FFT	Fast Fourier Transform
GB	Gigabyte
GA	Genetic Algorithm
GLCM	Gray Level Co-occurrence Matrix
Hz	Hertz
Max	Maximum
MMH	Maximum Marginal Hyperplane
MB	Megabyte
MFCC	Mel Frequency Cepstral Coefficients
Ms	Millisecond
MLP	Multiple Perceptron
MIR	Music Information Retrieval
MP3	Music Picture Expert Group Layer 3 Audio (Audio format/ file extension)
NLP	Natural Language Processing

NB	Naive Bayes
RAM	Random Access Memory
ReLU	Rectified Linear Unit
RGB	Red, Green, and Blue
St.	Saint
SYKZC	Saint Yared Kum Zema Classification
TDSN	Tensor Deep Stacking Network
STFT	Short Term Fourier Transformation
SVM	Support Vector Machine
Wav	Window Wave (Audio format/ file extension)
3D CNN	3 Dimensional Convolutional Network

Chapter One: Introduction

1.1. Background

Nowadays, technology spreads in different ways and provides valuable information for society by producing ideal solutions to existing problems. Artificial intelligence in machine learning is applicable in different fields of studies. Automatic music categorization is one application area of artificial intelligence which is grouped under the category of Music Information Retrieval. The development of the knowledge on Machine Learning, researchers applied different approaches to automatic genre classification. It involved audio analysis tasks like music genre classification, song identification, chord recognition, sound event detection, mood detection and feature extraction (Nasrulla and Zhao, 2019). Zema genre classification is one specific task of automatic audio genres classification technology which is included under the discipline of music information retrieval; It enables the machine to recognize and classify melody. The study focuses on kum Zema genres classification to distinguish the types of zema classes based on the input audio data and recognize what type of Zema is based on the feature that will be extracted and identified from each sample of zema genres.

Zema or melody is defined as the manner of tactical shouting or sound generation which makes people happy when it is heard. In Yared music, zema is one of the divisions Ethiopian sacred music in Ethiopia orthodox church and we used it interchangeably with pleasing sound, song, chant ,and melody when it consist zema notations (Woube, 2018). Zema has tactical and formula to be song, based on its tactical and formula it is possible to say every Zema can be sound but the reverse is not true (Tadese, 2018).

The source of Zema is GOD himself and provides for Saint Angels to give glory to their creator and obtain prestige. It became diversified after the war of angels. This sweet melody was reached in our generation by the greatest Ethiopian orthodox zema composer named Saint Yared. He told to us there was Zema before him and such zema was song with Saint Angels in the heaven • @ ó Ü p 0 Ð ` , 0ͥ that means the first song is listened from the heaven (Habtemaryam, 2012). Before him the scholars of the churches were used a reading style which is still applicable in the celebration of Crucifixion with wurd nibab, but it didn€t have well structured and formalized way to be called zema because they weren€t used Zema notations as well as song standards at the time. Saint Yared also told that •ËíÜ Ø0 ¥©`0 í¥ ¥-uEñ3•¥•Øíe Eñ5 Eñ5 Eñ5¥ Ú e-8c¦uM9 (Eõ3p5ẹ, which means The Holy of Holy our Lord I heard the angels singing Your praises saying, Holy Holy o Holy parise that filled the Earth and the Heavens (Abebe, 1986 E.C). Basically Zema can be categorized into two types. The first one is spiritual Zema and the second is secular Zema. They have their own characterization and have great differences between them. Our study is focused on the spiritual one especially after Saint Yared was born, because he put great pressure on the occurrence of every spiritual and secular zema with their singing techniques.

Saint Yared was born in the city of Aksum on April 25, 505 A.D. Adam/Abyude was his father's name, and Tauklia was his mother's. Yared was a descendant of Aksum's priesthood. When Yared was six years old, his parents entrusted him to the tutorship of Yishaq, an Aksum teacher. Yared finished his alphabet studies and began studying the Psalms with this teacher. His teacher, however, sent him back to his parents since he was having problems learning his lesson. In the meantime, his father died, and his mother, Tauklia, entrusted him to her brother, Abba Gedeon, the parish priest, with the request that he raise and educate. Abba Gedeon was an Old and New Testament teacher in the courtyard of St. Mary of Tsion's church, and he had begun translating the Holy Scriptures from Hebrew and Greek into Geez. Yared moved in with Abba Gedeon and began studying alongside the other kids, but he was continually admonished and chastised by the new teacher since he lagged behind the others in his academics for years. Yared wasn't a particularly brilliant student, and no matter how hard he tried, he couldn't seem to absorb his lectures. His peers teased and mocked him because of his slowness of mind. His uncle brutally thrashed Yared one day, telling him, "You must not fall behind your peers and must pay attention to your academics as the others do." (Chavis, 2011)

Yared grew enraged by his failure as a student and resolved to relocate and begin a new life. As a result, he ran away from school, and while traveling to his uncle's birthplace, Medebai welel, he was caught in a strong rainstorm and forced to seek shelter under a tree near a spot called Maikerah, about four kilometers outside of Aksum. He watched an occurrence that would change his life while sheltering behind the trees, contemplating and feeling guilt for his failure. His attention was drawn to a caterpillar attempting, despite numerous failures, to climb up the tree stem to consume its leaves. Six times the caterpillar failed, but on the seventh attempt it fought valiantly and succeeded. Yared sobbed as he watched the caterpillar persevere, comparing his weakness to the grub's strength. After witnessing the tiny creature's might, he decided to return to school and resume his studies. He reasoned that man was a superior creature to a caterpillar, and that since the caterpillar had achieved its goal and eaten the tree's leaves through repeated effort, he, too, should suffer the repercussions of whipping, study attentively, and succeed. After making his decision, he went back to Abba Gedeon, his spiritual instructor, and requested to be forgiven and resumed his studies

Abba Gedeon finally gave in and started teaching him the Psalms. Apart from his studies, Yared visited the church of St. Mary of Tsion every day and prayed to God, saying, "Oh, gracious Lord, grant me wisdom!" God answered the child's petition by bestowing understanding and wisdom on him. His teacher was taken aback by his unexpected brightness. As a result of his perseverance and hard work, he was able to complete the study of the Old and New Testaments in a short period of time. Because Yared was now a talented student, he completed his studies with flying colors and went on to become a deacon. He had learned Hebrew and Greek from his instructor Abba Gedeon and was fluent in both languages. He surpassed his teacher in his understanding of the Holy Scriptures and mastery of foreign languages. Even though he was only fourteen years old at the time of his untimely death, teacher's Yared assumed the chair and profession of his tutor and began providing lessons (Belai, 1991).

When we come to Saint Yared profession which is compositions of spiritual Zema prepared with the leader of the Holy Spirit. He introduced his first Zema by standing in front of Axum Tsion by saying

• ë e ë È õ ë È • H 5 E ñ5 @ ó =î • 0 è #((È ` ó - î 4Ø ¨ e- e+ ð e p+, which means Praise be to the Father Praise be to the Son Praise be to the Holy Ghost Prior to Tsion God created the Heavens God showed Moses the tent with Araray Zema, and he named it - ë Aryam (Abebe, 1986 E.C). During that time some church scholars said we didn€t accept him but some of them accepted and followed him because the rhythm of the sound was very interesting for the listeners. St. Yared compositions are the only and unique resource of Ethiopian Orthodox Tewahdo Church and our country Ethiopia which is not found in other countries and religions even if it is not found in the remaining sisterhood orthodox tewahdo churches.

The aim of this research is classification of Saint Yared kum zema genres from the recorded audio data after transformed into spectrogram images. The genres of zema are three in number and namely Geez, Ezil and Araray. They have their own characteristics which makes one different from other. So, classification of each genre of zema enables the flock to distinguish the classes of zema during the time of singing as well as from the visual representation of audio.

We will apply a machine learning approach specifically convolutional neural network to categorize the given kum zema collected from experts with audio form. The audio zema will be segmented with fixed size in seconds and each segmented audio is automatically converted into visual representation form named as spectrogram. The Convolutional neural network took the spectrogram images as an input, each layer of the network filters the image with different filter size and a fully connected layer puts the spectrogram with one raw. Finally, SoftMax classifies transformed spectrogram images into proper classes.

1.2. Statement of problem

Many years ago traditional schools were the only academic institution that offered different courses for disciples to enable them to be knowledgeable as well as become creative in several disciplines. During that time most of the flock sent their male child to attend those spiritual academic institutions, through time, the flock reduced their interest in traditional school (Yekolo timhrt) and started modern education even if its source was the traditional school.

Nowadays that trend is highly reduced and has little motivation in the town areas. Due to such events for the coming generation the traditional school scholar will be highly decreased to support and provide attention for these institutions this study needed especially for Zema scholars. The other one is that genres of St. Yared compositions are identified by the school experts and their disciples only. Most of the remaining flock are unable to identify the genres of zema. Additionally, it helps minimize the generation knowledge skill gap between the traditional school teaching particularly in zema bet and the modern education to have combined knowledge in both institutions. In zema Gubaebet, the main source of Zema who, told to the disciples is the expert/teacher and the expert doesn't teach for all the disciple's little gap will occur because the traditional school scholars teach their successor and then successor will teach disciples who are found below their level.

In order to address the above statement of problem we formulate some research questions that supplement and delimit what we will do in the study.

- Which tune from Zema Gubaebet song with Geez, Ezil, and Araray?
- What methods will be utilized to extract features from the audio spectrum in order to classify St. Yared zema?
- How to develop a classifier model that categorizes each zema genre?
- How to improve the accuracy of the zema classification using features that are extracted from visual representation of the audio?

1.3. Objectives

1.3.1. General objective

The general objective of this study will be classification of Saint Yared Kum Zema genres using machine learning approaches.

1.3.2. Specific objectives

- Preparing dataset from recorded zema from zema expert
- Select an appropriate deep learning algorithm that can construct a model to classify St Yared Kum zema.
- Develop a model using the selected deep learning algorithm.
- Test and evaluate the performance of the model.
- Compare the performance of our model with the existing models

1.4. Significance of study

The significance of the study will be described in two ways.

The first one is from the perspective of practical contribution. It will provide the following application.

- It will provide motivation for the existing zema scholars as well as for their disciples
- It offers supportive information for flocks who have interest in the traditional school.

It minimizes the generation knowledge gap between modern education students and the traditional school disciples to have nearly common understanding about St. Yared compositions.

It enables any interested group as well as foreign tourist such have some knowledge about Saint Yared zema types

The second one is from scientific and methodological perspective it will opening direction for the coming researcher to apply different approaches to enhance and obtain better result in Saint Yared zema classification like Aquakuam, Zimmare, Mewase€t, Deggwa, Tsome Deggwa and Kidasie with similar approach.

1.5. Scope and delimitation

This study will be bounded with classification of kum zema of Saint Yared. When we see the types of íu` /culture of zema there are around four types of culture. They have their own characteristics and ways which make one differ from others. These are dp /Betelhem, F /Qomie, +e-/Achabir and p /Tegulet each types of culture of zema have foundation area. In Ethiopia Orthodox Tewahdo Church the most dominant and provided in most scholar Gubaebet is Betelhem due to such reason our study concentrated on Betelhem kum zema classification. So, the study will only focus on classification of Saint Yared kum zema with visual representation of audio files after being transformed into spectrogram images. The compositions included Deggwa, Tsome Deggwa, Me€eraf, Zimmare, Mewase€t and Kidasie with zema. This study will include Deggwa, Tsome Deggwa and Me€eraf because each composition has nearly similar song gloss and the remaining zema have their own zema rhythm.

This study won€t concern the video and any textual types of input data in the data. The type of Saint Yared zema that is called Aquakuam is not included because it is not kum zema the song will be performed with zema instruments and have different behavior in zema classification. Additionally, Kidasie, Zimmare and Mewase€t are not included because they their own singing gloss

1.6. Organization of the Thesis

This thesis will be organized with different chapters that help us to study the details of the research, so it will be organized with five chapters, each chapter specifically focused on the main activity of the research. There are five parts remaining in this report. The following is exhibited a framework of the substance canvassed in every section requested by the chapter number:

Chapter 2- Review of Existing Literature: This Chapter explores previous research work in Saint Yared composition audio classification, music genre classification, audio event detection, audio feature extraction with machine learning and deep learning algorithms. It discusses the application of machine learning algorithms in music information retrieval, multimedia content management and retrieval. This chapter also will discuss related works: This section mainly focuses on the studies that were conducted before and have some relation with algorithm usage or method which follow. Additionally, the chapter also highlights and provides brief on strength and gaps in the previous work. The Chapter also points the way forward by summarizing the state-of-the-art techniques in audio classification.

Chapter 3- Model Design and Methodology: This section outlines the details of the research design approach regarding methodology, experimental setup, orderly information of work process and data preparing stages. It provides details on all the significant steps taken that structure the premise of this investigation and their precise execution. Specifically, it covers the data collection, description, preprocessing investigation, feature extraction and finally the classification.

Chapter 4- Result and Discussion: This chapter gives an inside out explanation of the experiments performed as part of this research work. It centers around the implementation of the model including details on model training, tuning and performance. This chapter briefly also describes the details of the comparative model built on specification from previous research work. All the more, the implementation of the machine learning algorithms, comparison between the models and conversion of audio files to images is exhibited in this section. There is also evaluation to evaluate the performance of testing and assessment of the methodologies utilized by analyzing the results of the experiments conducted to classify using different machine learning models trained on the same datasets. It reasons that the work done by this research work is able to classify audio as intended and the performance of the classifier can be measured in terms of various performance metrics, accuracy, precision, f1-score, confusion matrix. Average scores were also calculated for precision, accuracy and F1-score to estimate the overall performance of the model.

Chapter 5: Conclusion and future work: This chapter covers the general accomplishments of the research work and highlights the future work that could be developed later on. The section also gives a conclusion and review of the experiment conducted in this research work. The section moreover outlines the recommendations for heading of future work.

Chapter Two: Literature Review

2.1. Introduction

This section mainly includes two components of the research. The first one is literatures review section which focus on the detailed information about the study including background of St. Yared, Compositions, Types of zema, Zema notations used during song, representation of recorded audio zema into waveform as well as spectrogram image with DSP & DIP. The techniques applied to carry out the study, the overall flow of steps we follow and finally, the metrics we used to evaluate the accuracy and performance of the designed model. The second one is related works which were conducted before this study related to the approaches and techniques which the researcher used related to our study

2.2. Literature Review

Audio signals are something which involves voice, music, and ambient sounds which are important media forms of communication. Humans can easily differentiate various types of audio sounds by simply listening to a short section of an audio signal (Karthikeyan and Mala, 2018). Each category of audio signal is defined in sub-category detail and categorized as disciplines with their own unique characteristics. Classification is described as the way in which an individual object is automatically assigned to one of several categories or classes, based on its characteristics. Music genre is defined as a music style that has common characteristics shared by its members, and can be differentiated one from other music styles. Such characteristics are typically associated with the music's instrumentation, rhythm, harmony, and melody (Nasridinov1 and Park, 2014). This study will consider the additional features that enable us to clearly distinguish the form of the class of zema including use of notations zema.

Classification of music genres has been an exciting as well as difficult task in the discipline of music information retrieval and genre classification which can be useful in explaining some very interesting issues such as creating song references, finding similar songs, finding communities that want that particular song (Asim and Siddiqui, 2017)The Ethiopian orthodox tewahdo church traditional education bounded with two main streams and provided independently. These are Nibab which means reading and zema which is religious music (Woube, 2018)The first one Nibab bet or reading is one stream of the education which emphasizes on reading and learning by heart the prayers of St. Mary and Jesus Christ, the psalms of David and the Gospel of John whereas Zema bet or religious music consists of the following branches which are Me€eraf it means chapter and cannot be employed alone, but always with the other chant books, Tsome Deggwa or chants of the main fasting, Deggwa or the main chant book, Kidasie or liturgy ceremony of the holy communion, Zimmare or songs sung at the end of Eucharist and Mewasit or songs related to commemorative services and funerals, and Aquaquam or religious dance and movements in which drums and sistra are studied in this school.

In the schools of Zema bet or school of music includes Me€eraf, Tsome Deggwa, Deggwa are studied; in Kidasie bet or school of mass music liturgy, Zimmare Mewasit bet and Aquaquam bet (Woube, 2018)Our study is concentrate on the second categories of church education or religious music specifically kum zema classification by applying machine learning algorithms and recognizing each class of zema whether grouped under spiritual zema or secular we can focus on spiritual, since spiritual zema have their own classification category according to the rule of Saint Yared, who is the greatest zema composer.

There wasn€t song of hymns and structured spiritual zema that are song in loud voice with well-defined tunes before Saint Yared was introduced spiritual zema, but men murmured in a low voice and God wishing to raise up to himself a memorial sent into him three birds as we said before from the Garden of Edom. They kept conversation with Yard in man's tongue, and took him to heavenly Jerusalem. He learned the songs of the four and twenty priests of heaven and angels. Saint Yared composed a hymn in three modes for each season of the year, including summer and winter, spring and autumn, festivals and Sabbaths, and the days of angels, prophets, martyrs, and holy people (shelemay, 1982) Spiritual melody provided from God to Saint angels and Saint Yared to indicate his forgiveness to Adam's child and protect them from sin. After the angels€war this special zema was diversified and divided into two categories. There were secular zema and spiritual zema as the Bible told us we heard from the devil's about the secular zema. The sacred zema is told by the Holy Spirit with the song of this kind of sweet zema for holy Yared. Human beings sing this kind of sweet zema together with Saint angels at the time of Jesus Christ was born in Bethlehem. This is described with • 5eu ¥ Ú e-`0 ëuÈ0 È õ-5 (q 0e¥ that means Praise be to God in heaven, and let the peace of man be upon earth.

Various researchers around the world have conducted studies on secular and spiritual zema, concluding that the source of spiritual zema is St. Yared and all musical traditions around the world, with fruitful comparisons to medieval European zema possible. The Ethiopian Orthodox Church's zema tradition promises to be especially insightful (Shelemay et al., 1993). From the existence of the world up to the birth of the Saint Yared, there was no well structured and organized zema that the follower of faith, as well as the priest themselves, had no zema, simply they were used reading style like now we apply in the celebration of what we call it Siklet / Crucifixion with wurd nbab. This form of reading is still used in EOTC.

2.3. SaintYared

The Ethiopian Re'ese Liqawnt (head professor) Saint Yared was born from his father Abyude or Isaac and his mother Christina or Tawklia in Axum in 505 E.C. (Abebe, 1986 E.C). His father died when he was a baby and his mother sent him to a traditional school of Ethiopian orthodox tewahdo to attend the teachings which the church scholars offer. Since his father was the scholar for traditional school, his uncle named Gedyon took him and taught, but he was overwhelmed with the teaching because what he learned was incapable of understanding and passed through the next school level. The scholar instructed their disciple verbally according to the school rule, and they listened carefully to what their instructor was telling them and revised it based on what they had learned. He tried to attend teaching for seven years, but nothing was changed on his education stage without knowing something. Seven years after passing the challenging case, he got a new as well as an unusual form of zema and was named as Saint Yared zema because of his patency. Such a gift was provided from the Holy Spirit to him in the form of the three birds; these birds were Saint angels as shown below.

Figure 2.1 St. Yared when singing zema from Tsome Deggwa (Mengstu, 2008 E.C)

Yared was a composer and choreographer in Aksum during the sixth century AD. St. Yared is credited with the Ethiopian zema tradition, particularly the zema of the Ethiopian Orthodox Tewahdo Church. He is credited with originating the Church's song, and his zema has been used for about 1500 years (Ayele, 2007).

2.4. Saint Yared Zema compositions

Saint Yared offers Six basic compositions or services for followers of the Ethiopian Orthodox Church religion and our country. These are Tsome Deggwa, Deggwa, Me€eraf, Zimmare, Mewase€t and Zema Kidasie with their sweet Zema. He rearranges the time schedule in a structured form, the genre of zema and the zema notations which guide the scholars who follow his roadmap. Deggwa and Tsome Deggwa are books of Zema used for Church Festivals and Sundays. Whereas Tsome Deggwa books include zema for the Main Lent or fasting season particularly in Abiy Tsom, holidays and daily prayers, praise, and zema. Deggwa derives from the word Deggwa which means zema of sorrow and tearful songs are written. Deggwa is also often called Mahlete Yared, or Yared songs, remembering Yared authorship of the zema. Scholars write about the importance of Deggwa, while it was described in the general form of poetry, passages related to theology, philosophy, history and ethics. The Book of Me€eraf, Sabat Zema, essential holidays, daily prayers and praises; even zema for the fasting month. Zimmare, includes zema to be sung after Qurban or offerings after Mass. Zimmare was written at monastery Zur Amba and Mewase€t Poem, Zema to the dead, along with Zimmare, Yared wrote Mewase€t. Kidasie Novel, zema for blessing Qurban offering (Ayele, 2007).

In zema bet Deggwa scholars teaches their disciples starting from the introductory part of Wudasie Maryam zema called 0• ª¥ • Ø • 0 õ• e ª - ë, which ¥• means We bow to you, Mother Mary, peace be upon you up to Deggwa via the sequence of step with three form of zema. Wudasie Mariam is the first part of the curriculum to be studied in Zema bet (Woube, 2018). As we stated before the initial zema sings only with two forms Araray and Ezil (Habtemaryam, 2012). St. Yared sang Wudasie Maryam zema starting from Monday to Saturday with Araray and the Sunday with Ezil. The next traditional school disciple learned Mestegab, which means collection. Most of the songs under this category are taken from Psalms of David. It is usually sung during the main fasting season known as Abiytsom. This type of zema sings with Geez and Ezil zema. Aryam is said to be composed by Saint Yared after he went to Aryam, one of the heavens and listened to the singing of Angels while praising God by emphasing kidus, kidus, kidus. Most of the melodies of the songs are based on Geez and Araray. Selest means the 3rd. It is employed after the 3rd line of Psalms of David and songs with Geez and Araray modes zema. It consists of the hymnography proper to the Sundays and Weekdays of the great Lenten season and beyond. It includes Holy Week services and the night of the Resurrection as well. The main parts of Tsome Deggwa are: - Zewerede, Ze-kidist (of the Holy), Ze-mekurab (of the Temple), Ze-metsagu (of the paralytic), Ze-debrezeit (of the mount Olives), Ze-gebriher (of the good servant) and Ze-niqodimos (of Nicodemus), Ze-hosaena (of Hosanna - Palm Sunday) and the last one is Deggwa covers the whole year of church server. It consists of Johannes, Astemehiro and Fasika (Woube, 2018). There is a sequence of providing a zema course for disciples until they reach the final lesson of zema called Deggwa zema. Deggwa zema is sing with three form of zema we can see one example which is sing with three zema forms • c (- p, or bahire girmte, •p c, ;or tegabau and so on.

2.5. The three types of Saint Yared zema

Saint Yared categorized every spiritual zema with three groups out of three there is no song called zema chant. Deggwa supports this proof with • f ¥ E õ È f ¥ õ … , sí 0 e ¥ Ø ¨ ë , õ « • í ¬ Íó 4 ¥ • 5 3 È 0 e ¥ È + Ê è õ É ¥ 5 Ü that means No One before and after like him, Yared chanter and clergy man had sounded his compliment; animal, human being and wild lives will not get out from his three melodies (SENKORIS, 2018). He prepared Deggwa in the three chanting modes used in the church and known as Geez, Ezil, and Araray, respectively. To ordinary days, geez means simple chant; Ezil means a more measured rhythm to funerals; Araray means a lighter, relaxed mood for major festivals. During Emperor Gebre Meskel's reign in 505 E.C. Yared compiled the Megabi Deggwa meaning the hymn of sorrow that included three main modes: Ge'ez that is the first stage of the song, Ezil is the second stage to be sung along with the first and the last one is Araray a sad and plaintive song (Hazen and Daoud, 2014). We can see sample zema from Wudasie Maryam with Araray and Ezil for similar words but zema notation and types of zema are different as shown below.

Figure 2.2 Sample for zema notation in Wudasie Maryam (Tadese, 2018)

It is possible to define the types of Saint Yared zema with different forms of function that the songs have, Ge'ez, first and straight note. It is described as hard and imposing in its musical style. Sometimes, scholars refer to it as dry and devoid of sweet melody. Ezil, melodic, gentle and sweet note, which is often sung after Ge€ez. Araray is the third, melodious and melancholic note, often sung on somber moments, such as fasting and funeral mass, it is also described as an affective tone suggesting intimacy and tenderness (Assele, 2007).

Figure 2.3 Sample notation of zema (Tadese, 2018)

2.6. Names and signs of St. Yared zema Notations

St. Yared introduced and song his first form of zema by standing in front of Axum Tsion with Araray zema and calls it - ë/Aryam. Initially, one song is said to be zema when it follows various attributes that identify it as having unique characteristics as well as function, particularly Saint Yared zema. One feature that makes Saint Yared zema different from other forms of sound or music zema is zema notation as well as the model of song. At the first time St. Yared introduced eight types of zema notation which enabled him to guide the song within a formalized method. Almost all the chants are written by the Yaredic notation model. In this model, signs known as Milikets are put on the top of the lyrics. We can see from figure 2.2 and 2.3

Basically zema notations or Milikits are divided into two major categories: abbreviated words (Sirey) and basic Milikets. There are probably more than 900 abbreviated words. Milikets are accentual signs such as curves, lines, dots and other symbols that are usually helpful in directing the melodies. Sirey could be taken as the abbreviated letters that denote simple sounds or stand for groups of successive phrases of melodies. In other words, they designate melodic patterns i a kind of shorthand. Both appear in the manuscripts in combination (Woube, 2018). In other ways notations are named as Non-alphabetic notations and Alphabetic notations. As we said the first types of zema notation were introduced by St. Yared himself and next to him there were several disciples who followed his teaching. The alphabetic zema notations were formulated by his disciples next to St. Yared based on Non-alphabetic zema notation as well as several orthodox scholars adding different notational representation. At the first time Saint Yared incorporated the following notations. These are õK/Difat, ð/Hidet, E“/Qinat, íØ/Yizet, A -/%Qurt, -(u/Chiret, ð (/Deret and - - -/Rikrik & õ - 5/Dirs and • e/Anbir respectively. The notation and the symbol are given as below.

Figure 2.4 Non-alphabetic and alphabetical zema notations (Girma, 2014)

Any type of Saint Yared zema is singing with the above type of zema notation and each type of zema notation has their own description with regard to the Ethiopian orthodox tewahdo scholars.

ðØ í Ø/Yizet -- detached and accented tone. Derived from the word hold or meyaz, to be said equivalent to staccato.

ðØ ð (/Deret ---sing in a low, deep voice. The chest register also applied to singing with closed lips and deep chest resonance with clenched teeth. Humming at the male voice's lowest rage.

ðØ E “/Qinat ---upward raising of voice. The term derived from the verb makenat

ðØ -(u/Chiret----start high and proceed with downward glissando. The vocal melody of Chiret is also related to a cadence. The word is derived from chira, tail

ðØ õ K/Difat --- Drop the voice. Skip to a lower range. This also applies to singing an octave lower. Medfat, to throw down, is the root verb.

ðØ A -/%Qurt ---- Its root word is mekuret, to end or to cut. Equivalent to coda.

ðØ ---/Rikrik ---- rapid repeat of a single syllable tone. This type of singing usually creates a sense of tension in the high range. Equivalent to tremolo.

ðØ ð/Hidet ---gradually getting faster and louder. Sing each syllable distinctly. Equivalent to accelerando and crescendo at the same time. The remaining two are added after Saint Yared with traditional school scholars not only two more than it.

According to Ethiopia orthodox tewahdo church Saint Yared zema notes have their own interpretation for what purpose they are used and what event it shows related to God. Deret represents Jesus Christ's resurrection and ascension, Difat Jesus came to this earth, Qurt Jesus decided/promised to save Adam from death, Yizet Jesus was captured and beaten by the Jews, Qinat Judah gave Jesus to the Jews, Chiret Jesus Christ was beaten, Hidet Jesus was taken to Hanna and to Pilatos, Rikrik is the final which the prophecy told by David was realized on Friday specially the blood flew from his body (Abebe, 1986 E.C) The above eight types of Saint Yared zema notations are grouped under non-alphabetic that means each notation doesn€t relate with the name simply represented with symbols whereas the alphabetic notation are related with the name that is given for itself and next to Saint Yared several traditional school scholars add several types of zema notation inside the above eight notations and by adding this notation Saint Yared zema become more popular and acceptable.

The scholars used Sreyor ½(í zema notation types which means the root for singing zema and represent each notation with alphabetic symbol by relate the symbols with its name. Zema notation of every spiritual song is derived from Deggwa notation. During the time of St. Yared as we said there were only eight types of notation next to him his successor disciples Hawira/Ế +, Sawira/ 3 Ế,+ Eskndra/¥ 5 - • /õ+, Foesknedra/V¥ 5 - • /ã+ and Abidira/ bò + plays significant role on preparing zema notation for Deggwa and another traditional school scholar named as Û å +“ Û å+/Azzaz Gera and Azzaz Raguel, they were Ethiopia orthodox tewahdo church zema scholars in Tedbabe Mariam and lived in the regime of Atse Gelawdewos. The written history indicates the event by saying • È` ËÕ • 5 ÉòÎ5 p• 5 ¡ Û å +È Û å+ ¤ « “u ¥ +• Ü ÈÈ ‘ ÉE ¦ , (Habtemaryam, 2012) (Tadese, 2018) Generally, the added zema notations Anbir, and Dirs had near age with Saint Yared but notations like 0 ª/Selam leki(0)ª,

¥•Ø•0 õ•ª/Emzensegd nibleki(¥•)ª, ¥-Ì• /Emarwie neawi(¥-Ó), Ê a`í/Bubey(a), Ñ/Aqu(Ñ), í¨õ•ª=î/Yikednki tsiyon(• )ª, í'Ø- /Aeynu zergb(Ó)í, `Õ•F0/Enqo senper(T), • /Aneni(•), ¥ Ú¦p0 /Egzio tesehalene(Ú¦), 3/Selasa(), •p/Minte(•), % /Atimek(@), …Ù•Ø u/Hizune ewotkuze lib(e), €`Ø u«u /Habe zetkat menberu(«u), (/Semre( ), í5 /Yishalene(5), dp /Betelhem( ), Ë/Lehewan(Ë), • 0 /Seali lene(0), õƒ'u/medhanite(ƒ), u0 /tshali(u5), •rÉ¥/Anti wetu(•É¥), "•Ü/Nahu nizienu("), ®•ª/Konki(®•)ª, È@ó p õƒ/woqedamite medhanitne( ), @Èë5 t/ðastedelu(ô), etc. were added after Saint Yared ,but it doesn€t change the overall style of zema notations which was formulated by St. Yared and found in Wudasie Maryam zema name as Srey.

Any type of Saint Yared zema sings using the above zema notation and to sing zema from the beginning of one letter /word to another letter/word Saint Yared uses tonic vowels that are used for concatenating each word/letter of the song. These are mentioned with researchers by said the seven Ethiopic (G„z) vowels (referred to as 'orders' when combined with one of the thirty-three basic symbols in the „'G„z syllabary) are represented as„ u, i, a, e„ and o (Shelemay et al., 1993).

Several studies on music information retrieval have been conducted with different methods and the mechanisms that researchers classify each type by taking different classes from several attributes of the music. Similarly, we apply such techniques for our study to classification of Saint Yared zema especially on kum zema. Here when we say kum zema each type of zema doesn€t need any types of instruments during the time of singing. If there are instruments like Drum, flange and Tsinatsil during singing of zema it is the Mahlet and called Aquakuam zema.

2.7. Formation of Spiritual Zema

As we said early zema can be described as the method of producing sweet sound or shouting mechanism that allows the listener to respond to excited filling and has the ability to make the broken heart or feeling of sad people into suitable condition. The spiritual song used for physiotherapy purpose like Saint Dawit and it have their own schedules, that means it is time dependent for example Tsome Deggwa are often used in fasting time, Zimmare Mewase€t used when human beings die and Kidasie is used every day in the Ethiopian orthodox tewahdo church there, but some of it is date dependent example if date is 21 the Saint marry Kidasie will be sing. When zema sing it uses zema notation and the notation can be generated with the inner or outer part of our body like the image given below.

Speech or music is produced when air flows from lung to exterior through mouse and noise. There are plenty of physical components of speech production in human organs, mainly the following are listed. These are lung, trachea, larynx, pharynx, oral and nasal cavity and vocal folds and Human speech begins with the vocal cords (folds). Air forced up by the lungs passes over the vocal cords, causing them to vibrate at certain frequencies, depending on the force of the air and the position of the vocal cords. At this point the fundamental frequency of the speech is formed and then modified by the soft palate, tongue, lips, and other parts of the vocal tract, filtering out some frequencies and adding additional frequencies which are the integral product of the fundamental frequency and it is known as formant frequency (Girma, 2014). The way of production of sound is shown as follows.

Figure 2.5 Formation of zema and sound (Girma, 2014)

2.8. Formation of Secular Zema (Music)

The traditional school of Ethiopian orthodox tewahdo was used as ministry of education with different fields of study and fighting illiteracy over the past 3000 years (Mezmur, 2011). As we stated before different fields of study are there in the church from them zema is one field provided in the school. The foundation for each zema is Saint Yared specially for spiritual zema with gradual sequence of time some individual who had background knowledge on zema wants to express their personal feelings with zema, even the Ethiopian orthodox tewahdo church scholar uses a specific zema to admire the flock who invites them in zema like ceremony called Mahlete Genbo, but secular music is even different from Mahlete Genbo because it doesn€t contain as well as follows zema notation that used in Saint Yared. Spiritual zema and secular music have common/similar attributes that are shared together like timbre, rhythm, pitch and tone. Each genre has a separate approach for determining which features are used for which classes.

2.9. Digital signal processing

The mathematical algorithms, and techniques used to control signals after converting into a digital form is called digital signal processing (DSP). uses a digital form of signal to have communication in the environment with the usage of unique data named as signal (Smith, 1999). It uses digital processing to perform a wide range of signal processing operations, such as computers or more advanced digital signal processors. DSP is mostly used in audio signal arenas, speech synthesis, radar, seismology, audio, sonar, and voice recognition signals. Signals could be continuous (analog) as they exist in digital devices such as computers, either naturally or digitally. Computers can only store and process signals in digital form. Therefore, image, audio and video signals need to be converted to a digital form before they are stored and processed by computers. Digital signal processing is applicable in different fields of studies like space, medical, commercial, telephone, military, industry and scientific areas (Smith, 1999)

2.10. Audio signal

Audio or sound is one of the main sensory information we receive to perceive our environment with our sense organ with auditory and audio signals emitted from us with our mouth and released into the environment from the environment different receptors will perceive it. Nearly every activity or occurrence in our world has its own unique tone. There are three key properties of audio that allow us to differentiate between two or more sounds. The first is Amplitude, which implies the sound's loudness, the second is the frequency that implies the sound's pitch, and the third is Timbre, which implies the sound's consistency or identity. Deep learning is one of the most advanced methods for categorizing audio signals, and it uses several algorithms to classify anddistinguish sound, music, voice, and various environmental sounds (Purwins et al., 2019.)

Audio signal processing is the area of engineering which relies on analytical methods to intentionally modify auditory signals or sounds to achieve a particular target. Music Signal Processing is a digital Signal Processing branch and a very large and complex subject in its own right. In short, as the name implies, the analyzing, analysis and transformation of analog music signals, or effects signals in digital music. Signal Processing is the art and science of modifying, for analysis or improvement purposes, the data obtained from the time series. Examples include spectral analysis using Fast Fourier or other transformations and data acquisition enhancement using digital filtering and image processing due to the defined image signal and spectrogram. The efficiency of a set of characteristics depends on the application. Therefore, the key challenge in designing audio classification models is the creation of descriptive functionality for a particular application. In reality, audio tells a lot about the clip's mood, the music part, the noise, the speed or slowness of the pace, and the human brain can also classifyonly on the basis of audio (Karthikeyan and Mala, 2018.)

Figure 2.6Waveform representation of signal

2.10.1.Signal Terminologies

A signal or wave form is an amount that varies with time or space and that generally transmits data. The distinctions between analog versus digital and continuous time versus discrete time are also made when addressing waveform processing problems. These terms are sometimes used interchangeably; the two sets of terms should be credited with different definitions. Signals are emitted from the source with the form of sound and the sound has its own components like pitch, loudness and timber.

Analog signal:- This determines the waveform that is continuous in time and belongs to a class that takes on a continuous amplitude value spectrum. Analog wave forms or analog signals are derived from acoustic sources of data. The signals are represented mathematically as a function of continuous variables. Analog signals are continuous time with continuous amplitude (Smith, 1999).

Digital signal:- implies that both time and amplitude are quantized. In digital models the signals are represented as a sequence of numbers which takes only a finite set of values. These types of signals have continuous time. As we know computers understand any form of input with numeric value which means in the form of 0 and 1 (Smith, 1999).

ðØ Frequency:- it is used to measure the strength or loudness of audio with the given specified time

ðØ Pitch: - it is the frequency in the sound of the fundamental variable, which is the frequency with which the waveform is repeated.

ðØ Loudness:- is a sound wave volume measurement

ðØ Timber:-is more complicated, being determined by the harmonic content of the signal

ðØ Mel spectrogram:- is a spectrogram where the frequencies are converted to the Mel scale. Fourier transform:- is a mathematical representation of sound that takes a time domain signal as input and decomposes it into frequencies as output. It is a mathematical function that converts the shape of a signal into the time and frequency domains of representation.

2.10.2. Audio zema acquisition

Audio zema acquisition is the first phase to conduct study since our initial input is audio zema. It is the first step for audio signal processing and concerned with gaining of audio files used for the study with different audio file format and translating the signal to spectrogram image so it is the key part of the study, unless no processing is possible. It is the process of taking an Audio sound primary source of data by recording from traditional school experts using a sound recorder to record it properly in an uncontrolled environment as well as from secondary source of data that is recorded audio data.

In audio zema acquisition, we gathered audio records from Abay Mado Debre Abay Saint Gebreal Monastery Zema Gubaebet Deggwa scholars and from other scholars. The first scholar is Merigeta Libsework Alemayehu who teaches the disciples in the Monastery and provides every necessary information about kum zema. The second successor Zema expert is Merigeta Abrham Misganaw who graduated from Bethlehem with Deggwa and teaches in the Monastery and plays a significant role in our study by providing records for each type of zema. Finally, Merigeta Mengistu Fekadie who graduated from Aquakuam and Merigeta Sertse Wudu help me by providing any relevant data for our study. We recorded audio from Liketebebt Teklie Sirak who is an expert of zema and Aquaquam. Audio can also be acquired from a database or another source tailored for research purposes that enable us to get some sample data that support our study (church, 2019) and (EOTC, 2003). Most of the time an audio data taken is unprocessed and requires further processing and analysis to be used for specific purposes.

2.10.3 Preprocessing of Audio

Audio zema data always needs to be preprocessed to have a refined form of Audio signals to ensure an accurate output prediction of Saint Yared zema class. The existing techniques in Audio classification and recognition literature have a lack of focus on preprocessing steps that effectively refines the data and assists in boosting the accuracy of the final classifier. In this paper, we will present a preprocessing strategy in which noises are extracted via a novel adaptive thresholding technique followed by the removal of silent portions in aural data and the long audio provided is segmented with a fixed time interval. For noise reduction we apply spectral gating technique to reduce the noise that occurs our data. This technique required two inputs, the first input is a noise audio clip containing prototypical noise of the audio clip and the second one is a signal audio clip containing the signal and the noise intended to be removed. Our preprocessing approach will play a prominent role in the overall classifier model of Saint Yared kum zema.

2.10.4. Audio segmentation

For most applications of audio analysis, segmentation is a very significant processing step. The purpose is to break an uninterrupted audio signal into segments that are homogeneous. When we claim homogeneous consideration of time. Here the audio files are split with equal time intervals which enable uniform variation time intervals because the difference of time will lead us to generate unrelated results of the spectrogram. So, here we apply a thresholding technique simply to assign the required time interval to chuck the given audio data into homogeneous segments with the time interval that we used to equal split audio file is that 10 second enough to recognize each category of Saint Yared kum zema class. We have seen research conducted on music classification and they take a time interval of 10 second even if they didn€t put their justification why they take this amount of time. Other researchers conduct study on instrument classification and sound classification take the time interval of 2 or 3 sec. In our study this amount of time is not sufficient because each class has intra similarity, the model doesn€t easily distinguish the given input data with a small amount of time due to this problem we take the time of 10 second. So to segment we use an audio file cutter as well as an algorithm that easily segments the input long audio file into equal sized segments of audio file.

Figure 2.7 the overall flow of classification

2.11. Digital Image processing

An image is described as a two-dimensional function, g (x, y), where x and y are spatial (plane) co-ordinates, and the amplitude of g at any pair of coordinates (x, y) at that point is referred to as the image's intensity or gray level.

When x, y, and the amplitude values of g are all finite, discrete quantities, we call the image a digital image. The field of digital image processing is concerned with the processing of digital images using a digital machine. A digital image is made up of a finite number of elements, each of which has a unique position and value. Such elements are referred to as dimensions of images, elements of images, and pixels. Pixel is the most common term used to refer to the elements of a digital image (Gonzalez, June 2019). As we have mentioned in the above, after converting the audio file into spectrogram the input which is fed for the model is spectrogram image. So, it must be proceeding with the technique of image processing mechanism. As we have stated, the spectrogram is that the two dimensional image becomes three dimensional when we include colors and four dimensional when it consists of the variable of colors. The image spectrogram also represents the x-axis and y-axis in which the x axis is the width of the spectrum and always the time interval is described and the y-axis the height of the spectrogram and the frequency or pitch of the zema is depicted.

2.11.1. Spectrogram

A spectrogram is computed from each music clip (with 22050 Hz sampling rate) through the short-time Fourier transform (STFT) with a window size of 1024 samples. The horizontal and vertical axis of a spectrogram represents time and frequency, respectively (M. Wu et al., 2011). It is a visual illustration of a signal's frequency spectrum as it varies over time. It is a visual way of describing the signal intensity, or loudness, of a signal at different frequencies in a given waveform over time. Spectrograms are two-dimensional representations that depict spectra sequences with time on one axis, frequency on the other, and brightness or color signifying the strength of a frequency component at each time frame (Wyse, 2017). Not only can one observe if there is more or less energy at a given frequency, but one can also watch how energy levels change over time. In other sciences, spectrograms are commonly employed to describe microphone-recorded frequencies of sound waves produced by humans. They are essentially two-dimensional graphs, with colors reflecting a third dimension. The comprehensive audio view, capable of representing time, frequency and amplitude all on one graph, is also defined. At the different frequencies present in a waveform, a spectrogram shows signal intensity over time. Spectrograms may be two-dimensional graphs represented by color with a third variable, or three-dimensional graphs represented by a fourth color variable. The color scale is red-green-blue, where low amplitudes or loudness correspond to blue, and high amplitudes correspond to red. A Spectrogram graph of the signal's energy content expressed as a frequency and time function. A graph of a signal showing the frequency of the vertical axis, the time of the horizontal axis, and the amplitude is displayed on a gray scale.

Several parameters like FFT-Length, Frame Size, Window Type and Overlap are selected from the Spectrogram Parameters command and can be adjusted when a spectrogram is generated in order to obtain the required time/frequency resolution and spectrogram bandwidth. A matrix of amplitudes is the digital spectrogram. A single pixel (picture element) of the spectrogram image corresponds to each amplitude of the matrix. The frequency resolution is the height of such a pixel. The pixel width is the temporal resolution. The total height of the matrix of the spectrogram is equal to half of the FFT-Length. The bandwidth of the spectrogram is not the same as that of the digital spectrogram matrix's frequency resolution. The bandwidth is normally higher than the resolution and is affected by the size of the frame and the form of the window. The spectrographic image resolution depends on window size and is a trade-off between the time and frequency domains, which means longer time windows provide increased spectral resolution (narrowband spectrogram) while shorter time windows provide increased temporal resolution (wideband spectrogram) (C. Knight et al., 2019)

Bandwidth is often determined by the window form. With the rectangular window, the smallest bandwidth is defined. The rectangular and Bartlett window cannot be used for normal applications due to the undesirable leakage effect (bad selectivity and spurious frequency components depending on the signal frequency. The lowest bandwidth can be accomplished with the Hamming window. The FlatTop window has the largest bandwidth. The following list is sorted by ascending bandwidths:(Bartlett, Rectangular), Hamming, Hann, Blackman, 3.0 Gauss, Kaiser-Bessel, FlatTop.

In general, if the signal to be analyzed does not have fast frequency modulations and if there is no important information in the time domain, narrow bandwidths should be chosen. In addition, if there is any remarkable frequency modulation or if there are noticeable temporal trends, large bandwidths should be selected.

Figure 2.8 Ways of audio file classification

So, our study concerning classification of kum zema with frequency domain representation is called spectrogram image.

We're familiar with seeing a waveform in audio software that shows changes in the amplitude of a signal over time. However, a spectrogram reveals variations in frequencies in a signal over time. The waveform displays amplitude over time, but at individual frequencies, we can't really see what's happening. For the length of the file, we can see that the waveform is constant standard, but we can't say anything about how the pitch or frequency varies over time. In the spectrogram view, the vertical axis represents frequency in Hertz, the horizontal axis represents time (exactly like the waveform display), and amplitude is represented by brightness. Spectrograms contain detailed information for audio data relative to waveform representation. We can see their representation by taking one audio along their form of representation.

Figure 2.9 Waveform representation

Figure 2.10 Spectrogram representation

A comprehensive audio view, capable of reflecting time, frequency and amplitude all on one graph, is a spectrogram. It is also defined as a visual way to reflect the signal intensity, or loudness, of a signal over time at different frequencies in a specific waveform, and it can show us whether over time there is more or less energy. Spectrograms, with a third dimension expressed by colors, are essentially two-dimensional graphs.

Figure 2.11 Training audio data

2.12. Feature Extraction

In order to carry out recognition/classification, the neural network must carry out feature extraction. Features are the elements of the data that you care about which will be fed through the network. In the specific case of image recognition, the features are the groups of pixels, like edges and points, of an object that the network will analyze for patterns. Feature extraction is the mechanism of taking the required useful attribute of the audio file with a machine learning algorithm that allows one to be differentiated from another. Feature recognition or feature extraction is also defined as the process of pulling the relevant features out from an input image so that these features can be analyzed. Many images contain annotations or metadata about the image that helps the network find the relevant features. Generally, there are two key steps in the genre classification process of music: extraction and classification of features. The first step obtains details about the audio signal, while the second step classifies the music according to extracted features in different genres (Nasridinov1 and Park, 2014). So, feature extraction is the precondition for classification because the classifier model will identify each unique class depending on what types of features or characteristics the items will have.

2.12.1. Audio file Feature Extraction

Feature can be described as an attribute value that tells us the detailed information about the entity. For one entity there may be a number of attributes each attribute enables to uniquely identify from another related entity. So audio files also have their own characteristics that become different from other files. Specifically, audio files also have several subsections like music, sound and the like. Extraction of audio features is the process of translating an audio signal into a sequence of feature vectors that carry signal characteristic information. These vectors are used as the basis for several types of algorithms for audio processing. For audio analysis algorithms, it is common to be based on features measured on a window basis. These window-based characteristics can be regarded as a brief summary of the signal for that particular moment in time (Karthikeyan and Mala, 2018). If zema said to be Saint Yard zema it must fulfill the criteria, from the criteria the first one is that it must be sing with zema notations, as we discussed before Saint Yared introduce eight types of zema notation and after the sequence of time several traditional school scholars add different notation that helps them to sing such sweet zema easily and clearly. Generally, Saint Yared zema characterizes different features like music and it became unique with some features. A wide range of audio features exist for classification tasks. These fall under the following categories: Time Domain Features, Pitch Based Features, Frequency Domain Features, Energy Features and MFCC (Karthikeyan and Mala, 2018). Additionally, zema notations will also be included.

Like music Saint Yared zema has elements or features that describe it in detail information and enable the listener or user to easily identify its categorical class. Saint Yared zema classification is categorized within three forms of zema classes Geez, Ezil, Araray.

In music genre classification the researchers can be considered three main step processes. These are the extraction of acoustic features from short frames of the audio signal, the aggregation of the features into more abstract segment-level features and the prediction of the music genre using a classification algorithm that uses the segment-level features as input (N. Silla and L. Koerich, 2009). Similarly, we follow the same method to conduct our study. The audio must be translated into a spectrogram then after segmentation will be conducted. On each segment the feature will be extracted with the above step like that on music classification.

2.12.2 Feature extraction from the spectrogram

As we have stated before about the feature extraction process for classification of audio files in machine learning in particular we have two main possibilities. We have seen one way of simply taking the feature of audio without converting it into image form and the other mechanism is directly taking the audio file and transforming it into spectrogram image then apply image processing technique, so we will focus on this approach. Spectrogram is one technique of frequency domain representation that is used in audio file classification with audio feature extraction method. After conversion into spectrogram features will be automatically extracted from spectrogram since we used convolutional neural networks, it is more powerful for feature extraction because it has layers that can filter each image.

2.13. Techniques and approaches for classification

Basically, machine learning algorithms are used for classification and recognition of normal images as well as the spectrogram with two possible techniques or approaches. These are shallow learning approaches and deep learning approaches. The result of the study will depend on their accuracy rate as well as performance even if different parameters are there to evaluate the model. We will discuss each type of technique as follows.

2.13.1. K-nearest neighbors (K-NN)

It is a sort of supervised machine learning technique that can be used to classification and regression problems. However, predictive problems in industry are primarily used for classification. KNN can be well described by the following two properties.

Lazy learning algorithm - KNN is a lazy learning algorithm since it doesn't have a particular stage of training and instead uses all of the data collected during classification to train.

Non- Non-parametric algorithm for learning - KNN is also a non-parametric algorithm for learning since it assumes nothing about the underlying data.

The K-nearest neighbors (KNN) algorithm predicts the values of new data points using similarity characteristics, which means that a value is assigned to the new data point depending on how closely it resembles the points in the training set

2.13.2. Naïve Bayes (NB)

Consider the classification problem where a sample x belongs to one of two classes, denoted as C1 and C2. Assume the prior probabilities P (C1), and P (C2) are known. The density function, P(Ci|x), is obtained by:

P(Ci|x) = (|) ( ) / ( ) (2)

According to Bayes theorem, the probability of the classification error can be minimized by the following rule:

x is classified to C1, if P(C1|x) > P(C2|x)

x is classified to C2, if P(C2|x) > P(C1|x)

Naïve Bayes assumes that the attribute values are conditionally independent to one another. It ignores the possible dependencies among the inputs. It has a series of steps used for the classification of the information provided.

2.13.3. Support vector machine (SVM)

A group of similar supervised learning techniques used for classification and regression are support vector machines (SVMs). It belongs to a family of linear classifiers that are generalized. In other words, Support Vector Machine (SVM) is a prediction method for classification and regression that uses machine learning theory to optimize predictive accuracy while avoiding overfitting the data automatically. Support Vector machines are high-dimensional feature space models that use linear function hypothesis space and are trained with an optimization theory learning method that incorporates a learning bias derived from statistical learning theory. The following are significant ideas in the SVM.

Support Vectors ƒ Data points that are nearest to the hyperplane are called support vectors. With the aid of these data points, the separation line will be established.

Hyperplane:- it is a plane of choice or space that is divided between various classes of a group of objects.

Margin … The margin can be described as the distance between two lines of different classes on the cabinet data points. It can be measured as the distance from the line to the support vectors that is perpendicular. A broad margin is regarded as a good margin, and a small margin is regarded as a poor margin.

SVM's main purpose is to partition datasets into classes in order to find a maximum marginal hyperplane (MMH), which may be accomplished in two steps. First, SVM will iteratively construct hyperplanes that best separate the classes. Then it will choose the hyperplane that correctly divides the groups.

2.13.4 Artificial Neural Network (ANN)

Neural networks are parallel models for computation, and are actually an attempt to make the brain a computer model. The main goal is to build a model faster than conventional models to perform different computer tasks. A neural network consists of at least a layer of input and a layer of output. Some network architectures may include multiple hidden layers between the input and output layers. Each layer can have one or more nodes. Each neuron in the input layer is connected to every output neuron in the next layer.

Two operating phases, training and testing, are always encountered in neural networks. During the training phase, the neural network takes the training dataset as input and adjusts the connection weights to achieve the desired association or classification. During the testing phase, the neural networks are tested with the testing dataset (different from training dataset) to retrieve corresponding outputs based on the knowledge discovered from the training phase.

An input layer, one or more hidden layers, and a single output layer make up a neural network. Each layer might have a varied number of neurons and be fully connected to the layer above it. The behavior of neural networks is shaped by its network architecture. A network€s architecture can be defined in terms of:

'V Number of neurons

'V Number of layers

'V Types of connections between layers

For the input layer, the input is the raw vector input. The input to neurons of other layers is the output (activation) of the previous layer€s neurons. As data moves through the network in a feedforward fashion, it is influenced by the connection weights and the activation function type.

Input layer: shows how we get input data to our network. The number of neurons in an input layer is typically the same number as the input feature to the network. Input layers are followed by one or more hidden layers.

Hidden layer: There are one or more hidden layers in a feedforward neural network. The weight values on the connections between the layers are how neural networks encode the learned information extracted from the raw training data. Hidden layers are the key to allowing neural networks to model nonlinear functions.

Output layer: output (prediction or classification) of our model is answered from the output layer. The output layer gives us an output based on the input from the input layer. Depending on the setup of the neural network, the final output may be a real value output (regression) or a set of probabilities (classification). This is controlled by the type of activation function we use on the neurons in the output layer.

Connections between layers: In a fully connected feed-forward network, the connections between layers are the outgoing connections from all neurons in the previous layer to all of the neurons in the next layer. These weights are progressively changed as the algorithm finds the best solution with the backpropagation learning algorithm. The overall diagram that shows the above definition will be described as below.

Figure 2.12 Artificial neural network structure

2.13.5. Convolutional Neural Network (CNN)

Convolutional neural networks (ConvNets or CNNs) are one of the key groups for image recognition, image classification, in neural networks. Detections of objects, faces of identification, etc., are some of the places where CNNs are commonly used. The computational model of this neural network uses a variant of the multilayer perceptron.

It requires one or more convolutional layers that can be either fully linked or pooled. Such convolutional layers produce function maps that record an image area that is ultimately split into rectangles and sent out for nonlinear processing. The Convolutional Neural Networks are multilayer perceptron (MLP) regularized models. An input image is taken for image classifications with CNN, processed and categorized under those categories. Computers see an input image as a pixel array and this depends on the resolution of the image. You can see h x w x d (h = Height, w = Width, d = Dimension) based on the image resolution. An image of a 6 x 6 x 3 RGB matrix array (3 for RGB values) and an image of a 4 x 4 x 1 grayscale matrix array. Each input image can move through a series of convolution layers with filters (Kernels), pooling, completely connected layers (FC) and apply SoftMax to classify an object with probabilistic values between 0 and 1. Technically, deep learning models to train and evaluate.

The first layer of a neural network takes in all the pixels within an image. After all the data has been fed into the network, different filters are applied to the image, which forms representations of different parts of the image. This is feature extraction and it creates feature maps.

A convolutional layer is used to extract information from an image, and convolution is merely the formation of a representation of a portion of an image. By learning image characteristics using small squares of input data, Convolution maintains the relationship between pixels. It is a mathematical process that involves two inputs, such as an image matrix and a kernel or filter.

Figure 2.13 Image matrix multiplies kernel or filter matrix

From the above diagram image matrix(volume) (h x w x d), a filter (fh x fw x d) and output of volume dimension (h - fh +1) x (w - fw +1) x 1.

Let us take the above image size is 5 x 5 whose image pixels 0,1 and the size of filter is 3 x 3

*

Table 2.1 Image matrix multiplies kernel or filter matrix

Then the convolution of a 5 x 5 image matrix multiplied with a 3 x 3 filter matrix which is called Feature Map. Stride is the number of changes over the input matrix in pixels. If the stride is 1, we change the filters to 1 pixel at a time. For a non-linear operation, ReLU stands for Rectified Linear Unit. Its aim is to implement our ConvNets with non-linearity. Since, the real world data would want our ConvNets to learn would be non-negative linear values.

The output is †(x) = max (0, x) (3)

There are other nonlinear functions which can also be used instead of ReLU, such as tanh or sigmoid. Many data scientists use ReLU because it is better performance-wise than the other.two.

Figure 2.14 Basics of CNN architecture

2.13.5.1. CNN architectures

CNN architectures are formed by a stack of distinct layers that transform the input volume into an output volume through a differentiable function. A few distinct types of layers are commonly used. All the above elements of convolutional neural network such as convolution, pooling and padding are relatively direct. Some of the architectures are discussed here

ResNet

ResNet was introduced by considering it as a continuous deep network. It revolutionize the architectural hierarchy in CNN by incorporating the idea of residual learn in CNN and develop an efficient technique for deep network training (Khan et al., 2016). ResNet introduce 152 layers of deep CNN that won ILSVRC-2015 computation. The residual block of ResNet shown in fig below (taken from (Khan et al., 2016)) was 20 and 8 times deeper than AlexNet and VGG correspondingly. It shows the complexity of computations than other previous introduced networks. It gained 28% of improvement on image recognition.

AlexNet

The model was designed to address ILSVRC-2010 competition for the classification of object images into one of 1,000 different categories (Krizhevsky et al., 2012). The model has five convolutional layerin the feature extraction part and three fully connected layers in the classification part. In the feature extraction part, the first four layers are followed one on another sequentially. However, in the middle between the fourth and fifth layer there is pooling layer. The fifth layer is then followed by the three fully connected layer and finally there is SoftMax to classify the incoming image to their respective class. The neural network has 60 million parameters and 650,000 neurons. After each convolutional and fully connected layer AlexNet uses ReLU as the nonlinearly.

VGG

The model was designed to classify over 14 million images in to 1000 classes in 2014. It achieves 92.7% accuracy and is one of the famous model submitted to ILSVRC. It improves against AlexNet by replacing large kernel-sized filters. This improves on AlexNet by replacing large kernel-sized filters with multiple 33 kernel-sized filters one after the other The size of the input image to the convolutional layer is 224X224 RGB image. After the image is passed through a stack of convolutional layers where the filters used with a very small receptive field 3×3. Which is the smallest size to capture the notion of right or left, up or down, center. Three Fully-Connected layers follow a stack of convolutional layers. The first two has 4096 channels each, the third performs 1000-way ILSVRC classification and thus contains 1000 channels The last layer is the soft-max layer that produces a distribution over the 1000 class labels. VGG also uses ReLU as the nonlinearly (Khan et al., 2016)

2.14. Evaluation metrics

There are different performance metrics that have been used to evaluate the performance of the proposed solution or model. Among these, accuracy, precision, recall and f1-score are used extensively for measuring the performance of proposed solutions.

Accuracy: is the proportion of true positives (include both true positives and true negatives) against the whole population. Accuracy may mislead the quality of the model if the class is not balanced

Accuracy = (TP + TN) / (P + N) (4)

Precision: is the proportion of true positives against the whole positives. Mathematically, it is expressed as:

Precision = TP / P (5)

Recall or sensitivity: is the proportion of true positives against the whole true or correct data. It quantifies how well the model avoids false negatives. It is also known as true positive rate or hit rate.

Recall = TP / (TP + FN) (6)

F1-score: is the weighted average, the precision and recall. The relative contribution of precision and recall to the F1-score are equal.

F1-score = 2 * (precision * recall) / (precision + recall) (7)

Micro-average, macro-average, and weighted-average for all the aforementioned performance metrics can also be calculated and used for additional analysis of results.

Macro-average precision or recall is just the average of the precision and recall (respectively) of the model on different classes.

Macro-average precision = (P1 + P2 + ‡ + PN) / N (8)

Macro-average recall = (R1 + R2 + ‡ + RN) / N (9)

Micro-average precision or recall is calculated by summing up the individual true positives, false positives and false negatives for each class.

Micro-average precision = (TP1 + TP2 + ‡ + TPN) / (TP1 + TP2 + ‡ + TPN) +(FP1 + FP2 + ‡ + FPN)
(10)

Micro-average recall = (TP1 + TP2¨ï+TPN) /(TP1 + TP2+¨ï +TPN) +(TN1 + TN2 + ‡ + TNN) (11)

Figure 2.15 Model evaluation metrics for the given data

2.15. Related work

Saint Yared was the most famous composer of zema not only in the Ethiopian orthodox tewahdo church, but also the became base for traditional music before anyone via the world. All traditional as well as modern popular musicians come next to him, even if he does not teach music but replaces his spiritual task in music.

We haven€t seen research papers which were conducted on Saint Yared Kum zema classification. Thus, in order to conduct the research, we have used related works from audio music classification, sound classification and recognition, audio emotions classification and other research that are nearly related with some approaches. Even musical classification doesn€t have a similar attribute with our study because in music classification the classification is highly dependent on the instrument that the musician used, but here our study will only focus on vocal song since it is classification of Saint Yared Kum zema.

2.15.1. Zema Classification Methods

In this section we are going to review different research works which have related approaches with our study especially concerned on music genre classification, speech classification, sound recognition and so on with two techniques. The first one reviews related works done using a traditional machine learning approach and the second one reviews related works done using a modern approach or we call it deep learning approach to solve the given problem.

2.15.1.1. Classification Using Shallow Machine Learning

The research (Karthikeyan and Mala, 2018) try to classify the audio file based on the feature that the audio files have like time domain features, pitch domain features, frequency domain features, energy domain features and using Mel frequency Cepstral coefficient by applying the artificial neural network specifically multi layered feed forward neural network with back propagation learning algorithm with the total accuracy of 80%.

According (Costa et al., 2011) proposed music recognition using spectrogram by taking the audio data then converting the audio signal into spectrogram from the spectrogram extract local features which is used for classification with ten groups classes. The researchers used around 900 audio data and transformed it into spectrogram and finally performed the recognition using SVM and GLCM algorithm by extracting the texture descriptors as features. The classifier achieves the accuracy of recognition rate of 67.2%

According to (N. Silla and L. Koerich, 2009) Genetic Algorithms (GA)-based feature selection process for multiple feature vectors extracted from different sections of the music signal and analysis of the discriminatory power of the features according to the part of the music signal from which they were extracted and the effect of the selection of features on the classification of the music genre. The classifier was developed by different machine learning algorithms like Naive-Bayes, Decision Trees, Support Vector Machines and Multi-Layer Perceptron Neural Networks. Basically the audio file will be changed into spectrogram then after it will be segmented with some time interval and our final goal will be identifying in that class the segmented audio file will group based on the feature that audio file has.

2.15.1.2. Classification Using Deep Learning

According to (Jawaherlalnehru and Jothilakshmi, 2018) the researchers were trying to conduct the study on music instrument recognition with spectrogram image. It was the frequency domain feature extraction technique and enable them to obtain optimal accuracy by using input data audio files and applying CNN algorithm with better accuracy that is 97%, but this study only focused on the music instrumental recognition; it doesn€t include the vocal of the musician.

In our environment there are different sounds that are emitted from different objects as well as from human beings into the surrounding, so researchers are motivated to identify and clas this emitted sound into the environment with different categories. In (KHAMPARIA et al., 2019) The researchers carried out a study in which deep learning networks were utilized to classify environmental sounds based on their generated spectrograms. Their initial input is an audio file and changes it in the frequency domain feature that is spectrogram images of environmental sounds. They apply machine learning algorithms to train the data are convolutional neural networks (CNN) and the tensor deep stacking network (TDSN). The accuracy measured in this study with the above two algorithms were 77% and 49% in CNN and 56% in TDSN.

Researchers in (Bilal Er* and Aydilek, 2019) also stated many academics conduct their research with extracting acoustic aspects from music and investigating relationships between emotional tags corresponding to these features, Music Emotion Recognition Using Chroma Spectrogram and Deep Visual Features Recent research has used deep learning to analyze music spectrograms that include information from both the temporal and frequency domains. Recently, by using a pre-trained deep learning model with Chroma spectrograms derived from music recordings, a new approach for music emotion recognition has been introduced. The AlexNet architecture is used as the pre-trained network model. As the feature extraction layer, the AlexNet model's conv5, Fc6, Fc7 and Fc8 layers are picked, and deep visual features are extracted from these layers. For training and testing the Support Vector Machines (SVM) and the SoftMax classifiers, the extracted deep features are used in addition. Deep visual features are taken from the conv5-3, Fc6, Fc7, and Fc8 layers of the VGG-16 deep network model, and the same experimental applications are used to determine the success of pre-trained deep networks in the recognition of music emotions. The best result is obtained on their own dataset as 89.2% from the VGG-16 in the Fc7 layer. Several researchers are conducted research in MIR specially their dataset was audio then they apply the technique of converting the audio data into the spectrogram image on this image they will apply different preprocessing, segmentation and feature extraction methods like that of image processing method, so according to (Badshah et al., 2019) concerned on speech emotion recognition using spectrograms and deep convolutional neural network (CNN). Input to the deep CNN is spectrograms created from the speech signals. The proposed model consists of three layers of convolution and three entirely related layers.

Layers derive discriminative features for the seven emotions from spectrogram images and performance predictions. As we have seen in the above they used Deep CNN algorithm and additionally used AlexNet model for improving the accuracy as well as the performance of the classification and the recognition model and its overall accuracy was 84.3%.

According to the research titled text to hymn synthesis for Saint Yared hymn notation, that uses the NLP definition as a synthesis of text in zema but does not concentrate on the genres of zema and classification with regard to hymn notation (Girma, 2014).

Using features retrieved automatically from audio, the audio analysis process becomes easier and more accurate. Low-level audio characteristics are commonly used in audio categorization studies. Clustering studies have also used low-level audio features. For a better recommendation model, Li et al. investigated clustering based on timbral texture features and rhythmic content features derived automatically from audio. In clustering and classification, there are eleven sets of time domain and frequency domain characteristics: spectral centroid, spectral entropy, spectral flux, spectral rolloff, cepstral coefficient of Mel-Frequency, Harmonic, Chroma Vector, and spectral zone are all terms used to describe energy, entropy energy, zero crossing rate, spectral centroid, spectral entropy, spectral flux, spectral rolloff, cepstral coefficient of Mel-Frequency, and spectral zone (Jonya and Iswanto, 2017). Here the study performed on clustering using machine learning and it doesn't consider labeling of data as well as performed using unsupervised machine learning. A music genre description that converts audio signals into spectrograms and derives attributes from this visual representation. The concept is that by treating the time-frequency representation as a texture image, we can extract features to develop accurate music genre categorization algorithms (Costa et al., 2011). Similarly, to classify the given input audio zema it must be transformed into spectrogram form and then the feature is extracted from the spectrogram image.

The researcher also explains the task conducted in their study and explains the proposed method for automatic music genres classification, that consists of three steps: marking, matching genres and classification (Nasridinov1 and Park, 2014). There are several techniques that are used to extract or select features of the audio file. Audio data is a part of many new, multimedia and computer applications.

The need to identify automatically which class an audio sound belongs to makes audio classification and categorization a new and significant area of research (Karthikeyan and Mala, 2018). We will use different audio files with different file extensions like mp3, amr, wav. The audio signal will be processed and converted into a spectrogram image; this image can also be described as short time Fourier representation and also named as texture image. The features extracted from it are local features since the texture of the spectrogram is not uniform (Costa et al., 2011).

From the above researches were conducted in different areas as well as discipline with different algorithms, techniques and methodology to achieve better performance and accuracy on their study general it will be described as follows.

Table 2.2 Related works

No	Research title	Algorithm Used	Authors	Limitation
1.	Content basedaudioclassifier &feature extractionusingANN tecniques	Multi layered feed forward neural network with back propagation learning algorithm	(Karthikeyan and Mala, 2018)	Doesn€t consider th visual represented feature
2.	Music Instrument Recognition from Spectrogram Images Using Convolution Neural Network	Convolutional neural network	(Jawaherlalnehru and Jothilakshmi, 2019)	Only classifythe music instrument doesn€t consider the vocal sound
3.	Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network	Convolutional neural network (CNN) and the tensor deep stacking network (TDSN)	(KHAMPARIA et al., 2019)	Doesn€t have intra similarity between class easily distinguishable
4.	Music Emotion Recognition by Using Chroma Spectrogram and Deep Visual Features	Convolutional neural network (CNN) and support vector machine	(Bilal Er* and Aydilek, 2019)	Used combined method but the accuracy result is not better
5.	Speech emotion recognition using spectrograms and deep convolutional neural network (CNN	Convolutional neural network (CNN)	(Badshah et al., 2019)	Doesn€t have intra similarity between class easily distinguishable
6.	Indonesian€s Traditional Musi Clustering Based on Audio Features,	X-mean algorithm	(Jonya and Iswanto, 2017)	Doesn€t havelabeled data because it simply grouped in to some predefined cluster
7.	Music Genre Recognition Using Spectrograms	Support Vector Machine (SVM)	(Costa et al., 2011)	Manual feature extraction
8.	A Study on Music Genre Recognition and Classification Techniques	Hidden Markov models, Neural networks, dynami Bayesian network and	(Nasridinov1 and Park, 2014)	Only focus on acoustic feature

		Rule-based methods, and template matching methods		
9.	Automatic Music Genres Classification using machine learning algorithm	K-nearest neighbor (k-NN) and Support Vector Machine (SVM)	(Asim and Siddiqui, 2017)	Focus on acoustic feature and manual feature extraction
10.	Feature Selection in Automatic Music Genre Classification	Genetic algorithm for feature extraction and Naive-Bayes, Decision Trees, Support Vector Machines and Multi-Layer Perceptron Neural Network for classifier	(N. Silla and L. Koerich, 2009)	Unable to extract features automatically

Generally, several researches were conducted on audio file classification like classification of emotion with music, environmental sound, musical instrument, music and so on. Some of the study was only focused on Acoustic features of the audio. Some other studies only focus on instruments which don't consider the vocal. So, our study will mainly focus on classification of kum zema with classes of Geez, Ezil and Araray. Each class has high intra similarity during generation of spectrograms. So, by solving this problem we will get better results.

2.16. Summary

Audio signal processing is one key mechanism which is used to represent the audio data in digital form by applying different algorithms. St. Yared is the founder of zema who provides around six compositions. These are Me€eraf, Tsome Digua, Digua, zimare, Mewasit and Kidasie zema which are sung with three types of zema Geez, Ezil and Araray forms. The standard and structure of zema is formulated by St Yared named as zema notation. Zema notations are Eight in numbers after Saint Yared different traditional scholars add several notations that originated from the initial one. Kum zema is said to have no need of instruments available during the singing. To classify zema different approaches are used for classification with acoustic features and visual features representation. We used visual representation of audio after transforming audio into spectrogram image. To generate spectrogram image different parameters are used. The input for the convolutional network model is spectrogram image and features are extracted from the images. The classification is performed using the SoftMax classifier into appropriate classes Araray, Ezil and Geez. Researchers conducted their studies on genre classification of music, instruments, and environmental sound by applying shallow learning approaches and deep learning approaches. Most of the studies are focused on classification by taking acoustic features.

Chapter Three: Methodology

3.1. Introduction

This chapter will discuss the research methodology and the proposed classifier model which is used to show the classification of Saint Yared zema genre and used to indicate the sequential steps to implement the model. The data will be used in two ways: the training data which is initially given for the model to learn the available features required for the classification and the second one is the test data which is used for testing our model by taking some sample of data from the training data or out of the training data. The classifier model will classify the data into three distinct classes based on the features learned from the training data. Simply, the machine classifies appropriately depending on what it was training. In the real environment the classifier model must perform the classification by taking the test data which may not have related with what the machine was learning. So, during that time there may be some difficulty to classify if it is performed with such mechanism it is well standard as well as highly accepted but most nearly all of the classification perform by using split the related data with two groups and more than the half percent of data is allocated for training and the machine learn very well it is not difficulty to identify the class of the remain testing data.

3.2. Research methods

Research methods are specific procedures or guidelines for the study to conduct with the sequence of activity. We used experimental research method because we used the result which obtained from the experiment of our study with different working environments.

3.3. Model Architecture

When we say model architecture it means that the prototype used to represent the model with designing and implementing to classify Sainted Yared zema with their appropriate features. It also implies that the overall design of the model leads us to implement the prototype into the real application on model. In this research we mainly focus on classification of Saint Yared kum zema as we said kum zema means the types of zema that doesn€t use any type of musical instruments simply only the vocal sound that zema expert song basically such types of zema have three classes, and we call Tsewate wezema 8Ë•pÈ¸Ü. The proposed model will have different components like audio file reading, converting the audio file into spectrogram, preprocessing, segmentation, feature extraction and finally classification. A spectrogram defines signal strength of visual information at various frequencies available in input waveform. The spectrogram represents two-dimensional graphs contains horizontal and vertical axis for frequency and amplitude. These are basic components specified using by color in a particular time in the spectrogram. Low amplitudes indicated by dark blue. Strong amplitude indicated by red color(Jawaherlalnehru and Jothilakshmi, 2019). The feature extraction will be performed by the Convolutional neural network (CNN) since it has well-defined layers that be used to filtrate by applying different activation functions and the classification will be held with SoftMax classifier. The overall activity will be shows as follows. The proposed stride CNN architecture has input layers, convolutional layers, and fully connected layers followed by a SoftMax classifiers.

Figure 3.1 Proposed model architecture for Saint Yared kum zema classification

The SYKZC model is designed to classify Saint Yared zema genres into three proper classes. It includes different activities starting from input audio up to classification. The main sequence of the developed model consists of input audio data recorded from experts with Wave form, preprocessing input audio, transformation of audio, resizing of spectrogram image, extraction of relevant features with convolutional neural network layers, classification of the input data using SoftMax classifier.

3.4. Audio zema acquisition

Audio acquisition is the method of acquiring required audio zema from different resources and from the expert /scholar of traditional schools. Audio zema acquisition is the primary task of the study because without collecting audio data from different sources and from expert it is impossible to conduct the study. In order to identify and classify whether the given kum zema is grouped as Geez, Ezil and Araray first we try to collect the audio zema with two forms. The one way is recording the audio zema from zema Gubaebet and the second way is taking the annotated audio file.

3.5. Preprocessing

In order to achieve model accuracy and performance, pre-processing is an important part of preparing data. In this point, we need to clean the audio signals using adaptive threshold-based preprocessing to remove the background noises, silent portion and irrelevant song signal detail, and it also focuses on audio file segmentation with the same amount of time that allows us to properly and correctly convert spectrogram images. So preprocessing of audio data contains noise removal and segmentation of audio files.

3.5.1. Noise removal techniques

Sound is produced by vibrating objects and enters the listener's ears as waves in the air or other media. As an object vibrates, it causes minor changes in air pressure. Changes in air pressure travel through the air as waves, which produce sound when they move. There may be some interference. Noise interference is the term for sound interference. The mechanism which is used to reduce or remove this unwanted sound is named as noise removal or reduction. Noise reduction may simply be defined as the process of eliminating noise from a signal. For audio and pictures, noise reduction techniques exist. Algorithms for noise reduction aim to change signals to a greater or lesser degree. Both signal processing devices have characteristics that make them sensitive to noise, both analog and digital. So, algorithms are needed for the sack of removing these unwanted interferences of sound in the normal sound. Different methods will be applied for reduction and elimination of noise from that the common method for the removal of noise is optimal linear filtering method, and some algorithms in this method are Wiener filtering, Kalman filtering and spectral subtraction technique. Here a filter or transformation is passed through the noise signal(H.E.V et al., 2007). We applied the spectral gating techniques to reduce the background noise which may occur in our data.

We find the energy-amplitude relationship in waves in the next step and then measure the maximum amplitude in each frame and transfer from an acceptable threshold to eliminate the noise and salient portion and save it in an array. In the last step, we reconstruct a new audio with the same sample rates without any noise and silent signals. Additionally, we will apply the following list of steps. First, we need to come up with a method to represent audio clips (.wav files). The audio data should then be preprocessed in order to use the machine learning algorithms as inputs. Some useful functionalities for processing python audio are supported by the Librosa library. Using Librosa, the audio files are loaded into a numerical array. At a rate called Sampling rate, the list would consist of the amplitudes of the respective audio clip. (The sampling rate will normally be 22050 or 44100).

Figure 3.2 Audio file preprocessing and noise removal

3.5.2. Segmentation of audio

Audio segmentation is a method that separates the composite sounds of an audio file. A single sound that is acoustically distinct from other parts of the audio should consist of each section. The term refers to the problem of splitting an audio stream into homogeneous segments and classifying each segment as speech or music. The techniques used to segment the given long recorded audio into homogeneous segments using a thresholding method which means to assign fixed value of time interval based on the assigned time chunk audio. The time assigned to make the audio file to be segment 10 seconds used. Here is some sample pseudocode which shows how the longest audio files are segmented into several segments.

Table 3.1 Pseudocode for segmenting audio

```
Input: long size audio data
Output: segmented audio file
Begin:
    Read the long sized audio data from the folder
    Assign the size to be the audio segmented // equal 10 sec
    Cut audio equal second
    Return the segmented audio
    End
```

3.6. Transformation of Audio zema

In order to classify audio files, we will have different techniques, basically the two main mechanisms are mostly used. This mechanism is to convert the audio data in the signal and from the signal image the feature will be extracted whereas the other way will be changing the audio data in the spectrogram image and from the image extracted the required features that enables us to classify each zema with their proper class. Basically the audio file may be represented within image forms amplitude with respect to time and what we call waveform representation and frequency with time is called spectrogram representation, but our study will concentrate on spectrogram representation of the audio file. The spectrogram of the audio will have 2D representation of frequency with respect to time that has more information than text transcription words for recognizing the categorical class of song.

The fundamental idea is to train high-level discriminative features from audio signals using a CNN architecture, and the spectrogram is well suited for this task. Spectrogram and MFCC characteristics are used together using a CNN for identification and classification of speech emotions, according to the researchers, but the spectrogram characteristics are used to achieve good output in speech emotion recognition (Badshah et al., 2019). We must convert the one-dimensional representation of the speech signal into an acceptable 2D representation for 2D CNN because the major goal of this research is to learn high-level features from speech signals using the CNN model. Spectrogram is the best and suitable representation of audio speech signal in two dimensions that represent the strength of speech signals over frequency. For visual representation of frequencies over various periods, the short-term Fourier transformation (STFT) is applied to the speech signal. STFT is used to convert a longer time speech signal to a shorter section or frame of equivalent duration and then to measure the Fourier spectrum of that frame by applying rapid Fourier transformation FFT on the frame. The representation shown as follows. The first representation shows the waveform for the given input audio data. Here is some pseudocode that shows how each audio data is converted into spectrogram images.

Table 3.2 Pseudocode for transform audio to spectrogram

```
    Input: segmented audio data
```

```
Output: spectrogram image
Begin:
        Read the segmented audio data from the folder
        Assign the maximum value of frequency and time
        adjust the size window for the image
        Assign the proper type of window
        Number of Mel if Mel spectrogram
        Return the spectrogram image
End
```

3.7. Feature extraction

A very important part of evaluating and finding associations between different objects is the extraction of features. The audio data generated cannot be explicitly understood by the models in order to translate them into a comprehensible extraction of format features. It is a process that describes much of the details, but in a comprehensible manner. For classification, prediction and recommendation algorithms, feature extraction is required. To extract features different algorithms are used but for our study we will apply Gabor filters since these techniques are better for extracting the required features from the spectrogram. The study in (Kandre M. G. et al., January 2017) stated that the techniques used to extract the feature of music in music classification are Local Binary Patterns, Local Phase Quantization, and Gabor filters which leads the result to have better accuracy but for spectral images we used CNN as feature extractor as well as classifier

3.8. Classification

As we have seen, classification is the method of grouping similar data with one class and another data into another group depending on the feature extracted from the data using machine learning algorithms. Itis performed after the feature of each data is extracted. So, the algorithm learns features by passing different layers that enable us to know several features exist in the visual representation of an audio file. Finally, based on the feature of visual represented audio file or what we call it spectrogram. Saint Yared kum zema have three classes. These are Geez, Ezil and Araray.

3.8.1. Training phase

3.8.1.1. Feature extraction and learning phase

In this phase several activities are performed which enable the algorithm to classify the data properly. In this phase feature of the audio file extracted after audio data is represented with visual form. Basically there are two methods of representation as we said before, the waveform representation and the spectrogram method of representation. The second way of representation is better in several ways. So, the audio file used for generating the spectrogram then this transformed image is directly fed to CNN to learn features and with different layers the spectrogram will be filtered out to be classified with its appropriate class. In CNN layers several activities are performed so we will see it deeply.

We can describe the sequence training the data from the input image up to the last layers of convolutional neural networks this way.

Figure 33. Sequence input and activation function usage

Convolution layer: there are different convolution layers in the training phase. The input to the first convolution layer is 128 x 128 x 3 image. Here just neural networks that use Convolutional layers, also known as Conv layers, which are based on the mathematical operation of convolution. As we have mentioned in the above, these CNN layers take the input size of 128 size of width and 128 size of height and the next 3 indicate the filter size then total the input images have around 49152 input features which is fed for the Conv layers. Even if the size of the spectrogram image is determined by the algorithm that generates the spectrogram image from the audio file, most of the researchers take this size for their research. In our model, we have used 32, 64, 96, and 128 filters. The number of filters we have applied increased as we went down to the fully connected layers and the Softmax classifier. We have also used 3 x 3, and 1 x 1 filter size at a single layer and to determines the number of pixels skipped (horizontally and vertically) each time we make convolution operation this we have used stride size of two (2, 2) and one (1, 1) since the size of stride is determine the size of image if the size is two it reduce the size of image vertically as well as horizontally with half.

Activation layer: here the activation function is used even if there are different types of functions in our study we used ReLU activation function for generating output. Some of the activation functions are: Sigmoid, Hyper tangent, ReLU and Softmax. Nowadays ReLU is the most used activation function and Softmax is normally used in the last layer to obtain the output vector as a probability vector. The output of the activation function is always the same as the size (dimension) of the input. Hence, the width, height, and depth of the output layer is the same as the width, height, and depth of the input layer respectively. We have used ReLU activation functions in the activation layer throughout our model.

The ReLU activation function returns zero, if the value in the input layer is negative, otherwise it returns the existing value. Mathematically, it is defined as:

$$y = \max (0, x) \tag{11}$$

Pooling layers: this one is another layer of convolutional neural network which is used to change the volume of the input image by taking the minimum value, average value or maximum value of the image with the given number of kernel size. It is normal to insert a pooling layer periodically in a ConvNet architecture between successive Conv layers. Its function is to gradually lower the spatial size of the representation and thus check the overfitting in order to lower the quantity of parameters and computation in the network. On each depth slice of the input, the Pooling Layer works independently and resizes it spatially, using the MAX operation. It is used to reduce the volume of the input which means the height and width of the input.

Figure 34. how the spectrogram image is downsizing in CNN layers and Max pooling

Pooling layer down samples the volume spatially, independently in each depth slice of the input volume. Left: In this example, the input volume of size [126x126x3] is pooled with filter size 2, stride 2 into output volume of size [63x63x3]. Note that the depth of the volume is retained. Right: Max is the most common down sampling process, giving rise to max pooling, with a phase of 2 shown here. That is, 4 numbers are taken over each max (little square 2x2)

Fully connected layers: Neurons have total links to all activations in the previous layer, as seen in normal Neural Networks, in the last layers of convolutional neural networks. It holds three nodes that are directly added to the Softmax classifier (equal to the number of classes). The key thing about a fully linked layer is to take the convolution/pooling process results and use them to classify the picture into a name. The convolution/pooling output is flattened into a single value vector, each representing the probability that a certain characteristic belongs to a name.

Dropout layer: The dropout concept refers to the falling out of a neural network of units (neurons). The neurons are discarded during the training phase randomly with a certain probability; this is the parameter that we can change. This technique is used to avoid over-fitting, pushing the neural network to learn more stable characteristics that are useful in combination with various random subsets of other neurons (Boixeda, June 2019). Here when we see the range of dropout rate is between 0 and 1, which means if there is a dropout rate 0<X<1. There was a value between 0 and 1. where X is the value. The default interpretation of the dropout hyper parameter is that a given node is likely to be trained in a layer, where 1.0 means no dropout and 0.0 means no layer output. A strong dropout value is between 0.5 and 0.8 in a concealed sheet. A greater dropout rate, such as 0.8, is used by the input layers. The set of instructions which is used for performing the training phase of our model called SYKZC models is provided as follows.

Table 3.3 Pseudocode for general classification of SYKZC model

```
Input: preprocessed spectrogram image S
Output: extracted feature vector
Begin:
    Get preprocessed spectrogram image S
    Initialize the number of filters K, filter size F, stride size S, and zero-padding ZP, pool size PS, the number of nodes N, the number of classes C, and dropout probability P;
    Apply convolution operation, Convolution (K, F, ZP, S);
    Apply activation function, ReLU on the output of the previ convolution operation;
    Apply convolution operation, Convolution (K, F, ZP, S);
    End For
    // the first block of pooling module
    Apply max pooling operation, MaxPool (PS, S)
    Concatenate filter size
    // Similarly apply other pooling modules
    // the first block of convolution operation
    Apply 1 x 1 convolution operation, Convolution (K, (1,1), ZP, S);
    Apply 3 x 3 convolution operation, Convolution (K, (3,3), ZP, S);
    Concatenate filter size
    // similarly apply other convolution modules
    Apply dropout operation, Dropout (P) // drop around half of nodes, if P = 0.4
    Apply fully connected layer, FC (C); // takes only the number of classes
    which will be directly applied to the SoftMax classifier.
    Save (or return) extracted features
End
```

3.8.1.2. SoftMax

Here the values generated from the previous convolutional layers are given for full connected layers and FC gives the generated output for SoftMax classifier to classify into classes that are defined previously. These are Geez, Ezil and Araray zema. This classification method is described with a set of instructions as follows.

Table 3.4 Pseudocode for classification

```
Input: extracted or learned features
Output: class label
Begin:
    Get the extracted or learned features (from the above)
    Apply the SoftMax classifier on the learned features
    Return the class label
End
```

In order to increase the performance and accuracy of the model to classify the given data based on its basic features and to decrease the loss which may happen in our model we used different operational layers and additionally we used the following techniques that enable the model to perform well. From these:

Batch normalization: It's a method of standardizing the inputs to a layer while training very deep neural networks for each mini-batch. This has the effect of stabilizing the learning process and significantly reducing the number of training cycles needed for deep network training.

Batch size:-Since the databases are so big, the databases are split in batches. The number of the training examples present in this split is the batch size. This batch represents the input in a single iteration to the neural network. The forward and backward optimization of each batch against the labels of the actual prediction.

Epochs:- An epoch is when one time a whole dataset is moved through the neural network forward and backward. To train the model, the number of epochs should be greater than one, and as the number of epochs increases, the weight in the network changes more frequently, and the curve shifts from underfitting to optimal or even overfitting.

Optimizer:-Optimizer is an optimization algorithm that helps us to minimize the loss function towards changing and adapting the values of the weights and bias of network. There are many different types such as Stochastic Gradient Descent, Adam, Adamax and RMSprop. Most researchers prefer the Adam optimizer

Loss:- The Loss function is the most important unit to estimate the error from the prediction to the original value. To fit the estimated and expected values perfectly the training phase aims to have a loss of zero. To obtain it the weights of the neurons have to be adjusted using an optimization function until better predictions.Testing phase

Here we apply simply the prior methods which are used for the training phase in feature extraction and learning as well as for SoftMax classification of the given tested data.

3.8.2. Testing phase

Here we apply simply the prior methods which are used for the training phase in feature extraction and learning as well as for SoftMax classification of the given tested data. Initially the recorded audio data is prepared and preprocessed then in a similar way to what we applied in the training phase. If there is background noise reduce it by applying spectral gating technique and apply audio segmentation techniques to obtain homogeneous segment of audio with time interval and data size. The transformation of segmented audio into spectrogram image is performed and then the input for convolutional neural network becomes generated spectrogram image after resizing the image into the required data size. Basically we used downsizing technique because when the size of the spectrogram image increases the brightness of color becomes reduced. The CNN layers filter it with different filter size and strides finally the SoftMax classified the input audio data into appropriate classes Araray. Ezil and Geez

3.9. Summary

The SYKZC model is designed to classify St. Yared kum zema types into three main types. Initially the model takes input data audio data and this data needs further preprocessing, this includes noise reduction, segmentation to have uniform segment of audio with time interval and size. Each segmented form of audio is transformed into spectrogram images. The transformed spectrogram becomes an input for convolutional neural networks. From the input image features are extracted with the first layers of the CNN called convolution and the dimension reduction of image is performed with the second layers named as pooling and finally the pixels of the spec imagetransformed into vector form with full 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.

Chapter Four: Result and Discussion

4.1. Introduction

This chapter is focused on evaluation of the SYKZCNet model with respect to the design model that was proposed in the previous chapter three as well as the structure of the dataset which is used for conducting our study. Applying different techniques which leads our model to have better accuracy and performance and finally perform comparison of our model with other models using prepared data set and observe the result to know which algorithm has better performance as well as better accuracy.

4.2. Dataset

The main aim of this study is classification of Saint Yared kum zema. Data is needed for the research because without data it is impossible to perform anything. There is no prepared data which existed before due to the reason researchers are not conducting study in this area. So, to conduct this research we collected different types of Saint Yared kum zema from Ethiopian orthodox traditional schools specifically from zema bet Ü (d)u scholars with recording and take some sample of data from internet which are annotated by the scholars. Data, in recorded form is collected from three traditional schools F(u - u d) u scholars. The audio data were collected with different forms and transformed each audio file into the same audio file extension which is wave form. The collected data consist of Wudasie Maryam, ankeste birhan, Mestegabe, Selamta, Tsome Digua and Digua. The data needed rearrangement as well as preprocessing which means to represent the audio file with visual form or spectrogram as well as to take features in acoustic form the long recorded file must be segmented with equal size to have uniform time interval. Finally, the segmented Audio files changed into a visual representation form which is a spectrogram. The data fed for the convolutional network is the spectrogram image with the form of jpg or other image format. The total numbers of transformed visual representations form audio for each corresponding class Araray were 595, Ezil were 539 and Geez were 421 in number. Their summation was around 1555.

Table 4.1 Data set used for the study

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

4.3. Implementation Tools

The model developed using different implementation tools, the following tools (programming language, libraries and framework) were used:

Python 3: The programming language that is used to implement the models is python. We decided to use python because of the richness of libraries in data manipulation and frameworks in the deep learning and data processing area.

Keras 2.2.4: It is a deep learning framework or a library providing high-level building blocks for developing deep learning models.

Scikit-learn 0.23.2: It is a machine learning library with various features and tools.

Jupyter Notebook: Jupyter notebooks are a great way to run deep-learning experiments. It allows you to break up a long experiment into smaller pieces that can be executed independently which makes the development interactive. All the experiments in this research were run in Jupyter. NumPy 1.19.2: It is a multi-dimensional array (tensor) manipulation library. When doing deep learning every data must be represented in a tensor of different size and for storing and manipulating the arrays NumPy was used.

Librosa 0.8.0: It's a music and audio analysis package that gives you the tools you need to build musicinformation retrieval models. This library is used to extract features from audio.

PyDub ƒ It is a library to manipulate audio data with a simple high-level interface.

OpenCV ƒ It is a library to solve computer vision problems. We use the library to read video data from disk and dismantle it to the image pieces that constitute the video.

Matplotlib 3.3.2 ƒ It is a python 2D plotting library.

In addition to the above software package and library we used an intel ‰ core-i3 5005U CPU and RAM 4GB. 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 percent of the data is assigned for training the model and 30 percent of the data is allotted for testing.

4.4. Results

As we have said before, to measure the performance and accuracy of the model we used different metrics. Like F1-score, precision, recall and accuracy. There are also additional measuring techniques like macro-average, and weighted average.

4.4.1. SYKZC model in Training phase

We used different python libraries and programming languages to implement the proposed classifier model and working environments are needed to execute the source code. These environments include Anaconda with tensor flow and google colab. The model is executed with anaconda, needs more than two hours and we tried with another environment google colab. The second environment is connection oriented, uses GPU as processor and generate better in time usage. Relative to anaconda the google colab has better processing speed and we used it. When we come to the experiment to evaluate our model, the SYKZC model obtains 98% training accuracy and 88% testing accuracy. The overall average loss rate for the model is 0.1. 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.

Table 42 Classification Accuracy of training phase of SYKZCNet

Epoch	Time taken	Number loss	Accuracy	Val_ loss	Val_ accuracy
1/100	4sec 49ms/step	2.8684	0.5190	1.3675	0.3498
2/100	1sec 20ms/step	2.2378	0.6572	2.3252	0.3498
3/100	1sec18ms/step	1.5812	0.6772	1.6999	0.3498
4/100	1sec 18ms/step	1.3916	0.6654	1.5313	0.2811
5/100	1sec 18ms/step	1.0123	0.7224	3.4827	0.2790
. .	. .	. .	. .	. .	. .
95/100	1sec 20ms/step	0.0791	0.9742	1.2046	0.8734
96/100	1sec18ms/step	0.0540	0.9883	1.6185	0.7790
97/100	1sec 18ms/step	0.1978	0.9584	1.7029	0.8262

98/100	1sec 18ms/step	0.1112	0.9716	1.0422	0.8541
99/100	1sec 18ms/step	0.0765	0.9829	1.9113	0.8133
100/100	1sec 18ms/step	0.0597	0.9815	1.0141	0.8755

Class/metrics	Precision	Recall	F1-score	Support
Araray	0.82	0.95	0.88	173
Ezil	0.93	0.79	0.85	163
Geez	0.91	0.88	0.90	130
Accuracy			0.88	466
Macro avg	0.89	0.87	0.88	466
Weighted avg	0.88	0.88	0.88	466
Test result:87.554		Loss:1.014		

The diagram given in figure 4.1, 4.2 and 4.3 shows the accuracy and loss of the trained and tested phase of the proposed model with respect to the prepared dataset. When we have seen the trained phase it had uniform results whereas the tested phase had some up and down curve it doesn€t have uniform results even if it has better results. Generally, the classifier model had better accuracy results and low rate of losing rate relative to the related convolutional neural network models. The loss rate for this model is 1.014. The overall diagrammatic representation of the SYKZC model accuracy and loss for training and testing are shown below.

Figure 4.1 The training and testing accuracy and loss of SYKZC Model

Figure 4.2 Training accuracy curve of SYKZCModel

Figure 4.3 The training loss curve of SYKZCNet

4.4.2. Comparison of the proposed model with different activation function

An activation function is a function that is added into an artificial neural network to help the network learn complex patterns in the data. It takes the preceding cell's output signal and turns it into a format that may be used as input to the next cell. Basically three types of activation function are used. These are ReLU, Sigmoid and tanh. The proposed model has different results when the activation functions are interchanged. The above result is obtained using ReLU activation function and we will see the result obtained using sigmoid and tanh.

4.4.2.1. Comparison with Sigmoid activation function

Nonlinear activation functions are preferable because they enable nodes to learn more complicated data structures. The sigmoid and hyperbolic tangent activation functions are two often used nonlinear activation functions. The logistic function, often known as the sigmoid activation function, has long been a common activation function for neural networks. The function's input is converted to a value between 0.0 and 1.0. Inputs that are significantly bigger than 1.0 are changed to 1.0, and values that are significantly smaller than 0.0 are snapped to 0.0. The function's shape for all conceivable inputs is an S-shape ranging from zero to 0.5 to 1.0.

To execute the SYKZC model using Sigmoid activation function, it needs to be more than two hours and we tried with another environment google colab. Relative to anaconda the google colab has better processing speed and we used it. The experiment to evaluate our model, the SYKZC model with Sigmoid function, obtained 96% training accuracy and 84 % testing accuracy. The overall average loss rate for the model is 0.875. So, sigmoid activation function has better loss rate and the accuracy is less than ReLU with 4%.

Table 4.3 Classification Accuracy of training phase of SYKZCNet with sigmoid

Epoch	Time taken	Number loss	Accuracy	Val_ loss	Val_ accuracy
1/100	4sec 53ms/step	4.7598	0.4754	1.2973	0.3498
2/100	1sec 19ms/step	4.2667	0.6016	1.1858	0. 3498
3/100	1sec 19ms/step	2.1184	0.6031	1.1725	0. 3498
4/100	1sec 18ms/step	1.2727	0.6876	1.2406	0. 3498
5/100	1sec 19ms/step	1.4378	0.6581	1.3920	0. 3498
. .	. .	. .	. .	. .	. .
95/100	1sec 19ms/step	0.1098	0.9539	1.0108	0.8348
96/100	1sec 19ms/step	0.1103	0.9586	0.8575	0.7983
97/100	1sec 20ms/step	0.0995	0.9625	0.7573	0.8519
98/100	1sec 19ms/step	0.0925	0.9708	0.0505	0.8240
99/100	1sec 19ms/step	0.0868	0.9598	0.1996	0.8090
100/100	1sec 19ms/step	0.1265	0.9626	0.8753	0.8369

Class/metrics	Precision	Recall	F1-score	Support
Araray	0.75	0.95	0.84	173
Ezil	0.91	0.77	0.83	163
Geez	0.92	0.78	0.84	130

Accuracy			0.84	466
Macro avg	0.86	0.83	0.84	466
Weighted avg	0.85	0.84	0.84	466
Test result:83.691		Loss:0.875		

The diagram given in figure 4.4 and 4.5 shows the accuracy and loss of the trained and tested phase of the proposed model with respect to prepared dataset and Sigmoid activation function. When we have seen the trained phase it had uniform results whereas the tested phase had some up and down curve it doesn€t have uniform results even if it has better results. Generally, the classifier model had better accuracy results and low rate of losing rate relative to the related convolutional neural network models. The loss rate for this model is 0.875. The overall comparison of the SYKZC using ReLU activation function has better accuracy for training and testing as shown below.

Figure 4.4 Training accuracy curve of SYKZC Model using sigmoid

Figure 4.5 Training Loss curve of SYKZC Model with sigmoid

4.4.2.2. Comparison with tanh activation function

The hyperbolic tangent function, or tanh for short, is a nonlinear activation function with a similar structure that produces values ranging from -1.0 to 1.0. The tanh function was chosen over the sigmoid activation function in the late 1990s and early 2000s because it was easier to train and had superior predictive performance.

SYKZC model executed using tanh activation function, it needs more than two hours and we tried with another environment google colab. It is connection oriented but it uses GPU as processor and has better time usage. Relative to anaconda the google colab has better processing speed and we used it. The experiment to evaluate the model, the SYKZC model obtains 97% training accuracy and 84 % testing accuracy using tanh function. The overall average loss rate for the model is 0.875. The tanh activation function has a better loss rate but the accuracy is less than ReLU with 4%.

Table 4.4 Classification Accuracy of training phase of SYKZCNet with tanh

Epoch	Time taken	Number loss	Accuracy	Val_ loss	Val_ accuracy
1/100	4sec 53ms/step	4.8533	0.4954	1.5409	0.4013
2/100	1sec 18ms/step	4.5866	0.5855	1.5220	0. 4700
3/100	1sec 19ms/step	1.7569	0.6170	1.0385	0. 5494

4/100	1sec 18ms/step	1.4589	0.6312	2.3590	0. 4678
5/100	1sec 19ms/step	1.2978	0.6293	0.9418	0. 6824
. .	. .	. .	. .	. .	. .
95/100	1sec 19ms/step	0.1944	0.9369	1.9304	0.8112
96/100	1sec 19ms/step	0.1006	0.9587	0.7804	0.8262
97/100	1sec 21ms/step	0.1187	0.9593	0.8691	0.7790
98/100	1sec 19ms/step	0.0884	0.9665	0.9114	0.7897
99/100	1sec 19ms/step	0.0993	0.9685	0.9723	0.8262
100/100	1sec 19ms/step	0.0831	0.9725	0.9273	0.8369

Class/metrics	Precision	Recall	F1-score	Support
Araray	0.83	0.90	0.86	173
Ezil	0.92	0.72	0.81	163
Geez	0.77	0.90	0.83	130
Accuracy			0.84	466
Macro avg	0.84	0.84	0.83	466
Weighted avg	0.85	0.84	0.84	466
Test result: 83.691		Loss: 0.875		

The diagram given in figure 4.6 and 4.7 shows the accuracy and loss of the trained and tested phase of proposed model with respect to prepared dataset and activation function. When we have seen the trained phase it had uniform results whereas the test phase had some up and down curve it doesn€t have uniform results even if it has better results. Generally, the classifier model had better accuracy results and low rate of losing rate relative to the related convolutional neural network models. The loss rate for this model is 0.875.The overall comparison of the SYKZC using ReLU activation function has better accuracy for training and testing as shown below.

Figure 4.6Training accuracy curve of SYKZCModel using tanh

Figure 4.7Training Loss curve of SYKZCModel using sigmoid

Table 4.5 Comparison SYKZCModel with different Activation function

Model name	Activation function	Max Time taken Per each epoch	Training accuracy	Testing accuracy	Loss rate
SYKZC Model	ReLU	4sec 49ms/step	98%	88%	1.014
	Sigmoid	4sec 53ms/step	96%	84%	0.875
	Tanh	4sec 53ms/step	95%	84%	0.875

The diagram in figure 4.8 shows which activation function has maximum accuracy rate for the given dataset. When we saw the testing phase which had some up and down curve on the graph. All functions don€t have uniform results, but the ReLU function has better results. Generally, the classifier model with ReLU had better accuracy results relative to the related convolutional neural network models. The accuracy for training 98%, for testing 88% and loss rate is 1.014 which is greater than 0.139 from sigmoid and tanh as shown below.

Figure 4.8 comparison of SYKZCM with different activation function

4.5. Comparison of the Proposed Model with other models

We have seen related works that were conducted before this study particularly audio classification with visual representation. Researchers performed their studies which have relation with our study with two main approaches. The shallow machine learning approaches and the deep learning approaches. So, comparisons are performed with related deep learning classification algorithms which are mentioned below. To evaluate the proposed model, we have seen the result obtained from other related models which are performed on image classification with respect to our result and if it has better accuracy & performance well otherwise we must apply different techniques to make our model more accurate and to have better performance. The comparison is performed with the proposed model with other CNN models like AlexNet, VGGNet and ResNet models.

4.5.1. Comparison with ResNet Model

The performance (accuracy and loss value) of the ResNet model is shown in figure below. It takes nearly three hours to train the model in anaconda software and it is better to run in colab to execute within a few minutes even if it is connection based. As the table indicated in below, ResNet obtains 99% training and 85% testing accuracy on our data. It is (about 1% greater than the training phase and around 3% lower than the testing phase) with our model, SYKZC, which obtained 98% training and 88% testing accuracy. The total time required to train this model in anaconda takes more than the time which consumes our model not only in anaconda but also in google colab as shown below in the table for the first 100 epochs per second.

Table 46 Classification Accuracy of training phase of ResNet model

Epoch	Time taken	Loss	Accuracy	Val_ loss	Val_accuracy
1/100	65sec 307ms/step	1.4833	0.5885	1.5584	0.4421
2/100	6sec 171ms/step	1.0174	0.8038	1.5772	0.6438
3/100	6sec 172ms/step	1.0088	0.8035	1.1441	0.6288
4/100	6sec 172ms/step	1.0265	0.8072	1.4278	0.6459
5/100	6sec 173ms/step	0.9407	0.8353	1.1959	0.7146
. .	. .	. .	. .	. .	. .
95/100	6sec 180ms/step	0.2010	0.9979	0.9069	0.8433
96/100	6sec 179ms/step	0.1910	0.9980	0.9500	0.8670
97/100	6sec 180ms/step	0.1816	0.9985	0.9312	0.8648
98/100	6sec 178ms/step	0.1799	0.9985	0.9453	0.8519
99/100	6sec 178ms/step	0.1701	0.9994	0.9699	0.8455
100/100	6sec 180ms/step	0.1729	0.9973	0.9814	0.8519

Class/metrics	Precision	Recall	F1-score	Support
Araray	0.82	0.85	0.84	165
Ezil	0.83	0.82	0.82	165
Geez	0.92	0.89	0.90	136

Accuracy			0.85	466
Macro avg	0.86	0.85	0.86	466
Weighted avg	0.85	0.85	0.85	466
Test result: 85.193		Loss: 0.981		

The diagrams which are described below in figure 4.9 and 4.10 are the accuracy and loss of the ResNet model with prepared dataset and its results as below. The diagram describes the overall accuracy and loss of the training and testing phases using the ResNet model. The overall loss obtained from this model is 0.9814 which is less than the value obtained from our SYKZC model with the value of 0.0326. As clearly shown in the training loss and accuracy curve in figure below, the training accuracy was higher than testing accuracy throughout the curve. It shows that when the number of epochs are increased, accuracy of the model is high and loss of model decreases.

Figure 4.9 Training accuracy curve of ResNet model

Figure 410 Training loss curve of ResNet model

4.5.2. Comparison with VGGNet Model

The performance (accuracy and loss value) of the VGGNet model is shown in figure below. It takes nearly three hours to train model in anaconda software and it is better to run in colab to execute within a few minutes even if it is connection based. As the diagram indicates below, VGGNet obtained 95% for training and 75% testing accuracy on our data. It is lower (about 3% form training and 13% form the testing) than our model, SYKZC, which obtains 98% training and 88% testing accuracy. It also takes a few minutes in the colab as shown below by taking on average 100 epochs per second.

Table 47 Classification Accuracy of training phase of VGGNet model

Epoch	Time taken	Loss	Accuracy	Val_ loss	Val_accuracy
1/100	23sec 428ms/step	3.4110	0.5263	75.0381	0.35
2/100	7sec 198ms/step	1.4331	0.6221	12.2999	0.437
3/100	7sec 195ms/step	0.8476	0.7247	10.1018	0.429
4/100	7sec 198ms/step	0.8624	0.7251	3.2029	0.5365
5/100	7sec 192ms/step	0.8286	0.7330	2.3275	0.3948
. .	. .	. .	. .	. .	. .
95/100	7sec 193ms/step	0.0182	0.9910	1.0424	0.8541

96/100	7sec 192ms/step	0.0507	0.9884	1.4785	0.7876
97/100	7sec 192ms/step	0.1498	0.9621	1.5931	0.7554
98/100	7sec 192ms/step	0.2535	0.9331	11.6287	0.4506
99/100	7sec 192ms/step	0.2193	0.9249	3.6099	0.5236
100/100	7sec 192ms/step	0.1089	0.9562	1.9965	0.7511

Class/metrics	Precision	Recall	F1-score	Support
Araray	0.80	0.89	0.84	165
Ezil	0.65	0.82	0.73	165
Geez	0.93	0.49	0.64	136
Accuracy			0.75	466
Macro avg	0.79	0.74	0.74	466
Weighted avg	0.78	0.75	0.74	466
Test result: 75.107		Loss:1.997		

The diagrams which are described below in figure 4.11 and 4.12 are the accuracy and loss of the VGGNet model with prepared dataset and its results as below. The diagram described the overall accuracy and loss of the training and testing phases using the VGGNet model. The overall loss obtained from this model is 1.997% which is greater than the value obtained from our SYKZC model with the value of 0.983%. It indicates our model has better performance. As clearly shown in the training loss and accuracy curve in figure below, the training accuracy is higher than testing accuracy throughout the curve. It shows that when the number of epochs are increased, accuracy of the model is high and loss of model decreases

Figure 411 The training accuracy curve of VGGNet

Figure 412 The training loss curve of VGGNet model

4.5.3. Comparison with AlexNet Model

The performance (accuracy and loss value) of the AlexNet model is shown in figure below. It takes nearly an hour to train the model in anaconda software and it is better to run in colab to execute within a few minutes even if it is connection based. As the diagram indicated below, AlexNet obtains 98% training and 82% testing accuracy on our data. It is (same value from training and 6% lower for the testing) relative to our model, SYKZC, which obtains 98% training and 88% testing accuracy. It takes a few minutes in the google colab taking on average 100 epochs per second.

Table 4.8 Classification Accuracy of training phase of AlexNet model

Epoch	Time taken	Loss	Accuracy	Val_ loss	Val_accuracy
1/100	5sec 80ms/step	2.5324	0.5386	68.1952	0.3541
2/100	1sec 37ms/step	0.8011	0.6854	4.9687	0.5322
3/100	1sec 38ms/step	0.7202	0.7381	1.3780	0.6674
4/100	1sec 38ms/step	0.6258	0.7589	1.9512	0.5043
5/100	1sec 38ms/step	0.5370	0.7867	4.2035	0.4635
. .	. .	. .	. .	. .	. .
95/100	1sec 38ms/step	0.0449	0.9853	0.8664	0.8155
96/100	1sec 38ms/step	0.0709	0.9894	2.1165	0.6867
97/100	1sec 38ms/step	0.0799	0.9744	1.4272	0.8090
98/100	1sec 38ms/step	0.0395	0.9869	1.4815	0.8240
99/100	1sec 38ms/step	0.0323	0.9894	1.1787	0.7918
100/100	1sec 27ms/step	0.0470	0.9855	1.2620	0.8262

Class/metrics	Precision	Recall	F1-score	Support
Araray	0.81	0.92	0.86	165
Ezil	0.76	0.86	0.80	165
Geez	1.00	0.68	0.81	136
Accuracy			0.83	466
Macro avg	0.86	0.82	0.82	466
Weighted avg	0.85	0.83	0.83	466

Test result: 82.618	Loss: 1.262

The diagrams which are described below in figure 4.13 and 4.14 are the accuracy and loss of the AlexNet model with prepared dataset and its results as below. The diagram described the overall accuracy and loss of the training and testing phases using the AlexNet model. The overall loss obtained from this model is 1.262 which is greater than the value obtained from our SYKZC model with the value of 0.208. It indicates our model has better performance. As clearly shown in the training loss and accuracy curve in figure below, the training accuracy was higher than testing accuracy throughout the curve. It shows that when the number of epochs are increased, accuracy of the model is high and loss of model decrease.

Figure 413 The training accuracy curve of AlexNet

Figure 414 The training loss curve of AlexNet

4.5.4. Models comparison summary

Table 49 Model comparison

Model name	Max Time taken Per each epoch	Training accuracy	Testing accuracy	Loss rate	Size of model (MB)
SYKZC Model	4sec 49ms/step	98%	88%	1.014	8.73
ResNet Model	65sec 251ms/step	99%	85%	0.9814	25.49
VGGNet model	23sec 428ms/step	95%	75%	1.9965	745.31
AlexNet model	5sec 80ms/step	98%	82%	1.2620	343.56

The above table shows us the overall accuracy and loss of the training and testing phase of the developed model with relative to the other related models. It has best accuracy as compared to the remaining models especially for testing phase and also it has less percent of loss rate. The amount of time needed to execute the given input image data in anaconda and google colab environments required less time and the final one is the size of our model has less size relative to the other models.

4.6. Summary

Generally, the dataset which is appropriated for this study was collected from zema Gubae, bet specifically Deggwa scholars. We collected kum zema starting from Wudasie Maryam zema up to Deggwa since these courses are given by scholars of Deggwa. The total number of data taken for this study was more than 1555 and by segmenting with equal size of ten minute. The model stakes the converted data in image form. We used around 1555 images generated from the audio and this data with the size 70% of the data as training and of 30% for testing. In order to implement the coding part, we have used python with tensor flow and Keras as a backend and several libraries were imported. Specifically, Librosa was used for audio files since we applied audio signal processing as well as image processing together with sequence. The model runs on a core-i5 pc with 4 GB RAM and we tried with the first 100 epochs. We applied different techniques to maximize the classifier accuracy and performance. we have seen the result obtain from our model and other models so, our model performs better classification

Chapter Five: Conclusion and Future Work

5.1. Conclusion

Music information retrieval is researchable area and focused on the extraction of information from music audio and musical notes. It includes music genre classification, music transcription, instrument classification, beat detection, blind instrument separation, capturing musical features, such as melody, harmony and rhythm to name a few. St. Yared music is a part of this area which involves St. Yared zema. It is the technique of producing pleasing sound that makes the listeners. We used the word zema interchangeably, pleasing sound, chant, and melody.

The aim of this study was classification saint Yared kum zema classification using convolutional neural networks. To achieve this objective, we formulated three research questions which were answered by the research. The first question was which types of melody in zema Gubaebet grouped under the genres of Araray, Ezil and Geez. We provided answers for this question when we collected from experts like Wudasie Maryam zema song with Araray and Ezil, Mestegab zema song with Geez and Ezil, Selamta, Tsome Digua and Digua song with three genres of zema. We prepared the dataset with three folder name equivalent to the classes name

The second research question was the technique applied to classification of kum zema. We used a deep machine learning classification technique. The collected dataset was initially in audio form and applied different preprocessing techniques to have uniform transformation of spectrogram image. The input for our convolutional neural network was image with RGB and specified dimensions. The convolutional neural network filter, extract relevant features, reduction of dimension and finally classify into appropriate classes using SoftMax classifier.

This study provided significance with two perspectives. The first one was from the practical perspective, for the problems which stated before this study offered supportive information for flocks who have interest in the traditional school. It also minimized the generation gap between modern education students and the traditional school disciples to have nearly common understanding about St. Yared compositions. It enabled any interested group as well as foreign tourists to have some knowledge about Saint Yared zema types. The second one was from a methodological and scientific perspective, it opened a roadmap for researchers to perform in Saint Yared compositions. To classify the audio data into one of the three classes two methods were there. Those were Extracting acoustic features of the audio data and Visual representation of audio data with waveform and spectrogram. We applied the second method because it is better as compared to the first one with several ways.

The classifier designed using the proposed architecture has a total parameter of 750,403 from these parameters 747,139 trainable parameters and the remaining 3,264 parameters are non-trainable. It was compiled using Adam as an optimizer with a learning rate of 0.001. The loss function that was used was categorical cross entropy and it was trained for 100 epochs using 32 as a batch size. Data for this research was collected from internet repositories and from zema Gubaebet particularly from Deber abay St. Gebreal monastery Gubaebe. The collected data passes through pre-processing steps and is given to the neural network architecture so that the model could be trained. A total of 1555 audio segment zema with three classes (Geez, Ezil, and Araray) were collected. To develop the prototype, we have used python programming language and Keras deep learning framework with TensorFlow as a back-end. In addition, we have used Jupyter Notebook and google colaboratory to run all the experiments.

The results obtained from the experiments measured the performance of the proposed model using only visual features audio. The accuracy of the classifier model is 98% for the trained model and 88% for the tested model. The accuracy showed that the model had better classification performance and its loss rate was 1.014

5.2. Contribution of the Research

This research has the following contributions:

The focused on classification of St. Yared zema, initially our data was waveform audio data which is preprocessed by applying Audio signal processing then after transforming the preprocessed audio segmented file into spectrogram image. The audio transformed into spectrogram images and applied image processing technique because our model took spectrogram images with specified dimension as input and with different layers of the convolutional neural network filter out the image with the form of pixel, finally using SoftMax classifier grouped into appropriate classes

Some of the data needed for the study were collected from an uncontrolled Environment in this case several external noise were there so to make our data free from unnecessary interference of waves we have used noise reduction techniques.

There was no prepared dataset before we conducted this study so to accomplish our research work we collected zema from zema Gubaebet as well as recorded data from internet sources which were filtered by traditional school scholars.

Additionally, we showed that from the acoustic representation and classification of audio zema the visual representation and classification techniques will lead to an increase in accuracy of the audio data only classifier.

Lastly, this research work showed that it is possible to classify St. Yared zema using deep learning algorithm which reduces the time it would have taken extracting features manually.

5.3. Future Work

We have achieved good results in this research but that does not mean it could not be improved. To increase the accuracy or performance of the model we recommend trying different approaches such as:

Applying both acoustic feature extraction methods and visual representation of audio data may maximize the accuracy and performance of classifier model

Increasing the dataset also has a great impact on classifier model. When the number of input data increases, the ability of the classifier model becomes better.

Even maximization of the time interval for the audio segment may also have great effect so for the next study increase the audio segmentation time interval may lead better result

Applying data augmentation techniques on the audio signal such as addition of a noise, using different loudness range, time stretching and pitch shifting.

Adding textual information (features) other than audio and video such as metadata found in Zema itself such as the Singers name could give us additional information.

Using a pre-trained network like LSTM may maximize the accuracy rate for the model

Using more representational data and complex network structure such as 3D CNN that learns the visual and temporal features from the audio at the same time.

References

Ma et al. (2002). Iris Recognition Based on Multichannel Gabor Filtering. 2.

Abebe. (1986 E.C, 11 4). A brief history of Saint Yared. Retrieved 6 23, 2012, from Ethiopianorthodoxchurch.org.

Asim and Siddiqui. (2017). Automatic Music Genres Classification using Machine Learning. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, N.

Ayele. (2007, Nov 29). St. Yared€ the great Ethiopian composer. Retrieved 3 1, 2020, from Tadias megazine: http://www.tadias.com/11/29/2007/st-yared-the-great-ethiopian-composer/

Badshah et al. (2019). Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. Conference paper, 2-3.

Belai. (1991). Ethiopian civilization.

Bilal Er* and Aydilek. (2019). Music Emotion Recognition by Using Chroma Spectrogram and Deep Visual Features. International Journal of Computational Intelligence Systems, Vol. 12(2), 1622-1626.

BISANDU. (2016). Design science research methodology in Computer Science and Information System. 1.

Boixeda. (June 2019). URBAN SOUNDS CLASSIFICATION USING DEEP LEARNING. BARCELONATECH.

C. Knight et al. (2019). Pre-processing spectrogram parameters improve the accuracy of bioacoustic classification using convolutional neural networks. The International Journal of Animal Sound and its Recording, 8.

church, E. o. (2019). Ethiopia orthodox tewahido online spritual school. Retrieved from Debelo: http://debelo.org/

Costa et al. (2011). Music Genre Recognition Using Spectrograms. conference paper, 151-154.

EOTC. (2003). Ethiopian orthodox church. Retrieved from the Ethipoian orthodox tewahdo church faith and order: http://www.ethiopianorthodox.org/

Girma. (2014, march). Text-to-Hymn Synthesis for St Yared Hymn Notations. 3.

Gonzalez. (June 2019). Digital Image Processing Second Edition. University of Tennessee: Tom Robbins.

H.E.V et al. (2007). A Novel Automatic Noise Removal Technique for Audio and Speech Signals. Audio Engineering Society.

Habtemaryam. (2012, 10 25). è Ü C u .- u
(è ¢uî5ë¦-vö-5pË ödp--5uë• %Óu"ú–u). Retrieved 10 25, 2012, from
è ¢uî5ë¦-vö-5pË ödp--5uë• : ú–u"%–Óu www.ethiopianorthodox.org

Hazen and Daoud. (2014, 02 22). The Liturgy of the Ethiopian Orthodox. 4. Retrieved 03 1, 2020

Jawaherlalnehru and Jothilakshmi. (2019). Music Instrument Recognition from Spectrogram Images Using Convolution Neural Network. International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-9, 1076-1079.

Jonya and Iswanto. (2017). Indonesian€s Traditional Music Clustering Based on Audio Features. 2nd International Conference on Computer Science and Computational Intelligence, 175-178.

Karthikeyan and Mala. (2018). CONTENT BASED AUDIO CLASSIFIER and FEATURE EXTRACTION USING ANN TECNIQUES. International Journal of Innovative Research in Advanced Engineering, Volume V, 6-106, 107.

KHAMPARIA et al. (2019). Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network. SPECIAL SECTION ON NEW TRENDS IN BRAIN SIGNAL PROCESSING AND ANALYSIS, 7717-7721.

Khan et al. (2016). A Survey of the Recent Architectures of Deep Convolutional Neural Networks. Internatinal journal, 22-27.

Krizhevsky et al. (2012). ImageNet Classification with Deep Convolutional Neural Networks. dvances in neural information processing systems

L.Chavis. (2011, april 5). Saint Yared. Retrieved from BlackPast: https://www.blackpast.org/global-african-history/saint-yared-505-571/

M. Wu et al. (2011). Combining Visual And Acoustic Features For Music Genre Classification. International Conference on Machine Learning and Applications, 126.

Mengstu. (2008 E.C). > õ Ø E ñ5ë õ. addis abeba: Ethiopian orthodox tewahido church.

Mezmur, T. (2011). Traditional Education of Ethiopia orthodox Tewahido Church and Its potential for Tourism Development. 7.

Minhas and Javed. (2009). Iris Feature Extraction Using Gabor Filter. 253.

Mustaqeem and Soonil Kwon. (2019). A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition.

N. Silla and L. Koerich. (2009). Feature Selection in Automatic Music Genre Classification. Tenth IEEE International Symposium on Multimedia, 40.

Nasridinov1 and Park. (2014). A Study on Music Genre Recognition and Classification Techniques. International Journal of Multimedia and Ubiquitous Engineering, 32.

Nasrulla and Zhao. (2019). Music Artist Classification with Convolutional Recurrent Neural Networks. 1-2.

P. Dhanalakshmi et al. (2010). Classification of audio signals using AANN and GMM. Applied Soft Computing.

Peffers et al. (2006). The design science research process: A model for producing and presenting information systems research.

Purwins et al. (2019). Deep Learning for Audio Signal Processing. IEEE Journal of Selected Topics in Signal Processing, VOL. 14, 1.

SENKORIS. (2018). THE ETHIOPIC SEMIOSIS: HISTORY, APPLICATION AND INTERPRETATION WITH REFERENCE TO THE HYMNAL BOOKS OF ST. YARED. 7.

shelemay. (1982). concept of sacred music in ethiopia.

Shelemay et al. (1993). Oral and written transmission in Ethiopian Christian chant. Cambridge university press, Vol. 12 (1993), 59.

Smith. (1999). The Scientist and Engineer's Guide to Digital Signal Processing. San Diego, California: California Technical Publishing.

Tadese. (2018). è dp--5rë•e- •Eõ5ëõ. addis Ababa, Ethiopia: ethiopian orthodox tewahido church mahiber kidusan.

Woube. (2018). Education of the Ethiopian Orthodox Church: Personal Reflection on Nibab Bet and Zema bet. Journal of Ethiopian Church Studies, 1, 15.

Wyse. (2017). Audio spectrogram representations for processing with Convolutional Neural Network. Proceedings of the First International Workshop on Deep Learning and Music joint with IJCNN, 37-41.

Yandre M. G. et al. (January 2017). An Evaluation of Convolutional Neural Networks for Music Classification Using Spectrograms.

Appendices

A. The input image generated from audio

Sample of spectrogram image for Geez Zema

Sample spectrogram image of Ezil Zema

Sample spectrogram image of Araray zema

B. The result ofproposedmodel as well as other models comparison

SYKZCModel training and testing result

ResNet model result with our dataset

VGGNet model result with our dataset

AlexNet model result with our dataset