Intelligent Guitar Chord Recognition Using Spectrogram-Based Feature Extraction and AlexNet Architecture for Categorization

Dr. Nilesh B. Korade[1], Dr. Mahendra B. Salunke[2], Dr. Amol A. Bhosle[3], Dr. Sunil M. Sangve[4], Dhanashri M. Joshi[5], Gayatri G. Asalkar[6], Dr. Sujata R. Kadu[7], Dr. Jayesh M. Sarwade[8]

Assistant Professor, Department of Computer Engineering, JSPM's Rajarshi Shahu College of Engineering, Tathawade, Pune-411033, Maharashtra, India[1, 5, 6]

Assistant Professor, Department of Computer Engineering, PCET's, Pimpri Chinchwad College of Engineering and Research, Ravet, Pune-412101, Maharashtra, India[2]

Associate Professor, Department of Computer Science and Engineering, School of Computing, MIT Art, Design and Technology University, Loni Kalbhor, Pune-412201, Maharashtra, India[3]

Professor, Department of Artificial Intelligence and Data Science, Vishwakarma Institute of Technology, Bibwewadi, Pune-411037, India[4]

Assistant Professor, Department of Computer Engineering, Terna Engineering College, Mumbai, Maharashtra, India[7]

Associate Professor, Department of Information Technology, JSPM's Rajarshi Shahu College of Engineering, Tathawade, Pune-411033, Maharashtra, India[8]

***Abstract*—Chord prediction plays a key role in the advancement of musical technological innovations, such as automatic music transcription, real-time music tutoring, and intelligent composition tools. Accurate chord prediction can assist musicians, educators, and developers in constructing tools that help in learning, playing, and composing music. Background noise and audio distortions may have an impact on chord prediction accuracy, particularly in real-world situations. Chords can have distinct voicings or finger positions on the guitar, resulting in slight variations in audio representation. This study focuses on the classification of guitar chords using techniques of deep learning. There are eight major and minor guitar chords in the dataset. They have been turned into spectrograms, chromagrams, and Mel Frequency Cepstral Coefficients (MFCC) so that features can be extracted. Various deep learning architectures, including CNN, ResNet50, AlexNet, and VGG, were employed to classify the chords. Experimental results demonstrated that the spectrogram-based AlexNet model outperforms others, achieving good accuracy and robustness in chord classification. The proposed study demonstrates the efficiency of spectrograms and advanced deep learning models for audio signal processing in music applications. By automating chord detection, this study provides beneficial resources for music learners as well as educators, enabling more efficient learning and real-time feedback during practice sessions.**

***Keywords*—Chords; prediction; spectrogram; chromagram; Mel Frequency Cepstral Coefficients; AlexNet**

I. Introduction

Nowadays people prefer to learn music online, particularly through video lessons. There is a growing demand for systems that can provide real-time feedback and support. Many aspiring musicians struggle to assess their own skill level, especially while learning to play musical instruments such as the guitar. The demand for an automated system capable of effectively identifying and classifying guitar chords is rising, as it can assist learners in determining whether they are performing the chords correctly [1].

A guitar typically contains six wires or strings stretched across the neck and body, each tuned to a distinct pitch, which makes sound when plucked or strummed [2]. The standard tuning of a guitar, from the lowest (thickest) to the highest (thinnest) string, is EADGBE. Frets are metal strips that run horizontally across the guitar neck. The frets separate the neck into parts. When you press a string against a fret, it shortens its vibrating length and changes its pitch. Theoretically, there are endless chords on a guitar due to changes in tunings, fingerings, and voicings [3]. Typically, guitarists utilize a more manageable set of chords. Based on the 12 notes of the chromatic scale, there are 12 distinct major and minor chords. Major chords, which include the root note, major third, and perfect fifth, are known for their bright, happy sound. Minor chords are typically described as having a darker, sadder sound, and they are made up of the root note, minor third, and perfect fifth [4]. The list of 12 major and 12 minor chords is presented in Table I.

TABLE I. Guitar Basic Major and Minor Chords

Major Chords	Minor Chords
A major (A)	A minor (Am)
A# major (A#) or Bb major (Bb)	A# minor (A#m) or Bb minor (Bbm)
B major (B)	B minor (Bm)
C major (C)	C minor (Cm)
C# major (C#) or Db major (Db)	C# minor (C#m) or Db minor (Dbm)
D major (D)	D minor (Dm)
D# major (D#) or Eb major (Eb)	D# minor (D#m) or Eb minor (Ebm)
E major (E)	E minor (Em)
F major (F)	F minor (Fm)
F# major (F#) or Gb major (Gb)	F# minor (F#m) or Gb minor (Gbm)
G major (G)	G minor (Gm)
G# major (G#) or Ab major (Ab)	G# minor (G#m) or Ab minor (Abm)

In our research, we demonstrate the 12 major and minor guitar chords with a standard notation system that incorporates the 6-string arrangement, fret numbers, and finger locations. This diagram is commonly used by guitarists to learn how to play each chord on the guitar. For each chord, the string number (ranging from 1 to 6, with string 1 being the thinnest and string 6 being the thickest) is specified along with the fret numbers that correspond to where the fingers should press on the fretboard [5]. For example, in the A Major (A) chord, String 6 (E) is not played, String 5 (A) is played open, String 4 (D) is pressed at the second fret with the third finger, String 3 (G) is pressed at the second fret with the second finger, String 2 (B) is pressed at the second fret with the first finger, and String 1 (E) is played open [6]. Fig. 1 presents a representation of a few chords.

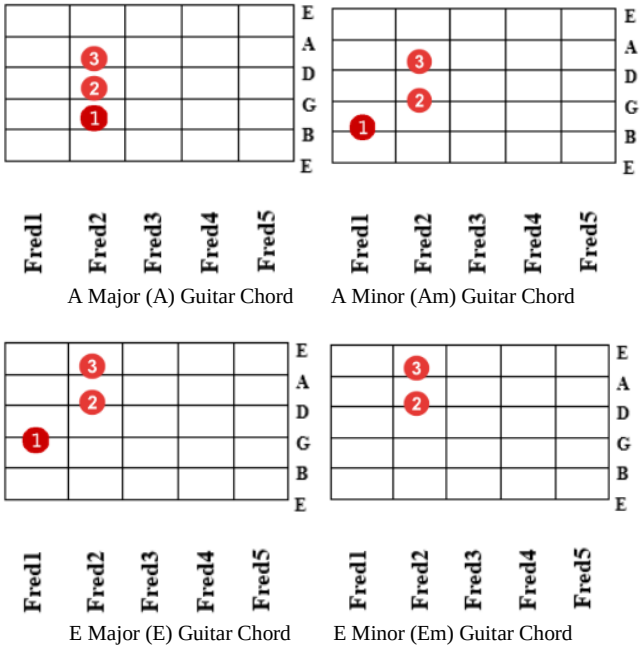


Fig. 1. Representation of chords.

To investigate the most effective representation of audio data for chord classification, we examined three extensively used feature extraction techniques: spectrogram [7], chromagram [8], and Mel-Frequency Cepstral Coefficients (MFCC) [9]. The model trained on spectrogram data outperformed the others in effectively predicting chords, demonstrating its capacity to capture pitch and harmonic relationships that are important for chord differentiation. Several deep learning architectures, such as CNN, ResNet50, AlexNet, and VGG-19, were used to perform the classification challenge [10, 11]. When trained on a limited dataset, AlexNet outperformed the other models, providing the highest accuracy while utilizing the fewest computational resources and training time. This makes AlexNet ideal for cases in which data availability is limited or rapid deployment is required. The approaches applied in the present research included preprocessing audio files to generate chromagrams, which were then scaled to standard dimensions. These representations were then utilized to train the classification models. The precision, recall, F1-score, and accuracy metrics were used to assess the performance of each architecture, ensuring an accurate assessment of its effectiveness. By evaluating the audio input from the learner's performance, it can provide immediate feedback, measure progress, and recommend areas for growth. This approach can be a useful aid for beginners studying guitar, allowing them to rapidly recognize and correct errors while practicing chords. It can function as a virtual tutor, providing assistance when a live instructor is unavailable.

The structure of this research article is outlined as follows: Section Two examines the most recent investigations on music categorization and discusses significant research gaps. Section Three illustrates the methodology, comprising details regarding the dataset, the feature extraction method, the architecture of the AlexNet model, and the assessment criteria used. Section Four discusses the performance of various feature extraction techniques and models in predicting chords based on the selected metrics. Finally, Section Five presents the conclusion, followed by a list of references used.

II. Literature Survey

Automated chord recognition is an essential aspect of music information retrieval (MIR) that assists with applications such as music transcription, evaluation, and production. The work describes a chord detection method that combines a revised Pitch Class Profile (PCP) feature with Support Vector Machines (SVM) for classification. The PCP feature enhances music chord recognition by efficiently capturing harmonic structures while reducing noise and octave-related ambiguity. SVM is employed as it has high categorization proficiency, allowing precise chord detection across varied musical works. The methodology was evaluated on four songs: *Good to Be Alive*, *Ghost*, *The Royal Wedding Song*, and *Trouble I'm In*, with accuracy rates of 91%, 93%, 95%, and 98%, respectively. Investigations reveal that the approach performs well at detecting chord progressions, with satisfactory outcomes. Employing PCP as a classifier allows for evaluating a song's emotional undertones quickly and accurately, with broader ramifications. It enhances the comprehension of musical concepts by providing essential insights into theoretical frameworks [12].

The traditional methods for characterizing audio signals use handmade features such as MFCC, spectral centroid, or zero-crossing rate (ZCR). The recent improvements have centered on DL approaches and intelligent feature extraction to boost performance. The investigation explores the implementation of textural features and Mel-spectrograms produced from the short-time Fourier transform to capture the frequency content of an audio signal for accurate music identification. Texture features, primarily utilized in processing images, produced promising results in audio analysis by capturing patterns and fluctuations in representations of spectrograms. A diversified song recording dataset of 404 audio files from four unique classes of Arabic music has been gathered and transformed into Mel-spectrograms, using which various texture features are extracted. Each Mel-spectrogram undergoes a two-dimensional Haar wavelet processing before feature extraction with Local Binary Patterns (LBP), Histogram of Oriented Gradient (HOG), and Gray Level Co-occurrence Matrix (GLCM). To assess classification performance, several machine learning methods were used, and the proposed approach was evaluated on two datasets: the recently collected Arabic music dataset and the commonly utilized GTZAN dataset. Using five-fold cross-

validation, the research findings demonstrated that the XGB classifier performed better, giving 97.8% accuracy, 97.7% F1- score, 97.7% recall, and 97.8% precision [13].

Folorunso et al. explore the understudied topic of automatic genre categorization for Nigerian traditional music utilizing the ORIN dataset, which comprises 478 music with five genres: Apala, Fuji, Waka, Juju, and Highlife. The Librosa Python package has been utilized to retrieve timbral texture and tempo information for 30-second portions of each song. The study employed the global mean (Tree SHAP) method to assess feature significance and its impact on the model for classification. Timbral features of texture allow for differentiation between comparable beats and melodies. The timbral texture can be used to compute a variety of properties such as MFCC, spectral centroid, tempo, flatness, bandwidth, contrast, sample-silence, zero-crossing-rate, and so on. After the extraction, information about the level of variance and dispersion is extracted, including mean, skewness, kurtosis, standard deviation, minimum, and maximum values. The methodology, which combines feature extraction with classifiers, delivers insights into the performance of several models for genre categorization. The Tree SHAP method is based on Shapley Additive Explanations (SHAP), a game-theoretic strategy for assigning importance scores to input data in interpreting tree-based model predictions. Four classifiers were implemented for genre classification, and among these, the XGBoost classifier has an outstanding accuracy of 81.9% and recall of 84.5% [14].

Categorizing music genres can be automated, and many approaches have been presented in recent years for accomplishing this objective; however, analysis shows that there is still an inequality between the observed outcomes and an optimal categorization approach. The Teng Li presented an approach that involves preprocessing the input signals and then demonstrating the properties of each signal using a combination of MFCC and STFT features. The proposed technique combines two independent CNN models optimized with black hole optimization (BHO) to evaluate MFCC and STFT data, and the results of the two models are integrated to identify the music genre by applying a SoftMax classifier. The GTZAN and Extended-Ballroom datasets were used for assessing the performance of the suggested technique for categorizing music genres. Test outcomes revealed that the suggested approach obtained a classification accuracy of 95.2% on GTZAN and 95.7% on Extended-Ballroom datasets [15].

Carsault et al. investigate the real-time evaluation and forecasting of musical chord patterns to improve innovative methods in composing music and performance. The chords have inherent hierarchical and functional linkages that are critical for comprehending and anticipating musical forms. The input is an audio real-time recording of a musician, which is processed in order to produce a time-frequency representation assisting in the identification of chord sequences. Each beat of the music is then tagged with a chord, and the prediction module makes use of these chords in recommending what might occur next in the sequence. Chord prediction employs several loss functions, such as Tonnetz and correct notes, for boosting accuracy in both diatonic and non-diatonic music. The Bi-LSTM model is utilized in chord prediction because it retains past and future interdependence in chord sequences, making it ideal for sequential music prediction services. This method provides a deeper understanding of the model's performance than typical accuracy measurements [16].

Deep learning models substituted earlier ML approaches that relied on hand-crafted features and transformed the area of music classification by allowing pattern features to be learned automatically. Researcher Jiyang Chen stated that CNNs face challenges in accurately simulating global features that are essential for identifying music signals with temporal properties due to the influence of the local receptive field. The proposed hybrid architecture based on CNN and Transformer encoder worked on transforming audio signals into mel spectrograms. The CNN architecture consists of four 2D convolutional layers, followed by batch normalization, max pooling, and ReLU as the activation function. The transformer encoder has two layers, each of which has multi-head attention, a multi-layer perceptron, and two normalizing layers. The CNN primarily captures low- level and localized features from the spectrogram, followed by the transformer encoder, which processes these features globally to extract high-level and abstract semantic information. The approach is evaluated on the GTZAN with 100,000 tracks and FMA datasets, contains 8000 music clips, and produces amazing outcomes with lower parameters and an increased inference speed, with accuracy 87.41, precision 87.93, recall 87.58, and F1 score 87.28 [17].

As music is a sort of time-series data, which makes it difficult to build a robust MGC, Zhiqiang Zheng introduced the DL-Enabled MGC approach, which aims at boosting the precision and effectiveness of genre categorization work. The proposed strategy extracts significant features from raw musical data by converting pitches from input Musical Instrument Digital Interface (MIDI) files into vector sequences employing the Pitch2vec method. A hybrid model employs bidirectional long short-term memory optimized using cat swarm optimization to successfully capture temporal dependencies in music data. CSO is a swarm intelligence-based optimization technique that fine-tunes hyperparameters, including learning rate, number of hidden layers, and activation functions, resulting in improved model convergence. BiLSTM analyzes data both forward and reversed, facilitating the model to capture past and future dependencies concurrently. The DLE-MGC technique was examined using the MIDI music dataset containing thirteen types of music utilizing 1000 and 2000 epochs. The results demonstrate that with 1000 epochs, the DLE-MGC methodology has offered a precision, recall, F1-score, and accuracy of 94.97%, 95.97%, 96.53%, and 95.42%, and with 2000 epochs, 95.84%, 95.93%, 96.71%, and 95.77%, respectively [18].

Yu-Huei Cheng presents an experimental approach for recognizing genres of music that employs graphical representations of sound data and effective ML algorithms. To achieve excellent classification accuracy, the visual Mel spectrum was employed as a feature representation, together with the YOLOv4 neural network architecture. The visual Mel spectrum, which captures both temporal and spectral aspects of the music, is provided as input for the classification model, allowing for the extraction of rich, discriminative properties relevant to several music genres. The YOLOv4 architecture has

been chosen as its potential to effectively handle visual data makes it suitable for interpreting visual Mel spectrograms, resulting in effective feature learning and categorization. The 1.6 GB GTZAN dataset was used for the assessments, and the proposed technique achieved 91.49% accuracy for training and 97.93% for testing. This study uses YOLO, which has never been utilized for music genre classification, and it has research significance [19].

A review of previous studies indicates that the majority of music analysis research focuses on genre classification rather than chord prediction. There is a significant gap in research on automatic chord recognition. The majority of studies use CNN for music classification tasks like genre and instrument recognition. While deep learning models are commonly used for classification, the effect of various feature extraction methodologies (such as chromagram, spectrogram, and MFCC) on model performance has not been thoroughly investigated. Most research does not examine how different feature representations impact classification accuracy, leaving a gap in determining the optimal characteristics for chord recognition.

III. Methodology

We loaded the chords dataset containing .wav audio files for different chord categories and extracted the Chromagram, spectrogram, and MFCC features. The normalization is performed on each extracted feature to ensure that all features are on the same scale [20]. Each feature set was used to train and evaluate several deep learning models in order to determine the most effective feature extraction technique and chord prediction model. PyAudio was used to capture live audio from the guitarist while he performed in order to figure out chords in real time. The selected deep learning model was then used to determine the correct chord [21-23]. The proposed methodology is described in Fig. 2.

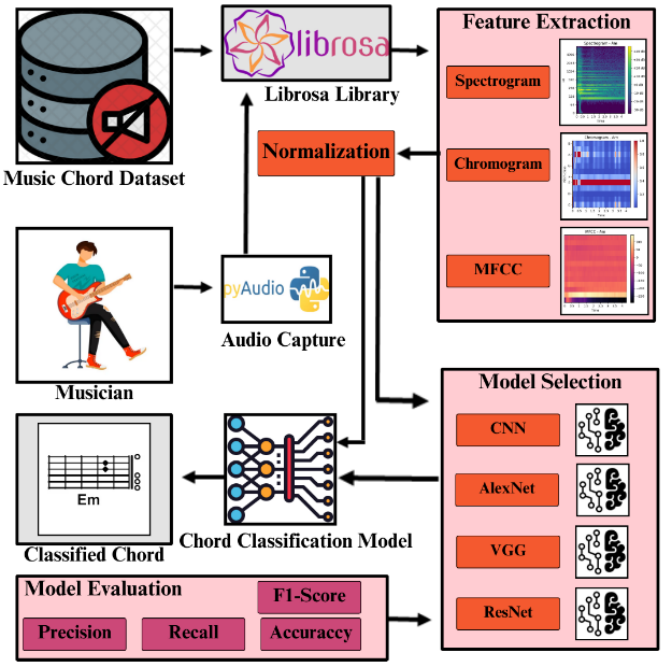


Fig. 2. Proposed methodology.

A. Dataset

We collected a dataset of 1440 audio files in .wav format, each representing eight distinct guitar chords: [Minor: {Em', 'Dm', 'Am', 'Bdim'}, Major: {G', 'C', 'Bb', 'F'}]. Each chord was played on a guitar by hand, with speed and duration variations to imitate several acoustic aspects associated with performances in real time. This variation in playing approach promises that the dataset covers a wide range of musical expressions, which enhances the robustness of the chord prediction model. The dataset establishes a robust foundation for constructing and evaluating deep learning models for real-time chord recognition in music. The 1,152 .wav files, representing 80% of the chord dataset, were employed to train the chord prediction model, while the remaining 288 files, which is 20%, were used to test its accuracy and effectiveness.

B. Feature Extraction

Feature extraction is a critical step in chord prediction because it transforms raw audio data into meaningful representations that can be used by machine learning algorithms. We employed three fundamental feature extraction strategies: chromagram, spectrogram, and MFCC, each of which captured a different aspect of the audio stream. Fig. 3 demonstrates the representation of the chromatogram, spectrogram, and MFCC for chords Am and C.

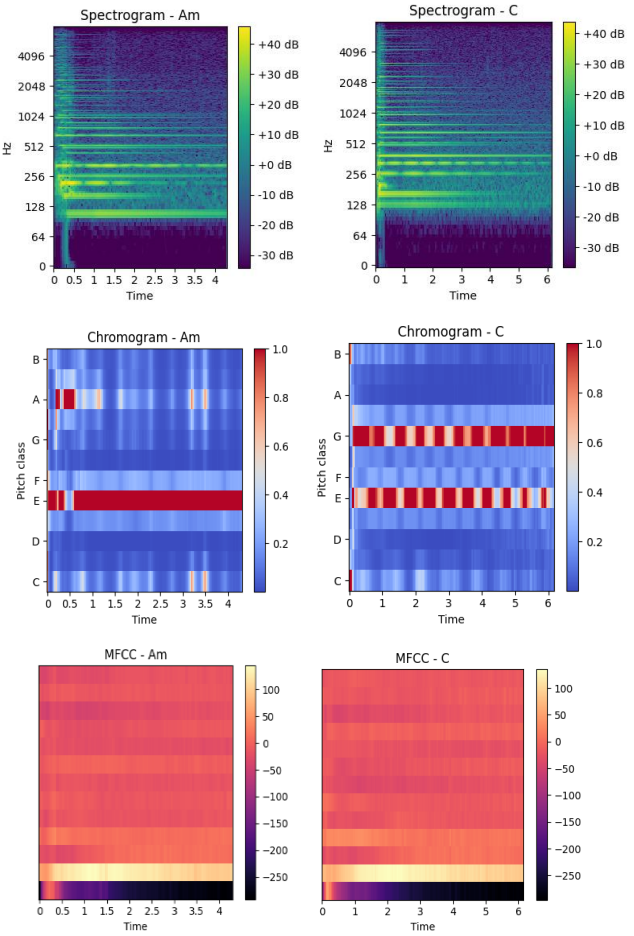


Fig. 3. Representation of chromagram, spectrogram, and MFCC for chord Am & C.

1) Spectrogram: Spectrograms are extensively utilized in domains such as machine learning, voice analysis, audio processing, and seismology, particularly in applications like environmental sound classification and recognition of speech. A spectrogram is a two-dimensional graph where a third dimension is represented by color. Time is represented by the horizontal X-axis, which moves from left to right, while frequency is represented by the vertical Y-axis, which goes from low at the bottom to high at the top [24]. A frequency's amplitude or loudness at a given moment is indicated by its color intensity; brighter colors imply higher, louder amplitudes, while darker shades suggest lower amplitudes. A signal is usually broken down into its frequency components over short time intervals using the Short-Time Fourier Transform (STFT) expressed in Eq. (1), which serves as a foundation for building spectrograms [25]. Eq. (2) uses the squared magnitude of the STFT to derive the spectrogram.

$$S(t,f) = \int_{-\infty}^{\infty} x(\tau)\omega(\tau - t)e^{-j2\pi f\tau}\, d\tau \quad (1)$$

$$Spectrogram(t,f) = |S(t,f)|^2 \quad (2)$$

The sliding window current time location represented by t and f represents the frequency being analyzed. The input signal as a function of time is represented by x(τ). A window function w(τ−t) centered at time t isolates a segment of the signal to analyze. The Fourier kernel $e^{-j2\pi f\tau}$ that transforms the signal into the frequency domain.

2) Chromagram: A chromagram is a representation in the form of an image of the intensity or energy of the pitch classes or musical notes in a signal, irrespective of their octave. As it can capture the melodic and harmonic elements of audio signals, it is frequently utilized in music information retrieval (MIR). The vertical Y-axis represents musical notes that are classified into twelve pitch classes (e.g., C, C#, D, ..., B) in the Western music system, and the horizontal X-axis represents the time evolution of the signal [26]. The color or brightness of a cell in the chromagram represents the energy or intensity of a specific pitch class at a given time. The audio signal x(n) (where n represents discrete time samples) is transformed into the frequency domain using the STFT using Eq. (1). The Eq. (3) translates frequencies in the spectrum to one of the twelve pitch classes [27].

$$p = mod\left(\left\lfloor 12.log_2\left(\frac{f}{f_{ref}}\right)\right\rfloor, 12\right) \quad (3)$$

where, p is Pitch class ranges from 0 to 11, corresponding to C, C#, ..., B. The f represents frequency in Hz, and f_{ref} refers to reference frequency, typically 440 Hz. The $\lfloor\cdot\rfloor$ symbol represents the floor function, which truncates to the largest integer less than or equal to the value. For each pitch class p, the total energy over all octaves at time t is calculated. The chromagram C(t,p) can be calculated as follows:

$$\mathrm{C(t,p)} = \sum_{f\in F(P)}|X(t,f)| \quad (4)$$

where, |X(t,f)| represents the magnitude of the STFT at time t and frequency f. The set F(p) contains all frequencies f that correspond to the pitch class p. To achieve uniform representation across time frames, normalization is employed, expressed in Eq. (5) [28].

$$NC(t,p) = \frac{C(t,p)}{max_p C(t,p)} \quad (5)$$

where, *NC(t,p)* is the normalized chromagram, and maximum chromagram intensity over all pitch classes for a particular time frame t is represented by $max_pC(t,p)$.

3) Mel-frequency cepstral coefficients (MFCC): MFCCs are created by translating an audio signal into a set of coefficients that represent the signal's short-term power spectrum on the Mel frequency scale, which is commonly utilized in speech and audio signal processing operations [28]. MFCCs give a concise representation of an audio signal's spectrum features and have been designed to represent the human auditory system's perception of sound. To amplify high frequencies and balance the spectrum, pre-emphasis is applied to the signal, assisting to retain key information for analysis.

$$Y[n] = X[n] - \alpha \cdot X[n-1] \quad (6)$$

where, X[n] represents the original signal, Y[n] represents the pre-emphasized signal, and α is the coefficient for pre-emphasis, usually 0.95. To examine short-term features, the signal is separated into small overlapping frames spanning about 20-40 milliseconds. Each frame represents a quasi-stationary segment of the signal. To eliminate edge effects, a window function w[n] (especially Hamming) gets applied to each frame 1 to N expressed in Eq. (7) [29].

$$w[n] = 0.54 - 0.46 \cdot cos\left(\frac{2\pi n}{N-1}\right) \quad (7)$$

The number 0.54 ensures that the window has a "base value" at its center, while -0.46 multiplied by the cosine function modifies the form of the window, regulating how quickly the window tapers to zero at its edges. The power spectrum is obtained using the Fourier transform, which transforms the windowed frame into the frequency domain.

$$X[k] = \sum_{n=0}^{N-1} x[n] \,.\, e^{-j2\pi kn/N} \quad (8)$$

where, X[k] is frequency components and k represents the frequency bin index. To map the power spectrum to the Mel scale m(f), it first passes through a Mel filter bank (MFB) expressed in Eq. (9). The scaling ratio 2595 was used to translate frequencies from the Hertz scale to the Mel scale, and the constant 700 is a reference frequency at which the Mel scale begins to diverge significantly from linear scaling. Triangular filters concentrate distinct frequency bands to replicate human hearing. The energy S[m] for each Mel filter is calculated using Eq. (10), where Weight of the k^{th} frequency in the m^{th} filter. Logarithmic scaling L[m] is used to simulate the human auditory system's perception of sound expressed in Eq. (11) [30].

$$m(f) = 2595.log_{10}\left(1 + \frac{f}{700}\right) \quad (9)$$

$$S[m] = \sum_k |X[k]|^2 . H_m[k] \quad (10)$$

$$L[m] = log(S[m]) \quad (11)$$

The Discrete Cosine Transform (DCT) is used to decorrelate log Mel energy and compress them into MFCC coefficients

using Eq. (12), where M is number of Mel filters and n is the coefficient index.

$$MFCC[n] = \sum_{m=0}^{M-1} L[m] \cdot cos\left(\frac{\pi n(2m+1)}{2M}\right) \quad (12)$$

The overall process for MFCC computation C(t, n) is represented in Eq. (13), where t is the time frame index and n is the MFCC coefficient index.

$$C(t,n) = DCT\left(log\left(MFB\left(FFT(x(t))\right)\right)\right) \quad (13)$$

C. Model Selection

Several deep learning architectures were employed to evaluate the chord prediction performance, including CNN, AlexNet, VGG-19, and ResNet-50. The best models were picked based on their demonstrated effectiveness in audio categorization tasks utilizing assessment measures. CNN is a fundamental deep learning model that can extract hierarchical features from input data while accurately capturing spatial patterns, making it suitable for a wide range of classification applications [31]. AlexNet is a deep CNN architecture designed for rapid feature extraction and classification. It employs stacked convolutional layers, ReLU activation, and dropout to improve generalization. A VGG-19 is a deeper convolutional network with 19 layers known for its identical design and potential to capture intricate patterns with small receptive fields and dense layers. ResNet-50 is a residual learning framework that uses skip connections to address the problem of vanishing gradient, resulting in deeper network training and improved classification accuracy [32].

1) AlexNet: AlexNet is a foundational baseline in deep learning for image classification. AlexNet consists of five convolutional layers for feature extraction and three fully connected layers for classification. AlexNet additionally uses max-pooling for spatial dimensionality reduction and GPUs for parallel processing, allowing for effective training on big datasets. It employs ReLU activation functions to incorporate nonlinearity, dropout to reduce overfitting, and data augmentation to improve generalization [33]. The acceptable input image shape is 224×224×3. The first convolution layer extracts low-level characteristics such as edges and corners using 96 filters of size 11×11 with a stride of 4 and an output of 55×55×96. The second convolution layer captures more detailed features and applies local response normalization (LRN). It uses 256 filters of size 5×5×96 with a stride of 1 and generates an output of 27×27×256. Fig. 4 presents an in-depth description of the AlexNet architecture layers, including filter size, filter number, stride, input and output sizes [34].

D. Model Evaluation

1) Various metrics for assessment, such as accuracy, precision, recall, and F1-score, were employed for evaluating the feature extraction technique and model's performance in music chord prediction. These metrics provide an extensive assessment of the model's potential to accurately categorize chords while balancing false positives and false negatives [35-37]. The TP represents the instances where the model correctly predicts the positive class, and TN are the Instances where the model accurately identifies the negative class. The FP is the number of instances for which the model incorrectly classifies it as positive, and the FN is the number of instances for which the model incorrectly classifies it as negative.

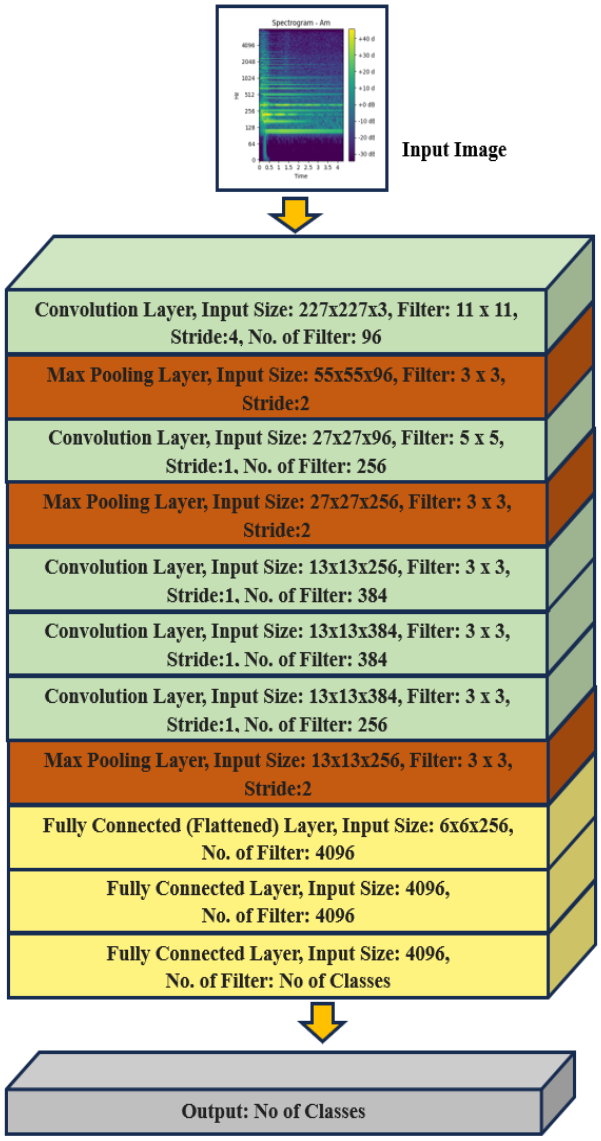


Fig. 4. AlexNet architecture.

$$\text{Accuracy} = \frac{Sum\ of\ Accurately\ Forecasted\ +ve\ samples\ [TP]\ and\ -ve\ Samples\ [TN]}{Total\ Samples(N)} \quad (14)$$

$$\text{Precision} = \frac{Accuately\ forecasted\ +ve\ samples\ [TP]}{sum\ of\ Accuatly\ forecasted\ +ve\ samples\ [TP]\ and\ incorrect\ forecasting\ as\ +ve\ [FP]} \quad (15)$$

$$\text{Recall} = \frac{Accuatly\ forecasted\ +ve\ samples\ [TP]}{Sum\ of\ Accuatly\ forecasted\ +ve\ samples [TP]\ and\ incorrect\ forecasting\ as\ -ve\ [FN]} \quad (16)$$

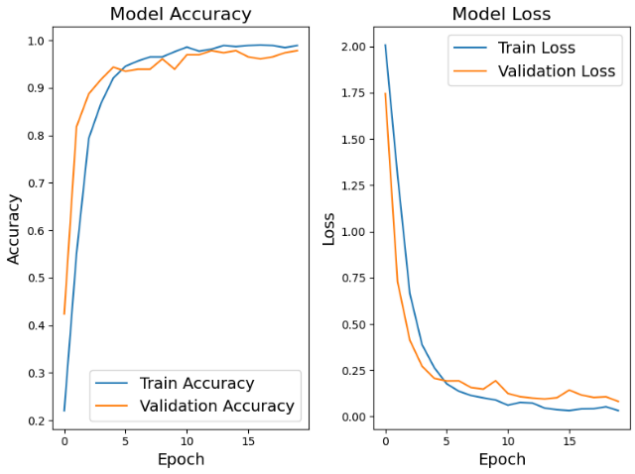
$$\text{F1Score} = 2 * \frac{[\text{Precision} * \text{Recall}]}{[\text{Precision} + \text{Recall}]} \quad (17)$$

IV. Result and Discussion

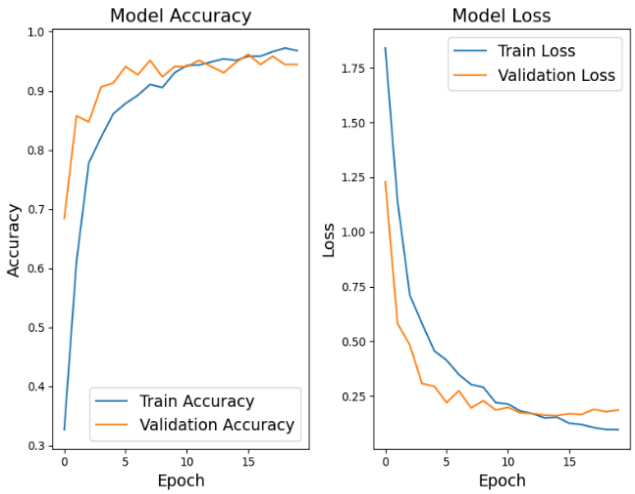
The research study has been carried out on Google Colab, using a T4 GPU for accelerated training. The chords dataset is stored on Google Drive and mounted with the Colab notebook for simple retrieval. Librosa is used for feature extraction, and Keras is utilized to build the CNN model and its variants.

A. Evaluation of Spectrogram, Chromagram, and MFCC Features

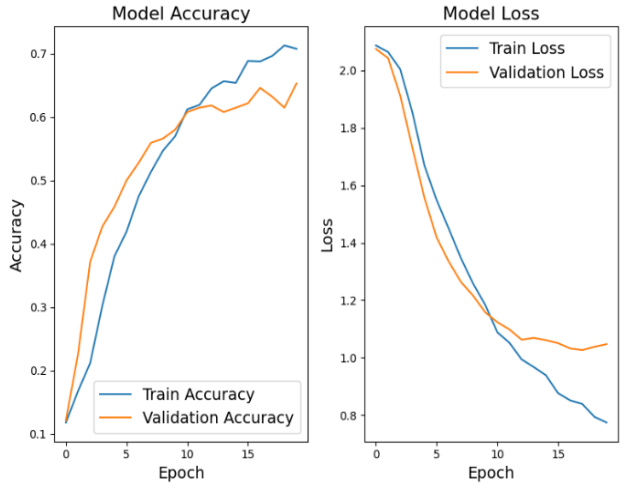
The dataset contains 1440 .wav audio files representing eight different major and minor chords, split into training and testing sets in an 80:20 ratio. To assess the effectiveness of feature extraction approaches, a CNN is trained on each method employing a similar training set. The effectiveness of CNN models trained using different feature extraction techniques was evaluated by plotting training & validation accuracy and loss against the number of epochs. Fig. 5 demonstrates the performance of each feature extraction strategy during CNN training.



Spectrogram



Chromagram



MFCC

Fig. 5. Performance of CNN model trained on spectogram, chromagram and MFCC.

CNN trained on MFCC features has less satisfactory accuracy (65%) than spectrogram and chromagram models. The loss reduction is not as smooth, demonstrating that the model has difficulty with feature representation. The model does not fully converge within 20 epochs, indicating the need for additional training or hyperparameter adjustment. CNN models based on spectrograms and chromagrams outperform MFCC-based CNN models. The spectrogram-based CNN is more accurate, converges in fewer epochs, and has closely aligned training and validation accuracy and loss values, indicating stable learning. In comparison, the chromagram-based CNN takes considerably more epochs to converge, achieves slightly reduced accuracy compared to the spectrogram-based model, and has a greater gap between training and validation accuracy and loss values. The performance of each feature extraction strategy for the test. WAV sound files are presented in Table II.

TABLE II. Performance of CNN on Different Feature Extraction Techniques

Model	Feature Extraction Techniques	Accuracy	Precision	Recall	F1-Score
CNN	Spectrogram	0.94	0.95	0.94	0.94
	Chromagram	0.93	0.94	0.93	0.93
	MFCC	0.65	0.67	0.65	0.65

The results demonstrate that the spectrogram-based CNN model delivers the highest accuracy of 94%, with other assessment measures at 0.94 each. This shows that spectrograms efficiently capture the time-frequency representation of audio inputs, resulting in reliable feature extraction for chord categorization. The chromagram-based CNN model likewise revealed strong performance, with an accuracy of 93%. However slightly less accurate than the spectrogram-based model, the chromagram technique accurately represents harmonic content, offering it as an acceptable alternative for chord detection tasks. MFCCs primarily capture spectral envelope information and are frequently employed in speech processing; however, due to their limited ability to represent harmonic structures, it may be inefficient for musical chord classification.

B. Evaluation of Spectrogram Based Deep Learning Architecture

Choosing an appropriate architecture capable of effectively capturing the extracted feature for accurate chord detection is a vital component of the chord prediction problem. In this experiment, we trained CNNs and their variants, such as AlexNet, VGG-19, and ResNet-50, on spectrogram-based features and evaluated their performance using standard metrics. Table III summarizes the performance of each spectrogram-based model.

TABLE III. PERFORMANCE OF SEVERAL CNN VARIANTS ON SPECTOGRAM

Model	Accuracy	Precision	Recall	F1-Score
CNN [38-40]	0.94	0.95	0.94	0.94
AlexNet	0.96	0.97	0.97	0.97
VGG-19	0.77	0.80	0.77	0.77
ResNet-50	0.91	0.92	0.91	0.91

AlexNet consistently outperformed the other models examined, revealing its potential to capture pertinent information in the spectrogram and being well-suited for the chord prediction task, offering a balanced performance in both detecting and properly classifying guitar chords. The result shows that ResNet-50 and VGG-19 are deeper architectures having more parameters. These deeper models are more suitable for larger datasets, where they may learn from a diverse set of features. With a limited dataset of 1440 samples, these architectures may struggle to generalize successfully, resulting in overfitting. On the other hand, AlexNet and CNNs are less likely to overfit due to their more compact structure, and they can extract useful information from a smaller dataset without becoming overly specialized in it. Fig. 6 demonstrates the performance of AlexNet trained on spectrograms. The confusion matrix demonstrates in Fig. 7 the performance of AlexNet on the testing dataset.

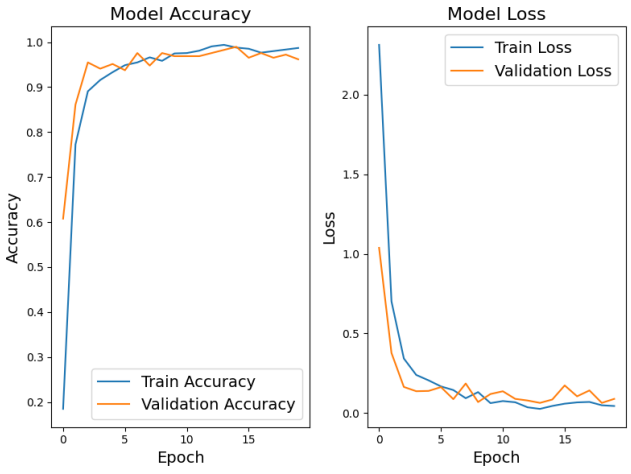


Fig. 6. Performance of AlexNet model trained on spectogram.

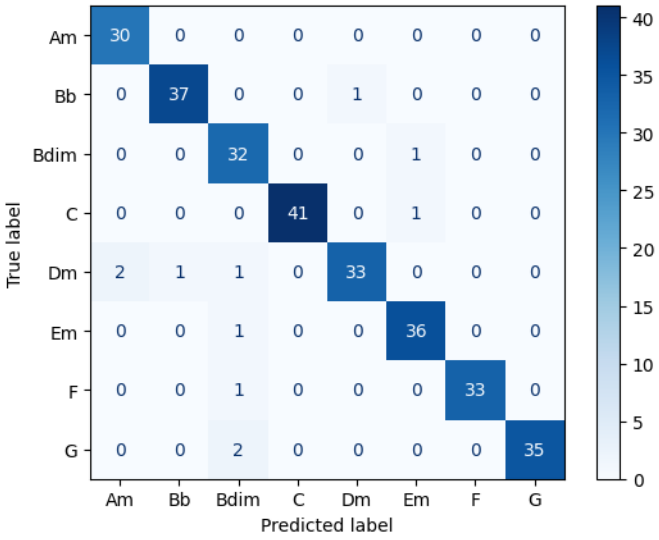


Fig. 7. Confusion matrix of AlexNet model trained on spectogram.

V. CONCLUSION

Intelligent guitar chord categorization is a significant advancement towards improved music transcription processes, assisting musicians and growing music learning tools. Implementing deep learning approaches can build a robust system that appropriately recognizes chords from audio signals, eliminating the need for manual intervention and strengthening real-time chord recognition. The experimental results reveal that spectrogram-based models provide the best classification performance, accurately discriminating distinct chords as compared to chromagram and MFCC-based approaches. AlexNet performed outstandingly on a smaller dataset with minimal training time, making it suitable for low-data scenarios in comparison with CNN, VGG-19, and ResNet-50 architectures. The investigated training and validation accuracy/loss curves, confusion matrices, and key classification metrics demonstrate that spectrogram-based AlexNet architectures outperform other architectures trained on several feature extraction methods for chord categorization. Future studies might concentrate on expanding the dataset to foster model generalization and incorporating transformer-based designs for better feature extraction. Furthermore, incorporating the trained model with real-time applications for live chord detection and music analysis can increase its practical applicability. Future research will also concentrate on further enhancing computing efficiency in mobile and embedded devices.

REFERENCES

[1] T. Daikoku, M. Tanaka, S. Yamawaki, “Bodily maps of uncertainty and surprise in musical chord progression and the underlying emotional response,” iScience 27, 109498, 2024, doi: 10.1016/ j.isci.2024.109498.

[2] R. M. French, “Structure of the Guitar,” Technology of the Guitar, Springer, Boston, doi: 10.1007/978-1-4614-1921-1_3.

[3] S. Bilbao, R. Russo, "Real-Time Guitar Synthesis," Proceedings of the 27th International Conference on Digital Audio Effects (DAFx24), Guildford, United Kingdom, pp. 163-170, 2024.

[4] T. Groves, J. A. Kemp, "Applicability of the Capstan Equation to Guitar Strings," *Archives of Acoustics*, vol. *44, no.* 3, pp. 459–465, 2019.

[5] J. M. Hjerrild, S. Willemsen and M. G. Christensen, "Physical Models For Fast Estimation Of Guitar String, Fret And Plucking Position," *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2019, pp. 155-159, doi: 10.1109/WASPAA.2019.8937157.

[6] Y. Bando, M. Tanaka, "A Chord Recognition Method of Guitar Sound Using Its Constituent Tone Information," IEEJ Transactions on Electrical and Electronic Engineering, vol. 17, no. 1, 2021, doi: 10.1002/tee.23492.

[7] J. Mycka, J. Mańdziuk, "Artificial intelligence in music: recent trends and challenges," Neural Computing and Applications, vol .**37**, pp. 801–839, 2025, doi:10.1007/s00521-024-10555-x.

[8] M. B. Er, I. B. Aydilek, "Music Emotion Recognition by Using Chroma Spectrogram and Deep Visual Features," International Journal of Computational Intelligence Systems, vol. 12, pp. 1622–1634, 2019, doi: 10.2991/ijcis.d.191216.001.

[9] M. W. Lakdari, A. H. Ahmad, S. Sethi, G. A. Bohn, D. J. Clink, "Mel-frequency cepstral coefficients outperform embeddings from pre-trained convolutional neural networks under noisy conditions for discrimination tasks of individual gibbons," Ecological Informatics, vol. 80, 2024, doi: 10.1016/j.ecoinf.2023.102457.

[10] N. B. Korade, M. B. Salunke, A. A. Bhosle, G. G. Asalkar, D. M. Joshi, A. S. Patil, S. M. Sangve, "Tomato Leaf Disease Detection with YOLOV8 Leaf Extraction, Resnet-50 Classification, and Gpt-3.5 for Treatment Recommendations," International Research Journal of Multidisciplinary Scope (IRJMS), vol. 6, no.1 ,2025, doi: 10.47857/irjms.2025.v06i01.02864

[11] N. B. Korade, and M. Zuber, "Stock Price Forecasting using Convolutional Neural Networks and Optimization Techniques", International Journal of Advanced Computer Science and Applications, vol. 13, no. 11, pp. 378-385, 2022, doi: 10.14569/IJACSA.2022.0131142.

[12] S. Kamonsantiroj, L. Wannatrong, L. Pipanmaekaporn, "Chord Recognition in Music Using a Robust Pitch Class Profile (PCP) Feature and Support Vector Machines (SVM)," International Journal of Informatics and Information Systems, vol. 7, no. 1, pp. 01-07, 2024, doi: 10.47738/ijiis.v7i1.191.

[13] M. E. ElAlami, S. M. K. Tobar, S. M. Khater, Eman. A. Esmaeil, "Texture Feature and Mel-Spectrogram Analysis for Music Sound Classification," International Journal of Advanced Computer Science and Applications, vol. 15, no. 9, 2024, doi: 10.14569/IJACSA.2024.0150918.

[14] S. O. Folorunso, S. A. Afolabi, A. B. Owodeyi, "Dissecting the genre of Nigerian music with machine learning models," Journal of King Saud University – Computer and Information Sciences, vol. 34, no. 8, pp. 6266–6279, Sep. 2022. doi:10.1016/j.jksuci.2021.07.009.

[15] T. Li, "Optimizing the configuration of deep learning models for music genre classification," Heliyon, vol. 10, no. 2, Jan. 2024. doi:10.1016/j.heliyon.2024.e24892.

[16] T. Carsault, J. Nika, P. Esling, and . G. Assayag, "Combining Real-Time Extraction and Prediction of Musical Chord Progressions for Creative Applications," Electronics, vol. 10, no. 21, 2021, doi: 10.3390/electronics10212634.

[17] J. Chen, X. Ma, S. Li , S. Ma, Z. Zhang, and X. Ma, "A Hybrid Parallel Computing Architecture Based on CNN and Transformer for Music Genre Classification," electronics, vol.13, no. 16, 3313, 2024, doi:10.3390/electronics13163313.

[18] Z. Zheng, "The Classification of Music and Art Genres under the Visual Threshold of Deep Learning," omputational Intelligence and Neuroscience, Article 4439738, 2022, doi: 10.1155/2022/4439738.

[19] Y-H. Cheng, and C. N. Kuo, "Machine Learning for Music Genre Classification Using Visual Mel Spectrum," Mathematics, vol. 10, no. 23, 4427, 2022, doi:10.3390/math10234427.

[20] A. Yadav, S. Gaikwad, T. Kuigade, A. Patil, "Music Chord Prediction Using Machine Learning," International Research Journal of Modernization in Engineering Technology and Science, vol. 05, no. 12, 2023, doi: 10.56726/IRJMETS46945.

[21] X. Riley, D. Edwards and S. Dixon, "High Resolution Guitar Transcription Via Domain Adaptation," ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, pp. 1051-1055, doi: 10.1109/ICASSP48485.2024.10446182.

[22] S. Koelsch, P. Vuust, K. Friston, "Predictive Processes and the Peculiar Case of Music," Trends in Cognitive Sciences, vol. 23, no. 1, pp. 63-77, 2019, doi: 10.1016/j.tics.2018.10.006

[23] Y. S. Chen, C. -S. Hsu and F. -Y. C. Chien, "A Music Generation Scheme with Beat Weight Learning," *2023 International Conference on Smart Applications, Communications and Networking (SmartNets)*, Istanbul, Turkiye, 2023, pp. 1-6, doi: 10.1109/SmartNets58706.2023.10216030.

[24] M. S.N.V. Jitendra, Y. Radhika, "Singer Gender Classification using Feature-based and Spectrograms with Deep Convolutional Neural Network," International Journal of Advanced Computer Science and Applications(IJACSA), vol. 12, no. 2, 2021, doi: 10.14569/IJACSA.2021.0120218.

[25] R. Chen, A. Ghobakhlou, A. Narayanan, "Hierarchical Residual Attention Network for Musical Instrument Recognition Using Scaled Multi-Spectrogram," Applied Sciences, vol. 14, no. 23, 2024, doi:10.3390/app142310837.

[26] J. Liu, C. Wang, L. Zha, "A Middle-Level Learning Feature Interaction Method with Deep Learning for Multi-Feature Music Genre Classification," *Electronics, vol.* 10, no. 18, 2021, doi: 10.3390/electronics10182206.

[27] B. S. Hameed, C. S. Bhatt, B. Nagaraj, A. K. Suresh, "Chromatography as an Efficient Technique for the Separation of Diversified Nanoparticles," Nanomaterials in Chromatography, Current Trends in Chromatographic Research Technology and Techniques, pp. 503-518, 2018, doi: 10.1016/B978-0-12-812792-6.00019-4.

[28] S. Jagjeet, L. B. Saheer, and O. Faust, "Speech Emotion Recognition Using Attention Model," *International Journal of Environmental Research and Public Health* , vol. 20, no. 6, 2023, doi:10.3390/ijerph20065140.

[29] E. Yücesoy, "Gender Recognition Based on the Stacking of Different Acoustic Features," *Applied Sciences, vol.*14, no. 15, 2024, doi:10.3390/app14156564.

[30] M. Ashraf, F. Abid, I. U. Din, J. Rasheed, M. Yesiltepe, S. F. Yeo, M. T. Ersoy, " A Hybrid CNN and RNN Variant Model for Music Classification," *Applied Sciences*, vol. 13, no. 3: 1476. 2023, doi:10.3390/app13031476.

[31] N. B. Korade, and M. Zuber, "Boost Stock Forecasting Accuracy Using The Modified Firefly Algorithm And Multichannel Convolutional Neural Network", Journal of Theoretical and Applied Information Technology, vol. 101, no. 7, pp. 2668- 2677, 2023.

[32] N. B. Korade, M. B, Salunke, A. A. Bhosle, et. al., "Proactive Soybean Disease Detection through YOLO Leaf Extraction and ResNet-50 Classification to Reduce Crop Loss and Boost Productivity," International Journal of Engineering Trends and Technology, vol. 73, no. 1, pp. 385-396, 2025, doi: 10.14445/22315381/IJETT-V73I1P133.

[33] H. Eldem, E. Ülker, O. Y. Işıklı, "Alexnet architecture variations with transfer learning for classification of wound images," Engineering Science and Technology, an International Journal, vol. 45, 2023, doi: 10.1016/j.jestch.2023.101490.

[34] I. Singh, G. Goyal, A. Chandel, "AlexNet architecture based convolutional neural network for toxic comments classification," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 9, pp. 7547-7558, 2022, doi: 10.1016/j.jksuci.2022.06.007.

[35] N. B. Korade, M. B. Salunke, A. A. Bhosle, G. G. Asalkar, B. Lal, P. B. Kumbharkar, "Elevating intelligent voice assistant chatbots with natural language processing, and OpenAI technologies," Indonesian Journal of Electrical Engineering and Computer Science, vol. 37, no.1, pp. 507-517, 2025, doi: 10.11591/ijeecs.v37.i1.pp507-517.

[36] N. B. Korade, and M. Zuber, "Forecasting Stock Price Using Time-Series Analysis and Deep Learning Techniques," Data Engineering and Applications: Proceedings of the International Conference, IDEA 2K22, vol.1, 2024, DOI: 10.1007/978-981-97-0037-0_31.

[37] N. B. Korade, and M. Zuber, "Stock Forecasting Using Multichannel CNN and Firefly Algorithm", Proceedings of the 2nd International

Conference on Cognitive and Intelligent Computing, pp. 447-458, 2023, doi: 10.1007/978-981-99-2742-5_46.

[38] N. M R and S. Mohan B S, "Music Genre Classification using Spectrograms," *2020 International Conference on Power, Instrumentation, Control and Computing (PICC)*, Thrissur, India, 2020, pp. 1-5, doi: 10.1109/PICC51425.2020.9362364.

[39] M. K. Abbas, K. Gupta, M. Mudassir and R. Jain, "A Comprehensive Analysis of Music Genre Classification with Audio Spectrograms using Deep Learning Techniques," *2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, Greater Noida, India, 2023, pp. 1139-1147, doi: 10.1109/ICAC3N60023.2023.10541392.

[40] S. Pasrija, S. Sahu and S. Meena, "Audio Based Music Genre Classification using Convolutional Neural Networks Sequential Model," *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, Lonavla, India, 2023, pp. 1-5, doi: 10.1109/I2CT57861.2023.10126446.