DOI: https://doi.org/10.15276/aait.06.2023.28
UDC 004.8

Music emotion classification using a hybrid CNN-LSTM model

Vitaliy S. Yakovyna[1)]
ORCID: https://orcid.org/0000-0003-0133-8591; yakovyna@matman.uwm.edu.pl. Scopus Author ID: 6602569305
Valentyn V. Korniienko[2)]
ORCID: https://orcid.org/0009-0000-2581-9133; valik.kornienko97@gmail.com
[1)] University of Warmia and Mazury in Olsztyn, 2, Oczapowskiego Str. Olsztyn, 10-719, Poland
[2)] Lviv Polytechnic National University, 12, Bandera Str. Lviv, 79000, Ukraine

ABSTRACT

The emotional content of music, interwoven with the intricacies of human affect, poses a unique challenge for computational recognition and classification. With the digitalization of music libraries expanding exponentially, there is a pressing need for precise, automated tools capable of navigating and categorizing vast musical repositories based on emotional contexts. This study advances music emotion classification in the field of music information retrieval by developing a deep learning model that accurately predicts emotional categories in music. **The goal of this research** is to advance the field of music emotion classification by leveraging the capabilities of convolutional neural networks combined with long short-term memory within deep learning frameworks. The contribution of this study is to provide a refined approach to music emotion classification, combining the power of convolutional neural networks and long short-term memory architectures with sophisticated preprocessing of the Emotify dataset for a deeper and more accurate analysis of musical emotions. The research introduces a novel architecture combining Convolutional Neural Networks and Long Short-Term Memory networks designed to capture the intricate emotional nuances in music. The model leverages convolutional neural networks for robust feature detection and Long Short-Term Memory networks for effective sequence learning, addressing the temporal dynamics of musical features. Utilizing the Emotify dataset, comprising tracks annotated with nine emotional features, the study expands the dataset by segmenting each track into 20 parts, thereby enriching the variety of emotional expressions. Techniques like the synthetic minority oversampling technique were implemented to counter dataset imbalance, ensuring equitable representation of various emotions. The spectral characteristics of the samples were analyzed using the Fast Fourier Transform, contributing to a more comprehensive understanding of the data. Through meticulous fine-tuning, including dropout implementation to prevent overfitting and learning rate adjustments, the developed model achieved a notable accuracy of 94.7 %. This high level of precision underscores the model's potential for application in digital music services, recommendation systems, and music therapy. Future enhancements to this music emotion classification system include expanding the dataset and refining the model architecture for even more nuanced emotional analysis.

Keywords: Deep learning; music emotion classification; neural network; spectrum analysis; convolutional neural network

For citation: **Yakovyna V.S., Korniienko V. V. "Music emotion classification using a hybrid CNN-LSTM model".** ***Applied Aspects of Information Technology.*** **2023; Vol.6 No.4: 418–430. DOI: https://doi.org/10.15276/aait.06.2023.28**

INTRODUCTION

In the evolving landscape of Music information retrieval (MIR) [1, 2], the classification of musical emotions through automated systems is a burgeoning domain of research that has captivated scholars and technologists alike. The emotional content of music, interwoven with the intricacies of human affect, poses a unique challenge for computational recognition and classification. With the digitalization of music libraries expanding exponentially, there is a pressing need for precise, automated tools capable of navigating and categorizing vast musical repositories based on emotional contexts. Music emotional recognition (MER) is not only applicable to music track navigation, search, and recommendation but is also widely used in the field of music therapy [3].

Traditional methods of MER, which typically involve the manual curation of audio features fed into machine learning algorithms, have shown limitations in their ability to capture the high-dimensional nature of music data [4, 5].

Deep learning, with its remarkable success in fields such as computer vision and natural language processing, presents an innovative frontier for addressing the challenges of MER.

Leveraging the capabilities of deep learning, this study introduces a hybrid Convolution neural network-long short-term memory (CNN-LSTM) model designed to extract and process visual audio features, transforming the way emotional content is discerned in music tracks. By treating audio signals as visual spectrograms, the model captures a comprehensive representation of the temporal and frequency aspects inherent in musical compositions.

A one-dimensional (1D) Convolution neural network (CNN) component excels in identifying patterns within these visual features, while the Long short-term memory (LSTM) layer interprets the sequential flow of the music, an essential factor in understanding emotional progression.

LITERATURE OVERVIEW

Music has long been recognized as a potent conduit for expressing a spectrum of human emotions. The quest to decode the emotional content within musical compositions has garnered significant interest across various disciplines. As the landscape of technology evolves, especially with the advent of deep learning, the methodologies for recognizing and categorizing musical emotions are being revolutionized [7].

The emotional impact of music is diverse and profound, yet it remains a deeply personal experience, with individual responses to music varying widely.

Music often reflects our emotions, and different features in music help us figure out how a song might make us feel. For instance, a song with a quick beat might make us feel excited or happy, while one with a slower beat might make us feel calm or sad [8]. To understand these emotions better, researchers use different methods to study music. They look at things like the Mel frequency cepstral coefficients (MFCC) [9], which help analyze the sound's pitch, and the Zero crossing rate (ZCR) [10], which tells us about the rhythm and the pitch analysis to dig deeper into the music's characteristics.

In the last few years, the CNN architecture has proven itself not only for image analysis but also for sound classification. In a recent study [11], authors present how CNN can accurately classify sounds from spectrogram images achieved notable success. One of the key findings of this study is the effectiveness of the CNN model (Fig. 1) in classifying environmental sounds.

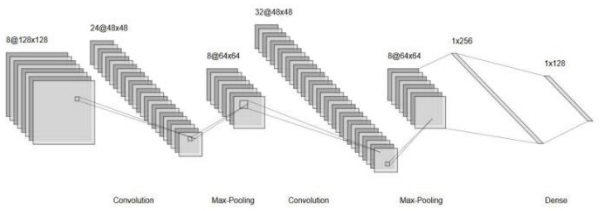

Fig. 1. **Convolutional neural network model structure**

Source: **compiled by the [11]**

The model achieved a classification accuracy of 77 % on the ESC-10 dataset, while the Tensor Deep Stacking Network model with which they were compared achieved only 56 % accuracy.

In the exploration of music emotion classification, recent advancements have centered around deep learning techniques that interpret music as a dynamic language of emotions, capable of expressing complex human feelings. Notably, a novel approach utilizing an Inception-GRU residual structure has been put forth, capturing the intricacies of musical expressions with significant efficacy. This methodology, grounded in the spectral matrix derived from logarithmic short-time Fourier transform, has showcased promising results on the Soundtrack dataset, achieving an accuracy surpassing traditional machine learning models [12]. In this paper, the researches presented an optimized structure of the Inception-V1 model which combines different convolution layers in parallel, and a deeper matrix is formed by concatenation the results processed by the convolution layers.

Simultaneously, advancements in music emotion classification have seen the integration of pitch frequency and band energy distribution features, reflecting the nuanced changes in a singer's emotional state through music. The innovation in this realm involves an enhanced Deep belief network (DBN) coupled with a support vector machine for classification, leading to a robust fusion classification algorithm. This improved DBN framework, by assimilating distinctive musical features, has demonstrated a considerable improvement in classifying music emotions, indicating a significant leap forward in the field [13].

Deep belief network is a typical deep learning model that can learn the corresponding input and obtain more abstract and higher-level features [14]. The structure of the typical model is shown in Fig. 2.

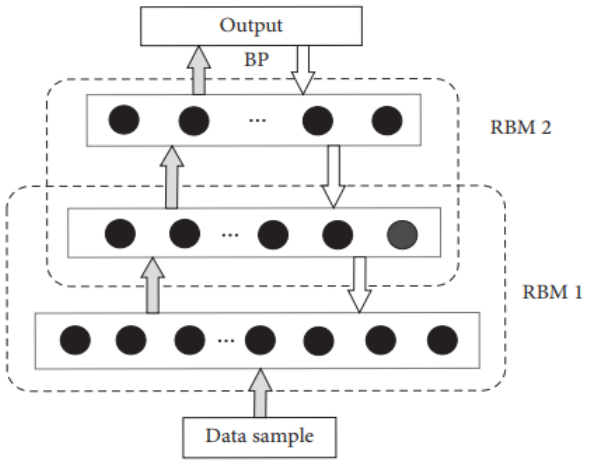

Fig. 2. **Typical deep belief network model structure**

Source: **compiled by the [14]**

The authors take a typical DBN model, which includes forward propagation and backpropagation processes and is improved by composed of an n-layer improved Restricted Boltzmann machine (RBM) model, a one-layer of traditional RBM model, and a Softmax layer (Fig. 3).

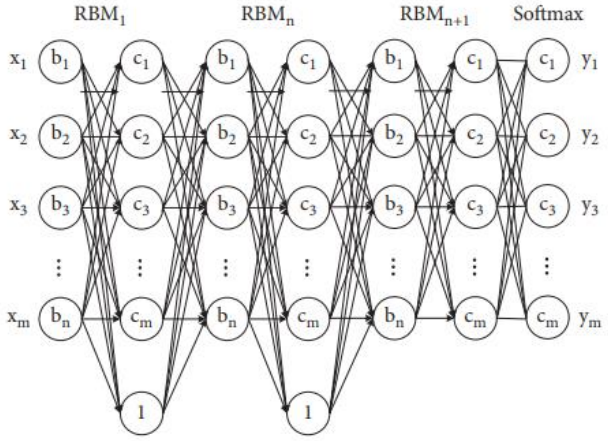

Fig. 3. **Improved deep belief network model structure**

Source: **compiled by the [14]**

After that, they integrated the Support Vector Machine (SVM) classification algorithm [15], which uses a small number of support vectors and thus can better represent the classification information of the whole training sample set to participate in training with the DBN network.

THE AIM AND OBJECTIVES OF THE RESEARCH

The primary goal of this research is to advance the field of music emotion classification by leveraging the capabilities of CNNs combined with LSTM within deep learning frameworks. To achieve this, the research is focused on optimizing the CNN architectures coupled with LSTM specifically for the intricate task of decoding emotional cues in music.

To develop deep learning technology, it is necessary to solve the following tasks:

– create a CNN-LSTM model for efficient and accurate classification of a wide range of emotions in music, leveraging the strengths of CNN and LSTM architectures to process complex audio data;

– employing the Emotify dataset for model training and testing, with a focus on preprocessing techniques to ensure data quality and relevance;

– train the model to achieve high accuracy with optimized computational resources;

– test the model across various musical genres to demonstrate its effectiveness and adaptability.

The contribution of this study is to provide a refined approach to music emotion classification, combining the power of CNN-LSTM architectures with sophisticated preprocessing of the Emotify dataset for a deeper and more accurate analysis of musical emotions.

METHODS OF AUDIO ANALYSIS

Audio analysis is a crucial component in the process of music emotion classification, where the objective is to extract meaningful information from raw audio that correlates with human emotional states. To achieve this, visualization of audio data is often employed, which not only aids in understanding the characteristics of the sound but also serves as a pre-processing step for further analysis and feature extraction.

Sound Wave Visualization

The sound wave or waveform display is a fundamental method of visual representation, illustrating the variations in air pressure or the audio signal amplitude over time. This visualization can reveal the temporal structure of a sound, including its rhythm, pauses, and energy fluctuations, which can be indicative of different emotions in music. The waveform provides an intuitive understanding of the loudness and dynamics of the audio track.

Fast Fourier Transform

Fast Fourier transform (FFT) is a popular algorithm in the signal processing field. Rather than the information, which we can gather from time-domain analysis, FFT supplies frequency or spectral-based information about the audio signals. Fast Fourier transform implies that any continuous signal can be expressed in terms of the sum of delicately chosen sinusoidal waves with appropriate frequency, amplitude, and phase [16].

Spectrum Analysis

Spectrum analysis (Fig. 4) transforms the audio signal from the time domain to the frequency domain using FFT. The resulting spectral plot shows the distribution of power across various frequency components. This analysis can uncover the harmonic content and the balance between different frequency ranges, which are essential attributes related to the perceived 'color' or 'texture' of the sound, often associated with specific emotional qualities.

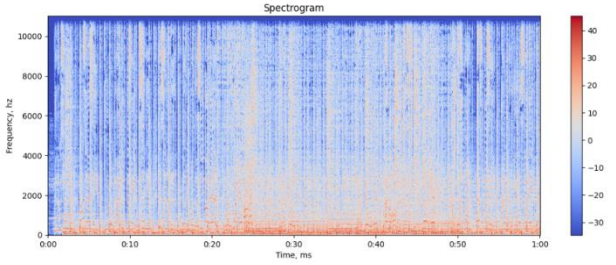

Fig. 4. **Spectrogram of audio**

Source: **compiled by the authors**

Mel Spectrogram

A Mel spectrogram (Fig. 5) is a more sophisticated visual representation that combines the concepts of spectrum analysis over time and applies a Mel scale to the frequency axis. The Mel scale is designed to mimic the human ear's response to different pitches, making it highly relevant for audio analysis in music emotion recognition. The Mel spectrogram provides a time-varying visual depiction of the sound's spectral content, highlighting the changes in energy across the Mel frequency bands over time. This can be particularly revealing of the timbral and textural shifts that accompany different emotional expressions in music.

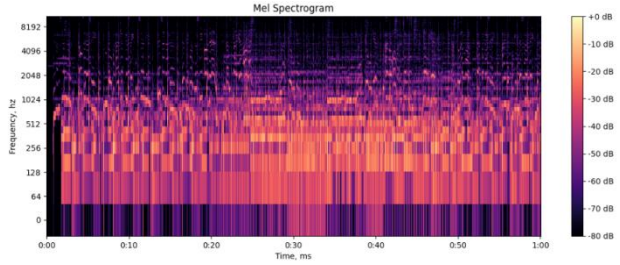

Fig. 5. **Mel spectrogram of audio**
Source: **compiled by the authors**

Mel-Frequency Cepstral Coefficients

Mel-Frequency Cepstral Coefficients (MFCCs) (Fig. 6) are derived from the Mel spectrogram and represent the power spectrum of a sound using a small number of features, which approximate the audio signal's overall shape. They are widely used in audio recognition tasks as they capture the key aspects of the Mel spectrogram that are perceptually important to humans. By focusing on these features, MFCCs provide a compact and informative representation that can be used for machine learning models to classify emotional content in music.

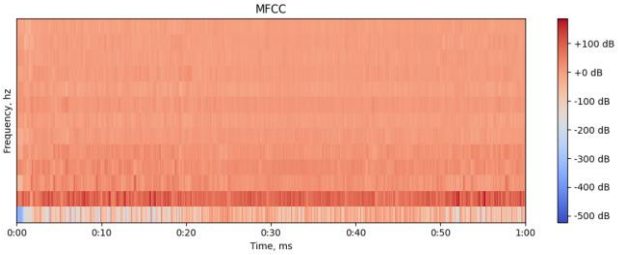

Fig. 6. **Mel-frequency cepstral coefficient of audio**
Source: **compiled by the authors**

DATASET DESCRIPTION

The Emotify [17] dataset consists of 400 song excerpts (1 minute long) in 4 genres (rock, classical, pop and electronic). The annotations were collected using the GEMS scale (Geneva Emotional Music Scales) [18]. Each participant could select maximally three items from the scale (the emotions that he felt strongly listening to this song. Below (Table 1) is the description of the emotional categories.

The annotations produced are spread unevenly among the songs, which is caused both by the design of the experiment and the design of the game. Participants could skip songs and switch between genres, and they were encouraged to do so because induced emotional response does not automatically occur on every music-listening occasion. Therefore, less popular (among our particular sample of participants) genres received fewer annotations, and the same happened to less popular songs.

Each line in the file corresponds to one participant (i.e., annotations are not averaged per song).

This is the description of information found in the file:

– id of the music file;
– genre of the music file;
– 9 annotations by the participant (whether emotion was strongly felt for this song or not). 1 means emotion was felt;
– participant's mood prior to playing the game;
– liking (1 if the participant decided to report he liked the song);
– disliking (1 if participant decided to report he disliked the song);
– age, gender and mother tongue of the participant (self-reported).

DATASET PREPROCESSING

The dataset initially contained diverse emotional feature values for each music file, as rated by different individuals. To create a unified representation, these values were averaged for each track, resulting in a consolidated emotional feature set per song.

The division of music tracks into segments was due to the relatively modest size of the data set of 400 tracks. By dividing each minute track into 10 and 20 shorter segments of approximately 6 and 3 seconds each, the size of the dataset was increased by a factor of 10 and 20, so that we ended up with 4000 and 8000 sample songs as two different datasets instead of 400.

Following the segmentation, using the python library librosa were produced feature extraction that included the chroma-stft, rms, spectral centroid, spectral bandwidth, spectral rolloff, zero-crossing rate, harmonic content, tempo, and Mel-frequency cepstral coefficients (MFCCs).

Table 1. **Description of the emotional categories**

Emotional category	Explanation
Amazement	Feeling of wonder and happiness
Solemnity	Feeling of transcendence, inspiration. Thrills
Tenderness	Sensuality, affect, feeling of love
Nostalgia	Dreamy, melancholic, sentimental feelings
Calmness	Relaxation, serenity, meditativeness
Power	Feeling strong, heroic, triumphant, energetic
Joyful activation	Feels like dancing, bouncy feeling, animated, amused
Tension	Nervous, impatient, irritated
Sadness	Depressed, sorrowful

Source: **compiled by the authors**

All features, except of MFCCs, represent temporal characteristics of sound. These features, therefore, are suitable for LSTM layers, as they can track and analyze the evolution of these temporal characteristics over time, crucial for understanding the structure and progression of music.

Mel-frequency cepstral coefficients capture the timbral characteristics of audio signals, representing the texture and quality of sounds. They encapsulate complex relationships between different frequencies, which make them well-suited for CNNs.

All these features encapsulate the core elements that convey emotion in music, such as rhythm, pitch, and timbre, and are instrumental for the subsequent machine learning tasks.

The dataset's class distribution was observed to be imbalanced (Fig 7).

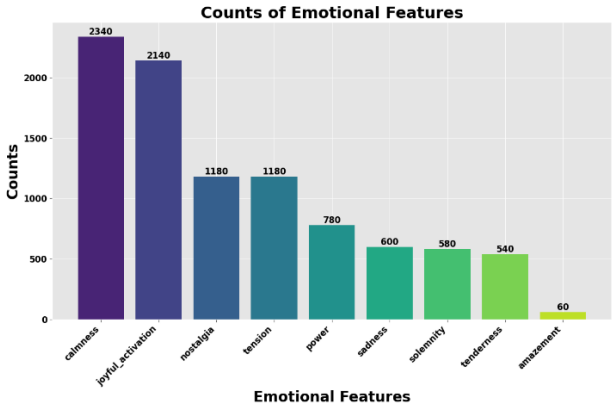

Fig. 7. **Count of imbalanced emotional features**
Source: **compiled by the authors**

To rectify this imbalance, was implemented Synthetic minority oversampling technique (SMOTE) [19]. SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. After applying SMOTE method, the amount of data in the classes was equalized to 2320.

This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.

This method not only augments the underrepresented classes by generating new, synthetic samples but does so in a manner that respects the underlying distribution of each class, thus preserving the authenticity of the dataset.

Once the features were extracted and the class balance addressed, were applied normalization, label encoding and dataset splitting to train, test and validation parts of the feature set (70%, 15%, 15%).

MODEL ARCHITECTURE

In the development of a music emotion classification model, a combined CNN-LSTM architecture (Fig. 8) was utilized to exploit both the spatial and temporal characteristics of the audio features.

The input to the CNN block consists of a multi-dimensional array where each dimension represents different extracted audio features, such as Chroma-STFT, RMS, Spectral Centroid, Spectral Bandwidth, Spectral Rolloff, Zero-Crossing Rate, Harmonic Content, Tempo, and MFCCs. These features are structured to form a consistent input shape suitable for convolutional processing, often resembling a time-frequency representation of the audio signal.

The model begins with a three Conv1D layers, each followed by batch normalization, max pooling, and dropout, which work together to extract and refine feature representations from the input data. These convolutional layers progressively increase in depth, starting from 128 filters and expanding to 512, allowing the network to learn a hierarchy of features with increasing complexity.

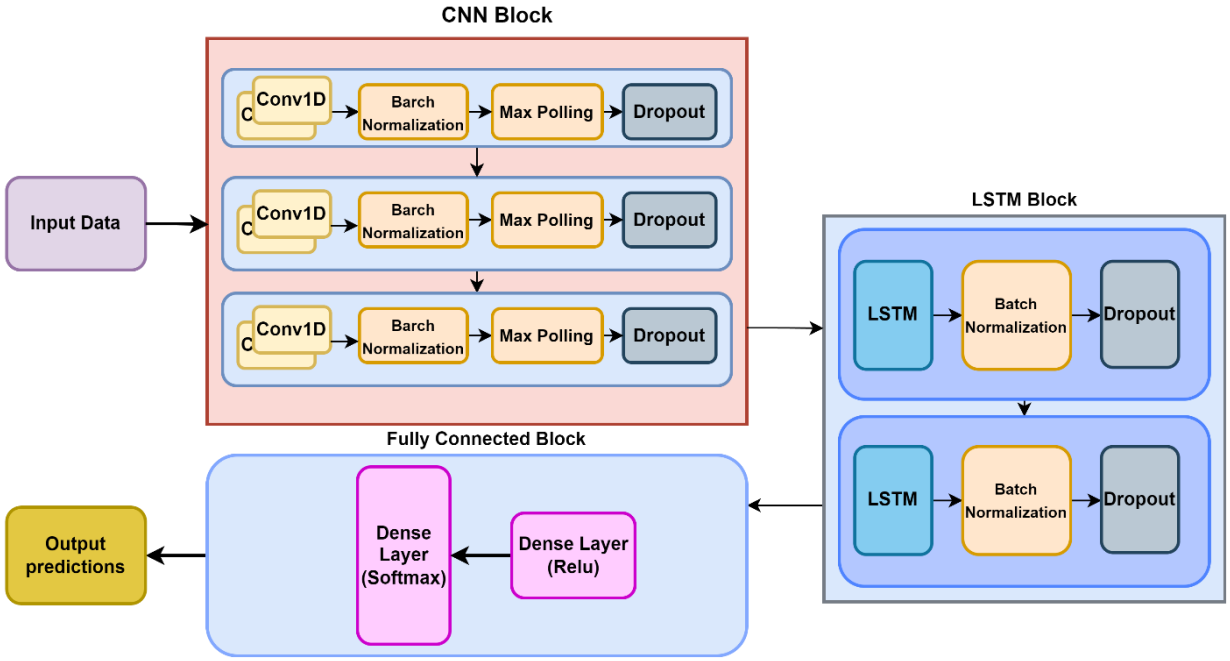

Fig. 8. **Convolutional neural network-Long short-term memory model architecture**
Source: **compiled by the authors**

Batch Norm is a normalization technique done between the layers of a Neural Network instead of in the raw data [20]. It is done along mini batches instead of the full data set.

It serves to speed up training and use higher learning rates, making learning easier. the normalization formula of Batch Norm is as follows:

$$z^N = \left(\frac{z - m_z}{s_z}\right), \quad (1)$$

where m_z is the mean of the neurons' output and s_z is the standard deviation of the neurons' output.

Max pooling significantly reduces the dimensionality of the data, ensuring that the most important features are preserved. The max pooling means moving the window along the matrix with data [21]. From the pixels falling into its field of view, the maximum is selected and moved to the resulting matrix.

To prevent overfitting was also used dropout. This method randomly disables a subset of neurons during training, forcing the network to learn more diverse features and, therefore, improving generalization [22].

The Rectified Linear Unit (ReLu) was chosen as the activation function for CNN layers. ReLU introduces non-linearity into the network, allowing it to learn complex patterns in the data [23]. Its simplicity and efficiency in computation make it a popular choice in deep learning architectures.

Formula for ReLu activation function:

$$f(x) = argmax(0, x). \quad (2)$$

Or it can be written as:

$$< ReLU'(x) = \begin{cases} 0, for\ x < 0 \\ 1, for\ x \geq 0 \end{cases}. \quad (3)$$

After the convolutional layers, two Long Short-Term Memory (LSTM) layers were incorporated to capture the temporal dynamics inherent in music tracks. LSTMs are particularly advantageous for this application due to their capacity to remember information over extended periods, making them suitable for sequence prediction problems such as time-series analysis found in music tracks.

The output from the LSTM layers is then fed into dense layers with ReLU and Softmax activations. The ReLU layer serves as a fully connected layer that introduces non-linearity and aids in learning complex patterns, while the Softmax layer maps the final output to a probability distribution over the predicted classes. Batch normalization and dropout are consistently used throughout the model to ensure generalization and prevent overfitting. It leveraged the features distilled by the previous layers to perform the emotion classification task, outputting a probability distribution across the predefined emotion labels.

The summary representation of the network presented at Table 3.

The total number of total parameters is 2507466, where trainable parameters is 2505034, and non-trainable – 2432.

Table 3. **Summary representation of the network**

Layer (type)	Output Shape	Param
Conv1d_170	(None, 28, 128)	640
Conv1d_171	(None, 28, 128)	65664
Batch_normalization_158	(None, 28, 128)	512
Max_pooling1d_86	(None, 14, 128)	0
Dropout_158	(None, 14, 128)	0
Conv1d_172	(None, 14, 256)	131328
Conv1d_173	(None, 14, 256)	262400
Batch_normalization_159	(None, 14, 256)	1024
Max_pooling1d_87	(None, 7, 256)	0
Dropout_159	(None, 7, 256)	0
Conv1d_174	(None, 7, 512)	524800
Vonv1d_175	(None, 7, 512)	104908
Batch_normalization_160	(None, 7, 512)	2048
Max_pooling1d_88	(None, 3, 512)	0
Dropout_160	(None, 3, 512)	0
Lstm_50	(None, 3, 128)	328192
Batch_normalization_161	(None, 3, 128)	512
Dropout_161	(None, 3, 128)	0
lstm_51	(None, 128)	131584
Batch_normalization_162	(None, 128)	512
Dropout_162	(None, 128)	0
Dense_46	(None, 64)	8256
Dropout_163	(None, 64)	0
Batch_normalization_163	(None, 64)	256
Dense_47	(None, 10)	650

Source: **compiled by the authors**

The final flow of the proposed method for music emotion classification is illustrated in Fig. 9.

The architecture represents a multi-stage process starting with a raw audio file. This file is first segmented into 20 3-second segments, allowing for more detailed and focused feature extraction. Features are then extracted from these segments, capturing both the spectral and temporal characteristics inherent to the audio. After scaling and normalization, these features are then fed into the CNN-LSTM model, a hybrid neural network., that employs convolutional layers to detect patterns and structures within the features and LSTM layers to understand the temporal progression of these patterns. The output from this model gives a prediction of the music's emotion.

EXPERIMENT AND RESULTS

The experiment was performed on the two types of preprocessed Emotify dataset. In the first type, the original audio files were divided into 10 segments, resulting in a data set of 4,000 audio files of 6 seconds each. In the second type – each audio file was divided into 20 segments; the result dataset is 8,000 audio files of 3 seconds each. Each track within the dataset was distinctly characterized by a prominent emotional label, as perceived by human listeners and devoid of lyrical content to ensure the focus remained on the music's instrumental and timbral properties. The audio files were maintained in a stereo mp3 format with a 44.1 kHz sampling rate.

The optimization method used was RMSprop. RMSProp is an unpublished adaptive learning rate optimizer proposed by Geoff Hinton [24, 25]. The motivation is that the magnitude of gradients can differ for different weights and can change during learning, making it hard to choose a single global learning rate. RMSProp tackles this by keeping a moving average of the squared gradient and adjusting the weight updates by this magnitude. The gradient updates are performed as:

$$E[g^2]_t = \gamma E[g^2]_{t-1} + (1-\gamma) g_t^2, \quad (4)$$

$$\theta_{t+1} = \theta_t - \frac{n}{\sqrt{E[g^2]_t + \epsilon}} g_t, \quad (5)$$

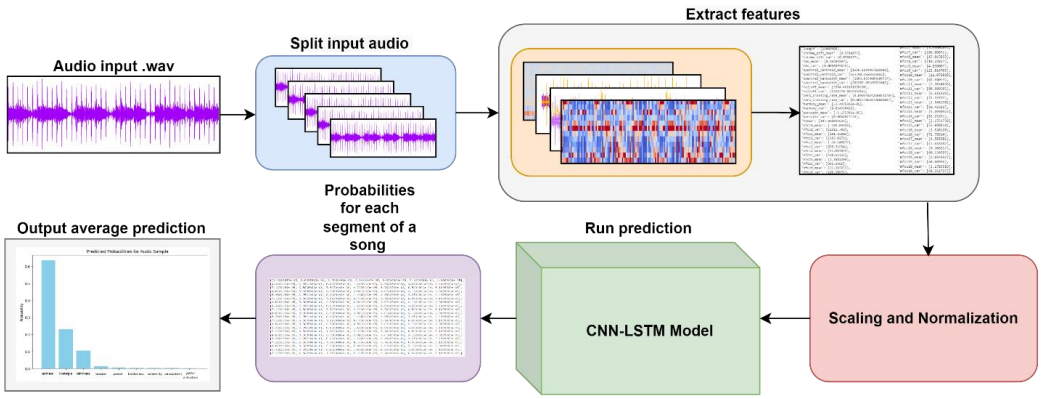

Fig. 9. **Full flow of music emotion classification based on the Convolution Neural Network – Long Short-Term Memory model**

Source: **compiled by the authors**

where $E[g]$ is the moving average of squared gradients, n is the learning rate. Hinton suggests γ=0.9, with a good default for n as 0.001.

RMSprop involved dynamic adjustments to the learning rate, implemented through the ReduceLROnPlateau callback [26], which methodically reduced the learning rate once learning stagnated, thereby enhancing the convergence process. The code was written in Python using the Librosa library [27], and the training model used the Keras library [28], also written in Python.

The proposed neural network was compared with recently proposed models, as represented in Table 4, including two types of preprocessed datasets. The inception-GRU Residual Structure method [12] achieved 84.23 % accuracy in music emotion classification on the Soundtrack dataset. Improved deep belief network [13] achieved 83.35% accuracy. The proposed method in this study achieved 74.8 % accuracy on the 10-segment dataset and 94.7 % – on the 20-segment dataset. The experimental results reflected that the proposed network model achieved a higher accuracy.

Table 4. **Comparison of accuracy of different models**

Method	Accuracy
Inception-GRU Residual Structure (Soundtrack dataset)	84.23 %
Improved deep belief network (FMA)	83.35 %
Proposed CNN-LSTM network (10 segments Emotify dataset)	74.8 %
Proposed CNN-LSTM network (20 segments Emotify dataset)	**94.7 %**

Source: **compiled by the authors**

Training and validation accuracy and loss are shown in Fig. 10 and Fig. 11, respectively.

As a loss function, sparse categorical cross-entropy was used. Variation of the categorical cross-entropy loss used for multi-class classification tasks where the classes are encoded as integers rather than one-hot encoded vectors. Given that the true labels are provided as integers, we directly select the correct class using the provided label index instead of summing over all possible classes. Thus, the loss for each example is calculated as

$$H(y, \hat{y}) = -\log(\hat{y}_i, y_i). \qquad (6)$$

Fig. 10. **Training and validation accuracy**
Source: **compiled by the authors**

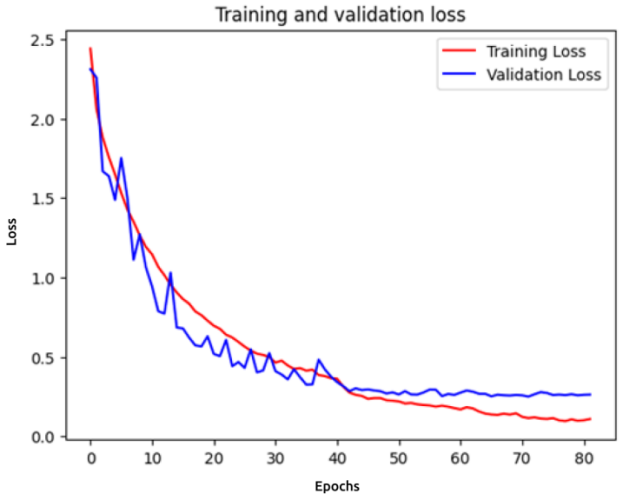

Fig. 11. **Training and validation loss**
Source: **compiled by the authors**

And the final sparse categorical cross-entropy loss is the average over all the samples:

$$H(Y, \hat{Y}) = -\frac{1}{n}\sum_{i=1}^{n} \log(\hat{y}_i, y_i), \qquad (7)$$

here y_i is the true class of the *i*-th sample and $\hat{y}_i, y_i$ is the predicted probability of the *i*-th sample for the correct class y_i.

Table 5 shows the results of precision, recall and F-1 [30] score with comparison between 10 segments and 20 segments datasets.

The confusion matrix is presented in Fig. 12. The confusion matrix is a tool for visualizing the performance of a classification algorithm on a data set for which the true values are known [29]. It helps to understand which classes the algorithm mixes with each other.

Table. 5 **Results of precision, recall and F-1 score for two types of dataset**

Class	10 segments/6 sec each			20 segments/3 sec each		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Amazement	0.80	0.67	0.73	1.00	1.00	1.00
Calmness	0.76	0.77	0.77	0.89	0.84	0.86
Joyful activation	0.78	0.82	0.80	0.95	0.90	0.92
Nostalgia	0.73	0.70	0.71	0.92	0.89	0.91
Power	0.82	0.74	0.78	0.95	0.98	0.96
Sadness	0.59	0.61	0.60	0.92	0.99	0.96
Solemnity	0.72	0.72	0.72	0.98	1.00	0.99
Tenderness	0.67	0.59	0.62	0.98	1.00	0.99
Tension	0.71	0.71	0.71	0.94	0.95	0.95

Source: **compiled by the authors**

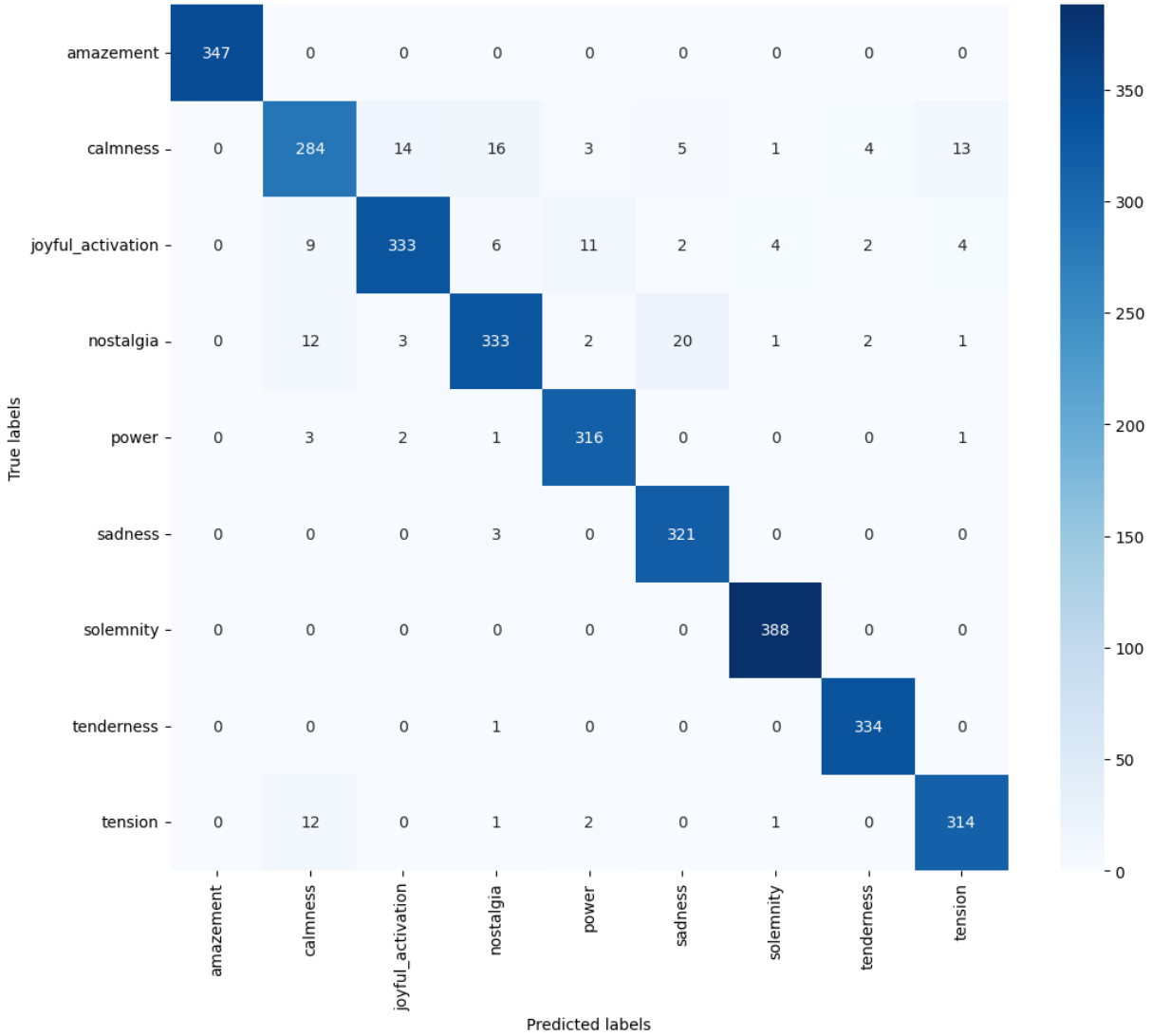

Fig. 12. **Confusion matrix**

Source: **compiled by the authors**

The main components of the confusion matrix:

– True Positive (TP): Number of positive samples correctly classified as positive.

– False Positive (FP): Number of negative samples incorrectly classified as positive.

– True Negative (TN): Number of negative samples correctly classified as negative.

– False Negative (FN): Number of positive samples incorrectly classified as negative.

CONCLUSIONS

This study represents a significant milestone in Music information retrieval (MIR), particularly in the classification of emotional content in music. By developing a CNN-LSTM network and utilizing the Emotify dataset, the research introduced innovative methodologies in data processing and model architecture. The segmentation of original music tracks into 4000 six-second clips and 8000 three-second clips was a key innovation, augmenting the data's diversity and offering the network a richer learning environment.

The architecture, combining convolutional layers and LSTM units, was adept at capturing both the subtle features and the temporal dynamics of musical emotions. This synergy resulted in the model achieving an exceptional accuracy rate of 94.7 % on the three-second segments. This performance not only demonstrates the effectiveness of the segmentation approach but also positions the CNN-LSTM framework as a leading contender in the domain of music emotion classification.

Compared to other recent models, such as the Inception-GRU Residual Structure model and the Improved deep belief network model, this study demonstrated higher accuracy, which emphasizes its reliability and effectiveness. The first comparable model scored an accuracy of 84.23 %, and the second model scored 83.35 %. This can be attributed in part to the significant preprocessing of the music data, which involved intricate feature extraction and careful consideration of the emotional attributes within the music.

The advances made in this research offer practical applications in the development of more nuanced and emotionally intelligent music recommendation systems and therapeutic interventions. The potential for enhancing user interaction through emotionally responsive technologies is enormous, opening new avenues for MIR research.

In summary, paves the way for future advancements in this field. The combination of detailed data preprocessing, innovative model architecture, and impressive classification accuracy highlights the potential of deep learning techniques to understand and interpret the rich emotional tapestry of music.

REFERENCES

1. He, N. & Ferguson, S. “Multi-view neural networks for raw audio-based music emotion recognition”. *IEEE International Symposium on Multimedia (ISM)*. 2020. p. 168–172, https://www.scopus.com/authid/detail.uri?authorId=57224619594.
DOI: https://doi.org/10.1109/ISM.2020.00037.

2. Jeon, B., Kim, C., Kim, A., Kim, D., Park, J. & Ha, J. “Music Emotion Recognition via End-to-End Multimodal Neural Networks”. *RecSys ’17 Poster Proceedings*. 2017, https://www.scopus.com/authid/detail.uri?authorId=57216435446.
DOI: https://doi.org/10.1109/JSTSP.2017.2764438.

3. Sanyal, S., Banerjee, A., Sengupta, R. & Ghosh, D. “Chaotic Brain, Musical Mind-A Non-Linear neurocognitive”. *Physics Based Study. J. Neurol. Neurosci*. 2016; 7: 1–10, https://www.scopus.com/authid/detail.uri?authorId=57213850454. DOI: https://doi.org/10.21767/2171-6625.100063.

4. Gharavian, D., Bejani M. & Sheikhan M. “Audio-visual emotion recognition using FCBF feature selection method and particle swarm optimization for fuzzy ARTMAP neural networks”. *Multimedia Tools and Applications*. 2017; 76: 2331–2352, https://www.scopus.com/authid/detail.uri?authorId=24075697400. DOI: https://doi.org/10.1007/s11042-015-3180-6.

5. Han, B. J., Rho, S., Jun, S. & Hwang, E. “Music emotion classification and context-based music recommendation”. *Multimedia Tools and Applications*. 2010; 47: 433–460, https://www.scopus.com/authid/detail.uri?authorId=23094368600. DOI: https://doi.org/10.1007/s11042-009-0332-6.

6. Aljanaki A., Wiering F. & Veltkamp R. C. “Studying emotion induced by music through a crowdsourcing game”. *Information Processing & Management*, 2015; 52: 115–128,

https://www.scopus.com/authid/detail.uri?authorId=56410808700.
DOI: https://doi.org/10.1016/j.ipm.2015.03.004.

7. Han D., Kong Y., Han J. & Wang G. “A survey of music emotion recognition”. *Frontiers of Computer Science*. 2022; 16, https://www.scopus.com/authid/detail.uri?authorId=55642123800. DOI: https://doi.org/10.1007/s11704-021-0569-4.

8. Gabrielsson, A. & Lindström, E. “The influence of musical structure on emotional expression”. *Music and emotion: Theory and research*. 2001. p. 223–248, https://www.scopus.com/authid/detail.uri?authorId=6603700763.

9. Huang, Z., Xue, W., Mao, Q. & Zhan, Y. “Unsupervised domain adaptation for speech emotion recognition using PCANet”. *Multimedia Tools and Applications*. 2017; 76: 6785–6799, https://www.scopus.com/authid/detail.uri?authorId=55796259300. DOI: https://doi.org/10.1007/s11042-016-3354-x.

10. Lin, Y. C., Yang, Y. H. & Chen, H. H. “Exploiting online music tags for music emotion classification”. *ACM Transactions on Multimedia Computing, Communications and Applications*. 2011; 7: 26, https://www.scopus.com/authid/detail.uri?authorId=57203772039. DOI: https://doi.org/10.1145/2037676.2037683.

11. Khamparia A., Gupta D., Nguyen N. G., Khanna A., Pandey B. & Tiwari P. “Sound classification using convolution neural network and tensor deep stacking network”. *IEEE Access*. 2019; 7: 7717–7727, https://www.scopus.com/authid/detail.uri?authorId=55811315600.
DOI: https://doi.org/10.1109/ACCESS.2018.2888882.

12. Han X., Chen F. & Ban J. “Music emotion recognition based on a neural network with an Inception-GRU residual structure”. *Electronics (Switzerland)*. 2023; 12: 978, https://www.scopus.com/authid/detail.uri?authorId=58116879700.
DOI: https://doi.org/10.3390/electronics12040978.

13. Tong G. “Music emotion classification method using improved deep belief network”. *Mobile Information Systems*. 2022, https://www.scopus.com/authid/detail.uri?authorId=57463836200. DOI: https://doi.org/10.1155/2022/2715765.

14. Liu, J., Wu, N. Q., Qiao, Y. & Li, Z. “Short-Term traffic flow forecasting using ensemble approach based on deep Belief networks”. *IEEE Transactions on Intelligent Transportation Systems*, 2022, 23 (1): 404–417, https://www.scopus.com/authid/detail.uri?authorId=57221597704.
DOI: https://doi.org/10.1109/TITS.2020.3011700.

15. Sun, S., Xie, X. & Dong, C. “Multiview learning with generalized eigenvalue proximal support vector machines”. *IEEE Transactions on Cybernetics*, 2019, 49 (2): 688–697, https://www.scopus.com/authid/detail.uri?authorId=8892785100.
DOI: https://doi.org/10.1109/TCYB.2017.2786719.

16. Hladnik, A., Muck, T., Stanic, M. & Cernic, M., “Fast Fourier transform in papermaking and printing: Two application examples”. *Acta Polytechnica Hungarica*. 2012, 9 (5): 155–166, https://www.scopus.com/authid/detail.uri?authorId=6602421874.

17. Aljanaki, A., Wiering, F. & Veltkamp, R. C. “Studying emotion induced by music through a crowdsourcing game”. *Information Processing and Management*. 2015, 52: 115–128, https://www.scopus.com/authid/detail.uri?authorId=56410808700.
DOI: https://doi.org/10.1016/j.ipm.2015.03.004.

18. Zentner, M., Grandjean, D., & Scherer, K. R.. “Emotions evoked by the sound of music: Characterization, classification, and measurement”. *Emotion*. 2008; 8 (4): 494–521, https://www.scopus.com/authid/detail.uri?authorId=7004225336.
DOI: https://doi.org/10.1037/1528-3542.8.4.494.

19. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. “SMOTE: Synthetic minority over-sampling technique”. *Journal of Artificial Intelligence Research*. 2002, 16: 321–357, https://www.scopus.com/authid/detail.uri?authorId=35077581400. DOI: https://doi.org/10.1613/jair.953.

20. Ioffe, S. & Szegedy, C. “Batch normalization: accelerating deep network training by reducing internal covariate shift”. *32nd International Conference on Machine Learning, ICML*. 2015; 1: 448–456, https://www.scopus.com/authid/detail.uri?authorId=7005880161.

21. Wu, H. & Gu, X. “Max-pooling dropout for regularization of convolutional neural networks neural information processing”. *Lecture Notes in Computer Science*. 2015; 9489; 46–54,

https://www.scopus.com/authid/detail.uri?authorId=56784559400. DOI: https://doi.org/10.1007/978-3-319-26532-2_6.

22. Goodfellow, I., Bengio, Y. & Courville, A. “DeepLearning”. *MIT Press*. 2016. p. 802. – Available from: http://www.deeplearningbook.org. – [Accessed: June 2023].

23. Agarap, A. F. “Deep learning using rectified linear units (ReLU)”. *arXiv*. 2019. – Available from: https://arxiv.org/abs/1803.08375. – [Accessed: June 2023].

24. Ruder, S. “An overview of gradient descent optimization algorithms”. *arXiv*. 2017. – Available from: https://arxiv.org/abs/1609.04747. – [Accessed: June 2023].

25. Graves A. “Generating sequences with recurrent neural networks”. *arXiv*. 2014. – Available from: https://arxiv.org/abs/1308.0850. – [Accessed: June 2023].

26. PyTorch: ReduceLROnPlateau – PyTorch 1.9.0 documentation. – Available from: https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html. – [Accessed: June 2023].

27. McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E. & Nieto, O. “Librosa: Audio and music signal analysis in Python”. *14th Python in Science Conference*, Austin: USA. 2015. p. 18–25, https://www.scopus.com/authid/detail.uri?authorId=34875379700. DOI: https://doi.org/10.25080/Majora-7b98e3ed-003.

28. Keras-team/keras. – Available from: https://github.com/keras-team/keras. – [Accessed: June 2023].

29. Stehman, S. V. “Selecting and interpreting measures of thematic classification accuracy”. *Remote Sensing of Environment*. 1997; 62 (1): 77–89, https://www.scopus.com/authid/detail.uri?authorId=7006807562. DOI: https://doi.org/10.1016/S0034-4257(97)00083-7.

30. Powers, D. “Evaluation: From precision, recall and F-Factor to ROC”. *Informedness, Markedness & Correlation*. 2008; 2. – Available from: https://doi.org/10.48550/arXiv.2010.16061. – [Accessed: June 2023].

Conflicts of Interest: the authors declare no conflict of interest

Received 11.09.2023
Received after revision 08.12.2023
Accepted 14.12.2023

DOI: https://doi.org/10.15276/aait.06.2023.28
УДК 004.8

Класифікація музичних емоцій за допомогою гібридної CNN-LSTM моделі

Яковина Віталій Степанович[1)]
ORCID: https://orcid.org/0000-0003-0133-8591; yakovyna@matman.uwm.edu.pl. Scopus Author ID: 6602569305
Корнієнко Валентин Валерійович[2)]
ORCID: https://orcid.org/0009-0000-2581-9133; valik.kornienko97@gmail.com
[1)] Вармінько-Мазурський університет в Ольштині, вул. Очаповського, 2. Ольштин, 10-719, Польща
[2)] Національний університет «Львівська політехніка», вул. С. Бандери, 12. Львів, 79000,Україна

АНОТАЦІЯ

Емоційний зміст музики, переплетений із тонкощами впливу на людину, створює унікальний виклик для систем комп’ютерного розпізнавання та класифікації. Оскільки оцифрування музичних бібліотек експоненціально розширюється, існує нагальна потреба в точних автоматизованих інструментах, здатних навігації та класифікації величезних музичних сховищ на основі емоційного контексту. Це дослідження покращує класифікацію музичних емоцій у сфері пошуку музичної інформації шляхом розробки моделі глибокого навчання, яка точно передбачає емоційні категорії в музиці. Метою цього дослідження є розвиток класифікації музичних емоцій шляхом використання можливостей згорткових нейронних мереж у поєднанні з довготривалою короткочасною пам’яттю в рамках глибокого навчання. Внесок цього дослідження полягає в тому, щоб забезпечити вдосконалений підхід до класифікації музичних емоцій, поєднуючи потужність згорткових нейронних мереж і архітектур довготривалої короткочасної пам’яті зі складною попередньою обробкою набору даних

Emotify для глибшого та точнішого аналізу музичних емоцій. Дослідження представляє нову архітектуру, що поєднує згорткові нейронні мережі та мережі довготривалої короткочасної пам'яті, призначені для вловлювання складних емоційних нюансів у музиці. Модель використовує згорткові нейронні мережі для надійного виявлення функцій і мережі довготривалої короткочасної пам'яті для ефективного навчання послідовності, звертаючись до часової динаміки музичних особливостей. Використовуючи набір даних Emotify, що включає доріжки з дев'ятьма емоційними характеристиками, дослідження розширює набір даних, сегментуючи кожну доріжку на 20 частин, таким чином збагачуючи різноманітність емоційних проявів. Для протидії дисбалансу набору даних, забезпечуючи рівномірне представлення різних емоцій, було застосовано такі методи, як техніка передискретизації синтетичної меншості. Спектральні характеристики зразків аналізували за допомогою швидкого перетворення Фур'є, що сприяло більш повному розумінню даних. Завдяки ретельному тонкому налаштуванню, включаючи реалізацію відсіву для запобігання надмірному оснащенню та коригування швидкості навчання, розроблена модель досягла помітної точності 94,7 %. Цей високий рівень точності підкреслює потенціал моделі для застосування в цифрових музичних службах, системах рекомендацій і музичній терапії. Майбутні вдосконалення цієї системи класифікації музичних емоцій включають розширення набору даних і вдосконалення архітектури моделі для ще більш тонкого емоційного аналізу.

Ключові слова: глибоке навчання; класифікація емоцій; нейронна мережа; спектральний аналіз; згорткова нейронна мережа

ABOUT THE AUTHORS

Vitaliy Yakovyna - Dr Hab., Associate Professor. Faculty of Mathematics and Computer Science. University of Warmia and Mazury in Olsztyn, 2, Oczapowskiego Str. Olsztyn, 10-719, Poland
ORCID: http://orcid.org/0000-0003-0133-8591; yakovyna@matman.uwm.edu.pl. Scopus Author ID: 6602569305
***Research field*:** Software reliability and safety; machine learning; computational intelligence

Яковина Віталій Степанович - доктор технічних наук, професор. Професор факультету Математики та комп'ютерних наук. Вармінськo-Мазурський університет в Ольштині, вул. Очаповського, 2. Ольштин, 10-719, Польща

Valentyn V. Korniienko – Master, Lviv Polytechnic National University, 12, Bandera Str. Lviv, 79000, Ukraine
ORCID: https://orcid.org/0009-0000-2581-9133; valik.kornienko97@gmail.com
***Research field*:** Artificial intelligence; deep learning models

Корнієнко Валентин Валерійович – Магістр, Національний університет «Львівська політехніка»,
вул. С. Бандери, 12. Львів, 79000,Україна