Advances in Science and Technology Research Journal

Advances in Science and Technology Research Journal, 2026, 20(1), 354–372
https://doi.org/10.12913/22998624/210059
ISSN 2299-8624, License CC-BY 4.0

Received: 2025.07.10
Accepted: 2025.09.29
Published: 2025.11.21

A convolutional neural network-driven model with adaptive feature fusion for Polish national dance music recognition

Kinga Chwaleba[1*], Weronika Wach[1]

[1] Faculty of Electrical Engineering and Computer Science, Department of Computer Science, Lublin University of Technology, Nadbystrzycka 38D, 20-618 Lublin, Poland
* Corresponding author's e-mail: k.chwaleba@pollub.pl

ABSTRACT

Mel spectrograms have been widely applied in music identification, often yielding successful results when combined with well-known pre-trained classification methods such as VGG16, DenseNet121, or ResNet50. However, the acquired performance may still be improved by employing fusion techniques and proposing a dataset consisting of more samples, which generally demonstrate superior results. Thus, a novel approach employing these methods with the formerly pre-trained classifiers has been introduced. The core innovation of our study is feature fusion utilizing Mel spectrograms, spectrograms, scalograms, and Mel-Frequency Cepstral Coefficients plots, created based on audio recordings from the created dataset encompassing Polish national dance music. The adaptive model is suggested as a mechanism adjusting the highly relevant features for Polish national dance music identification. Furthermore, the use of SHapley Additive exPlanations makes it possible to visualize which parts of the input feature maps are crucial to the model fusion decisions. Subsequently, the most prevalent classification metrics were employed including accuracy, precision, recall, and F1-score to compare the obtained results with state-of-the-art. Hence, the present method yields highly satisfactory results, exceeding 94% accuracy. Consequently, this study not only sets a new benchmark for Polish national dance recognition but also underscores the broader potential of multi-representation fusion as a general blueprint for next-generation audio classification systems.

Keywords: machine learning, convolutional neural networks, Polish national dance music identification, SHapley Additive exPlanations, feature fusion.

INTRODUCTION

Artificial intelligence (AI) and machine learning (ML) are currently undergoing rapid development, impacting nearly every aspect of daily life. One of the affected areas is music recognition, which serves as a valuable element of music information retrieval (MIR), focusing on identifying crucial features and patterns within audio signals [1]. It also supports a wide range of applications, such as genre classification or song detection, which can be performed using cutting-edge technologies involving convolutional neural networks (CNNs). This may entail utilizing widely applied pre-trained classifiers or incorporating various visual representations of sound as input to the network. These up-to-date techniques can often be employed not only to contribute to entertainment purposes but also to fulfill more educational or scientific objectives. Some other areas where machine learning may be applicable include audio signal processing in pipe organ systems [2–4].

One of the aspects that may be overlooked by the majority is related to intangible cultural heritage (ICH) which includes traditions, practices, or knowledge transmitted across generations. It helps shape every nation's identity, distinguishing them from one another. Music and dances can also be characterized as an essential element of ICH revealing unique characteristics of various cultures [5–7]. Notably, Polish national dances (the Krakowiak, the Kujawiak, the Mazurek, the Oberek, and the Polonez) reflect

Poland's cultural heritage. Yet, each of them consists of distinct qualities and features that help differentiate them [8]. In addition, these dances and their music serve as a significant element of Polish ICH and history [9]. Thus, AL and ML may help preserve them and make them accessible to younger generations.

MOTIVATION OF THE STUDY

The main motivation for this study was the lack of scientific works regarding Polish national dance music that applied some of the recent trends that have developed in the area of music classification such as the fusion of various features, attention mechanisms, or explainable artificial intelligence (XAI) techniques. Some prevalent databases were utilized to verify the lack of suitable literature search such as Scopus, Web of Knowledge, IEEE Xplorer, and Google Scholar. It was decided to use some of the following keywords: 'Polish national dance music recognition', 'folk music', 'machine learning', 'attention mechanisms'. Hence, this research aims to present a novel multimodal CNN that fuses Mel spectrograms, spectrograms, scalograms, and MFCCs using attention mechanisms. The network architecture is based on popular pre-trained classifiers and is applied to a dataset of audio recordings of five Polish national dances.

Scientific novelty of the proposed work

The key contributions can be summarized as follows:

1. Altering the available dataset of Polish national dance music [10] by reducing the length of audio samples from 10 seconds to 3 seconds. Contrasting both datasets and their performance proves the superiority of the datasets with shorter recordings. This demonstrates that even 3-second pieces are valuable for music recognition.
2. Developing the innovative multimodal CNN that performs a fusion of spectral features by employing the Adaptive Attention Module (AAM) with the adaptive trunk branch, the adaptive mask branch, and the adaptive gate. The proposed structure strengthens the network's performance in detecting salient sound characteristics and mitigating over-suppression in subsequent layers. Applying modalities such as Mel spectrograms, spectrograms, scalograms, and MFCC plots.
3. Introducing up-to-date formerly pre-trained classifiers, including EfficientNetB0, Xception, VGG16, VGG19, ResNet50, MobileNetV2, and DenseNet121, as a core of each CNN's modality. Comparing their efficiency utilizing one of the most prevalent metrics such as testing accuracy, testing loss, precision, recall, and F1-score for both of the audio recording collections used.
4. Visualizing the Shapley Additive explanations (SHAP) for each of the analyzed sound visualization techniques to localize areas that the developed classifier utilizes to predict Polish national dance music.
5. Performing an ablation study by evaluating testing accuracy and testing loss when one, two, or three of the proposed modalities are reduced for various classification methods and two datasets.
6. Comparing state-of-the-art techniques across various datasets, implemented models, and commonly used deep learning methods in music recognition.

Notably, previously researchers focused mainly on feature-based approaches where features such as Mel-Frequency Cepstral Coefficients (MFCC) vectors were extracted and fed to the classifiers [11]. There were some drawbacks to this method, as it might not capture the temporal and spatial dependencies present in audio signals and often requires careful feature engineering. Moreover, spectrograms were predominantly chosen in many music classification works incorporating CNNs while they might overlook some other aspects of the audio signal. Thus, our CNN-driven model with Adaptive Feature Fusion is introduced to address these limitations. It incorporates multimodality by combining four inputs that visualize signals in various ways such as spectrograms, Mel spectrograms, MFCC plots, and scalograms. Moreover, applying attention mechanisms and pre-trained classification methods can also enhance classification accuracy and the extraction of complementary features. In comparison with traditional approaches, the proposed method yields superior results through the integration of multiple audio representations and advanced convolutional neural network architecture.

RELATED WORKS

Many recent papers focus mainly on analyzing visual representations of audio recordings employing various presentation techniques such as Mel spectrograms, spectrograms, scalograms, and Mel-Frequency Cepstral Coefficients plots. They facilitate the application of convolutional neural networks to music identification tasks. In [11] Mel spectrograms were obtained from 30-second audio recordings originating from the GTZAN dataset [12]. They represent a range of 10 various music genres e.g. country, rock, and pop music. Then, it was decided to employ pre-trained on ImageNet [13] models namely ResNet34, ResNet50, VGG16, and AlexNet. Following the training stage, each classifier was independently assessed utilizing confusion matrices and accuracy metrics ranging from 71 to 79%. Spectrograms and Mel spectrograms with the utilization of other pre-trained classification methods such as AlexNet and LeNet-5 were also presented in [14] demonstrating favourable outcomes. Moreover, spectrograms were also generated from the GTZAN and 10GenreGram subsets, as detailed in [15]. Then, ResNet18 and NNet2 classifiers were applied, trained on 50 epochs, and evaluated by contrasting their accuracies and confusion matrices. It is noteworthy that the results varied significantly, with accuracy ranging from 40% to nearly 80%. To address the task of music genre classification, Mel spectrograms combined with self-adjusted convolutional neural networks (CNNs) were also employed in [16–18] yielding sufficient results in terms of obtained metrics comprising precision, recall, F1-score, and accuracy. Alternative methods for sound visual representations include spectrograms [19–22] and scalograms [23,24] generation which are also utilized as an input to the CNNs contributing to achieving satisfactory outcomes. MFCCs were extracted from audio signals in [25–27] and subsequently utilized in classification models such as support vector machines (SVM), convolutional recurrent neural networks (CRNN), and Convolutional Neural Networks, respectively. In [28], MFCC plots were generated using the extracted coefficients and employed to train the self-adjusted CNN.

While many papers tackle music genre classification employing some well-known datasets such as the GTZAN, the FMA [29,30], or EMA, there is still a lack of scientific works relating to folk music classification. However, the recent rise in awareness about the importance of preserving nations' intangible cultural heritage (ICH) provides research with some valuable insights. In [31] an ethnic music dataset was obtained representing ten various genres with each genre containing 100 audio recordings lasting 20 seconds each. Then Mel sound spectrum and short-time Fourier spectroscopy were utilized as an input into a self-adjusted CNN and compared in terms of obtained accuracy. There are some other works combining folk music classification with machine learning techniques concerning different nations and cultures including Chinese culture [32, 33], Turkish [34], Greek [35], Indian [36–38], Hungarian [39], Vietnamese [40], Assamese [41], Bengali [42], Irish [43], Korean [44], Arabic [45], and Nigerian [46]. Noteworthy, there is also some work presenting Polish national dance music recognition [10] where audio samples demonstrating Polish national dances such as the Krakowiak, the Kujawiak, the Mazur, the Oberek, and the Polonez were collected yielding a dataset encompassing 137 recordings in the MP3 format. Then, the data preprocessing stage was performed where each audio sample was carefully listened to remove some unnecessary parts including noise, etc. Afterwards, each audio was converted to the WAV format and divided into 10-second pieces obtaining a final dataset consisting of over two thousand samples. From each sample Mel spectrograms were generated, split into training, validation, and testing sets in an 8:1:1 ratio, and utilized as an input into the pre-trained ImageNet classifiers such as VGG16, ResNet50, DenseNet121, and MobileNetV2. The acquired outcomes were compared with the attained metrics such as testing accuracy, testing loss, precision, recall, and F1-score resulting in a testing accuracy of approximately 90%.

Recently, attention mechanisms have gained more and more recognition concerning computer vision issues such as image classification helping achieve improved outcomes [47,48]. Music genre recognition could also be handled employing attention mechanisms which were utilized in [49] where Mel spectrograms and MFCC plots were generated based on the GTZAN dataset. Subsequently, three distinct architectures were chosen and trained such as ResNet18, Bi-LSTM, and ResNet18-BiLSTM with ResNet18 and ResNet18-BiLSTM implementing the convolutional block attention module (CBAM). Then, each classifier was trained with and without it, and the impact of each feature and its combination

was also evaluated in terms of the obtained testing accuracy. The combination of Mel spectrograms and CNNs with attention mechanisms utilized for the music genre classification was also implemented in the following papers [50–52] providing acceptable outcomes. Spectrograms are fairly commonly selected methods employed to visualize audio signals when combined with a range of attention mechanisms confronting music genre recognition issues. The residual attention network (RAN) applying residual blocks and attention modules was selected in [53], while [54] presented the CNN with NetVlad and self-attention. Additionally, MFCC extraction was utilized with the attention-based CNN showcasing accuracy at the level of 85% in [55].

It was recognized that it might be beneficial to enhance sound recognition by developing classifiers based on the fusion of several music features in CNNs. In [28], it was stated that it could be more valuable to use the late-fusion strategy to combine features such as MFCC plots, Mel spectrograms, and spectrograms due to the possible information redundancy. The aforementioned features are strictly connected with each other as Mel spectrograms derived from short-time Fourier transform (STFT) [56] which is the backbone of spectrograms, and MFCC plots are connected to Mel spectrograms. The presented solution yielded promising accuracy. The late-fusion strategy was also demonstrated in [57] with the same set of combined features.

It is essential to provide AI models with suitable resources to be able to properly assess their effectiveness. Thus, XAI has emerged and provides some proper techniques to address this concern. Regarding the sound recognition approaches such as Explain like I am 5 (ELi5), SHAP, and local interpretable model-agnostic explanations (LIME) were selected in [58] providing some impactful insights into CNNs' decision-making process. The SHAP technique was also elected in [59,60] to enable more transparent interpretation of the classification methods' results.

The conducted literature research facilitated the assessment of the current state of music recognition and the identification of up-to-date techniques utilized in this area. It could be observed that many studies remained focused on music genre classification based on commercially popular music. They typically employed their own self-employed CNN or some pre-trained classifiers with various types of inputs including Mel spectrograms, spectrograms, MFCC plots, or scalograms. Some of the works provided more complex solutions requiring the implementation of various attention mechanisms or fusion of features. To gain more comprehension of the chosen classification methods some XAI tools have been implemented, too. Nevertheless, more and more works concerning folk or national songs relating to distinct cultures around the world have been revealed. Consequently, it was determined to undertake research pertaining to Polish music recognition combined with the widely known previously pre-trained classifiers with the Adaptive Feature Fusion enabling the classification of Polish national dance music. To assess the performance of the provided classification methods the SHAP technique was employed.

MATERIALS AND METHODS

This section presents the general research methodology. Some main elements of this may be retrieved. Firstly, the overall dataset containing music of five Polish national dances in WAV audio recordings was collected. Secondly, the collected samples were segmented into 3-second and 10-second clips, resulting in the creation of two separate datasets. Then, distinct sound features such as Mel spectrograms, spectrograms, scalograms, and MFCC plots were extracted and saved as JPG images. Ultimately, developed classifiers were trained and evaluated utilizing chosen classification metrics.

Dataset

It was decided to utilize the formerly developed Polish national dances music dataset presented in [10] comprising music representing five Polish national dances such as the Krakowiak, the Kujawiak, the Mazur, the Oberek, and the Polonez. The demonstrated collection consists of 137 audio recordings in the MP3 format. Audio samples were previously manually checked to remove parts that did not contain music including crowd noises or silence and converted into the WAV format. Although this step was performed manually in the current study, future research could automate the process. For example, pyAudioAnalysis, a Python library, can be introduced to remove silence periods from audio recordings [61]. In [10] audio

recordings were then divided into 10-second pieces. In our study, this process was enhanced by creating two sets where one reflected 10-second samples and the other 3-second samples. It was selected to evaluate the impact of employing audio recordings that lasted only 3 seconds due to similar studies presenting this issue [62,63] where favourable results were exhibited. It increases the overall size of the dataset which may result in achieving enhanced outcomes. It can also be used to evaluate whether shorter data samples affect music recognition performance. Notably, during the data splitting stage, it is essential to select only those samples that match the specified length, while shorter recordings must be excluded.

As might be observed in Table 1, an over 3 times larger dataset was obtained due to the splitting of audio samples into 3-second pieces. Yet, each class which is represented by each Polish national dance presents the same data distribution with the Kujawiak still being the most numerous class, and the Mazur the least. The dataset continues to exhibit a slight imbalance in the number of samples among the classes.

Data preprocessing

The analysis of audio signals may be conducted using a variety of techniques encompassing time-domain analysis, frequency-domain analysis, and wavelet analysis. Each method facilitates the process of capturing essential spectral characteristics of the signal.

By applying the Fourier Transform to localized time intervals, the short-time Fourier transform (STFT) reveals variations in the frequency characteristics of sound signals over time [64]. STFT may be analyzed by Equation 1, where t denotes the time parameter of the signal, u represents the frequency parameter, $f(t)$ is the input signal, and W refers to the windowing function [65].

$$STFT_f^u(t',u) = \\ = \int_t [f(t) \cdot W(t-t')] \cdot e^{-j2\pi ut} dt \quad (1)$$

Spectrograms help depict this technique where the x-axis refers to time and the y-axis to the frequency (Figure 1) [66]. Often, the frequency on the vertical axis is visualized using a logarithmic scale due to the fact that people typically have a better perception of low-frequency sounds compared to high-frequency ones.

A Mel spectrogram can be defined as a spectrogram where frequencies are transformed according to the Mel scale [22] presented by Equation 2. The scale is designed to align with human auditory perception of sound frequencies.

$$Mel\ Scale = 2595\, log_{10} \\ \left(1 + \frac{frequency}{700}\right) \quad (2)$$

An example of Mel spectrogram with the time on the x-axis and Mel-frequency bins on the y-axis was generated for the Mazur dance (Figure 1). Unlike spectrograms with the linear or logarithmic scale, Mel spectrograms map frequencies onto the Mel scale that aligns better with human auditory perception as it compresses higher frequencies and expands lower frequencies. That is why it may be a better tool utilized in sound analysis due to the fact that it focuses on how humans hear.

The Mel scale is incorporated into the calculation of Mel-Frequency Cepstral Coefficients, too. The transformation of the input signal is processed, and then Mel filter banks are employed to compute the amplitude across the frequency bands defined by the Mel scale. Eventually, the cepstral coefficients are determined by taking the logarithm of these amplitudes.

A sample MFCC plot was depicted in Figure 1 with the horizontal axis representing time and the

Table 1. The aggregate number of WAV-format audio samples

Dance	Samples number before splitting	10-second samples number	3-second samples number
Krakowiak	23	444	1501
Kujawiak	34	588	1993
Mazur	25	410	1388
Oberek	38	428	1472
Polonez	17	426	1432
Overall	137	2296	7782

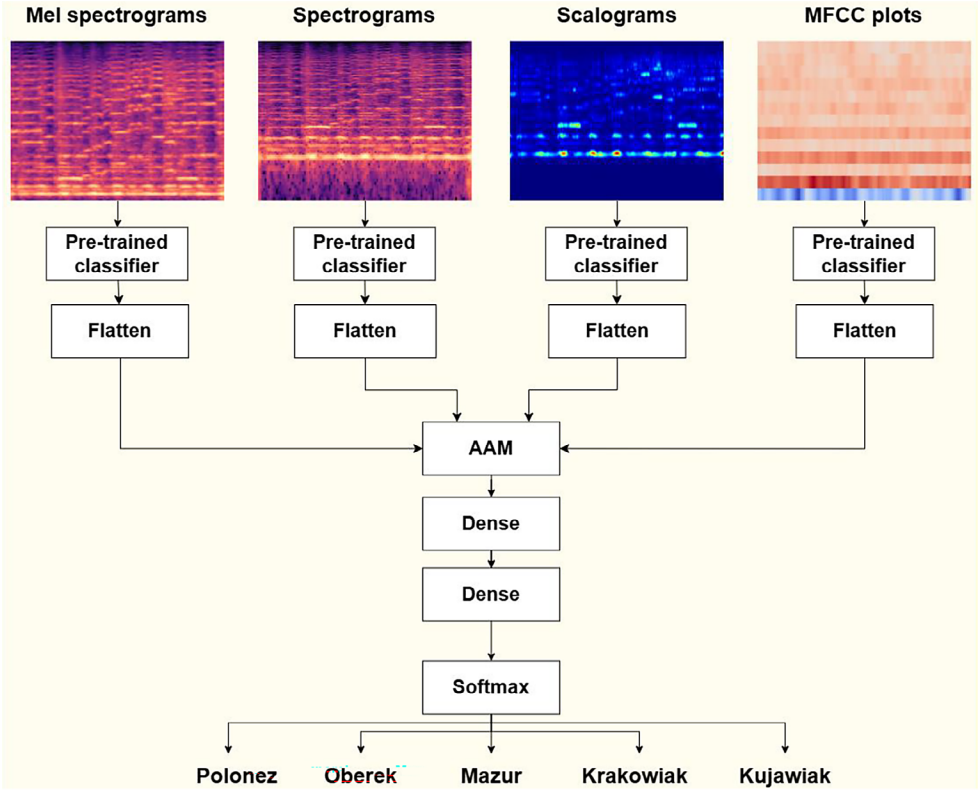

Figure 1. The CNN-driven model with adaptive feature fusion

calculated MFCC on the vertical axis. The magnitude of the associated coefficient is demonstrated by the intensity, potentially indicating the presence of distinctive audio characteristics.

Continuous wavelet transform (CWT) utilizes wavelets to analyze changes in a signal's frequency content over time, capturing both short- and long-term features by adapting the size of the analysis windows. It is defined by Equation 3 [65] where ψ denotes the mother wavelet (Equation 4). One of the most widely adopted wavelets is the Morlet wavelet, developed as a combination of a sine wave with a Gaussian function [67].

$$CW(a,b) = \int_{-\infty}^{\infty} f(t)\psi_{a,b}(t)dt \tag{3}$$

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \tag{4}$$

Scalograms are an effective approach to the visualization of CWTs (Figure 2). The horizontal axis represents time, whereas the vertical axis corresponds to scale.

Model architecture

The innovative model combining multiple spectral features of sound with the late-fusion strategy has been proposed to tackle music identification (Figure 1). The designed network has four separate inputs pertaining to spectral characteristics derived from audio including Mel spectrograms, spectrograms, scalograms, and MFCC plots which have been merged into a multimodal CNN using a late fusion strategy to mitigate the presence of redundant features [28]. Each outlined modality encompasses one of the most prevalent pre-trained ImageNet classifiers such as EfficientNetB0, Xception, VGG16, VGG19, ResNet50, MobileNetV2, and DenseNet121. It is essential to ensure that every branch incorporates the same classification method. Each time the last layer from these classifiers is removed and it is connected to the flattened layer. Then, the up-to-date solution named Adaptive Attention Module is adopted. This mechanism consists of an adaptive trunk branch, an adaptive mask branch, and an adaptive gate. After feature fusion, two Dense blocks are utilized and Softmax is employed to classify one of the dance classes.

Pre-trained classifiers

After comprehensive research, it was determined to select some of the most prevalent

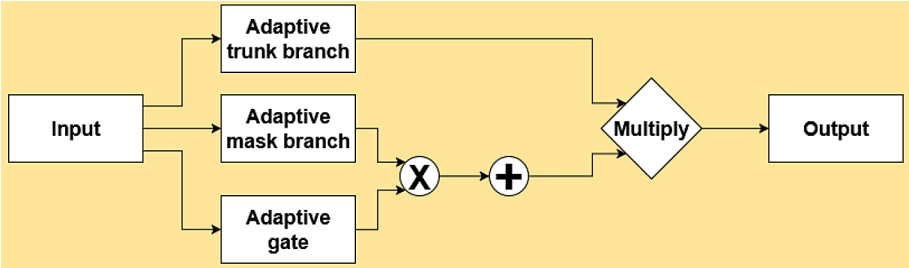

Figure 2. The adaptive feature fusion mechanism

classification methods like Xception, VGG16, VGG19, ResNet50, MobileNetV2, DenseNet121, and EfficientNetB0. Their unique characteristics may be valuable to increase the effectiveness of the introduced classifier for Polish national dance music identification:

1. Xception is a novel classifier where an Inception module is substituted with depthwise separable convolutions resulting in achieving better performance [68].
2. VGG introduced deeper convolutional neural networks by employing more convolutional layers [69]. This became feasible due to leveraging 3 × 3 convolutional filters, contributing to improved accuracy. VGG16 and VGG19 represent 16 and 19-weight layers, respectively.
3. ResNet50 provides training of deeper neural networks without their degradation due to the application of shortcut connections and residual blocks [70]. Moreover, it maintained a lower level of complexity compared to the aforementioned VGG classifiers.
4. The decrease in memory consumption was obtained as a result of an innovative design of MobileNetV2 [71]. It developed a cutting-edge structure established on inverted residuals with linear bottlenecks and depthwise separable convolutions.
5. ResNet classification methods are surpassed by DenseNet121 by enhancing feature propagation while minimizing the number of parameters [72]. This classifier also addresses the vanishing gradient issue by establishing dense connections between layers.
6. EfficientNets contrast with formerly mentioned classifiers such as MobileNets and ResNet yielding higher accuracy due to the compound scaling [73]. It adjusts the depth, width, and resolution of the network evenly while the stem layer conducts initial convolutions. EfficientNetB0 is a classifier belonging to the EfficientNet family.

Adaptive attention module

It may be challenging to recognize even slight disparities between audio signals by some prevalent classification methods. Thus, an adaptive attention structure called AAM has been proposed to handle feature fusion (Figure 2). This facilitates the dynamic adjustment of the number of attention modules based on the input class or type of image. Extra gate layers in the proposed modules dynamically influence the feature maps by optimizing the attention mask structure according to the complexity of the detected sound features. This design mitigates the over-suppression of features in deeper network layers while enhancing the model's ability to capture important sound characteristics. The suggested adaptive attention module consists of three main components: the adaptive trunk branch, the adaptive mask branch, and the adaptive gate.

Adaptive trunk branch

The proposed adaptive trunk branch serves as the primary feature extraction path, containing two convolutional layers with a *3x3* kernel size $Conv_{3x3}$, followed by a rectified linear unit (ReLU) activation function and batch normalization (BN) (Equation 5, 6).

$$AT_1(x) = ReLU(BN(Conv_{3\times 3}(x))) \quad (5)$$

$$AT(x) = BN(Conv_{3x3}(AT_1(x))) \quad (6)$$

Adaptive mask branch

The adaptive masking branch enhances salient regions in the input of the classifier by forming a spatial-channel mask. This process includes channel reduction, application of a standard *3x3* convolutional kernel, and concludes with a sigmoid activation function. It is presented by Equation 7 with $Conv_{1x1}$ noting a convolutional layer

which is responsible for reducing the channels' number (which can serve as a bottleneck). Conversely, σ refers to the sigmoid activation function limiting the output to the 0–1 range.

$$M_a(x) = \sigma\left(\begin{matrix} Conv_{3\times 3} \\ (ReLU(Conv_{1\times 1}(x))) \end{matrix}\right) \quad (7)$$

Adaptive gate

The adaptive gate is employed to perform the process of dynamically scaling the attention mask to allow the classifier to adjust the impact of attention based on the presence of sound features. The gating function, defined by Equation 8, uses Global Average Pooling (AvgPool) to reduce the input tensor from $\mathbb{R}^{CxHxW}$ to $\mathbb{R}^{Cx1x1}$ where C is the number of channels, H is the height, and W is the width.

$$G(x) = \sigma\left(\begin{matrix} Conv_{1\times 1} \\ (ReLU(Conv_{1\times 1}(AvgPool(x)))) \end{matrix}\right) \quad (8)$$

This attention module adjusts weights based on global audio characteristics – distinct acoustic patterns like harmonic structures result in higher gating weights, enhancing their influence. The final output of the Adaptive Attention Module is calculated utilizing Equation 9 where the output is calculated through element-wise multiplication of the trunk features and attention weights.

$$\begin{aligned} Output(x) = \\ = AT(x) \odot \left(1 + G(x) \cdot M_a(x)\right) \end{aligned} \quad (9)$$

Employing a gating mechanism $G(x)$ avoids over-suppressing features, especially in noisy or low-quality audio, making it valuable in variable sound environments.

Classification metrics

The CNN-driven model with adaptive feature fusion was assessed by applying some of the most prevalent classification metrics involving accuracy, precision, recall, and F1-score.

In the mentioned equations true positive (TP) represents correctly identified positives, false positive (FP) – incorrectly identified positives, and false negative (FN) – missed positives. In addition, accuracy reflects the ratio of correctly predicted instances to the total number suggesting the overall performance of the proposed classifier across all classes [74]. Precision and recall may also be utilized for multi-class classification when precision evaluates how many predicted instances of a class are correct and recall reflects how well the classifier identifies all actual instances of that class [75]. A harmonic mean of recall and precision is F1-score [76] which may be particularly advantageous when dealing with uneven class distributions.

As another classification metric loss is also computed utilizing the Categorical Cross Entropy [77]. It is a widely applied function in multiclass classification that examines the difference between predicted and actual probability distributions.

Experiments

Mel spectrograms, spectrograms, scalograms, and MFCC plots were generated for the samples both from the 3-second and 10-second datasets. Subsequently, every catalog consisting of various sound features was randomly divided into training, validation, and testing sets. Specifically, 80% of the data was allocated to the training set, with 10% each assigned to validation and testing. During this process, it was vital to ensure that each directory representing diverse auditory characteristics followed the same labeling and number of samples.

It was established to employ several prevalent classification methods like EfficientNetB0, Xception, VGG16, VGG19, ResNet50, MobileNetV2, and DenseNet121 to examine their impact on the proposed CNN-driven model. The training process encompassed the following steps for 3- and 10-second audio segments separately:

1. Change the core classifier of the proposed model.
2. Load images of Mel spectrograms, spectrograms, scalograms, and MFCC plots for training, validation, and testing sets.
3. Preprocess images according to the specific needs of the selected classification method.
4. Set the number of epochs to 30 based on the former studies [15].
5. Evaluate the performance of the model by determining the value of the following metrics: testing accuracy, testing loss, precision, recall, and F1-score.

The aforementioned training stage was repeated each time an alternative primary classifier was employed.

RESULTS

The performance of the proposed classifier was demonstrated separately for both the 3-second and 10-second datasets. Then, an ablation study was performed and SHAP visualization was presented.

The CNN-driven model with adaptive feature fusion

The performance of the proposed CNN-driven model, incorporating the adaptive feature fusion module, was evaluated based on testing accuracy and testing loss (Table 2). The results were calculated for different core classification methods across both 10-second and 3-second audio datasets. It might be noted that EfficientNetB0 and ResNet50 achieved the highest testing accuracy (over 94%) on the 3-second dataset, with EfficientNetB0 also yielding the lowest loss (around 0.16) while VGG16 performed the worst on this dataset, with 88% accuracy and 0.75 loss. For the 10-second dataset, DenseNet121 performed the best (91% accuracy and around 0.25 loss), while ResNet50 showed the weakest results (86% accuracy and around 1.73 loss). The results obtained for the remaining classifiers demonstrated relatively consistent performance. In the 10-second sample dataset, testing accuracy fluctuated within a range of approximately 2–3 percentage points, with testing loss varying by about 1.2. For the 3-second dataset, accuracy differences spanned roughly 4–5 percentage points, while testing loss varied by around 0.5.

One of the most prevalent classification metrics such as precision, recall, and F1-score has been computed for every classifier, dataset and class representing music from one of the Polish national dances such as the Krakowiak, the Kujawiak, the Mazur, the Oberek, and the Polonez. They were presented in Tables 3–9.

Regarding DenseNet121, the 3-second dataset yielded to a minor extent better outcomes. However, for the Krakowiak and the Oberek dance F1-score was higher for the 10-second dataset. The Polonez class presented the highest precision (over 95%) with slightly lower recall for the 3-second sample dataset. Notably, the Oberek dance gained the highest precision (100%) with significantly lower recall (over 86%) for the 10-second dataset. In general, all dances (for both datasets) might be characterized by relatively high metrics with precision and recall above 86% and F1-score above 88%.

It could be observed that considering the EfficientNetB0 classifier in almost every scenario precision, recall, and F1-score yielded notably higher outcomes for the database with 3-second audio recordings than with the 10-second pieces. Especially for the Kujawiak, the Oberek, and the Polonez classes F1-score acquired over 95%. Correspondingly, MobileNetV2 performed better on the dataset with shorter samples resulting in an F1 score between 91–94% while the dataset with 10-second pieces obtained the highest F1 score for the Kujawiak dance (around 92%).

The ResNet50 classification method demonstrated even better for the 3-second dataset with the F1-score between 93-95% and recall and precision mainly over 90%. Substantial disparities might be noticed for the 10-second collection where differences between precision and recall were around 20 percentage points for almost every class. For example, for the Polonez class precision is 100%, however, recall was only around 72%. In addition, the obtained F1-score was around 82% for the Mazur, while the highest was for the Kujawiak and it was equal to 90%.

Table 2. The testing accuracy (in %) and testing loss for the selected classifier on both datasets (TA – testing accuracy, TL – testing loss)

Classifier	Dataset type	TA	TL	Dataset type	TA	TL
DenseNet121	10 s samples	91.553	0.254	3 s samples	91.826	0.333
EfficientNetB0	10 s samples	90.871	0.375	3 s samples	94.636	0.165
MobileNetV2	10 s samples	91.008	0.362	3 s samples	93.614	0.287
ResNet50	10 s samples	86.512	1.735	3 s samples	94.252	0.415
VGG16	10 s samples	88.692	1.549	3 s samples	88.505	0.750
VGG19	10 s samples	89.373	1.006	3 s samples	89.399	0.638
Xception	10 s samples	88.692	0.356	3 s samples	92.337	0.317

Table 3. The obtained metrics (in %) for DenseNet121 regarding the type of dataset

Type	10-s dataset			3-s dataset		
Dance/Metric	F1-score	Precision	Recall	F1-score	Precision	Recall
Krakowiak	90.526	86.000	90.526	89.864	91.724	88.079
Kujawiak	93.548	90.625	96.666	93.796	93.103	94.500
Mazur	89.156	88.095	90.243	89.679	89.361	90.000
Oberek	92.682	100.000	86.363	90.066	88.311	91.891
Polonez	90.476	95.000	86.363	95.774	97.142	94.444

Table 4. The obtained metrics (in %) for EfficientNetB0 regarding the type of dataset

Type	10-s dataset			3-s dataset		
Dance/Metric	F1-score	Precision	Recall	F1-score	Precision	Recall
Krakowiak	85.106	81.632	88.888	91.582	93.15	90.066
Kujawiak	95.798	96.610	95.000	95.477	95.959	95.000
Mazur	92.500	94.871	90.243	94.244	94.927	93.571
Oberek	86.021	81.632	90.909	95.145	91.304	99.324
Polonez	92.682	100.000	86.363	95.774	97.142	94.444

Table 5. The obtained metrics (in %) for MobileNetV2 regarding the type of dataset

Type	10-s dataset			3-s dataset		
Dance/Metric	F1-score	Precision	Recall	F1-score	Precision	Recall
Krakowiak	90.526	86.000	95.555	92.255	93.835	90.728
Kujawiak	92.561	91.803	93.333	94.865	92.822	97.000
Mazur	89.156	88.095	90.243	94.366	93.055	95.714
Oberek	89.156	94.871	84.090	91.408	93.006	89.864
Polonez	93.023	95.238	90.909	96.140	97.163	95.138

Table 6. The obtained metrics (in %) for ResNet50 regarding the type of dataset

Type	10-s dataset			3-s dataset		
Dance/Metric	F1-score	Precision	Recall	F1-score	Precision	Recall
Krakowiak	86.000	78.181	95.555	94.520	97.872	91.390
Kujawiak	90.598	92.982	88.333	94.285	90.000	99.000
Mazur	82.105	72.222	95.121	94.076	91.836	96.428
Oberek	87.500	97.222	79.545	95.890	97.222	94.594
Polonez	84.210	100.000	72.727	93.090	97.709	88.888

Surprisingly, there were notable imbalances of the obtained metrics regarding VGG16 for both datasets. F1-score was between 82–92% for the 10-second collection, and 79–94% for the 3-second collection. It could be recognized that the Mazur dance performed considerably low with a precision of around 69% and 79% for the F1-score for the 3-second dataset. Similarly, for the collection with longer samples, the Mazur yielded the worst outcomes. Although the F1 score was higher (around 82%). Concerning the VGG19 classifier, the 3-second datasets yielded generally higher outcomes. However, for the Mazur and the Oberek dance, the F1 score was slightly lower. Overall, the notable disparities between obtained metrics might be spotted for the dances including the Mazur, the Oberek, and the Polonez. The 3-second dataset outperformed the 10-second dataset pertaining to the Xception with the F1-score between 89–95%. Precision and recall also presented more concise

Table 7. The obtained metrics (in %) for VGG16 regarding the type of dataset

Type	10-s dataset			3-s dataset		
Dance/Metric	F1-score	Precision	Recall	F1-score	Precision	Recall
Krakowiak	87.500	82.352	93.333	88.339	94.696	82.781
Kujawiak	92.561	91.803	93.333	92.783	95.744	90.000
Mazur	82.051	86.486	78.048	79.635	69.312	93.571
Oberek	89.887	88.888	90.909	86.545	93.700	80.405
Polonez	90.476	95.000	86.363	94.158	93.197	95.138

Table 8. The obtained metrics (in %) for VGG19 regarding the type of dataset

Type	10-s dataset			3-s dataset		
Dance/Metric	F1-score	Precision	Recall	F1-score	Precision	Recall
Krakowiak	84.210	80.000	88.888	88.054	90.845	85.430
Kujawiak	93.548	90.625	96.666	94.320	93.170	95.500
Mazur	88.372	84.444	92.682	83.280	74.576	94.285
Oberek	90.243	97.368	84.090	89.377	97.60	82.432
Polonez	88.888	97.297	81.818	92.086	95.522	88.888

Table 9. The obtained metrics (in %) for Xception regarding the type of dataset

Type	10-s dataset			3-s dataset		
Dance/Metric	F1-score	Precision	Recall	F1-score	Precision	Recall
Krakowiak	85.057	88.095	82.222	91.836	94.405	89.403
Kujawiak	89.230	82.857	96.666	93.137	91.346	95.000
Mazur	87.500	89.743	85.365	89.208	89.855	88.571
Oberek	88.372	90.476	86.363	93.069	90.967	95.270
Polonez	91.764	95.121	88.636	95.406	95.406	93.750

results between each class while for the collection with longer audio recordings recall was around 96% with precision only 83% for the Kujawiak.

Ablation study

An ablation study might be perceived as an up-to-date method utilized to evaluate the impact of individual components within a neural network by systematically removing or altering them and assessing results. Nowadays, it has gained notable acknowledgment in the music recognition area [78–80]. Hence, it was determined to perform an ablation study within our research with the introduced scenarios such as:

1. Ablation Scenario 1: The proposed classifier with Mel spectrograms, spectrograms, and scalograms.
2. Ablation Scenario 2: The proposed classifier with Mel spectrograms, scalograms, and MFCC plots.
3. Ablation Scenario 3: The proposed classifier with Mel spectrograms, spectrograms, and MFCC plots.
4. Ablation Scenario 4: The proposed classifier with scalograms, spectrograms, and MFCC plots.

The acquired results were presented in Table 10 for the 10-second dataset and in Table 11 for the 3-second dataset where testing accuracy and testing loss were presented for every assessed pre-trained classifier. Generally, the dataset with shorter audio samples yields better outcomes both regarding testing accuracy and testing loss. It might be perceived that ResNet gained the highest testing accuracy around 94.5% for the 3-second dataset when the spectrograms were reduced. Similarly, in the same dataset, MobileNetV2 achieved the lowest testing loss equal to 0.204. However, it was for this scenario that scalograms were reduced. As it might be noticed the VGG16 classifier performed worst in terms of the acquired testing loss for the longer audio recordings

as in every scenario the value of the testing loss was over 1. For the third and second scenario it was even higher than 1.7. Similarly, VGG19 had a testing loss above 1 in the fourth scenario. This indicates that VGG16 might be a less effective classifier compared to others.

The outcomes gained by the proposed CNN-driven model were compared with every ablation experiment for both datasets. Regarding the collection with longer audio samples, it was revealed that interestingly ablation scenarios sometimes outperformed the proposed model. For example, DenseNet121, EfficientNetB0, MobileNetV2, ResNet50, and Xception showed the highest accuracy when applied in the classification method with scalograms, spectrograms, and MFCC plots. For VGG16 and VGG19 the highest testing accuracy was obtained in this scenario without scalograms. Moreover, the obtained testing loss was only superior for DenseNet121. Concerning the 3-second database, only for EfficientNetB0 and Xception the highest testing accuracy was yielded. However, there is no noticeable pattern relating to the acquired accuracy across all scenarios and classifiers. The testing loss was the lowest for EfficientNetB0 and Xception for the CNN-driven model, too. Notably, EfficientNetB0 surpassed other classifiers in every scenario. On the 10-second datasets, EfficientNetB2 also exceeds both in terms of testing accuracy and testing loss. Although it was obtained for the experiment where scalograms were not employed.

The analysis of the ablation study may suggest that the unique compound scaling feature utilized in the EfficientNetB0 led to superior results relating to both datasets. Moreover, it is profitable to employ shorter audio samples as it leads to a general higher efficiency regardless of the experiment.

Shapley additive exPlanations

Nowadays, it is essential not only to generate accurate predictions but also to understand and interpret the underlying factors contributing to those predictions. Thus, XAI techniques are gaining increasing attention, also in music identification. One of the proposed techniques is SHAP which was presented in some research [58–60]. SHAP provides a unified method for explaining predictions by assigning an importance value to each feature. It combines several existing

Table 10. The testing accuracy (in %) and testing loss regarding the selected classifier and number of ablation scenarios for the 10-second samples dataset (TA – testing accuracy, TL – testing loss, SN – scenario number)

SN	Metric/Classifier	DenseNet121	EfficientNetB0	MobileNetV2	ResNet50	VGG16	VGG19	Xception
1	TA	91.553	89.509	89.645	90.599	88.692	89.237	90.190
	TL	0.318	0.481	0.916	0.903	1.890	0.999	0.380
2	TA	91.416	89.373	91.961	92.098	86.239	86.784	87.329
	TL	0.295	0.459	0.327	0.895	1.712	0.659	0.531
3	TA	89.782	93.188	92.098	89.645	90.054	90.190	89.373
	TL	0.342	0.340	0.340	1.006	1.428	0.996	0.357
4	TA	91.689	94.005	92.370	92.506	89.100	89.100	91.280
	TL	0.389	0.212	0.473	0.799	1.296	1.012	0.344

Table 11. The testing accuracy (in %) and testing loss regarding the selected classifier and number of ablation scenarios for the 3-second samples dataset (TA – testing accuracy, TL – testing loss, SN – scenario number)

SN	Metric/Classifier	DenseNet121	EfficientNetB0	MobileNetV2	ResNet50	VGG16	VGG19	Xception
1	TA	91.826	94.125	94.636	93.869	87.994	92.081	91.060
	TL	0.378	0.246	0.320	0.3125	0.991	0.713	0.406
2	TA	92.337	91.698	92.720	94.508	92.848	90.804	86.845
	TL	0.348	0.308	0.366	0.255	0.472	0.681	0.509
3	TA	93.486	91.315	93.869	94.252	92.975	88.505	88.888
	TL	0.262	0.355	0.204	0.230	0.409	0.785	0.423
4	TA	91.315	92.464	93.869	91.570	89.527	88.250	90.166

methods into one tool ensuring consistent, interpretable results. SHAP is supposed to better align with human understanding, too.

EfficientNetB0, which was utilized as a core classification method for every branch of the proposed model and the 3-second dataset, presented the most superior results across all introduced scenarios. Hence, it was determined to prepare the SHAP visualization based on this model. It is presented in Figure 3 when outcomes have been computed for the Kujawiak sample. Respectively, it is depicted for the Mel spectrogram, spectrogram, scalogram, and MFCC plot. For every horizontal axis, the calculated SHAP values were pictured with a greater red color symbolizing the greater influence of suggested features on the model's decision influence. Simultaneously, blue highlights features that negatively impact the prediction for that specific class. Moreover, gray areas suggest almost zero contribution to the prediction of the classifier.

It may be noticed that the proposed model was highly positive for the Kujawiak class regarding the Mel spectrogram. Since the strong red color is visible in this sample across the whole duration of the audio recording for the lower frequencies. There is some ambiguity regarding the spectrograms since the red color was computed for the first half of the Kujawiak sample, with the other half presented as some blue color. It might suggest less confident predictions regarding this visualization method. Strong feature importance is again present for the Kujawiak in the scalogram. However, there is some area indicating a negative impact of the scalogram feature while the same area implies a greater impact for the Oberek and the Krakowiak. Consequently, this may suggest that the classifier confused these classes within scalograms due to similar features presented across various classes. The Kujawiak was undoubtedly predicted positive for the MFCC plot across almost horizontal bands.

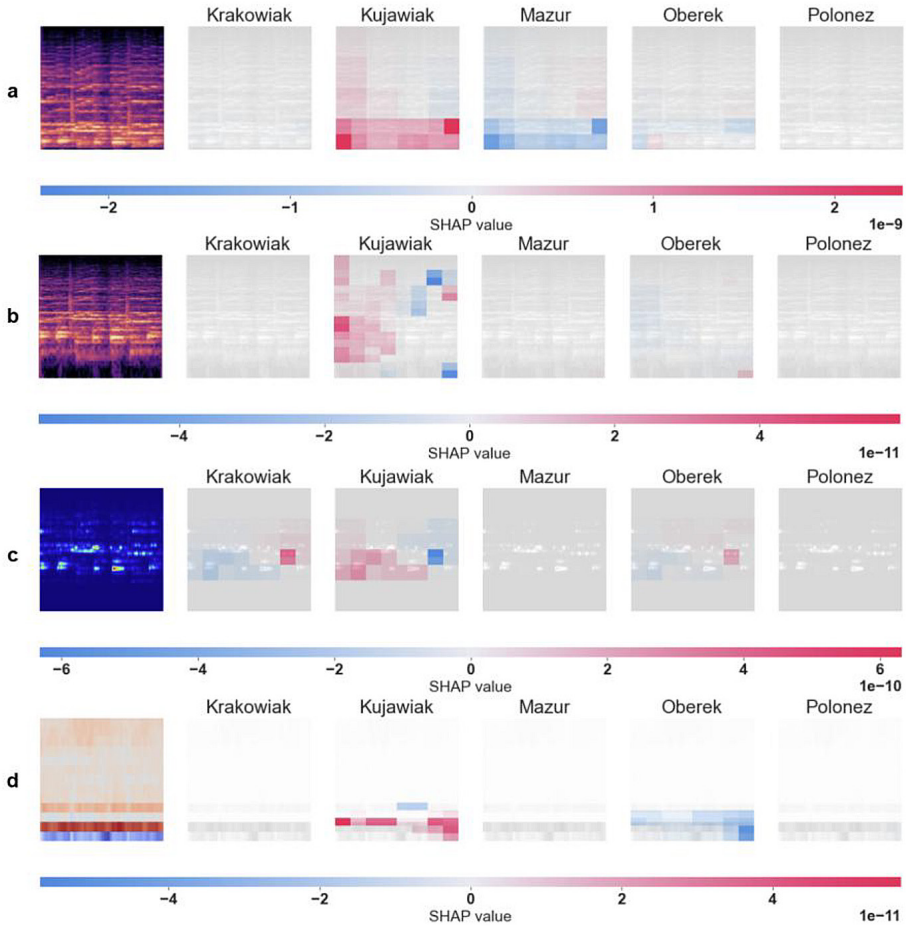

Figure 3. The SHAP visualization generated for the Kujawiak sample for the CNN-driven model based on EfficientNetB0 and for the 3-second dataset regarding the Mel spectrogram (a), spectrogram (b), scalogram (c), and MFCC plot (d)

DISCUSSION

In our study, the existing collection of 10-second audio recordings in WAV format representing music from Polish national dances has been altered by introducing new datasets with shorter, 3-second pieces. This enlarged the existing dataset over three times while ensuring the same class balance. Both databases were utilized for visualizing sound as a variety of techniques including Mel spectrograms, spectrograms, scalograms, and MFCC plots. They were employed as input into an introduced CNN-driven model with Adaptive Feature Fusion. This novel structure developed a multimodal network that employed a late-fusion strategy to reduce feature redundancy. Each of the above-mentioned sound visualizations represented one modality of this architecture. Moreover, for each modality, a separate branch was constructed using one of the following pre-trained classifiers such as DenseNet121, EfficientNetB0, MobileNetV2, ResNet50, VGG16, VGG19, and Xception. To perform feature fusion, a cutting-edge Adaptive Attention Mechanism was introduced containing a trunk branch, mask branch, and gate branch. AAM dynamically adjusted attention modules based on the input type, utilizing gate layers to optimize mask generation relative to feature complexity. This structure addressed the struggle of detecting subtle musical features. The proposed network was evaluated on the two distinct datasets consisting of 3- and 10-second audio recordings for the aforementioned classification methods. The gained outcomes were contrasted using testing accuracy, testing loss, precision, recall, and F1-score. Then to evaluate the impact of each modality an ablation study was prepared. Four experiments representing networks without the subsequent branches were performed. Additionally, the plot with SHAP visualizations calculated for the scenario with the highest outcomes was depicted.

It was observed that the CNN-driven model performed the best (testing accuracy 94.636% and testing loss 0.165) when EfficientNetB0 was applied as a key classifier and the larger dataset was utilized among all tested classification methods and within the ablation study's scenarios. This may be due to the distinctive compound scaling approach utilized in the classifiers from this family. Additionally, using shorter audio samples yielded generally higher efficiency regardless of the experimental setup. Acquired precision, recall, and F1-score generally prove the superiority of the proposed method among all classes. Although it may be noted that minor data imbalance could affect the overall outcome for some classes. For example, precision decreases from 86.49% in the 10-second dataset to 69.31% in the 3-second dataset, while recall improves from 78.05% to 93.57% regarding the Mazur with VGG19. This implies that shorter samples enhance the model's ability to recognize true Mazur cases but reduce precision due to more misclassifications. It is worth mentioning that the achieved testing loss is predominantly less than 1.00, or even 0.5 suggesting that models were learning effectively. In addition, depicted SHAP visualizations elucidate the model's decision. They help to recognize the areas that models struggle the most with correct

Table 12. Comparison with the state-of-the-art regarding the music identification (TA – testing accuracy in %)

Ref.	Dataset	Input	Classifier	TA
[11]	GTZAN	Mel spectrograms	ResNet24, VGG16, ResNet50, AlexNet	79.00
[19]	FMA, GTZAN, EMA	Spectrograms	ResNet50, VGG16, MobileNetV2, NASNetMobile, DenseNet121	81.00
[15]	GTZAN, 10GenreGram	Spectrograms	ResNet18, NNet2	80.00
[31]	A dataset with 10 ethnic music genres relating to ethnic music	Mel-sound spectrum, the short-time Fourier sound spectrum	A self-adjusted CNN 90.30-	92.80
[32]	A dataset with Chinese traditional folk music	Mel spectrograms	A self-adjusted CNN, ResNet18, ShuffleNet	89.00
[10]	A dataset with 2292 audio samples of Polish national dance music	Mel spectrograms	ResNet50, DenseNet121, VGG16, MobileNetV2	90.59
[28]	Ballroom, ISMIR04, GTZAN	MFCC plots, Mel spectrograms	A self-adjusted CNN	85.00
Our work	A dataset with 7782 audio samples of Polish national dance music	Mel spectrograms, spectrograms, scalograms. MFCC plots	The CNN-driven model with Adaptive Feature Fusion based on EfficientNetB, MobileNetV2, ResNet50, VGG16, VGG19, Xception	94.63

predictions in terms of music identification. They also facilitate the identification of the most suitable spectrum feature considering Polish national dance music recognition.

As a part of the research, a comparison with the state-of-the-art methods for music identification was conducted in terms of applied datasets, inputs, and classifiers (Table 12). They were evaluated in terms of acquired testing accuracy. Most studies employ song collections based on popular music genres. What is more, Mel spectrograms and spectrograms were the most prevalent inputs. However, some studies utilized MFCC plots, too. Most studies used pre-trained classifiers like DenseNet121 or ResNet50, though some introduced self-adjusting CNN architectures. Achieved testing accuracy varied between 79.00–92.80% which was lower than the one acquired in our study (94.63%). In addition, it was also superior to the research that employed Mel spectrograms for Polish national dance music classification where the highest accuracy was 90.59%. Noteworthy, compared to various state-of-the-art methods, the proposed CNN-driven model with adaptive feature fusion demonstrates promising outcomes.

CONCLUSIONS

The majority of the available research on music recognition topics is related to correctly classifying distinct genres of popular music such as pop or rock. This work includes applying formerly pre-trained classification methods like DenseNet121 or EfficientNetB0 with diverse inputs in particular Mel spectrograms, spectrograms, scalograms, and MFCC plots. Recently, a feature fusion of sound features has been addressed yielding promising effectiveness with an ablation study evaluating the impact of each feature. Furthermore, the decision-making process of a CNN has been visualized using many XAI techniques including SHAP and LIME. Despite the growing interest in folk music datasets, this area still needs to be adequately addressed. Thus, a CNN-driven model with an Adaptive Attention Module for four various inputs such as Mel spectrograms, spectrograms, scalograms, and MFCC plots has been proposed. The proposed research focused on the recognition of Polish national dance music namely the Krakowiak, the Kujawiak, the Mazur, the Oberek, and the Polonez. It evaluated the performance of distinct classifiers (DenseNet121, EfficientNetB0, MobileNetV2, ResNet50, VGG16, VGG19, Xception) that were chosen as a core structure for each modality regarding two datasets with 3- and 10-second audio samples. Their effectiveness was assessed utilizing popular classifications such as testing accuracy and testing loss while EfficientnetB0 yielded the most superficial results for the 3-second dataset with testing accuracy equal to 94.636 % and testing loss 0.165. Calculated precision, recall, and F1-score for each class represented by each of the Polish national dances such as the Krakowiak, the Kujawiak, the Mazur, the Oberek, and the Polonez also presented that the obtained results were consistent. In addition, an ablation study helped assess how each spectral feature impacted the proposed CNN-driven model. It also proved the superiority of EfficientNetB0. Additionally, the 3-second audio recordings might be a better choice since they generally yielded higher outcomes, also for the ablation study. SHAP visualizations were also generated to help understand the introduced model predictions revealing that it might be a profitable technique to assess what part of sound was considered mostly by the classifiers or which classes were supposed to be misclassified. There are several limitations, mainly due to the limited availability of data representing Polish national dance music. Overall, the conducted study improves Polish dance recognition and highlights multi-representation fusion as a promising approach for future audio classification.

Some alterations of the proposed network may be applied as a part of future works concerning Polish national dance music recognition. Firstly, replacing pre-trained classifiers applied in each branch by entailing the self-adjusted structure can be a proper solution. Secondly, it may be beneficial to apply some of the prevalent layers such as Dropout or GlobalAveragePooling2D, and evaluate their impact on the general performance of the proposed CNN-driven model. Employing some other classifiers from the EfficientNet family could also be a profitable resolution. In addition, a lack of sufficient input data can be addressed by introducing some of the data augmentation techniques employed on the images and the audio recordings by applying noise or proposing a solution that involves generative adversarial networks (GAN). This will help increase the overall size of the dataset and enable the model's performance to be contrasted with real-world scenarios. The

implemented SHAP visualizations may be contrasted with other up-to-date XAI tools including Gradient-weighted Class Activation Mapping (Grad-CAM) or LIME, too. A comparison of the CNN-driven model's efficiency on distinct folk and mainstream music collections will also be valuable in assessing model generalization.

REFERENCES

1. Schedl M, Gómez E, Urbano J. Music information retrieval: recent developments and applications. Foundations and Trends® in Information Retrieval 2014; 8: 127–261. https://doi.org/10.1561/1500000042
2. Kowol P, Nowak P, Lo Sciuto G. A control strategy for mechatronic action of a pipe organ using a VCM actuator. Electronics (Basel) 2023; 12: 4754. https://doi.org/10.3390/electronics12234754
3. Kowol P, Nowak P, Banaś W, Lo Sciuto G. Innovative design technologies of a miniaturized organ instrument. International Journal on Interactive Design and Manufacturing (IJIDeM) 2022; 16: 1551–8. https://doi.org/10.1007/s12008-022-00853-w
4. Kowol P, Nowak P, Di Nunzio L, Cardarilli GC, Capizzi G, Lo Sciuto G. Pipe organ design including the passive haptic feedback technology and measurement analysis of key displacement, pressure force and sound organ pipe. Applied System Innovation 2024; 7: 37. https://doi.org/10.3390/asi7030037
5. Yu T, Wang X, Xiao X, Yu R. Harmonizing Tradition with Technology: Using AI in Traditional Music Preservation. 2024 International Joint Conference on Neural Networks (IJCNN), IEEE; 2024; 1–8. https://doi.org/10.1109/IJCNN60899.2024.10651124
6. Puxia Li. The mediating effect of artificial intelligence on the relationship between cultural heritage preservation and opera music: A case study of Shanxi Opera. Evol Stud Imaginative Cult 2024: 249–67. https://doi.org/10.70082/esiculture.vi.880
7. Skublewska-Paszkowska M, Powroznik P, Smolka J, Milosz M, Lukasik E, Mukhamedova D, et al. Methodology of 3D scanning of intangible cultural heritage—the example of Lazgi dance. Applied Sciences 2021; 11: 11568. https://doi.org/10.3390/app112311568
8. Shevchuk O. Elements of classic choreography at the academization of Polish folk-stage dance. Global Prosperity 2023; 3: 34–41.
9. National List of Intangible Cultural Heritage n.d. https://niematerialne.nid.pl/krajowa-lista-niematerialnego-dziedzictwa-kulturowego/ (accessed June 6, 2025).
10. Chwaleba K, Wach W. Polish dance music classification based on mel spectrogram decomposition. Advances in Science and Technology Research Journal 2025; 19: 95–113. https://doi.org/10.12913/22998624/195506
11. Mehta J, Gandhi D, Thakur G, Kanani P. Music Genre Classification using Transfer Learning on log-based MEL Spectrogram. Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021 2021: 1101–7. https://doi.org/10.1109/ICCMC51019.2021.9418035
12. gtzan | TensorFlow Datasets n.d. https://www.tensorflow.org/datasets/catalog/gtzan (accessed June 6, 2025).
13. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis 2015; 115: 211–52. https://doi.org/10.1007/S11263-015-0816-Y/FIGURES/16
14. Dhall A, Srinivasa Murthy Y V., Koolagudi SG. music genre classification with convolutional neural networks and comparison with F, Q, and Mel spectrogram-based images 2021: 235–48. https://doi.org/10.1007/978-981-33-6881-1_20
15. Hassen AK, Janßen H, Assenmacher D, Preuss M, Vatolkin I. Classifying music genres using image classification neural networks. Archives of Data Science, Series A (Online First) 2018; 5: 20. https://doi.org/10.5445/KSP/1000087327/20
16. Rawat P, Bajaj M, Vats S, Sharma V. A comprehensive study based on MFCC and spectrogram for audio classification. Journal of Information and Optimization Sciences 2023; 44: 1057–74. https://doi.org/10.47974/JIOS-1431
17. Matocha M, Zielinski SKZ. Music genre recognition using convolutional neural networks. Advances in Computer Science Research 2018; 14. https://doi.org/10.24427/ACSR-2018-VOL14-0008
18. Putta N, Srinivas R, Ravi Prasad DV, Rawoof SM, Suvarna Kumari T. Audio Content Processing System for Automatic Music Classification using Mask R-CNN. 2024 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies, 3ICT 2024 2024: 196–201. https://doi.org/10.1109/3ICT64318.2024.10824338
19. Li J, Han L, Li X, Zhu J, Yuan B, Gou Z. An evaluation of deep neural network models for music classification using spectrograms. Multimed Tools Appl 2022; 81: 4621–47. https://doi.org/10.1007/S11042-020-10465-9/TABLES/6
20. Jang BY, Heo WH, Kim JH, Kwon OW. Music detection from broadcast contents using convolutional neural networks with a Mel-scale kernel. EURASIP J Audio Speech Music Process 2019; 2019: 1–12. https://doi.org/10.1186/S13636-019-0155-Y/TABLES/5
21. Powroźnik P, Czerwiński D. Spectral methods in

polish emotional speech recognition. Advances in Science and Technology Research Journal 2016; 10: 73–81. https://doi.org/10.12913/22998624/65138

22. Powroźnik P. Polish emotional speech recognition using artificial neural network. Advances in Science and Technology Research Journal 2014; 8: 24–7. https://doi.org/10.12913/22998624/562
23. Xu K, Alif M Al, He G. A novel music genre classification algorithm based on Continuous Wavelet Transform and Convolution Neural Network. ACM International Conference Proceeding Series 2021: 1269–73. https://doi.org/10.1145/3501409.3501632;PAGE:STRING:ARTICLE/CHAPTER
24. Powroznik P, Wojcicki P, Przylucki SW. Scalogram as a representation of emotional speech. IEEE Access 2021; 9: 154044–57. https://doi.org/10.1109/ACCESS.2021.3127581
25. Wang X. Research on recognition and classification of folk music based on feature extraction algorithm. Informatica 2020; 44: 521–5. https://doi.org/10.31449/INF.V44I4.3388
26. Setiorini E, Widjaja M, Wicaksana A. Reduced convolutional recurrent neural network using MFCC for music genre classification on the GTZAN Dataset. Informatica 2025; 49. https://doi.org/10.31449/INF.V49I17.6885
27. Vishnupriya S, Meenakshi K. Automatic Music Genre Classification using Convolution Neural Network. 2018 International Conference on Computer Communication and Informatics, ICCCI 2018 2018. https://doi.org/10.1109/ICCCI.2018.8441340
28. Seo W, Cho SH, Teisseyre P, Lee J. A short survey and comparison of CNN-based music genre classification using multiple spectral features. IEEE Access 2024; 12: 245–57. https://doi.org/10.1109/ACCESS.2023.3346883
29. mdeff/fma: FMA: A Dataset For Music Analysis n.d. https://github.com/mdeff/fma (accessed June 6, 2025).
30. Defferrard M, Benzi K, Vandergheynst P, Bresson X. FMA: A Dataset For Music Analysis. Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017 2016: 316–23.
31. Ning Q, Shi J. Artificial neural network for folk music style classification. Mobile Information Systems 2022; 2022: 9203420. https://doi.org/10.1155/2022/9203420
32. Mi D, Qin L. Classification system of national music rhythm spectrogram based on biological neural network. Comput Intell Neurosci 2022; 2022: 2047576. https://doi.org/10.1155/2022/2047576
33. Yang S, Li Y. Research on Music Feature Extraction and Machine Learning Classification Algorithm for Anhui Folk Songs. 2025 International Conference on Digital Analysis and Processing, Intelligent Computation (DAPIC) 2025: 255–60. https://doi.org/10.1109/DAPIC66097.2025.00053
34. Mirza FK, Gürsoy AF, Baykaş T, Hekimoğlu M, Pekcan Ö. Residual LSTM neural network for time dependent consecutive pitch string recognition from spectrograms: a study on Turkish classical music makams. Multimed Tools Appl 2024; 83: 41243–71. https://doi.org/10.1007/S11042-023-17105-Y/TABLES/2
35. Papaioannou C, Valiantzas I, Giannakopoulos T, Kaliakatsos-Papakostas M, Potamianos A. A dataset for Greek traditional and folk music: Lyra n.d. https://doi.org/https://doi.org/10.48550/arXiv.2211.11479
36. 3Patel A, Shah A, Gor K, Mankad SH. IFSC: A Database for Indian Folk Songs Classification 2021:171–86. https://doi.org/10.1007/978-981-33-6881-1_15
37. Puttegowda K, Keoy KH, Deepak R, Armoogum V, Parameshachari BD. Automated Music Classification using Machine Learning for Indian Songs. 2nd IEEE International Conference on Networks, Multimedia and Information Technology, NMITCON 2024. https://doi.org/10.1109/NMITCON62075.2024.10698871
38. Kalapatapu P, Lakshmi Sravani J, Gupta S, Sharma A, Malapati A. Genre Classification using Spectrograms as input to CNN on Indian Music. Proceedings of the 2021 International Conference on Emerging Techniques in Computational Intelligence, ICETCI 2021 2021: 12–5. https://doi.org/10.1109/ICETCI51973.2021.9574060
39. Kiss A, Sulyok C, Bodó Z. Region Prediction from Hungarian Folk Music Using Convolutional Neural Networks. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2019; 11730 LNCS: 581–94. https://doi.org/10.1007/978-3-030-30490-4_47/FIGURES/4
40. Thanh CB, Loan T Van, Thuy DT Le. Automatic identification of some Vietnamese folk songs Cheo and Quanho using convolutional neural networks. Journal of Computer Science and Cybernetics 2022; 38: 63–84. https://doi.org/10.15625/1813-9663/38/1/15961
41. Bora K, Pratim Barman M, Patowary AN, Bora T. Indian journal of science and technology classification of Assamese folk songs' melody using supervised learning techniques. Indian Journal of Science and Technology 2023; 16: 89–96. https://doi.org/10.17485/IJST/v16i2.1686
42. Riad; SGF. Attention-based CNN-BiGRU for Bengali music emotion classification. International Journal of Computer Science & Network Security 2023; 23: 47–54. https://doi.org/10.22937/IJCSNS.2023.23.9.6
43. Navarro-Cáceres JJ, Carvalho N, Bernardes G,

Jiménez-Bravo DM, Navarro-Cáceres M. Exploring Mode Identification in Irish Folk Music with Unsupervised Machine Learning and Template-Based Techniques. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2024; 14639 LNCS: 412–20. https://doi.org/10.1007/978-3-031-60638-0_34

44. Han D, Repetto RC, Jeong D. Finding Tori: Self-supervised Learning for Analyzing Korean Folk Song 2023.
45. Almazaydeh L, Atiewi S, Al Tawil A, Elleithy K. Arabic music genre classification using deep convolutional neural networks (CNNs). Computers, Materials and Continua 2022; 72: 5443–58. https://doi.org/10.32604/CMC.2022.025526
46. 4Folorunso SO, Afolabi SA, Owodeyi AB. Dissecting the genre of Nigerian music with machine learning models. Journal of King Saud University - Computer and Information Sciences 2022; 34: 6266–79. https://doi.org/10.1016/J.JKSUCI.2021.07.009
47. Guo MH, Xu TX, Liu JJ, Liu ZN, Jiang PT, Mu TJ, et al. Attention mechanisms in computer vision: A survey. Comput Vis Media (Beijing) 2022; 8: 331–68. https://doi.org/10.1007/S41095-022-0271-Y/METRICS
48. Sun J, Jiang J, Liu Y. An Introductory Survey on Attention Mechanisms in Computer Vision Problems. Proceedings - 2020 6th International Conference on Big Data and Information Analytics, BigDIA 2020 2020: 295–300. https://doi.org/10.1109/BIGDIA51454.2020.00054
49. Setyoko DESDE. Resnet-18 with attention mechanism-bidirectional LSTM hybrid approach for music genre classification using stacking MFCC and Mel-spectogram features. The Indonesian Journal of Computer Science 2024;13. https://doi.org/10.33022/IJCS.V13I6.4456
50. Wen Z, Chen A, Zhou G, Yi J, Peng W. Parallel attention of representation global time–frequency correlation for music genre classification. Multimed Tools Appl 2024; 83: 10211–31. https://doi.org/10.1007/S11042-023-16024-2/TABLES/7
51. Lin Y-X, Lin J-C, Wei W-L, Wang J-C. Learnable counterfactual attention for music classification. IEEE Trans Audio Speech Lang Process 2025; 33: 570–85. https://doi.org/10.1109/TASLPRO.2025.3527143
52. Xie C, Song H, Zhu H, Mi K, Li Z, Zhang Y, et al. Music genre classification based on res-gated CNN and attention mechanism. Multimed Tools Appl 2024; 83: 13527–42. https://doi.org/10.1007/S11042-023-15277-1
53. Nguyen QH, Do TTT, Chu TB, Trinh L V., Nguyen DH, Phan C V., et al. Music Genre Classification using Residual Attention Network. Proceedings of 2019 International Conference on System Science and Engineering, ICSSE 2019 2019: 115–9. https://doi.org/10.1109/ICSSE.2019.8823100
54. Ng WWY, Zeng W, Wang T. Multi-level local feature coding fusion for music genre recognition. IEEE Access 2020;8:152713–27. https://doi.org/10.1109/ACCESS.2020.3017661
55. Ajay A, Rajan R. Music genre classification using attention-based CNN-feature fusion paradigm. 2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems, AICERA/ICIS 2023 2023. https://doi.org/10.1109/AICERA/ICIS59538.2023.10420267
56. Kozieł G, Harasim D, Dziuba-Kozieł M, Kisała P. Fourier transform usage to analyse data of polarisation plane rotation measurement with a TFBG sensor. Metrology and Measurement Systems 2024; 31: 369–81. https://doi.org/10.24425/MMS.2024.149698
57. Cho SH, Park Y, Lee J. Effective Music Genre Classification using Late Fusion Convolutional Neural Network with Multiple Spectral Features. 2022 IEEE International Conference on Consumer Electronics-Asia, ICCE-Asia 2022 2022. https://doi.org/10.1109/ICCE-ASIA57006.2022.9954732
58. Murindanyi S, Hamza K, Kagumire S, Marvin G. Responsible music genre classification using interpretable model-agnostic visual explainers. SN Comput Sci 2025; 6: 1–26. https://doi.org/10.1007/S42979-024-03584-9/FIGURES/17
59. Murindanyi S, Nakate A, Kyebambe MN, Nakibuule R, Marvin G. Responsible Artificial Intelligence for Music Recommendation. Lecture Notes in Networks and Systems 2024; 818: 291–306. https://doi.org/10.1007/978-981-99-7862-5_22
60. Hasib KM, Tanzim A, Shin J, Faruk KO, Mahmud J Al, Mridha MF. BMNet-5: A novel approach of neural network to classify the genre of bengali music based on audio features. IEEE Access 2022; 10: 108545–63. https://doi.org/10.1109/ACCESS.2022.3213818
61. Giannakopoulos T. pyAudioAnalysis: An open-source python library for audio signal analysis. PLoS One 2015;10:e0144610. https://doi.org/10.1371/journal.pone.0144610
62. Zhang W, Lei W, Xu X, Xing X. Improved music genre classification with convolutional neural networks 2016. https://doi.org/10.21437/Interspeech.2016-1236
63. Ceylan HC, Hardalaç N, Kara AC, Hardalaç F. Automatic music genre classification and its relation with music education. World Journal of Education 2021; 11: 36. https://doi.org/10.5430/WJE.V11N2P36
64. Borisagar KR, Thanki RM, Sedani BS. Fourier transform, short-time Fourier transform, and wavelet transform. Speech Enhancement Techniques

for Digital Hearing Aids 2019: 63–74. https://doi.org/10.1007/978-3-319-96821-6_4

65. Ortiz-Echeverri CJ, Rodríguez-Reséndiz J, Garduño-Aparicio M. An approach to STFT and CWT learning through music hands-on labs. Computer Applications in Engineering Education 2018; 26: 2026–35. https://doi.org/10.1002/CAE.21967;WGROUP:STRING:PUBLICATION
66. Mika D, Józwik J. Advanced time-frequency representation in voice signal analysis. Advances in Science and Technology Research Journal 2018; 12: 251–9. https://doi.org/10.12913/22998624/87028
67. Kumar N, Kumar R. Wavelet transform-based multipitch estimation in polyphonic music. Heliyon 2020;6:e03243. https://doi.org/10.1016/J.HELIYON.2020.E03243
68. Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE; 2017; 1800–7. https://doi.org/10.1109/CVPR.2017.195
69. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition 2015.
70. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE; 2016; 770–8. https://doi.org/10.1109/CVPR.2016.90
71. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks 2018: 4510–20.
72. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE; 2017; 2261–9. https://doi.org/10.1109/CVPR.2017.243
73. Tan M, Le Q V. EfficientNet: Rethinking model scaling for convolutional neural networks. Appl Math (Irvine) 2019; 11:1204–18.
74. Skublewska-Paszkowska M, Powroznik P, Barszcz M, Dziedzic K, Aristodou A. Identifying and animating movement of Zeibekiko sequences by spatial temporal graph convolutional network with multi attention modules. Advances in Science and Technology Research Journal 2024; 18: 217–27. https://doi.org/10.12913/22998624/193180
75. Correia M, José L, Machado R, Skublewska-Paszkowska M, Powroznik P. Temporal pattern attention for multivariate time series of tennis strokes classification. Sensors 2023; 23: 2422. https://doi.org/10.3390/S23052422
76. Skublewska-Paszkowska M, Powroznik P, Lukasik E. Learning three dimensional tennis shots using graph convolutional networks. Sensors 2020; 20: 6094. https://doi.org/10.3390/S20216094
77. Terven J, Cordova-Esparza DM, Romero-González JA, Ramírez-Pedraza A, Chávez-Urbiola EA. Loss Functions and Metrics in Deep Learning. Artif Intell Rev 2023; 58. https://doi.org/10.1007/s10462-025-11198-7
78. Zhang J. Music feature extraction and classification algorithm based on deep learning. Sci Program 2021; 2021: 1651560. https://doi.org/10.1155/2021/1651560
79. Mao Y, Zhong G, Wang H, Huang K. MusicCRN: an efficient content-based music classification and recommendation network. Cognit Comput 2022; 14: 2306–16. https://doi.org/10.1007/S12559-022-10039-X/TABLES/8
80. Zhang H, Karystinaios E, Dixon S, Widmer G, Cancino-Chacón CE. Symbolic Music Representations for Classification Tasks: A Systematic Evaluation 2023.