Advances in Science and Technology Research Journal

Advances in Science and Technology Research Journal, 2025, 19(2), 95–113
https://doi.org/10.12913/22998624/195506
ISSN 2299-8624, License CC-BY 4.0

Received: 2024.08.21
Accepted: 2024.12.15
Published: 2025.01.01

Polish dance music classification based on mel spectrogram decomposition

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

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

ABSTRACT

Folk dances and music are essential aspects of intangible cultural heritage identifying the history and traditions of nations. Due to dynamic changes in the social structure, many national aspects are not cultivated and therefore forgotten. There is a need to develop methods to preserve these valuable aspects of culture. There are five Polish national dances: the Polonez, the Oberek, the Mazur, the Krakowiak, and the Kujawiak that reflect key elements of Polish intangible cultural heritage. They can be observed both in the way of performing dances as well as in music. There are many preserved audio and video files that differ depending on the multiple features such as composers or versions. The primary objective of this study was to apply machine learning approaches in order to distinguish the above-mentioned music of Polish traditional dances. The audio recordings dataset consisting of 137 dances in mp3 format was created. Each recording was divided into ten-second files reflecting the characteristic elements of each dance. The transformation of sound to the Mel scale improves human auditory perception. Thus, from every recording the Mel-spectrograms were generated. For the purpose of this study the most applied classification tools were compared such as VGG16, ResNet50, DenseNet121, and MobileNetV2. To compare the performance of the selected models, the following measures were applied: accuracy, precision, recall, and F1 score. ResNet50 achieved the best testing accuracy (over 90%), while DenseNet121 had the best testing loss (0.38).

Keywords: Mel-spectrograms, folk music classification, Polish national dances recognition, VGG16, ResNet50, DenseNet121, MobileNetV2.

INTRODUCTION

Dances have been present in the culture of almost every nation for ages reflecting its history and traditions [1]. These arts are key elements of their intangible cultural heritage identifying communities, preserving their cultural diversity, and serving as a fundamental link for intergenerational knowledge transfer. Polish national dances have been officially a part of the Polish national list of intangible cultural heritage since 2015 with five dances: the Polonez, the Mazurek, the Krakowiak, the Kujawiak, and the Oberek [2]. Each dance presents its individual traits and characteristics that distinguish it from each other [3]. Dances are an integral part of Polish folklore among other elements such as music and singing. Nowadays, folk dancing is vanishing – the transmission of dance styles across generations is fragile, even unpredictable. For this reason, there is a great need to preserve it for future generations [4]. Additionally, music serves as an integral part of Polish National Dances retaining various features such as rhythm.

Artificial intelligence (AI) offers solutions to the challenges of preserving musical heritage [5, 6]. By facilitating large-scale digitization and analysis of traditional music, AI makes it accessible to a wider audience and promotes deeper understanding [7, 8, 9]. Many studies have shown that machine learning algorithms can be applied to the analysis of musical patterns, enabling more effective documentation and classification of different musical styles and techniques [10]. Natural

language processing aids in transcribing lyrics, particularly in endangered languages, thereby contributing to both language preservation and the safeguarding of musical traditions. AI-powered applications can create interactive platforms that engage users in learning about traditional music, enhancing education and appreciation.

Motivation of the study

The main motivation for undertaking this research was the urgent need to preserve and promote Polish intangible cultural heritage by nurturing the essence of Polish national dance music. In the face of globalization and rapid technological advancement, traditional forms of music are often at risk of marginalization and oblivion. This concern is heightened by the significant similarities between Polish national dance music and that of other Slavic nations, which can lead to confusion and the blending of distinct musical identities. Polish national dance music, being a living testimony of our history, traditions, and national identity, deserves special attention and protection.

Therefore, the aim of this article was to apply modern classification methods based on machine learning to recognize the music of Polish national dances. Classifying these musical pieces is particularly urgent due to their resemblance to music from other Slavic national groups. By developing accurate classification systems, we can distinguish and highlight the unique features of Polish national dance music, ensuring it is correctly identified and appreciated.

Utilizing advanced technologies in cultural heritage research allows not only for documentation and analysis but also for active promotion and adaptation in a changing world. By integrating innovative tools with traditional forms of music, we strive to prevent their decline and strengthen their presence in social consciousness. We believe that such an approach will contribute to revitalizing interest in Polish national dance music, both among the younger generation and on the international stage, thereby supporting the continuity and development of our intangible cultural heritage.

Scientific novelty of the proposed work

In this study the comprehensive analysis of up-to-date machine learning methods and their ability to recognize Polish national dance music is performed. Thus, the main contribution of this study is as follow: A unique dataset consisting of music from five Polish national dances was created. The dataset consisted of 2296 10-second WAW files.

The study methodology consisted of a five-stage process. MP3 audio recordings of Polish national dance music were collected, converted into WAV format, and segmented into 10-second samples. Then, every audio recording was transformed into a Mel-spectrogram. The created Mel-spectrograms were split into training, validation, and testing datasets, with an 80%, 10%, and 10% division, respectively. Each classification method was trained and assessed. The performance of each classifier was compared and evaluated.

Recognition of 5 national Polish dances (the Krakowiak, the Kujawiak, the Mazur, the Oberek and the Polonez) using architectures such as VGG16, ResNet50, DenseNet121 and MobileNetV2. Recognition of 5 national Polish dances (the Krakowiak, the Kujawiak, the Mazur, the Oberek, and the Polonez) using architectures such as VGG16, ResNet50, DenseNet121, and MobileNetV2. Each classification method is characterized by its distinctive features: VGG16 utilizes 3x3 convolution filters, ResNet50 incorporates shortcut connections and residual blocks, DenseNet121 connects densely each layer to every subsequent layer, and MobileNetV2 leverages architecture based on the inverted residual with a linear bottleneck and depthwise convolutions, which enables to achieve better performance.

Evaluation the most adequate deep learning method for preserving Polish national dances utilizing accuracy, precision, recall and F1-score measures. In this study, the following neural networks models were applied: VGG16, ResNet50, DenseNet121 and MobileNetV2. What is more, a unique dataset consisting of music from five Polish national dances was created independently.

The rest of the paper is organized as follows. Related Works presents a review of current up-to-date methods of music classification techniques and studies that focus on folk music. Materials and methods demonstrate the process of data gathering and preprocessing and explain utilized classification methods and metrics. Results reveal the obtained outcomes for each classifier and display summarized results. The discussion compares obtained performances with state-of-the-art studies. Conclusions and future works consist of final afterthoughts and further studies on following Polish national dance music recognition.

RELATED WORKS

Music Information Retrieval (MIR) is an area of study that mainly focuses on extracting and inferring meaningful features from music [11]. Moreover, it includes indexing music utilizing these features and constructing search as well as retrieval systems. One of the most essential subfields is sound classification which has been explored in many scientific papers [12–16].

In [12] it was noticed that previously a feature-based approach was predominantly applied in music classification. Thus, the researchers decided to employ Mel-spectrograms and utilize them as input to the Convolution Neural Networks (CNNs). The GTZAN dataset [17] was the study material consisting of audio samples belonging to 10 classes. Each class represented a diverse musical genre, such as classical, blues, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. Each file was 30 seconds long. The sampling rate of 44,100 Hz was applied during the conversion of audio files into spectrograms. Ultimately, the size of images set for 224×224 was utilized in pretrained ImageNet models. In this study, data augmentation techniques were also considered as key elements in improving the size of the final dataset – the spectrograms were segmented into different pieces in a vertical cut. Then, the most prevalent classification methods such as ResNet24, ResNet50, VGG16, and AlexNet were evaluated by the testing accuracy, loss, and confusion matrices. The highest accuracy achieved by ResNet24 was 79%. The accuracy of other classifiers could be listed as follows: 76.93% for VGG16, 75.9% for ResNet50, and 71.3% for AlexNet.

Furthermore, spectrograms were also employed as study materials in [13]. They were generated from three datasets: the FMA [18, 19], the GTZAN, and the EMA. To obtain consistent results files were converted from MP3 format to WAV format, and audio samples longer than 5 seconds were split into 5-second pieces. Then, a short-time Fourier transform function was applied to generate spectrograms. Moreover, the sampling rate was reconstructed from 44,100 Hz to 22,050 Hz utilizing the Nyquist-Shannon sampling theorem. Generally, it managed to gather almost 22,000 images of spectrograms. Five state-of-the-art classification models such as DenseNet121, NASNetMobile, MobileNetV2, VGGNet16, and ResNet50 were compared on various datasets using accuracy as a metric. To acquire more reliable results, it was decided to undertake experiments on every model and every dataset five times, and then calculate the average as a final accuracy. ResNet50 managed to obtain the highest score which was over 77–81%, the second was VGG16 with 72–79% accuracy. One of the most prevalent classifiers, such as AlexNet, and LeNet-5, were also examined in [14] with the GTZAN dataset as a study material. Furthermore, three transformations were employed to generate the spectrograms: Fourier transform-based (FS), Q transform-based (QS), and Mel-frequency transform-based (MS) spectrograms. Both classification methods managed to acquire a satisfactory level of accuracy. In addition, ResNet18 and NNet2 were opted for classification methods in [20]. The GTZAN dataset and the 10GenreGram were utilized once again with the standard parameters. The second dataset was self-created as a part of the research. It consisted of audio recordings saved from a popular streaming platform called SoundCloud. Afterward, 3-second spectrograms were created and utilized in evaluated classifiers. For the GTZAN dataset, it was determined to divide it using the following ratio – 80% for training, 10% for testing, and 10% for the validation set. Then, utilizing tree-structured Parzen estimators on the GTZAN dataset, with 30 epochs and 50 evaluations, optimal hyperparameters (like learning rate and activation function) were selected for the ResNet model. These model parameters were then transferred to the 10GenreGram dataset and the performance of both classifiers was assessed using accuracy, loss, and confusion matrix. All in all, appropriate results were achieved. Image classification models performed well on both datasets compared to related works, and its architecture was less susceptible to overfitting.

In [21] authors compared the application of Mel-spectrograms utilized by the CNN classifier with the Mel-frequency Cepstral Coefficients (MFCC) features employed in training and Artificial Neural Network model to classify audio recording from the obtained dataset. The analyzed data collection incorporates 6,000 audio samples from 10 balanced music genres. Each recording was a 30-second long WAV file. The evaluated approaches were compared using a binary classification, as well as a multiclass classification for 5 and 10 classes utilizing accuracy and loss. In every case, the CNN with Mel-spectrograms method obtained more acceptable results. Another study also reveals that Mel-spectrograms could

be applied in the area of music classification and recognition [22], while images of Mel-spectrograms could be treated as input in the CNN model. The GTZAN dataset was utilized as study material among other datasets such as FMA-small and JUNO. As a classification method a MapReduce Based Deep Convolutional Neural Network (MR-DCNN) model was proposed which should overcome some limitations related to traditional music recommendation systems. After generating Mel-spectrograms, researchers chose an 8 to 2 ratio for the training and testing set. Then, obtained metrics such as accuracy, the Mean Absolute Error (MAE) value, and Root Mean Square Error (RMSE) were compared for each music dataset. It could be stated that acquired accuracies such as 63%, 78%, and 89.7% for the respective datasets: FMA, JUNO, and GTZAN, were high enough to depict that CNN and Mel-spectrogram could be combined as a sufficient method for music classification. Mel-spectrograms and the CNNs model were also applied in [23] where they were employed in music genre recognition. As a study material the FMA dataset was utilized. Consisting of 8,000 30-second audio samples in eight following genres, such as music, experimental, folk, hip-hop, instrumental, international, pop, and rock. As a part of the experiment, Mel-spectrograms were generated. For the study, the self-developed CNN classifier was implemented and contrasted with a classic approach dependent on audio metrics and support vector machines. Both techniques were characterized by a similar level of overall performance when the Receiver Operating Curves (ROC) were compared. However, confusion matrices depicted that the prediction accuracy varied between classes. In the final analysis, it was discerned that the proposed classification method employed by the CNNs could exhibit notable results, and might be valuable in music recognition.

In [24–26], the main objective was the application of Mel-spectrograms and a Mel-scale kernel in detection of such elements as speech, singer's voice or music in the noisy audio recordings. Based on these studies, it could be stated that Mel-spectrograms might be employed as a successful approach in feature extractions and might serve well as input in the classification models based on convolutional neural networks.

Although the majority of studies focus on Western types of music, there are several academic works where data material is strictly linked to folk music. In [27], a dataset consisting of 10 genres related to ethnic music was created. Each category has 100 recordings which last 30 seconds. 22,050 Hz was chosen as a sampling rate. The following ratios were chosen to split their data: 80% for the training dataset, 10% for validation, and 10% for testing. It was also ensured to get balanced categories for each dataset. In addition, the tenfold cross-validation was utilized for this division. As a classification method, a self-adjusted convolutional neural network was applied. Such factors as various convolution structures, Global Pooling Feature Aggregation, Audio Segmentation, and Music Data Enhancement were compared using accuracy. Overall, the Mel-sound spectrum attained higher accuracy (92.8%) over the short-time Fourier sound spectrum (90.3%).

In [28] a dataset consisting of 6 types of folk music was created. Each file was in WAV format and as a part of data processing, they were sliced into 10-second segments. Additionally, a sampling frequency of 16 KHz was chosen. As a part of the research, 13-dimensional Mel-frequency cepstral coefficient features and 4-dimensional features were extracted, then the average value and the standard deviation were considered so that each segment managed to obtain 36-dimensional features. For the training dataset 80% of features were selected, and 20% for testing. In the experiment, Back Propagation Neural Network (BPNN), decision tree, and Support Vector Machine (SVM) were compared in terms of recognition rate. It was noticed that SVM obtained a higher result (over 92%) compared to 73% for BPNN and 64% for decision tree.

Chinese national music was a crucial study element in [29]. Gathering data was an important stage of the survey. It consisted of crawling the Internet in order to build the dataset of Chinese traditional folk music. Then every audio sample was splitted into 30 seconds long recording, in WAV format, and with a 16,000 sampling rate. Eventually 2608 Mel-spectrogram images were gathered. The self-adjusted CNN model was utilized as a classification method and it was compared using accuracy, sensitivity, and specificity with widely known pre-trained models such as ResNet18, and ShuffleNet. Every method was distinguished by obtaining favorable outcomes, including an accuracy of over 89%.

The article [30] tries to recognize 'Makam' which is part of Turkish classical music and could

be characterized as a peculiar melodic configuration defined by sequences of pitches and intervals. An extensive dataset comprising 1,154 samples with 15 rare Makams was created. Each Makam was converted to an audio file in WAV form, and 6-second samples were employed to generate logarithmic scale spectrograms for input to the classifier. The final dataset was divided into 9 to 1 ratios to create training and validation sets. During the study, it was discovered that a spectrogram classification diverged from each other since spectrograms involved certain patterns evolving across time and frequency dimensions. Consequently, the self-adjusted residual long short-term memory (LSTM) neural network model was utilized, which combined the spatial capabilities of two-dimensional convolutional layers and temporal capabilities of one-dimensional convolutional and LSTM mechanisms. As evaluation metrics, accuracy, precision, recall, and F1 score were accessed and compared. Additionally, confusion matrices were generated. All in all, the model demonstrated an appropriate accuracy: almost 90% for the dataset consisting of 15 Makams, and over 95% for 9 Makams.

A great number of examples of the utilization of artificial intelligence models, especially CNNs, can be found in various contexts of the tangible and intangible culture heritage dedicated for protection and preservation of valuable arts. Machine learning models are used to recognize texts, images, or architectural styles [9]. They are also applied in image segmentation to localize important features, e.g. architectural ones [31]. Models are widely used to recognize folk dances, music, and speech. An interesting research aspect is also the performance of 3D reconstruction of dance costumes [32]. The ResNet50 classifier was utilized to perform heritage image classification containing ancient artifacts and buildings [33]. Evaluated classifiers presented a high level of testing accuracy after fine-tuning for ResNet50 – 98.58%. Classification methods such as DenseNet121, VGG16, MobileNet, and ResNet50 were employed to perform feature extraction for identifying archeological sites, frescoes, and monasteries [34]. They managed to be beneficial in obtaining sufficient accuracies around 71–84%. VGG16 classified was also employed in heritage coin identification and categorization obtaining over 90% accuracy [35]. Pre-trained classification methods are also used to preserve the Intangible Cultural Heritage. ResNet architecture was employed in automatic recognition of Greek folk dances to obtain over 80% accuracy [36]. Indian Classical Dance Forms classification was performed utilizing ResNet50 and surpassed current state-of-the-art approaches achieving an accuracy score of 91.1% [37].

It might be noticed that the assessed literature review reveals that recent music classification employing Mel-spectrograms and the CNN models is a popular study subject. Nonetheless, it focuses mainly on genre classification using current mainstream music. However, only a few studies focused on folk music. According to the authors' knowledge, there are no studies referencing Polish national music, which is a crucial element of Polish intangible cultural heritage. Therefore, the conducted study presented in this paper is innovative in terms of a self-created dataset containing Polish national dance music, applying Mel-spectrograms for traditional Polish music classification.

MATERIALS AND METHODS

In this section, the proposal for Polish national dance music classification is explored. It was introduced following a previous literature review which revealed some common aspects based on topics related to Music Information Retrieval. Figure 1 presents five steps of the methodology: collecting Polish national dance music audio recordings in MP3 format, converting them into WAV format, slicing them into smaller 10-second pieces, generating Mel-spectrogram pictures, and the application of chosen classification methods.

Dataset

Audio recordings consisting of Polish national dance music were collected from online

Figure 1. The study methodology

sources such as YouTube. They were reviewed and some redundant segments, such as conversations or silences, were reduced manually. Every audio sample was converted from MP3 to WAV format to facilitate further music processing. Then the dataset consisting of 137 files was obtained. Every audio sample was read using 22050 Hz as a sampling rate and then they were cut into smaller 10-second pieces. It is important to mention that every last piece of every audio sample was removed if it was shorter than 10 seconds. That is why, it was possible to achieve a dataset consisting of 2296 audio recordings in WAV format. The final number of attained files is presented in Table 1. The final dataset consists of 5 classes that represent each Polish national dance. It could be observed that the examined dataset is unbalanced: the Kujawiak is the most numerous set, and the Mazur is the least.

Data preprocessing

Audio spectrograms are state-of-the-art methods implemented in analyzing sound as they offer valuable information about the time-frequency characteristics of a music [38].

Humans tend to detect lower frequency variations faster compared to high ones due to the non-linear perception of the frequency [39]. Applying Mel-scale and Mel-spectrograms could help display and read audio signals. A Mel-spectrogram is a spectrogram scaling frequency to the Mel scale [40, 41] where time is on the x-axis, and Mel-frequency bins are on the y-axis [42]. Mel-spectrograms were generated for every 10-second audio sample. Then Mel-spectrograms were saved as images, and used as input into CNN classifiers. An example Mel-spectrogram created for the Krakowiak fragment is shown in Figure 2.

As a final stage of data preprocessing the obtained Mel-spectrograms were split into directories which could be leveraged into Polish national dance music recognition. Based on the study research it was decided to employ the following ratios when splitting the dataset into train, validation, and test datasets: 0.8, 0.1, and 0.1.

Table 1. Total number of utilized audio samples

Dance	Number of audio files before cutting	Number of 10 second samples
Krakowiak	23	444
Kujawiak	34	588
Mazur	25	410
Oberek	38	428
Polonez	17	426
Overall	137	2296

Classification methods

After a comprehensive study, it was decided to select the most up to date classifiers nowadays such as VGG16, ResNet50, DenseNet121, and MobileNetV2 because they have been widely used in studies where Mel-spectrograms were fed forward as an input to the CNNs.

VGG was proposed as a novel architecture for commonly utilized convolutional networks (ConvNets) which helped add more convolutional

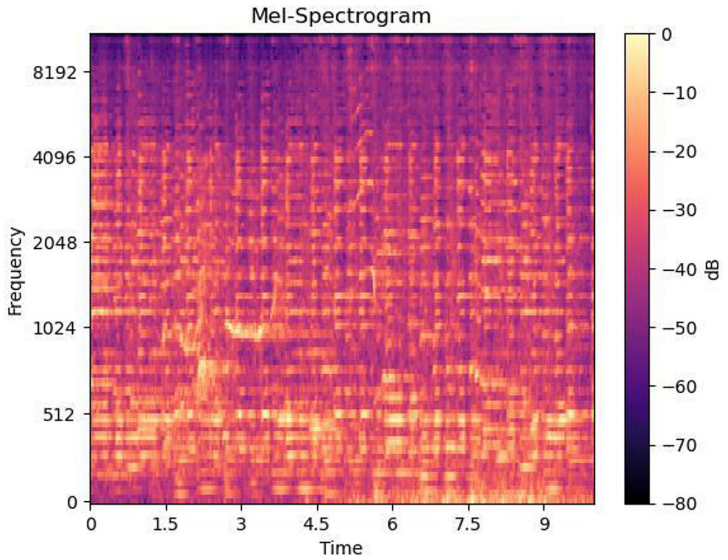

Figure 2. The Mel-spectrogram generated for the Krakowiak sample

layers to achieve a deeper network [43]. It was possible due to applying tiny (3x3) convolution filters in all layers. It resulted in accomplishing excellent accuracy in ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) classification and localization tasks [44]. In [43] various configurations of the ConvNets were evaluated that differed in depth. The 11, 13, 16, and 19 weight layers were assessed.

Residual Neural Networks (ResNets) were introduced as a solution to the degradation problem that was encountered. It was found that when the architecture of a model became deeper and accuracy saturated with increasing depth it deteriorated quickly [45]. The proposed architecture introduced shortcut connections and the residual block which enabled the training of deeper networks without facing the previously mentioned issue. Thus it is feasible to train CNNs consisting of many more layers. As a part of the study [45] ResNets with 18, 34, 50, 101, and 152 layers were evaluated. It was proved that ResNet50/101/152 outperformed ResNet18 and ResNet34 in terms of accuracy. Moreover, despite being deeper than VGG16/19, these networks still had lower complexity.

Dense Connected Convolutional Networks (DenseNets) have been developed as a response to increasing depth of CNNs such as ResNets which could encounter the vanishing gradient issue [46]. It proposed an innovative architecture establishing connections between each layer in a feed-forward way, not utilizing identity connections. The feature-maps of all proceeding layers were employed as inputs for every layer and its feature-maps were utilized as inputs for all subsequent layers. That is why there are $\frac{L(L+2)}{2}$ connections in the network with L layers, not L as it used to be previously. This resulted in a network that could be characterized by its dense connectivity. DenseNets help with the vanishing-gradient issue, minimizing the number of parameters, and feature propagation. They managed to outperform known CNNs such as ResNets, yet needed less computation. DenseNet121, DenseNet169, DenseNet201, and DenseNet264 were introduced as various DenseNet architectures.

The architecture of the MobileNetV2 network relies on a novel structure: the inverted residual with a linear bottleneck [47]. What is more, filtering features in the intermediate expansion layer are utilized by the lightweight depthwise convolutions and elimination of non-linearities in the narrow layers. It helped maintain the same level of accuracy while reducing the number of operations and memory requirements.

Overall classification methods were selected due to their unique features which could help handle Polish national dance music classification well such as:

1. VGG16 employs 3x3 convolution filters which enable higher accuracy [43].
2. ResNet50 utilizes shortcut connections and residual blocks which allows for training deeper networks without the degradation issue [45]. It also performs better compared to VGG16.
3. DenseNet121 tackles the vanishing gradient problem by densely connecting each layer to every subsequent layer, improving feature propagation, reducing parameters, and surpassing ResNets in performance [46].
4. A MobileNetV2 innovative architecture based on the inverted residual with a linear bottleneck and depthwise convolutions reduces memory usage while maintaining accuracy [47].

Pre-trained networks were used for training. The input data were pre-processed, adapting them to the requirements of specific classifiers. Additionally, the architecture of each model has been individually adjusted by adding several Flatten and Dense layers. Then each model was trained and 50 epochs were chosen based on similar studies [20, 23, 24]. However, EarlyStopping callback was implemented to prevent overfitting. If the validation error is not improved in the next 5 epochs, the training process will be interrupted.

Classification metrics

The aforementioned classifiers' performance were evaluated based on popular metrics used in former studies such as accuracy (1), precision (2), recall (3) and F1-score (4) as a harmonic mean of precision and recall [48, 49]:

$$Accuracy = \frac{Number\ of\ correct\ classifications}{Total\ number\ of\ classifications} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1 = \frac{2\times Precision\times Recall}{Precision\times Recall} \quad (4)$$

where: TP indicates the true positive fraction, FP – the false positive fraction, and FN – the false negative fraction.

Usually, metrics such as precision, recall, and F1-score are calculated for binary classification problems, however they could be quickly extended to multiple classes problems [50]. In other words, for instance, the Kujawiak precision could be described as correctly classified Kujawiak Mel-spectrograms for every classified Kujawiak Mel-spectrogram. Moreover, recall could be characterized as the number of successfully predicted Kujawiak Mel-spectrograms out of all input Kujawiak Mel-spectrograms.

The confusion matrices were also presented in order to assess the effectiveness of chosen classification methods. The confusion matrix is a table where rows picture the actual (true) class, and columns picture the class produced by the classifier. The correctly predicted instances are represented in the diagonal of the matrix [51]. In this study, the rows contain actual dance classes, whereas predicted dances are in the columns. The diagonal contains the correctly predicted dances Mel-spectrograms.

In addition, the loss using the Categorical Cross Entropy loss was also calculated. It is a common function used for multiclass classification tasks. It helps assess the discrepancy between the predicted probability distribution and the actual distribution [52].

RESULTS

The obtained results were presented for every classification method independently, while testing accuracy and testing loss were presented in a grouped form for all classifiers.

VGG16

Figure 3 depicts the achieved results for the VGG16 classifier: training accuracy compared to validation accuracy in the left plot and training loss versus validation loss in the right plot. It could be stated that the learning process stops on 12 epochs as validation loss has not improved. Training accuracy starts low from 0, although it starts growing rapidly, reaches almost 1 within a few epochs, and then plateaus. Conversely, validation accuracy starts quite high at almost 0.8, drops, and starts growing slowly till it becomes stable around 0.9. A huge gap between training and validation accuracy could be observed. Training loss almost immediately drops and reaches a plateau near 0.0 while validation loss is higher and stagnates around 1.2 value. There is also some gap between validation and training loss.

Some other examined metrics such as precision, recall, and F1-score are presented in Table 2. The best precision was reached by the Mel-Spectrograms representing the Krakowiak dance, while the best recall – by Mel-spectrograms representing the Kujawiak dance. This might indicate that the VGG16 classification method performs accurately concerning the prediction of the Krakowiak although it could miss some actual instances of it, and it might well identify the Kujawiak instances. On the other hand, it could be stated that the Kujawiak has the worst precision

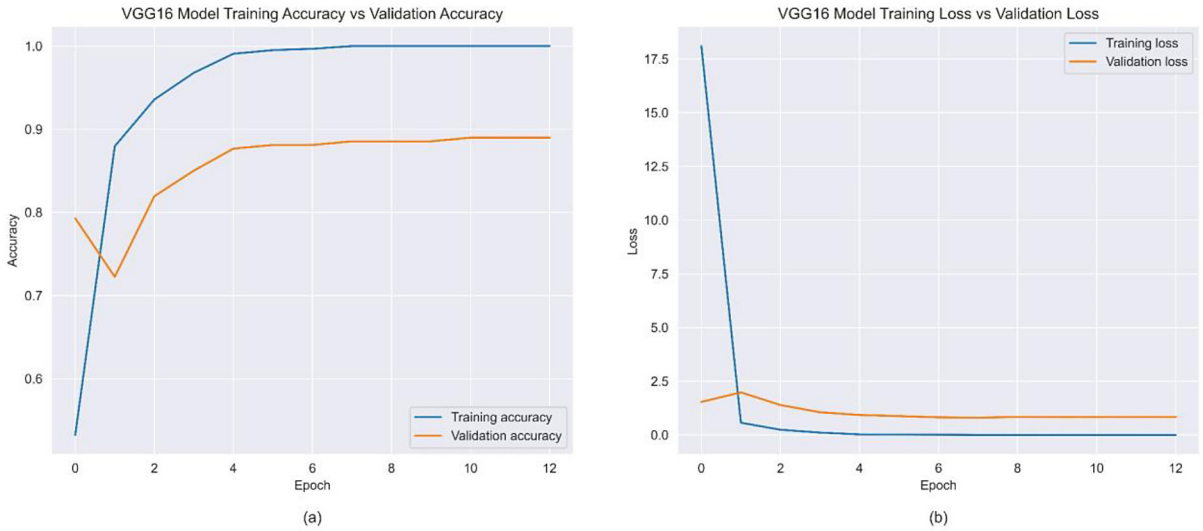

Figure 3. Obtained results for the VGG16 model: (a) training accuracy compared to validation accuracy, (b) training loss compared to validation loss

(79%). It could mean that other dances could be mislabeled as it. Overall, for every assessed dance F1-score was around 85–88%. It demonstrates that the selected classifier performs well and produces consistent outcomes. Figure 4 pictures the confusion matrix for the VGG16 model. The actual classes of Polish national dances are shown on the vertical axis (named True Label) while the classes predicted by the VGG16 classifier are presented on the horizontal axis (named Predicted Label) (Table 2). The whole confusion matrix is normalized and presents correct classifications of the diagonal axis and misclassifications off it. Each cell pictures the proportion of how well the actual dance is classified by the model. Based on these results it could be stated that 95% of the Kujawiak music was correctly predicted compared to only almost 82% for the Polonez. The Polonez was misclassified as the Kujawiak in over 18% of cases while the Kujawiak was improperly categorized as the Polonez in only 5%.

ResNet50

The training process for the ResNet50 classifier lasts 16 epochs while the obtained curves for validation accuracy and training are uneven (Figure 5). Accuracy fluctuates between 0.75 and 0.9 and stabilizes after 12 epochs below 0.9 value. Training loss rises suddenly and falls between 8 and 10 epochs. There is also a disparity between validation and training accuracy and loss.

The accomplished F1-score is between 89% and 91% which could mean that there is enough balance between precision and recall (Table 3). Figure 6 displays the confusion matrix generated for the ResNet50 classification model with true labels of Polish national dances (the vertical axis), classes predicted by the selected classifier (the horizontal axis), correctly recognized dances at the diagonal axis, and incorrectly outside it. Once more the Kujawiak music was classified properly with the best precision (over 98%). Only 1.67% of the Kujawiak music was wrongly recognized as being from the Polonez. However, with this classification method, the Oberek reached the worst precision (almost 82%). The Mazur was most misclassified with the Kujawiak reaching 6.82%.

DenseNet121

Training accuracy versus validation accuracy and training loss versus validation loss plots for the DenseNet121 model are pictured in Figure 7. It could be observed that training lasts 14 epochs. Validation accuracy rises from 0.6 to over 0.85 around

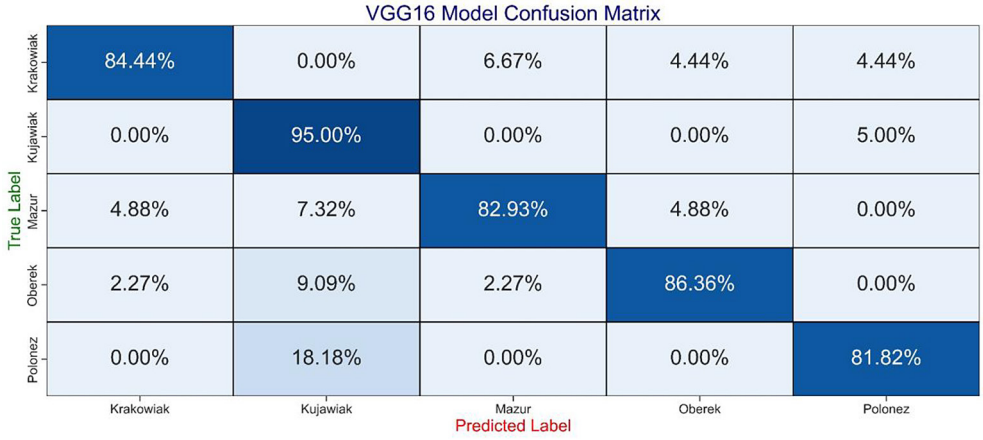

Figure 4. Confusion matrix (in %) for the VGG16 model

Table 2. Precision, recall, and F1-score support for the VGG16 model

Dance/Metric	Precision	Recall	F1-score
Krakowiak	93%	84%	88%
Kujawiak	79%	95%	86%
Mazur	89%	83%	86%
Oberek	90%	86%	88%
Polonez	88%	82%	85%

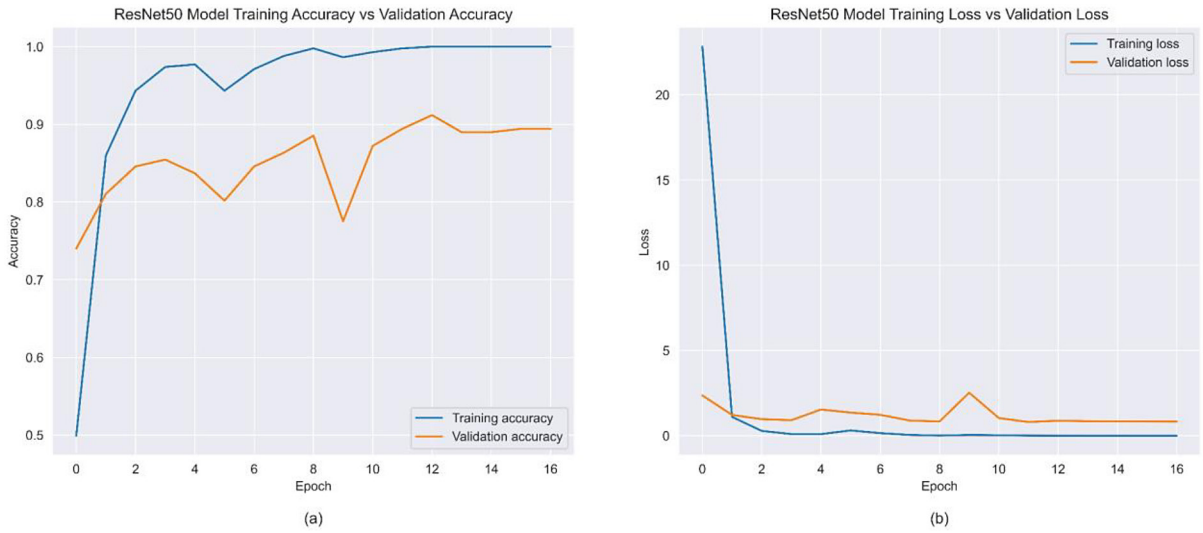

Figure 5. Obtained results for the ResNet50 model: (a) training accuracy compared to validation accuracy, (b) training loss compared to validation loss

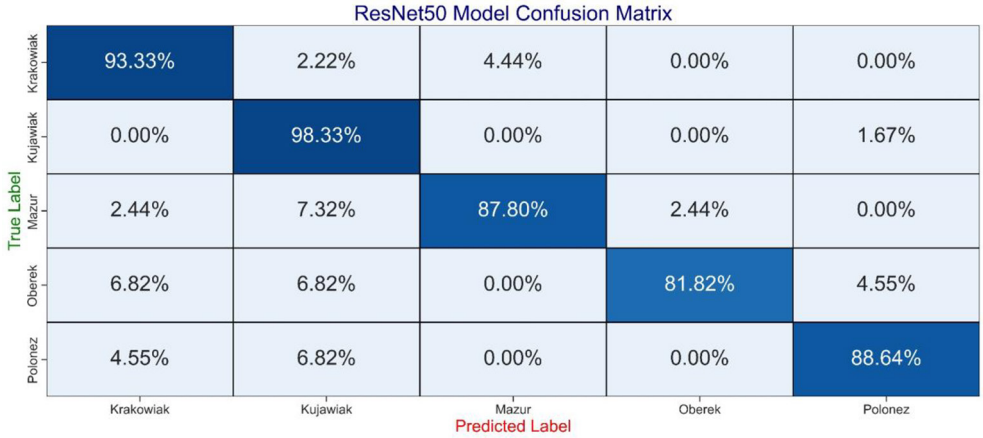

Figure 6. Confusion matrix (in %) for the ResNet50 model

Table 3. Precision, recall, and F1-score for the ResNet50 model

Dance/Metric	Precision	Recall	F1-score
Krakowiak	88%	93%	90%
Kujawiak	86%	98%	91%
Mazur	95%	88%	91%
Oberek	97%	82%	89%
Polonez	93%	89%	91%

the third epoch. Then it starts fluctuating and stabilizes after the 8th epochs below much below 0.9 while training accuracy is more constant. It has been rising till it reaches almost 1.0 after the 5th epoch. Validation loss plateaus after the 7th epoch being around 0.5 value while training loss is near 0. There is a gap between training and validation accuracy as well as between training and validation loss. Table 4 shows that the F1-score is the highest for the Kujawiak (92%) and the lowest for the Krakowiak and the Oberek (88%). Moreover, the Kujawiak has the best recall (over 98%). This might imply that the DenseNet121 also shows consistent results and it has an exceptional outcome relating to identifying the Mel-Spectrograms of the Kujawiak music. Conversely, high precision for the Oberek might suggest

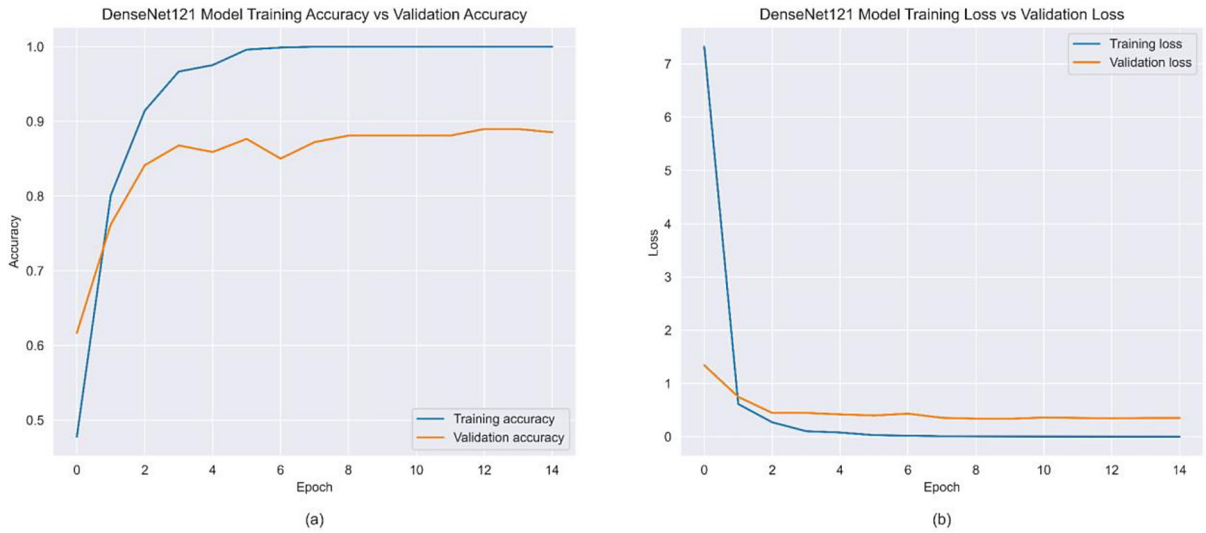

Figure 7. Obtained results for the DenseNet121 model: (a) training accuracy compared to validation accuracy, (b) training loss compared to validation loss

Table 4. Precision, recall, and F1-score for the DenseNet121 model

Dance/Metric	Precision	Recall	F1-score
Krakowiak	87%	89%	88%
Kujawiak	87%	98%	92%
Mazur	89%	83%	86%
Oberek	95%	82%	88%
Polonez	89%	89%	89%

that in most instances of predicting the Oberek Mel-Spectrograms are correct. The confusion matrix for the DenseNet121 classification method is presented in Figure 8. Correctly classified dances are indicated along the diagonal, while misclassified instances appear outside of this diagonal. Surprisingly, the results reached for this classifier are similar to those obtained by the ResNet50 classification method. The same outcomes were acquired by the Kujawiak dance – 98.33% of Mel-Spectrograms were predicted accordingly, while only 1.67% were mislabeled as the Polonez. The Oberek was also the least classified dance with only almost 82% of Mel-Spectrograms recognized correctly. The Polonez was again misclassified as the Kujawiak the most time (nearly 12%) as with the VGG16 classifier.

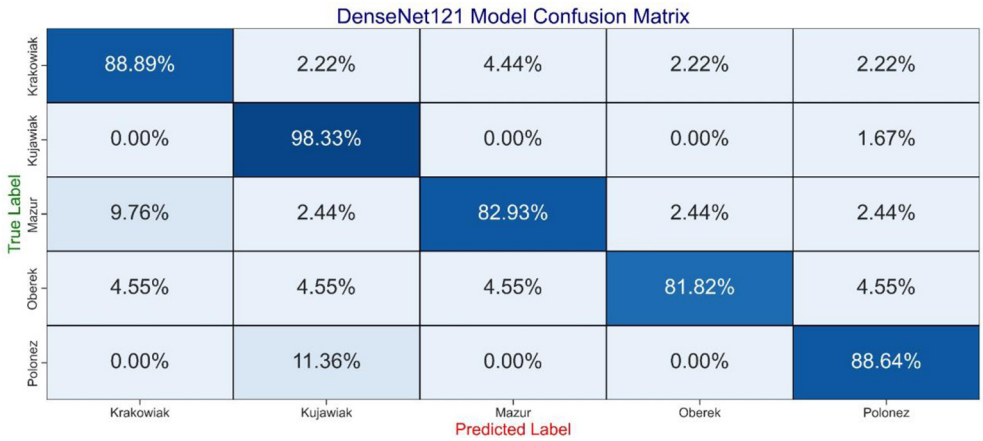

Figure 8. Confusion matrix (in %) for the DenseNet121 model

MobileNetV2

Obtained metrics for the MobileNetV2 classification method such as training and validation accuracy and loss are presented in Figure 9. The training continues only for 7 epochs. Training accuracy rises from 0 to over 0.9 in 2 epochs and then fluctuates slightly and plateaus after the 5th epoch not reaching 1.0 value. On the other hand, validation accuracy rises from over 0.7 to over 0.8 and drops sharply reaching less than 0.8, and starts rising to over 0.85. Although, it does not stabilize. Training loss reaches almost 0 after the 3rd epoch while validation loss fluctuates between less than 1 and over 1.5 for 3 epochs then it regulates. There are differences in both the training and validation accuracy and loss.

Table 5 displays precision, recall, and F1-score obtained for the MobileNetV2 classifier. It could be stated that the F1-score is less consistent and reaches around 79–93%. This might imply that the MobileNetV2 classification method performs less consistently. This could be perceived with the difference between precision (91%) and recall (71%) reached for the Mazur dance which means that the evaluated classifier mislabeled other dances quite often. The confusion matrix for the MobileNetV2 classification method is illustrated in Figure 10. The percentage of dances recognized correctly by the examined classification method is shown on the diagonal axis whereas the off-diagonal elements point to how much each dance was misclassified as the predicted dance. It could be outlined that concerning the Mel-Spectrograms of the Krakowiak, the Kujawiak, and the Polonez the accomplished results exhibited considerable similarity with the proportion of correctly labeled elements. For each dance, it reached over 93% of correct predictions. However, the Mazur and the Oberek display decreased performance with correct predictions at the level of 70–77%. The most Mel-Spectrograms of the Mazur were also misclassified as the Oberek (over 12%).

Combined results

Table 6 shows acquired testing accuracy and loss for every utilized classifier. ResNet50 outperforms other classification methods obtaining over 90% accuracy while testing loss is the worst and reaches 1.00. In contrast, DenseNet121 is characterized by the best testing loss (0.38). The

Figure 9. Obtained results for the MobileNetV2 model: (a) training accuracy compared to validation accuracy, (b) training loss compared to validation loss

Table 5. Precision, recall, and F1-score for the MobileNetV2 model

Dance/Metric	Precision	Recall	F1-score
Krakowiak	91%	96%	93%
Kujawiak	89%	93%	91%
Mazur	91%	71%	79%
Oberek	87%	77%	82%
Polonez	79%	95%	87%

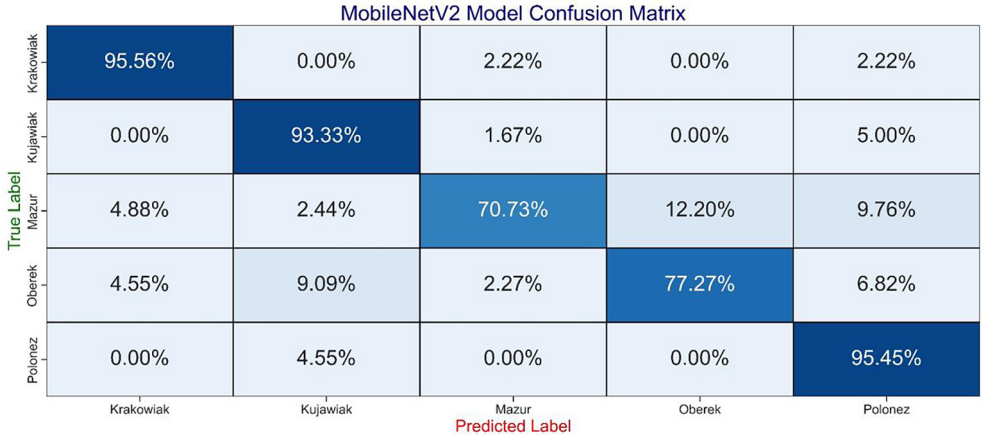

Figure 10. Confusion matrix (in %) for the MobileNetV2 model loss

Table 6. Testing accuracy and loss obtained for each classifier

Classification method/Classification metric	Testing accuracy	Testing loss
VGG16	86.75%	0.94
ResNet50	90.59%	1.00
DenseNet121	88.88%	0.38
MobileNetV2	87.17%	0.60

lowest accuracy is achieved by VGG16 and it is near 87%. Generally, it might be noticed that each classifier reaches quite a similar accuracy over 87% and could be perceived as a sufficient result.

DISCUSSION

A comparison of the state-of-the-art studies with obtained results is presented in Table 7. Audio recordings utilized in conducted studies were mainly obtained from the available online datasets such as the GTZAN or FMA. Only a few studies focused on folk music. None of them were related to Polish folk music. It could be noticed that many studies focused on applying Mel-spectrograms as an input to the CNNs. Some of them developed their CNNs however many applied transfer learning utilizing one of the most prevalent classifiers nowadays such as VGG16 or ResNet50. Obtained results were sometimes compared with some other methods of classification music such as extracting MFCC features and feeding them forward into Artificial Neural Networks (ANNs). Then attained outcomes were evaluated employing some well-known metrics such as accuracy or loss. Few studies analyzed precision, recall, F1-score, or confusion matrices, too. Based on the acquired results it could be stated that approaches that applied Mel-spectrograms and CNNs to classify music obtain sufficient outcomes which were usually better compared to other methods.

In this study, Mel-spectrograms were generated relying on Polish national dance music such as the Polonez, the Kujawiak, the Oberek, the Mazur, and the Krakowiak. Then, formerly applied classifiers (VGG16, ResNet50, DenseNet121, MobileNetV2) in studies from Music Information Retrieval were utilized in Polish National Dances Recognition. Obtained results were compared using classification metrics such as accuracy, loss, precision, recall, and F1-score. Confusion matrices were also created. It was possible to obtain testing accuracy of over 86% for every classification method. Utilized models were able to classify dance music well, too.

Based on the authors' current knowledge there are four studies where datasets consisted of folk or ethnic music. They employed various arguments as the input such as Mel-frequency cepstral coefficient features, the Mel-sound spectrum, the short-time Fourier sound spectrum, or logarithmic scale spectrograms. One of them utilized Mel-spectrograms. It could be remarked that two of them employed self-adjusted CNNs, while one introduced transfer learning classification methods such as ResNet18 and ShuffleNet. Generally, studies consisting of CNNs performed better considering testing accuracy allowing it to reach 89–93% which is partly

Table 7. Comparison with the state of the art

Paper	Dataset	Input	Classification method	Testing accuracy	Other metrics
[12]	GTZAN	Mel-spectrograms	ResNet24	79%	-
			VGG16	76.93%	
			ResNet50	75.9%	
			AlexNet	71.3%	
[13]	FMA, GTZAN, EMA	spectrograms	ResNet50	77-81%	-
			VGG16	72-79%	
			MobileNetV2	57-66%	
			NASNetMobile	65-70%	
			DenseNet121	52-62%	
[14]	GTZAN	Fourier transform-based (FS), Q transform-based (QS), Mel-frequency transform-based (MS) spectrograms	AlexNet, LeNet-5	-	-
[20]	GTZAN, 10GenreGram	spectrograms	ResNet18	51.4% and 80.4%	loss
			NNet2	41.3% and 80.4%	
[21]	a dataset containing 6000 audio files from 10 genres	Mel-spectrograms, MFCC features	CNNs and ANN	76.2% for Mel-spectrograms, and 61.4% for MFCCs	loss
[22]	GTZAN, FMA-small, JUNO	Mel-spectrograms	MR-DCNN	63-89.7%	MAE, RMSE
[23]	FMA	Mel-spectrograms	self-developed CNN classifier	60.5%	F1-score, ROC
[24]	a dataset with mixed music, speech, and noise	Spectrograms	a CNN with a Mel-scale convolutional layer	-	precision, recall, F1-score
[25]	a dataset of music and speech from “All India Radio” news archives overlapping in different languages	Sobel edge spectrograms	CNN	98.91%	-
			ANN	95.56%	
			SVM	96.94%	
			RNN	99.1%	
			CNN	99.44%	
			ANN [24]	94.08%	
[26]	nine different singers, 20 different songs for each singer	MFCC, octave-based spectral contrast	Extra Tree	89.4%	-
	Artist20	MFCC, octave-based spectral contrast	KNN	85.4%	
[27]	a dataset consisting of 10 genres related to ethnic music	Mel-sound spectrum, the short-time Fourier sound spectrum	a self-adjusted CNN	90.3-92.8%	-
[28]	a dataset consisting of 6 types of folk music	Mel-frequency cepstral coefficient features	SVM	92%	-
			BPNN	73%	
			decision tree	64%	
[29]	a dataset of Chinese traditional folk music	Mel-spectrograms	self-adjusted CNN, ResNet18, ShuffleNet	over 89%	sensitivity, specificity
[30]	a dataset comprising 1,154 samples with 15 Makams	logarithmic scale spectrograms	self-adjusted LSTM	90-95%	precision, recall, F1 score
Our work	a dataset consisting of 2296 audio samples of Polish national dance music	Mel-spectrograms	ResNet50	90.59%	loss, precision, recall, F1-score
			DenseNet121	88.88%	
			VGG16	86.75%	
			MobileNetV2	87.17%	

higher than the obtained testing accuracy in Polish national music dance recognition (87–91%). However, utilized classification methods could be contrasted with other studies where popular classifiers within the transfer learning area were employed such as VGG16 or ResNet50. These works were usually based on the available online datasets such as GTZAN or FMA. The acquired testing accuracy

was significantly lower attaining 41–81%. What is more, it could be mentioned that the study where Sobel Edge spectrograms from a dataset of music and speech from "All India Radio" news archives overlapping in different languages were the input to the CNN, ANN, SVM, ANN, and RNN outperformed acquired results by reaching testing accuracy of 94–99%.

As a part of this study, the dataset consisting of 137 audio recordings of Polish national dances such as the Polonez, the Krakowiak, the Kujawiak, the Oberek, and the Mazur was created. This dataset was utilized in the novel methodology for recognizing Polish national dance music based on Mel-spectrograms. Generated Mel-spectrograms were fed forward into pre-trained classification methods such as VGG16, ResNet50, DenseNet121, and MobileNetV2. These classifiers were chosen due to their application in general music classification where they outperformed previously employed methods. The performance of the used classification methods was compared using the following metrics: accuracy, loss, precision, recall, and F1-score. Acquired results were also presented on confusion matrices. Obtained results might be considered sufficient. ResNet50 acquired the best testing accuracy (over 90%) and DenseNet121 was characterized by the best testing loss (0.38). However, based on the comparison of training and validation accuracy and loss it could be stated that there are some aspects requiring modification. There are gaps in every classifier between validation and training loss and accuracy which could suggest model overfitting. Overfitting is a situation where the model performs well on the training set, but its performance on the new dataset is not satisfactory [53]. The applied pre-trained classifiers and the EarlyStopping callback were utilized to mitigate overfitting. However, it might not be sufficient due to the small size of the dataset. It could hinder the model's ability to establish patterns. The presence of noise could also affect the performance of the classification methods since it might lead the classifier to learn from it and act as a prediction basis [54]. According to [55] other methods might be applied to handle overfitting such as regularization techniques or the cross-validation. L2 Regularization is one of the most popular forms of the regularization techniques [56]. It might be achieved by adding the squared magnitude of all the parameters, including weights and biases, into the cost/loss function. In this type of regularization, every weight is decreased linearly towards zero. Alongside with L2 Regularization, there is L1 regularization where the absolute value of the weights is added to the cost function. They could be both employed together. Usually, L2 regularization performs better than L1 regularization. A Dropout could be also applied which refers to removing units from both the hidden layer and the input layer. Based on [57] the Dropout technique performs better than L1 and L2 regularization. However, it might achieve even greater results when the fusion of these techniques is applied rather than being employed independently. Applying the cross-validation method could also prevent the classification method from overfitting. In cross-validation, the dataset is divided into k folds, where k-1 folds are utilized for training the classifier, and the remaining part is employed for testing [58]. Then, the folds rotate allowing all folds to be applied in the training and testing processes. Then the final performance metrics are calculated by averaging over the k estimates from each test fold. This approach ensures that k independent sets are utilized to evaluate the model. The test set remains completely unavailable during the model's training phase to prevent overfitting. In some classification methods (ResNet50, MobileNetV2) validation accuracy fluctuates and stabilizes later. This could mean that there is some instability in the created models. Due to the implemented EarlyStopping callback, each model stops training between 7 and 16 epochs which could imply that the created dataset might be too small. The obtained results could be also analyzed regarding how each classification method performed on each of the Polish national dances. It is important to note that the sample sizes varied across classes, with the Kujawiak being the most numerous and the Mazur the least. For VGG16, ResNet50, and DenseNet121 the Kujawiak demonstrated that over 95-99% of Mel-Spectrograms were correctly classified as this dance. A slightly worse outcome was reached for MobileNetV2 (over 93%). However, all classification methods were generally worse in identifying the Kujawiak Mel-Spectrograms to be the Kujawiak Mel-Spectrograms. It could be mentioned that two of the least numerous classes, Mazur and the Oberek, performed worst in classifying actual dances. Although classifiers mainly predicted Mel-Spectrograms as these dances, in the majority are these dances. By analyzing the F1-score it might be noticed that every classifier displays quite a

balance of outcomes, only MobileNetV2 fluctuates more, for the Krakowiak and Kujawiak the F1-score is around 91–93% while for the Mazur only 79%. As a result of evaluating the confusion matrices and parameters such as precision, recall, and F1-score for each Polish national dance it could be stated that the unbalanced dataset generally does not skew acquired outcomes. Through various classes obtained results are similar and there are only a few occurrences when the obtained result is significantly worse. However, to improve the results the class imbalance problem in utilized classification methods could be tackled by employing oversampling [59]. It should entirely obliterate the imbalance and the optimal undersampling ratio should depend on the extent of the imbalance. According to [59] it could help with the overfitting issue, too. Another solution to handle the small dataset could include employing chosen data augmentation techniques to enlarge the training set size. It helps create variability in order to improve the model's ability to generalize and reduce the risk of overfitting. It could be done by shifting, flipping, and zooming the images [60]. Another approach to addressing the problem of small and unbalanced datasets is to use generative models focused on musical data [61, 62, 63]. They can learn the underlying musical structures and patterns present in fragments of folk songs.

In the context of a dataset containing fragments of folk music, these models can be trained to understand characteristics such as melody, harmony, rhythm, and phrasing that are typical of folk songs. Once trained, the generative model can produce new musical fragments that are stylistically similar to the originals. By generating additional musical pieces, the dataset can be augmented to increase its size and diversity. This is particularly beneficial for balancing the dataset if certain musical styles, regional variations, or instrumental arrangements are underrepresented.

The synthetic musical data produced by generative models can then be combined with the original dataset to train other machine learning models. This augmentation can lead to improved performance by providing a richer and more varied set of training examples, helping models to generalize better to new, unseen musical data. Additionally, using synthetic data can help mitigate issues related to overfitting that often occur with small datasets.

Potential future works include acquiring a larger dataset by applying data augmentation techniques based on data warping and oversampling [64]. This could be also beneficial in terms of handling overfitting. However, some regularization techniques such as L1/L2 or Dropout could be also applied [65]. Additionally, more enhanced outcomes could be obtained by utilizing various popular pre-trained classifiers such as AlexNet or modifying the architecture of the developed models. A late fusion strategy could be employed to integrate features derived from multiple spectral features such as short-time Fourier transform, Mel-spectrograms, and MFCC into a CNN model [66]. Some ensemble methods such as voting mechanisms (hard and soft voting) could also be applied [67].

The study encountered several challenges, including the following: a limited size of gathered data due to the deficient sources where the music of Polish national dances might be found; a slight imbalance in the number of samples in each class representing each dance; potential overfitting, likely resulting from the small size of the dataset.

However, gathered data and proposed methodology might still signify that an undertaken study obtained state-of-the-art results in Polish national dance music recognition. They could be employed in various areas such as music-driven dance generation [68] or creating tools for song auto-tagging [69].

CONCLUSIONS

Based on the results could be stated that the selected methodology was the right choice for Polish national dance music recognition. It allowed the creation of a unique dataset and cutting-edge solution for classifying audio samples from Polish national dances. The utilized classification methods performed better when evaluated against similar classification methods employed in music genre recognition. However, the obtained outcomes were slightly worse compared to models that used other datasets based on folk or ethnic music which utilized self-adapted CNNs. It is noteworthy that Sobel Edge spectrograms could yield considerably better results.

REFERENCES

1. Wargowska-Dudek, A. Ochrona dziedzictwa kulturowego na przykładzie polskich tańców narodowych – praktyki depozytariusza Henryka Dudy (in Polish). Perspektywy Kultury, 2023; 40(1): 95–108. https://doi.org/10.35765/pk.2023.4001.08.

2. NID. (2024, April 23). Krajowa Lista Niematerialnego Dziedzictwa Kulturowego (in Polish) – NID. https://niematerialne.nid.pl/en/niematerialne-dziedzictwo-kulturowe/krajowa-lista-niematerialnego-dziedzictwa-kulturowego/ (Accessed: 08.06.2024).
3. Olha, S. Elements of classic choreography at the academization of Polish folk-stage dance. Zenodo (in Ukrainian) (CERN European Organization for Nuclear Research). 2023. https://doi.org/10.5281/zenodo.7806696.
4. Skublewska-Paszkowska, M., Powroznik, P., Smolka, J., Milosz, M., Lukasik, E., Mukhamedova, D., & Milosz, E. Methodology of 3D scanning of intangible cultural heritage – the example of Lazgi Dance. Applied Sciences, 2021; 11(23): 11568. https://doi.org/10.3390/app11231156.
5. Li, N.P. The mediating effect of artificial intelligence on the relationship between cultural heritage preservation and opera music: A case study of Shanxi Opera. Evolutionary Studies in Imaginative Culture, 2024; 249–267.
6. Yu, T., Wang, X., Xiao, X., Yu, R. Harmonizing tradition with technology: Using AI in traditional music preservation. Conference: 2024 International Joint Conference on Neural Networks (IJCNN), 2024; 15: 1–8. https://doi.org/10.1109/ijcnn60899.2024.10651124.
7. Rallis, I., Voulodimos, A., Bakalos, N., Protopapadakis, E., Doulamis, N., Doulamis, A. machine learning for intangible cultural heritage: a review of techniques on dance analysis. Springer Series on Cultural Computing, 2020; 103–119. https://doi.org/10.1007/978-3-030-37191-3_6.
8. Huang, L., Song, Y. Intangible cultural heritage management using machine learning model: a case study of northwest folk song huaer. Scientific Programming, 2022; 1–9. https://doi.org/10.1155/2022/1383520.
9. Stacchio, L., Garzarella, S., Cascarano, P., De Filippo, A., Cervellati, E., & Marfia, G. DanXe: An extended artificial intelligence framework to analyze and promote dance heritage. Digital Applications in Archaeology and Cultural Heritage, 2024; 33: e00343. https://doi.org/10.1016/j.daach.2024.e00343.
10. Skublewska-Paszkowska, M., Powroźnik, 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(8).
11. Schedl, M., Gómez, E., Urbano, J. Music information retrieval: recent developments and applications. Foundations and Trends in Information Retrieval, 2014; 8(2–3): 127–261. https://doi.org/10.1561/1500000042.
12. Mehta, J., Gandhi, D., Thakur, G., Kanani, P. Music Genre Classification using Transfer Learning on log-based MEL Spectrogram. Conference: 2021 5th International Conference on Computing Methodologies and Communication (ICMC), 2021. https://doi.org/10.1109/iccmc51019.2021.9418035.
13. Li, J., Han, L., Li, X., Zhu, J., Yuan, B., Gou, Z. An evaluation of deep neural network models for music classification using spectrograms. Multimedia Tools and Applications, 2021; 81(4): 4621–4647. https://doi.org/10.1007/s11042-020-10465-9.
14. Dhall, A., Murthy, Y.V.S., Koolagudi, S.G. Music genre classification with convolutional neural networks and comparison with f, q, and mel spectrogram-based images. In Advances in intelligent systems and computing, 2021; 235–248. https://doi.org/10.1007/978-981-33-6881-1_20.
15. Powroznik, P., Wojcicki, P., Przylucki, S.W. Scalogram as a representation of emotional speech. IEEE Access, 2021; 9: 154044-154057. https://doi.org/10.1109/ACCESS.2021.3127581.
16. Czerwinski, D., Powroznik, P. Human emotions recognition with the use of speech signal of polish language. 2018 Conference on Electrotechnology: Processes, Models, Control and Computer Science (EPMCCS), 2018; 1-6. https://doi.org/10.1109/EPMCCS.2018.8596404.
17. gtzan. (n.d.). TensorFlow. A gtzan dataset. https://www.tensorflow.org/datasets/catalog/gtzan (Accessed: 08.06.2024).
18. Mdeff. (n.d.). GitHub – mdeff/fma: FMA: a dataset for music analysis. GitHub. https://github.com/mdeff/fma (Accessed: 08.06.2024).
19. Defferrard, M., Benzi, K., Vandergheynst, P., Bresson, X. FMA: a dataset for music Analysis. arXiv (Cornell University), 2017; 316–323. https://doi.org/10.48550/arXiv.1612.01840.
20. Hassen, A.K., 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(1), 20. https://doi.org/10.5445/ksp/1000087327/20.
21. Rawat, P., Bajaj, M., Vats, S., Sharma, V. A comprehensive study based on MFCC and spectrogram for audio classification. Journal of Information & Optimization Sciences, 2013; 44(6): 1057–1074. https://doi.org/10.47974/jios-1431.
22. Yin, T. Music track recommendation using DeepCNN and MEL Spectrograms. Journal on Special Topics in Mobile Networks and Applications/Mobile Networks and Applications, 2023. https://doi.org/10.1007/s11036-023-02170-2.
23. Matocha, M., Zielinski, S.K. Music genre recognition using convolutional neural networks. Advances in Computer Science Research, 2015; 14: 125–142. https://doi.org/10.24427/acsr-2018-vol14-0008.
24. Jang, B., Heo, W., Kim, J., & Kwon, O. Music detection from broadcast contents using

convolutional neural networks with a Mel-scale kernel. EURASIP Journal on Audio, Speech and Music Processing, 2019; 1. https://doi.org/10.1186/s13636-019-0155-y.

25. Sawant, O., Bhowmick, A., Bhagwat, G. Separation of speech & music using temporal-spectral features and neural classifiers. Evolutionary Intelligence, 2023. https://doi.org/10.1007/s12065-023-00828-0.
26. Hizlisoy, S., Arslan, R.S., Çolakoğlu, E. Singer identification model using data augmentation and enhanced feature conversion with hybrid feature vector and machine learning. EURASIP Journal on Audio, Speech and Music Processing, 2024: 1. https://doi.org/10.1186/s13636-024-00336-8.
27. Ning, Q., Shi, J. Artificial neural network for folk music style classification. Journal of Mobile Information Systems, 2022; 1–9. https://doi.org/10.1155/2022/9203420.
28. Wang, X. Research on recognition and Classification of folk music based on feature extraction Algorithm. Informatica, 2020; 44(4). https://doi.org/10.31449/inf.v44i4.3388.
29. Mi, D., & Qin, L. Classification System of National Music rhythm Spectrogram based on biological neural network. Computational Intelligence and Neuroscience, 2022: 1–10. https://doi.org/10.1155/2022/2047576.
30. Mirza, F.K., Gürsoy, A.F., 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. Multimedia Tools and Applications, 2023; 83(14): 41243–41271. https://doi.org/10.1007/s11042-023-17105-y.
31. Abed, M.H., Al-Asfoor, M., Hussain, Z.M. Architectural heritage images classification using deep learning with CNN. Conference: VIPERC 2020 Visual Pattern Extraction and Recognition for Cultural Heritage Understanding, 2020; 2602: 1–12.
32. Zou, Z., Yu, Z., Guo, Y., Li, Y., Liang, D., Cao, Y., Zhang, S. Triplane Meets Gaussian Splatting: Fast and Generalizable Single-View 3D Reconstruction with Transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024; 10324–10335. https://doi.org/10.1109/cvpr52733.2024.00983.
33. Vu, M., Beurton-Aimar, M., Le, V. Heritage Image Classification by Convolution Neural Networks. IEEE 2018. https://doi.org/10.1109/mapr.2018.8337517.
34. Babić, R.J. A comparison of methods for image classification of cultural heritage using transfer learning for feature extraction. Neural Computing and Applications, 2023; 36(20): 11699–11709. https://doi.org/10.1007/s00521-023-08764-x.
35. Mehta, S., Kukreja, V., Bordoloi, D. Heritage Coin Identification using Convolutional Neural Networks: A Multi-Classification Approach for Numismatic Research. IEEE 2023. https://doi.org/10.1109/icaiss58487.2023.10250481.
36. Loupas, G., Pistola, T., Diplaris, S., Ioannidis, K., Vrochidis, S., Kompatsiaris, I. Comparison of deep learning techniques for video-based automatic recognition of greek folk dances. In Lecture notes in computer science, 2023; 325–336. https://doi.org/10.1007/978-3-031-27818-127.
37. Jain, N., Bansal, V., Virmani, D., Gupta, V., Salas-Morera, L., Garcia-Hernandez, L. An enhanced deep convolutional neural network for classifying Indian classical dance forms. Applied Sciences, 2021; 11(14): 6253. https://doi.org/10.3390/app11146253.
38. Powroznik, P., Czerwiński, D. Spectral methods in Polish emotional speech recognition. Advances in Science and Technology. Research Journal, 2016; 10(32). https://doi.org/10.12913/22998624/65138.
39. Kumar, C.S.A., Maharana, A.D., Krishnan, S.M., Hanuma, S.S.S., Lal, G.J., Ravi, V. Speech Emotion Recognition using CNN-LSTM and Vision Transformer. In Lecture notes in networks and systems, 2013; 86–97. https://doi.org/10.1007/978-3-031-27499-2_8.
40. Powroznik, P. Polish emotional speech recognition using artifical neural network. Advances in Science and Technology. Research Journal, 2014; 8(24): 24-27. https://doi.org/10.12913/22998624/562.
41. Utebayeva, D., Ilipbayeva, L., Matson, E.T. Practical study of recurrent neural networks for efficient real-time drone sound detection: a review. Drones, 2013; 7(1): 26. https://doi.org/10.3390/drones7010026.
42. Bahuleyan, H. Music Genre Classification using Machine Learning Techniques. arXiv (Cornell University). 2018. https://doi.org/10.48550/arxiv.1804.01149.
43. Simonyan, K., & Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv (Cornell University), 2014. https://doi.org/10.48550/arxiv.1409.1556.
44. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2015; 115(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y.
45. He, K., Zhang, X., Ren, S., Sun, J. Deep residual learning for image recognition. arXiv (Cornell University). 2015. https://doi.org/10.48550/arxiv.1512.03385.
46. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q. Densely connected convolutional networks. arXiv (Cornell University), 2016. https://doi.org/10.48550/arxiv.1608.06993.

47. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L. MobileNetV2: Inverted residuals and linear bottlenecks. arXiv (Cornell University), 2018. https://doi.org/10.48550/arxiv.1801.04381.
48. Skublewska-Paszkowska, M., Powroznik, P., Lukasik, E. Learning three dimensional tennis shots using graph convolutional networks. Sensors, 2020; 20(21): 6094. https://doi.org/10.3390/s20216094.
49. Hossin, M., Sulaiman, M.N. A review on evaluation metrics for data classification evaluations. International Journal of Data Mining and Knowledge Management Process, 2015; 5(2): 01–11. https://doi.org/10.5121/ijdkp.2015.5201.
50. Skublewska-Paszkowska, M., & Powroznik, P. Temporal pattern attention for multivariate time series of tennis strokes classification. Sensors, 2023; 23(5), 2422. https://doi.org/10.3390/s23052422.
51. Marom, N.D., Rokach, L., Shmilovici, A. Using the confusion matrix for improving ensemble classifiers. 2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel, 2010. https://doi.org/10.1109/eeei.2010.5662159.
52. Terven, J., Cordova-Esparza, D.M., Ramirez-Pedraza, A., Chavez-Urbiola, E.A. Loss functions and metrics in deep learning. arXiv (Cornell University), 2023. https://doi.org/10.48550/arxiv.2307.02694.
53. Santos, C.F.G.D., Papa, J.P. Avoiding Overfitting: A survey on regularization methods for convolutional neural networks. ACM Computing Surveys, 2022; 54(10s): 1–25. https://doi.org/10.1145/3510413.
54. Alamri, N.M. Reducing the overfitting in convolutional neural network using nature-inspired algorithm: a novel hybrid approach. Arabian Journal for Science and Engineering, 2024. https://doi.org/10.1007/s13369-024-08998-4.
55. Rajvanshi, S., Kaur, G., Dhatwalia, A., Arunima, N., Singla, A., Bhasin, A. Research on Problems and solutions of overfitting in machine learning. In Lecture notes in electrical engineering, 2024; 637–651. https://doi.org/10.1007/978-981-97-2508-3_47.
56. Gupta, S., Gupta, R., Ojha, M., Singh, K.P. A comparative analysis of various regularization techniques to solve overfitting problem in artificial neural network. In Communications in computer and information science, 2018; 363–371. https://doi.org/10.1007/978-981-10-8527-7_30.
57. Thakkar, A., Lohiya, R. Analyzing fusion of regularization techniques in the deep learning-based intrusion detection system. International Journal of Intelligent Systems, 2021; 36(12): 7340–7388. https://doi.org/10.1002/int.22590.
58. Santos, M.S., Soares, J.P., Abreu, P.H., Araujo, H., Santos, J. Cross-Validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [Research Frontier]. IEEE Computational Intelligence Magazine, 2018; 13(4): 59–76. https://doi.org/10.1109/mci.2018.2866730.
59. Buda, M., Maki, A., & Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 2018; 106: 249–259. https://doi.org/10.1016/j.neunet.2018.07.011.
60. Alshomrani, S., Aljoudi, L., Arif, M. Arabic and American sign languages alphabet recognition by convolutional neural network. Advances in Science and Technology – Research Journal, 2021; 15(4): 136–148. https://doi.org/10.12913/22998624/142012.
61. Zhang, J., Fazekas, G., Saitis, C. (2023). Fast diffusion GAN model for symbolic music generation controlled by emotions. arXiv (Cornell University). 2023. https://doi.org/10.48550/arxiv.2310.14040.
62. Zhang, H., Xie, L., Qi, K. Implement music generation with GAN: a systematic review. 2021 International Conference on Computer Engineering and Application (ICCEA), 2012. https://doi.org/10.1109/iccea53728.2021.00075.
63. Dong, H., Hsiao, W., Yang, L., Yang, Y. MUSEGAN: multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. arXiv (Cornell University). 2017. https://doi.org/10.48550/arxiv.1709.06298.
64. Khosla, C., Saini, B.S. Enhancing performance of deep learning models with different data augmentation techniques: a survey. 2020 International Conference on Intelligent Engineering and Management (ICIEM). 2020. https://doi.org/10.1109/iciem48762.2020.9160048.
65. Ying, X. An Overview of overfitting and its solutions. Journal of Physics. Conference Series, 2019; 1168: 022022. https://doi.org/10.1088/1742-6596/1168/2/022022.
66. Seo, W., Cho, S., Teisseyre, P., Lee, J. A Short Survey and Comparison of CNN-based Music Genre Classification using Multiple Spectral Features. IEEE Access, 2024; 12: 245–257. https://doi.org/10.1109/access.2023.3346883.
67. Assiri, A.S., Nazir, S., Velastin, S.A. Breast tumor classification using an ensemble machine learning method. Journal of Imaging, 2020; 6(6): 39. https://doi.org/10.3390/jimaging6060039.
68. Zhuang, W., Wang, C., Chai, J., Wang, Y., Shao, M., Xia, S. Music2Dance: dancenet for music-driven dance generation. ACM transactions on multimedia computing, communications and applications/acm transactions on multimedia computing communications and applications, 2022; 18(2): 1–21. https://doi.org/10.1145/3485664.
69. Song, G., Wang, Z., Han, F., Ding, S., Iqbal, M.A. Music auto-tagging using deep recurrent neural networks. Neurocomputing, 2018; 292, 104–110. https://doi.org/10.1016/j.neucom.2018.02.076.