Scientia Iranica Journal

Iranian Traditional Music Subgenre (Dastgah) Recognition Using Ensemble Learning And Graph–Based Representation By Introducing New Database

Sina Ghazanfaripour[a], Morteza Khademi[a]*, Abbas Ebrahimi-Moghadam[a]

* Corresponding author, Email: khademi@um.ac.ir, Tel: +989153156497

[a] Department of Electrical Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

Appendix

The selection of the 1DCNN architecture was based on evaluating the effects of different configurations on the model's accuracy, as illustrated in Table 1. This evaluation included varying the number of convolutional layers, the number of filters, kernel sizes, and the use of fully connected layers. As observed, reducing the number of layers or filters decreased the model's accuracy, while increasing the number of filters in certain layers improved accuracy. However, excessive increases led to overfitting and reduced generalizability. Proposed architecture achieved the best balance between accuracy and computational complexity. The use of five convolutional layers was necessary to extract more complex features at different levels. Filters of 32, 64, and 128 were incrementally increased to first capture local features and then model broader relationships in the musical data. Kernel sizes of 3 and 5 were chosen to accurately model subtle changes in note sequences and better identify Iranian music patterns.

The number of LSTM layers and the dropout rate were determined through extensive experimentation and analysis on the Iranian music dataset. Initially, models with varying numbers of LSTM layers were evaluated, comparing their performance in terms of accuracy and their ability to capture complex patterns. The results showed that using four LSTM layers provides an optimal balance between model accuracy and prevention of overfitting. The Figure 1 shows the classification accuracy based on different LSTM layer counts and dropout rates. Based on various experiments and comparing the models' accuracy, it can be concluded that using 4 LSTM layers and a dropout rate of 0.3 provides the best performance on the Iranian music dataset. These choices seem to offer a good balance in terms of accuracy, generalizability, and prevention of overfitting. Lower dropout rates led to overfitting, while higher rates resulted in the loss of valuable information. Ultimately, the dropout rate was set based on the best performance of the model on the test dataset, ensuring maximum accuracy while maintaining generalization capability.

Table 1. 1DCNN architectures

Architecture	#Conv layers	#Filters	Kernel size	#FC layers	Accuracy(%)
Model 1	3	32, 32	3	2	65.16
Model 2	4	32, 32, 64, 64	3, 5	2	69.31
Model 3	4	64, 64, 128, 128	3, 5	3	70.56
Model 4	**5**	**32, 32, 64, 64, 128**	**3, 5**	**3**	**73.33**
Model 5	6	32, 32, 64, 64, 128, 128	3, 5	3	73.20

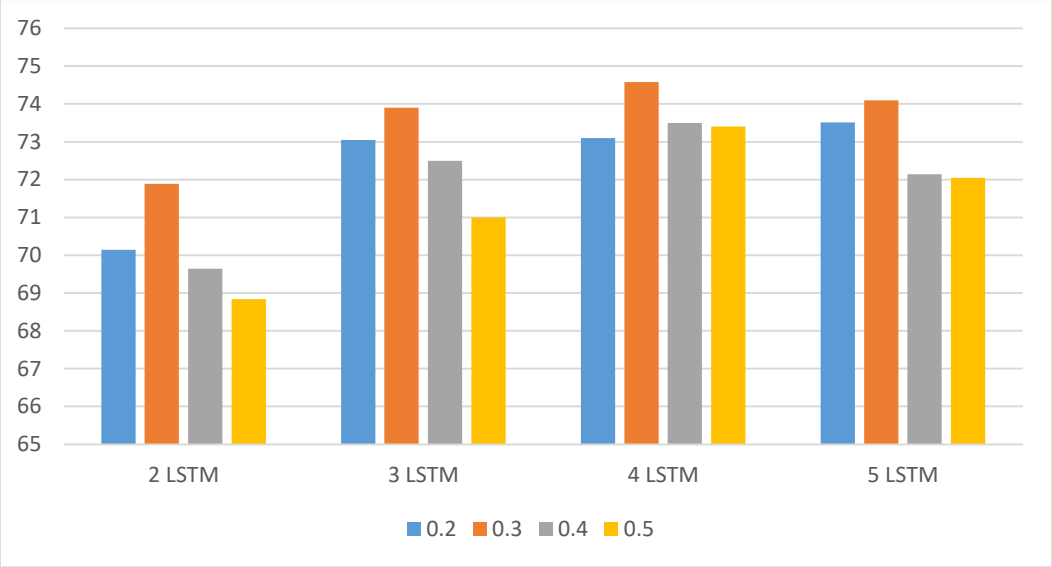

Fig. 1 Accuracy based on different LSTM layer counts and dropout rates

To evaluate the performance of different CNN architectures in recognizing Iranian music dastgahs, several architectures with varying numbers of convolutional layers, filter counts, kernel sizes, dropout rates, and the accuracy of each model were tested, and the results are presented in the table 2. The combination of 64 filters in the initial layers helped in identifying simpler features, while larger filters in the middle and final layers extracted more complex features. The use of larger kernels in the initial layers and smaller kernels in the middle and final layers allowed for the simultaneous extraction of both detailed and general information. These experiments demonstrate that the proposed architecture provides the best performance in terms of accuracy and generalizability. The precise hyperparameters led to a significant improvement in model performance and more accurate feature extraction from dastgahs. These choices were made based on a thorough analysis and comprehensive evaluation of the model's performance.

Three models are compared (1DCNN, CNN, and LSTM) using the Wilcoxon signed-rank test. The dataset was randomly divided into training and testing sets 20 times, generating 20 accuracy samples for each model. Due to the limitation of having only five folds, the test was performed based on these 20 samples for each model. As shown in table 3, both LSTM and 1DCNN demonstrate a statistically significant difference compared to CNN. However, when comparing LSTM and 1DCNN, the result is close to the threshold of statistical significance, indicating that the difference between these two models is not as pronounced as the difference between each of them and CNN.

Table 2. CNN architectures

Architecture	#Conv layers	#Filters	Kernel size	Dropout rate	Accuracy(%)
Model 1	4	64, 128, 256, 512	3×3, 5×5	0.3	71.58
Model 2	5	32, 64, 128, 256, 512	3×3, 5×5	0.3	72.10
Model 3	**7**	**64, 64, 64, 128, 128, 256, 512**	3×3, 5×5	**0.3**	**72.54**
Model 4	7	64, 64, 128, 128, 256, 256, 512	3×3	0.4	71.39
Model 5	6	64, 64, 128, 256, 256, 512	5×5	0.4	71.55

Table 3. Models Wilcoxon signed-rank test

Models	P value
LSTM & CNN	0.03623
1DCNN & CNN	0.04864
LSTM & 1DCNN	0.05851