Improvement of CNN Network Parameters in Turkish Music Emotion Recognition
Türk Müziği Duygu Tanımasında CNN Ağ Parametrelerinin İyileştirilmesi

Murat Surucu[1,*]
[1]Republic of Turkey Ministry of National Education, Ankara, Turkey
ORCIDs: 0000-0002-9889-9952
E-mails: msurucu@gmail.com
*Corresponding author.

Abstract—**Music has been an integral part of humanity throughout history. People have conveyed their emotional expressions through music, and musical styles have evolved alongside communities. Despite the diversity of styles, music has always existed within an emotional context. Therefore, measuring the emotional expressions conveyed by music has given rise to a broad field of study encompassing art, science, history, and sociology. Additionally, with the proliferation of electronic music platforms, the ability to automatically identify the emotional genres of music has become a prominent feature sought after by end users. In this context, while numerous studies have been conducted in various languages, there is a scarcity of research specifically tailored to the Turkish language. For successful execution of processes that can be automated through machine learning, several factors need to be considered: the proper selection of data preprocessing methods, determination of the structure and complexity of the model to be trained, accurate selection of training and testing data, and more. Optimal performance cannot be achieved solely through the correct choice of a model, as flawed data preprocessing can hinder results, and conversely, accurate data preprocessing cannot compensate for a faulty model. This article aims to enhance the performance of a rare music emotion recognition study conducted in the Turkish language by constructing a "problem-specific network model." To achieve this goal, data subjected to various normalization techniques were analyzed using Convolutional Neural Network (CNN) models of different dimensions and complexities. The achievements were compared with two different classifiers to establish a reference point in comparison with previous studies. At the end of the study, it was observed that for data subjected to MinMax normalization, a success rate of 86.67% was achieved with the Softmax classifier and 80% with the SVM classifier. Similarly, with Z-Score normalization, success rates of 84.17% and 81.67% were obtained, respectively. These values are higher than the highest achievement value of 74.2% obtained for the same data group in the reference study. Furthermore, it is believed that applying the additional performance-enhancing procedures used in the reference study to the models in this study would lead to even higher achievements.**

Keywords—CNN; model selecting; hyperparameters; normalization

Özetçe—**Müzik, tarih boyunca insanlığın ayrılmaz bir parçası olmuştur. İnsanlar duygusal ifadelerini müziğin aracılığıyla aktarmış ve topluluklarla birlikte müzik tarzları da evrimleşmiştir. Farklı tarzlarda olmalarına rağmen, müzik her zaman duygusal bir bağlamda var olmuştur. Bu nedenle, müziğin hangi duygusal ifadeleri taşıdığının ölçülmesi, sanattan bilime, tarihten sosyolojiye geniş bir çalışma alanı oluşturmuştur. Ayrıca, elektronik müzik platformlarının yaygınlaşmasıyla birlikte, müziğin duygusal türlerini otomatik olarak belirleyebilmek, son kullanıcıların aradığı özellikler arasında öne çıkmaktadır. Bu bağlamda, farklı dillerde bu konuda birçok çalışma yapılmış olsa da, Türkçe diline özgü çalışmalar oldukça sınırlıdır. Makine öğrenmesi sayesinde otomatikleştirilebilen işlemlerin başarılı bir şekilde gerçekleştirilebilmesi için, veri ön işleme yöntemlerinin doğru bir şekilde seçilmesi, eğitilecek modelin yapısının ve karmaşıklığının belirlenmesi, eğitim ve test verilerinin doğru bir şekilde seçilmesi gibi faktörler üzerinde çalışmak gerekmektedir. Doğru bir model seçimi ile hatalı veri ön işlemesi sonucunda en yüksek başarı elde edilemeyeceği gibi, tersi durumda doğru veri ön işlemesi ile hatalı bir model de başarılı sonuçlar üretemeyecektir. Bu makalede, Türkçe dilinde yapılan nadir müzik duygu tanıma çalışmalarından birine yönelik olarak, "problem özgü ağ modeli" oluşturarak başarımın arttırılması amaçlanmıştır. Bu amaç doğrultusunda, farklı veri normalizasyon yöntemlerine tabi tutulmuş veriler, farklı boyut ve karmaşıklıkta Evrişimli Sinir Ağı (CNN) modelleri kullanılarak analiz edilmiş ve önceki çalışma ile referans olması adına iki farklı sınıflandırıcı ile olan başarımları incelenmiştir. Çalışmanın sonucunda, MinMax normalleştirmeye tabi tutulmuş veriler için Softmax sınıflandırıcının %86,67 ve SVM sınıflandırıcının %80 başarı elde ettiği gözlenmiştir. Benzer şekilde, Z-Skor normalleştirme ile elde edilen sonuçlar ise %84,17 ile %81,67 olarak bulunmuştur. Bu değerler, referans çalışmasında aynı veri grubu için elde edilen en yüksek başarı değeri olan %74,2'den daha yüksektir. Ayrıca, referans çalışmasında kullanılan diğer performans artırıcı işlemlerin bu çalışmanın modellerine uygulanmasıyla daha yüksek başarılar elde edilebileceği düşünülmektedir.**

Anahtar Kelimeler—CNN; model seçimi; hiperparametre; normalleştirme

I. Introduction

The history of written music dates back to even before the 19th century, reaching as far as mythological legends. While there isn't a unanimous consensus on the identity of the first musical instrument, a more intriguing question arises: why did humanity develop musical instruments? All creatures express their emotions through actions, except for humans. Humans, on the other hand, can convey their emotions through rhythm, or in other words, through music. Thus, music has retained its significance for humanity as a means of expressing emotions from ancient times to the present [1].

As humanity conveyed emotions through music, music evolved, giving rise to various region-specific and culturally unique music genres. However, regardless of the genre, music has always possessed an underlying emotional foundation. In today's world, where music is easily accessible, the need arises to categorize it based on its genres, the emotions it evokes, and similar characteristics. With the proliferation of digital music platforms, algorithms capable of distinguishing between different emotional qualities of music have become popular.

Even in contemporary times, recognizing emotions from music remains challenging. This is due to the fact that emotions can vary from person to person [2]. Therefore, databases for classifying music emotions with the participation of numerous individuals are being created [3]–[6]. However, access to most of these databases is limited. Furthermore, due to the presence of language- or region-specific nuances in music databases, achieving a universal classification of music emotions is challenging. Moreover, many studies in this field tend to be predominantly focused on the English language [7], [8]. Open libraries for regional studies are also quite limited. Addressing the deficiency in a Turkish music emotion labeling database, a valuable dataset for the Turkish language, Er and Aydilek's work provides researchers with an essential resource [9].

The initial step with the dataset involves determining the features. If not performed using machine learning techniques [10]–[15], feature selection is a critical process. Various types of features are utilized for detecting emotions in music, categorized into groups like energy, rhythm, temporal, spectrum, and harmony [16]. While the range of emotional states conveyed through music can be further segmented, they are generally conceptualized and modeled using the 2D Arousal-Valence emotion plane, commonly referred to as Thayer's Model or Russell's Model [17], [18]. The categories addressed in this study are the four labels, "angry," "sad," "happy," and "relax," highlighted in bold on Figure X.

Following the determination of target labels and utilizing feedback from participants, a dataset is created. This dataset is then processed for the selected features. To this end, various tools are employed in the literature, with one of the most popular being the MIRToolbox in MATLAB [19]. This toolbox facilitates the extraction of features such as RMS energy, Chromagram, Mel-Frequency Cepstral Coefficients (MFCCs), and Spectrum information from music.

Subsequently, classification or regression processes are modeled using the chosen feature set. The compatibility and complexity of the selected model and feature set play a pivotal role in determining the system's performance. In the literature, a variety of machine learning methods are employed for classification tasks. For instance, Feng et al. [20] employed a 3-layer ANN, while Song et al. [21] used SVM to categorize a dataset of 2904 songs collected by Last FM [4] into 4 categories. Liu et al. [13] utilized CNN for both feature extraction and classification tasks.

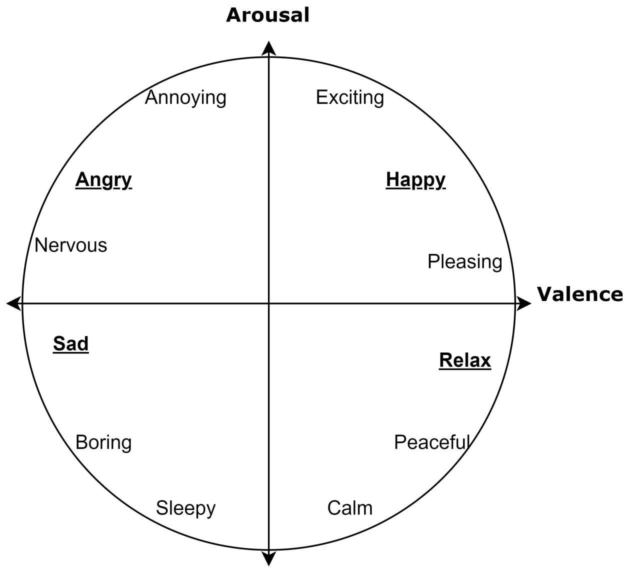

Figure 1: The 2D valence-arousal emotion space

In this study, the feature set presented by Er and Aydilek [9] was employed [22]. The primary focus of this work was on examining the impact of hyperparameter optimization for the selected classifier model on the outcomes.

II. Materials & Methods

A. Dataset

In this study, the feature set created by Er and Aydilek has been utilized. The feature set contains attributes from 400 music tracks, each lasting for 30 seconds, present in the Turkish Music Emotion Recognition database also developed by Er and Aydilek [22]. These attributes encompass a total of 50 distinct measurements, falling within the general categories of energy, MFCCs, Attack Time, Spectral, Chromagram, and Harmonic. For each sound file, the chosen 50 attributes constitute a total of 400 instances, with 100 samples per category. The values within the dataset have not undergone normalization. In this study, to observe the impact of normalization on performance, both raw attributes and normalized attributes using MinMax and Z-score techniques were employed.

The generated feature sets were divided into 70% training and 30% testing data, ensuring an equal distribution within each class. The resulting training set underwent 5-fold Cross-Validation to be applied to the trained model. Performance results were computed using the test data that had not been incorporated into network training or validation.

Given the inability to generate similar data for comparison with the reference study, recommended data augmentation techniques were not employed in this work. Additionally, to establish similarity with the reference study, Softmax and SVM classifiers were used at the output of the CNN model.

The proposed methodology involves determining parameters of the CNN model, such as its depth, complexity, and filter size, through hyperparameter tuning to ensure problem-specific suitability. To achieve this, the Python library named "hyperopt" has been utilized [23]. This approach aims to optimize the architecture of the network in order to align with the intricacies of the given problem. The "hyperopt" library facilitates a methodical exploration of various combinations of hyperparameters, enabling the identification of an optimal configuration that enhances both the performance and generalization capabilities of the model. By customizing the parameters of the CNN to the specific characteristics of the task at hand, this methodology seeks to achieve superior results in terms of accuracy, efficiency, and overall effectiveness.

B. Convolutional Neural Network

The process of convolution involves traversing one matrix over another matrix, calculating the sum of element-wise multiplications at overlapping positions. As observed in Figure 2, when convolution is applied to the two matrices, for each element of the resulting matrix Y, matrix W is slid over matrix X as depicted in equations 1 and 2. To prevent dimension reduction, padding can be applied by adding rows and columns to the outer edges of matrix X, thereby increasing the size of matrix Y, if desired. In fields such as image processing, dealing with large matrices, the reduction of output matrix dimensions is often sought. However, in cases where relatively small-sized data is used, as in this study, padding is commonly employed to prevent excessive reduction in matrix dimensions. In this study, we conducted operations with padding to ensure that the matrix dimensions remained unchanged.

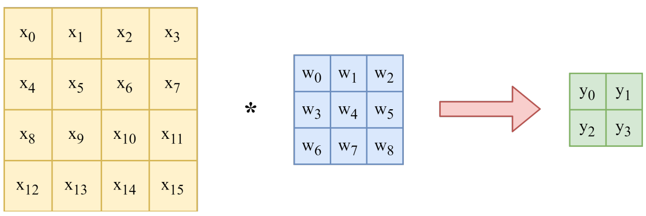

Figure 2: A convolution process example

$$y_0 = x_0w_0 + x_1w_1 + x_2w_2 + x_4w_3 + x_5w_4 + x_6w_5 + x_8w_6 + x_9w_7 + x_{10}w_8 \tag{1}$$

$$y(n) = x(n) \cdot w(n) = \sum_{k=0}^{n} w(k)x(n-k) \tag{2}$$

C. Classification Layer

The CNN model can have a classification layer at its output, or its raw outputs can be passed as feature inputs to another classifier. In this study, in order to compare with the reference work, we obtained results using both an artificial neural network model with a Softmax activation function and an SVM classifier.

Support Vector Machine, also known as Support Vector Networks, is a machine learning method that aims to find the best decision boundary between classes. To achieve this, input vectors need to be transformed using kernel functions that employ nonlinear mapping to a high-dimensional feature space. In this feature space, a linear decision surface can be created to perform classification [24].

For data that cannot be linearly classified, different kernel functions can be used. Polynomial and radial basis functions (RBF) are frequently employed in such processes. The margin boundary represents the distance of the decision boundary set by the SVM to the nearest features. When the decision surface cannot correctly classify all components, some components are allowed to be on the wrong side of the decision surface. This gives rise to a new margin boundary known as the Soft Margin when compared to the rigid margin boundary. Figure 3 illustrates an example SVM decision boundary.

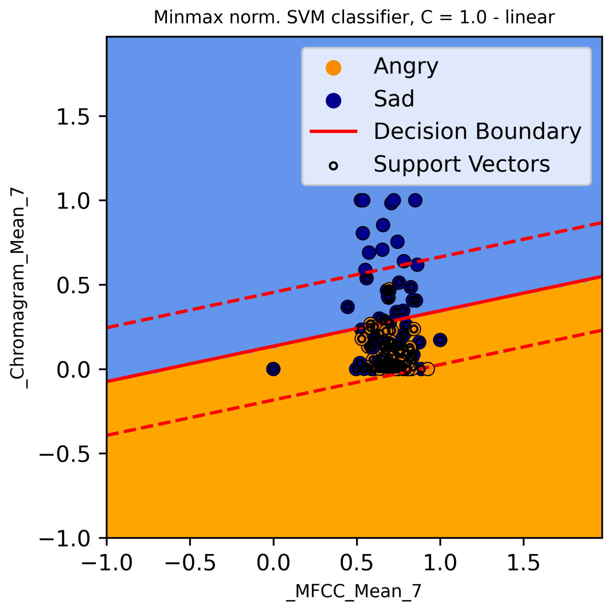

Figure 3: An example of SVM with two features

III. Results & Discussion

A. Results

In this study, a total of six different model performances were investigated using two distinct classifiers and three different sets of normalized attributes. Each model was hyperoptimized for the data applied to its input and the classifier layer at its output. During the model training, the attribute sets were divided into 70% training and 30% test data, and the training data were folded using 5-fold cross-validation. For each model, a randomly selected 4-layered CNN model was utilized as a starting point in the hyperparameter space. The models were trained using the "hyperopt" Python library, and the models that achieved the highest accuracy values in the complexity matrix obtained with the test data were saved.

Following the training and optimization processes, the network model parameters were summarized in Table I based on the utilized classifier and normalization method. The number of filters and kernel size of the four 1D convolution layers, as well as the size of the pooling layer after the convolution layers, are shown in the first nine columns. The empty spaces in the pooling layer column indicate that the pooling process was not selected for that particular model. The last two columns of the table show the number of neurons in the neural network layers.

Complexity matrices for the tested networks after hyperparameter optimization are presented in Figures 4 and 5 for Softmax and SVM classifiers, respectively. These figures demonstrate that data normalization and the choice of normalization method significantly influence classifier performance.

In Table II, the accuracy achievements of our study are presented alongside the accuracy achieved by the reference article for the same dataset. Our study reached an accuracy of 81.7% for the SVM classifier, indicating a 3.1% improvement, and an accuracy of 86.7% for the Softmax classifier, showing a substantial increase of 10.7%.

B. Discussion

In this study, a hyperparameter-optimized CNN network is compared with pre-trained deep learning models' performances on a sample dataset. For this purpose, the work of Er and Aydilek [9], which shares an attribute set and a Turkish music emotion recognition database prepared for music emotion recognition, is taken as a reference.

The reference article demonstrates excellent performance using pre-trained models such as AlexNet and VGG-16 in image processing methods. As emphasized in the study, the advantages of selecting pre-trained models and the potential for improvements in various parameters should be acknowledged. Additionally, it is observed that the suggested hyperparameter optimizations we propose contribute to performance enhancement. For future work, apart from the improvements on raw data suggested by Er and Aydilek [9], investigating performance improvements with different classifier layers could be considered.

Acknowledgements

The author wishes to extend their sincere appreciation to Mehmet Bilal Er and Ibrahim Berkan Aydilek for generously sharing their expertise, which played an instrumental role in shaping the direction of this research.

References

[1] Sachs C. The History of Musical Instruments. Mineola: Dover Publications. 2006

[2] Kim J, Andre E. Emotion recognition based on physiological changes in music listening. IEEE Transactions on Pattern Analysis and Machine Intelligence 2008; 30(12): 2067-2083.

[3] Music Information Retrieval Evaluation eXchange (MIREX) Wiki [Internet]. Available from: https://www.music-ir.org/mirex/wiki/MIREX_HOME

[4] Last.fm [Internet]. Last.fm. Available from: https://www.last.fm/

[5] Eerola T, Vuoskoski JK. A comparison of the discrete and dimensional models of emotion in music. Psychology of Music 2010; 39(1): 18-49.

[6] Mo S, Niu J. A novel method based on OMPGW method for feature extraction in automatic music mood classification. IEEE Transactions on Affective Computing 2019; 10(3): 313-324.

[7] Youngmoo E. Kim, Schmidt EM, Migneco R, Morton BG, Richardson P, Scott JJ, Speck JA, Turnbull D. Music emotion recognition: A state of the art review. International Society for Music Information Retrieval Conference 2010; 255-266.

[8] Yang Y-H, Chen HH. Machine Recognition of Music Emotion: A Review. Association for Computing Machinery Transactions on Intelligent Systems and Technology. 2012;3(3):1-30.

[9] Bilal Er M, Aydilek IB. Music emotion recognition by using chroma spectrogram and deep visual features. International Journal of Computational Intelligence Systems 2019; 12(2): 1622.

[10] Surucu M, Isler Y, Perc M, Kara R. Convolutional neural networks predict the onset of paroxysmal atrial fibrillation: Theory and applications. Chaos: An Interdisciplinary Journal of Nonlinear Science 2021; 31(11): 113119.

[11] Surucu M, Isler Y, Kara R. Diagnosis of paroxysmal atrial fibrillation from thirty-minute heart rate variability data using convolutional neural networks. Turkish Journal of Electrical Engineering and Computer Sciences 2021; 29(SI-1): 2886-2900.

[12] Narin A, Isler Y. Detection of new coronavirus disease from chest x-ray images using pre-trained convolutional neural networks. Journal of the Faculty of Engineering and Architecture of Gazi University 2021; 36(4): 2095-2107.

[13] Liu X, Chen Q, Wu X, Liu Y, Liu Y. CNN based music emotion classification. arXiv. 2017.

[14] Altan G, Kutlu Y, Pekmezci AO, Nural S. The diagnosis of asthma using Hilbert-Huang transform and deep learning on lung sounds. Journal of Intelligent Systems with Applications 2019; 2(2): 100-105.

[15] Balli O, Kutlu Y. Effect of deep learning feature inference techniques on respiratory sounds. Journal of Intelligent Systems with Applications 2020; 3(2): 134-137.

[16] Yang YH, Chen HH. Music Emotion Recognition. CRC Press. 2011

[17] Russell JA. A circumplex model of affect. Journal of Personality and Social Psychology 1980; 39(6): 1161-1178.

[18] Thayer RE. The Biopsychology of Mood and Arousal. Oxford University Press, 1990.

[19] Lartillot O, Toiviainen P, Eerola T. A Matlab toolbox for music information retrieval. Conference paper in Studies in Classification, Data Analysis, and Knowledge Organization 2008; 261-268.

[20] Feng Y, Zhuang Y, Pan Y. Popular music retrieval by detecting mood. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 2003; 375-376.

[21] Yading S, Dixon S, Pearce M. Evaluation of musical features for emotion classification. International Society for Music Information Retrieval Conference 2012; 523-528.

[22] Turkish Music Emotion Dataset [Internet]. [cited 2021 Aug 30]. Available from: https://www.kaggle.com/datasets/blaler/turkish-music-emotion-dataset

Classifier	Norm	Hyperparameters										
		Conv 1		Conv 2		Conv 3		Conv 4		Pool	Dense 1	Dense 2
		Kernel	Filters	Kernel	Filters	Kernel	Filters	Kernel	Filters	Size	Units	Units
Softmax	No Norm.	2	24	2	10	2	28	2	28	-	96	96
Softmax	MinMax	2	24	2	26	2	10	2	34	2	64	96
Softmax	Z_score	2	27	2	22	2	46	2	52	-	448	32
SVM	No Norm.	2	30	2	18	2	34	2	22	2	512	32
SVM	MinMax	2	24	2	14	2	34	2	16	2	384	128
SVM	Z_score	2	30	2	10	2	22	2	34	-	512	96

Table I: CNN Model parameters after hyperparameters optimization

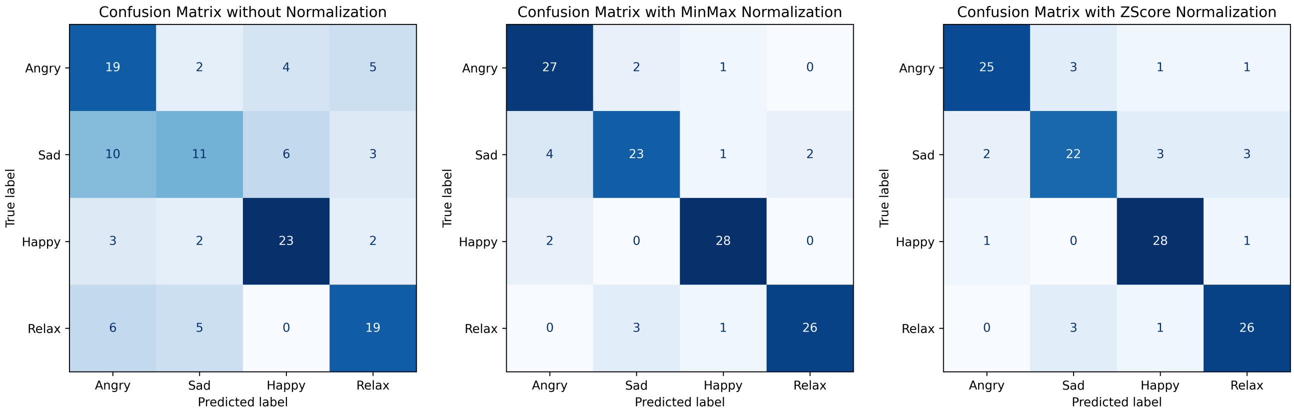

Figure 4: Confusion Matrices for Softmax Classifier

[23] Bergstra JA, Daniel Y, David DC. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. International Conference on Machine Learning, Proceedings of Machine Learning Research 2013; 28(1): 115-123.

[24] Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995; 20(3): 273-297.

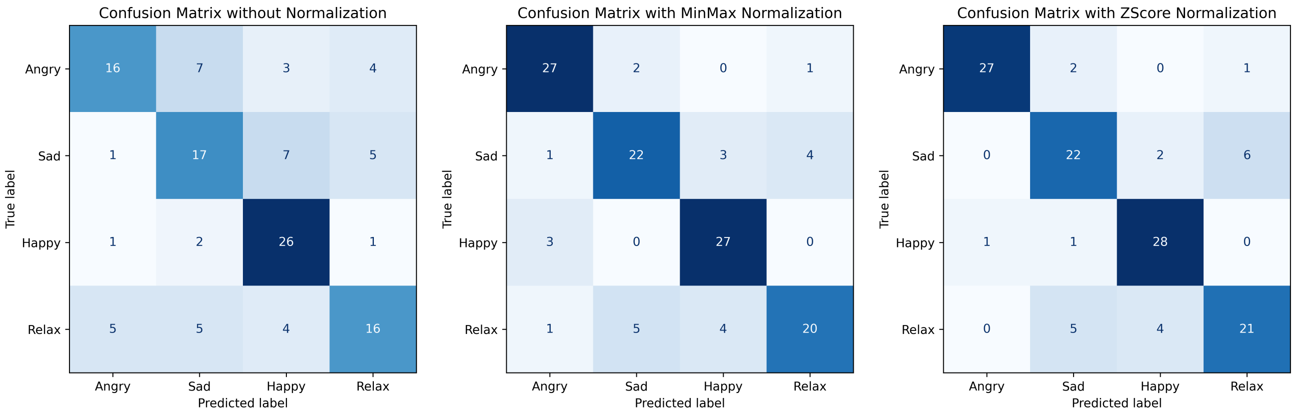

Figure 5: Confusion Matrices for SVM Classifier

Model	Layer	Normalization	Classifier	Train-Test Ratio	Accuracy
Hypertuned CNN	-	No Norm.	SVM	70%–30%	62.5
Hypertuned CNN	-	MinMax Norm.	SVM	70%–30%	80
Hypertuned CNN	**-**	**Z_score Norm.**	**SVM**	**70%–30%**	**81.7**
Hypertuned CNN	-	No Norm.	Softmax	70%–30%	60
Hypertuned CNN	**-**	**MinMax Norm.**	**Softmax**	**70%–30%**	**86.7**
Hypertuned CNN	-	Z_score Norm.	Softmax	70%–30%	84.2
AlexNet	Conv5	-	SVM	70%–30%	58.3
AlexNet	Conv5	-	Softmax	70%–30%	57.5
AlexNet	Fc6	-	SVM	70%–30%	74.0
AlexNet	Fc6	-	Softmax	70%–30%	74.2
AlexNet	Fc7	-	SVM	70%–30%	72.5
AlexNet	Fc7	-	Softmax	70%–30%	73.3
AlexNet	Fc8	-	SVM	70%–30%	68.8
AlexNet	Fc8	-	Softmax	70%–30%	70.8
VGG-16	Conv5_3	-	SVM	70%–30%	61.6
VGG-16	Conv5_3	-	Softmax	70%–30%	58.3
VGG-16	**Fc6**	**-**	**SVM**	**70%–30%**	**78.6**
VGG-16	**Fc6**	**-**	**Softmax**	**70%–30%**	**76.0**
VGG-16	Fc7	-	SVM	70%–30%	73.2
VGG-16	Fc7	-	Softmax	70%–30%	73.3
VGG-16	Fc8	-	SVM	70%–30%	70.0
VGG-16	Fc8	-	Softmax	70%–30%	72.5

Table II: The classification results of both this and the reference work