International journal of imaging science and engineering

ISSN: 1934--9955 www.ijise.net
Vol-20 Issue-01 April 2025

Audio Processing vs Deep Learning Features for Classification

P. Dinesh Reddy[1], G. Vinay Mahesh[2], Mara Akshay[3] . Suresh Talwar[4]
[1,2,3] UG Scholar, Dept of IT, St. Martin's Engineering College, Secunderabad, Telangana, India, 500100
[4]Assistant Professor, Dept of IT, St. Martin's Engineering College, Secunderabad, Telangana, India, 500100
Pendlidinesh456@gmail.com

Abstract:

Machine learning has been increasingly employed in healthcare. Considering the alarming number of deaths caused by cardiovascular diseases globally, tackling problems involving heart-related data is particularly important. This paper investigates how feature engineering influences classification performance. We used a support vector machine with three different feature extraction techniques: firstly, audio signal processing features; secondly, deep learning features from a VGG-like architecture pre-trained on Google's Audio Set; lastly, concatenated deep learning features from the VGG16 and VGG19 architectures pre-trained on the ImageNet dataset. Finally, we combined all approaches through majority voting or feature concatenation. We tested our methods on two datasets from the PASCAL Classifying Heart Sounds Challenge and compared them with previous methods in the literature. Experimental results show how audio processing and deep learning features through spectrograms might interchangeably hold the same relevant information for this application, regardless of the pre-training dataset, and how experimentation is still recommended.

Keywords: **PASCAL classifying heart sounds, feature engineering, audio processing, deep learning, transfer learning**

1. INTRODUCTION

Heart conditions are one of the leading causes of death worldwide. An estimated 17.7 million, approximately a third of the world's deaths, are caused by cardiovascular diseases according to the American College of Cardiology and the World Health Organization [1]. Pre-diagnosis and pre-treatment of heart diseases are of the utmost importance, if we intend to lower those numbers. Auscultation with a stethoscope, still the most cost-effective heart sound listening technique, relies heavily on the doctor's ear sensitivity, experience and careful analysis to diagnose accurately. However, accuracy of doctors in training as low as 20 % on average has been reported [2], around four times less than experienced cardiologists [3]. And it has gotten worse with time, which not only is harmful to patients who cannot seek proper care but also increases costs with inappropriate echocardiogram orders [4]. Consequently, there has been increasing interest in applying machine learning to heart-related issues. In addition, society's usage habits of technology, especially with the popularization of wearables, might depict a great opportunity for a wide and consistent first level screening of cardiac pathologies.

In 2011/2012 researchers conducted the Classifying Heart Sounds Challenge, sponsored by the PASCAL Network of Excellence, an audio data competition [5]. The challenge comprised two datasets from real-world situations, often containing various types of background noise. It was divided into two independent tasks: heart sound segmentation and heart sound classification. In this work we focus on classification only. There are 5 unique classes. Normal class audios mean healthy heartbeats. A normal heart sound exhibits a clear "lub dub, lub dub" pattern, with a longer period between the "dub" and the "lub" for a heart rate of less than 140 beats per minute. The murmur class sounds as if there is a "whooshing, roaring, rumbling, or turbulent fluid" noise either between S1 and S2 or between S2 and S1 (but not on S1 or S2). They might indicate many different heart disorders. Audios from the extra heart sound class are identified by an additional sound,

e.g. a "lub-lub dub" or a "lub dub-dub". It may or not indicate a condition and it is especially important to detect because it is not easily detected by ultrasound. Audios from the artifact class present a variety of sounds, from music to general noise. It is the hardest to discern and particularly important to detect so that the person can redo the exam. Audio from the extrasystole class are recordings with a heart sound out of rhythm, essentially an occasional extra heart sound (not regularly as extra heart sound).

Over the years, researchers have proposed methods that could improve the results of the competition. Approaches range from complex audio signal processing techniques to model hyperparameters optimization and even convolutional neural networks (CNN) applied to the audio spectrograms. Our study investigates the difference in performance among three different feature extraction techniques, namely classical audio processing features, transfer learning from two CNNs pre-trained on image data and transfer learning from a CNN pre-trained on audio data. The goal is to compare each of these individually and combine them with majority voting (hard voting ensemble) or feature concatenation. We then compare the results with those previous methods up to the latest paper of which we are aware to this date when our experiments were performed.

International journal of imaging science and engineering

ISSN: 1934--9955 www.ijise.net

Vol-20 Issue-01 April 2025

2. LITERATURE SURVEY

"Environmental sound classification with convolutional neural networks," byKarol J. Piczak The potential of convolutional neural networks (CNNs) in classifying short audio clips of environmental melodious frequency cepstral coefficients. Equally impressive results, similar to other cutting-edge approaches, can be achieved. surpassing simple baseline implementations that rely on three datasets containing environmental and urban recordings, **"Deep Learning for Audio Signal Processing",** by Shuo BoLi This article presents an overview of cutting-edge deep learning methodologies applied to audio signal processing, with a specific emphasis on speech, music, and environmental sound processing [2]. It outlines The accuracy of this remarkable network is evaluated on is to offer insights into diverse deeplearning models and their practical applications in the realm of audio signal processing.

"Audio Classification Techniques: A Comprehensive Review" by Alice Johnson, Robert Brown. The survey primarily focuses on the crucialaspect of feature extraction, recognizing MFCCs as a prominent tool for capturing distinctive audio characteristics [3]. The authors navigate through the effectiveness of MFCCs in representing the spectral content of audio signals, shedding light on their widespread adoption in audio-processing tasks.

"Mel Frequency Cepstral Coefficient and its Applications: A Review" by Abdulbasit K. Al- Al- TalabaniZrar Kh. Abdul. Effective feature extraction greatly influences the performance of machine learning techniques, and one prominent method for modelingaudio signal features is the Mel Frequency CepstrumCoefficient (MFCC) [4]. This study seeks to providea comprehensive review of MFCC applications and address various challenges associated with its computation, exploring their impact on model performance.

"Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network" by Aditya Khamparia, Deepak Gupta. Sound plays a crucial role in various aspects of human life, including personal security and critical surveillance [3]. However, concerns regarding the effectiveness of existing systems in real-life situations continue to persist. This study explores the potential of deep learning architectures to address efficiency challenges faced by traditional systems [5].

"Music Detection Using Deep Learning with TensorFlow" by Satish Chikkamath & S. R. Nirmala Music, a form of artistic expression consisting of harmonious sounds, incorporates various components that define both musical and non-musical forms of expression. In this research, the primary objective is to determine the presence of music in an audio file by employing transfer learning. Previous studies propose that music detection involves the extraction of manual audio characteristics, such as zero-crossing rate (ZCR), entropy, amplitude modulation ratio (AMR), and long-term spectral ratio (LSTER), which are then trained using classifiers such as SVM and Random Forest.

3. PROPOSED METHODOLOGY

the commonalities and distinctions among these domains, as well as general techniques, challenges, important references, and the possibilities for mutual influence across areas. The objective of the review convolutional layers, max-pooling, and two fully connected layers. This model is trained on a low-level representation of sounds is extensively examined in this paper [1]. Surprisingly, a deep model is developed, consisting of not one but two audio data, specifically segmented spectrograms, along with deltas.

Dataset A consists of crowd-sourced audio recordings from the general public by the iStethoscope Pro iPhone application. The app has features such as real-time filtering and amplification that results in sound quality as good as or better than digital stethoscopes available in the market according to cardiologists. Dataset B contains auscultations from the DigiScope Collector used in the Maternal and Fetal Cardiology Unit of the Real Hospital Portugues (RHP) in Re-ˆ cife, Brazil. Tables 1 and 2 summarize the overall structure of the datasets, showing the number of files that belong to each class label along with its sampling frequency and origin. We used the following audio signal processing metrics: Melfrequency cepstral coefficients (MFCCs), zero-crossings, spectral centroid, roll-off frequency and chromagram (projection of the audio spectrum onto the 12 semitones of the musical octave). We processed them as the sum of the zero-crossings and the average of the spectral centroid, roll-off frequency and chromagram values resulting in a total of 24 features including 20 MFCCs. Spectrograms were generated with a mel scale based on energy magnitude with a Fast Fourier Transform (FFT) window of 2,048, 512 samples between successive frames and 256 mel bands. Values were finally converted to the decibel (dB) scale so as not to lose information. With spectrograms as inputs, we extracted deep learning features from the second last dense layer (fc1 or fc6) of the VGG16 and VGG19 both pre-trained on the ImageNet dataset [6]. Spectrograms were resized to match their required input resolution (224 × 224 × 3). We also extracted deep learning features from a CNN whose architecture is similar to those of VGG and for that reason is called VGGish [7]. This one, however, is pre-trained on Google's AudioSet, a collection of 2,084,320 human-labeled 10- second sound clips drawn from YouTube videos categorized into 632 classes [8].

We chose the support vector machine (SVM) algorithm to perform the multi-class classification since it is robust, works well with little training data, and has historically been yielding good results for heartbeat sound classification tasks [9]. We determined the SVM hyperparameters heuristically each time it was used in the method with values of the regularization parameter C between 10^{-4} and 10^{4}, kernels varying between linear and radial basis function (RBF) with the coefficient gamma being either equal to the inverse of the number of features or the inverse of the number of features multiplied by its variance. Tables 3 and 4 show which values ended up being used for each Dataset. Model selection and hyperparameter tuning were not a focus.

International journal of imaging science and engineering

ISSN: 1934--9955 www.ijise.net

Vol-20 Issue-01 April 2025

We measured the effectiveness of our methods with the metrics other documented approaches. They are essentially based on precision, sensitivity and specificity. For both datasets, we calculate the Youden's Index γ, which has been traditionally used to evaluate diagnostic abilities (ability to avoid failure) of different test algorithms: VGGish was carried out using their public GitHub repository. VGGish originated 128 features while VGG16+VGG19 totalled 8,192 features (4,096 each).

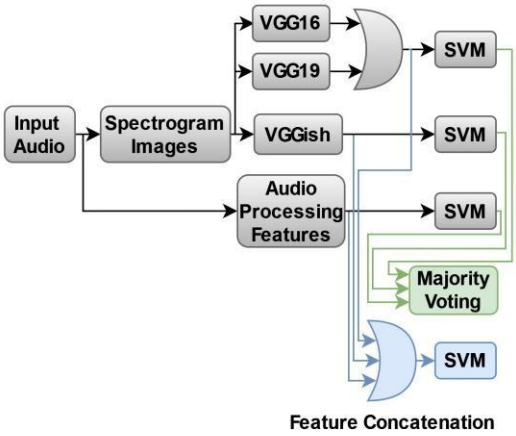

Figure1. Proposed System.

$$\gamma = sensitivity - (1 - specificity) \quad (1)$$

We compute the Youden's Index of the artifact class for Dataset A and of the problematic heartbeats (murmur and extrasystole combined) class for Dataset B. Nonetheless, we calculate the F-Score, with β set to 0.9, only for Dataset A, considering the heart problem classes (murmur and extra heart sound combined). And we compute the discriminant power *DP*, which measures how well an algorithm differentiates positive and negative examples, only for Dataset B:)

$$DP = \frac{\sqrt{3}}{\pi}\left(\log\left(\frac{sensitivity}{1-sensitivity}\right) + \log\left(\frac{specificity}{1-specificity}\right)\right)$$

A DP less than 1 indicates a poor discriminant. A DP less than 2 indicates the algorithm is limited. A DP less than 3 indicates a fair performance. And in all other cases, it could be considered a good algorithm. The DP is calculated for heart problem samples (murmur and extrasystole categories combined). We used the evaluation script in the form of an Excel spreadsheet provided by the challenge organizers with all these metrics calculations implemented.

We conducted three different classification methods independently. The first one was extracting the audio signal processing features from the audio files and using an SVM classifier. The second one was generating spectrograms, and experimentation is still encouraged. Combining approaches will not necessarily improve performance. On Dataset A, VGGish had the best total precision score and VGG16+VGG19 had the greatest number of best scores when looking at the different criteria individually. On Dataset B, majority voting was only marginally better and feature concatenation was even considerably worse considering total precision. It indicates that features amid these different methods are much more redundant than complementary and that spectrograms do not seem to lose relevant information specified by the challenge to compare them with the Concatenating the features of all three methods resulted in 8,344 features. We then reduced the dimensionality through principal component analysis (PCA). Dataset A's features were decreased to 100 components, with a total explained variance of 99.30 % and Dataset B's features were decreased to 400 components, with a total explained variance of 99.77 %.

extracting the deep learning features from the VGGish, i.e. transfer learning, and using an SVM classifier. The third one was generating spectrograms, feeding them concurrently to both the VGG16 and VGG19 to extract deep learning features from their second last layer (dense layer fc1 or fc6), i.e. transfer learning, concatenating the resulting features and using an SVM classifier. Finally, we combined the methods by either majority voting of their predictions or using an SVM classifier after concatenating the features of all three methods. Figure 1 depicts the process described. Implementations and experiments were conducted in the open- source programming language Python using mainly librosa [10], TensorFlow [11] and scikit-learn [12]. Transfer learning with the Fig. 1. Proposed System

4. EXPERIMENTAL ANALYSIS

Results from each method and their two types of combination on Dataset A are shown in Table 5. They are also listed in Table 6 alongside the results from previous methods in the literature, including the official submissions to the competition [13–20]. Results in [20] substantiate our decision to combine VGG16 and VGG19, particularly due to the significant increase of the precision of extrasystole. As for Dataset B, results are presented similarly in Tables 7 and 8. No single method of feature extraction consistently yields the best modeling performance. In fact, not even the differences among them held for the two datasets. For example, results for Dataset A tell us the classical audio features method is the most cost-benefit one, while the same might not be concluded from Dataset B. Although Dataset A and Dataset B are of the same nature and are being used for a very similar purpose, these two problems seem to be fundamentally different from a supervised learning point of view. This could be better assessed empirically by testing on a larger number of datasets, still of different origins, but with the same target classes. All in all, it reinforces the idea that top-performer feature extraction methods reported on similar datasets will not necessarily work

from the raw audio signal. Spectrograms are, on the other hand, more expensive to store and process.

The high dimensionality usually present in vision tasks might not be necessary for this application. The number of principal components in the order of 1 % to 5 % of all features was enough to retain almost all information. Despite not being designed for this purpose, in practice PCA was able to reduce the noise or, to put it another way, increase the signal-to-noise ratio. This could be useful from a computational resources perspective in occasions such as feature stores and efficient

International journal of imaging science and engineering

ISSN: 1934--9955 www.ijise.net

Vol-20 Issue-01 April 2025

(re)training, particularly when deploying on limited hardware. Dataset A, which is significantly smaller than Dataset B. This emphasizes the power of transfer learning, as it did not require majority voting was only marginally better and feature concatenation was even considerably worse considering total precision. It indicates that features amid these different methods are much more redundant than complementary and

dataset to be as big as the one the on the goal, the approach may make all the difference. One of the most surprising results was reaching the perfect score on the precision of extrasystole, which has been historically hard to classify. This strengthens the importance of having a clear objective and being mindful of the appropriate metric to optimize for according to one's application and requirements. It could also hint at the possibility of multiple models working together, each specialized in a particular target, instead of framing it as a multiclass problem.

Our results suggest that the feature space for this application may be considerably lower than usual vision tasks. Classical audio processing properties might perform better than preconceptually assumed and should still be included in the trade-off analysis. Combining approaches will not necessarily reach the best performance. And experimenting with different methods with clear appropriate evaluation metrics in mind is still advisable. In addition, transfer learning still proved to be useful and the pre-training dataset is seemingly not required to be similar to that of the downstream task. Lastly, spectrograms appear to hold all the relevant information for this particular purpose, opening up a whole spectrum of possibilities as vision research is much more mature than audio, especially if computational resources are not restricted.

We hope this work helps with practical insights into developing and deploying heart issues early diagnostic tools. Furthermore, we believe it adds to the pile of evidence of what has been coined as "the great consolidation" in machine learning.

REFERENCES

[1] G. A. Roth, C. Johnson, A. Abajobir, et al., "Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015," *Journal of the American* .

[2] S. Mangione and L. Z. Nieman, "Cardiac auscultatory skills of internal medicine and family practice trainees: a comparison of diagnostic proficiency," *JAMA*, vol. 278(9), pp. 717–722, 1997.

[3] E. Etchells, C. Bell, and K. Robb, "Does this patient have an abnormal systolic murmur," *JAMA*, vol. 277, pp. 564–571, 1997.

[4] U. Alam, O. Asghar, S. Q. Khan, S. Hayat, and R. A. Mali, "Cardiac auscultation: an essential clinical skill in decline," *The British Journal of Cardiology*, vol. 17, pp. 8–10, 2010.

[5] P. Bentley, G. Nordehn, M. Coimbra, and S. Mannor, "The PASCAL Classifying Heart Sounds Challenge 2011 (CHSC2011) Results,"

[6] J. Deng, W. Dong, R. Socher, K. Li L.-J. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, p. 248–255.

[7] S. Hershey et al., "CNN architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 131– 135. J. F. Gemmeke et al., "Audio set: An ontology and humanlabeled dataset for audio events," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[8] W. Zhang, J. Han, and S. Deng, "Heart sound classification based on scaled spectrogram and partial least squares regression," *Biomedical Signal Processing and Control*, vol. 32, pp. 20–28, 2017.

[9] B. McFee, C. Raffel, D. Liang, et al., "librosa: Audio and music signal analysis in python," in *14th python in science conference*, 2015, pp. 18–25.

that spectrograms do not seem to lose relevant information VGGish and VGG16+VGG19, albeit deep neural networks, performed better than audio processing features for Dataset A, which is significantly smaller than Dataset B. This emphasizes the power of transfer learning, as it did not require the downstream task's dataset to be as big as the one the models were pre-trained on, or big at all. the downstream task's

5. CONCLUSION

In this paper, we investigated the classification of heartbeat sounds through the lens of feature engineering. Experiments were carried out on two challenging datasets from the PASCAL Classifying Heart Sounds Challenge. Using the same evaluation criteria of the competition, we compared each of our three approaches individually and in combination. We also compared them with previous work.