International Journal of Advanced Research and Review

www.ijarr.org

Evaluating and Integrating Deep Learning and Audio Processing Capabilities for Heartbeat Sound Classification

[1] Azim, [2] V. Prathyusha,

[1] Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.
[2] MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

ABSTRACT

The use of machine learning in healthcare has been on the rise. It is of utmost importance to address issues linked to heart-related statistics in light of the concerning number of fatalities worldwide caused by cardiovascular illnesses. The effect of feature engineering on classification accuracy is explored in this work. A support vector machine was equipped with three distinct feature extraction methods: first, features extracted from audio signal processing; second, features extracted from a VGG-like architecture that had been pre-trained on Google's AudioSet; and lastly, features extracted from the ImageNet dataset that had been concatenated with features extracted from the VGG16 and VGG19 architectures. Last but not least, we used feature concatenation or majority vote to merge all methods. We compared our approaches to those in the literature and ran tests on two datasets from the PASCAL Classifying Heart Sounds Challenge. The experimental findings demonstrate that spectrograms used in deep learning and audio processing might potentially store the same pertinent information for this application, independent of the pre-training dataset. It is encouraged to do experiments to confirm this. Classification of cardiac sounds using PASCAL, feature engineering, deep learning, and transfer learning are all terms included in the index.

1. INTRODUCTION

One of the main reasons people die all across the globe is heart disease. The World Health Organization and the American College of Cardiology report that cardiovascular illnesses account for over one-third of all deaths globally, which amounts to around 17.7 million people [1]. If our goal is to reduce those numbers, it is crucial to prioritize the early detection and treatment of cardiac disorders. The most economical method of listening to heart sounds—auscultation using a stethoscope—depends significantly on the doctor's ear sensitivity, expertise, and meticulous analysis for an accurate diagnosis. On the other hand, compared to seasoned cardiologists, medical students' accuracy might be as low as 20% on average [2]. Not only has it worsened with time, but it also raises expenses because to improper echocardiography orders, which is bad for patients since they can't get the treatment they need [4]. Consequently, the use of machine learning to problems involving the heart has garnered more and more attention. Another potential option for widespread and consistent first-level screening of cardiac diseases may lie in society's digital use patterns, particularly with the rise of wearables. The Classifying Heart Sounds Challenge was an audio data competition that took place in 2011 and 2012 and was sponsored by the PASCAL Network of Excellence [5]. Two datasets representing real-world scenarios, each with its own unique kind of background noise, made up the task. Heart sound segmentation and classification were the two separate tasks. Only classification is addressed in this study. Each of the five classes is distinct. A normal class audio indicates a heartbeat that is within a healthy range. At a heart rate below 140 beats per minute, a typical heart sound has a longer time between the "dub" and the "lub" sound, forming a distinct "lub dub, lub dub" pattern. Between S1 and S2 or S2 and S1 (but not on S1 or S2), the murmur class produces an acoustic signature like a "whooshing, roaring, rumbling, or turbulent fluid"

sound. They could be symptoms of a variety of cardiac conditions. A "lub-lub dub" or "lub dub-dub" is an example of an extra sound that identifies an audio file as belonging to the extra heart sound class. Because ultrasonography has a hard time picking it up, finding it is crucial, even if it might be a sign of a problem or not. The audios provided by the artifact class include a wide range of noises and musical compositions. Finding it so the individual may retake the test is crucial, but it is also the most difficult. Records belonging to the extrasystole category include a heart sound that is not in sync with the rest of the recording, or what is effectively an extra heart sound that occurs from time to time but is not present consistently. Scientists have been trying to find ways to make the competition better for a long time now. Various methods are used, including convolutional neural networks (CNN) applied to the audio spectrograms, optimization of model hyperparameters, and complicated audio signal processing techniques. Classical audio processing features, transfer learning from two convolutional neural networks (CNNs) pre-trained on picture data, and transfer learning from a CNN pre-trained on audio data are the three feature extraction strategies that are compared in our research. The plan is to evaluate them one by one and then merge them using feature concatenation or majority vote ensemble. Afterwards, we evaluate the outcomes in relation to those earlier approaches, all the way up to the most recent publication that came to light during the time our trials were conducted.

2. METHODOLOGY

2.1. Datasets

The iStethoscope Pro iPhone app collected audio recordings from the general population and is part of Dataset A. Cardiologists have found that the app's sound quality is on par with or even better than that of commercially available digital stethoscopes, thanks to features like real-time filtering and amplification. In the Maternal and Fetal Cardiology Unit of the Real Hospital Portugũes (RHP) in Recife, Brazil, auscultations were recorded using the DigiScope Collector, which are included in Dataset B. The datasets are summarized in Tables 1 and 2, which provide the total number of files for each class label, together with the sampling frequency and provenance of those files.

Table 1. Dataset A structure

Class	Quantity	Audio Information
Normal	31	iStethoscope (iPhone) 44.1 kHz
Murmur	34	
Extra Heart Sound	19	
Artifact	40	
Unlabeled	52	
Total	176	

Table 2. Dataset B structure

Class	Quantity	Audio Information
Normal	320	Digital Stethoscope 4 kHz
Murmur	95	
Extrasystole	46	
Unlabeled	195	
Total	656	

2.2. Audio Processing Features

The following metrics were used in the analysis of the audio signals: spectral centroids, zero-crossings, midfrequency cepstral coefficients (MFCCs), roll-off frequency, and chromogram, which is a projection of the audio spectrum onto the 12 semitones of the musical octave. We averaged the spectral centroid, roll-off frequency, and chromogram values, added up the zero-crossings, and got 24 features, 20 of which were MFCCs. 2.3. Advanced Models A 256-mel band spectrogram was produced using an FFT window of 2,048 samples, 512 samples between each frame, and an energy-based mel scale. In the end, values were transformed to the decibel (dB) scale to ensure no data was lost. By feeding spectrograms into the VGG16 and VGG19, which were both pre-trained on the ImageNet dataset, we were able to extract deep learning features from the second to last dense layer (fc1 or fc6) [6]. In order for the spectra to fit their input resolution specifications (224 × 224 × 3), they were scaled. We also used a

CNN dubbed VGGish to extract deep learning features since its design is comparable to VGG's [7]. But this one has already been trained using data from Google's AudioSet, which contains 2,084,320 10-second audio snippets annotated by humans and extracted from 632 different types of YouTube videos [8]. Classifiers (2.4) Because of its reliability, ability to function with little training data, and track record of success with heartbeat sound classification tasks, we choose to use the support vector machine (SVM) method for the multi-class classification [9]. Every time we used the SVM in our method, we heuristically set its hyperparameters. The regularization parameter C could have values between 10−4 and 104, and the kernels could be either linear or radial basis function (RBF). The coefficient gamma could be either equal to the inverse of the number of features or the inverse of the number of features multiplied by its variance. You can see the final values utilized for each dataset in Tables 3 and 4. Hyperparameter tweaking and model selection were not priorities.

Table 3. SVM hyperparameters for Dataset A

Method	C	Kernel	Gamma
Audio Features	100	rbf	auto
VGGish	1	rbf	scale
VGG16+VGG19	0.001	linear	-
Feature Concatenation	1	linear	-

Table 4. SVM hyperparameters for Dataset B

Method	C	Kernel	Gamma
Audio Features	9	rbf	auto
VGGish	1	rbf	scale
VGG16+VGG19	5	rbf	auto
Feature Concatenation	0.001	linear	-

2.5. Evaluation Criteria

In order to compare our techniques to the other described alternatives, we used the metrics stated by the challenge to assess their efficacy. Their foundation is a focus on detail, sensitivity, and accuracy. In order to assess the diagnostic capabilities (i.e., the capacity to prevent failure) of various test methods, we compute the Youden's Index γ for both datasets.

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

Dataset A's artifact class and Dataset B's problematic heartbeats (murmur and extrasystole combined) classes are each given by Youden's Index, which is computed. However, we just apply the F-Score calculation to Dataset A, taking into account the heart issue classifications (murmur and additional heart sound combined), with β set to 0.9. The discriminant power (DP), which is a measure of an algorithm's ability to distinguish between positive and negative samples, is computed only for Dataset B:

$$DP = \frac{\sqrt{3}}{\pi}\left(\log\left(\frac{sensitivity}{1 - sensitivity}\right) + \log\left(\frac{specificity}{1 - specificity}\right)\right) \qquad (2)$$

A discriminant with a DP below 1 is not very effective. The algorithm is restricted if the DP is less than 2. An acceptable performance is indicated by a DP lower than 3. And it may be said to be an excellent algorithm in any other scenario. For samples including cardiac issues (including both murmur and extrasystole categories combined), the DP is computed. The challenge organizers supplied us with an Excel document that included the evaluation script, which included all of these metrics computations. 2.6. Controlled Trials On their own, we ran three distinct categorization algorithms. The first step was utilizing a support vector machine (SVM) classifier to extract characteristics related to audio signal processing from the original audio recordings. The second one included creating spectrograms, using a support vector machine (SVM) classifier to extract deep learning features from a VGGish (transfer learning), and so on. Third, we used an SVM classifier to combine the information we extracted

from the deep learning layers of VGG16 and VGG19—their second-to-last layers, dense layers fc1 and fc6, respectively—after creating spectrograms. The last step was to merge the algorithms by combining their features or by utilizing a majority vote for their forecasts. We then used an SVM classifier. The procedure is shown in Figure 1. The open-source programming language Python was used for the implementations and experiments, with librosa, TensorFlow, and scikit-learn being the major tools used. The VGGish was used for transfer learning via their public GitHub repository. In contrast to VGGish's 128 features, VGG16 and VGG19 combined for 8,192 features (4,096 each). There were 8,344 features total after merging the data from all three approaches. After that, we used principal component analysis (PCA) to lower the dimensionality. After reducing the number of features in Dataset A to 100 components, the total explained variance was 99.50%, whereas the number of features in Dataset B was reduced to 400 components, with a total explained variance of 99.997%.

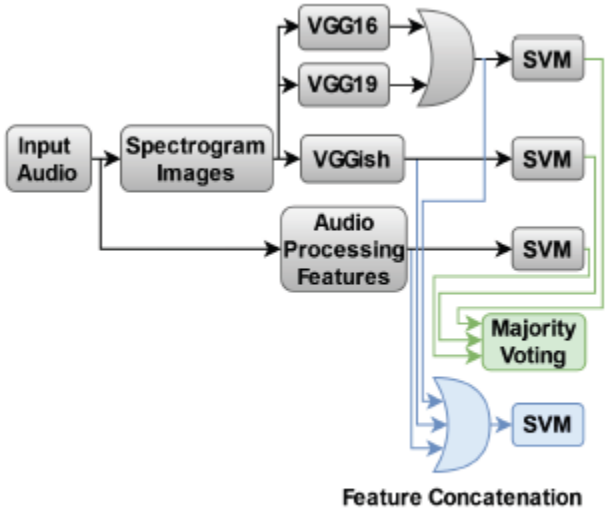

Fig. 1. Schematic diagram of our experiments.

3. RESULTS AND DISCUSSION

Table 5 displays the results obtained from all methods and the two combinations of them on Database A. Along with the findings from earlier approaches in the literature, including the official entries to the competition [13–20], they are also given in Table 6. The substantial improvement in extrasystole accuracy, as shown in [20], supports our choice to merge VGG16 and VGG19. Tables 7 and 8 provide the findings in a manner similar to that of Dataset B. When it comes to modeling, there is no clear winner when it comes to feature extraction methods. Actually, not even their differences applied to the two sets of data. The classical audio features technique may not be the most cost-effective option, according to Dataset B, but according to Dataset A, it is. From a supervised learning perspective, these two issues seem to be essentially distinct, despite the fact that Datasets A and B are identical in nature and serve very comparable purposes. Testing on a larger number of datasets, still from diverse sources but with the same goal classes, might provide a more accurate empirical assessment of this. In sum, it supports the premise that one should still experiment with feature extraction approaches as top-performing methods published on comparable datasets may not always work so well for one's application. It is not guaranteed that performance will be improved by combining methodologies. Among the datasets tested, VGGish achieved the highest overall precision score, while VGG16+VGG19 topped the charts for the most number of best scores across all criteria. When looking at overall accuracy, feature concatenation performed far worse on Dataset B than majority voting, which performed just slightly better. Spectrograms seem to retain all of the pertinent information from the original audio signal, and characteristics across these various approaches are more overlapping than complementing. The storage and processing costs of spectrograms are higher, however. This use case may not need the high dimensionality often associated with visual activities. The amount of principle components, which ranged from 1% to 5% of all characteristics, was sufficient to preserve almost all of the data. In spite of PCA's lack of intended use, it was able to

improve the signal-to-noise ratio by decreasing the amount of background noise. From the standpoint of computing resources, this might be helpful for situations like efficient (re)training and feature storage, especially when working with constrained hardware. Deep neural networks VGGish and VGG16+VGG19 outperformed audio processing features on the much smaller Dataset A compared to Dataset B. This highlights the efficacy of transfer learning, as it was not necessary for the dataset used for the downstream job to be as large as the one used for pre-training the models, or even large at all. The strategy could be essential depending on the objective. Obtaining a flawless score on the extrasystole precision—a finding that has been historically difficult to categorize—was one of the most unexpected outcomes. This highlights the need of knowing what you want out of an optimization effort and keeping in mind which statistic is most relevant to your specific use case and needs. As an alternative to seeing it as a multiclass issue, it may suggest the use of a combination of models, with each model focusing on a different objective.

Despite what was said, the proximity of the pre-training dataset to the downstream task's dataset is not necessary. Compared to VGG16+VGG19, which was pretrained on image data, we anticipated that VGGish, which was trained on audio data, would perform better. This might indicate that after the spectrograms transform it into a visual task, the model's pattern-spotting abilities are the most important factor, as it was not the case for Dataset B and not universally for Dataset A. In conclusion, in Dataset B, majority voting outperformed CNN-SVM, however on Dataset A, neither a combination nor the top scorer VGGish were able to overcome it. To further compare the methods, it would be interesting to record other metrics like training and prediction runtime and memory use. We anticipate transformer-based designs to provide state-of-the-art performance, if that is the goal, even if we also suggest extending this study by concentrating on model optimization.

4. CONCLUSIONS

Here, we used feature engineering as a prism to examine the categorization of heartbeat sounds. Two difficult PASCAL Classifying Heart Sounds Challenge datasets were used for the experiments. We evaluated all three of our strategies separately and in tandem using the same standards as the contest. Additionally, we contrasted them with prior efforts. Our findings imply that compared to typical vision tasks, the feature space for this application would be much smaller. It is important to incorporate classical audio processing qualities in the trade-off analysis even if they may not perform as well as first thought. Even if you combine methods, you may not get the optimum results. It is nevertheless recommended to experiment with various ways while keeping clear and acceptable assessment criteria in mind. It seems that there is no need for the pre-training dataset to be identical to the downstream task, yet transfer learning still showed usefulness. Finally, spectrograms seem to include all the pertinent data for this project, which opens up a world of possibilities since visual research is much more advanced than audio research, particularly when it comes to the availability of computer resources. Our goal in doing this research was to provide useful information for creating and implementing early diagnostic tools for cardiac problems. What has been dubbed "the great consolidation" in machine learning is further supported by this, in our opinion.

Table 5. Results for Dataset A

Dataset A	Method				
Evaluation Criterion	**Audio Features**	**VGGish**	**VGG16+VGG19**	**Majority Voting**	**Feature Concatenation**
Precision of Normal	0.64	0.53	**0.73**	0.57	0.69
Precision of Murmur	**0.79**	0.71	0.77	0.77	0.77
Precision of Extrasound	0.71	**1.00**	0.50	0.80	0.50
Precision of Artifact	0.80	0.89	**1.00**	0.80	**1.00**
Sensitivity of Artifact	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Specificity of Artifact	0.64	0.61	**0.67**	0.61	**0.67**
Sensitivity of Heart Problem	**0.73**	0.59	**0.73**	0.64	0.68
Precision of Heart Problem	0.76	0.76	0.64	**0.78**	0.65
Youden Index of Artifact	0.64	0.61	**0.67**	0.61	**0.67**
F-Score of Heart Problem	**0.33**	0.30	0.30	0.32	0.30
Total Precision	2.94	**3.13**	3.00	2.94	2.96

Values in bold are the best scores among methods for each criterion.

Table 6. Comparison of obtained results with other methods on Dataset A

Dataset A	Previous Methods in the Literature									Our Combined Methods	
Evaluation Criterion	J48 [13,14]	MLP [13,14]	CS-UCL [13,15]	SS [16]	SS-PLSR [16]	2D-PCA [17]	SS-TD [18]	SVM-DM [19]	CNN-SVM [20]	Majority Voting	Feature Concatenation
Precision of Normal	0.25	0.35	0.46	0.67	0.60	0.56	0.67	0.62	0.59	0.57	**0.69**
Precision of Murmur	0.47	0.67	0.31	0.91	0.91	0.91	**1.00**	**1.00**	0.77	0.77	0.77
Precision of Extra Heart Sound	0.27	0.18	0.11	0.37	0.44	0.30	0.43	**1.00**	0.83	0.80	0.50
Precision of Artifact	0.71	0.92	0.58	0.76	0.94	0.94	0.80	0.64	**1.00**	0.80	**1.00**
Sensitivity of Artifact	0.63	0.69	0.44	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Specificity of Artifact	0.39	0.44	0.44	0.58	0.64	0.58	0.64	0.58	**0.69**	0.61	0.67
Youden Index of Artifact	0.01	0.13	-0.09	0.58	0.64	0.58	0.64	0.58	**0.69**	0.61	0.67
F-Score of Heart Problem	0.20	0.20	0.14	0.28	0.30	0.26	0.30	0.31	**0.33**	0.32	0.30
Total Precision	1.71	2.12	1.47	2.71	2.89	2.80	2.90	3.17	**3.19**	2.94	2.96

Values in bold are the best scores among methods for each criterion.

Table 7. Results for Dataset B

Dataset B	Method				
Evaluation Criterion	Audio Features	VGGish	VGG16+VGG19	Majority Voting	Feature Concatenation
Precision of Normal	0.76	0.76	**0.78**	0.77	**0.78**
Precision of Murmur	0.61	0.85	0.79	**0.86**	0.79
Precision of Extrastole	0.50	0.50	**1.00**	**1.00**	0.50
Sensitivity of Heart Problem	**0.34**	0.31	**0.34**	0.32	**0.34**
Specificity of Heart Problem	0.90	0.97	0.97	**0.98**	0.96
Youden Index of Heart Problem	0.24	0.28	**0.31**	0.30	0.30
Discriminant Power	0.38	0.64	0.68	**0.73**	0.62
Total Precision	1.87	2.11	2.57	**2.63**	2.07

Values in bold are the best scores among methods for each criterion.

Table 8. Comparison of obtained results with other methods on Dataset B

Dataset B	Previous Methods in the Literature									Our Combined Methods	
Evaluation Criterion	J48 [13,14]	MLP [13,14]	CS-UCL [13,15]	SS [16]	SS-PLSR [16]	2D-PCA [17]	SS-TD [18]	SVM-DM [19]	CNN-SVM [20]	Majority Voting	Feature Concatenation
Precision of Normal	0.72	0.7	0.77	0.74	0.76	0.78	**0.83**	0.77	0.81	0.77	0.78
Precision of Murmur	0.32	0.3	0.37	0.66	0.65	0.57	0.7	0.76	0.76	**0.86**	0.79
Precision of Extrasystole	0.33	0.67	0.17	0.24	0.33	0.23	0.15	0.5	0.56	**1.00**	0.50
Sensitivity of Heart Problem	0.22	0.19	0.51	0.24	0.34	0.41	0.49	0.34	**0.54**	0.32	0.34
Specificity of Heart Problem	0.82	0.84	0.59	0.84	0.9	0.84	0.84	0.95	0.91	**0.98**	0.96
Youden Index of Heart Problem	0.04	0.02	0.01	0.13	0.24	0.24	0.33	0.29	**0.45**	0.30	0.30
Discriminant Power	0.05	0.04	0.09	0.24	0.36	0.3	0.39	0.54	**0.6**	0.73	0.62
Total Precision	1.37	1.67	1.31	1.57	1.75	1.58	1.68	2.03	2.15	**2.63**	2.07

Values in bold are the best scores among methods for each criterion.

5. 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 College of Cardiology, vol. 70(1), pp. 1–25, 2017.

[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," http://www.peterjbentley.com/heartchallenge/index.html.

[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.

[8] 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.

[9] 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.

[10] 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.

[11] M. Abadi, A. Agarwal, P. Barham, et al., "Tensorflow: Largescale machine learning on heterogeneous distributed systems," http://download.tensorflow.org/paper/whitepaper2015.pdf, 2015.
[12] L. Buitinck, G. Louppe, M. Blondel, et al., "API design for machine learning software: experiences from the scikit-learn project," in European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2013, p. 108–122.
[13] E. F. Gomes, P. J. Bentley, M. Coimbra, E. Pereira, and Y. Deng, "Classifying heart sounds: Approaches to the pascal challenge," in International Conference on Health Informatics, 2013, p. 337–340.
[14] E. F. Gomes and E. Pereira, "Classifying heart sounds using peak location for segmentation and feature construction," in Workshop Classifying Heart Sounds, 2012, p. 480–492.
[15] Y. Deng and P. J. Bentley, "A robust heart sound segmentation and classification algorithm using wavelet decomposition and spectrogram," in Workshop Classifying Heart Sounds, 2012, p. 1–6.
[16] S. Deng and J. Han, "Towards heart sound classification without segmentation via autocorrelation feature and diffusion maps," Future Generation Computer Systems, vol. 60, pp. 13–21, 2016.
[17] L. D. Avendano-Valencia, J. I. Godino-Llorente, M. Blanco- Velasco, and G. Castellanos-Dominguez, "Feature extraction from parametric time-frequency representations for heart murmur detection," Annals of Biomedical Engineering, vol. 38(8), pp. 2716–2732, 2010.
[18] S. C. Oliveira, E. F. Gomes, and A. M. Jorge, "Heart sounds classification using motif based segmentation," in 18th International Database Engineering & Applications Symposium. Association for Computing Machinery, 2014, p. 370–371.
[19] W. Zhang, J. Han, and S. Deng, "Heart sound classification based on scaled spectrogram and tensor decomposition," Expert Systems with Applications, vol. 84, pp. 220–231, 2017.
[20] F. Demir, A. Sengur, V. Bajaj, et al., "Towards the classification of heart sounds based on convolutional deep neural network," Health Information Science and Systems, vol. 7(16), 2019.