Unimodal Methods for DeepFake Audio and Video Classification with Spectral Features

Derrick Liu, Jack Smith, Kevin So, Andi Xu, Jiayi Zhang
{dsliu, jacktsmi, kvso, andixu, jiayizzz}@umich.edu

Abstract

The work of Durall et al [1] showed that spectral features make deepfake images highly separable from real images. Simple classification methods thus achieve near perfect accuracy when trained on enough high-resolution data. In this paper, we extend this idea to show its viability for both audio and video deepfake classification. We then build and compare two different models for multiclass deepfake classification, each with two unimodal classifiers (audio and video). One of the models uses spectral feature extraction and simple classification techniques for each modality while the other uses deep learning. The spectral featurization model achieves only slightly worse accuracy with a dramatically smaller computational cost.

1 Introduction

Photo manipulations have existed for a long time, but the introduction of Generative Adversarial Networks (GANs) brought about a new type of image manipulation known as "deepfakes". Trained on massive amounts of data, GANs are able to produce novel images that could easily fool humans at first glance. While deepfakes have useful applications in fields such as film production, the same technology can be used in nefarious ways. Public personalities, such as celebrities and politicians, are especially vulnerable to having their likeness used as training data for GANs.

Deepfake detection and classification is still an active research area, and most existing methods are based on deep learning techniques, where classifiers are trained on large amounts of images. Durral et al [1] demonstrate an alternative method by exploiting artifacts in the spectral domain. Their work showed that deepfake images tend to have a lot of power in higher spatial frequencies, which allows them to be easily distinguished from natural images.

In this paper, we will discuss our following contributions:

- Reimplementing Durral et al's spectral image classfier and examining its failure cases.
- Extending the idea of classifying with spectral data into the audio and video domains.
- Combining spectral audio and video classifiers, and performing a 4-class prediction task on deepfake videos between all permutations of real/fake video and audio.

2 Related Work

2.1 Image Classification

Deepfake image classification has been extensively studied, and state of the art methods generally utilize very deep convolutional networks for feature extraction and classification. Some researchers have developed more sophisticated architectures to explicitly target certain image characteristics. For example, the authors of MesoNet [2] designed their architecture to focus on mesoscopic (middle-level) properties of input images, after noticing failure cases in image forensics caused by compression which degrades image quality.

In contrast, Wang et al [3] take a more brute force approach to deepfake classification. Rather than designing a new architecture, the authors attempted to create a "universal" deepfake detector by transfer training an ImageNet-pretrained vanilla ResNet-50 model on a collection of augmented GAN generated images from 11 different well known GANs, such as CycleGAN [4] and ProGAN [5].

2.2 Audio Classification

Similar to work in the image classification domain, current deepfake audio classification techniques rely on deep neural networks. There are two popular methods, both based on existing audio classification techniques.

The first technique uses the Mel-frequency cepstral coefficients (MFCC) as input features to a recurrent neural network. MFCCs are found from a cepstral representation of an audio clip, and are very commonly used in tasks such as speech recognition or music classification. Ali et al's fake-audio classifier [6] extracts 26 MFCCs from an input audio clip, before passing the features into an LSTM model. Similarly, Phan et al [7] use MFCCs combined with a Gated Recurrent Unit for their architecture.

An alternative to directly calculating the MFCC of an audio file is to utilize 1D convolutional layers as feature extractors. Dai et al [8] demonstrate that very deep convolutional networks which take in time-domain audio waveforms can also achieve high accuracies in audio classification. Similar to image convolutional networks, Dai et al's architecture consists of many stacked Conv-1D, BatchNorm, and MaxPooling layers, with a final Softmax layer at the end.

2.3 Multimodal Video Classification

Zhou et al [9] propose a sync-stream that models the synchronization pattern of visual and auditory modalities to enhance the performance of methods that classify the video and audio separately. Their proposed method learns the intrinsic synchronization between the audio and video (seeing if lip movements match the words being said) to better predict if the media contains both real/fake video and audio, real video and fake audio or fake video and real audio.

3 Approach and Implementation

In this section, we will discuss how we tested Durral et al's spectral image classifier, how we extended the spectral-based approach into audio and video classification, and the structure of our classifiers for the 4-class deepfake classification task.

As a performance benchmark for our novel spectral classifiers, we create a deep learning baseline for each task based on a state of the art architecture as described in Section 2.

Please also note the following notation: our spectral methods for image, audio, and video classification will be referred to as I1, A1, and V1. If clarification is needed, there will also be an acronym in parentheses that indicates the specific classifier used. For example, A1(LR) would be a logistic regression classifier that uses audio spectral features. In contrast, our deep learning baseline models will be indicated as I2, A2, and V2.

3.1 Image Classification

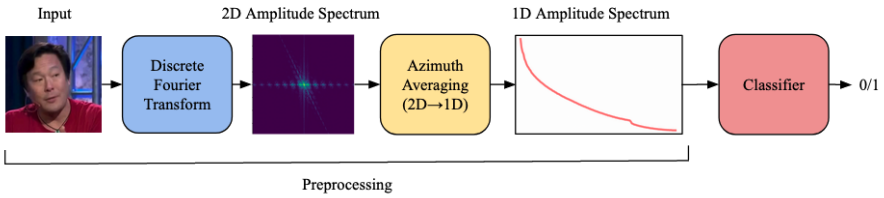


Figure 1: Spectral Image Classifier

Following the pipeline as described by Durral et al, we recreated their spectral-based classifier. There are three steps for the classifier pipeline. First, the 2D amplitude spectrum of an input image is

computed with a discrete Fourier transform. Then, the 2D amplitude spectrum is transformed into a 1D vector via Azimuth averaging, which essentially gathers similar frequencies radially. Finally, the 1D vector is passed in as the feature into a simple classifier.

We tested a variety of classifiers, including logistic regression, a degree-3 polynomial SVM, and K-Nearest Neighbors. Like Durral et al, we trained and tested on the Faces-HQ dataset [1].

To test the classifier's robustness, we examined potential failure cases via data augmentation. We compare the performance of Durral et al's method against a vanilla MesoNet architecture under various input image alterations. The results of these experiments will be reported in Section 4.

3.2 Audio Classification

3.2.1 Spectral Audio Classifier

Despite how well Durral et al's method performs in the image domain, it does not translate directly to the audio domain. Audio signals have two key differences: they are 1D signals, and they have a temporal aspect to them so simply taking an FFT would not work. Thus, we wanted to find a feature that is similar to that used by Durral, but could work on audio.

We discovered a rough equivalent named the spectral centroid. The spectral centroid provides frequency data of the audio signal while maintaining the temporal nature by calculating a weighted mean of the frequencies present in a signal. The calculation is as follows:

$$\text{Spectral Centroid} = \frac{\sum_{n=1}^{N} nF(n)}{\sum_{n=1}^{N} F(n)}$$

Here, $F(n)$ represents the frequency amplitude of bin n. In our implementation, a frame size of 2048 was used and we computed spectral centroid for the first 100 frames of each audio clip.

Following preprocessing, the audio features are passed into a simple classifier for prediction. As with image classification, we test the performance of three classifiers: Logistic Regression, SVM, and KNN.

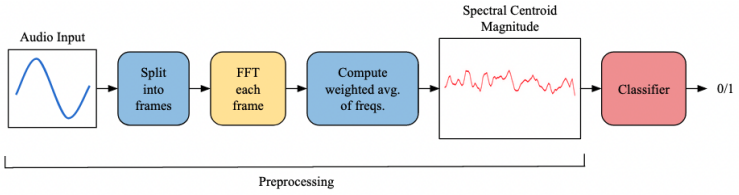


Figure 2: Spectral Audio Classifier

3.2.2 Baseline Audio Classifier

As a benchmark for our spectral audio classifier, we elected to use a model based on the MFCC+RNN architecture as discussed in Section 2.2. Our benchmark extracts an MFCC vector of dimension 400, before passing it through two LSTMs with hidden dimensions of 256 and 128 respectively. Following that, the tensor is passed through three fully connected layers with ReLU nonlinearities with dimensions (128, 64), (64, 48), and (48, 2).

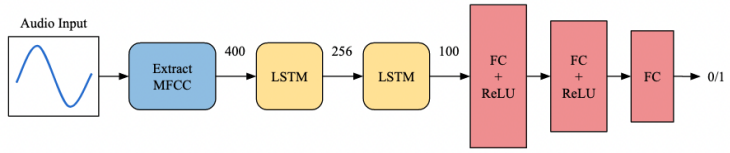


Figure 3: Baseline Audio Classifier

3.3 Video Classification

3.3.1 Spectral Video Classifier

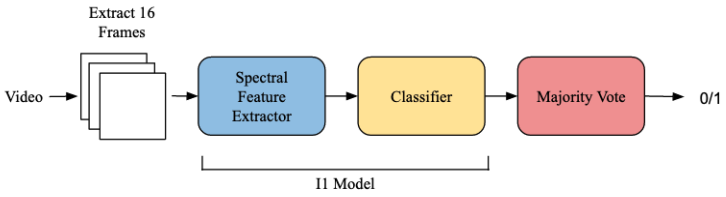


Figure 4: Spectral Video Classifier

Similar to audio signals, videos have a temporal aspect that cannot be captured by a direct fourier transform. To resolve this, we extract 16 uniformly spaced frames from each video, then pass each frame separately into our I1 model. At the output of the I1 model, we have 16 different predictions. To predict whether the video is real or fake, we simply take the majority vote from the 16 frame predictions. Note this approach has an obvious limitation: if a given input video is only partially deepfaked, this method will likely fail. However, in our testing we only used entirely real or entirely fake videos.

3.3.2 Baseline Video Classifier

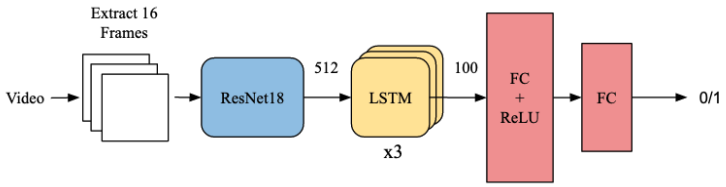


Figure 5: Baseline Video Classifier

Our baseline video classifier follows an LSTM architecture which are commonly found in many other video classification applications. Just like our spectral classifier, we extract 16 uniformly spaced frames from an input video as the input tensor. Each frame is featurized with a ResNet18 model that has been pre-trained on ImageNet, and then passed into three LSTMs. Finally, the LSTM output is passed into two fully connected layers with dimensions (256, 50) and (50, 2).

3.4 Multiclass Classification

Unlike previous work in deepfake video classification, we decided to test our spectral classifiers on a more granular test. Instead of a binary prediction task between fake and real, we perform a 4-class prediction task between real video-real audio, real video-fake audio, fake video-real audio, and fake video-fake audio.

We created and tested three models: a unimodal spectral-based model, a unimodal deep learning based model, and a multimodal deep learning model.

3.4.1 Unimodal Classifier

Both unimodal models have similar structure. We split an input video into two modalities: image frames and raw audio waveforms. Then, each modality is passed separately into their respective unimodal classifier. The output from both classifiers are concatenated and determines the final class of the video.

For our spectral approach, we use V1 with a logistic regression classifier and A1 with an SVM. For our deep learning baseline, we use V2 and A2, both LSTM based models.

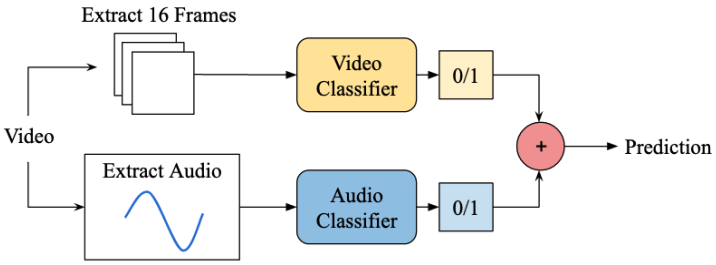


Figure 6: Unimodal Classifier

3.4.2 Multimodal Classifier

Our multimodal deep learning model differs from the baseline unimodal deep learning model in that the concatenation occurs before a prediction is made. We remove V2 and A2's final fully connected layers and concatenate the features, then pass in the combined feature into a new fully connected layer. The goal of the multimodal model was to attempt to exploit concurrency in the video data. We hypothesized that if our model could learn some correspondence between the shape of a person's mouth and the associated audio snippet, it may result in better performance than the unimodal approach.

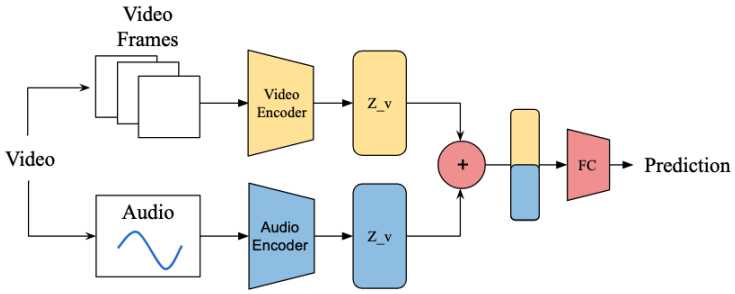


Figure 7: Multimodal Classifier

4 Experimental Setup and Results

This section details the binary classification results for each unimodal model as well as the multiclass classification results for the combined models. Each binary classifier was evaluated using precision, recall, and accuracy, while the multiclass models were evaluated using accuracy alone.

4.1 Data

We used two datasets. The first dataset is Faces-HQ [1]. This dataset contains 40,000 1024×1024 images from a variety of sources, including www.thispersondoesnotexist.com, CelebA-HQ, and more. Half of the faces in this dataset are real, and the other half are generated by GANs.

The second dataset we used is the FakeAVCeleb dataset, created by Khalid et al [10]. It consists of roughly 10,000 224×224 videos that are split evenly into one of four classes: real video and real audio, real video and fake audio, fake video and real audio, and fake video and fake audio.

Due to compute constraints, we only used a subset of both datasets. We randomly sampled 8000 images from Faces-HQ, and 1600 videos from FakeAVCeleb. For both datasets, we used an 80/20 split for training and testing.

4.2 Spectral Classifier Robustness Testing

We performed various data augmentation to our dataset to test the robustness of Durall et al's spectral method. These data augmentations include adding salt and pepper noise, Gaussian noise, and random

occlusions. In addition, we examined how downsampling input images would affect the spectral method's performance. For comparison, we used MesoNet as a benchmark.

4.2.1 Accuracies on Augmented Data

Table 1: Augmented Data Accuracy Results

Augmentation	Method	Precision	Recall	Accuracy
Original Image	MesoNet	0.994	0.999	0.993
	Spectral w/ SVM	0.995	1.0	0.994
Salt and Pepper Noise (1%)	MesoNet	0.9996	0.6994	0.8496
	Spectral w/ SVM	0.7471	0.9750	0.8225
Gaussian Noise $\mu = 10$, $\sigma = 100$	MesoNet	0.8673	0.8496	0.9164
	Spectral w/ SVM	0.6428	0.72	0.8575
Occlusion (500px x 500px)	MesoNet	0.9963	0.6382	0.8179
	Spectral w/ SVM	0.99502487	1.0	0.9975

The introduction of noise or occlusion lowers accuracy for both methods, but the spectral-based approach does seem to show resilience towards occlusion.

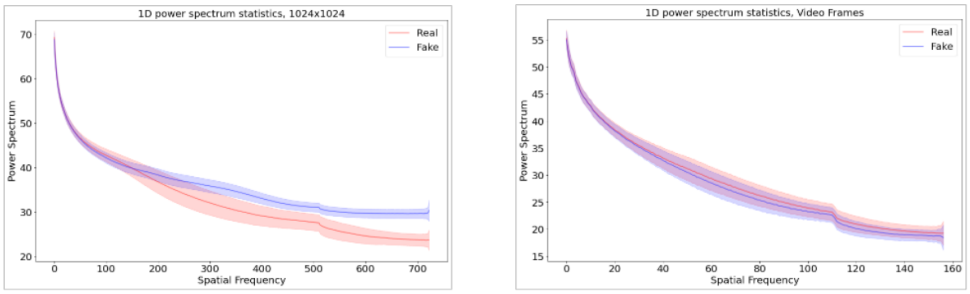
4.2.2 Effects of Downsampling

Since FacesHQ and FakeAVCeleb contain images of drastically different dimensions, we performed downsampling on images in the FacesHQ dataset to see whether the spectral classifier would adapt to images of smaller sizes. We hypothesized a loss in performance, as downsampling operations typically remove high frequency data. The testing results are as follows:

Table 2: Downsampling Results

Image Dimension	Precision	Recall	Accuracy
1024px x 1024px	1.0	1.0	1.0
512px x 512px	0.885	0.885	0.885
256px x 256px	0.7085	0.705	0.7075
128px x 128px	0.6432	0.64	0.6425

The figure below shows the image spectral features for the Faces-HQ dataset on the left and FakeAVCeleb dataset on the right.



(a) Faces-HQ (1024px) (b) FakeAVCeleb (224px)

Figure 8: Image Size Impact on Separability

Note the greatly reduced separability for the FakeAVCeleb dataset due to smaller image size. Although this decreases accuracy, the videos provide many frames which assists with accuracy when using the majority vote technique described in Section 3.3.1.

4.3 Audio Classification

We compiled our own audio dataset by extracting .wav files from the FakeAVCeleb dataset. In total, we used 1600 audio files, evenly split between real and fake audio.

For our spectral classifiers, we pre-featurized each audio file to find its spectral centroid, then passed the audio features into Scikit-learn's logistic regression, SVM, and KNN models. We used default hyperparameters for training.

For our MFCC+LSTM model, we trained the model using the Adam optimizer with learning rate 1e-5 for 25 epochs.

The spectral centroid statistics and classification results are provided below.

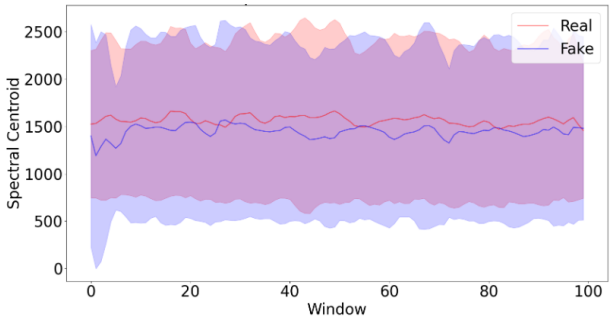


Figure 9: Spectral Centroid Statistics

In Figure 9 we see high variances caused by the variability in vocal frequency. But, separability looks roughly the same as the image spectra for the FakeAVCeleb dataset.

Table 3: Audio Classification Results

Method	Classifier	Precision	Recall	Accuracy
A1	LR	0.7514	0.6985	0.7343
	SVM (deg 3)	0.8474	**0.8090**	**0.832**
	KNN	0.8601	0.6181	0.7594
A2	N/A	0.835	0.923	**0.88**

4.4 Video Classification

For both V1 and V2, we extract frames from each video to create tensors of size (16, 224, 224, 3) before passing into our models as part of our preprocessing procedure.

Note that our spectral video classifiers are essentially the spectral image classifiers, except with a majority vote at the end. As such, V1 uses frozen weights obtained from I1.

For the CNN+LSTM model, we trained with the Adam optimizer with a learning rate of 5e-4 for 8 epochs. The training time for this was considerably faster than A2, as ResNet-18 was pretrained on ImageNet.

Table 4: Video Classification Results

Method	Classifier	Precision	Recall	Accuracy
V1	LR	0.9755	0.795	0.8875
	SVM (deg 3)	0.975	0.775	0.8775
	KNN	0.9412	**0.88**	**0.9125**
V2	N/A	1.0	0.94	**0.97**

4.5 Multiclass Classification

Our combined method takes an instance of each unimodal model (A1 and V1 for spectral, and A2 and V2 for the deep learning baseline) with frozen weights and forward passes a modal-separated video into its respective models before concatenating the predictions at the end. There is no additional training time required.

The multimodal model is also loaded in with pre-trained weights, although we fine-tune the model's new fully connected layer by training for an additional 15 epochs with the Adam optimizer and a learning rate of 5e-4.

Despite its complexity and the hypothesized benefits of mixing modalities, our multimodal model performs the worst in prediction accuracy out of all three models.

Table 5: Multiclass Classification Results

Combined Method	Accuracy
A1 (SVM) + V1 (LR)	0.72
A2 + V2	0.81
Multimodal	0.61

5 Discussion

5.1 Accuracy vs Training Time and Model Size

A2 has 882k trainable parameters and took 3 hours to train, whereas V2 took 11 minutes for transfer-learning and has 11M trainable parameters. In contrast, A1 (SVM) used a degree-3 polynomial kernel and V1 (LR) had 157 parameters. Both spectral models took less than 10 seconds to train. Although the baseline models had an accuracy of 81% over the 72% from our spectral model, it came at a cost of significantly greater training time and vastly more parameters. Our spectral model only sees a 9% decrease in accuracy but requires a fraction of the training time and memory required.

5.2 Comparison against Khalid et al's work

The creators of the FakeAVCeleb dataset, Khalid et al, also performed similar experiments on deepfake video classification [11]. Their experimental setup is slightly different: like most previous work in deepfake classification, Khalid et al perform binary classification on videos, instead of the 4-class task we demonstrated.

However, their work is extremely relevant to ours as both of us performed our experiments with the same dataset. Khalid et al used a pure deep learning approach, and utilized two model variations: ensemble voting by two unimodal classifiers, and a multimodal model. Below are their experimental results.

Table 6: Comparison Accuracy Results

Method	Accuracy
Multimodal	0.674
Ensemble	0.7804

We achieved comparable performance with our spectral-based method, but for a more difficult multiclass classification task, demonstrating the viability and effectiveness of spectral approaches in deepfake classification tasks.

6 Conclusion

We demonstrate that spectral-based classifiers can achieve reasonable accuracies not only for deepfake image classification, but for audio and video deepfake classification as well. These models are smaller in size, less complex, and have much faster training times compared to state of the art deep learning approaches. While accuracies are somewhat sacrificed, our lightweight model shines in its portability and simplicity.

References

[1] Durall, R., Keuper, M., Pfreundt, F., Keuper, J.: Unmasking DeepFakes with simple Features. arXiv:1809.00888

[2] Afchar, Darius, et al. "Mesonet: a compact facial video forgery detection network." 2018 IEEE international workshop on information forensics and security (WIFS). IEEE, 2018.

[3] Wang, Sheng-Yu et al. “CNN-generated images are surprisingly easy to spot...for now”. CVPR. N.p., 2020. Print.

[4] Jun-Yan Zhu*, Taesung Park*, Phillip Isola, and Alexei A. Efros. "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks", in IEEE International Conference on Computer Vision (ICCV), 2017.

[5] Karras, Tero, et al. "Progressive growing of gans for improved quality, stability, and variation." arXiv preprint arXiv:1710.10196 (2017).

[6] M. Ali, A. Sabir and M. Hassan, "Fake audio detection using Hierarchical Representations Learning and Spectrogram Features," 2021 International Conference on Robotics and Automation in Industry (ICRAI), 2021, pp. 1-6, doi: 10.1109/ICRAI54018.2021.9651401.

[7] Phan, Huy, et al. "Audio scene classification with deep recurrent neural networks." arXiv preprint arXiv:1703.04770 (2017).

[8] Dai, Wei, et al. "Very deep convolutional neural networks for raw waveforms." 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2017.

[9] Y. Zhou and S. -N. Lim, "Joint Audio-Visual Deepfake Detection," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14780-14789, doi: 10.1109/ICCV48922.2021.01453.

[10] Khalid, H., Tariq, S., Kim, M., Woo, S.: FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. arXiv:2108.05080

[11] Khalid, Hasam, et al. "Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors." Proceedings of the 1st workshop on synthetic multimedia-audiovisual deepfake generation and detection. 2021.

Appendix A Author Contributions

All authors contributed equally.

Appendix B Source Code

https://github.com/k2so3/eecs545