J-KBDI: Jurnal Kecerdasan Buatan dan Data Science Industri
Vol. I, No. 1, 14 Maret 2026, pp. 12-17

Rancang Bangun Sistem Deteksi Audio Deepfake Menggunakan Mel-Spectrogram dan ConvNext

Muhammad Fadhli[1*], Rudy Kurniawan[2]

[1,2] Teknik Elektro, Universitas Bangka Belitung, Bangka 33172, Indonesia
[1]muhammad.fadhli14032004@gmail.com, [2] rudy14k@gmail.com
*Penulis Korespondensi: muhammad.fadhli14032004@gmail.com

(Diterima 05 Maret 2026, Disetujui 11 Maret 2026, Tersedia Online 14 Maret 2026)

Abstract: Audio deepfake is audio or voice created by utilizing AI that resembles human speech. Along with the development of AI technology, audio deepfakes are increasingly difficult to distinguish from original audio, thus requiring a system that can be used to facilitate humans in classifying between deepfake and original audio. This research aims to create an audio classification system that utilizes signal processing techniques and deep learning technology, namely ConvNext. The signal processing technique itself is performed to transform raw audio into a spectrogram; the formation of the spectrogram itself goes through several stages, starting from sampling the signal by applying Short Time Fourier Transform (STFT) to projecting it into the form of a Mel-Spectrogram. In the training process, a model that has been pre-trained using ImageNet1k is retrained (fine-tuned) using audio data that has been transformed into Mel-Spectrograms. The dataset used to conduct the training and evaluation of the model is taken from ASVSpoof 2021. The research results show that the system trained using Mel-Spectrograms with a total of 128 Mel and 15 epochs became the model with the highest accuracy level, reaching 88.4%

Keywords: AsvSpoof 2021, Audio Deepfakes, Classification, ConvNext, Mel-Spectrogram, STFT.

Abstrak: Audio deepfake merupakan audio atau suara yang dibuat dengan memanfaatkan AI yang mirip dengan suara manusia. Seiring dengan berkembangnya teknologi AI, audio deepfake semakin sulit untuk di bedakan dengan audio asli sehingga membutuhkan suatu sistem yang dapat digunakan untuk memudahkan manusia dalam melakukan klasifikasi antara audio deepfake dan asli. Penelitian ini bertujuan untuk membuat suatu sistem klasifikasi audio yang memanfaatkan teknik pemrosesan sinyal (signal processing) dan teknologi deep learning yaitu ConvNext. Teknik pemrosesan sinyal sendiri dilakukan untuk mengubah audio mentah menjadi spectrogram, pembentukan spectrogram sendiri melalui beberapa tahapan mulai dari melakukan pengambilan sampling (mencuplik) sinyal dengan menerapkan Short Time Fourier Transform (STFT) hingga memproyeksikannya kedalam bentuk Mel-Spectrogram. Dalam proses pelatihan, model yang sudah dilakukan pre-training menggunakan ImageNet1k dilakukan pelatihan kembali (Fine-tune) menggunakan data audio yang telah diubah bentuknya menjadi Mel-Spectrogram. Dataset yang digunakan untuk melakukan pelatihan dan evaluasi model diambil dari ASVSpoof 2021. Hasil penelitian menunjukkan bahwa sistem yang dilatih menggunakan Mel-Spectrogram dengan jumlah Mel 128 dan epoch sebanyak 15 menjadi model dengan tingkat akurasi yang paling tinggi, mencapai 88.4%

Keywords: AsvSpoof 2021, Audio Deepfake, ConvNext, Klasifikasi, Mel-Spectrogram, STFT.

1. Pendahuluan

Audio deepfake merupakan audio atau suara yang dihasilkan dengan memanfaatkan AI yang mirip dengan suara manusia, sehingga sulit untuk dibedakan [1]. Modern text-to-speech merupakan hasil dari perkembangan teknologi yang memungkinkan untuk menirukan suara seseorang dengan sangat mirip yang sering disebut dengan audio deepfake[2]. Audio deepfake awalnya digunakan sebagai media hiburan, namun setiap perkembangan teknologi akan selalu ada sisi positif dan negatif, saat ini teknologi AI (Artificial Intelligence) mampu menghasilkan suara yang sangat menyerupai suara manusia asli, sehingga berpotensi disalahgunakan untuk kegiatan kriminal seperti penyebaran berita palsu (disinformasi), hingga penipuan . Hal itu dilakukan untuk berbagai macam alasan mulai dari menjatuhkan harga diri seseorang, merampas barang orang lain hingga kasus penculikan [3]. Ancaman ini menjadi semakin nyata seiring dengan semakin mudahnya akses untuk membuat suara tiruan (audio deepfake) dengan menggunakan sampel audio berdurasi singkat serta kemudahan dalam menyebarluaskannya. Tanpa adanya sistem yang andal, aksi kriminal yang dapat merugikan orang lain akan semakin marak terjadi di masyarakat.

Beberapa penelitian terdahulu tentang deteksi Audio Deepfake telah banyak dilakukan dan dibahas pada berbagai literatur dengan menggunakan berbagai metode seperti menggunakan Mel-Spectogram dan ResNet34 [4], Mel-Frequency Cepstral Coefficients (MFCC) dan Support Vector Machine (SVM) [5], Mel-Spectogram dan Convolutional

Neural Network (CNN) [6], Mel-Frequency Cepstral Coefficients (MFCC) dan Siamese Convolutional Neural Network [7].

Akan tetapi pada model-model tersebut terdapat beberapa kekurangan, seperti penggunaan RELU dan Batch Normalization pada ResNet34 [8] yang masing-masing memiliki kekurangan seperti rentan terhadap noise pada data yang dapat menyebabkan penurunan akurasi [9] dan pelatihan model tidak optimal ketika batch size terlalu kecil [10], model Support Vector Machine (SVM) juga memiliki kekurangan seperti pelatihan model tidak terlalu baik jika dataset yang digunakan tidak seimbang, pemanfaatan model CNN siamese juga pernah dilakukan dalam penelitian terdahulu, model ini juga memiliki kekurangan yaitu komputasi yang berat sehingga pelatihan model akan memakan waktu yang cukup lama walaupun dataset yang digunakan dalam pelatihan sangat kecil [11].

Pada penelitian deteksi audio deepfake kali ini menggunakan model ConvNext, karena model ini memiliki keunggulan yang dapat menutupi kekurangan dari model-model yang digunakan pada penelitian-penelitian sebelumnya, seperti penggunaan GELU, drop path dan Layer Normalization yang membuat model ConvNext menjadi lebih tahan terhadap noise, pelatihannya yang tidak bergantung pada batchsize serta komputasi yang lebih ringan. ConvNext merupakan model yang memiliki arsitektur konvolusi yang sederhana seperti neural network akan tetapi strukturnya menyerupai vision transformer, sehingga model ini memiliki akurasi yang tinggi dalam melakukan klasifikasi suatu objek menyerupai model vision transformer tetapi dengan tingkat komputasi yang lebih rendah.

Penelitian ini bertujuan untuk membuat sistem yang dapat melakukan klasifikasi audio deepfake dan asli dengan memanfaatkan fitur Mel-spectrogram sebagai data pelatihan. Mel-Spectrogram sendiri merupakan fitur yang dihasilkan dari teknik pemrosesan sinyal untuk mengubah sinyal suara menjadi kumpulan warna yang berupa objek 2 dimensi sebagai objek pelatihan dan evaluasi model. Proses pembentukan fitur Mel-Spectrogram dari audio mentah menjadi sangat penting karena memungkinkan model deep learning dapat melakukan pembelajaran dari pola-pola yang terbentuk dari suara asli dan suara tiruan (audio deepfake). semakin pekat warna dari Mel-Spectrogram maka tinggi frekuensi suara yang sedang diproses, dan semakin pudar warnanya maka semakin rendah frekuensi suaranya. Melalui penelitian ini, diharapkan sistem deteksi dapat memiliki akurasi yang tinggi dan dapat mengenali pola anomali yang terbentuk pada audio deepfake, sehingga dapat menjadi salah satu solusi dari pencegahan terhadap aksi penyebaran berita bohong dan disinformasi di masa depan.

2. Metode dan Eksperimen

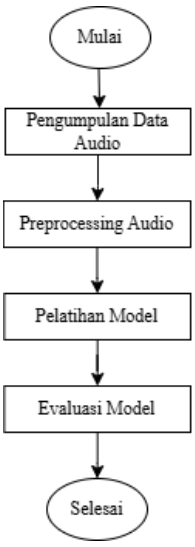

Fig. 1 Alur Penelitian

Penelitian ini menggunakan Kaggle Notebook untuk melakukan *preprocessing* audio, pelatihan dan evaluasi model. Metode yang dilakukan pada penelitian ini adalah sebagai berikut:

1. Pengumpulan Data Audio: Pengumpulan Data Audio: Pengumpulan data audio diambil secara acak dari dataset ASVSpoof 2021 sebanyak 500 data audio deepfake dan 500 data audio asli yang dimasukkan ke dalam 2 kelas yaitu DF dan PA.

2. Preprocessing Audio: Audio yang telah dikumpulkan di lakukan pra proses yang memiliki 2 tahapan yaitu augmentasi dan mengubahnya ke dalam bentuk Mel-Spectrogram. Augmentasi audio dilakukan agar dataset audio yang dimiliki menjadi lebih banyak dan variatif, augmentasi yang di terapkan pada audio pelatihan ini berupa penambahan white noise, perubahan nada (pitch shifting), perubahan waktu (time shifting), dan spec augmentation. Setelah dilakukan tahapan augmentasi audio diubah menjadi Mel-spectrogram yang disimpan dalam bentuk numpy array berukuran 128×87 berdimensi 1 sebagai input dari model. Dalam pembuatan Mel-Spectrogram format audio diubah menjadi .wav, lalu diterapkan Short Time Fourier Transform (STFT) pada sinyal audio, yang secara matematis didefinisikan sebagai berikut:

$$X(m,\omega)=\sum_{n=-\infty}^{\infty} x[n].\omega[n-m].\, e^{-j\omega n} \quad (1)$$

Hasil dari STFT tersebut digunakan untuk menghitung periodogram yang berfungsi untuk mengetahui distribusi pada tiap komponen-komponen dari pembagian sinyal. Persamaan periodogram dihitung dengan persamaan berikut:

$$P(\omega)=\frac{1}{N}|X(\omega)|2 \quad (2)$$

Dengan N adalah panjang sinyal setiap frame. Selanjutnya periodogram akan diproyeksikan kedalam Mel-Filterbank, Filterbank sendiri terdiri dari sekumpulan segitiga yang terdistribusi secara non-liniar terhadap skala frekuensi yang dirancang untuk mendekati persepsi pendengaran manusia, hubungan skala mel dan frekuensi dinyatakan sebagai berikut:

$$m=2595.\log_{10}(1+\frac{f}{700}) \quad (3)$$

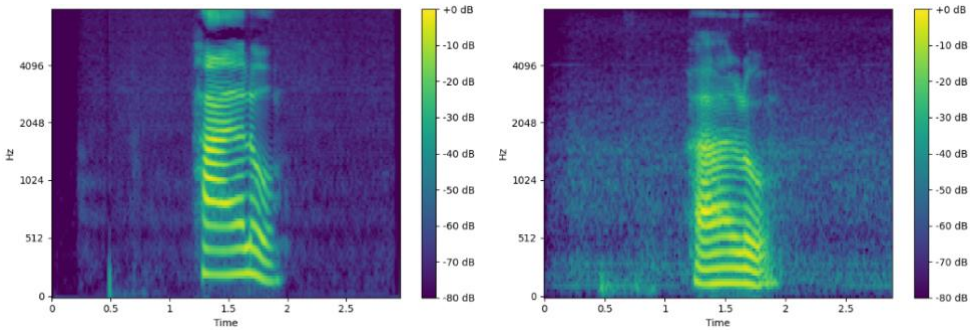

Fig. 2 Mel-Spectogram Suara Asli dan Audio Deepfake

3. Pelatihan Model: Model yang dilatih menggunakan mel-spectrogram adalah model yang telah di lakukan latih menggunakan imageNet1k, sehingga pada tahap ini adalah tahapan yang melatih ulang model agar dapat melakukan klasifikasi pada Mel-spectrogram. Tahapan ini terdapat 2 proses yaitu pembangunan model dan penentuan parameter seperti berikut.

   a. Pembangunan Model: Untuk melakukan klasifikasi Mel-Spectrogram, dilakukan beberapa penyesuaian pada arsiterktur model seperti penambahan layer untuk melakukan resize terhadap spectrogram menjadi 224×224 dan replikasi 1 channel menjadi 3 channel agar sesuai dengan model ConvNext yang telah dilatih dengan 3 channel (RGB). Pada lapisan akhir model dilakukan penyesuaian dengan menambahkan lapisan Dense, Batch Normalization, dan Dropout untuk meningkatkan kemampuannya pada tugas klasifikasi.

Table 1 Arsitektur Model ConvNext yang Telah Disesuaikan

Layer	Output shape
Input Layer	None, 128, 87, 1
Lambda	None, 224, 224, 1
Concatenate	None, 224, 224, 3
ConvNext Tiny	None, 768
Dense	None, 512
Batch Normalization	None, 512
Dropout	None, 512
Dense 1	None, 256
Dropout 1	None, 256
Dense 2	None, 1

b. Penentuan HyperParameter: Hyperparameter yang digunakan pada proses pelatihan model yaitu menggunakan AdamW sebagai Optimizer, One-cycle Learning Rate dengan Batch size 32 yang masing-masing dilatih dengan epoch 15, 30 dan 45 untuk mencari model terbaik.

4. Evaluasi Model: Setelah model dilatih, model akan diuji menggunakan sampel audio yang belum pernah diketahui oleh model berupa *Mel-spectrogram* dengan jumlah *mel* 64 dan 128 yang masing-masing terdiri dari 250 audio data untuk melihat ketahanan dari model dalam mendeteksi audio dengan jumlah *mel* yang berbeda pada saat model dilatih.

3. Hasil dan Pembahasan

Dari evaluasi model yang telah dilakukan, didapatkan hasil sebagai berikut.

Table 2 Hasil Pengujian

Kelas	Epoch	Mels Pelatihan	Mels Pengujian	Presisi	Recall	F1 Score	Akurasi
DF	15	64	64	84.33%	90.4%	87.26%	86.8%
PA				89.66%	83.2%	86.31%	
DF			128	79.51%	77.6%	78.54%	78.8%
PA				78.12%	80%	79.05%	
DF	30	64	64	87.3%	88%	87.65%	87.6%
PA				87.9%	87.2%	87.55%	
DF			128	83.04%	74.4%	78.48%	79.6%
PA				76.81%	84.8%	80.61%	
DF	45	64	64	86.05%	88.8%	87.4%	87.2%

PA				88.43%	85.6%	86.99%	
DF			128	80.67%	76.8%	78.69%	79.2%
PA				77.86%	81.6%	79.69%	
DF			64	93.64%	82.4%	87.66%	88.4%
PA	15	128		84.29%	94.4%	89.06%	
DF			128	86.67%	72.8%	79.13%	80.8%
PA				76.55%	88.8%	82.22%	
DF			64	95%	60.8%	74.15%	78.8%
PA	30	128		71.18%	96.8%	82.03%	
DF			128	92.39%	68%	78.34%	81.2%
PA				74.68%	94.4%	83.39%	
DF			64	97.37%	29.6%	45.4%	64.4%
PA	45	128		58.49%	99.2%	73.59%	
DF			128	95.95%	56.8%	71.36%	77.2%
PA				69.32%	97.6%	81.06%	

Pada tabel 2 dapat dilihat bahwa model yang dilatih menggunakan Mel-spectrogram yang dibentuk dengan jumlah mel 128 dan epoch 15 menjadi model dengan akurasi terbesar yakni 88.4% untuk pengujian menggunakan data audio dengan jumlah mel 64 serta menunjukkan f1 skor yang cukup baik di kelas DF dan PA yang masing-masing mencapai angka 87.66% dan 89.06% dan 80.8% untuk data audio dengan jumlah mel 128, juga mencapai angka 79.13% dan 82.22% di masing-masing kelas. Hal ini menunjukkan model memiliki kemampuan yang cukup kuat dalam melakukan prediksi audio walaupun kemampuan dalam memprediksi kelas PA lebih tinggi dibandingkan dengan kelas DF.

Pada tabel 2 juga dapat dilihat bahwa model yang dilatih menggunakan Mel-spectrogram yang dibentuk dengan jumlah mel 128 dan epoch 45 menjadi model yang paling lemah dalam memprediksi antara kelas DF dan PA yakni 64.4% dengan f1 score hanya 45.4% dan 73.59% di setiap kelasnya pada data uji yang memiliki jumlah mel 64. model hanya mampu mencapai akurasi sebesar 77.2% dengan f1 score hanya 71.3% dan 81.06% di setiap kelas.

4. Conclusion

Berdasarkan hasil penelitian yang telah dilakukan, model ConvNext mampu melakukan deteksi audio deepfake dengan cukup baik dengan akurasi mencapai 88.4% untuk audio dengan jumlah Mel 64 dan 80.8% untuk audio dengan jumlah Mel 128 pada model yang dilatih dengan epoch 15 dan jumlah Mel 128. Secara keseluruhan sistem ini dapat digunakan sebagai alat deteksi awal untuk mendeteksi audio deepfake secara cepat dan objektif tanpa bergantung pada penilaian manual oleh manusia.

Acknowledgment

Penulis berterima kasih kepada pihak penyedia dataset AsvSpoof 2021 yang digunakan dalam penelitian ini. Penulis juga berterima kasih kepada jurusan Teknik Elektro, Universitas Bangka Belitung atas fasilitas yang diberikan.

References

[1] J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, and Y. Zhao, "Audio Deepfake Detection: A Survey," arXiv (Cornell University), Aug. 2023, doi: https://doi.org/10.48550/arxiv.2308.14970.

[2] N. Müller, P. Czempin, F. Diekmann, A. Froghyar, and K. Böttinger, "Does Audio Deepfake Detection Generalize?," Interspeech 2022, Sep. 2022, doi: https://doi.org/10.21437/interspeech.2022-108.

[3] A. Alali and G. Theodorakopoulos, "Partial Fake Speech Attacks in the Real World Using Deepfake Audio," Journal of Cybersecurity and Privacy, vol. 5, no. 1, pp. 6–6, Feb. 2025, doi: https://doi.org/10.3390/jcp5010006.

[4] Rahul, T. P, P. R. Aravind, Ranjith C, Usamath Nechiyil, and Nandakumar Paramparambath, "Audio Spoofing Verification using Deep Convolutional Neural Networks by Transfer Learning.," arXiv (Cornell University), Aug. 2020.

[5] A. Hamza et al., "Deepfake Audio Detection via MFCC Features Using Machine Learning," IEEE Access, vol. 10, pp. 134018–134028, 2022, doi: https://doi.org/10.1109/access.2022.3231480.

[6] Fathima G, Kiruthika S, Malar M, Nivethini T, "Deepfake Audio Detection Model Based On Mel Spectrogram Using Convolutional Neural Network", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.12, Issue 4, pp.p208-p216, April 2024, Available at :http://www.ijcrt.org/papers/IJCRT24A4745.pdf

[7] O. A. Shaaban and R. Yildirim, "Audio Deepfake Detection Using Deep Learning," Engineering Reports, vol. 7, no. 3, Mar. 2025, doi: https://doi.org/10.1002/eng2.70087.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," arXiv, Dec. 2015, doi: https://doi.org/10.48550/arxiv.1512.03385.

[9] D. Hendrycks and K. Gimpel, "Gaussian Error Linear Units (GELUs)," Jun. 2016, doi: https://doi.org/10.48550/arxiv.1606.08415.

[10] S. Ioffe, "Batch Renormalization: Towards Reducing Minibatch Dependence in Batch-Normalized Models," arXiv (Cornell University), Jan. 2017, doi: https://doi.org/10.48550/arxiv.1702.03275.

[11] F.Bajić, and J.Josip, "Chart Classification Using Siamese CNN" Journal of Imaging 7, no. 11: 220, 2021, https://doi.org/10.3390/jimaging7110220.