# ANALISIS KINERJA MODEL VGGISH DENGAN INTEGRASI CONVOLUTIONAL BLCOK ATTENTION MODULE PADA DETEKSI AI-GENERATED VOICE

**Dominic Naufal Ramadhan Herwanto, Eddy Muntina Dharma, Ni Made Satvika Iswari**
Teknik Informatika, Universitas Primakara, Jalan Tukad Badung No 135, Denpasar, Bali
*dominic.naufal11@gmail.com*

**ABSTRAK**

Perkembangan teknologi *generative* AI telah memungkinkan pembuatan audio sintetis yang sangat mirip dengan suara manusia. Fenomena ini menimbulkan tantangan baru, terutama dalam membedakan suara asli dan suara yang dihasilkan oleh AI, yang berpotensi meningkatkan risiko misinformasi, penipuan digital, serta manipulasi opini publik. Penelitian ini bertujuan untuk menganalisis kinerja model VGGish dalam mendeteksi suara sintetis serta mengevaluasi peningkatan performa melalui integrasi *Convolutional Block Attention Module* (CBAM). Metode yang digunakan mencakup pemrosesan data audio menggunakan representasi *Mel-Spectrogram*, pelatihan model VGGish dan VGGish+CBAM, serta evaluasi performa melalui metrik akurasi, *confusion matrix*, dan ROC-AUC. Hasil penelitian menunjukkan bahwa integrasi CBAM meningkatkan kemampuan model dalam menyoroti fitur akustik penting secara lebih adaptif, sehingga menghasilkan performa klasifikasi yang lebih baik dibandingkan model VGGish standar. Dengan demikian, pendekatan ini berpotensi menjadi solusi efektif dalam deteksi *AI-Generated Voice* pada berbagai implementasi keamanan digital.

***Kata kunci :*** *Deepfake Voice, VGGish, CBAM, Mel-Spectrogram, Deteksi Suara AI.*

## 1. PENDAHULUAN

Peningkatan penggunaan media sosial dalam beberapa tahun terakhir telah memberikan dampak besar bagi masyarakat, baik secara positif maupun negatif. Media sosial memudahkan akses informasi serta mempercepat komunikasi antarindividu, namun pada saat yang sama juga membuka peluang terjadinya kejahatan siber seperti penipuan online, phishing, dan serangan *ransomware* yang dapat menyebabkan pencurian data serta kerugian finansial bagi pengguna [1] [2].

Kemajuan teknologi *Artificial Intelligence* (AI) semakin memperluas kompleksitas ancaman tersebut, terutama melalui munculnya *Deepfake Audio*, yaitu teknologi yang mampu menghasilkan suara sintetis dengan karakteristik yang sangat mirip suara manusia melalui pemodelan jaringan saraf tiruan [3]. Penggunaan *Deepfake Audio* pada platform digital berpotensi memicu penyebaran misinformasi, manipulasi opini publik, dan penipuan berbasis suara yang sulit diidentifikasi secara manual [4]. Penyebaran hoaks melalui media sosial juga semakin meningkat seiring rendahnya literasi digital dan algoritma distribusi konten yang mempercepat perputaran informasi tidak valid [5].

Masalah utama dari fenomena ini adalah semakin sulitnya membedakan suara asli dan suara yang dihasilkan oleh AI, sehingga dibutuhkan metode deteksi yang lebih akurat dan adaptif. Berbagai penelitian menunjukkan bahwa pendekatan *Deep Learning*, khususnya model berbasis CNN, telah digunakan dalam mendeteksi suara sintetis. Lim et al. (2022) menunjukkan performa tinggi dalam deteksi suara sintetis dengan akurasi mencapai 99,97% pada dataset ASVspoof 2021 [6]. Namun, beberapa arsitektur CNN seperti VGG16 dinilai kurang optimal dalam mengekstraksi fitur audio secara mendalam [7].

Untuk mengatasi keterbatasan tersebut, Google Research memperkenalkan VGGish, yaitu modifikasi arsitektur VGG yang diadaptasi khusus untuk analisis audio [8]. Penelitian terbaru menemukan bahwa integrasi *Attention Mechanism* seperti *Convolutional Block Attention Module* (CBAM) dapat meningkatkan performa model secara signifikan dengan cara menyoroti fitur akustik penting pada sinyal suara, sehingga menghasilkan akurasi deteksi mencapai 99,78% [9].

Berdasarkan kondisi tersebut, penelitian ini bertujuan untuk menganalisis dan mengevaluasi performa model VGGish, baik secara mandiri maupun setelah diintegrasikan dengan CBAM, dalam mendeteksi *AI-Generated Voice*. Rencana penyelesaian masalah dilakukan melalui proses ekstraksi fitur suara menggunakan *Mel-Spectrogram*, pelatihan model VGGish dan VGGish+CBAM, serta evaluasi akurasi menggunakan metrik seperti ROC-AUC dan *confusion matrix*. Penelitian ini diharapkan dapat memberikan kontribusi terhadap pengembangan teknologi deteksi suara sintetis yang lebih akurat dan efektif dalam mendukung keamanan digital.

## 2. TINJAUAN PUSTAKA

### 2.1. Penelitian Terdahulu

Penelitian mengenai deteksi suara sintetis (*AI-Generated Voice*) telah mengalami perkembangan signifikan dalam beberapa tahun terakhir. Ali et al. [10] mengusulkan model hybrid CNN–LSTM untuk mendeteksi *deepfake audio* melalui analisis fitur spektral dan menunjukkan akurasi tinggi dalam klasifikasi suara asli dan sintetis. Martínez-Serrano et al. [2] menekankan bahwa peningkatan realisme suara sintetis menuntut metode deteksi yang lebih kuat, terutama melalui pendekatan deep learning yang mampu mengenali ketidakselarasan akustik halus

dalam sinyal audio.Penelitian lain oleh McUba et al. [7]. juga menunjukkan bahwa metode *deep learning* berbasis CNN memiliki kemampuan yang baik dalam mendeteksi *deepfake audio*, meskipun performanya sangat bergantung pada kualitas ekstraksi fitur yang digunakan.

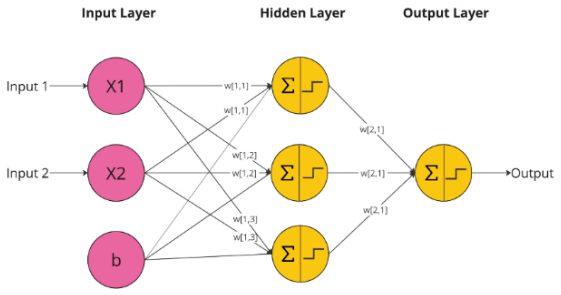
Lim et al. [6] menunjukkan bahwa pendekatan *deep learning* yang dikombinasikan dengan teknik *explainable AI* mampu mencapai akurasi tinggi dalam mendeteksi suara sintetis pada dataset ASVspoof. Secara keseluruhan, penelitian terdahulu menunjukkan bahwa penggunaan arsitektur CNN, VGGish, dan *attention mechanism* memiliki potensi besar dalam meningkatkan performa deteksi suara sintetis. Oleh karena itu, penelitian ini berfokus pada evaluasi model VGGish dan integrasinya dengan CBAM untuk meningkatkan akurasi dalam mendeteksi *AI-Generated Voice*.

### 2.2. Kecerdasan Buatan dan Generative Audio

Kecerdasan buatan (*Artificial Intelligence*) berkembang pesat dan memunculkan teknologi generative models yang mampu menghasilkan konten audio dengan kualitas yang menyerupai suara manusia. Penelitian yang dilakukan oleh Ali et al. menunjukkan bahwa kombinasi arsitektur CNN dan LSTM mampu mendeteksi audio *deepfake* dengan akurasi tinggi melalui pemanfaatan fitur spektral pada dataset suara asli dan sintetis, di mana model *deep learning* dapat mengidentifikasi pola akustik yang sulit dideteksi oleh metode konvensional [11].

### 2.3. *Neural Network*

*Neural network* adalah model komputasi yang meniru cara kerja otak manusia dalam memproses informasi. Model ini terdiri dari beberapa lapisan yaitu input layer, hidden layer, dan *output layer*. Setiap neuron menerima input yang dikalikan dengan bobot (*weight*) dan ditambahkan bias, kemudian diproses menggunakan fungsi aktivasi untuk menghasilkan *output* [12].

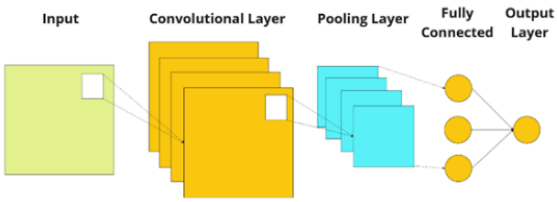



Gambar 1. Gambar arsitektur neural network

### 2.4. *Convolutional Neural Network* (CNN)

*Convolutional Neural Network* (CNN) merupakan pengembangan dari *neural network* yang banyak digunakan dalam pengolahan citra dan audio. CNN memiliki tiga komponen utama yaitu *convolutional layer*, *pooling layer*, dan *fully connected layer* [13]

a. *Convolutional layer* berfungsi untuk mengekstraksi fitur dari data input menggunakan kernel
b. *Pooling layer* digunakan untuk mengurangi dimensi data
c. *Fully connected* layer digunakan untuk proses klasifikasi




Gambar 2. Gambar arsitektur CNN

### 2.5. *Activation Function*

Fungsi aktivasi digunakan untuk menentukan apakah suatu neuron akan diaktifkan atau tidak. Fungsi ini memungkinkan model mempelajari hubungan non-linear. Beberapa fungsi aktivasi yang umum digunakan adalah ReLU dan Sigmoid [14]

ReLU digunakan pada proses pelatihan karena mampu mempercepat konvergensi.
Rumus ReLU:

$$f(x)_{Relu} = max(0, x) \quad (1)$$

Sedangkan Sigmoid digunakan pada output karena menghasilkan nilai antara 0 dan 1.
Rumus Sigmoid:

$$f(x)_{sigm} = \frac{1}{1 + e^{-z}} \quad (2)$$

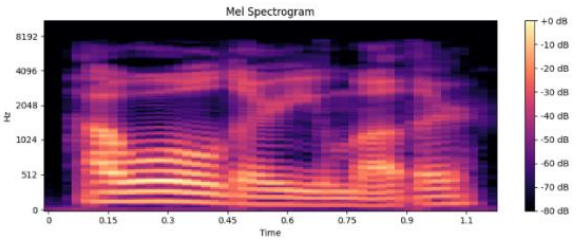
### 2.6. *Loss Function*

*Loss function* digunakan untuk mengukur tingkat kesalahan model terhadap data sebenarnya. Pada penelitian ini digunakan *Binary Cross Entropy* karena sesuai untuk klasifikasi biner.
*Rumus Binary Cross Entropy*:

$$L(y) = -\frac{1}{N}\sum_{i=1}^{n}(y_i) \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i)) \quad (3)$$

### 2.7. Mel-Spectrogram

Mel-Spectrogram merupakan representasi visual dari sinyal audio dalam domain frekuensi yang disesuaikan dengan persepsi pendengaran manusia. Proses pembentukannya melibatkan beberapa tahapan seperti windowing, *Short-Time Fourier Transform* (STFT), dan konversi ke skala mel [7].



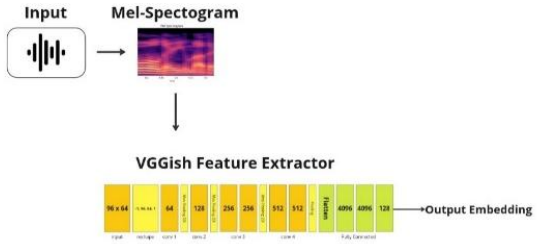

Gambar 3. Gambar mel-spectrogram

Mel-Spectrogram terbukti mampu memberikan representasi fitur audio yang lebih akurat dibandingkan metode lainnya.

## 2.8. VGGish

VGGish merupakan model CNN yang dikembangkan oleh Google untuk ekstraksi fitur audio. Model ini menggunakan input berupa Mel-Spectrogram dan menghasilkan embedding berdimensi 128 yang merepresentasikan karakteristik audio [8]

Model ini banyak digunakan dalam klasifikasi audio karena kemampuannya dalam menangkap pola spektral dan temporal.

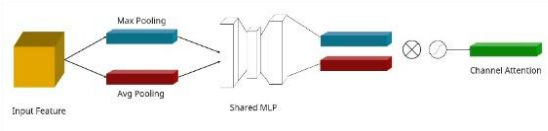



Gambar 4. Gambar arsitektur VGGish

## 2.9. *Convolutional Block Attention Module* (CBAM)

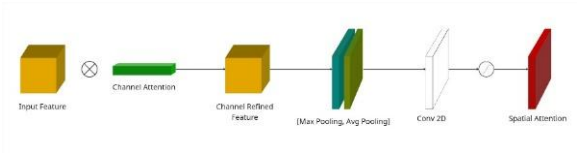
CBAM adalah modul perhatian (*attention module*) yang digunakan untuk meningkatkan performa *Convolutional Neural Network* (CNN) dengan memfokuskan pada fitur yang paling relevan dari data. Mekanisme *attention* memungkinkan model untuk menyesuaikan bobot fitur secara adaptif sehingga meningkatkan representasi fitur yang dihasilkan [14]. Modul ini terdiri dari dua komponen utama yaitu *channel attention* dan spatial attention, yang bekerja secara berurutan untuk memperbaiki kualitas feature map.

a. *Channel attention* berfungsi untuk menentukan fitur apa yang paling penting.




Gambar 5. Gambar channel attention

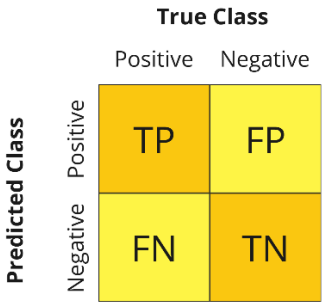
b. Sedangkan *spatial attention* digunakan untuk menentukan lokasi fitur penting pada data input.




Gambar 6. Gambar spatial attention

## 2.10. Confusion Matrix

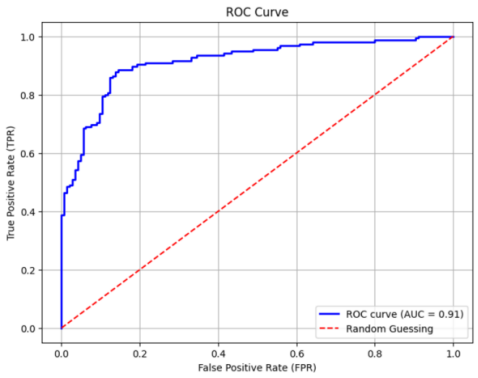
Confusion matrix digunakan untuk mengevaluasi performa model klasifikasi dengan membandingkan hasil prediksi dan nilai sebenarnya. Metode ini memberikan gambaran detail mengenai kinerja model melalui komponen seperti *true positive*, *false positive*, *false negative*, dan *true negative*, serta dapat digunakan untuk menghitung berbagai metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan F1-*score* [15]




Gambar 8. Gambar confusion matrix

## 2.11. ROC-AUC

ROC-AUC merupakan metode evaluasi model klasifikasi yang digunakan untuk mengukur kemampuan model dalam membedakan kelas. Kurva ROC menggambarkan hubungan antara *True Positive Rate* (TPR) dan *False Positive Rate* (FPR) pada berbagai nilai ambang (*threshold*), sedangkan AUC menunjukkan luas area di bawah kurva yang merepresentasikan kinerja keseluruhan model [16].

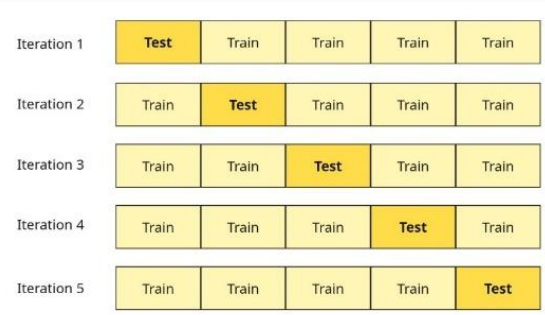



Gambar 9. Gambar confusion matrix

## 2.12. K-Fold Cross Validation

K-Fold Cross Validation adalah teknik evaluasi model yang digunakan untuk mengestimasi kemampuan generalisasi model dengan membagi dataset menjadi beberapa bagian (*fold*). Pada setiap iterasi, model dilatih menggunakan sebagian data dan

diuji pada bagian lainnya secara bergantian, sehingga menghasilkan evaluasi yang lebih stabil serta mengurangi risiko *overfitting* [17].

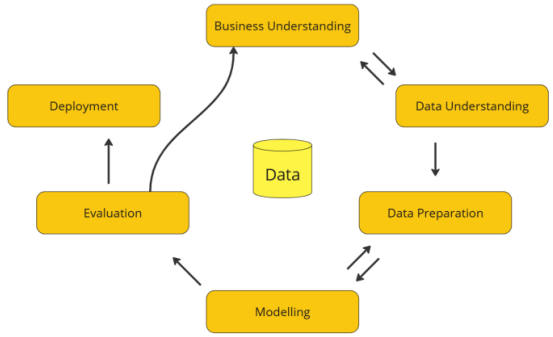



Gambar 10. Gambar K-Fold Cross Validation

## 3. METODE PENELITIAN

### 3.1. Metode Penelitian

Metode penelitian yang digunakan dalam penelitian ini adalah *Cross Industry Standard Process for Data Mining* (CRISP-DM). Metode ini merupakan kerangka kerja yang banyak digunakan dalam pengembangan proyek berbasis data karena memiliki pendekatan yang terstruktur, sistematis, dan fleksibel. CRISP-DM terdiri dari beberapa tahapan utama yaitu *business understanding*, *data understanding*, *data preparation*, *modelling*, *evaluation*, dan *deployment*.




Gambar 10. Gambar alur penelitian

### 3.2. Alur Penelitian

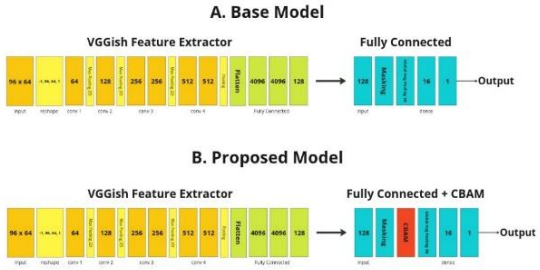
a. *Business Understanding*
Tahap ini berfokus pada pemahaman permasalahan penelitian yaitu analisis kinerja model VGGish yang diintegrasikan dengan CBAM dalam mendeteksi AI-generated voice. Selain itu, dilakukan studi literatur untuk mengidentifikasi metode, dataset, serta pendekatan yang relevan.

b. *Data Understanding*
Dataset yang digunakan dalam penelitian ini berasal dari platform Kaggle yang berisi rekaman suara untuk deteksi suara sintetis. Dataset terdiri dari beberapa variasi, yaitu:
- *for-original*: rekaman asli dari berbagai sumber
- *for-norm*: data hasil preprocessing
- *for-2seconds*: potongan audio berdurasi 2 detik
- *for-rerecorded*: rekaman ulang menggunakan perangkat berbeda.

c. *Data Preparation*
Tahap ini bertujuan untuk mempersiapkan data sebelum proses pemodelan. Data audio yang awalnya berada pada domain waktu dikonversi ke domain frekuensi agar informasi lebih mudah dipelajari oleh model. Proses ini dilakukan dengan mengubah audio menjadi representasi mel-spectrogram.
Selain itu, dataset dibagi menjadi beberapa bagian yaitu *training set*, *validation set*, dan *test set* untuk kebutuhan pelatihan dan evaluasi model.

d. *Modelling*
Tahap ini bertujuan untuk membangun model yang digunakan dalam penelitian. Terdapat dua pendekatan model yang digunakan, yaitu:
- Base Model: model VGGish tanpa integrasi CBAM
- Proposed Model: model VGGish dengan integrasi CBAM

Perbandingan kedua model ini digunakan untuk mengetahui pengaruh penggunaan *attention module* terhadap performa model dalam klasifikasi audio.




Gambar 10. Gambar base model dan proposed model

e. *Evaluation*
Tahap evaluasi dilakukan untuk mengukur kinerja model yang telah dibangun. Metrik yang digunakan meliputi:
- ROC-AUC untuk mengukur kemampuan model dalam membedakan kelas
- Confusion Matrix untuk melihat distribusi prediksi benar dan salah

Hasil evaluasi ini digunakan untuk membandingkan performa antara *base model* dan *proposed model*.

f. *Deployment*
Tahap *deployment* merupakan tahap akhir dimana model yang telah dilatih dan dievaluasi diimplementasikan agar dapat digunakan dalam lingkungan nyata. Model dapat diintegrasikan ke dalam aplikasi atau sistem tertentu melalui API atau platform lainnya sehingga dapat dimanfaatkan secara langsung oleh pengguna.

### 3.3. Hipotesis Penelitian

Penelitian ini mengajukan hipotesis bahwa penggunaan mel-spectrogram sebagai representasi audio mampu meningkatkan kemampuan model dalam mendeteksi perbedaan antara suara asli manusia dan

suara yang dihasilkan oleh AI. Hal ini dikarenakan mel-spectrogram mampu merepresentasikan karakteristik frekuensi suara secara lebih detail.

Selain itu, integrasi model VGGish dengan *Convolutional Block Attention Module* (CBAM) diperkirakan dapat meningkatkan performa model. Mekanisme *attention* memungkinkan model untuk lebih fokus pada fitur-fitur penting dalam data audio, sehingga diharapkan mampu meningkatkan akurasi dan efisiensi dalam proses klasifikasi.

## 4. HASIL DAN PEMBAHASAN

### 4.1. Hasil Penelitian

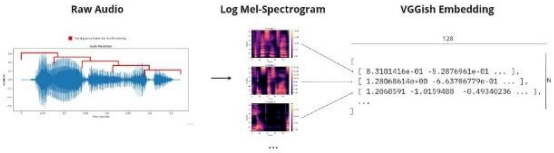
Penelitian ini berfokus pada analisis performa model dalam mendeteksi AI-*generated voice* dengan membandingkan dua pendekatan, yaitu *base model* (VGGish tanpa CBAM) dan *proposed* model (VGGish dengan integrasi CBAM). Permasalahan yang diangkat berasal dari meningkatnya penggunaan teknologi *deepfake voice* yang berpotensi menimbulkan penyalahgunaan seperti penyebaran informasi palsu dan manipulasi suara.

Berdasarkan studi literatur, metode *Convolutional Neural Network* (CNN), khususnya arsitektur VGGish, banyak digunakan dalam klasifikasi audio. Namun, performa model masih dapat ditingkatkan dengan menambahkan mekanisme perhatian seperti *Convolutional Block Attention Module* (CBAM). Oleh karena itu, penelitian ini mengusulkan integrasi CBAM untuk meningkatkan kemampuan model dalam memahami fitur penting pada data audio.

Tabel 1. Spesifikasi Dataset

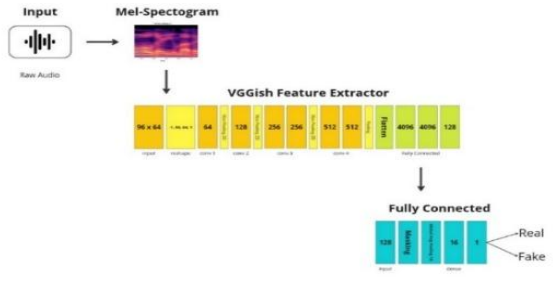
| No | FoR Dataset | Versi | Jumlah Data | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|
| | | | **Real** | | | **Fake** | | | |
| | | | **Train** | **Val** | **Test** | **Train** | **Val** | **Test** | |
| 1 | for-original | Balanced version | 26,941 | 5,400 | 2,264 | 26,941 | 5,400 | 2,370 | 69,316 |
| 2 | for-norm | Normalized version | 26,941 | 5,400 | 2,264 | 26,927 | 5,398 | 2,370 | 69,300 |
| 3 | for-2second | Truncated at two seconds | 6,978 | 1,413 | 544 | 6,978 | 1,413 | 544 | 17,870 |
| 4 | for-rerec | Re-recorded version | 5,104 | 1,101 | 408 | 5,104 | 1,143 | 408 | 13,268 |

Dataset yang digunakan adalah *Fake-or-Real* (FoR) Dataset yang diperoleh dari Kaggle, yang terdiri dari data audio *real* dan *fake*. Data kemudian diproses menjadi representasi mel-spectrogram dan diekstraksi menggunakan VGGish menjadi embedding berdimensi 128 sebagai input model.




Gambar 11. Gambar ilustrasi ekstraksi audio

### 4.2. Pengujian Model




Gambar 12. Gambar alur pelatihan base model

Proses pelatihan dilakukan dengan berbagai skenario, yaitu penggunaan jumlah *epoch* yang berbeda serta penerapan K-*Fold Cross Validation* dengan 3 dan 5 fold. Hasil pelatihan menunjukkan bahwa kedua model memiliki performa yang tinggi pada data training dan validation, dengan nilai akurasi berada pada rentang 0.99.

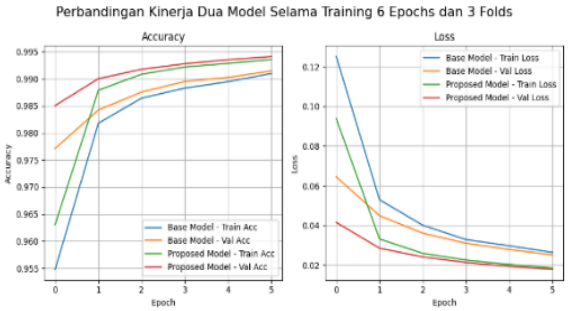
Model: "functional"

| Layer (type) | Output Shape | Param # | Connected to |
|---|---|---|---|
| input_layer (InputLayer) | (None, None, 128) | 0 | - |
| not_equal (NotEqual) | (None, None, 128) | 0 | input_layer[0][0] |
| masking (Masking) | (None, None, 128) | 0 | input_layer[0][0] |
| any (Any) | (None, None) | 0 | not_equal[0][0] |
| global_average_poo… (GlobalAveragePool… | (None, 128) | 0 | masking[0][0], any[0][0] |
| dense (Dense) | (None, 16) | 2,064 | global_average_p… |
| dense_1 (Dense) | (None, 1) | 17 | dense[0][0] |

Total params: 2,081 (8.13 KB)
Trainable params: 2,081 (8.13 KB)
Non-trainable params: 0 (0.00 B)

Gambar 13. Gambar arsitektur base model

Hasil pengujian menggunakan data uji menunjukkan bahwa *proposed* model secara konsisten memberikan performa yang lebih baik dibandingkan *base model*. Berdasarkan hasil evaluasi, peningkatan performa terlihat pada *metrik accuracy*, *precision*, *recall*, F1-score, dan ROC-AUC.




Gambar 14. Gambar model dengan 6 epochs dan 3 folds

Sebagai contoh, pada skenario pelatihan 10 epoch:

a. Base model menghasilkan AUC sebesar 0.87
b. Proposed model meningkat menjadi 0.94

Selain itu, proposed model juga menunjukkan peningkatan pada F1-score yang mencerminkan keseimbangan antara *precision* dan *recall* dalam klasifikasi.

Tabel 2. Perbandingan Performa Model

| Experiment | | Acc | Precision | Recall | F1-Score | AUC |
|---|---|---|---|---|---|---|
| Train 10 Epochs | Base | 0.80 | 0.73 | 0.93 | 0.82 | 0.87 |
| | Proposed | 0.85 | 0.79 | 0.92 | 0.85 | 0.94 |
| Train 20 Epochs | Base | 0.78 | 0.71 | 0.96 | 0.81 | 0.88 |
| | Proposed | 0.81 | 0.75 | 0.93 | 0.83 | 0.92 |
| Train 6 Epochs + 3 Fold | Base | 0.76 | 0.71 | 0.86 | 0.78 | 0.83 |
| | Proposed | 0.82 | 0.80 | 0.85 | 0.82 | 0.90 |
| Train 12 Epochs + 3 Fold | Base | 0.81 | 0.76 | 0.89 | 0.82 | 0.88 |
| | Proposed | 0.83 | 0.81 | 0.86 | 0.83 | 0.91 |
| Train 10 Epochs + 5 Fold | Base | 0.73 | 0.65 | 0.94 | 0.77 | 0.81 |
| | Proposed | 0.75 | 0.68 | 0.90 | 0.78 | 0.85 |
| Train 20 Epochs + 5 Fold | Base | 0.77 | 0.70 | 0.92 | 0.80 | 0.83 |
| | Proposed | 0.80 | 0.75 | 0.90 | 0.81 | 0.88 |

### 4.3. Pembahasan

Hasil penelitian menunjukkan bahwa penggunaan mel-spectrogram sebagai representasi data audio mampu membantu model dalam menangkap karakteristik frekuensi yang membedakan suara manusia dan suara hasil AI. Hal ini terbukti dari tingginya performa model pada berbagai skenario pelatihan.

Integrasi CBAM pada model VGGish memberikan peningkatan performa yang signifikan. Mekanisme *channel attention* dan *spatial attention* memungkinkan model untuk lebih fokus pada fitur-fitur penting dalam data, sehingga meningkatkan kemampuan klasifikasi. Hal ini terlihat dari peningkatan nilai AUC dan F1-*score* pada hampir seluruh skenario pengujian.

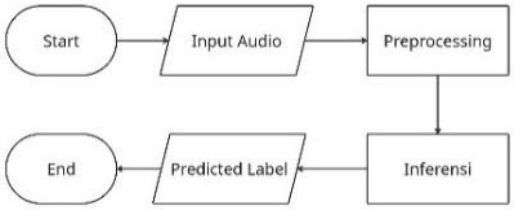
Namun demikian, peningkatan performa tersebut diikuti dengan peningkatan waktu inferensi. Pada beberapa skenario, *proposed* model membutuhkan waktu yang lebih lama dibandingkan base model, serta memiliki deviasi waktu yang lebih besar. Hal ini menunjukkan adanya *trade-off* antara akurasi dan efisiensi komputasi.

Dari seluruh skenario yang diuji, konfigurasi pelatihan 20 epoch tanpa K-Fold pada proposed model memberikan hasil terbaik dengan performa yang seimbang antara akurasi dan efisiensi. Model ini mampu mencapai nilai AUC sebesar 0.91 dengan waktu inferensi yang relatif stabil, sehingga menjadi pilihan optimal untuk implementasi.

### 4.4. *Deployment*

Tahap *deployment* dilakukan dalam bentuk prototipe untuk mensimulasikan penggunaan model dalam skenario nyata. Proses ini mencakup input berupa audio yang kemudian diproses melalui tahapan preprocessing, ekstraksi fitur, hingga menghasilkan output berupa label klasifikasi (*real* atau *fake*).

Meskipun belum diimplementasikan dalam sistem berbasis aplikasi atau API, prototipe ini menunjukkan bahwa model dapat digunakan secara langsung untuk melakukan klasifikasi audio. Hal ini memberikan gambaran bahwa model memiliki potensi untuk dikembangkan lebih lanjut ke dalam sistem deteksi *deepfake voice* secara *real-time*.




Gambar 15. Gambar flowchart deployment


Gambar 16. Gambar output prediction

## 5. KESIMPULAN DAN SARAN

Penelitian ini bertujuan untuk menganalisis pengaruh integrasi *Convolutional Block Attention Module* (CBAM) pada model VGGish dalam klasifikasi AI-*generated voice*. Berdasarkan hasil pengujian, kedua model menunjukkan performa yang baik dengan nilai AUC yang tinggi, namun *proposed model* secara konsisten memberikan peningkatan performa dibandingkan *base model*, terutama pada metrik *accuracy*, F1-*score*, dan ROC-AUC yang lebih seimbang. Integrasi CBAM terbukti mampu meningkatkan kemampuan model dalam memfokuskan fitur penting pada data audio sehingga menghasilkan klasifikasi yang lebih stabil dan akurat. Meskipun demikian, peningkatan performa tersebut diikuti dengan sedikit peningkatan waktu inferensi, sehingga terdapat *trade-off* antara akurasi dan efisiensi

komputasi. Secara umum, *base model* lebih unggul dari sisi kecepatan dan cocok untuk sistem dengan keterbatasan sumber daya, sedangkan *proposed model* lebih optimal untuk kebutuhan yang memerlukan akurasi tinggi seperti sistem keamanan atau forensik digital. Untuk penelitian selanjutnya, disarankan untuk melakukan pengujian pada data *real-world* yang lebih kompleks, mengeksplorasi arsitektur model lain seperti LSTM atau GRU untuk menangkap informasi temporal, serta mengembangkan metode attention yang lebih lanjut seperti *self-attention* atau *transformer* guna meningkatkan performa dan generalisasi model.

## DAFTAR PUSTAKA


[1] Y. K. Dwivedi, E. Ismagilova, N. P. Rana, and R. Raman, "Social Media Adoption, Usage And Impact In Business-To-Business (B2B) context: A state-ff-the-art literature review", *Information Systems Frontiers*, vol. 24, no. 5, pp. 1501–1523, 2023, doi: 10.1007/s10796-021-10106-y/Published.

[2] Anisa Sahara and Kuswandi Kuswandi, "Penipuan Online sebagai Bentuk Kejahatan Siber dalam Perspektif Kriminologi," *Parlementer : Jurnal Studi Hukum dan Administrasi Publik*, vol. 2, no. 4, pp. 91–105, Dec. 2025, doi: 10.62383/parlementer.v2i4.1425.

[3] Z. Almutairi and H. Elgibreen, "A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions," Applied Sciences, vol. 12, no. 5, pp. 1–25, 2022,. doi: 10.3390/a15050155.

[4] A. Mutmainnah, A. M. Suhandi, and Y. T. Herlambang, "Problematika Teknologi Deepfake sebagai Masa Depan Hoax yang Semakin Meningkat: Solusi Strategis Ditinjau dari Literasi Digital," *UPGRADE : Jurnal Pendidikan Teknologi Informasi*, vol. 1, no. 2, pp. 67–72, Feb. 2024, doi: 10.30812/upgrade.v1i2.3702.

[5] M. R. Juwita, N. R. Anggraini, M. S. Ulum, M. D. Prananta, M. F. Hidayat, and N. A. Purnama, "Hoaxes in the Digital Era: An Analysis of Social Media Users," *Jurnal Lemhannas RI*, vol. 12, no. 4, Feb. 2025, doi: 10.55960/jlri.v12i4.944.

[6] S. Y. Lim, D. K. Chae, and S. C. Lee, "Detecting Deepfake Voice Using Explainable Deep Learning Techniques," *Applied Sciences (Switzerland)*, vol. 12, no. 8, Apr. 2022, doi: 10.3390/app12083926.

[7] M. McUba, A. Singh, R. A. Ikuesan, and H. Venter, "The effect of deep learning methods on deepfake audio detection for digital investigation," in *Procedia Computer Science*, Elsevier B.V., 2023, pp. 211–219. doi: 10.1016/j.procs.2023.01.283.

[8] E. I. A. El-Latif, N. E. El-Sayad, K. K. Mohammed, A. Darwish, and A. E. Hassanien, "VGGish transfer learning model for the efficient detection of payload weight of drones using Mel-spectrogram analysis," *Neural Comput. Appl.*, vol. 36, no. 21, pp. 12883–12899, Jul. 2024, doi: 10.1007/s00521-024-09661-7.

[9] T. Kanwal, R. Mahum, A. M. AlSalman, M. Sharaf, and H. Hassan, "Fake speech detection using VGGish with attention block," *EURASIP J. Audio Speech Music Process.*, vol. 2024, no. 1, Dec. 2024, doi: 10.1186/s13636-024-00348-4.

[10] A. Heidari, N. Jafari Navimipour, H. Dag, and M. Unal, "Deepfake detection using deep learning methods: A systematic and comprehensive review," Mar. 01, 2024, *John Wiley and Sons Inc*. doi: 10.1002/widm.1520.

[11] H. Muhammad Sharafat Ali, S. Muhammad Muslim Rizvi, H. Tariq, S. Majeed, A. Tariq, and M. Munwar Iqbal, "AI-Based Deepfake Audio Detection Technique from Real and Fake Audio Dataset", International Journal of Engineering Research, vol. 8, no. 2, pp. 55–64, 2025, doi: 10.56979/802/2025.

[12] E. Yaghoubi, E. Yaghoubi, A. Khamees, and A. H. Vakili, "A systematic review and meta-analysis of artificial neural network, machine learning, deep learning, and ensemble learning approaches in field of geotechnical engineering," Jul. 01, 2024, *Springer Science and Business Media Deutschland GmbH*. doi: 10.1007/s00521-024-09893-7.

[13] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-021-00444-8.

[14] Y. Wu, H. Huang, Z. Li, and S. Zhang, "CBAM-ResNet: A Lightweight ResNet Network Focusing on Time Domain Features for End-to-End Deepfake Speech Detection," *Electronics (Switzerland)*, vol. 14, no. 12, Jun. 2025, doi: 10.3390/electronics14122456.

[15] S. Riyanto, I. S. Sitanggang, T. Djatna, and T. D. Atikah, "Comparative Analysis using Various Performance Metrics in Imbalanced Data for Multi-class Text Classification." Electronics, vol. 14, no. 12, pp. 1–15, Jun. 2025, doi: 10.3390/electronics14122456.

[16] S. Roumeliotis *et al.*, "ROC curve analysis: a useful statistic multi-tool in the research of nephrology," International Urology and Nephrology, vol. 56, pp. 1623–1639, Aug. 2024, doi: 10.1007/s11255-024-04022-8.

[17] T. M. Dutschmann, L. Kinzel, A. ter Laak, and K. Baumann, "Large-scale evaluation of k-fold cross-validation ensembles for uncertainty estimation," *Journal of Cheminformatics, vol. 15, no. 1, pp. 1–15, Dec. 2023, doi: 10.1186/s13321-023-00709-9*