

# BAB I

## PENDAHULUAN

### 1.1 Latar Belakang

*Deepfake* merupakan teknik manipulasi visual dan audio yang bertujuan membuat konten rekayasa agar terlihat dan terdengar sama seperti aslinya [1]. Dalam domain audio, salah satu pendekatan *deepfake* adalah *voice cloning*, yang digunakan untuk meniru suara seseorang mulai dari intonasi sampai ritme bicara, kemudian menyusun kembali menjadi suara baru yang belum pernah diucapkan oleh pemilik suara [2]. Perkembangan *audio deepfake* dapat menghasilkan suara tiruan yang mirip sampai sulit dibedakan dari suara aslinya [3].

Penyalahgunaan *deepfake* kini menjadi salah satu ancaman dalam kejahatan siber yang dapat menimbulkan kerugian ekonomi di Indonesia. Berdasarkan data Otoritas Jasa Keuangan (OJK), kasus penipuan finansial meningkat hingga 1.550% sejak tahun 2022 [4]. Kerugian ini terus bertambah, didukung oleh data *Indonesia Anti-Scam Centre* (IASC) yang mencatat 343.402 laporan penipuan dengan total kerugian mencapai Rp7,8 triliun sepanjang November 2024 hingga November 2025 [4]. *Audio deepfake* berbahasa Indonesia menyebabkan terjadinya disinformasi konten menyesatkan di media social yang meningkat sebesar 43% selama tahun 2020 hingga 2023 [5]. Lonjakan ini terjadi seiring meluasnya penyalahgunaan *Artificial Intelligence* (AI), termasuk *audio deepfake* untuk memengaruhi masyarakat dan menyebarkan informasi palsu [5]. Ancaman dari *audio deepfake* semakin berbahaya karena kemampuan manusia dalam membedakan suara sintetis makin sulit, apalagi suara yang dihasilkan memiliki kualitas tinggi [6]. Meningkatnya penyalahgunaan *audio deepfake* menunjukkan adanya kebutuhan mendesak akan pendekatan model klasifikasi yang efektif.

Untuk mengatasi permasalahan tersebut, berbagai arsitektur *deep learning* telah banyak diusulkan untuk mengklasifikasi *audio deepfake* [6]. *Convolutional Neural Network* (CNN) menjadi salah satu pendekatan yang umum digunakan karena mampu mengekstraksi fitur spektral melalui operasi konvolusi pada representasi audio seperti spektrogram [7]. Tantangannya, proses sintesis pada *audio deepfake* sering meninggalkan jejak manipulasi yang tersembunyi pada karakteristik spektral dan temporal [8], [9]. *Pooling operation* pada CNN membuat informasi posisi antarfitur cenderung hilang, sehingga konfigurasi jejak yang tersebar sulit dipertahankan [10]. *Capsule Network* (CapsNet) menutupi keterbatasan

tersebut dengan merepresentasikan fitur sebagai vektor dan mempertahankan hubungan spasial hierarkis antarfitur melalui mekanisme *dynamic routing* [11]. Di sisi lain, *Long Short-Term Memory* (LSTM) efektif dalam memodelkan pola temporal pada sinyal suara [12].

Beberapa penelitian telah mengusulkan pendekatan yang berbeda dalam mengatasi tantangan ini. Muruganandham et al. (2025) memanfaatkan LSTM untuk memodelkan pola temporal pada sinyal suara [12]. Wani et al. (2024) menerapkan ABC-CapsNet yang memanfaatkan mekanisme *dynamic routing* untuk mempertahankan hubungan spasial dan hierarki antarfitur pada representasi spektral audio [13]. Abisha et al. (2024) merancang arsitektur hibrida CapsNet-LSTM agar kedua jenis karakteristik tersebut dimanfaatkan bersamaan, namun rancangan itu hanya pada tahap konseptual tanpa hasil evaluasi empiris [14]. Almutairi dan Elgibreen (2022) mencatat bahwa mayoritas model fokus pada dataset berbahasa Inggris dan mengidentifikasi hal tersebut sebagai celah penelitian yang memerlukan pengujian pada bahasa lain [6]. Kajian yang secara khusus menguji model untuk klasifikasi *audio deepfake* bahasa Indonesia masih sangat terbatas [6], [15], [16].

Penelitian ini mengimplementasikan arsitektur hibrida CapsNet-LSTM untuk klasifikasi *audio deepfake* Bahasa Indonesia, yang diuji menggunakan *Speech Dataset of Human and AI-Generated Voices* berisi rekaman suara manusia dan suara sintetis berbasis AI dalam Bahasa Indonesia. Kajian ini diangkat dalam skripsi berjudul "**Implementasi Arsitektur CapsNet dan LSTM untuk Klasifikasi Audio Deepfake Bahasa Indonesia**".

## 1.2 Rumusan Masalah

Berdasarkan latar belakang tersebut, ancaman audio deepfake yang dapat disalahgunakan untuk penipuan dan penyebaran informasi palsu menunjukkan urgensi akan pendekatan klasifikasi *audio deepfake* berbahasa Indonesia.

## 1.3 Tujuan

Tujuan dalam penelitian ini adalah mengimplementasikan arsitektur CapsNet dan LSTM dalam mengklasifikasikan *audio deepfake* berbahasa Indonesia.

## 1.4 Manfaat

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Memberikan kontribusi teoretis pada kajian klasifikasi *deepfake* audio berbahasa Indonesia, khususnya melalui eksplorasi arsitektur *hybrid* CapsNet-LSTM yang masih sangat terbatas.
2. Memberikan kerangka dasar bagi pengembangan sistem klasifikasi *audio deepfake*, yang relevan dengan ancaman keamanan siber pada platform komunikasi digital di Indonesia.

## 1.5 Ruang Lingkup

Agar penelitian ini lebih terfokus, maka ditetapkan batasan-batasan masalah sebagai berikut:

1. *Dataset* yang digunakan bersumber dari *Speech Dataset of Human and AI-Generated Voices*. Link: <https://www.kaggle.com/datasets/kambingbersayaphitam/speech-dataset-of-human-and-ai-generated-voices> yang terdiri dari 42 file audio dalam Bahasa Indonesia, dikategorikan menjadi dua kelas klasifikasi:
  - i. Kelas asli (label 0) terdiri dari 21 file rekaman suara manusia asli dalam format WAV.
  - ii. Kelas palsu (label 1) terdiri dari 21 file rekaman suara sintetis yang dihasilkan menggunakan teknologi AI berbasis *voice cloning*.Seluruh file pada kedua kelas berasal dari satu penutur. File kelas palsu dihasilkan melalui *voice cloning* atas sampel suara penutur tersebut, sehingga kelas asli dan kelas palsu memiliki identitas penutur yang sama. Penelitian ini dibatasi pada klasifikasi audio asli dan audio hasil kloning pada satu identitas suara [17].
2. Seluruh file audio disegmentasi menjadi potongan berdurasi 4 detik tanpa *overlap*. Durasi tersebut dipilih karena telah digunakan pada penelitian sebelumnya dan dinilai cukup untuk membedakan suara sintetis [18], [19], sekaligus menghasilkan ukuran matriks input yang seragam pada tahap pelatihan. Untuk menguji pengaruh keheningan pada bagian awal rekaman terhadap performa klasifikasi, penelitian ini menyiapkan dua varian dataset dari sumber audio dan pembagian file yang sama:
  - i. Dataset A: file disegmentasi tanpa memotong keheningan pada bagian awal audio.
  - ii. Dataset B: keheningan pada bagian awal tiap audio dipotong sebelum segmentasi.Kedua varian diuji menggunakan kombinasi hyperparameter yang sama agar pengaruh pembuangan keheningan dapat dibandingkan secara adil.

3. Ekstraksi fitur audio dilakukan menggunakan metode *Mel-Frequency Cepstral Coefficients* (MFCC) untuk menghasilkan representasi matriks 2D sebagai *input* model.
4. *Output* model berfokus pada klasifikasi audio menjadi kelas "Asli" dan "Palsu". Penelitian ini tidak mencakup pengukuran tingkat kemiripan suara (*speech similarity*), verifikasi identitas penutur, pengujian generalisasi model terhadap penutur di luar dataset yang digunakan, serta pembuatan antarmuka pengguna (GUI/Web/Mobile).

