BAB 1
PENDAHULUAN

1.1 Latar Belakang

Deepfake adalah konten multimedia yang tampak meyakinkan yang dimodifikasi secara digital atau dihasilkan secara sintetis melalui keterlibatan model pembelajaran mendalam (*deep learning*) [1]. *Audio deepfake*, khususnya *speech audio deepfake*, melibatkan manipulasi atau pembuatan audio sintetis yang baik mengubah data asli maupun menghasilkan konten audio yang sepenuhnya baru dengan meniru suara individu yang menjadi data pelatihan model tersebut [2].

Terdapat beberapa jenis *audio deepfake*, yaitu sintesis ucapan, konversi suara, manipulasi audio, penerjemahan bahasa, dan *half deepfake*.

Sintesis ucapan merujuk pada pembuatan ucapan yang menyerupai manusia dari teks tertulis, yang umum dikenal sebagai *text-to-speech* (TTS) [3].

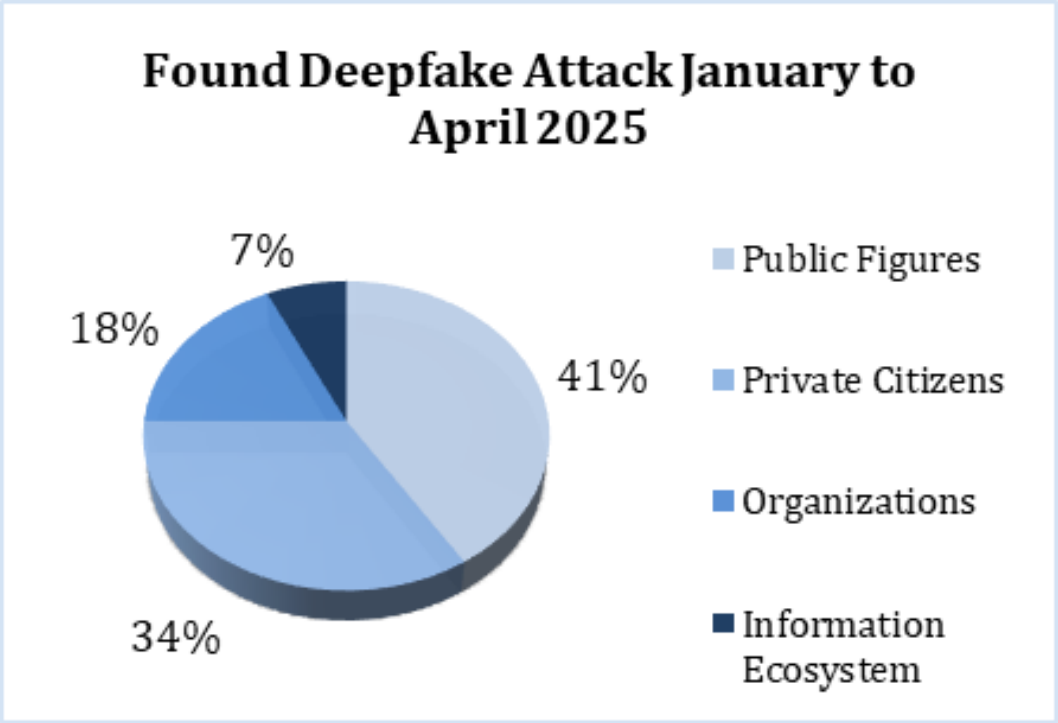
Konversi suara melibatkan modifikasi karakteristik vokal seorang pembicara agar terdengar seperti orang lain, sambil mempertahankan isi ucapan aslinya—hal ini juga dikenal sebagai *audio deepfake* berbasis peniruan atau imitasi. Teknik ini sering digunakan dalam AI cover atau cover lagu yang meniru gaya penyanyi tertentu dengan bantuan alat berbasis AI [4].

Manipulasi audio mencakup pengubahan audio yang sudah ada dengan menyesuaikan nada, pitch, tempo, atau ekspresi emosi; menyusun ulang atau menambah serta menghapus kata atau kalimat; mengubah aksen atau gender; maupun memodifikasi suara latar.

Deepfake penerjemahan bahasa mengubah ucapan dari satu bahasa ke bahasa lain dengan tetap mempertahankan karakteristik vokal pembicara asli.

Terakhir, *half deepfake* (atau *deepfake* parsial) hanya memanipulasi atau mensintesis sebagian dari audio, sementara bagian rekaman lainnya dibiarkan tidak berubah [5].

Selama beberapa dekade terakhir, *Audio Language Model* (ALM), khususnya dalam TTS, telah mengalami peningkatan yang sangat pesat. Integrasinya ke dalam aplikasi yang ramah pengguna memungkinkan masyarakat umum untuk menghasilkan *deepfake* yang meyakinkan dari individu lain hanya dengan beberapa detik rekaman audio [6].



Gambar 1.1. Target Konten Deepfake

Meningkatnya popularitas deepfake secara umum terlihat dari lonjakan penyebarannya di media sosial. Pada awal tahun 2023, Google Trends menunjukkan peningkatan minat terhadap kata kunci pencarian "*deepfake*" [7]. Pada tahun yang sama, terdapat sekitar 500.000 video dan audio deepfake yang dibagikan di media sosial di seluruh dunia [8]. Jumlah ini diperkirakan akan terus tumbuh secara eksponensial dan dapat mencapai 8 juta pada akhir tahun 2025 [9].

Namun, karena AI kini mampu menciptakan konten yang tampak sepenuhnya autentik dan hampir tidak dapat dibedakan dari konten buatan manusia, isu-isu seperti misinformasi, manipulasi, dan menurunnya kepercayaan terhadap media digital telah menjadi perhatian utama [10].

Gambar 1.1 menunjukkan temuan dalam laporan oleh [11] yang menganalisis 163 kasus *deepfake* yang terdokumentasi. Target utama serangan *deepfake* adalah figur publik, yang mencakup 41% dari total serangan. Meskipun demikian, pertumbuhan yang paling mengkhawatirkan terjadi pada serangan terhadap warga sipil, khususnya dalam lingkungan pendidikan, di mana *deepfake* digunakan sebagai alat perundungan siber dalam bentuk pelecehan dan penghinaan.

Konten yang dihasilkan oleh AI memengaruhi audiens pada tingkat psikologis, teknologi, dan sosial. Secara psikologis, konten tersebut memicu

respons emosional—semakin mirip dengan manusia, semakin kuat reaksi yang dapat ditimbulkannya. Sebagai contoh, video deepfake yang realistis atau gambar yang tampak hidup dapat memicu rangsangan emosional yang memengaruhi persepsi dan sikap penonton. Efek ini menjadi sangat kuat ketika berkaitan dengan informasi palsu, karena konten yang sarat emosi dapat memengaruhi orang terlepas dari seberapa rasional seseorang berpikir.

Salah satu insiden yang menunjukkan bagaimana *audio deepfake* dapat digunakan untuk tujuan jahat terjadi menjelang hari-hari menuju pemilihan presiden Amerika Serikat pada November, tepatnya pada Januari 2024 [12]. Puluhan ribu pemilih Partai Demokrat menerima panggilan suara hasil AI yang menirukan Presiden Biden dan menginstruksikan mereka untuk tidak memberikan suara dalam *primary election* New Hampshire yang akan datang.

Tersangka, Paul Carpenter, menggunakan platform ElevenLabs yang menawarkan layanan berlangganan untuk memanfaatkan teknologi kloning suara. Dengan bantuan rekannya, Kramer, serta Lingo Telecom, mereka menyamar sebagai seorang kandidat guna menekan partisipasi pemilih. Kasus ini hanyalah satu dari sekian banyak contoh bagaimana *audio deepfake* dapat dijadikan senjata, mulai dari intervensi pemilu, penyebaran disinformasi, hingga penipuan finansial berskala besar.

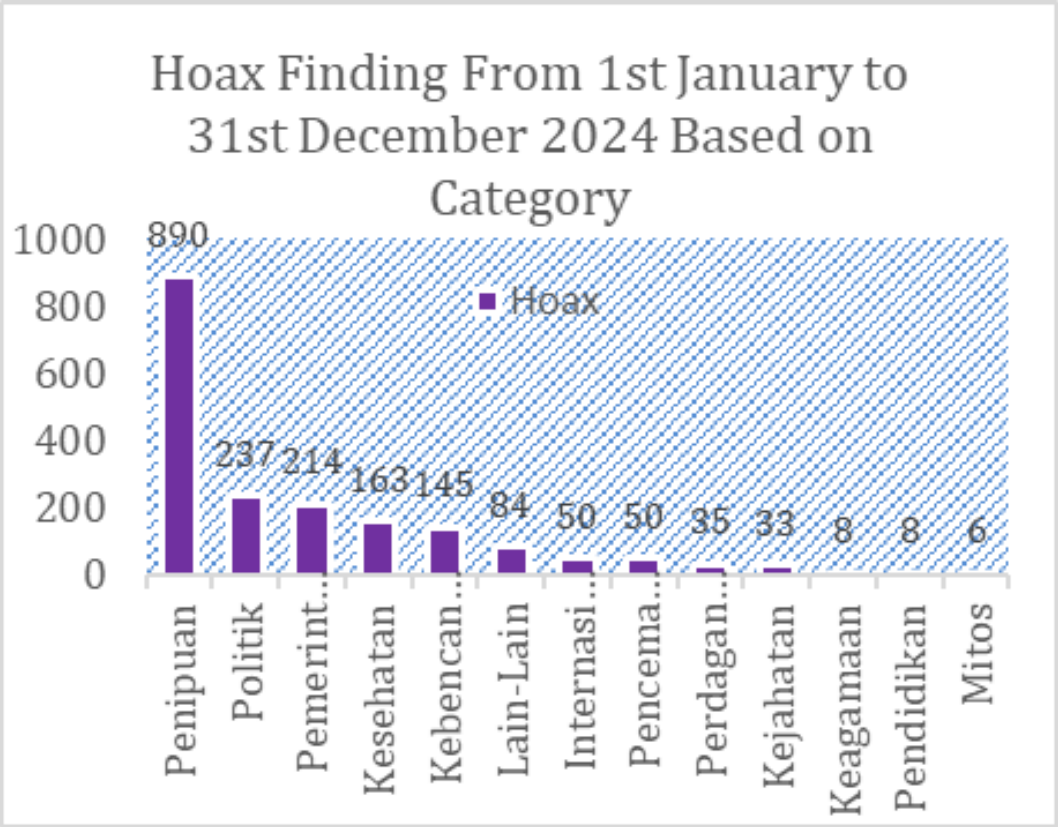
Penelitian yang dilakukan oleh [12]. mengkaji kemampuan masyarakat umum dalam mendeteksi apakah suatu ucapan dihasilkan oleh AI atau tidak. Hasil penelitian menunjukkan bahwa para partisipan hanya mampu mengidentifikasi suara buatan AI dengan benar sekitar 60% dari waktu, sementara dalam 80% kasus, partisipan cenderung menganggap identitas suara hasil AI sama dengan versi aslinya.

Temuan ini mengungkap kekhawatiran nyata bahwa suara *deepfake* akan segera menjadi tidak dapat dibedakan dari suara asli, baik dari segi kealamian maupun identitas. Jika kondisi ini terjadi dan kemudian dimanfaatkan sebagai senjata, maka proses investigasi secara manual oleh manusia akan menjadi sangat berat. Oleh karena itu, perlu dilakukan persiapan dengan mengeksplorasi alternatif lain untuk mengotomatisasi tugas tersebut, seperti dengan memanfaatkan teknik *deep learning*.

Hoaks adalah informasi palsu yang disebarkan dengan niat jahat untuk menyesatkan audiens [13]. Sementara itu, deephoax merupakan rekayasa digital yang lebih canggih, seperti skema phishing dan berita palsu yang direkayasa. Pada tahun 2018, terungkap betapa mudahnya algoritma pembelajaran mendalam

generatif digunakan untuk tujuan jahat. Sejak saat itu, deephoax semakin marak dan menjadi tantangan besar bagi keamanan siber karena memicu misinformasi dan merusak kepercayaan publik [14], [15].

Di Indonesia, berita hoaks merupakan masalah mendesak yang memerlukan perhatian segera. Data yang disediakan oleh Kementerian Komunikasi dan Digital (Komdigi) menunjukkan bahwa secara total terdapat 1.923 berita hoaks yang ditemukan dalam rentang waktu satu tahun (2024) [16].



Gambar 1.2. Temuan Hoax di 2024

Seperti ditunjukkan pada Gambar 1.2, kategori yang paling rentan adalah Penipuan (*Fraud*), yang kemungkinan besar berdampak langsung pada kehidupan sehari-hari masyarakat dan menimbulkan risiko terhadap mata pencaharian individu atau, dalam skala yang lebih besar, terhadap perekonomian nasional.

Kategori kedua yang paling rentan adalah hoaks berbasis politik. Hal ini sejalan dengan kasus terbaru terkait deephoax yang melibatkan mantan Menteri Keuangan Indonesia, Sri Mulyani, pada Agustus 2025, di mana sebuah video viral memperlihatkan Sri Mulyani seolah-olah menyatakan bahwa para guru merupakan beban bagi negara [17].

Gambar 1.3. Deephoax Mantan Menteri Keuangan Indonesia

Sumber: Rabbani (2025)

Peristiwa tersebut sempat memicu kemarahan publik hingga akhirnya terungkap bahwa video tersebut merupakan deephoax. Namun demikian, pandangan masyarakat terhadap korban telah berubah secara signifikan. Tidak lama setelah itu, terjadi pergantian di Kementerian Keuangan. Hal ini menunjukkan betapa besar dan berdampaknya pengaruh deephoax terhadap kehidupan korban. Oleh karena itu, diperlukan sistem deteksi deephoax berbasis multimodal.

Tujuan dari penelitian ini adalah untuk melakukan ablation berlapis (*layer-wise ablation*) pada *pre-trained model* sebelum membangun dan membandingkan beberapa MLP *classifier* guna menentukan arsitektur terbaik untuk implementasi publik di masa mendatang dalam deteksi deephoax. Oleh karena itu, fokus utama diberikan pada akurasi dan waktu eksekusi, karena keduanya sangat krusial untuk implementasi secara *real-time*.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana perancangan model deteksi deephoax audio?

2. Bagaimana perbandingan performa dari tiga bagian lapisan dari model wav2vec 2.0 dilihat dari *ablation study* yang dilakukan?
3. Bagaimana perbandingan performa model wav2vec 2.0 dan Multi-Layer Perceptron (MLP) dalam mendeteksi perbedaan antara suara asli dan suara *deepfake*?
4. Seberapa besar akurasi dan kecepatan model yang dibangun menggunakan dataset Fake-or-Real dalam mengklasifikasikan audio sebagai *fact* atau *hoax*?
5. Faktor-faktor apa saja yang memengaruhi hasil deteksi?

1.3 Batasan Permasalahan

Agar penelitian ini tetap fokus dan terarah, maka ditetapkan beberapa batasan penelitian sebagai berikut:

1. Dataset yang digunakan adalah Fake-or-Real Dataset yang tersedia di Kaggle, tanpa melakukan penambahan data eksternal.
2. Model ekstraksi fitur yang digunakan adalah wav2vec 2.0 versi pra-latih (*pre-trained model*) dari Hugging Face, tanpa melakukan *fine-tuning* mendalam terhadap model dasar.
3. Arsitektur *classifier* yang digunakan adalah Multilayer Perceptron (MLP) dengan beberapa *hidden layer* sederhana.
4. Klasifikasi hanya dibatasi pada dua kelas, yaitu *Fact* dan *Hoax*
5. Evaluasi dilakukan berdasarkan metrik umum seperti *accuracy*, *precision*, *recall*, dan *F1-score*.
6. Ablation study hanya dilakukan pada masing-masing satu representasi dari layer bagian bawah *(bottom)*, tengah *(middle)*, dan atas *(top)*.

1.4 Tujuan Penelitian

Penelitian ini dilakukan dengan tujuan sebagai berikut:

1. Merancang model deteksi deephoax audio.

2. Untuk menganalisis dan membandingkan performa tiga bagian lapisan pada model wav2vec 2.0 melalui *ablation study* yang dilakukan.

3. Untuk membandingkan performa model wav2vec 2.0 dan Multilayer Perceptron (MLP) dalam mendeteksi perbedaan antara suara asli dan suara *deepfake*.

4. Untuk mengevaluasi tingkat akurasi dan kecepatan model yang dibangun menggunakan dataset Fake-or-Real dalam mengklasifikasikan audio sebagai *Fact* atau *Hoax*.

5. Untuk mengidentifikasi faktor-faktor yang memengaruhi hasil deteksi dalam proses klasifikasi *audio deepfake*.

1.5 Urgensi Penelitian

Urgensi penelitian ini didasari oleh pesatnya perkembangan teknologi pembelajaran mendalam yang memungkinkan pembuatan *audio deepfake* dengan tingkat realisme yang semakin tinggi, sehingga semakin sulit dibedakan dari suara asli oleh manusia. Kondisi ini diperparah oleh meningkatnya penyebaran *deepfake* dan deephoax di media sosial yang berdampak serius pada misinformasi, manipulasi opini publik, penipuan finansial, hingga gangguan proses demokrasi, baik di tingkat global maupun nasional. Temuan bahwa manusia hanya mampu mendeteksi suara buatan AI dengan tingkat akurasi yang relatif rendah menunjukkan keterbatasan pendekatan manual dalam menghadapi ancaman ini. Di Indonesia sendiri, tingginya jumlah hoaks—terutama pada kategori penipuan dan politik—serta munculnya kasus deephoax yang melibatkan figur publik menegaskan perlunya solusi teknis yang andal, cepat, dan dapat diimplementasikan secara real-time. Oleh karena itu, penelitian mengenai perancangan dan evaluasi model deteksi deephoax audio berbasis pembelajaran mendalam menjadi sangat mendesak sebagai upaya mitigasi risiko dan perlindungan kepercayaan publik terhadap media digital.

1.6 Manfaat Penelitian

Luaran dari penelitian ini meliputi perolehan arsitektur model deteksi deephoax audio yang optimal melalui proses *layer-wise ablation* pada *pre-trained model* dan evaluasi beberapa konfigurasi MLP *classifier*. Selain itu, penelitian

ini diharapkan menghasilkan analisis komprehensif terkait pengaruh tiap lapisan fitur terhadap akurasi dan waktu eksekusi model, sehingga dapat menjadi dasar pemilihan arsitektur yang paling efisien untuk penggunaan publik. Luaran lainnya berupa model deteksi yang memiliki keseimbangan antara performa dan kecepatan, sehingga layak untuk diterapkan dalam skenario deteksi real-time, serta kontribusi ilmiah berupa referensi metodologis bagi penelitian lanjutan di bidang deteksi *audio deepfake* dan deephoax berbasis multimodal.