BAB 2
LANDASAN TEORI

2.1 Perbandingan Studi Terdahulu dan Kesenjangan Riset (*Research Gap*)

Untuk memperjelas posisi penelitian ini dibandingkan dengan penelitian terdahulu, dilakukan perbandingan terhadap beberapa penelitian yang membahas deteksi *audio deepfake*. Perbandingan tersebut meliputi metode yang digunakan, dataset, kelebihan, keterbatasan, serta kontribusi penelitian yang diusulkan. Ringkasan perbandingan disajikan pada Tabel 2.1.

Tabel 2.1. Perbandingan penelitian terdahulu dan posisi penelitian usulan

Peneliti dan Tahun	**Model / Metode**	**Dataset**	**Kelebihan / Performa**	***Keterbatasan / Research Gap***	**Kontribusi Penelitian Usulan**
Tak *et al.* (2021) [9]	RawNet2 (CNN 1D + *Sinc Filters*)	ASVspoof 2019	Memproses sinyal mentah secara langsung tanpa FFT dan menghasilkan nilai *Equal Error Rate* (EER) yang rendah.	Model dilatih dari awal (*from scratch*) sehingga membutuhkan komputasi tinggi dan rentan mengalami *overfitting*.	Menggunakan *transfer learning* pada *Wav2Vec 2.0* dengan membekukan lapisan *CNN Feature Extractor* untuk mengurangi *overfitting* dan mempercepat konvergensi.
Yi *et al.* (2023) [7]	CNN / LCNN Baseline	*Audio Deepfake Survey*	Mampu mendeteksi serangan *audio spoofing* konvensional.	Fitur spektral tradisional seperti MFCC kurang mampu menangkap manipulasi fase gelombang dan frekuensi tinggi.	Menggunakan *Wav2Vec 2.0* berbasis *Transformer* yang mengekstraksi representasi laten secara langsung dari *raw waveform*.
Boldisor *et al.* (2025) [10]	*Self-Supervised Learning* (SSL)	Audio-Visual Deepfake	Representasi SSL lebih tahan terhadap *noise* dan kompresi data.	Kemampuan generalisasi lintas dataset masih terbatas.	Menerapkan strategi *fine-tuning* dengan membekukan enam lapisan *Transformer* terbawah untuk meningkatkan kemampuan generalisasi pada dataset *Fake-or-Real*.
Hao *et al.* (2025) [11]	Wav2DF-TSL (*Two-Stage Learning*)	ASVspoof 2019/2021	Mekanisme *expert fusion* meningkatkan akurasi deteksi.	Penelitian berfokus pada evaluasi teknis menggunakan dataset *benchmark* dan belum mengaitkannya dengan mitigasi ancaman nyata.	Melakukan evaluasi mendalam terhadap metrik klasifikasi, khususnya *Recall* kelas *FAKE*, untuk mengkaji penerapan model dalam mitigasi serangan *voice phishing*.

2.2 Metode *Baseline* dan Fitur Akustik Tradisional

Pada tugas deteksi *AI-generated speech* (*audio deepfake*), pendekatan konvensional umumnya mengandalkan proses ekstraksi fitur akustik tradisional (*hand-crafted features*) sebelum fitur tersebut diproses oleh algoritma klasifikasi. Ekstraksi fitur bertujuan memperoleh representasi karakteristik sinyal suara yang mampu membedakan antara suara asli dan suara sintetis. Beberapa fitur akustik tradisional yang banyak digunakan dalam penelitian deteksi *audio deepfake* dijelaskan sebagai berikut.

2.2.1 *Mel-Frequency Cepstral Coefficients* (MFCC)

Mel-Frequency Cepstral Coefficients (MFCC) merupakan metode ekstraksi fitur yang merepresentasikan amplop spektral jangka pendek (*short-term spectral envelope*) berdasarkan karakteristik persepsi pendengaran manusia. Tahapan pembentukan fitur MFCC meliputi:

a. ***Framing* dan *Windowing***

Sinyal audio dibagi menjadi beberapa bingkai berdurasi sekitar 20–30 ms agar dapat diasumsikan bersifat stasioner. Selanjutnya setiap bingkai dikalikan dengan fungsi jendela (*window*) untuk mengurangi efek diskontinuitas pada tepi sinyal.

b. ***Fast Fourier Transform* (FFT)**

Setiap bingkai dikonversi dari domain waktu ke domain frekuensi menggunakan algoritma *Fast Fourier Transform* (FFT) sehingga diperoleh spektrum daya (*power spectrum*).

c. ***Mel Filterbank***

Spektrum daya diproyeksikan ke dalam skala Mel menggunakan sekumpulan filter segitiga (*triangular filterbank*). Skala Mel dirancang untuk meniru sensitivitas pendengaran manusia yang lebih peka terhadap frekuensi rendah dibandingkan frekuensi tinggi.

d. **Logaritma dan *Discrete Cosine Transform* (DCT)**

Nilai keluaran dari *Mel Filterbank* diubah ke dalam skala logaritmik, kemudian diterapkan *Discrete Cosine Transform* (DCT) untuk menghasilkan koefisien cepstral. Umumnya digunakan 13 koefisien MFCC yang merepresentasikan karakteristik timbre atau warna suara.

Meskipun MFCC banyak digunakan pada berbagai aplikasi pengenalan suara, metode ini memiliki keterbatasan dalam mendeteksi *audio deepfake*. MFCC menghilangkan informasi fase (*phase information*) dari sinyal audio dan melakukan proses *smoothing* terhadap spektrum frekuensi tinggi. Padahal, model pembangkit suara berbasis kecerdasan buatan modern sering menghasilkan artefak berupa riak frekuensi tinggi (*high-frequency spectral ripples*) dan ketidaksesuaian fase (*phase misalignment*) yang menjadi indikator penting keberadaan manipulasi suara.

Akibatnya, sistem yang hanya mengandalkan fitur MFCC cenderung memiliki kemampuan yang terbatas dalam mendeteksi suara sintetis berkualitas tinggi.

2.2.2 *Linear Frequency Cepstral Coefficients* (LFCC)

Linear Frequency Cepstral Coefficients (LFCC) merupakan metode ekstraksi fitur yang memiliki tahapan perhitungan serupa dengan MFCC. Perbedaan utama terletak pada penggunaan *filterbank* dengan distribusi frekuensi linier, bukan skala Mel.

Karena menggunakan *filterbank* linier, LFCC memberikan bobot yang relatif sama pada seluruh rentang frekuensi, termasuk frekuensi tinggi. Karakteristik ini membuat LFCC lebih mampu mempertahankan informasi frekuensi tinggi dibandingkan MFCC sehingga sering digunakan pada sistem deteksi *speaker spoofing* maupun *audio deepfake*.

Meskipun memiliki representasi frekuensi tinggi yang lebih baik, LFCC tetap mengabaikan informasi fase sinyal suara. Selain itu, fitur yang dihasilkan bersifat statis sehingga kurang mampu beradaptasi terhadap perubahan karakteristik akustik yang dihasilkan oleh model generatif modern berbasis *deep learning*.

2.2.3 *Constant-Q Cepstral Coefficients* (CQCC)

Constant-Q Cepstral Coefficients (CQCC) dikembangkan untuk mengatasi keterbatasan resolusi frekuensi pada MFCC. Berbeda dengan MFCC yang menggunakan *Fast Fourier Transform* (FFT), CQCC memanfaatkan *Constant-Q Transform* (CQT) sebagai dasar transformasi sinyal.

Transformasi CQT menghasilkan resolusi frekuensi yang tinggi pada frekuensi rendah dan resolusi waktu yang tinggi pada frekuensi tinggi. Karakteristik tersebut memungkinkan CQCC menangkap pola akustik yang lebih rinci sehingga efektif digunakan dalam mendeteksi serangan *voice conversion* maupun *speech synthesis* pada sistem deteksi *speaker spoofing*.

Meskipun memiliki representasi frekuensi yang lebih baik dibandingkan MFCC dan LFCC, CQCC tetap hanya memanfaatkan informasi amplitudo sinyal dan mengabaikan informasi fase gelombang. Selain itu, proses ekstraksi menggunakan *Constant-Q Transform* memiliki kompleksitas komputasi yang lebih tinggi sehingga membutuhkan waktu pemrosesan yang lebih lama dibandingkan metode ekstraksi fitur tradisional lainnya.

2.3 Arsitektur *Baseline* RawNet2

RawNet2 merupakan salah satu arsitektur *deep learning* berbasis *end-to-end* yang dirancang khusus untuk memproses sinyal audio mentah (*raw waveform*) secara langsung tanpa memerlukan proses ekstraksi fitur manual, seperti spektrogram maupun koefisien cepstral, pada tahap awal. Melalui pendekatan ini, model mempelajari representasi fitur akustik secara otomatis dari sinyal suara mentah selama proses pelatihan. Secara umum, arsitektur RawNet2 terdiri atas beberapa komponen utama sebagai berikut.

a. ***Sinc-Convolution* (*SincNet*)**

Lapisan pertama pada RawNet2 menggunakan operasi konvolusi berbasis fungsi *sinc* (*SincNet*) sebagai pengganti filter konvolusi konvensional. Pendekatan ini memungkinkan model mempelajari *filterbank* frekuensi secara langsung dari sinyal audio satu dimensi (*1D raw waveform*), sehingga karakteristik akustik penting dapat diekstraksi secara adaptif selama proses pelatihan.

b. ***Residual Blocks* (*ResNet*)**

Setelah proses ekstraksi awal menggunakan *SincNet*, fitur yang dihasilkan diproses melalui beberapa *Residual Block* yang mengadopsi konsep *Residual Network* (*ResNet*). Setiap blok memiliki *shortcut connection* yang membantu mengatasi permasalahan *vanishing gradient* sekaligus memungkinkan pembelajaran representasi fitur akustik yang lebih mendalam secara hierarkis.

c. ***Gated Recurrent Unit* (GRU)**

Keluaran dari blok residual selanjutnya diteruskan ke lapisan *Gated Recurrent Unit* (GRU). Lapisan ini bertugas memodelkan hubungan temporal antarsegmen suara sehingga model mampu menangkap pola perubahan karakteristik suara sepanjang waktu yang penting dalam proses identifikasi suara sintetis.

Meskipun RawNet2 mampu mempelajari representasi akustik langsung dari sinyal mentah, arsitektur ini memiliki keterbatasan karena seluruh parameter model dilatih dari awal (*supervised learning from scratch*). Akibatnya, RawNet2 memerlukan dataset berlabel dalam jumlah yang sangat besar agar mampu menghasilkan model dengan kemampuan generalisasi yang baik. Pada kondisi

dataset yang terbatas, model cenderung mengalami *overfitting*, membutuhkan waktu pelatihan yang lebih lama, serta memerlukan sumber daya komputasi yang lebih tinggi dibandingkan pendekatan *transfer learning* berbasis model *self-supervised learning*, seperti *Wav2Vec 2.0*. Oleh karena itu, penelitian ini memilih menggunakan *Wav2Vec 2.0* sebagai model utama karena lebih efisien dalam memanfaatkan pengetahuan hasil *pre-training* dan lebih sesuai untuk proses *fine-tuning* pada dataset *Fake-or-Real*.

2.4 Voice Phishing

Voice phishing atau yang sering disebut sebagai *vishing* merupakan salah satu bentuk serangan *social engineering* yang memanfaatkan komunikasi suara untuk memperoleh informasi sensitif dari korban melalui manipulasi psikologis dan penyamaran identitas. Dalam serangan ini, pelaku biasanya menghubungi korban melalui panggilan telepon atau layanan komunikasi berbasis internet dengan mengaku sebagai pihak yang memiliki otoritas atau kredibilitas tertentu, seperti petugas perbankan, lembaga pemerintah, atau penyedia layanan digital. Tujuan utama dari serangan tersebut adalah untuk meyakinkan korban agar secara sukarela memberikan informasi rahasia seperti data identitas pribadi, nomor rekening, kata sandi, maupun kode autentikasi sekali pakai yang dapat digunakan untuk melakukan akses tidak sah terhadap sistem atau layanan finansial. Dalam literatur keamanan siber, voice phishing dikategorikan sebagai salah satu varian dari serangan phishing yang memanfaatkan media komunikasi suara sebagai sarana utama untuk mengeksploitasi faktor kepercayaan manusia dalam proses komunikasi [12] [13].

Berbeda dengan metode phishing konvensional yang menggunakan email atau pesan teks, voice phishing memiliki karakteristik interaksi yang bersifat sinkron dan interpersonal sehingga memungkinkan pelaku memanfaatkan teknik persuasi secara lebih efektif selama percakapan berlangsung. Interaksi langsung melalui komunikasi suara memungkinkan pelaku untuk menyesuaikan strategi manipulasi secara dinamis berdasarkan respons korban, misalnya dengan menciptakan rasa urgensi, tekanan psikologis, maupun ancaman keamanan yang mendorong korban untuk segera mengambil keputusan tanpa melakukan verifikasi lebih lanjut. Studi terbaru menunjukkan bahwa komunikasi berbasis suara memiliki tingkat keberhasilan serangan *social engineering* yang lebih tinggi dibandingkan dengan komunikasi berbasis teks karena adanya unsur kepercayaan interpersonal

serta keterbatasan kemampuan pengguna dalam memverifikasi identitas penelepon secara real time [14]. Selain itu, meningkatnya penggunaan layanan komunikasi digital dan sistem perbankan daring juga memperluas permukaan serangan (*attack surface*) yang dapat dimanfaatkan oleh pelaku kejahatan siber untuk melakukan penipuan berbasis komunikasi suara.

Dampak dari serangan voice phishing tidak hanya terbatas pada kerugian finansial individu, tetapi juga dapat mempengaruhi keamanan informasi pada tingkat organisasi. Informasi sensitif yang diperoleh melalui serangan vishing dapat dimanfaatkan untuk melakukan berbagai bentuk kejahatan lanjutan seperti pengambilalihan akun (*account takeover*), pencurian identitas digital, hingga akses tidak sah terhadap sistem internal organisasi. Penelitian dalam bidang keamanan siber menunjukkan bahwa sektor perbankan, layanan finansial, dan layanan digital merupakan target utama dari serangan voice phishing karena memiliki potensi keuntungan ekonomi yang tinggi bagi pelaku [15]. Selain kerugian ekonomi secara langsung, serangan ini juga dapat menimbulkan dampak jangka panjang berupa penurunan kepercayaan pengguna terhadap sistem komunikasi dan layanan digital yang digunakan oleh organisasi [16].

Perkembangan teknologi kecerdasan buatan dalam beberapa tahun terakhir semakin meningkatkan kompleksitas dan tingkat ancaman dari serangan voice phishing. Teknologi *speech synthesis* modern memungkinkan pembuatan suara sintetis yang sangat realistis melalui pendekatan *deep learning* seperti *neural text-to-speech* dan *voice cloning*. Teknologi ini memungkinkan pelaku untuk menghasilkan suara yang menyerupai karakteristik akustik seseorang secara spesifik, termasuk intonasi, aksen, dan pola bicara. Fenomena ini dikenal sebagai *AI-generated speech* atau *deepfake audio*. Berbagai penelitian menunjukkan bahwa model generatif modern mampu menghasilkan suara sintetis dengan kualitas yang sangat tinggi sehingga sulit dibedakan oleh manusia dalam kondisi percakapan normal [17]. Bahkan dalam beberapa eksperimen, tingkat keberhasilan manusia dalam membedakan suara asli dengan suara sintetis pada skenario komunikasi telepon relatif rendah, yang menunjukkan bahwa pendekatan berbasis persepsi manusia tidak lagi cukup untuk mendeteksi potensi penyalahgunaan teknologi tersebut [18].

Kemajuan teknologi sintesis suara ini menimbulkan tantangan baru dalam bidang keamanan siber karena memungkinkan pelaku kejahatan untuk melakukan serangan voice phishing yang jauh lebih meyakinkan dibandingkan metode tradisional. Dalam beberapa kasus, teknologi *voice cloning* dapat digunakan

untuk meniru suara individu tertentu seperti anggota keluarga, pimpinan organisasi, atau pejabat perusahaan guna meningkatkan kredibilitas serangan. Kondisi ini mendorong munculnya kebutuhan terhadap sistem deteksi otomatis yang mampu menganalisis karakteristik akustik dari sinyal suara untuk mengidentifikasi apakah suatu audio merupakan suara manusia asli atau suara sintetis yang dihasilkan oleh sistem kecerdasan buatan [19]. Oleh karena itu, penelitian mengenai metode deteksi *AI-generated speech* menjadi sangat penting sebagai upaya mitigasi terhadap ancaman voice phishing berbasis kecerdasan buatan serta untuk meningkatkan keamanan sistem komunikasi suara di era digital saat ini.

2.5 Artificial Intelligence Generated Speech

Artificial Intelligence Generated Speech (AI-generated speech) merupakan teknologi sintesis suara yang dihasilkan melalui model kecerdasan buatan, khususnya pendekatan *deep learning*, untuk menghasilkan sinyal audio yang menyerupai karakteristik suara manusia secara alami. Berbeda dengan pendekatan sintesis suara tradisional yang bergantung pada aturan akustik atau *concatenative synthesis*, sistem modern memanfaatkan arsitektur *neural network* yang mampu mempelajari representasi kompleks dari pola fonetik, prosodi, serta karakteristik akustik pembicara langsung dari data audio berskala besar [20]. Pendekatan ini memungkinkan model untuk menghasilkan suara sintetis dengan tingkat naturalness yang tinggi, sehingga semakin sulit dibedakan dari suara manusia asli.

Perkembangan model *neural text-to-speech* (TTS) berbasis *transformer* dan *self-supervised learning* telah meningkatkan kualitas sintesis suara secara signifikan dalam beberapa tahun terakhir. Model generatif modern seperti VITS dan sistem TTS berbasis pembelajaran end-to-end mampu menghasilkan suara dengan kualitas mendekati rekaman manusia asli serta mempertahankan konsistensi karakteristik pembicara (*speaker identity*) [21]. Selain itu, kemajuan pada teknologi *voice cloning* memungkinkan sistem untuk mereplikasi karakteristik suara individu tertentu hanya dari sampel audio berdurasi pendek, sehingga membuka kemungkinan pembuatan suara sintetis yang sangat spesifik terhadap identitas seseorang [22].

Kemampuan model AI dalam menghasilkan suara yang realistis menimbulkan tantangan baru dalam bidang keamanan informasi karena teknologi ini dapat disalahgunakan untuk membuat *deepfake audio*. Dalam konteks kejahatan siber, AI-generated speech memungkinkan pelaku untuk melakukan penyamaran

identitas suara secara meyakinkan dalam komunikasi berbasis telepon atau layanan suara daring. Beberapa studi menunjukkan bahwa sistem sintesis suara berbasis *neural speech generation* mampu menghasilkan audio dengan kualitas yang cukup tinggi sehingga sulit dibedakan oleh pendengar manusia tanpa bantuan sistem analisis otomatis [23]. Kondisi ini meningkatkan risiko penyalahgunaan teknologi tersebut dalam serangan *social engineering* seperti voice phishing, di mana pelaku dapat memanfaatkan suara sintetis untuk meniru individu terpercaya guna memanipulasi korban.

Seiring meningkatnya kualitas dan ketersediaan teknologi sintesis suara berbasis kecerdasan buatan, penelitian mengenai metode deteksi AI-generated speech menjadi semakin penting. Pendekatan deteksi modern umumnya berfokus pada analisis karakteristik akustik dan representasi fitur spektral dari sinyal suara untuk mengidentifikasi artefak yang dihasilkan oleh model generatif. Dalam beberapa penelitian terbaru, model representasi audio berbasis *self-supervised learning* seperti wav2vec 2.0 menunjukkan potensi yang signifikan dalam mengekstraksi fitur akustik yang lebih robust untuk membedakan antara suara manusia asli dan suara sintetis yang dihasilkan oleh sistem kecerdasan buatan [24].

2.6 Speech Processing

Speech Processing merupakan cabang dari pemrosesan sinyal digital yang berfokus pada analisis, pemodelan, dan interpretasi sinyal suara manusia menggunakan metode komputasional. Bidang ini mencakup berbagai teknik untuk mengolah sinyal audio sehingga informasi akustik yang terkandung di dalamnya dapat dianalisis oleh sistem komputer. Dalam beberapa tahun terakhir, perkembangan metode berbasis *deep learning* telah meningkatkan kemampuan sistem speech processing dalam memahami karakteristik kompleks dari sinyal suara, sehingga banyak digunakan pada berbagai aplikasi seperti pengenalan ucapan, verifikasi pembicara, sintesis suara, serta deteksi keaslian audio [25]. Dalam konteks penelitian ini, speech processing digunakan untuk mengekstraksi representasi akustik dari sinyal suara yang kemudian dimanfaatkan untuk mendeteksi apakah suatu audio merupakan suara manusia asli atau suara sintetis yang dihasilkan oleh sistem kecerdasan buatan.

2.6.1 Speech Acquisition

Speech acquisition merupakan tahap awal dalam sistem speech processing yang bertujuan untuk memperoleh sinyal suara dari sumber audio menggunakan perangkat perekam seperti mikrofon. Pada tahap ini, sinyal suara analog dikonversi menjadi sinyal digital melalui proses *analog-to-digital conversion* sehingga dapat diproses oleh sistem komputer. Kualitas proses akuisisi suara sangat mempengaruhi performa analisis selanjutnya karena faktor seperti noise lingkungan, kualitas perangkat perekam, serta kondisi akustik dapat mempengaruhi karakteristik sinyal suara yang dihasilkan [26].

2.6.2 Preprocessing

Tahap preprocessing bertujuan untuk meningkatkan kualitas sinyal suara sebelum dilakukan analisis lebih lanjut. Proses ini umumnya meliputi normalisasi amplitudo, pengurangan noise, serta segmentasi sinyal untuk memisahkan bagian suara yang relevan dari komponen yang tidak diperlukan. Tahap ini penting untuk memastikan bahwa sinyal audio yang dianalisis memiliki kualitas yang lebih konsisten sehingga dapat meningkatkan keandalan fitur yang dihasilkan pada tahap berikutnya [27] [28].

2.6.3 Feature Extraction

Feature extraction merupakan proses mengekstraksi karakteristik penting dari sinyal suara [29]. Pendekatan berbasis self-supervised learning seperti wav2vec 2.0 memungkinkan model mempelajari representasi audio secara langsung dari sinyal mentah.

2.6.4 Classification

Tahap classification menggunakan algoritma machine learning maupun deep learning untuk mengklasifikasikan sinyal audio [30]. Model seperti wav2vec 2.0 terbukti efektif dalam berbagai tugas pemrosesan suara termasuk deteksi spoofing dan AI-generated speech [31].

2.7 Deepfake Audio

Deepfake audio merupakan teknologi sintesis suara yang memanfaatkan model *deep learning* untuk menghasilkan atau memanipulasi sinyal suara manusia sehingga menyerupai karakteristik akustik pembicara tertentu [32] [33]. Teknologi ini berkembang pesat seiring dengan kemajuan metode pembelajaran mendalam yang mampu mempelajari pola kompleks dari sinyal audio, termasuk prosodi, intonasi, dan timbre suara manusia [34]. Akibatnya, suara sintetis yang dihasilkan oleh model modern sering kali memiliki tingkat naturalness yang tinggi dan semakin sulit dibedakan dari suara manusia asli [35].

Dalam implementasinya, pembuatan deepfake audio umumnya memanfaatkan berbagai arsitektur model generatif seperti *Generative Adversarial Networks* (GAN), *Variational Autoencoder* (VAE), serta sistem *neural text-to-speech* berbasis pembelajaran end-to-end. Model generatif tersebut memungkinkan proses pembangkitan suara sintetis yang dapat meniru karakteristik pembicara tertentu melalui teknik yang dikenal sebagai *voice cloning*. Kemampuan ini memungkinkan sistem untuk menghasilkan suara yang menyerupai identitas vokal seseorang bahkan hanya dari sampel audio berdurasi pendek.

Meskipun teknologi deepfake audio memiliki berbagai manfaat pada aplikasi seperti asisten virtual, produksi media digital, serta teknologi aksesibilitas, penyalahgunaannya menimbulkan berbagai risiko keamanan yang signifikan. Deepfake audio dapat dimanfaatkan untuk melakukan impersonasi suara secara meyakinkan sehingga berpotensi digunakan dalam berbagai bentuk kejahatan siber seperti penipuan suara, manipulasi informasi, dan serangan *social engineering* berbasis komunikasi suara [36]. Dalam beberapa kasus, teknologi ini bahkan digunakan untuk meniru suara individu terpercaya guna meningkatkan keberhasilan serangan *voice phishing* [36].

Seiring dengan meningkatnya kualitas teknologi sintesis suara berbasis kecerdasan buatan, penelitian mengenai metode deteksi deepfake audio menjadi bidang yang berkembang pesat dalam keamanan siber. Berbagai pendekatan telah dikembangkan untuk membedakan suara manusia asli dengan suara sintetis dengan menganalisis karakteristik akustik serta artefak yang dihasilkan oleh model generatif. Pendekatan modern banyak memanfaatkan teknik *deep learning* serta representasi audio berbasis *self-supervised learning* untuk mengekstraksi fitur yang lebih robust dari sinyal suara [37]. Pendekatan ini menjadi penting dalam mendukung sistem deteksi *AI-generated speech* sebagai upaya mitigasi terhadap

ancaman voice phishing berbasis manipulasi suara.

2.8 Transfer Learning

Transfer learning dalam konteks pembelajaran mesin merujuk pada pendekatan yang memanfaatkan representasi pengetahuan yang telah dipelajari oleh suatu model pada tahap pelatihan awal (*pretraining*) untuk meningkatkan kinerja model pada tugas target yang berbeda namun masih berada dalam domain yang berkaitan [38]. Berbeda dengan pendekatan pelatihan konvensional yang menginisialisasi parameter model secara acak, transfer learning memungkinkan parameter model yang telah mempelajari struktur representasi data pada skala besar digunakan kembali sebagai dasar untuk pembelajaran pada dataset yang lebih terbatas. Pendekatan ini terbukti mampu meningkatkan kemampuan generalisasi model terutama pada skenario di mana distribusi data pada tugas target tidak cukup besar untuk melatih model *deep learning* secara efektif dari awal [39].

Dalam bidang *speech processing*, transfer learning banyak diimplementasikan melalui pemanfaatan *pretrained speech models* yang dilatih menggunakan pendekatan *self-supervised learning* pada korpus audio berskala besar tanpa anotasi [40]. Model tersebut mempelajari representasi akustik tingkat tinggi dari sinyal suara, seperti pola temporal, struktur spektral, serta karakteristik fonetik yang bersifat umum pada berbagai jenis data audio [41]. Representasi yang diperoleh pada tahap pretraining kemudian dapat disesuaikan kembali melalui proses *fine-tuning* menggunakan dataset yang lebih kecil untuk menyelesaikan tugas spesifik seperti pengenalan ucapan, verifikasi pembicara, maupun analisis karakteristik suara.

Pendekatan ini menjadi sangat relevan dalam penelitian yang berkaitan dengan deteksi suara sintetis atau *AI-generated speech*. Model yang telah dilatih pada korpus audio besar mampu menangkap representasi akustik yang lebih kaya dibandingkan metode ekstraksi fitur konvensional, sehingga dapat mengidentifikasi artefak halus yang dihasilkan oleh sistem sintesis suara berbasis kecerdasan buatan [42]. Oleh karena itu, transfer learning dengan memanfaatkan model representasi audio modern seperti wav2vec 2.0 menjadi salah satu pendekatan yang banyak digunakan dalam penelitian deteksi *deepfake speech* dan analisis keaslian suara dalam konteks keamanan siber [43].

2.9 Wav2Vec 2.0

Pendekatan *self-supervised learning* dalam pemrosesan suara berkembang pesat melalui pemanfaatan data audio tidak berlabel untuk mempelajari representasi akustik yang bersifat umum dan dapat digunakan kembali pada berbagai tugas *downstream*. Salah satu pendekatan yang relevan adalah model berbasis prediksi unit tersembunyi seperti HuBERT, yang menunjukkan bahwa representasi laten yang dipelajari dari sinyal suara mampu menangkap struktur fonetik dan pola akustik tanpa memerlukan anotasi manual [44]. Pendekatan ini menjadi landasan bagi pengembangan model seperti Wav2Vec 2.0 dalam memanfaatkan pembelajaran representasi berbasis konteks.

Arsitektur model berbasis *self-supervised speech representation* umumnya mengombinasikan*convolutional feature encoder* untuk mengekstraksi representasi awal dari sinyal audio mentah, serta *Transformer-based context network* untuk menangkap ketergantungan jangka panjang dalam sinyal suara. Studi menunjukkan bahwa representasi kontekstual yang dihasilkan oleh model berbasis Transformer secara signifikan meningkatkan performa pada berbagai tugas seperti pengenalan ucapan dan klasifikasi audio [45].

Dalam proses pembelajaran, pendekatan *contrastive learning* digunakan untuk membedakan representasi yang benar dari sampel negatif, sehingga model dapat mempelajari struktur distribusi data audio secara lebih efektif. Mekanisme ini memungkinkan pembelajaran representasi yang robust terhadap variasi sinyal suara dan noise [46]. Selain itu, pemanfaatan kuantisasi untuk menghasilkan target diskrit juga menjadi komponen penting dalam meningkatkan kualitas pembelajaran representasi audio pada model berbasis self-supervised.

Dalam konteks keamanan, representasi yang dihasilkan oleh model self-supervised seperti Wav2Vec 2.0 telah dimanfaatkan dalam deteksi *deepfake audio*. Penelitian terbaru menunjukkan bahwa fitur berbasis representasi laten mampu mengidentifikasi artefak sintetis yang tidak mudah dikenali oleh fitur akustik tradisional, terutama pada serangan berbasis *AI-generated speech* yang semakin kompleks [47].

2.10 Fine-Tuning Wav2Vec 2.0 untuk Klasifikasi Audio

Tahap *fine-tuning* merupakan proses adaptasi model Wav2Vec 2.0 yang telah melalui tahap *self-supervised pretraining* agar dapat digunakan pada tugas

downstream seperti klasifikasi audio dan deteksi *AI-generated speech*. Dalam pendekatan ini, representasi akustik yang telah dipelajari dari data tidak berlabel dimanfaatkan sebagai fitur awal yang kemudian disesuaikan menggunakan data berlabel pada domain target [48].

Pada tahap implementasi, *fine-tuning* dilakukan dengan menambahkan *classification head* pada bagian akhir model, yang menggunakan representasi kontekstual dari *Transformer* sebagai input. Parameter model kemudian diperbarui melalui proses optimasi berbasis *backpropagation*, baik secara parsial maupun penuh, sehingga model dapat menyesuaikan representasi terhadap karakteristik data target [49].

Sejumlah penelitian menunjukkan bahwa pendekatan *fine-tuning* pada model berbasis *self-supervised learning* seperti Wav2Vec 2.0 memberikan peningkatan performa yang signifikan pada berbagai tugas speech processing, termasuk klasifikasi audio dan deteksi spoofing suara [50] [51]. Dalam konteks deteksi *deepfake audio*, representasi yang dihasilkan memungkinkan identifikasi artefak akustik halus yang sulit dideteksi menggunakan fitur tradisional, sehingga meningkatkan kemampuan sistem dalam membedakan suara asli dan sintetis.

2.11 Fungsi Loss

Dalam tugas klasifikasi audio berbasis *deep learning*, fungsi loss digunakan untuk mengukur tingkat kesalahan antara prediksi model dan label sebenarnya pada data pelatihan [52]. Nilai loss ini kemudian digunakan dalam proses optimasi untuk memperbarui parameter model melalui algoritma *backpropagation*.

Salah satu fungsi loss yang paling umum digunakan pada masalah klasifikasi multikelas adalah *Cross-Entropy Loss*. Fungsi ini mengukur perbedaan antara distribusi probabilitas prediksi yang dihasilkan model dan distribusi label sebenarnya [53]. Cross-entropy banyak digunakan karena mampu memberikan penalti yang lebih besar ketika model memberikan probabilitas tinggi pada kelas yang salah.

Secara matematis, fungsi *Cross-Entropy Loss* dinyatakan pada Persamaan 2.1.

$$L = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{2.1}$$

Nilai loss yang lebih kecil menunjukkan bahwa distribusi probabilitas prediksi model semakin mendekati label sebenarnya. Oleh karena itu, tujuan proses pelatihan model adalah meminimalkan nilai fungsi loss agar model mampu menghasilkan prediksi yang lebih akurat [54].

2.12 Metrik Evaluasi

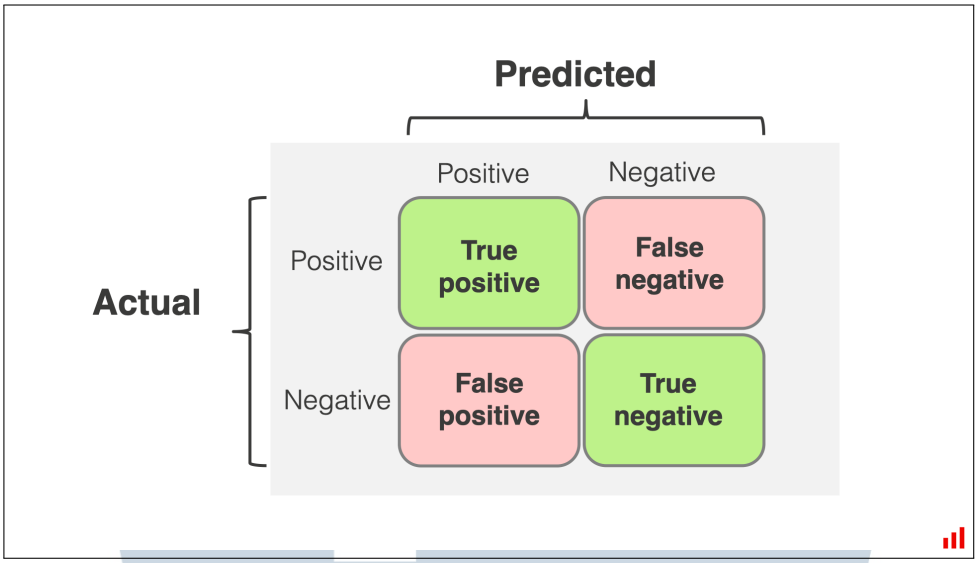
Evaluasi model dilakukan untuk menilai performa sistem dalam mendeteksi apakah suatu sinyal suara merupakan suara manusia asli atau suara yang dihasilkan oleh kecerdasan buatan. Pada penelitian klasifikasi, performa model umumnya dianalisis menggunakan *confusion matrix* serta beberapa metrik evaluasi turunan seperti accuracy, precision, recall, dan F1-score [55].

Metrik accuracy digunakan untuk mengukur proporsi prediksi yang benar terhadap seluruh data, sedangkan precision dan recall memberikan gambaran yang lebih spesifik terkait performa model dalam mendeteksi kelas tertentu. F1-score merupakan harmonisasi antara precision dan recall yang sering digunakan ketika terdapat ketidakseimbangan data [56].

Dalam konteks deteksi *AI-generated speech*, penggunaan berbagai metrik evaluasi menjadi penting untuk memastikan bahwa model tidak hanya memiliki akurasi tinggi, tetapi juga mampu mendeteksi kelas minoritas secara efektif, seperti kasus suara sintetis yang jumlahnya lebih sedikit dibandingkan suara asli [57].

2.12.1 Confusion Matrix

Confusion matrix merupakan tabel yang digunakan untuk menggambarkan kinerja model klasifikasi dengan membandingkan hasil prediksi model terhadap label sebenarnya. Pada kasus klasifikasi biner seperti deteksi *AI-generated speech*, confusion matrix terdiri dari empat komponen utama yaitu *True Positive (TP)*, *True Negative (TN)*, *False Positive (FP)*, dan *False Negative (FN)* [54].



Gambar 2.1. Confusion Matrix pada Klasifikasi Biner
Sumber: [58]

2.12.2 Accuracy

Accuracy merupakan metrik yang mengukur proporsi prediksi yang benar dibandingkan dengan seluruh jumlah data pengujian [54]. Nilai *Accuracy* dihitung menggunakan Persamaan 3.3.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.2}$$

Nilai *Accuracy* yang tinggi menunjukkan bahwa model memiliki tingkat ketepatan prediksi yang baik secara keseluruhan. Namun, metrik ini dapat menjadi kurang representatif apabila digunakan pada dataset yang memiliki distribusi kelas tidak seimbang (*imbalanced dataset*) [54].

2.12.3 Precision

Precision mengukur kemampuan model dalam menghasilkan prediksi positif yang benar dibandingkan dengan seluruh prediksi positif yang dihasilkan model [54]. Nilai *Precision* dihitung menggunakan Persamaan 3.4.

$$Precision = \frac{TP}{TP + FP} \tag{2.3}$$

Nilai *Precision* yang tinggi menunjukkan bahwa model memiliki tingkat kesalahan prediksi positif (*False Positive*) yang rendah. Oleh karena itu, metrik ini penting pada aplikasi yang menuntut minimnya kesalahan dalam mengklasifikasikan data negatif sebagai positif [54].

2.12.4 Recall

Recall mengukur kemampuan model dalam mendeteksi seluruh data yang benar-benar termasuk ke dalam kelas positif [54]. Nilai *Recall* dihitung menggunakan Persamaan 3.5.

$$Recall = \frac{TP}{TP + FN} \tag{2.4}$$

Nilai *Recall* yang tinggi menunjukkan bahwa model mampu mendeteksi sebagian besar data positif pada dataset. Metrik ini sangat penting pada sistem deteksi karena dapat meminimalkan jumlah *False Negative* [54].

2.12.5 F1-Score

F1-Score merupakan rata-rata harmonik antara nilai *Precision* dan *Recall*. Metrik ini digunakan ketika diperlukan keseimbangan antara kemampuan model dalam mendeteksi data positif dan ketepatan prediksi positif [54]. Nilai *F1-Score* dihitung menggunakan Persamaan 3.6.

$$F1\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2.5}$$

Nilai *F1-Score* yang tinggi menunjukkan bahwa model memiliki keseimbangan yang baik antara *Precision* dan *Recall*, sehingga metrik ini banyak digunakan untuk mengevaluasi model klasifikasi pada dataset yang memiliki distribusi kelas tidak seimbang [54].
Keterangan:

- TP (*True Positive*) adalah jumlah data positif yang berhasil diklasifikasikan dengan benar.

- TN (*True Negative*) adalah jumlah data negatif yang berhasil diklasifikasikan dengan benar.
- FP (*False Positive*) adalah jumlah data negatif yang salah diklasifikasikan sebagai positif.
- FN (*False Negative*) adalah jumlah data positif yang salah diklasifikasikan sebagai negatif.