BAB I
PENDAHULUAN

1.1. Latar Belakang

Perkembangan teknologi kecerdasan buatan telah mendorong munculnya audio *deepfake*, yaitu suara sintetis yang dihasilkan untuk meniru suara manusia dengan sangat realistis. Teknologi ini pada awalnya dimanfaatkan untuk tujuan positif seperti *audiobook* [1]. Namun, seiring kemajuan teknologi kecerdasan buatan, fenomena tersebut menimbulkan ancaman serius terhadap keamanan publik [2]. Penyebaran audio *deepfake* memberikan risiko besar berupa potensi penyalahgunaan untuk penipuan, pencurian identitas, dan manipulasi informasi yang dapat merusak kepercayaan masyarakat serta mengancam stabilitas sosial [3]. Selain itu, keberadaan audio *deepfake* juga menjadi ancaman terhadap *voice assistant* dan sistem berbasis pengenalan suara yang banyak digunakan dalam layanan sehari-hari, sebab kelemahan autentikasi suara membuka peluang untuk peretasan, penipuan finansial, hingga penyebaran disinformasi secara masif [4].

Fenomena audio *deepfake* menjadi perhatian global karena kemampuannya meniru suara manusia dengan tingkat kemiripan yang hampir sempurna sehingga sulit dibedakan dari suara asli. Disamping itu, kemajuan model generatif telah membuat proses pembuatan suara sintetis semakin mudah diakses publik. Kondisi ini menimbulkan ancaman serius terhadap keamanan dan privasi, karena audio sintetis dapat digunakan untuk menipu autentikasi berbasis suara, menyebarkan disinformasi, serta melakukan penipuan identitas [5]. Oleh karena itu, dibutuhkan sistem yang mampu membedakan antara suara asli dengan suara hasil sintetis. Di sisi lain, sistem deteksi audio *deepfake* yang ada masih menghadapi kendala dalam kemampuan generalisasi, sehingga efektivitasnya dalam mengenali variasi manipulasi suara di situasi nyata masih terbatas [6]. Situasi ini menunjukkan pentingnya pengembangan metode klasifikasi yang lebih adaptif dan mampu diterapkan secara praktis untuk menghadapi penyalahgunaan teknologi AI generatif.

Sistem klasifikasi audio *deepfake* pada umumnya masih bergantung pada teknik ekstraksi fitur konvensional seperti *Mel-Frequency Cepstral Coefficients* (MFCC), *Linear Frequency Cepstral Coefficients* (LFCC), dan *spectrogram*. Meskipun metode tersebut mampu merepresentasikan karakteristik dasar dari sinyal suara, sejumlah penelitian menunjukkan bahwa pendekatan ini memiliki kelemahan dalam menghadapi kompleksitas audio *deepfake*. Model berbasis ekstraksi fitur konvensional memiliki kemampuan generalisasi yang rendah, di mana performanyan menurun drastis saat diuji pada *dataset* yang berbeda dari data pelatihan [7]. Hal ini menunjukkan bahwa ekstraksi fitur konvensional cenderung mengabaikan fitur penting dan berpotensi menimbulkan bias [8].

Berdasarkan beberapa kelemahan yang telah disebutkan, penelitian ini mengusulkan YAMNet sebagai metode ekstraksi fitur berbasis *transfer learning*. YAMNet dirancang untuk menghasilkan representasi audio yang lebih kaya dibandingkan pendekatan konvensional. Model ini menggunakan arsitektur MobileNetV1 dengan *depthwise separable convolution* untuk mempelajari pola semantik dari sinyal suara secara mendalam. Pendekatan ini memungkinkan sistem mengenali karakteristik audio secara lebih adaptif terhadap variasi data, sehingga meningkatkan kemampuan generalisasi pada proses klasifikasi. Pada penelitian sebelumnya, penggunaan YAMNet mampu menghasilkan peningkatan akurasi pada berbagai tugas klasifikasi audio, sekaligus mengurangi risiko *overfitting* [9].

Namun, ekstraksi fitur yang unggul saja tidak cukup untuk memastikan akurasi sistem klasifikasi audio *deepfake*. Oleh karena itu, diperlukan model klasifikasi yang mampu mengolah representasi fitur tersebut secara optimal. *Deep Neural Network* (DNN) dipilih karena memiliki kemampuan untuk mempelajari hubungan nonlinier antar fitur melalui proses pelatihan berbasis banyak *hidden layer*. Struktur berlapis ini memungkinkan DNN untuk mempelajari pola-pola kompleks dari fitur yang dihasilkan oleh YAMNet, sehingga mampu membedakan perbedaan kecil namun signifikan dalam struktur sinyal audio. Berdasarkan penelitian sebelumnya, metode klasifikasi berbasis DNN terbukti memberikan performa yang baik dalam mendeteksi audio *deepfake* dengan tingkat kesalahan rendah pada berbagai *dataset* uji [10].

Meskipun kombinasi antara YAMNet dan DNN menunjukkan potensi yang menjanjikan, masih terdapat tantangan teknis terkait representasi hasil *embedding* yang dihasilkan oleh YAMNet. *Embedding* tersebut umumnya berbentuk sekuens dengan jumlah frame yang tidak tetap, bergantung pada durasi masing-masing audio. Kondisi ini menjadikan proses komputasi menjadi lebih berat. Penelitian terdahulu pernah menggunakan operasi *reduce mean* yang secara praktis dapat meringankan beban komputasi dengan menghasilkan satu vektor *embedding* tunggal [11]. Namun, pendekatan tersebut belum mempertimbangkan kombinasi statistik yang lebih komprehensif untuk meningkatkan kemampuan model dalam mengklasifikasikan audio.

Oleh karena itu, penelitian ini berkontribusi dengan mengeksplorasi metode agregasi statistik yang lebih luas, mencakup nilai rata-rata (*mean*), simpangan baku (*standard deviation*), nilai minimum (*min*), dan nilai maksimum (*max*), untuk mentransformasi sekuens *embedding* menjadi representasi yang lebih informatif. Pendekatan ini diharapkan mampu menangkap karakteristik suara dengan lebih efektif dibandingkan metode yang hanya menggunakan rata-rata, sehingga menghasilkan sistem klasifikasi audio *deepfake* yang lebih andal dan *robust*.

1.2. Rumusan Masalah

Berdasarkan paparan latar belakang di atas, rumusan masalah dalam penelitian ini dapat diuraikan sebagai berikut:

1. Bagaimana penerapan metode agregasi statistik terhadap *embedding* hasil ekstraksi YAMNet dapat menghasilkan representasi fitur yang efektif untuk membedakan audio *deepfake* dan audio asli?
2. Bagaimana efektivitas kombinasi YAMNet dan DNN dalam melakukan generalisasi lintas *dataset* pada tugas klasifikasi audio *deepfake*?
3. Bagaimana model klasifikasi yang telah dilatih dapat diimplementasikan ke dalam sebuah antarmuka pengguna grafis (GUI) dengan menggunakan Flask?

1.3. Batasan Masalah

Agar penelitian ini lebih terfokus, diperlukan adanya batasan masalah yang mendefinisikan ruang lingkup studi. Batasan-batasan ini ditetapkan untuk menghindari perluasan topik yang dapat mengganggu pencapaian tujuan penelitian.

1. *Dataset* sekunder yang digunakan dibatasi pada audioset yang disediakan oleh In-The-Wild, dengan kategori *bona-fide* dan *spoof*.
2. Fokus penelitian terbatas pada klasifikasi biner antara audio asli dan audio palsu, tanpa membedakan jenis atau teknik manipulasi suara yang digunakan dalam pembuatan *deepfake*.
3. Proses ekstraksi fitur dibatasi pada penggunaan model *pre-train* YAMNet yang disediakan oleh TensorFlow, tanpa menerapkan atau membandingkan metode ekstraksi fitur lain seperti MFCC, spectrogram, atau model *embedding* alternatif.
4. Teknik agregasi *embedding* dibatasi pada penggunaan operator reduksi dasar yang tersedia di TensorFlow, dengan menggunakan fungsi *reduce mean*, *reduce std*, *reduce min*, dan *reduce max*.
5. Arsitektur klasifikasi dibatasi pada model *Deep Neural Network* (DNN) berbasis TensorFlow, tanpa melibatkan perbandingan dengan arsitektur model lain.

1.4. Tujuan Penelitian

Tujuan utama dari penelitian ini adalah untuk menguji efektivitas model YAMNet yang dikombinasikan dengan DNN dalam mengklasifikasikan audio *deepfake*. Secara spesifik, tujuan penelitian ini adalah:

1. Menganalisis efektivitas penerapan metode agregasi statistik terhadap *embedding* YAMNet dalam menghasilkan fitur untuk membedakan audio *deepfake* dan audio asli.
2. Mengukur tingkat efektivitas kombinasi YAMNet dan DNN dalam melakukan generalisasi lintas *dataset* untuk tugas klasifikasi audio *deepfake*.
3. Mengimplementasikan model klasifikasi audio *deepfake* yang telah dilatih ke dalam sebuah prototipe berbasis web dengan GUI yang menggunakan Flask.

1.5. Manfaat Penelitian

Dalam pelaksanaannya, penelitian ini ditujukan untuk menghasilkan manfaat teoritis dan praktis, di antaranya:

1. Manfaat Teoritis

Penelitian ini diharapkan dapat memberikan kontribusi pada pengembangan ilmu pengetahuan di bidang pengolahan sinyal audio dan kecerdasan buatan, khususnya dalam penerapan arsitektur *transfer learning* YAMNet yang dipadukan dengan DNN untuk klasifikasi audio *deepfake*. Hasil penelitian ini juga dapat memperkaya literatur terkait keamanan digital dan deteksi manipulasi suara berbasis AI, serta menjadi rujukan bagi penelitian selanjutnya di bidang klasifikasi audio sintetis.

2. Manfaat Praktis

a. Bagi Penyusun

Penelitian ini memberikan pengalaman langsung dalam mengimplementasikan arsitektur YAMNet dengan DNN untuk klasifikasi audio *deepfake*, meningkatkan keterampilan dalam pengolahan data suara, serta memperdalam pemahaman mengenai penerapan AI dalam bidang keamanan audio digital.

b. Bagi Pembaca

Penelitian ini dapat menjadi referensi dalam pengembangan teknologi pengolahan sinyal suara dan sistem keamanan digital. Selain itu, penelitian ini memberikan gambaran tentang bagaimana teknologi kecerdasan buatan, khususnya YAMNet dan DNN, dapat dimanfaatkan untuk mendeteksi dan mengklasifikasikan audio *deepfake*, sehingga mendukung upaya peningkatan keamanan komunikasi digital.