BAB V
PENUTUP

5.1. Kesimpulan

Berdasarkan hasil penelitian dan pengujian yang telah dilakukan, diperoleh beberapa kesimpulan sebagai berikut:

1) Penerapan metode agregasi statistik terhadap *embedding* hasil ekstraksi YAMNet berhasil mengurangi kompleksitas data melalui perubahan dimensi fitur dari bentuk *embedding* awal [n_frames, 1024] menjadi representasi yang lebih ringkas seperti [1, 1024], [1, 2048], dan [1, 4096], serta menurunkan rata-rata waktu pemrosesan dibandingkan metode *embedding* awal. Seluruh metode agregasi yang diuji juga mampu menghasilkan performa klasifikasi yang tinggi dengan rentang akurasi antara 97,58% hingga 99,24%, sehingga menunjukkan bahwa informasi penting pada *embedding* tetap dapat dipertahankan setelah proses agregasi statistik dilakukan dan dapat digunakan secara efektif untuk proses pemodelan klasifikasi audio *deepfake* dan audio asli. Berdasarkan hasil evaluasi, tiga model terbaik diperoleh pada metode Mean + Std Model 1 dengan akurasi 99,24%, Mean Model 1 dengan akurasi 99,06%, dan Mean Model 2 dengan akurasi 99,06%.
2) Kombinasi ekstraksi fitur YAMNet dan model DNN menunjukkan bahwa pengujian langsung lintas *dataset* pada tiga model terbaik berdasarkan hasil pemodelan data sekunder masih menghasilkan performa yang rendah, dengan akurasi primer (ACC_P) hanya berada pada rentang 61,00% hingga 62,37%. Hasil tersebut menunjukkan bahwa model generasi suara saat ini memiliki kemampuan yang semakin canggih dalam mengikuti pola suara manusia asli sehingga karakteristik audio *deepfake* pada *dataset* primer menjadi lebih sulit dibedakan oleh model yang dilatih menggunakan *dataset* sekunder. Oleh karena itu, dilakukan *fine-tuning* menggunakan sebagian data primer melalui skenario pembagian data 20:80 dan 40:60. Pada proses *fine-tuning* tanpa *freeze layer*, akurasi primer meningkat hingga 94,84%, namun akurasi sekunder menurun menjadi 88,17%, yang menunjukkan gejala *catastrophic forgetting*

karena model menjadi lebih menyesuaikan diri terhadap *dataset* primer dan kehilangan sebagian kemampuan pada *dataset* sekunder. Oleh karena itu, pendekatan *freeze layer* dipilih karena mampu memberikan keseimbangan performa yang lebih stabil, dengan Mean + Std Model 1 memperoleh akurasi primer sebesar 92,81% dan akurasi sekunder sebesar 91,41%. Setelah proses *fine-tuning* menggunakan *freeze layer*, hasil evaluasi pada data uji primer menunjukkan nilai *precision* sebesar 0,93, *recall* sebesar 0,93, dan *F1-score* sebesar 0,93, sedangkan pada data uji sekunder diperoleh *weighted average precision* sebesar 0,93, *recall* sebesar 0,91, dan *F1-score* sebesar 0,92. Hasil tersebut menunjukkan bahwa model mampu mempertahankan performa klasifikasi yang baik pada kedua *dataset* setelah proses *fine-tuning* dilakukan.

3) Model klasifikasi yang telah dilatih berhasil diimplementasikan ke dalam antarmuka pengguna grafis (GUI) berbasis Flask melalui sistem VocalSentry untuk mendukung proses deteksi audio *deepfake* secara interaktif. Implementasi GUI mencakup fitur unggah file audio, perekaman suara secara langsung, pemutaran audio, hasil prediksi klasifikasi secara *real-time*, interpretasi hasil menggunakan model Qwen3 1.7B, serta fitur tanya jawab berbasis AI. Berdasarkan hasil implementasi, sistem mampu menyajikan proses deteksi audio deepfake dalam satu antarmuka yang sederhana, informatif, dan mudah digunakan sehingga dapat membantu pengguna memahami hasil klasifikasi tanpa memerlukan pemahaman teknis yang kompleks.

5.2. Saran Pengembangan

Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa saran pengembangan yang dapat dilakukan untuk meningkatkan performa model pada penelitian selanjutnya, antara lain:

1) Menambahkan variasi *dataset* audio *deepfake* dan audio asli dari berbagai sumber, bahasa, serta teknologi generasi suara terbaru untuk meningkatkan

kemampuan generalisasi model terhadap karakteristik audio yang lebih beragam.

2) Mengembangkan metode ekstraksi fitur dan arsitektur model lain, seperti transformer atau *attention-based* model.
3) Mengembangkan sistem deteksi yang tidak hanya berfokus pada audio, tetapi juga mengintegrasikan informasi visual dari video (*audio-visual deepfake detection*) sehingga sistem dapat mendeteksi *deepfake* secara lebih komprehensif pada konten multimedia.
4) Mengembangkan dan mengevaluasi model pada audio dengan tingkat kebisingan (*noise*) yang tinggi, seperti adanya musik pada latar, suara lingkungan, atau gangguan lainnya, guna meningkatkan ketahanan sistem terhadap kondisi penggunaan di dunia nyata.
5) Melakukan evaluasi lebih lanjut menggunakan audio *deepfake* dengan durasi yang lebih panjang untuk mengetahui konsistensi dan stabilitas performa model dalam mendeteksi *deepfake* pada berbagai variasi durasi audio.
6) Mengembangkan sistem agar mampu melakukan deteksi audio *deepfake* secara real-time selama proses perekaman berlangsung, sehingga hasil deteksi dapat diperoleh secara langsung tanpa harus menunggu proses perekaman selesai terlebih dahulu.