# BAB 1
# PENDAHULUAN

## 1.1 Latar Belakang Masalah

Dalam era digital saat ini, perkembangan teknologi kecerdasan buatan semakin pesat, terutama dalam bidang pemrosesan gambar dan video. Teknologi ini memungkinkan komputer untuk menganalisis, memanipulasi, dan mengekstraksi data dari suatu gambar. Salah satu inovasi yang menarik perhatian dalam bidang tersebut adalah teknologi *deepfake*. *Deepfake* dapat diartikan sebagai rekayasa digital berbasis kecerdasan buatan yang memanipulasi wajah, ekspresi, atau suara dalam suatu gambar, video, dan rekaman audio sehingga menampilkan seseorang seolah-olah melakukan atau mengatakan sesuatu yang sebenarnya tidak pernah terjadi [1]. Terminologi *deepfake* berasal dari gabungan kata "*Deep Learning*" dan "*Fake*" yang pertama kali diperkenalkan oleh seorang pengguna anonim situs *online* Reddit pada akhir tahun 2017 [2]. Pengguna tersebut menggunakan metode *deep learning* untuk mengubah wajah seseorang dalam sebuah video pornografi dengan wajah orang lain dan menghasilkan video yang terlihat realistis.

Pada prinsipnya, *deepfake* dibuat menggunakan teknologi kecerdasan buatan *neural network*, terutama *autoencoder* dan GAN atau *generative adversarial network* [1]. *Autoencoder* bekerja dengan melakukan *encoding* pada gambar atau video ke dalam bentuk yang lebih sederhana dengan menyimpan informasi penting dan membuang detail yang tidak diperlukan, lalu melakukan *decoding* pada bentuk sederhana tersebut untuk mengembalikannya menjadi gambar atau video yang mirip dengan aslinya [3]. Di sisi lain, GAN terdiri dari dua *neural network*, yaitu *generator* dan *discriminator*. *Generator* berfungsi untuk menghasilkan data baru yang menyerupai data asli, sedangkan *discriminator* bertugas untuk membedakan antara gambar asli dan hasil manipulasi yang dilakukan oleh *generator*, sehingga keduanya saling berkompetisi dalam proses pelatihan untuk meningkatkan kualitas *deepfake* yang dihasilkan semakin realistis dan sulit untuk dibedakan dari gambar atau video asli.

Hasil realistis yang dihasilkan *deepfake* dari manipulasi media gambar, video, atau audio dapat memiliki banyak kegunaan yang bersifat positif dan menguntungkan. Hal ini meliputi berbagai hal, seperti menampilkan adegan dalam film di mana seorang aktor terlihat lebih muda atau tua, atau bahkan

seseorang yang sudah meninggal seperti Paul Walker pada film Fast & Furious 7 serta pembuatan karakter *video game* yang terlihat realistis. Namun, di samping manfaatnya, *deepfake* juga memiliki potensi penyalahgunaan yang dapat memiliki berbagai dampak negatif. Hal ini terutama disebabkan oleh tersedianya *tools* seperti Face2Face dan FaceSwap, yang memungkinkan masyarakat luas untuk menghasilkan *deepfake* dengan mudah [3]. Salah satu risiko yang timbul adalah penyebaran misinformasi dan disinformasi, terutama dengan menggunakan kemiripan tokoh-tokoh terkenal seperti politisi dan selebritas untuk memberikan citra negatif dalam persepsi publik [4].

Terdapat juga kasus di mana *deepfake* digunakan sebagai wadah penipuan dengan menggunakan wajah tokoh-tokoh publik, seperti penggunaan video *deepfake* dengan wajah Baim Wong yang digunakan untuk menipu para korban untuk mengirimkan uang kepada penipu [5]. Selain itu, *deepfake* juga marak disalahgunakan untuk membuat gambar atau video pornografi non-konsensual. Pada tahun 2019, 96 persen dari video *deepfake* yang terdapat di internet merupakan pornografi yang dibuat tanpa izin dari individu yang wajahnya digunakan [6]. Fenomena-fenomena tersebut menunjukkan betapa berbahaya apabila teknologi *deepfake* disalahgunakan, terutama dengan betapa sulitnya membedakan media *deepfake* dengan yang asli menggunakan mata telanjang [7].

Berdasarkan permasalahan yang telah dipaparkan sebelumnya, diperlukanlah suatu teknologi untuk mendeteksi *deepfake*. Salah satu pendekatan yang dapat digunakan untuk mendeteksi *deepfake* adalah dengan menggunakan teknik *deep learning*, khususnya Convolutional Neural Network atau CNN. CNN adalah suatu metode *supervised deep learning* yang dapat mengekstraksi fitur spasial penting untuk kemudian memprediksi dan mengklasifikasikan suatu gambar [8]. Metode ini sesuai untuk tugas-tugas yang berkaitan dengan analisis data visual, seperti klasifikasi gambar dan deteksi objek [8].

Dalam konteks deteksi *deepfake*, CNN berperan dalam menemukan pola visual yang mungkin tidak dapat terlihat oleh mata manusia, seperti ketidaksesuaian tekstur, pencahayaan yang tidak alami, serta distorsi pada wajah. Namun, karena suatu video memiliki komponen spasial dan temporal, di mana informasi temporal mencerminkan hubungan antar *frame*, pendekatan CNN saja tidak cukup untuk menangkap pola pergerakan serta perubahan ekspresi wajah dalam suatu video *deepfake*. Untuk mengatasi permasalahan ini, metode RNN dapat digunakan untuk mengekstraksi data temporal pada video *deepfake* [8].

Berbagai penelitian terdahulu telah menggunakan berbagai varian metode

CNN dan RNN untuk mendeteksi video *deepfake*. Salah satu contohnya adalah penelitian oleh Antad et al. yang menggabungkan CNN dan LSTM, salah satu bentuk RNN, dan menggunakan *dataset Deepfake Detection Challenge* (DFDC) [9]. Model tersebut mencapai akurasi sebesar 92,11% pada video dengan 100 *frame*, namun menurun secara signifikan menjadi 84,66% pada video dengan hanya 10 *frame*. Hal ini menunjukkan pentingnya jumlah *frame* dalam mendeteksi manipulasi temporal pada video.

Permasalahan lain yang juga sering muncul adalah lemahnya kemampuan generalisasi model deteksi *deepfake* terhadap variasi teknik manipulasi. Brodarič, Štruc, dan Peer melakukan penelitian menggunakan Xception, sebuah *pre-trained model* CNN buatan Google, untuk mendeteksi video *deepfake* [10]. *Pre-trained model* merupakan model yang telah dilatih sebelumnya pada *dataset* besar seperti ImageNet dan dapat digunakan kembali melalui teknik *transfer learning*, sehingga menghemat waktu dan sumber daya komputasi [11]. Hasil penelitian tersebut menunjukkan bahwa model cenderung mengalami *overfitting* terhadap metode pembuatan tertentu dan mengalami penurunan akurasi yang signifikan saat diuji pada metode manipulasi lain, menandakan kelemahan generalisasi *cross-method*.

Berdasarkan permasalahan tersebut serta temuan-temuan dari penelitian terdahulu, penelitian ini bertujuan untuk merancang dan menguji enam model deteksi video *deepfake* menggunakan arsitektur gabungan antara *pre-trained model* CNN (InceptionV3) dan RNN (LSTM). InceptionV3 dipilih karena performanya yang baik dalam klasifikasi gambar, ditunjukkan dengan gelar *runner-up* pada kompetisi *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) 2015 [12]. Sementara itu, *Long-Short Term Memory* (LSTM) digunakan bersama InceptionV3 untuk menangkap pola temporal dalam urutan *frame* video, sehingga model dapat mengenali inkonsistensi gerakan wajah yang khas pada video *deepfake*.

Lima model akan dilatih secara terpisah menggunakan masing-masing dari lima metode pembuatan *deepfake* yang tersedia dalam *dataset* FaceForensics++ (yaitu Deepfakes, Face2Face, FaceSwap, NeuralTextures, dan FaceShifter), sedangkan satu model tambahan dilatih dengan data gabungan dari kelima metode tersebut. Pendekatan pelatihan ini memungkinkan evaluasi performa dan generalisasi model terhadap berbagai teknik manipulasi wajah, baik secara spesifik maupun lintas-metode, melalui *cross-subset testing*.

## 1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut.

1. Bagaimana implementasi model berbasis InceptionV3 dan LSTM dapat digunakan untuk mendeteksi video *deepfake*?
2. Bagaimana kinerja dan kemampuan generalisasi model tersebut dalam mendeteksi video *deepfake* yang dibuat dengan berbagai metode pembuatan, berdasarkan metrik evaluasi akurasi, presisi, *recall*, *F1-score*, dan AUC-ROC?

## 1.3 Batasan Permasalahan

Penelitian ini memiliki beberapa batasan sebagai berikut.

1. *Dataset* yang digunakan adalah FaceForensics++ dengan lima metode pembuatan *deepfake*: Deepfakes, Face2Face, FaceSwap, NeuralTextures, dan FaceShifter.
2. Hanya digunakan video dengan tingkat kompresi c23.
3. Penelitian ini mengimplementasikan enam model deteksi: lima model dilatih secara spesifik pada masing-masing metode pembuatan *deepfake*, serta satu model tambahan yang dilatih menggunakan data gabungan dari kelima metode tersebut.
4. Evaluasi dilakukan menggunakan skenario *cross-subset testing*, yaitu dengan menguji model baik pada *subset* data dari metode pembuatan yang sama maupun berbeda dengan data pelatihannya.
5. Metrik evaluasi yang digunakan mencakup akurasi, presisi, *recall*, *F1-score*, dan AUC-ROC.

## 1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut.

1. Mengimplementasikan enam model deteksi video *deepfake* berbasis arsitektur InceptionV3 dan LSTM.

2. Mengevaluasi kinerja dan kemampuan generalisasi model dalam mendeteksi video *deepfake* dari berbagai metode pembuatan, berdasarkan metrik akurasi, presisi, *recall*, *F1-score*, dan AUC-ROC.

## 1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut.

1. Menyumbangkan pengetahuan baru mengenai penggunaan model InceptionV3 dan LSTM dalam deteksi video *deepfake*, khususnya dalam konteks generalisasi lintas metode manipulasi.
2. Menjadi acuan dan referensi bagi penelitian selanjutnya untuk mengembangkan sistem deteksi video *deepfake* yang lebih akurat dan efisien.

## 1.6 Sistematika Penulisan

Sistematika penulisan laporan ini adalah sebagai berikut.

- Bab 1 PENDAHULUAN
  Menjelaskan latar belakang, rumusan masalah, batasan masalah, tujuan, dan manfaat penelitian.
- Bab 2 LANDASAN TEORI
  Menyajikan landasan teori yang berkaitan dengan *deepfake*, CNN, LSTM, dan teknik evaluasi model.
- Bab 3 METODOLOGI PENELITIAN
  Membahas metodologi penelitian, termasuk arsitektur model, skenario pelatihan, dan pembagian *dataset*.
- Bab 4 HASIL DAN DISKUSI
  Menampilkan hasil pelatihan model, analisis performa pada masing-masing subset, dan pembahasan hasil.
- Bab 5 KESIMPULAN DAN SARAN
  Menyampaikan kesimpulan dari penelitian serta saran untuk penelitian selanjutnya.