

# PENDAHULUAN

## 1.1 Latar Belakang

Deepfake merupakan teknologi berbasis *Artificial Intelligence* (AI) yang memanfaatkan teknik *deep learning* untuk menghasilkan citra maupun video hasil manipulasi digital sehingga objek yang ditampilkan tampak nyata meskipun sebenarnya telah dimodifikasi atau dihasilkan secara sintetis, dan seiring perkembangan teknologi AI generatif, kualitas citra *deepfake* menjadi semakin realistis sehingga semakin sulit dibedakan dari citra asli oleh pengamatan manusia [1]. Kondisi tersebut menimbulkan berbagai tantangan, terutama pada keamanan informasi digital, karena *deepfake* dapat dimanfaatkan untuk melakukan penyebaran hoaks visual, pencurian identitas, manipulasi informasi, hingga penipuan digital yang merugikan individu maupun organisasi [2]. Media sosial menjadi salah satu sarana utama penyebaran konten *deepfake* karena informasi dapat tersebar dengan cepat tanpa melalui proses verifikasi yang memadai, sehingga meningkatkan risiko penyebaran berita palsu dan ujaran kebencian di masyarakat [3][4]. Oleh karena itu, diperlukan suatu sistem yang mampu mendeteksi citra *deepfake* secara otomatis agar penyebaran hoaks visual dapat diminimalkan.

Salah satu metode *deep learning* yang banyak digunakan dalam klasifikasi citra adalah *Convolutional Neural Network* (CNN) karena mampu melakukan ekstraksi fitur secara otomatis melalui lapisan *konvolusi*, *pooling*, dan *fully connected*, sehingga tidak memerlukan proses ekstraksi fitur secara manual seperti pada beberapa metode klasifikasi konvensional [5]. Kemampuan tersebut memungkinkan CNN mempelajari karakteristik visual, seperti tekstur, bentuk, pola, warna, serta detail citra secara langsung dari data pelatihan sehingga memberikan performa yang baik dalam berbagai permasalahan klasifikasi citra [6]. Berbagai penelitian menunjukkan bahwa CNN berhasil diterapkan pada berbagai bidang pengolahan citra, seperti klasifikasi nominal uang [7], klasifikasi jenis sampah [8], deteksi warna objek [9], serta identifikasi wajah asli dan citra *deepfake* [10]. Hasil penelitian tersebut menunjukkan bahwa CNN mampu menghasilkan tingkat akurasi yang tinggi karena dapat mengenali pola visual yang kompleks dan melakukan proses pembelajaran fitur secara otomatis tanpa bergantung pada *feature engineering* yang rumit [11]. Berdasarkan keunggulan tersebut, metode *Convolutional Neural Network* (CNN) dipilih pada penelitian ini untuk membangun model deteksi citra *deepfake* serta mengevaluasi kemampuannya dalam membedakan citra asli dan citra hasil manipulasi menggunakan *dataset* yang memiliki karakteristik visual beragam berdasarkan variasi visual

yang terdapat pada *dataset* GRAVEX-200K, seperti perbedaan wajah, pose, ekspresi, pencahayaan, dan latar belakang.

Berbagai penelitian telah dilakukan untuk mendeteksi citra *deepfake* menggunakan pendekatan berbasis *Convolutional Neural Network* (CNN). Pada Penelitian [12] menunjukkan bahwa CNN mampu mengekstraksi karakteristik visual citra hasil manipulasi secara otomatis melalui proses pembelajaran fitur (*feature learning*), sehingga menghasilkan performa deteksi yang tinggi pada berbagai *dataset deepfake*. Kemampuan tersebut menjadikan CNN sebagai salah satu metode yang efektif dalam membedakan citra asli dan citra hasil manipulasi. Pada penelitian [12] mengembangkan sistem deteksi *deepfake* berbasis CNN dan menunjukkan bahwa model mampu mencapai nilai *Area Under Curve* (AUC) yang tinggi serta memiliki kemampuan generalisasi yang baik pada pengujian lintas dataset. Hasil penelitian tersebut menunjukkan bahwa CNN dapat mempelajari pola manipulasi digital yang kompleks sehingga tetap mampu mempertahankan performa klasifikasi meskipun data uji berasal dari sumber yang berbeda. Pada penelitian [12] peneliti melakukan evaluasi terhadap beberapa pendekatan deteksi *deepfake* berbasis CNN pada berbagai dataset publik. Hasil penelitian menunjukkan bahwa CNN masih menjadi salah satu metode yang kompetitif dalam mendeteksi citra hasil manipulasi wajah karena mampu mengenali karakteristik spasial yang tidak mudah diamati oleh penglihatan manusia. Selain itu, penelitian tersebut juga menunjukkan bahwa peningkatan kualitas citra *deepfake* menuntut model CNN untuk memiliki kemampuan ekstraksi fitur yang semakin baik agar akurasi deteksi tetap terjaga. Penelitian lain pada [13] menyatakan bahwa sebagian besar metode deteksi *deepfake* modern masih menggunakan CNN sebagai arsitektur dasar karena kemampuannya dalam melakukan ekstraksi fitur secara otomatis tanpa memerlukan proses *feature engineering* secara manual. Berdasarkan berbagai penelitian tersebut, CNN terbukti mampu memberikan performa klasifikasi yang baik sehingga dipilih sebagai metode utama dalam penelitian ini untuk mendeteksi citra *deepfake*.

Berdasarkan uraian tersebut, penelitian ini menerapkan metode *Convolutional Neural Network* (CNN) untuk membangun model deteksi citra *deepfake* serta mengevaluasi kinerjanya dalam membedakan citra asli dengan citra hasil manipulasi. Hasil penelitian diharapkan dapat memberikan kontribusi dalam pengembangan sistem deteksi *deepfake* yang mampu membantu mengurangi penyebaran hoaks visual di media sosial serta meningkatkan keamanan informasi digital. Atas dasar tersebut, penelitian ini diajukan dengan judul "Deteksi *Deepfake* Citra Berbasis CNN untuk Pencegahan Hoaks Visual di Media Sosial."

## 1.2 Rumusan Masalah

Bagaimana menerapkan metode *Convolutional Neural Network* (CNN) untuk membangun sistem deteksi citra *deepfake* menggunakan *dataset* yang memiliki karakteristik visual beragam serta mengevaluasi kinerja model berdasarkan nilai *accuracy*, *precision*, *recall*, *F1-score*, AUC, *loss*, dan *confusion matrix* dalam membedakan citra asli dan citra hasil manipulasi.

## 1.3 Tujuan

Tujuan penelitian ini meliputi:

1. Membangun dan melatih model *Convolutional Neural Network* (CNN) untuk membedakan citra asli (*real*) dan citra hasil manipulasi (*deepfake*) menggunakan dataset “GRAVEX-200K (*Real and Fake Face Detection Dataset*)”.
2. Mengevaluasi kinerja model CNN terbaik dalam mendeteksi citra *deepfake* berdasarkan nilai *accuracy*, *precision*, *recall*, *F1-score*, AUC, *loss*, dan *confusion matrix*.

## 1.4 Manfaat

Penelitian ini diharapkan dapat memberikan beberapa manfaat sebagai berikut:

1. Penelitian ini dapat menjadi referensi tambahan dalam pengembangan studi tentang penerapan *deep learning* untuk deteksi manipulasi citra digital.
2. Penelitian ini diharapkan menjadi dasar bagi pengembangan sistem deteksi otomatis berbasis CNN yang dapat diterapkan pada skala lebih luas, seperti verifikasi citra di media daring dan keamanan informasi digital.

## 1.5 Ruang Lingkup

Ruang lingkup dalam penelitian ini dibatasi pada beberapa hal berikut:

1. Penelitian difokuskan pada pendeteksian citra wajah pada gambar dan video dan tidak mencakup analisis dalam bentuk audio.
2. Dataset yang digunakan adalah GRAVEX-200K (*Real and Fake Face Detection Dataset*) yang diperoleh dari Kaggle, yang terdiri dari 200.000 gambar, dengan rincian 100.000 citra asli (*real*) dan 100.000 citra manipulasi (*fake/deepfake*). Citra *real* merupakan gambar wajah asli yang berasal dari individu nyata dan tidak mengalami proses manipulasi atau rekayasa digital yang mengubah identitas maupun karakteristik wajah. Sementara itu, citra *fake/deepfake* merupakan gambar wajah yang telah dibuat atau dimanipulasi menggunakan teknologi kecerdasan buatan sehingga menyerupai

wajah atau identitas seseorang. Setiap gambar memiliki ukuran yang sama, yaitu  $256 \times 256$  piksel dengan format warna RGB.

3. Data pada penelitian ini dibagi menjadi data latih (*training data*), data validasi (*validation data*) dan data uji (*testing data*) dengan perbandingan 70:15:15.
4. Proses implementasi dilakukan menggunakan bahasa pemrograman *Python* dengan bantuan pustaka seperti *TensorFlow* dan *Keras*.
5. Evaluasi kinerja model dilakukan menggunakan metrik akurasi, loss, dan confusion matrix, yang meliputi nilai *True Positive (TP)*, *True Negative (TN)*, *False Positive (FP)*, dan *False Negative (FN)* untuk mengukur performa klasifikasi secara lebih detail.
6. Penelitian ini tidak membahas proses pembuatan *deepfake*, melainkan hanya berfokus pada proses deteksi keaslian citra digital hasil manipulasi.

