

BAB I

PENDAHULUAN

1.1 Latar Belakang

Data Mining merupakan sebuah proses analitis yang bertujuan untuk menggali informasi berharga, tersembunyi, dan konsisten dari kumpulan data dalam jumlah besar (Nurdin, 2017). Salah satu teknik yang umum digunakan dalam *data mining* adalah klasifikasi. Secara sederhana, *Data Mining* yang juga dikenal dengan istilah *knowledge discovery in databases* (KDD) merupakan rangkaian tahapan yang dirancang untuk menemukan pola-pola penting atau menarik dari data berskala besar, yang umumnya sulit diidentifikasi secara manual.

Data Mining merupakan bagian integral dari proses penemuan pengetahuan dalam basis data (*Knowledge Discovery in Databases/KDD*). Proses ini terdiri dari beberapa tahapan yang saling berurutan. *Data Mining* tidak dapat dipisahkan dari sejarah penggunaan istilah seperti "*Data Mining*" itu sendiri. Pada era 1960-an, ketika komputer mulai banyak dimanfaatkan untuk analisis data, ditemukan bahwa jika data dianalisis cukup lama, akan selalu ada model yang tampak cocok dengan data tersebut (Abdullah, 2021). *Data Mining* sering juga disebut dengan istilah *knowledge-in-database* atau *recognition* (Rahayu *et al.*, 2023). Istilah KDD merujuk pada proses penemuan pengetahuan dari data, karena inti dari *Data Mining* adalah memanfaatkan data yang ada dalam basis data untuk menghasilkan informasi baru yang bermanfaat. Sementara itu, *pattern recognition* atau pengenalan pola bertujuan untuk mengekstraksi pengetahuan dari kumpulan data yang sedang dianalisis (Mughnyanti & Ginting, 2023).

Klasifikasi merupakan proses untuk mengevaluasi suatu data objek dan menempatkannya ke dalam salah satu kelas yang telah ditentukan sebelumnya. Dalam proses ini, terdapat dua langkah utama yang dilakukan, pertama, membangun *prototype* sebagai representasi yang disimpan dalam memori, dan kedua, menggunakan model tersebut untuk mengenali data objek baru guna menentukan kelasnya berdasarkan model yang telah dibentuk. Secara umum, klasifikasi adalah tahapan untuk menemukan sebuah model atau fungsi yang dapat

menggambarkan sekaligus membedakan berbagai kelas data atau konsep, dengan tujuan agar model tersebut dapat digunakan untuk memprediksi kelas dari data yang belum diketahui. Model klasifikasi sering kali dianalogikan seperti kotak hitam, karena model ini mampu menerima *input*, memprosesnya secara internal, dan kemudian menghasilkan *output* berupa prediksi atau keputusan berdasarkan hasil pemrosesan tersebut.

Naive Bayes adalah algoritma klasifikasi yang didasarkan pada prinsip probabilitas dan statistik, yang dikembangkan dari teori yang dikemukakan oleh ilmuwan asal Inggris, *Thomas Bayes*. Metode klasifikasi ini termasuk dalam kategori klasifikasi statistik, yang bertujuan untuk memperkirakan probabilitas suatu data termasuk ke dalam kelas tertentu, serta menghitung kemungkinan terjadinya suatu atribut berdasarkan informasi yang tersedia (Pratama *et al.*, 2022).

Algoritma ini memanfaatkan pengalaman masa lalu untuk memprediksi kemungkinan yang akan terjadi di masa depan, sehingga dikenal dengan nama *teorema Bayes*. Menurut buku yang ditulis oleh Muhammad Arhani dan Muhammad Nasir pada tahun 2020, algoritma *Naive Bayes* dapat diterapkan pada masalah klasifikasi biner maupun multikelas. Keunggulan algoritma ini terletak pada kemampuannya dalam membangun dan mengevaluasi model secara cepat, serta skalabilitasnya yang linier terhadap jumlah fitur prediktor dan jumlah data yang dianalisis.

Karena pengenalan gambar visualnya yang unggul dan akurasinya yang tinggi, metode *Convolutional Neural Network* (CNN) digunakan secara luas dalam pemrosesan gambar (Yuliany *et al.*, 2022). *Convolutional Neural Network* (CNN) merupakan salah satu algoritma dalam *deep learning* yang dirancang untuk memproses data berupa gambar. CNN mampu mengenali objek-objek yang terdapat dalam gambar dan memanfaatkannya sebagai dasar untuk mengklasifikasikan serta membedakan setiap gambar yang dianalisis. (Iswantoro & Handayani UN, 2022).

Deep learning berbeda dengan metode pembelajaran mesin tradisional karena melakukan representasi data secara otomatis, seperti gambar, video, atau teks, tanpa menggunakan aturan yang dikodekan atau keahlian domain manusia

(Bartlett *et al.*, 2021). *Deep learning* juga dikenal sebagai metode pembelajaran yang memanfaatkan jaringan saraf tiruan (*neural network*) berlapis-lapis atau banyak *layer* (*multi layer*) (Roberts *et al.*, 2022).

Convolutional Neural Network (CNN) merupakan salah satu pendekatan *deep learning* yang paling umum diterapkan dalam pengklasifikasian gambar. Algoritma ini dipilih karena kemampuannya yang unggul dalam meniru mekanisme Algoritma ini mampu mengolah informasi visual secara efisien dengan cara yang menyerupai proses pengenalan gambar di korteks visual manusia. menginterpretasikan visual, sehingga memungkinkan pengenalan dan analisis citra yang lebih akurat dan efisien (Nugroho *et al.*, 2020).

Algoritma CNN dirancang khusus untuk mengenali dan mengklasifikasikan gambar dengan cara mengekstraksi informasi penting dari gambar tersebut. Proses ini melibatkan beberapa lapisan bertingkat, dimana setiap lapisan memiliki fungsi spesifik dalam mengolah data visual. Setelah melalui tahapan ini, algoritma akan menghasilkan skor klasifikasi yang digunakan untuk menentukan kategori atau klasifikasi akhir dari gambar tersebut (Oktaviani *et al.*, 2024).

Adapun beberapa penelitian yang menggunakan CNN dalam klasifikasi citra diantaranya yaitu penelitian yang dipimpin oleh Mardiyah, (2020) Dengan judul “Penerapan *Deep Learning* untuk Klasifikasi Gambar Menggunakan Algoritma *Convolutional Neural Network* (CNN) pada Citra Lahan Perkebunan dan Persawahan”. Dataset yang dimanfaatkan mencakup dari 100 gambar. Data ini dipisahkan dengan proporsi 80% dialokasikan untuk proses *training*, sementara sisanya 20% untuk validasi, karena pembagian ini dianggap memberikan tingkat akurasi validasi tertinggi sebesar 75%, dengan akurasi pelatihan mencapai 96,25%.

Pada tahun 2020, Febian Fitra Maulana dan Naim Rochmawati melakukan sebuah penelitian dengan judul “Klasifikasi Citra Buah Menggunakan *Convolutional Neural Network*”. Penelitian ini menggunakan metode *deep learning*, yaitu *Convolutional Neural Network* (CNN) dalam pemrosesan citra. Dataset yang digunakan terdiri dari berbagai gambar buah-buahan yang berasal dari kumpulan data *Fruit-360*, di mana penelitian ini memilih untuk fokus pada 15 kelas buah dari total 111 kelas yang ada dalam dataset tersebut. Proses pembelajaran yang

dilakukan berhasil menghasilkan sebuah model CNN yang sangat efektif, dengan mencapai tingkat akurasi sempurna sebesar 100% dan nilai *loss* yang minimal, hanya sebesar 0,012. Keberhasilan ini menunjukkan bahwa model CNN mampu secara akurat memahami dan menginterpretasikan fitur-fitur penting dari gambar-gambar buah yang rumit dan bervariasi. Selanjutnya, saat model CNN ini diuji dengan menggunakan 45 sampel gambar buah yang berbeda, penelitian menemukan bahwa tingkat akurasi yang dicapai mencapai angka yang sangat memuaskan, yaitu sebesar 91,42%. Hal ini mengindikasikan bahwa model yang dibangun tidak hanya efektif dalam mempelajari data selama proses *training*, tetapi juga memiliki kemampuan generalisasi yang baik terhadap data baru pengetahuannya dengan baik saat dihadapkan pada data uji yang baru (Maulana & Rochmawati, 2020).

Riset yang telah dilakukan oleh (Khoiruddin *et al.*, 2022), berjudul "Klasifikasi penyakit pada daun padi dengan menggunakan algoritma *Convolutional Neural Network* (CNN)" bertujuan untuk memberikan solusi kepada petani dalam menangani gangguan yang ditimbulkan oleh penyakit daun pada tanaman padi. Penelitian ini berfokus pada pengelompokan beberapa jenis penyakit daun padi, yakni bercak daun bakteri, penyakit blas, dan tungro. Sebanyak 6000 data citra digunakan dalam penelitian ini, yang kemudian dibagi dengan proporsi 80% untuk pelatihan, 10% untuk validasi, dan 10% sisanya untuk pengujian. Hasil penelitian menunjukkan bahwa akurasi yang dicapai bervariasi tergantung pada jumlah *epoch* yang digunakan, yaitu 25, 50, 75, dan 100 *epoch*. Akurasi pelatihan terbaik tercatat sebesar 98% pada 100 *epoch*, sementara pengujian dengan matriks campuran memberikan rata-rata akurasi sebesar 98%. Hal ini menunjukkan bahwa algoritma CNN sangat akurat dalam mengklasifikasikan penyakit pada daun padi.

Penelitian oleh Putra *et al.*, (2022) bertujuan untuk memanfaatkan *Convolutional Neural Network* (CNN) dengan *desain Resnet 50* dalam mengatur penyakit daun pada tanaman jagung, dengan memanfaatkan strategi pengembangan seperti Adam (*Adaptive Moment Estimation*), Nadam dan SGD. Dataset yang dipakai mencakup 4225 gambar, yang kemudian dipisahkan menjadi 3380 gambar untuk interaksi persiapan dan 845 gambar untuk sistem pengujian,

dengan ukuran gambar masing-masing adalah 224×224 *pixel*. Berdasarkan hasil eksperimen, arsitektur *Resnet 50* yang dioptimalkan menggunakan metode Adam mencapai akurasi tertinggi, yaitu 98,4%.

Berdasarkan hasil penelitian sebelumnya, metode CNN terbukti efektif dalam menghasilkan performa sistem yang sangat baik dan menjanjikan. Oleh karena itu, penelitian ini memilih untuk menerapkan CNN dalam mengklasifikasi penyakit pada tanaman timun suri. Studi ini akan mencakup proses pelatihan dengan berbagai variasi dalam pembagian data untuk pelatihan dan pengujian, serta penerapan beberapa parameter untuk mengatasi masalah *overfitting* yang mungkin timbul. Teknik augmentasi data juga diterapkan untuk meningkatkan performansi sistem.

1.2 Rumusan Masalah

Mengacu pada penjelasan latar belakang yang telah disampaikan, pertanyaan penelitian yang akan dibahas dalam studi ini adalah sebagai berikut:

1. Bagaimana mengembangkan model CNN untuk mengklasifikasi penyakit pada daun timun suri?
2. Seberapa efektif model CNN dalam mengenali kondisi daun timun suri?
3. Bagaimana pengaruh argumentasi data terhadap peningkatan akurasi model CNN?

1.3 Tujuan Penelitian

Berdasarkan latar belakang dan rumusan masalah di atas maka tujuan penelitian ini adalah sebagai berikut:

1. Untuk mengembangkan model CNN untuk mengklasifikasi penyakit pada daun timun suri.
2. Untuk mengetahui seberapa efektif model CNN dalam mengenali kondisi daun timun suri.
3. Untuk mengetahui pengaruh argumentasi data terhadap peningkatan akurasi model CNN.

1.4 Manfaat Penelitian

Diharapkan bahwa penelitian ini dapat memberikan kontribusi yang bermanfaat. Beberapa manfaat yang diharapkan dari penelitian ini adalah:

1. Membantu petani mengenali penyakit pada timun suri dengan lebih cepat dan akurat tanpa harus bergantung pada ahli atau metode manual.
2. Menjadi referensi untuk pengembangan lebih lanjut dalam klasifikasi penyakit pada tanaman lain menggunakan *Deep Learning* yang lebih kompleks.
3. Mendukung upaya peningkatan produksi pertanian melalui teknologi berbasis AI untuk pengelolaan hama dan penyakit.

1.5 Ruang Lingkup dan Batasan Penelitian

Agar penelitian ini fokus dan Untuk mencapai tujuan yang telah ditetapkan, penelitian ini memiliki batasan masalah yang ditentukan sebagai berikut:

1. Dataset yang diterapkan dalam studi ini mencakup foto asli yang diperoleh langsung dari kebun timun suri di Desa Matang Sagoe serta gambar tambahan yang diambil dari *platform* daring seperti *iStock*, *Kaggle*, dan *Pinterest*.
2. Adapun jenis-jenis kondisi timun suri yang digunakan di sini yaitu: Antraknosa, Hawar Batang Bergetah, Jamur Berbulu Halus (Bulai), Layu Bakteri dan Daun Segar.
3. Dataset yang digunakan sebanyak 60 dataset untuk masing-masing gejala. Total data yang digunakan yaitu 300 dataset dengan rincian: 60 data Antraknosa, 60 data Hawar Batang Bergetah, 60 data Jamur Berbulu Halus (Bulai), 60 data layu bakteri dan 60 data daun segar.
4. Metode yang diterapkan untuk klasifikasi dalam penelitian ini adalah *Convolutional Neural Network (CNN)* dengan arsitektur tertentu, dan tidak membahas metode klasifikasi lainnya.
5. Pengembangan dan pengujian dilakukan menggunakan *Google Colaboratory* dengan spesifikasi perangkat keras tertentu dan *library TensorFlow*.