BAB II
TINJAUAN PUSTAKA

2.1 Tinjauan Pustaka

Sejumlah penelitian telah dilakukan mengenai klasifikasi tanaman terong berdasarkan citra daun menggunakan metode *K-Nearest Neighbors*. Salah satu penelitian yang relevan dilakukan oleh Eko Hari Rachmawanto dan Heru Pramono Hadi berjudul "Optimasi Ekstraksi Fitur Pada *K-NN* Dalam Klasifikasi Penyakit Daun Jagung". Penelitian ini bertujuan untuk mengidentifikasi daun yang tidak sehat dengan mengekstraksi fitur dan warna pada citra untuk mendeteksi penyakit daun tanaman jagung seperti hawar, bercak, dan karat. Proses klasifikasi citra dilakukan dengan mengakuisisi citra menjadi data latih dan uji, kemudian menghitung nilai fitur ekstraksi warna dan ekstraksi fitur. *GLCM (Gray-Level Cooccurrence Matrix*) digunakan sebagai ekstraksi fitur dan *HSV* digunakan sebagai ekstraksi warna. *K-NN* (*K-Nearest Neighbors*) dengan jarak *Euclidean* digunakan untuk klasifikasi. Dari 160 data citra latih dan 40 citra uji yang menggunakan algoritma *K-NN-HSV-GLCM*, hasilnya menunjukkan akurasi terbaik sebesar 85% dengan nilai k adalah 3 dan jarak piksel 1, sedangkan akurasi terendah sebesar 70% dengan nilai k adalah 3 dan jarak piksel 3.(Rachmawanto & Hadi, 2021).

Agus Yuliani, Ause Labellapansa, dan Ana Yulianti melakukan penelitian dengan judul "Klasifikasi Citra Daun Kelapa Sawit Yang Terkena Dampak Hama Menggunakan Metode *K-Nearest Neighbors*". Dalam proses deteksinya, langkah-langkah yang digunakan meliputi *preprocessing*,

segmentasi, ekstraksi fitur *zoning,* dan klasifikasi dengan metode *K-Nearest Neighbors* untuk mengklasifikasikan jenis hama yang terjadi. *Preprocessing* mencakup akuisisi citra, *resize,* konversi citra ke *grayscale,* dan segmentasi *threshold.* Kemudian dilakukan ekstraksi fitur *zoning* dengan membagi citra menjadi 4 bagian, dan selanjutnya dilakukan klasifikasi jenis hama yang menyerang daun kelapa sawit. Hasilnya, pengujian teknik *zoning* menunjukkan akurasi pendeteksian hama *limacodidae* sebesar 55% dan hama *psychidae* sebesar 72.5%. Hal ini menunjukkan kemampuan sistem dalam mendeteksi jenis hama. Oleh karena itu, untuk meningkatkan kinerja sistem, perbaikan pada proses *preprocessing* citra dapat dilakukan berdasarkan hasil yang diperoleh.(Yuliani, Labellapansa, 2019).

Rifqi Hakim Ariesdianto, Zilvanhisna Emka Fitri, Abdul Madjid, dan Arizal Mujibtamala Nanda Imron melakukan penelitian dengan judul "Identifikasi Penyakit Daun Jeruk Siam Menggunakan *K-Nearest Neighbors*". Dua penyakit yang umum menyerang daun jeruk siam adalah kanker yang disebabkan oleh patogen *Xanthomonas axonopodis pv.* citri dan ulat peliang. Pengamatan penyakit daun jeruk siam selama ini dilakukan secara manual dengan mata manusia, yang membuat penentuan kondisi daun bersifat subjektif. Untuk mengatasi hal tersebut, dibuatlah sistem identifikasi otomatis menggunakan teknik *computer vision* untuk membedakan daun jeruk siam yang sehat dan yang terinfeksi penyakit. Langkah-langkah penelitian meliputi pengumpulan citra daun jeruk, konversi warna, ekstraksi fitur warna dan tekstur, serta klasifikasi menggunakan metode *K-Nearest Neighbors (K-NN).*

Parameter fitur yang digunakan mencakup fitur warna *GB* dan fitur tekstur seperti *ASM, entropi*, dan kontras. Melalui metode *K-NN*, sistem mampu mengklasifikasi dan mengidentifikasi penyakit daun jeruk siam dengan akurasi sebesar 70% dengan variasi nilai K = 21.(Ariesdianto et al., 2021).

Eva Agustina Ompusunggu, Dian Eka Ratnawati, dan Lailil Muflikhah melakukan penelitian yang berjudul "Identifikasi Penyakit Tanaman Jarak Pagar Menggunakan Metode *Fuzzy K-Nearest Neighbor (FK-NN)*". Sebagian masyarakat bergantung pada tanaman jarak pagar sebagai sumber penghidupan utama mereka. Namun, kualitas tanaman jarak pagar mengalami penurunan karena berbagai penyakit yang menyerangnya. Keterbatasan pengetahuan tentang penyakit tanaman jarak pagar dan kurangnya informasi mengenai cara penanggulangannya menjadi faktor utama. Oleh karena itu, diperlukan sistem untuk memudahkan diagnosis penyakit pada tanaman jarak pagar. Dalam penelitian ini, metode *k-nearest neighbor* dan *fuzzy* digunakan untuk membantu dalam proses diagnosis. Langkah awal metode ini melibatkan penggunaan data latih yang berisi gejala penyakit. Setelah itu, klasifikasi dilakukan menggunakan *k-nearest neighbor* dan *fuzzy*. Sehingga, sistem ini menghasilkan diagnosis penyakit pada tanaman jarak pagar dari sembilan penyakit yang umumnya terjadi. Pengujian penelitian ini menunjukkan tingkat akurasi tertinggi sebesar 80%.(Eva Agustina Ompusunggu, Dian Eka Ratnawati, 2018).

2.2 Klasifikasi

Klasifikasi merupakan tahap pengelompokan objek ke dalam kategori tertentu berdasarkan fitur-fiturnya. Algoritma klasifikasi adalah suatu fungsi yang mengaitkan objek ke label kelas tertentu. Algoritma klasifikasi, atau disebut juga *classifier,* berguna untuk melakukan prediksi. Tujuan dari klasifikasi adalah menentukan kategori objek yang belum memiliki label kelasnya, berdasarkan ciri-ciri yang dimiliki objek tersebut, sehingga dapat dikelompokkan ke dalam label kelas yang sudah ditetapkan sebelumnya. Dalam tugas klasifikasi, terdapat dua proses utama yang dilakukan: pertama, membangun *classifier* atau model untuk disimpan dalam memori, dan kedua, melakukan pengenalan. Klasifikasi objek untuk memprediksi kategorinya dilakukan berdasarkan model yang telah disimpan dalam memori tersebut. Model dapat berupa aturan, pohon keputusan, jaringan saraf tiruan, atau formula matematika.(Rahayu Marlis et al., 2021).

2.3 Tanaman Terong

Terong *(Solanum melongena L.)* banyak ditanam di berbagai daerah di Indonesia. Tanaman terong memiliki kemampuan untuk berproduksi hingga dua tahun dan memiliki tingkat produktivitas yang tinggi. Karena alasan tersebut, terong menjadi salah satu jenis sayuran yang menawarkan prospek yang cerah. Stabilitas harga terong yang relatif konstan dapat menjadi pertimbangan untuk melakukan penanaman secara luas.(Oktaviani, 2020).

Terong memiliki popularitas yang signifikan di kalangan masyarakat Indonesia, yang mengenalnya dengan luas. Berbagai olahan dan konsumsi mentah dari terong telah menjadi bagian dari kebiasaan masyarakat dalam konsumsinya. Kandungan gizi yang terdapat dalam terong dimanfaatkan untuk memenuhi kebutuhan gizi. Tanaman terong dapat tumbuh dengan baik di berbagai ketinggian, mulai dari dataran rendah hingga mencapai 1000 meter di atas permukaan laut. Tanaman terong membutuhkan suhu lingkungan yang berkisar antara 22-30^{o} C, dengan tingkat keasaman tanah ideal berada pada kisaran 5-6.(Musthafa, 2022).

2.4 Citra Digital

Representasi digital adalah penafsiran gambar ke dalam format digital dengan dua dimensi, yang terdiri dari sumbu horizontal dan vertikal serta nilai intensitas atau keabuan. Setiap nilai ini ditetapkan dalam rentang tertentu yang mencerminkan citra yang ditampilkan. Meskipun citra aslinya mungkin memiliki tiga dimensi, namun dalam pengolahan komputer, citra tersebut direduksi menjadi titik-titik data diskrit dalam dua dimensi karena keterbatasan perangkat keras. Titik-titik ini, dikenal sebagai piksel, merepresentasikan citra dua dimensi tersebut(Hartono et al., 2020).

2.5 Orange Data Mining Tools

Alat penambangan data yang dipakai untuk mendukung penelitian ini adalah *Orange. Orange* adalah perangkat lunak yang dirancang untuk analisis *data* dan eksplorasi *visual* terbuka untuk melakukan pengolahan *data mining.* Perangkat lunak *Orange* dilengkapi dengan berbagai *widget* yang berperan sebagai unit komputasi untuk membaca, memproses, memvisualisasikan, menganalisis, dan mengeksplorasi *data,* serta melakukan fungsi lainnya. *Widget-widget* ini disusun sedemikian rupa sehingga membentuk alur kerja *(workflow)* dan dapat berinteraksi satu sama lain di lingkungan *Orange.* Salah satu jenis *widget* yang disediakan adalah *widget data,* yang memungkinkan *Orange* untuk memanipulasi *data* teks atau gambar. Ketika melakukan analisis gambar dengan *Orange,* diperlukan tambahan *add-ons* yang disebut *image analytics.* Dengan menggunakan *image analytics, Orange* dapat mengubah *data* gambar menjadi representasi vektor menggunakan *deep neural network* yang telah dilatih dengan banyak gambar. Hal ini menghasilkan *data* yang dapat diproses dan memungkinkan penerapan *Machine learning* pada gambar.(Hartono et al., 2020).

2.6 Machine learning

Machine learning adalah teknologi yang dirancang untuk belajar secara mandiri tanpa arahan langsung dari pengguna. Pembelajaran mesin dibangun berdasarkan konsep disiplin ilmu lain seperti statistika, matematika, dan *data mining,* yang memungkinkan mesin untuk belajar dengan menganalisis data.

Dalam konteks ini, *Machine learning* memiliki kemampuan untuk mengumpulkan dan menganalisis data tanpa perintah langsung dari pengguna. Selain itu, *Machine learning* dapat mempelajari data yang telah ada serta data baru yang diperoleh, sehingga dapat melakukan berbagai tugas tertentu tergantung pada apa yang telah dipelajarinya. Istilah *"Machine learning"* pertama kali diperkenalkan oleh ilmuwan matematika seperti Adrien Marie Legendre, Thomas Bayes, dan Andrey Markov pada tahun 1920-an dengan membahas dasar-dasar dan konsep-konsepnya. Sejak saat itu, pengembangan *Machine learning* terus berkembang pesat. Salah satu contoh penerapan *Machine learning* yang terkenal adalah *Deep Blue* yang dikembangkan oleh *IBM* pada tahun 1996. *Deep Blue* dirancang agar dapat belajar dan bermain catur. Mesin ini telah diuji dengan bermain melawan juara catur profesional, dan berhasil memenangkan pertandingan catur tersebut.(Herdiana, 2022)

Peran *Machine learning* sangat membantu manusia di berbagai bidang, dan saat ini penerapannya mudah ditemukan dalam kehidupan sehari-hari. Contohnya adalah fitur *face unlock* yang memungkinkan pengguna untuk membuka perangkat *smartphone* mereka, atau ketika menjelajah di internet atau media sosial dan sering ditemui dengan iklan yang disesuaikan dengan minat pribadi. Ada banyak contoh penerapan *Machine learning* yang sering ditemui dalam kehidupan sehari-hari. Bagaimana *Machine learning* dapat belajar? *Machine learning* dapat belajar dan menganalisis data berdasarkan informasi yang diberikan pada awal pengembangan dan data yang diakses ketika *Machine learning* sudah digunakan. *Machine learning* akan mengikuti

teknik atau metode yang telah diterapkan selama pengembangan. Ada beberapa teknik yang digunakan dalam *Machine learning*, tetapi secara umum terdapat dua teknik dasar belajar, yaitu *supervised* dan *unsupervised*..

a. *Supervised Learning*

   Supervised learning merupakan teknik dalam pembelajaran mesin yang mampu menggunakan informasi yang tersedia dalam data dengan label yang telah ditentukan. Teknik ini bertujuan untuk memberikan target terhadap hasil keluaran dengan membandingkannya dengan pengalaman belajar di masa sebelumnya..

b. *Unsupervisd Learning*

   Unsupervised learning merupakan teknik dalam *Machine learning* yang dapat diterapkan pada data yang tidak memiliki informasi yang dapat diberikan secara langsung.

2.7 K-Nearest Neighbors

K-Nearest Neighbors (KNN) termasuk dalam kategori pembelajaran berbasis instansi. Algoritma ini juga merupakan salah satu teknik pembelajaran malas *(lazy learning). KNN* bekerja dengan mencari kelompok k objek dalam data pelatihan yang paling dekat atau mirip dengan objek yang ada dalam data baru atau pengujian. Algoritma *K-Nearest* Neighbors adalah suatu pendekatan untuk mengklasifikasikan objek berdasarkan data pelatihan yang memiliki jarak terdekat dengan objek yang sedang dianalisis.(Rofiq et al., 2020).

Algoritma *K-Nearest Neighbors* Operasinya terdiri dari memeriksa jarak antara data latihan yang telah ada dan set k tetangga terdekat dalam data pengujian yang baru. Saat data baru masuk untuk diklasifikasikan ke dalam kategori yang belum diketahui, penentuan kategori dilakukan dengan membandingkan karakteristiknya dengan contoh-contoh lain dalam data pengujian yang telah ada. Fitur-fitur dari data yang akan diklasifikasikan diekstraksi dan dibandingkan dengan fitur-fitur dari setiap data dalam kategori yang sudah dikenal dalam data pengujian. Selanjutnya, k tetangga terdekat dipilih dari data pengujian untuk menentukan kategori yang paling umum di antara mereka. Gabungan nilai-nilai ini, berdasarkan pada prosedur waktu yang diharapkan, digunakan sebagai perkiraan nilai di masa depan. Prediksi dengan metode *K-Nearest Neighbors* bergantung pada pola urutan yang diamati dari waktu ke waktu. Jika pola perilaku sebelumnya dapat diidentifikasi yang mirip dengan pola perilaku dari deret waktu saat ini, nilai-nilai berikutnya dapat diestimasi berdasarkan nilai sebelumnya untuk memprediksi pola atau nilai perilaku pada masa mendatang.(Desfiani et al., 2021).

Penentuan nilai *k* dalam algoritma klasifikasi *K-Nearest Neighbors* dapat dilakukan berdasarkan nilai dari k tetangga terdekat, yaitu $k1$, $k2$, … , ks. Semakin besar jumlah data yang ada, semakin kecil nilai k yang biasanya dipilih, namun jika dimensi data lebih besar, maka nilai *k* yang dipilih harus lebih tinggi. Saat menentukan nilai *k*, disarankan untuk menggunakan angka ganjil seperti $k = 1, 3, 5, \ldots$, dst. Nilai *k* harus mematuhi syarat bahwa *k* harus

kurang dari jumlah *N*, di mana *N* adalah jumlah dari *dataset* latih. Hal ini karena nilai *k* digunakan untuk mencari mayoritas kelas/*label* pada data latih, sehingga nilai *k* tidak boleh melebihi jumlah *dataset* latih. Dalam mencari tetangga terdekat atau jarak pada algoritma *K-NN*, terdapat lima metode yang umum digunakan, yaitu Jarak *Euclidean*, Jarak *Manhattan*, Jarak *Cosine*, Jarak Korelasi, dan Jarak *Hamming*. Jarak antara dua tetangga *k* terdekat berdasarkan nilai kemiripan dapat dihitung menggunakan jarak *Euclidean*, yang didefinisikan sebagai berikut:

$$Dist\,(X, Y) = \sqrt{\sum_{i=1}^{D}(X_i - Y_i)^2} \qquad (2.1)$$

dengan:

$Dist\,(X, Y)$: jarak antar objek (*Euclidean Distancing*)

Xi : *sampel data*

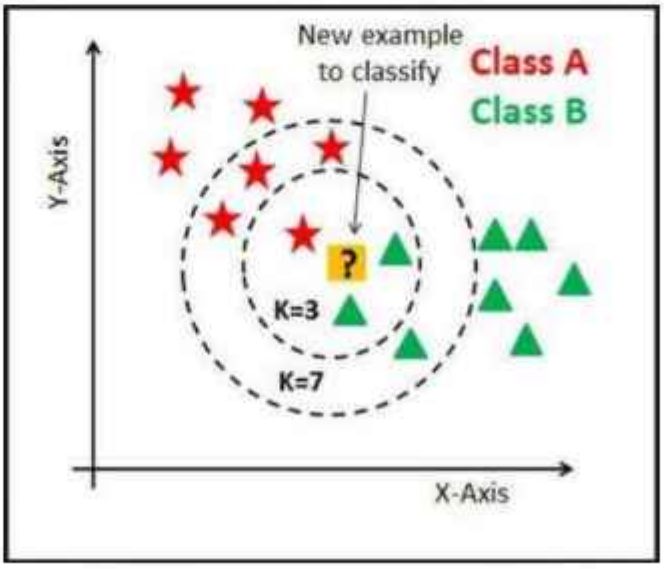
Yi : data uji

D : dimensi data

i : *variabel data*

Kinerja dari masing-masing model *K-Nearest Neighbors* dapat dievaluasi melalui optimasi parameter, seperti yang dilakukan dengan metode *cross validation*. *Cross validation* adalah teknik yang digunakan untuk memvalidasi kinerja dan akurasi suatu model. *K-Nearest Neighbors* memiliki keunggulan dan kelemahan sebagai berikut:

1. Kelebihan *K-Nearest Neighbors*
    a. Sederhana untuk diterapkan
    b. Tidak diperlukan pembangunan model, penyesuaian parameter yang rumit, atau pembuatan asumsi tambahan
    c. Kuat terhadap *training* data yang *noise*
    d. Efektif Jika ukuran data pelatihannya besar
2. Kekurangan *K-Nearest Neighbors*
    a. *K-Nearest Neighbors* perlu Memilih parameter *k* (jumlah tetangga terdekat) yang optimal.
    b. Pembelajaran jarak menjadi fokus. Tetapi kejelasan tentang penggunaan jarak dan atribut dalam mencapai hasil terbaik masih belum pasti.
    c. Biaya cukup tinggi karena diunakan untuk menghitung jarak setiap sampel uji keseluruhan sampel *training*.



Gambar 2. 1 Contoh Algoritma *KNN*

2.8 K-Fold Cross Validation

Mengevaluasi model *machine learning* dapat menjadi tantangan yang signifikan. Secara umum, kita membagi *dataset* menjadi *set* pelatihan dan pengujian. Kemudian, kita menggunakan *set* pelatihan untuk melatih model dan *set* pengujian untuk menguji model. Evaluasi kinerja model dilakukan berdasarkan matriks kesalahan untuk menilai akurasi model. Namun, metode ini tidak selalu dapat diandalkan karena akurasi yang diperoleh untuk satu set pengujian bisa sangat berbeda dengan akurasi yang diperoleh untuk set pengujian yang lainnya. *K-fold Cross Validation (CV*) memberikan solusi untuk masalah ini dengan membagi data menjadi *fold* dan memastikan bahwa setiap *fold* digunakan sebagai set pengujian di beberapa titik *CV*.(Peryanto et al., 2020)

Cross validation adalah metode pengujian standar yang digunakan untuk memperkirakan kesalahan. Prosesnya melibatkan pembagian data latihan secara acak menjadi beberapa bagian dengan perbandingan yang seragam. Kemudian, kesalahan dihitung untuk setiap bagian secara terpisah, dan akhirnya, rata-rata kesalahan dari seluruh bagian dihitung untuk mendapatkan tingkat kesalahan secara keseluruhan.(Sumarlin & Anggraini, 2018).